Shahid Rajaee
Teacher Training University

Journal of
Electrical and Computer Engineering Innovations (JECEI)

Vol. 11 No. 1 Winter-Spring 2023

Journal of Electrical and Computer Engineering Innovations Vol. 11 No. 1 Winter-Spring 2023 JECEI

Shahid Rajaee
Teacher Training University

pISSN 2322-3952
eISSN 2345-3044

Journal of
Electrical and Computer Engineering Innovations (JECEI)

Semiannual Publication

Volume 11, Issue 1, Winter-Spring 2023

Journal of Electrical and Computer Engineering Innovations
http://jecei.sru.ac.ir
EISSN: 2345-3044 ISSN: 2322-3952
JECEI
SRTTU
JECEI

License Holder: Shahid Rajaee Teacher Training University (SRTTU)

Address: Lavizan, 16788-15811, Tehran, Iran.

Journal of Electrical and Computer Engineering Innovations

Vol. 11; Issue 1: 2023

Contents

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

A Multi-Aspect Semi-Automated Service Identification Method

S. Hekmat[1], S. Parsa[2,], B. Vaziri[1]*

[1]*Department of Computer Engineering, Central Tehran Branch, Islamic Azad University, Tehran, Iran.*

[2]*Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran.*

Article Info

Article History:
Received 05 December 2021
Reviewed 15 January 2022
Revised 17 February 2022
Accepted 25 April 2022

Keywords:
Service identification
Business process model
Process mining
Service-oriented architecture
Goal model
Data model

*Corresponding Author's Email Address: parsa@iust.ac.ir

Abstract

Background and Objectives: Several service identification methods have been proposed to identify services using business process-based strategy. However, these methods are still not accurate enough and adequately automated and thus need improvements. The present study addresses this gap by proposing a new semi-automated combinational method that applies process mining techniques and simultaneously considers different aspects of the business domain (e.g., goal and data). We argue that this method facilitates service identification more comprehensively and accurately and helps enhance organizational performance and lower cost structure.

Methods: Our method includes three Phases. In the first phase, the system log is inspected, and the running business process is extracted using process mining techniques. After validating this model, we create a goal and data model in the next phase. In the third phase, we establish connections between the introduced models by defining some matrices. These connections are of two types: structural and conceptual. Finally, we propose a couple of algorithms that lead to the identification of services.

Results: We evaluate the utility of our proposed method by conducting a case study and using the experts' opinions from different perspectives as follows: (1) assessing the accuracy and reusability of the identified services, (2) appraising the efficiency of employing this method in more complex processes, (3) calculating the cohesion to coupling ratio, and (4) assessing the performance of the method and other service quality measures. The results indicate that the average accuracy of this method is about 12 % higher than the previously identified methods for both simple and complex processes. Additionally, it empirically proves that using the process mining techniques improves the service identification considerably (8%). Moreover, according to the experts' opinions, the combination of goal and data model and process mining has increased the performance by 8%. In comparison, cohesion to coupling ratio demonstrated a 7% increase compared to other methods. In sum, we conclude that this method is an advanced and reliable way of service identification regardless of the process size and the complexity.

Conclusion: The findings reveal that considering different aspects of business processes together and using process mining techniques improves the ratio of cohesion to coupling and accuracy of the identified services. Adherence to this approach enables companies to mine their business processes, modify them, and quickly identify services with higher performance. Besides, using this method provides a semi-automated and more effective way of service identification.

Introduction

In each organization, several business processes exist that need to be reviewed and changed continuously to comply with the organization's goals [1], [2]. Using the service-oriented architecture (SOA) is very helpful for the

required changes due to creating reusable and easy-to-change business services.

The first step in moving toward SOA is service identification. There are different strategies for this purpose, such as business processes analysis. Many methods have adopted this strategy and identified services by decomposing business process models defined by experts in the design phase. Recognizing potential services in the strategy relies on understanding and identifying relations between activities, especially those with loose coupling and high cohesion dependencies in the business process models [3]-[5]. Using models in the strategy has caused simplicity and understandability. Besides, the output of the identified services will address the business needs [6]. Nevertheless, service identification based on the strategy in the run-time and dynamic environment might face some challenges.

First, the major problem with relying on predefined BPMs to identify services is that such models are not always available, especially for legacy software systems [7]. In addition, there are some differences in most companies between the process models that are supposed to be run and the one that actually runs [8]. This discrepancy is because some necessary tasks, internal control flows, and dependencies may not be included in the business processes models defined in the design phase. Therefore, if the future services are identified based on the decomposition of the predefined processes and not the ones that really run in systems, the identified services may not meet the business requirements. Besides, some tasks may never be considered as a potential service and are ignored.

Second, if the company's business requirements are changed, some parts of the business process may be affected and need to be altered. Therefore, having deep knowledge and understanding of the ongoing processes seems necessary to facilitate the changing process [9]. However, most companies do not have such sufficient knowledge about running processes or, if they have, it is not documented sufficiently or at least not presented explicitly. Hence, the changing process would be complicated, and following that may lead to some problems in finding new suitable services.

Third, suppose services are only identified based on the decomposition of process models without paying enough attention to other aspects of the business domain (e.g., business goals and data). In that case, the final identified services may not have enough precision and not be applicable.

Finally, the conceptual relations are the other subject that has not been considered sufficiently compared to structural ones in the strategy. Paying attention to such connections between tasks, goals, and data models and having data-aware business processes can lead to finding or defining new services [10]-[11].

All above mentioned challenges can effect on accuracy of the Method.

One solution to tackle these issues is combining process mining concepts and business process-driven strategy as a service identification method. In this regard, some open questions need to be addressed [9]. One of these questions is how to bundle the activities discovered from process mining as a potential service. The other one is recognizing reusable services that can be used out of context [9]. This study presents a new semi-automated service identification method by combining process mining techniques, goal model, and data model to answer these questions and the previously mentioned challenges. To this end, we choose data and goal models because they play a vital role in every process such that other aspects of a business domain would be meaningless without considering them. The proposed method includes three following phases:

In the first phase, the running business process and related knowledge are extracted from the information systems' event log. This phase, in turn, has three steps: (1) identifying workflow (without data), which is modeled by Petri-net, (2) validating the model by conformance checking techniques, and (3) enriching the model with finding related data to obtain a Business Process Model and Notation (BPMN) with the data. These steps are performed using process mining techniques and algorithms. These techniques help organizations to have a comprehensive view of the current business processes and associated activities [12]. Also, they provide various monitoring and improvement analysis tools. Furthermore, they help companies reengineer their processes to gain better performance, identify potential activities to be a service, change or reuse them with reasonable cost, and have the processes that address their business needs.

Phase 2 is dedicated to creating the goal and data model that helps extract the relations between tasks and identifying reusable services.

Finally, in the third phase, the mentioned models are linked by introducing different matrices. These matrices play a pivotal role in answering the mentioned challenges and service identification. Also, they help recognize semantic and structural relationships between tasks to identify other potential services. In the following, services are identified by presenting two algorithms. These algorithms use the mentioned matrices to find high cohesion and loose coupling tasks in their structural or conceptual dependencies, targeted common goals, or operating on mutual data. These tasks are assigned to different services based on their features. By following the above phases, the introduced setup questions are

addressed. All the phases are covered in a case study, followed by listing the services and their assigned methods.

The remaining of this paper is structured as follows: Section 2 presents a quick background related to the research area. Section 3 describes details of our proposed approach through a case study. Finally, Section 4 is dedicated to the evaluation and conclusion. In the evaluation section, our method is compared with state-of-the-art service identification baselines based on the criteria introduced by some previous studies [7], [13], [14].

Related Work

Several service identification methods with a different strategy, input, or output have been presented in recent years [12]-[14]. Because of the diversity and multiplicity of these methods, our paper has limitations to inspect all of them. Therefore, only well-known methods that use business process identification strategies are discussed in this section.

Wang et al. [15] introduced an approach for service identification by making rules to design relations between business activities and potential services. They also designed ports, messages, and interfaces for identified services by considering both business and technical aspects. They presented a new algorithm that decomposes the BPMN model. However, their approach is not automated and has not been validated by a case study. Also, it has ignored other aspects of task dependency, such as the conceptual one.

Inaganti et al. [16] used some guidelines and technology to identify services by decomposing strategy. However, they did not introduce their measurement for recognizing potential business activities. Besides, there is no validating case study in their work. Indeed, their research work focuses on bid business-level services.

Jamshidi et al. [17] presented a cluster-based method that combines business entity strategies with business process decomposition. After modeling the business process, they identified and categorized the service model elements. The method was evaluated based on its use, users' analysis, and comparing with other existing methods. It is noteworthy that automation is not considered in this approach.

Dwivedi et al. [18] introduced a method that uses a heuristics algorithm for service identification. These researchers used the UML diagram as a business process model and applied it in a real example and model-driven development. Detailed information about the heuristics is not provided in this paper.

Bianchini et al. [19] used BP decomposition and ontology to recognize potential services and investigated process annotation semantically. Also, these researchers considered cohesion and coupling to determine services. Finally, they validated their method using a case study without considering automation.

BPA Onto SOA is the name of a method introduced by Yousef et al. [20]. Business process analysis and considering functional and non-functional business needs are the basis of this method. Also, Ontology, BP decomposition, and clustering techniques are applied to identify services; however, automation is missing in this method.

Azevedo et al. [21] chose candidate services using semantic point of view and heuristics algorithms. But detailed information and automation were not enough taken to account.

In [22], a clustering-based method is introduced. It considers services as many activities with high cohesion and loose coupling in their functionality. In this paper, BP decomposition and clustering techniques and algorithms are used. However, it does not consider other aspects of the business domain.

In [23], a method was proposed based on clustering and BP decomposition.

Moreover, in [20], 2PSIM was presented in which partitioning algorithms are applied to the BP model to identify services [24].

Kazemi et al. [25] introduced a method in the scope of automatic service identification using decomposition of business processes. But, it does not consider a different relation between tasks such as conceptual ones. In [26], the authors exhibited an automated model-driven service identification approach. This method uses the business model as an input and, after running heuristic algorithms, gives the service model as an output. AMSI is another automatic model-driven service identification method that identifies services by applying a multi-objective evolutionary algorithm. A high-level BP model is the input of this method [27].

El Amine et al. presented a method for service identification that uses both BP decomposition strategy and particle swarm optimization (PSO) algorithm [28]. A hybrid PSO algorithm was also applied in [29] for service identification. This algorithm addresses service design principles such as reusability, granularity, high cohesion, and loose coupling.

Alwis et al. [30] presented a heuristic algorithm for service identification based on business objects and their relationships. This method is considered a semi-automated approach that employs business processes. However, they did not assess the accuracy of their method, cohesion, and coupling metrics.

These authors also proposed another semi-automated heuristic algorithm that splits tasks into different categories based on their service identification functionality. This approach has a good performance for very large enterprises.

Nevertheless, due to the complexity of the method, it is not successful in addressing the needs of small to medium enterprises (SMEs) [31].

In [32], the authors provided a customized heuristic algorithm for service identification based on analyzing the business process codes. To this end, they developed a tool that converts a business model to a service model. However, this approach does not pay sufficient attention to service measurements like cohesion and coupling.

In [33], the authors provided an automatic tool that maps the BPMN component to software code for service identification. These methods only consider structural relations between tasks and ignore conceptual ones.

In [34], the authors proposed a framework that decomposes process and related logs with clustering algorithms. Then, an expert should modify the recognized services based on service features. This approach primarily depends on the process code and experts' opinions.

In [35], Leopold et al. introduced a new automated identification method that extracts a list of service candidates, including microservice, composite, and in hierarchy services. This method did not provide any solution for identifying the relationship between tasks.

Taibi et al. [36] proposed a method for service identification using business process flow extraction and clustering. After finding business process flow, they applied clustering to find the services. This approach does not consider system goals and data to find relations between tasks.

Lshob et al. [37] proposed a framework that produces a SOA model from the business process model. This approach employs some interfaces to convert a business process-based information system to an SOA-based one. Business requirements and the relation between activities are the missing aspects of this framework.

All the above methods employ a business process decomposition strategy to identify services. These process models are supposed to be run in the organization. But, in the run time and dynamic environments, most of them do not accurately and sufficiently address business needs [38]-[39]. Also, they may not show all tasks and related semantic and structural relationships that run in running time. Besides, since most of these methods do not have automated solutions to identify services, they are often descriptive and not sufficiently accurate.

Our literature review reveals that using running processes and taking other business processes into account can be a powerful technique for a more accurate service identification. However, the studies conducted in this regard are still scarce.

Proposed Approach

In this section, the proposed method is presented in three phases.

In the first phase, the ongoing business process model is extracted from the event log using process mining techniques and tools. After validating this model, it is applied as one of the inputs for the service identification phase.

In the second phase, the goal and data model are defined. These models help identify business and entity services. Moreover, using them allows finding conceptual relations and potentially reusable services.

Finally, in the third phase, services are identified by introducing some matrices and proposed algorithms. These matrices make a connection between BP, goal, and data model. Also, structure and semantic relations between tasks are identified using them. In this phase, the tasks that address the common goals or use the mutual data are considered the potential services method because of their internal conceptual relations.

Having real data-aware business processes, considering conceptual and structural interconnection between tasks (with the help of goal and data models and presenting semi-automated methods for service identification), leads to more accurate services.

Phase 1: Process Mining

Discovering an accurate business process model by process mining techniques needs different activities. They include collecting event logs and converting them to an acceptable format for process mining tools. Besides, applying discovery techniques and validating the output model is necessary in this regard [8]. In this phase, we investigate the mentioned activities.

In general, event logs generated by ERP systems, workflow management systems, or other information systems can be used as a primary input for process mining [8]. However, according to [40], they should be converted to the most common mining formats, i.e., extensible Markup Language (MXML) and extensible Event Stream (XES). There are different tools and methods for this purpose [8], [9], [40].

Since this paper focuses on service identification methods with the help of process mining techniques, we adopt a valid real-life event log format that is ready to use and exists in the process mining website[1]. This event log is related to the fine management process. The system's duty is to support and manage driving fines.

It contains activities, timestamps, and data related to the activities. We use the Prom tools for processing because it supports different process mining algorithms and related plug-ins [41].

[1] www.data.4tu.nl

Prom is one of the most popular and comprehensive tools that researchers use. The Manual and tool base/automatic part of this phase is shown in Table 1.

Table 1: Manual vs. automated part of Phase 1

steps	User(manual)	Automated/tool support
1	Import the log to Prom for process	Processes the log to find activities and data in the output
2	Import the step 1 output to the mining algorithm	Processes the inputs, create process flow
3	Import both previous outputs to conformance checking algorithm	Processes both inputs, inspects and assesses the validation of the model
4	Applying data-aware algorithm	Generates process model with assigned data

Step 1: Finding Log Information

In this step, we inspect the log to discover activities and their assigned data using Prom tools. The results show that the event log contains 150,370 events. There are at least two and at most 20 trails of events in each set of events (called case). Also, the log demonstrates 11 different activities that happen in one process. These activities are created fine, send fine, add a penalty, fine notification, payment, credit collection, appeal to the prefecture, appeal to judge, send and receive appeal and notify result. Fig. 1 shows the extracted information about the log.

Step 2: Mining Process

In this step, we try to find the process flow in the event log by applying the inductive miner algorithm, which is one of the best techniques for finding business processes [42], [43]. There are different mining algorithms in the Prom. However, studies show that the Inductive miner produces more accurate and comprehensive results than others in dealing with noises and making models [42], [44]. The output is presented as a Petri net shown in Fig. 2.

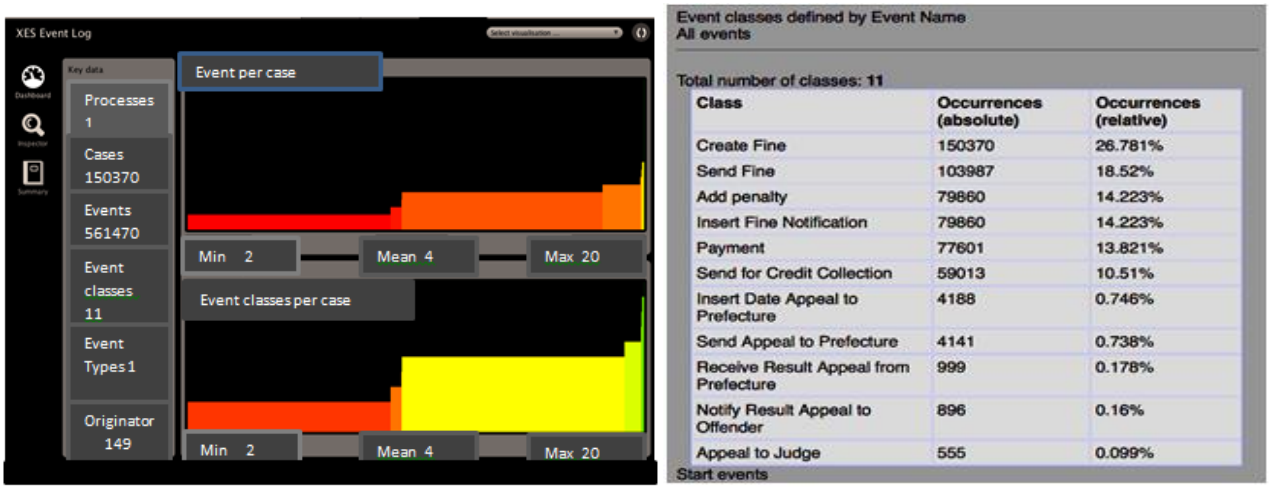


Fig. 1: Log inspection.

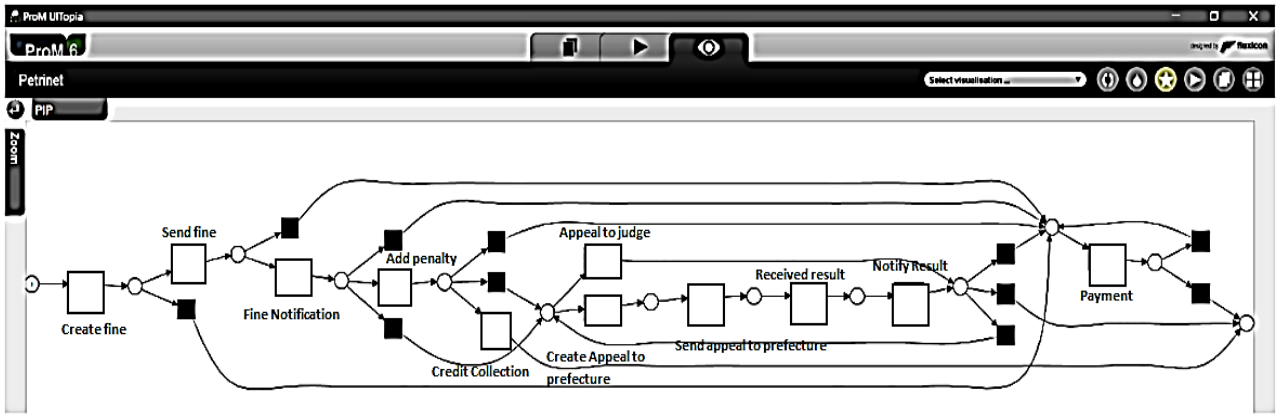


Fig. 2: The workflow created by the inductive miner algorithm.

Some black squares are placed parallel to other events in the figure, meaning that this event can be avoided, and another route can be taken. However, this rule does not apply to 'create a fine' and 'add the penalty' events. The output model needs to be checked for complying with the event log. This procedure, called conformance checking, can be performed using different algorithms and tools [45]. We use the "reply a log on Petri-net for conformance analysis" algorithms and plug-in, which is simple and the most accurate algorithms in Prom tools for this. Purpose [46] Conformance checking tries to answer the question of "what is the probability of re-establishing this process with this event log?". Fig. 3 presents a 91% fitness for this process, suggesting the validity of the extracted model.

Step 3: Data-Aware Process Mining

In this step, the log and the validated primary Petri net are imported to the Prom for detecting data. The output Data Petri net is shown in Fig. 4. In the resulting Data Petri net, each task is presented by a transition. Variables demonstrate the data associated with each transition. Here, transitions perform CRUD operations on data variables and identified data present attributes of the four main entities which flow in the process. For the sake of simplicity, this paper considers the entity instead of the attributes. The entities and the corresponding data models are shown in Phase 2 (Step 2).

Step 4: Mining BPMN

With the previous step's output and utilizing converting data Petri net to BPMN plug-in in Prom tool, the final BPMN model with data is mined. To check the validity of the outcome, we use BPMN to Petri net converter plug-in in Prom to see whether it generates the same data Petri net from obtained BPMN or not. The results analysis reveals that the BP model completely represents the same data as Petri net. If the model needs to be re-engineered for any reason (e.g., having better performance), it can be done in this step. For transforming As-is to the To-Be model, following the best practices proposed in [47] will be helpful. Following these rules, we create To-Be Models and the final output is demonstrated in Fig. 5.

Many plug-ins and algorithms in Prom directly produce the BPMN model from the event log [44], [48]. However, their output just shows workflows without presenting data, data streams, and their effect on decision gates. Having such a process model does not address the questions of this paper.

In this respect, by following our solution to reach the data-aware BPMN model, not only the accuracy and validation of the model in each step are checked but also data entities play a key role in discovering a more accurate BPMN model. Moreover, the Create, Read, Update and Delete operation (CRUD) on each entity is recognizable, which will be used in the service identification phase to find conceptual relations between tasks and entity services. It will be done by defining the entity-task relation matrix.

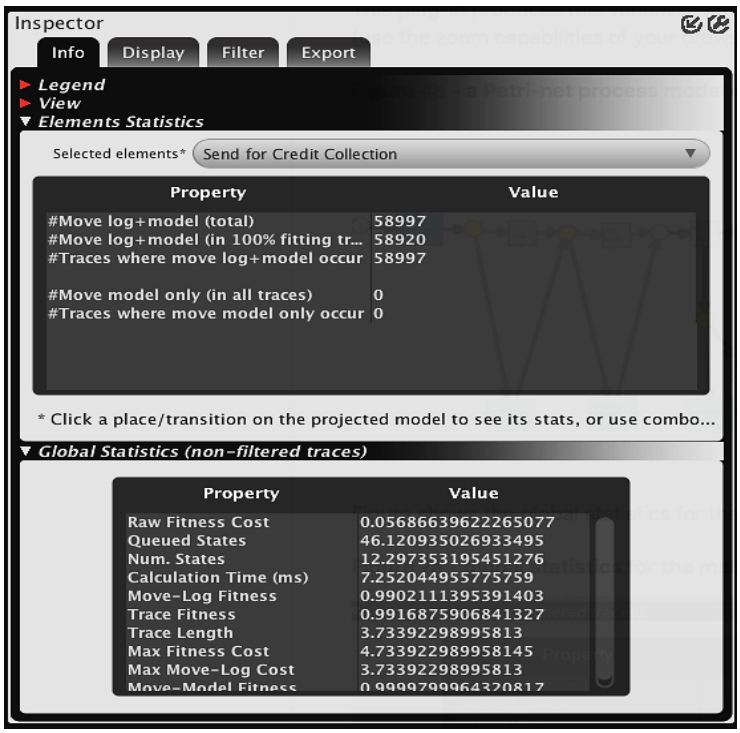


Fig. 3: Conformance checking by Prom.

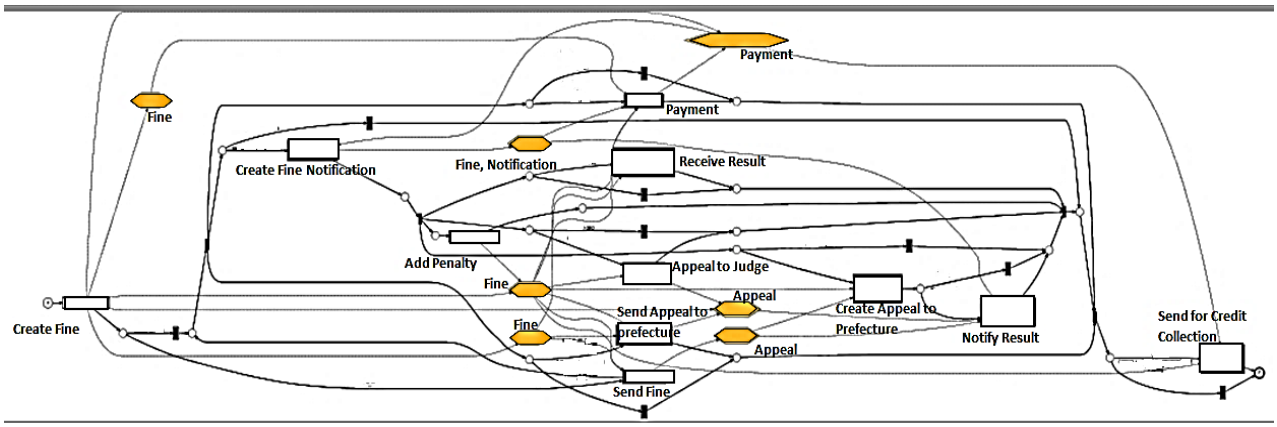


Fig. 4: Petri net with data.

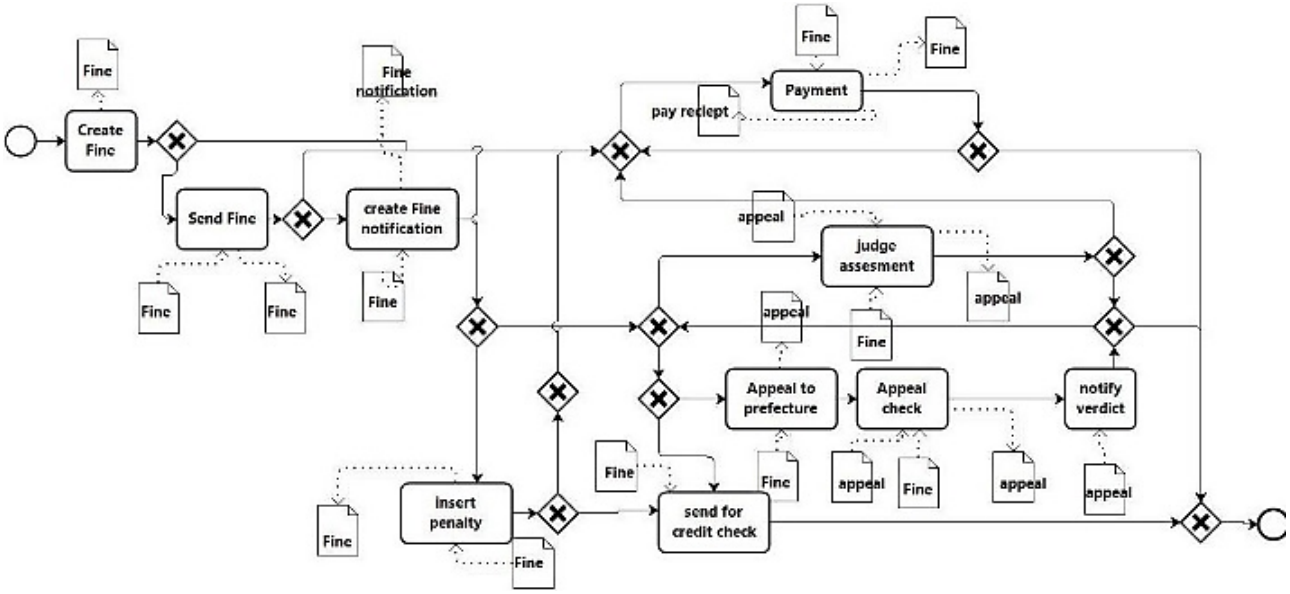


Fig. 5: The Business process model after applying process mining.

Phase 2: Modeling Goal and Data

This phase contains two steps: creating a goal and a data model. Table 2 presents the automated vs. manual parts of this phase.

Table 2: Manual vs. Automate parts of Phase 2

steps	Manual (user)	Automate
1	Create Goal Model	
2	Assess the output of the previous phase to create a data model	

Step 1: The Goal Model

In this phase, a goal model is created for the case study. Using this model helps find business requirements. Then, in the service identification phase, we will establish a relationship between the BPMN model's task and related requirements in the goal model by introducing the task-requirement matrix.

Such a relationship helps us to find the tasks with reusability features. The tasks addressing the same goals can be reused in similar service-oriented systems. There are different methods to creating the goal model [49]-[52]. We adopt GBRAM [50] in this research regarding its simplicity. In this model, only one identifier is considered for each requirement, and other characteristics of each need are omitted. The goal model in this method has different levels. The general goals are located at the highest level. These goals are decomposed into smaller ones to extract the requirements.

It is necessary to mention that if the goal model is documented and exists in the organization, it can be used as a basic model in this step, and there is no need to be recreated.

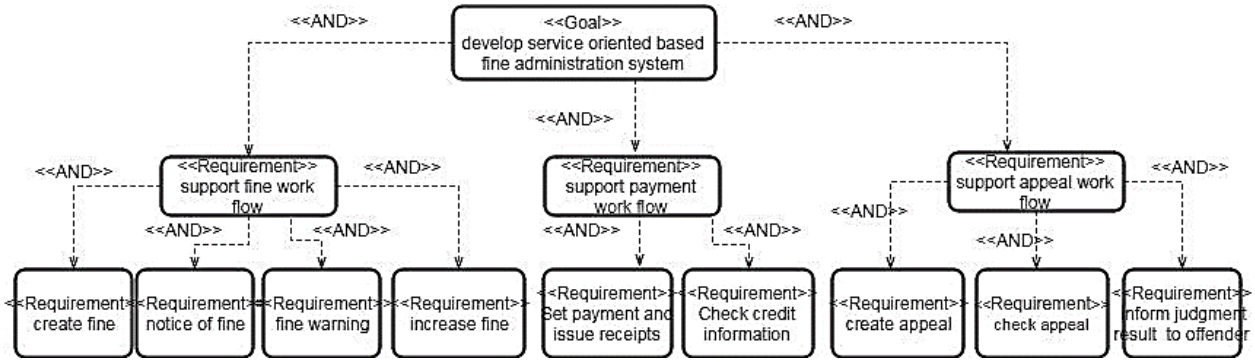


Fig. 6: Goal model.

Step 2: Data Model

In this step, we create the data model by utilizing data that flow in the system. As mentioned in the last step of Phase 1, this model helps find entity services. The data model for the case study is depicted in Fig. 7.

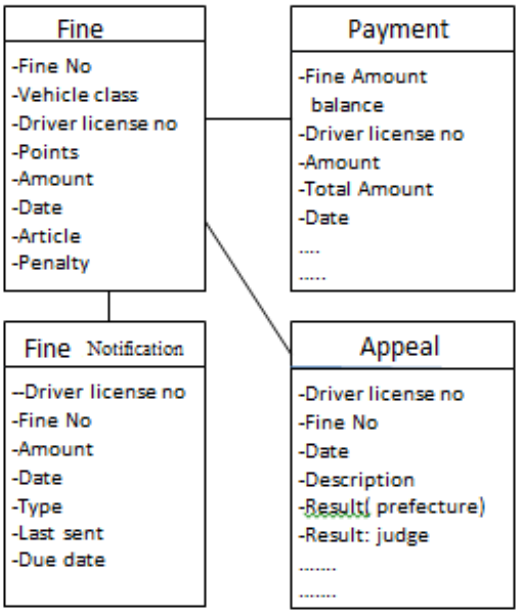


Fig. 7: Data model.

Phase 3: Service Identification

In this phase, services are identified based on data and goal models in the BPMN. Combining these models and using process mining techniques helps find both semantic and structural relations between tasks. In this paper, semantic relation is defined when either one or both of the following conditions occur: 1) two or more tasks either directly or indirectly (collaboratively) address the common goals and 2) when tasks do the CRUD operation on common entities. Considering these conceptual interrelations between tasks helps have highly coherent methods and more accurate services. Automated and manual parts of this phase are depicted in Table 3.

Table 3: Manual vs. Automated part of Phase 3

Steps	Manual (user)	Automate
1	Can be created by user or Tools (Visual Paradigm)	
2	Import required input (BPMN)	Process input, find relations Generate task-entity matrix Generate First relational matrix
3	Import requires input	Process input, find relations Generate a second relational matrix
4		Generate final relational matrix
5, 6	Import required inputs/Analyze the results	Finding services Merging services

Step 1: Create Requirements Task Array

In this step, a requirement- tasks two-dimensional array is created by adopting the goal model (Phase 2) and the BPMN model (Phase 1). This array shows the relationship between each task and related business needs.

To fill out the cells of this array, for each requirement set existing in the last level of the goal model, we investigate the tasks on the BPMN model; if the tasks support the needs, tasks are written in the related cell the array.

Having this array helps find reusable tasks. If organizations have the same goals, the tasks that address these goals can be reusable.

Therefore, identified services with this attitude meet the reusability factor, which is one of the service design principles. We use this array later to identify candidate services.

This matrix is indicated in Table 4.

Table 4: Requirements task array

Requirements	Create fine	Notice of fine	Fine warning	Increase fine	Set payment	Check credit	Create appeal	Check appeal	Inform verdict to the offender
Tasks	t1	t3	t4	t5	t2	t10	t6, t9	t7	t8

Step 2: Create First Task-Task Matrix

This step is divided into two-part. First, the relationship between tasks and related data in the BPMN model is investigated. For this purpose, we explore the data and associated tasks in the BPMN model and determine the type of operation that each task does on the related entities. As mentioned before, each task can perform different CRUD operations on entities. Then, an entity-task matrix is created. The matrix rows are tasks, and the columns are entities.

The cells show each task's type of operation on the related entities.

To quantify the relationship rate between tasks and the relevant entities, we need to convert their degree of relativity to a numerical value. For this purpose, we adopt this concept from [26]. Accordingly, it is assumed that CRUD operations have different degrees of importance. Their strength order is as C>U>D>R and their values (between 0 and 1) are determined as Create(c)=1, Update (U)=0.75, Delete= (D)=0.5 and Read(R)=0.25. As shown, the highest degree of importance is attributed to Create, and the lowest to Read [26]. The entity-task matrix for our case study is presented in Table 5.

Table 5: The task entity matrix

E1: Fine, E2: payment, E3: Notification, E4: Appeal

	E1	E2	E3	E4
t1	C			
t2	U	C		
t3	U			
t4	R		C	
t5	U			
t6	R			C
t7	R			R
t8				R
t9	R			U
t10	R			

In our case study, we consider each entity as a service to find entity services. Since each service comprises some methods, to determine them, in the task-entity matrix (Table 5), we look for the CRUD operations presented in the column associated with that entity. Entity services are derived from a business data model and can be reused to automate different business processes. Table 6 shows the entity services and related methods:

Table 6: Identified entity services

Methods	Entity services
Update (), Read (), Create ()	Fine
Create ()	Payment
Create ()	Notification
Create (), Read(), Update()	Appeal

In the second part, using the matrix shown in Table 5, the following algorithm created the first task-task matrix (Fig. 8).

Algorithm 1

1. **foreach** pair of tasks(i,j) in To-Be Model
2. sum = 0
3. **foreach** Entity e in Entity Set
4. **if** $[\text{Task-Entity}]_{i,e}$ & $[\text{Task-Entity}]_{j,e}$ have value
5. sum+ =1/2(Value$[\text{Task-entity}]_{i,e}$ + Value$[\text{Task-Entity}]_{j,e}$)
6. **end if**
7. **end for**
8. sum=sum*(count(shared entities)/count(available entities))
9. $[\text{TT-First}]_{i,j} = [\text{TT-First}]_{j,i} = \text{sum}$
10. **end for**

Fig. 8: The first task-task matrix algorithm.

This matrix illustrates the relations between tasks based on their access to entities. Thus, it helps determine semantic relations between tasks. To develop this matrix, for every pair of tasks in the business process model, if they operate on one or some similar entities, the average of their accessing types to each common entity will be calculated. Afterward, these values will be added together. In the next step, the number of shared entities by these tasks is divided by the total number of entities accessed by these tasks. The result is multiplied by the value calculated in the previous step. This output is placed in the cell, i.e., at the intersection of the two tasks in the matrix. Applying this algorithm has two benefits. First, it helps understand tasks' relationships considering their impact on shared entities. Second, it allows finding the hidden and semantic relations between tasks. For example, if task x and task y impact similar data, they are conceptually related.

Applying the above algorithm and identifying such relations lead to identifying tasks that can be potentially considered for service detection (Table 7).

Step 3: Finding Structural Relation by second task-task matrix

In this step, the second task-task matrix is generated to illuminate the structural relations between tasks in the business process model.

Some analysis and design tools such as Visual Paradigm can automatically generate this matrix using the BPMN model; however, the matrix created by such tools has some shortcomings. For instance, it only considers the direct relationship between two tasks. Also, if a gate or event exists between them, they will not be considered as related tasks. Additionally, this tool neglects some series of patterns in business processes, such as tasks placed before or after a gateway that can be considered as services. We develop new tools that get the XML version of the BPMN model as an input to tackle these problems and analyze them. Then, it tries to find and quantify the structural relation between tasks by applying the following algorithm. In our proposed algorithm, if the connection between the two tasks of Ti and Tj is direct in the business process model, the value of the cell (Ti, Tj) in the second task-task matrix will be equal to 1. There should be a gateway between Ti and Tj, 0.5 will be assigned as the value. The value will be set to 0.25 if two gate ways are placed in a row and between the tasks. The value will be equal to 0.75 under the condition that Ti and Tj appear on branches positioned before or after a gateway (Fig. 9).

Algorithm 2

```
foreach pair of tasks in To-Be model like Ti and Tj
    if there is a direct edge between Ti and Tj
        [TT_second]i,j = 1
    else if there is a gateway between Ti and Tj
        [TT_Second]i,j = 0.5
    else if there are two gateways between Ti and Tj
        [TT_Second]i,j =0.25
    else if Ti and Tj are tasks in branches after/before a gateway
        [TT_Second] i,j = 0.75
    end if
endfor
```

Fig. 9: Quantify second task-task matrix procedure.

Table 7: First task-task matrix

	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
t1	0	0.437	0.875	0.312	0.875	0.312	0.312	0	0.312	0.625
t2	0.473	0	0.375	0.166	0.375	0.166	0.166	0	0.166	0.25
t3	0.875	0.375	0	0.25	0.75	0.25	0.25	0	0.25	0.5
t4	0.312	0.166	0.25	0	0.25	0.083	0.083	0	0.083	0.125
t5	0.875	0.375	0.75	0.25	0	0.25	0.25	0	0.25	0.5
t6	0.312	0.166	0.25	0.083	0.25	0	1.125	0.312	1.125	0.125
t7	0.312	0.166	0.25	0.083	0.25	1.125	0	0.825	0.75	0.25
t8	0	0	0	0	0	0.312	0.125	0	0.25	0
t9	0.312	0.166	0.25	0.083	0.25	1.125	0.75	0.25	0	0.125
t10	0.625	0.25	0.5	0.125	0.5	0.125	0.25	0	0.125	0

Table 8: Second task-task matrix

	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
t1	0	0.25	0.5	0	0	0	0	0	0	0
t2	0	0.25	0	0	0	0	0	0	0	0
t3	0	0.25	0	0.5	0	0	0	0	0	0
t4	0	0.25	0	0	0.5	0.25	0	0	0.25	0.25
t5	0	0.25	0	0	0	0.25	0	0	0.25	0.5
t6	0	0	0	0	0	0	1	0	0.75	0
t7	0	0	0	0	0	0	0	1	0	0
t8	0	0	0	0	0	0	0	0	0.25	0.25
t9	0	0.5	0	0	0	0.75	0	0	0	0.25
t10	0	0	0	0	0	0	0	0	0	0

Table 9: Final task-task matrix

	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
t1	0	0.723	1.375	0.312	0.875	0.312	0.312	0	0.312	0.625
t2	0.473	0.25	0.375	0.166	0.375	0.166	0.166	0	0.166	0.25
t3	0.875	0.625	0	0.75	0.75	0.25	0.25	0	0.25	0.5
t4	0.312	0.416	0.25	0	0.75	0.333	0.083	0	0.083	0.375
t5	0.875	0.625	0.75	0.25	0	0.5	0.25	0	0.5	1
t6	0.312	0.166	0.25	0.083	0.25	0	2.125	0.312	1.875	0.125
t7	0.312	0.166	0.25	0.083	0.25	1.125	0	1.125	0.75	0.25
t8	0	0	0	0	0	0.312	0.125	0	0.5	0.25
t9	0.132	0.666	0.25	0.083	0.25	1.875	0.75	0.25	0	0.375
t10	0.625	0.25	0.25	0.125	0.5	0.125	0.25	0	0.125	0

The output matrices generated by the suggested algorithms facilitate the identification of structural relations between tasks. Thus, the challenge of insufficient attention to the structural relationships between tasks in other service identification methods is overcome by creating this matrix. The second task-task matrix is presented in Table 8.

Step 4: Final task-task Matrices

The final matrix is equal to the sum of the first and the second communication task matrices, which shows the relationship between tasks from both structural and conceptual points of view. Therefore, each cell's value in the first and second communication matrix is summed and written in the corresponding cell in the final task-task matrix. Table 9.

Step 5: Recognizing candidate services

In this part, we create a service based on the business needs identified in Step 1 of Phase 1. For each business requirement, a new service is defined. Then, related methods for each service are determined by applying an algorithm presented in Fig. 10. According to this algorithm, the requirement-task array is first reviewed, and the tasks that address each business need are recognized and assigned to the related services. Suppose the algorithm faces some tasks previously assigned to another service. The first service with those tasks will be the owner of such methods in this situation. However, they are called in when their functions are needed to complete another service. Second, in this step, the algorithm reviews the final task-task matrix to find the remaining method for each service that is not assigned to any services.

Moreover, the algorithm looking for the tasks that may not be directly aligned with any business requirements, but are semantically related to the other methods, is identified for a particular service. This semantic relationship can be of two types: implicit control flow or data flow, both prerequisites for executing the identified service. Then, each task that remains unassigned to any service is allocated to a new service. Finally, the algorithm elucidates the relationships between the identified task services and other types of services such as entity and utility services. The candidate services for our example are shown in Table 10.

Table 10: Primary services for the case study

Methods	Services	
t1	Create fine	S1
t3	Notice of fine	S2
t5	Increase fine	S3
t4	Fine warning	S4
t2	Payment	S5
t10	Check credit	S6
t6	Appeal to prefecture	S7
t9	Appeal to judge	S8
t7	Appeal check	S9
t8	inform result	S10

After recognizing primary services, we investigate whether merging services is possible or not. To do so, we define variable RD as a ratio of dependency between services. It shows the ratio of the average internal correlation between each service's methods to the average connection degree between the methods of different services (1).

In some business processes, like our case study, only one task is suitable to consider as a method of the identified services. In such a circumstance, having fewer services with more abilities would be helpful to decrease the costs.

$$\mathrm{RD} = \text{average cohesion}/\text{average coupling} \quad (1)$$

Variables of the above relation are defined in (2) and (3).

$$\text{Average Cohesion} = \frac{\sum \text{Cohesion}(S_i)}{N} \quad (2)$$

where N is the total number of services; and

$$\text{Cohesion}(S) = \begin{cases} 1 & |S| = 1 \\ \sum_{i,j} \text{task} - \text{task}(t_i, t_j)\ \forall t_i, t_j \in S & |S| > 1 \end{cases}$$

*If the service has one method, its cohesion is equal to 1. If the number of service methods is greater than 1, the formula calculates the value of cohesion.

and

$$\text{Average Coupling}(S) = \frac{\sum_{i,j} \text{Coupling}(S_i, S_j)}{D} \quad (3)$$

where D is the total number of connections between services; and

$$\text{Coupling}(S1, S2) = \sum_{i,j} \text{task} - \text{task}(t_i, t_j)\ \forall t_i \in S1, t_j \in S2$$

Using above relations, we present a complementary algorithm (Fig. 11) to refine identified services based on these relations.

This algorithm needs a service dependency matrix as an input. The rows and columns of this matrix are the services. The matrix's main diagonal indicates the cohesion of methods within each service that can be calculated by cohesion(s) relation, which exists in (2). Other cells of this matrix are filled by the coupling variable's value, which is shown in relation (3). For our case study, this matrix is depicted in Table 11.

Table 11: Dependency matrix

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	1	1.375	0.875	0.312	0.723	0.625	0.312	0.312	0.312	0
S2	0.875	1	0.75	0.75	0.625	0.5	0.25	0.25	0.25	0
S3	0.875	0.75	1	0.25	0.625	1	0.5	0.5	0.25	0
S4	0.312	0.25	0.75	1	0.416	0.375	0.333	0.083	0.083	0
S5	0.437	0.375	0.375	0.166	1	0.25	0.166	0.166	0.166	0
S6	0.625	0.5	0.5	0.125	0.25	1	0.125	0.125	0.25	0
S7	0.312	0.25	0.25	0.083	0.166	0.125	1	1.875	2.125	0.312
S8	0.312	0.25	0.25	0.083	0.666	0.375	1.875	1	0.75	0.25
S9	0.312	0.25	0.25	0.083	0.166	0.25	1.125	0.75	1	1.125
S10	0	0	0	0	0	0.25	0.312	0.5	0.125	1

After creating the dependency matrix, in this step, the value of the RD variable is calculated for the services using relation (1). If this ratio increases. Two services will be merged, and the set of services will be updated. The other situation is depicted in the algorithm

The final services for our case study are shown in Table 12.

Considering the definition of utility services from the identified services, each one can handle automatic tasks can be considered utility services.

For example, t3 can be considered a notification service, and when merged with t1, they can be identified as a composite service.

Table 12: Final identified services

services	Methods
S1	t1, t3
S2	t4
S3	t5
S4	t2
S5	t10
S6	t6, t7, t8, t9

Input: Task-Requirement matrix, Task-Task matrix
Output: the set M of candidate services

```
1.  Begin
2.  M=0 // the set of candidate services
3.  T=0 // the set of business process tasks
4.  flag=0; //temporary variable
5.  Used[T,S]=0; //matrix that shows each task belongs to which service
6.  foreach (requirement R ∈ requirement set)
7.      Create a new service Si and add Si to M
8.      foreach task tj in T that supports requirement R //check from Task-Requirement
    matrix
9.          for (k=0; K<M; K++)
10.             if (Used[tj, Sk] = 1){
11.                 Connect Si to Sk in service model;
12.                 Flag=1;
13.                 Break;
14.             end if
15.         end for
16.         if (flag=0)
17.             Add tj to Si;
18.             Set Used[tj, Si] to 1;
19              Remove tj from T;
20.         end if
21.         foreach task tP in task-task matrix
22.             if (task-task[tP , tj] has value and tP doesn't support any R)
23.                 Add tP to Si;
24.             end if;
25.             foreach simple tasks not yet assigned to any service
26.                 Create a new service Sn and add Sn to M;
27.             end for
28.         end for
29.     end for
30. end for
31. foreach task service St that performs CRUD operation on entity e //check from task-entity
    matrix
32.     Find entity service Se in entity service model that performs CRUD operation on
    entity e
33.     Connect service St to Se in service model
34. end for
35. foreach task service St that use utility U
36.     Find utility service SU in Utility service model that performs Utility function U
37.     Connect service St to SU in service model
38. end for
        return M;
39. end
```

Fig. 10: Service identification algorithm.

Refine Services

Input: Services dependency matrix (SD)
Output: the refined set M of candidate services

```
1.  begin
2.  Compute RD; (From (1))
3.  Count=0.
4.  New service set=M; //buffer
5.  foreach row in SD matrix correspond to Si service
6.      Cohesion [Si] = SD [Si, Si];
7.      foreach column in SD matrix correspond to Sj service
8.          Coupling [Si, Sj] = SD [Si, Sj]
9.          If Coupling [Si, Sj]> Cohesion [Si]
10.             Aggregate Si and Sj into a new service Sij
11.             Compute the AverageCohesion/AverageCoupling ratio RD2;
12.             If (RD2>RD)
13.                 Add Sij to new service set;
14.                 Remove Si and Sj from new service set;
15.                 M= new service set;
16.                 RD= RD2;
17.             end if
18.         end if
19      end for
20.  end for
21.  end
```

Fig. 11: Aggregation algorithm.

Results and Discussion

To evaluate our method, we sought to reach four different data. First, we needed to assess the accuracy of identified services, i.e., the number of services identified correctly based on their correct assigned tasks to the total number of services. Also, this kind of service should highly comply with business needs. Second, it was necessary to know how the complexity of the process can affect our method's accuracy. Here, complexity included both process size and the number of gateways. Then, we needed to compare our proposed method with other service identification methods in case of common service quality metrics. Finally, we needed to examine the performance of the proposed method compared to other methods. In this paper, performance could be defined as the ability of the method to provide services at an acceptable level of accuracy and time. The ratio of accuracy to consumed time was calculated, and the results were expressed as a percentage in Table 14

In this study, ten different processes were given to five experts two times and asked to identify services based on their selected method. The processes had various degrees of complexity and sizes. On the first try, we gave them predefined business processes, and on the second try, we presented them with ongoing business processes resulting from process mining techniques.

Ultimately, we compare their results with the results arising from our proposed method. Average results are shown in Figs. 12, 13, and 14. As shown in the figures, the results of our method are more accurate than others (≈12%). and this superiority does not depend on the size or degree of complexity of the processes. Also, the results keep superiority when the basis process results from process mining techniques, and it shows about 8% improvement. In the next comparison, we compared our combination method with each goal, data, and business process-driven method separately as the basis of the proposed method. Next, we calculated the RD ratio for these service identification approaches on different processes. Simultaneously, we asked three experts to apply each approach to different given processes and rate them between 1 and 10 based on the quality of identified service and related assigned tasks. If they faced each misaligning task with a service identified by each method, they had to deduct one mark from their score. Finally, the average of scores was calculated for each approach. The result showed that our method's cohesion to coupling ratio was higher than others (≈7%). Moreover, results showed that the identified services were more compliant with the business needs whenever the base business process applied process mining techniques. Table 13 demonstrates the results.

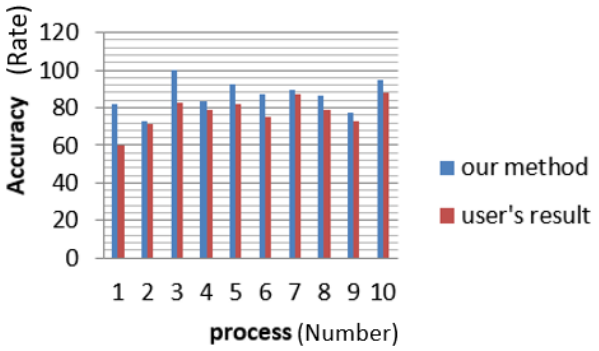


Fig. 12: Investigate the accuracy of the proposed method.

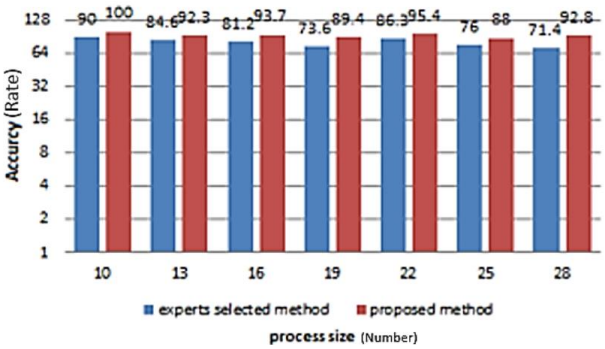


Fig. 13: Investigate accuracy and process size.

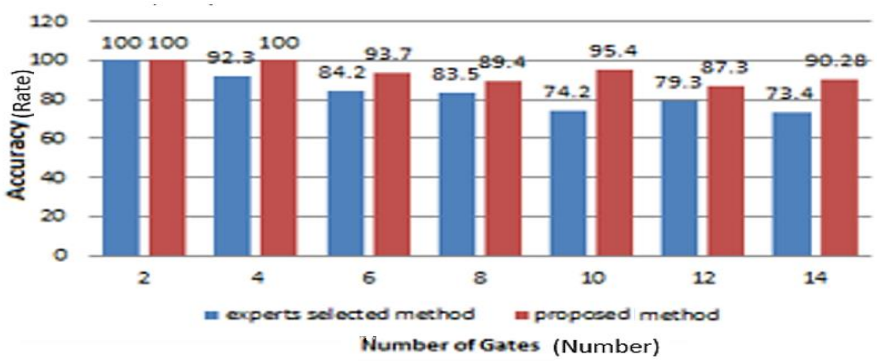


Fig. 14: Investigate accuracy and complexity.

Table 13: Comparing methods based on RD variable (PM: Process mining)

Method /Service	Business processes without PM	Goals	Entities	Business processes with PM	Proposed method
RD variable	2.146	1.092	1.726	2.274	2.389
Experts' evaluation	7.23	5.73	6.92	7.78	8.5

Table 14: Performance comparison between methods

Accuracy to time Ratio Average of Performance ≈ %	Business processes without PM	Goals	Entities	Business processes with PM	Proposed method
Process size <10	95.6%	92.24%	94.69%	97.47%	98.54%
Process size >10	77.25%	72.47%	76.33%	78.69%	79.96%
Number of gates <10	86.59%	83.59%	85.76%	88.36%	91.02%
Number of gates >10	76.59%	69.98%	74.49%	77.81%	82.67%

As mentioned earlier, in the Next step, the performance of our method has been calculated and compared with other methods. Table 14 shows that when the complexity of the process increases, the proposed method keeps the superiority over others to identify acceptable services in a more efficient time. In total, the average performance for the proposed method is almost 86% which is about 8% more than others.

In Table 15, our method is compared with other related methods from different perspectives, such as service guideline features, semantic and structural behavior, and automation.

As this table shows most of the service quality metrics have been considered in our method in comparison with others.

Table 15: Compare proposed method with others

	Strategy of identification	Identification technology	Type of service	Input	Standard s of model	Attention to structural behavior	Attention to semantic behavior	Attention to data structure	Level of automation	Cohesion	Coupling	Reusability	Granulity level
[21] Azevedo	Process Decomposition	Heuristic	Task, data, Composite	Process model	EPC FAD	✓	✓	✓	Non automative	-	-	✓	✓
[26] Jamshidi	Entities	Clustering Matrix	Software	CRUD process	-	✓	✓	✓	Automative	✓	✓	✓	✓
[53] Fareghzadeh	Process Decomposition, goal, Use-case, Legacy data	Analysis	Task, data, Composite, software	Use case, legacy system	UML	✓	✓	✓	Non-automative	✓	✓	✓	✓
[54] Kim	Goals and Scenario	Instruction	Software	Goal model	-	-	✓	-	Non-automative	✓	-	✓	-
[16] Inaganti	Process Decomposition	Instruction	Task	Process, Organization data model,	-	✓	-	✓	Non-automative	✓	-	✓	✓
[17] Jamshidi	Entities	Algorithm	Data, collaboration	Process	UML	✓	✓	✓	Semi-automative	✓	✓	✓	✓
[55] Chang	Process Decomposition	Analysis	Task	Process	UML	-	✓	-	Non-automative	-	-	✓	-
[56] Levi	Process Decomposition and Goals	Analysis	Task	Process, Goals	UML	✓	✓	-	Non-automative	-	✓	✓	✓
[57] Kazemi	Process Decomposition	Genetic algorithm	Task	Process	BPMN	✓	✓	-	Automative	✓	✓	-	✓
[58] Amiri	Process Decomposition	Heuristic-GA	Software	Process and goals	BPMN			✓	Semi-automative	✓		✓	✓
[30] Alwis	Process Decomposition	Heuristic	Task Composite Software	Process model	BPMN	✓	✓	✓	Non automative	-	-	✓	✓
[31] Alwis	Process functionality Decomposition	Clustering Heuristic	Software Task software	process	BPMN	✓			Automative			✓	✓
[32] Eric	Process Decomposition, Legacy code	Heuristic Algorithm	Composite, software	legacy system code	BPMN	✓	-	-	Semi-automative	-	-	✓	-
[33] Zafar	Process decomposition legacy code	Instruction	Software	legacy code	BPMN	✓	-	-	Semi-automative	-	-	✓	-
[34] Giseli	Process Decomposition	Clustering, Mining log	Task	Process,	BPMN	✓	-	✓	Non-automative	-	-	✓	✓
[35] Leopold	Process Decomposition	Algorithm	Task, software	Process, Legacy code	BPMN	✓	-	✓	automative		✓	✓	-
[36] Taibi	Process Decomposition	Clusteing, Mining process flow	Task, software	Process flow,	BPMN	✓	-	-	Semi-automated	-	-	✓	✓
[37] Leshob	Process Decomposition, legacy code	instruction	Task	Process, Legacy code	BPMN	✓	✓	-	automative	-	-	✓	✓
[59] Al Shereiqi	Process Decomposition	Clustering, Process mining	Task	Process	BPMN	✓			Non-automative	✓	✓	-	✓
Proposed method	Process Decomposition, Goals and entities	Algorithm, process mining	Task data utility, software	Goals, Process, Data	BPMN	✓	✓	✓	Semi-automative	✓	✓	✓	✓

Conclusion

This paper offers a new combined semi-automated method for service identification. The method considers different aspects of the business, including goals, data, and ongoing business processes, to recognize closely relevant tasks as a service. Since the Business process plays a key role in the method, we use process mining techniques to have deep and accurate knowledge about it. In the first phase, using such techniques, we extract the ongoing processes and associated data that flow in the organization from the real-life system event log. Then, we check the fitness of the output model with the verified conformance checking algorithms using Prom tools.

After validating the ongoing data-aware business process model, we create a goal and data model that helps find both structural and conceptual relations between tasks. These relationships' degree of correlation and coupling determines which tasks can be reused or recognized as services.

As a result, the process of moving to SOA will be easier, more flexible, and more accurate.

For this reason, the mentioned relations will be found by defining three matrices and considering different points of view: 1) addressing the same business needs (goal view and conceptual), 2) working on the same entity (data view and conceptual), and 3) structural interrelation between tasks in the business process model. Then, we link these matrices by introducing first, second, and final relational task-task matrices. These matrices' cells show the dependency rate between tasks and use to understand cohesion and coupling relation between them by emphasizing conceptual and structural relations.

In the end, we propose two algorithms for identifying and merging candidate services that use the relational matrices mentioned earlier. These algorithms introduce a variable to measure the internal and external relationship between tasks. It means that if the cohesion value between tasks is more than the value of the coupling relation, the related tasks should be assigned to one service.

In this way, the cost of additional calls between services is decreased. The reusability feature (a crucial quality factor in service-oriented systems) is covered by tasks that address the same business needs. Whenever the business goals are the same in other processes, the related services can be replaced, or if the needs are changed, only its related services will be changed. Besides, working on common entities not only increases the cohesion between tasks but also helps increase the ability to reuse services.

To do all the subjects mentioned above, we received help from experts and different processes.

Our method is evaluated from different points of view. The results indicate the superiority of the proposed method in the case of accuracy, performance, and cohesion to cohesive ratio. Also, other comparing metrics states the same result.

Using the proposed method helps company to have more knowledge about the conceptual and structural relations between the tasks in the business processes. This capability enables them to reengineer these processes and improve them to have more accurate services.

As the method considers multiple aspects of the business domain, it can be extended.

Here, the potential path to extend this method for the future is suggested:

This method can be extended by considering other aspects of the business domain between tasks to reach high-quality services. Also, the method can be extended to cover other phases of service software development. In addition, the automated part of the method can be improved by decreasing human intervention; for example, by integrating the whole process through developing a tool that supports both the mining and identification phases.

Author Contributions

All the Authors contributed to all part of preparing and writing of this paper.

Acknowledgment

The authors would like to thank both the Editor and anonymous reviewers of JECEI for their valuable and constructive Comments and feedbacks.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

BPMN	Business Process Management
SOA	Service Oriented Architecture

References

[1] Y. Baghdadi, "A business model for B2B integration through Web services," in Proc. IEEE International e-Commerce Technology Conf.: 187-194, 2004.

[2] Y. Baghdadi, "Modelling business process with services: towards agile enterprises," Int. J. of Business Inf. Syst., 15(4): 410-433, 2014.

[3] Y. Baghdadi, W. Al-Bulushi, "A guidance process to modernize legacy applications for SOA," Serv. Oriented Comput. App., 9(1): 41-58, 2015.

[4] Q. Gu, P. Lago, "Service identification methods: a systematic literature review," in Proc. European Conference on a Service-Based Internet: 37-50, 2010.

[5] F. Kramer, C. Görling, S. Wind, "Service identification--An explorative evaluation of recent methods," in Proc. IEEE in 2014 47th Hawaii International Conference on System Sciences: 1285-1295, 2014.

[6] J. W. Hubbers, A. Ligthart, L. Terlouw, "Ten ways to identify services," The SOA Magz., (48), 2007.

[7] M. Abdellatif et al., "A taxonomy of service identification approaches for legacy software systems modernization," J. Syst. Soft., 173: 110868, 2021

[8] W. Van Der Aalst, "Process mining: Overview and opportunities," ACM Trans. Mng. Info. Syst. (TMIS), 3(2): 1-17, 2012.

[9] W. Van Der Aalst et al., "Process mining manifesto," in Proc. International Conference on Business Process Management: 169-194, 2011.

[10] M. Reichert, "Process and data: Two sides of the same coin?," in Proc. Springer OTM Confederated International Conferences, On the Move to Meaningful Internet Systems: 2-19, 2012.

[11] R. S. Huergo, P. F. Pires, F. C. Delicato, "A method to identify services using master data and artifact-centric modeling approach," in Proc. 29th Annual ACM Symp on Applied Computing: 1225-1230, 2014.

[12] W. Van Der Aalst, "Service mining: Using process mining to discover, check, and improve service behavior," IEEE Trans. Serv. Comput., 6(4): 525-535, 2012.

[13] B. Bani-Ismail, Y. Baghdadi, "A literature review on service identification challenges in service oriented architecture," in Proc. International Conference on Knowledge Management in Organization(Springer): 203-214, 2018.

[14] S. Cai, Y. Liu, X. Wang, "A survey of service identification strategies," in Proc. 2011 IEEE Asia-Pacific Services Computing Conference: 464-470, 2011.

[15] Z. Wang, X. Xu, D. Zhan, "Normal forms and normalized design method for business service," in Proc. IEEE International Conference on e-Business Engineering (ICEBE'05): 79-86, 2005.

[16] S. Inaganti, G. K. Behara, "Service identification: BPM and SOA handshake," BPTrends, 3: 1-12, 2007.

[17] P. Jamshidi, M. Sharifi, S. Mansour, "To establish enterprise service model from enterprise business model," in Proc. 2008 IEEE International Conference on Services Computing: (1): 93-100, 2008.

[18] V. Dwivedi, N. Kulkarni, "A model driven service identification approach for process centric systems," in Proc. 2008 IEEE Congress on Services Part II (services-2 2008): 65-72, 2008.

[19] D. Bianchini, C. Cappiello, V. De Antonellis, B. Pernici, "P2S: A methodology to enable inter-organizational process design through web services," in Proc. International Conference on Advanced Information Systems Engineering, Springer: 334-348, 2009.

[20] R. Yousef, M. Odeh, D. Coward, A. Sharieh, "BPAOntoSOA: A generic framework to derive software service oriented models from business process architectures," in Proc. Second International Conference on the Applications of Digital Information and Web Technologie (IEEE) : 50-55, 2009.

[21] L. G. Azevedo , "A method for service identification from business process models in a SOA approach," in Entrp. Business-Process and Info. Syst. Modeling. Spr.,: 99-112,2009

[22] Y. Kim, K. G. Doh, "Formal identification of right-grained services for service-oriented modeling," in Proc. International Conference on Web Information Systems Engineering (Springer): 261-273, 2009.

[23] M. Ren, Y. Wang, "Rule based business service identification using UML analysis," in Proc. 2nd IEEE Information Management and Engineering International Conf.,: 199-204, 2010.

[24] A. Nikravesh, F. Shams, S. Farokhi, A. Ghaffari, "2PSIM: two phase service identifying method," in Proc. OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"(Springer): 625-634, 2011.

[25] A. Kazemi, A. Rostampour, A. N. Azizkandi, H. Haghighi, F. Shams, "A metric suite for measuring service modularity," in Proc. CSI international symposium on Computer Science and Software Engineering (CSSE,IEEE): 95-102, 2011.

[26] P. Jamshidi, S. Mansour, K. Sedighiani, S. Jamshidi, F. Shams, "An automated service identification method," Technical report, TR-ASER-2012-01, Automated Software Engineering Research, TR-ASER-2012-01, 2012.

[27] M. Soltani, S. M. Benslimane, "From a high level business process model to service model artifacts-a model-driven approach," in Proc. ICEIS (3): 265-268, 2012.

[28] C. M. El Amine, S. M. Benslimane, "Using combinatorial particle swarm optimization to automatic service identification," in Proc. 13th International Arab Conference on Information Technology ACIT: 17-19, 2013.

[29] M. Mohamed, B. S. Mohamed, M. E. A. Chergui, "A hybrid particle swarm optimization for service identification from business process," in Proc. IEEE Second World Conference on Complex Systems (WCCS): 122-127, 2014.

[30] D. Alwis, A. Anuruddha, A. Barros, C. Fidge, A. Polyvyanyy, "Discovering microservices in enterprise systems using a business object containment heuristic," in OTM Confederated International Conferences, On the Move to Meaningful Internet Systems, Springer, 60-79, 2018.

[31] D. Alwis, A. Anuruddha, A. Barros, C. Fidge, A. Polyvyanyy, C. Fidge, "Function-splitting heuristics for discovery of microservices in enterprise systems," in Proc. International Conference on Service-Oriented Computing: 37-53, 2018.

[32] S. Eric, A. Moreira, C. Faveri, "An approach to align business and IT perspectives during the SOA services identification," in Proc. 2017 17th International Conference on Computational Science and Its Applications (ICCSA, IEEE): 1-7, 2017

[33] I. Zafar, F. Azam, M. W. Anwar, B. Maqbool, W. H. Butt, A. Nazir, "A novel framework to automatically generate executable web services from BPMN models," IEEE Access, 7: 93653–93677, 2019.

[34] M. Gysel, L. Kölbener, W. Giersche, O. Zimmermann, "Service cutter: A systematic approach to service decomposition," in Proc. European Conference on Service-Oriented and Cloud Computing, Springer: 185-200, 2016.

[35] H. Leopold, P. Fabian, M. Jan, "Automatic service derivation from business process model repositories via semantic technology," J. Syst. Software, 108: 134-147, 2015.

[36] D. Taibi, K. Systä, " From monolithic systems to microservices: A decomposition framework based on process mining," in Proc. 8th International Conference on Cloud Computing and Services Science, 2019

[37] A. Leshob, R. Blal. H. Mili, P. Hadaya, O. Hussain "From BPMN models to SoaML models," in Proc. Conference on Complex, Intelligent, and Software Intensive Systems: 123-135,2019

[38] L. Jiang, J. Wang, N. Shah, H. Cai, C. Huang, R. Farmer, "A process-mining-based scenarios generation method for SOA application development," Serv. Oriented Comput. App., 10(3): 303-315, 2016.

[39] I. Zafar, F. Azam, M. W. Anwar, W. H. Butt, B. Maqbool, A. K. Nazir, "Business process models to Web services generation: A systematic literature review," in Proc. IEEE, IEMCON. Conf . on Information Technology, Electronics and Mobile Communication: 789–794, 2018.

[40] O. AlShathry, "Process mining as a business process discovery technique," Compt. Eng. & Info. Tech., 5(1), 2016.

[41] B. F. Van Dongen, A. K. A. de Medeiros, H. Verbeek, A. Weijters, W. M. van Der Aalst, "The ProM framework: A new era in process mining tool support," in Proc. International conference on application and theory of petri nets: 444-454, 2005.

[42] S. J. Leemans, D. Fahland, W. M. Van Der Aalst, "Discovering block-structured process models from event logs-a constructive approach," in Proc. International conference on applications and theory of Petri nets and concurrency: 311-329, 2013.

[43] A. A. Kalenkova, W. M. Van Der Aalst, I. A. Lomazova, V. A. Rubin, "Process mining using BPMN: relating event logs and process models," Soft. Syst. Modeling, 16(4): 1019-1048, 2017.

[44] R. Ghawi, "Process discovery using inductive miner and decomposition," arXiv preprint arXiv:1610.07989, 2016.

[45] J. Carmona, B. van Dongen, A. Solti, M. Weidlich, Conformance Checking, Springer, 2018.

[46] W. Van Der Aalst, A. Adriansyah, B. van Dongen, "Replaying history on process models for conformance checking and performance analysis," Wiley Interdiscip. Rev.: Data Min. Knowl. Discovery, 2(2): 182-192, 2012.

[47] H. A. Reijers, S. L. Mansar, "Best practices in business process redesign: an overview and qualitative evaluation of successful redesign heuristics," Omega, 33(4): 283-306, 2005.

[48] A. A. Kalenkova, M. De Leoni, W. M. Van Der Aalst, "Discovering, analyzing and enhancing BPMN models using ProM," in BPM (Demos): 36, 2014.

[49] A. Van Lamsweerde, R. Darimont, E. Letier, "Managing conflicts in goal-driven requirements engineering," IEEE Trans. Softw. Eng., 24(11): 908-926, 1998.

[50] C. Rolland, C. Souveyet, C. B. Achour, "Guiding goal modeling using scenarios," IEEE Trans. software engineering, 24(12): 1055-1071, 1998.

[51] A. I. Anton, "Goal-based requirements analysis," in Proc. the second international requirements engineering Conf.: 136-144, 1996.

[52] H. Kaiya, H. Horai, M. Saeki, "AGORA: Attributed goal-oriented requirements analysis method," in Proc. IEEE joint international requirements engineering Conf.: 13-22, 2002.

[53] N. Fareghzadeh, "Service identification approach to SOA development," in Proc. World Academy of Science, Engineering and Technology, 35: 258-266, 2008.

[54] S. Kim, M. Kim, S. Park, "Service identification using goal and scenario in service oriented architecture," in Proc. IEEE APSEC'08. 15th Asia-Pacific Software Engineering Conf.: 419-426, 2008.

[55] S. H. Chang, S. D. Kim, "A service-oriented analysis and design approach to developing adaptable services," in Proc. IEEE Services Computing International Conf.: 204-211, 2007.

[56] K. Levi, A. Arsanjani, "A goal-driven approach to enterprise component identification and specification," Commu. ACM, 45(10): 45-52, 2002.

[57] A. Kazemi, A. Rostampour, P. Jamshidi, E. Nazemi, F. Shams, A. N. Azizkandi, "A genetic algorithm based approach to service identification," in Proc. IEEE (SERVICES), World Cong.: 339-346, 2011.

[58] M. J. Amiri, S. Parsa, A. M. Lajevardi, "Multifaceted service identification: process, requirement and data," Comp. Sci. Info. Syst., 13(2): 335-358, 2016.

[59] A. Al Shereiqi, Y. Baghdadi, "Business process mining for service oriented architecture," in ICT for an Inclusive World: (Springer): 3-19, 2020.

Biographies

Shahrzad Hekmat received her B.S. degree from Azad University, Iran, and her M.S. degree from Tehran University, in 2009 and 2011, respectively, both in Computer Engineering. She is currently studying her Ph.D. in the Department of Computer Engineering, Azad University Central Tehran Branch. Her research interests are in the areas of software engineering, service-oriented systems, and process mining.

- Email: shahrzad.hekmat@gmail.com
- ORCID: 0000-0002-3412-4576
- Web of Science Researcher ID: AEF-4134-2022
- Scopus Author ID: 57489672700
- Homepage: NA

Saeed Parsa received his B.Sc. in mathematics and computer science from Sharif University of Technology, Iran, his M.S. degree in computer science from the University of Salford in England, and his Ph.D. in computer science from the University of Salford, England. He is an associate professor of computer science at Iran University of Science and Technology. His research interests include software engineering, soft computing and algorithms.

- Email: parsa@iust.ac.ir
- ORCID: 0000-0003-4381-2773
- Web of Science Researcher ID: S-9536-2018
- Scopus Author ID: 8407441400
- Homepage: http://www.iust.ac.ir/page/717/Saeed-Parsa,-Ph.D.

Babak Vaziri received his B.S. degree in computer engineering from Shahid Beheshti University, Iran, and M.S. and Ph.D. degree in computer engineering from Islamic Azad University, Iran. He is currently an Assistant Professor of Computer Engineering in the Department of Computer Engineering, Central Tehran Branch, Islamic Azad University, Tehran, Iran. His research interests include software and process re-engineering and data mining.

- Email: vaziribabak@gmail.com
- ORCID: 0000-0002-8255-2794
- Web of Science Researcher ID: AGU-4445-2022
- Scopus Author ID: 57190976954
- Homepage: http://faculty.iauctb.ac.ir/b-vaziri-comp/fa

How to cite this paper:
S. Hekmat, S. Parsa, B. Vaziri, "A multi-aspect semi-automated service identification method," J. Electr. Comput. Eng. Innovations, 11(1): 1-20, 2023.

DOI: 10.22061/JECEI.2022.8151.526

URL: https://jecei.sru.ac.ir/article_1703.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Reinforcement Learning-based Load Controller in IP Multimedia Subsystems

M. Khazaei*

Computer Engineering Department, Kermanshah University of Technology, Kermanshah, Iran.

Article Info

Article History:
Received 29 January 2022
Reviewed 07 March 2022
Revised 22 April 2022
Accepted 01 May 2022

Keywords:
IMS
SIP
Overload
Multi-agent system
Reinforcement learning

*Corresponding Author's Email Address: m.khazaei@kut.ac.ir

Abstract

Background and Objectives: IP multimedia subsystems (IMS) have been introduced as the Next Generation Network (NGN) platform while considering Session Initiation Protocol (SIP) as the signaling protocol. SIP lacks a proper overload mechanism. Hence, this challenge causes decline in the multimedia QoS. The main propose of overload control mechanism is to keep the network throughput at the same network capacity with overload.

Methods: NGN distributed with IMS is a complex innovative network consisting of interacting subsystems. Hence, multi-agent systems (MAS) receiving further attention for solving complex problems can solve the problem of overload in these networks. To this end, each IMS server is considered as an intelligent agent that can learn and negotiate with other agents while maintaining autonomy, thus eliminating the overload by communication and knowledge transfer between the agents. In the present research, using MAS and their properties, the intelligent hop by hop method is provided based on Q-learning and negotiation capability for the first time.

Results: In the proposed method, parameters of overload controller are obtained by reinforcement learning. In order to check the validity of controller performance, a comparison is made with the similar method in which the optimal parameters are achieved based on trial and error. The result of the comparison confirms the validity of the proposed method. In order to evaluate the efficiency of the learner method, it is compared with similar and standard methods, for which the results are compared to show performance. The results show, the proposed method has approximately improved the throughput by 13%, the delay by 49% and the number of rejected sessions by 17% compare with methods, passing control messages through the network such as CPU occupancy methods. While compare with external controller methods like holonic, throughput is improved by 1% and the number of rejected requests is decreased by 10%, but delay is increased by 6% due to the convergence time of the learning and negotiation process.

Conclusion: To overcome overload, complex IMS servers are considered as learner and negotiator agents. This is a new method to achieve the required parameters without relying on expert knowledge or person as well as, heterogeneous IMS entities can be inserted into the problem to complete study in future.

Introduction

Packet switching allows service providers to provide multimedia applications to their subscribers. The future approach of the telecommunications industry is towards NGN, in which all types of fixed and mobile access networks are integrated based on the IP platform. NGN can provide multimedia services based on the standard IMS platform. IMS has been used by most mobile

network operators since the third generation. In this regard, the SIP protocol has been adopted by 3GGP as the basic IMS architecture. Most cell phones and wireless devices support SIP as a multimedia session protocol [1]-[3].

SIP lacks a proper overload mechanism despite its positive features such as text-based, IP-based, data-independent, relocation support, and end-to-end properties. It uses the retransmitting mechanism to deal with packet loss when running on the UDP protocol. This is done by keeping different timers for each request to be sent. When the relevant response is not received within a specified time, the requests will be resent, which makes the situation worse in case of overload [4].

When the SIP encounters an overload, the overload control mechanism is activated rejecting the new session request messages by sending a 503 response. In this method, the messages must first be analyzed to generate a rejection response for new requests. The capacity devoted by the server to analyze messages and generate reject responses is wasted. All the server capacity is spent on rejecting repeated requests in addition to the negative effect on the overload. Therefore, it is inevitable to design an efficient overload controller for IMS. The overload controller must be able to gather the information needed to decide about overload. Based on this information, it must determine the appropriate response to the overload and apply it to the network. Depending on the type of overload and the location of the controller, different methods and policies have been proposed to deal with the overload [5].

Due to the complexity and heuristic nature of the overload problem in IMS, approximate mathematical or heuristic methods are used in designing the controller leading to occasionally a huge deal of errors and not acceptable answers. Therefore, in this paper, a new machine learning method is proposed to design an overload controller. In this regard, IMS servers are considered as intelligent learning agents able to negotiate with other network agents. These agents learn the amount of tolerable load by interacting with the dynamic environment and through unsupervised trial and error, and they control the overload in the network by negotiating with other agents.

In the following, the background and related works are given. Then, the proposed method is presented and the performance evaluation and analysis of the results are provided. Finally, deals with conclusions are presented.

Related Backgrounds

Since the papers' objective is to design an overload controller based on intelligent agents in IMS, it is essential to explain the related concepts and works briefly.

A. Multi-Agent Systems (MAS)

An agent is defined as a software or hardware located in the environment and acts autonomously to achieve its goals. The agent perceives the environment through its sensors and affects the environment through its stimulants. Everything around the agent except itself is called the agent environment. Today, the use of MAS to solve complex problems has received a huge deal of attention. MAS is made up of several agents trying to solve problems that are sometimes difficult and sometimes impossible to solve for a centralized and integrated system. On the other hand, communication is an important concept in MAS. Without communication, agents should rely only on decisions based on their observations, while communication enables agents to make more coordinated decisions. Negotiation is the dominant way to reach an agreement without the involvement of others to gain mutual benefit in common areas of interest to agents. A negotiation protocol is a set of rules that all agents know. In negotiation, an agreement is obtained when there is a common ground between the proposals. Agents have a personal preference for the outcome of the negotiation and may also have limitations on the use of the proposals [6], [7].

In learning the agents, there are problems for which scarce or incomplete resources exist for solving, thus, unsupervised learning has been considered. Reinforcement learning is unsupervised learning in which the system tries to optimize interaction with the dynamic environment through trial and error without specifying the action for the agent. Reinforcement learning is indeed to map different situations into actions to get the best results with the most reward. The two characteristics of "trial and error" and "reward with delay" are the most important characteristics of reinforcement learning [8].

Standard reinforcement learning, or Q-learning, is a model-independent method, in which the agent has no access to the transfer model. Q-learning based on Markov's decision-making process, with a delayed reward function, can learn the optimal policy. In this method, the agent estimates the pair (action, state) by continuous interaction with the environment as trial and error. The Q-learning steps are based on Algorithm 1.

The algorithm starts from an initial state and reaches the goal state by performing a series of actions and receiving a reward. In this situation, the agent state is not changed with each action while not receiving any reward from the environment. The action can be selected by exploration and exploitation. Selection of action by exploration means selecting the action randomly regardless of the values in Q-Table. This may discover optimized actions not selected yet and add

them to the table. In the exploitation form, the best action is selected based on Q-Table [9].

Algorithm 1: The Q-learning method algorithm

```
Begin
Q[states][actions]=0;
S=Get_Current_State ();

While (S! = absorption state)
   a= Select_Action ();
   r= Calculate_Reward ();
   s/= Get_New_state ();
   max=For_All_Possible_Action_Find_Max_Q (s/, a) ();
   Q (s, a) =α*(r+max) +(1-α) *Q (s, a);
   Update_Q_Table ();
   S= s/;
End While
End
```

In the Q-learning algorithm, $\alpha \in [0, 1]$ is the agent learning rate. A value of one causes the agent to consider only the most recent information, while zero leads the agent to have no learning. Parameter $\gamma \in [0, 1]$ is called the discount factor. Zero means that the agent only considers the current reward, however, the values close to one cause the agent to wait a long time to reach a higher reward. When all pairs (action, state) are experienced repeatedly while reducing the learning rate over time, Q-learning converges with a probability of one to the optimal value of $Q^*(s,a)$. Some applications of reinforcement learning include continuity of services in IMS [10], [11], unsupervised learning in next-generation networks [12], [13], intelligent transportation systems and urban traffic control [9], [14], [15], management of wireless distributed sensor networks [16]-[18], and complex software systems [19].

B. IP Multimedia Subsystems (IMS)

Unlike traditional applications, the purpose of IMS is to integrate different types of multimedia services and applications and converge between wired, wireless, and mobile networks. Other features of IMS are control of sessions, development of services, Quality of services (QoS), and the possibility of calculating costs under a single standard. IMS is a packet switching network extending over the IP platform and has a three-tier architecture. The users connect to IMS using IP-based networks. The simplified IMS core architecture is shown in Fig. 1 [20], [21].

There are two databases in the IMS architecture including Home Subscriber Server (HSS) and Subscription Locator Function (SLF). HSS is the place for storing subscribers' information and related services, and SLF is used to find subscribers' HSS addresses. ASs provide multimedia value-added services. Call Session Control Functions (CSCF) are SIP proxy servers each with a specific function. Their common role is during the registration, session creation, and routing process.

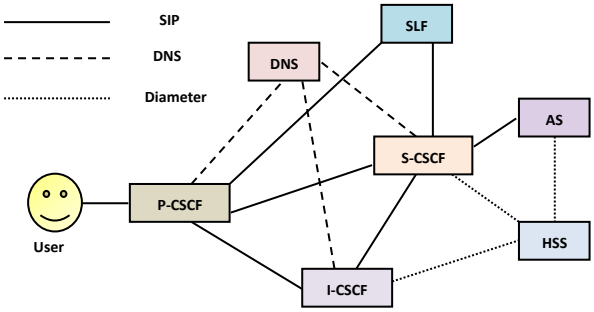


Fig. 1: The IMS core architecture.

The servers can be configured stateful or stateless, depending on the situation and needs. The stateful server stores transaction information, however, in the stateless server, no transaction is created on the server, and the server is solely responsible for receiving and routing messages. P-CSCF is a stateful SIP server. This server is the first point of users' contact with IMS. All user traffic is transferred to this server and also network traffic is transferred to users through this server. S-CSCF is also a stateful SIP server always located on a home network domain. This server is the central point of the IMS responsible for managing the registration process, routing, maintaining session status, and storing service information. All SIP signaling packets pass through the S-CSCF to determine the next action by processing. I-CSCF is a stateless SIP server, contacting point of an operator to connect with subscribers within that operator. I-CSCF allocates an S-CSCF server based on the defined policies when registering by receiving information from the HSS. Another task of this server is to get the next step in routing through HSS as well as directing requests to the assigned S-CSCF or AS.

While receiving the IP, the user receives the P-CSCF address and has to register in the IMS. After registration, the relevant P-CSCF knows the S-CSCF assigned to each user according to the response package. The S-CSCF also knows the P-CSCF to contact to reach the user. This information is used to establish the sessions.

As shown in Fig. 2, when user A wants to make a session with user B through SIP, it sends an Invite to P-CSCF. P-CSCS sends the packet to the S-CSCF assigned to A in the registration process. Based on ID B, S-CSCF finds the corresponding I-CSCF in domain B and delivers the package to it.

By contacting the HSS, the I-CSCF finds the S-CSCF assigned to B and delivers the package to it. The S-CSCF, with the information obtained at the time of registration, delivers the Invite package to the P-CSCF and then to the B. After receiving Invite by B, the response is generated and sent to A of the same route as Invite reached. After exchanging messages, a session is formed between A and B [1].

A SIP transaction is a request and all related responses exchanged between two adjacent SIP entities. Since in most cases the 503 mechanisms embedded in the SIP cannot cope with overload, an overload controller is inevitably required in IMS. The overload controller consists of three main components. The monitoring unit collects information from the specified parameters and provides them to the control function. The control function determines the policy and the amount of load received based on a defined algorithm and provides it to the stimulator unit, which rejects the overload based on the policies received [22], [23].

If all the controller components are on one server, this is called the internal method otherwise, it is termed external control. External controls are divided into two end to end and hop by hop categories.

In the end to end method, the edge server is responsible for regulating the load sent to the overloaded server. The challenge in this method is how to inform the edge server about the server with overload and how the edge server detects the request passing through the overloaded server. In the hop by hop method, two tandem servers determine the amount of load sent from the server upstream to downstream by different policies [24], [25].

In the window policy, the downstream server allows the upstream server to send the request in the specified window size without receiving confirmation. The size of the window on the upstream server can be determined using incoming messages, acknowledgments, 503 messages, timers expire, or calls delay. The overloaded server can also dynamically and continuously estimate its response capacity and notify upstream servers as the number of windows available [24], [26]-[29]. A window-based holonic mechanism (WHOC) is used holonic multi-agent system to control and manage overload in SIP networks.

Based on past observations, the normalized least mean square algorithm is used to estimate each agent window size. The size of the windows is adjusted in the way that no overload will occur in network paths, which could be fulfilled through using holonification properties, negotiation process and communications. WHOC offers an appropriate window size for edge servers to control the load from the beginning of the network and prevent network overload [30].

In the loss-based policy, the downstream server specifies the percentage of reduced load sent by the upstream servers. In this method, the rate of retransmissions can be reduced by modeling the interactions between the downstream and upstream servers as a controller [31]. Also, sometimes the upstream server predicts overload on the downstream server by methods such as 503 received message rates or uses mechanisms such as the leaky-buckets technique [32], [33]. Intelligent methods are provided to monitor the overloaded server and then prevent overload by classifying packets, intelligently deleting repeated packets, controlling active sessions and obtaining thresholds [34], [35].

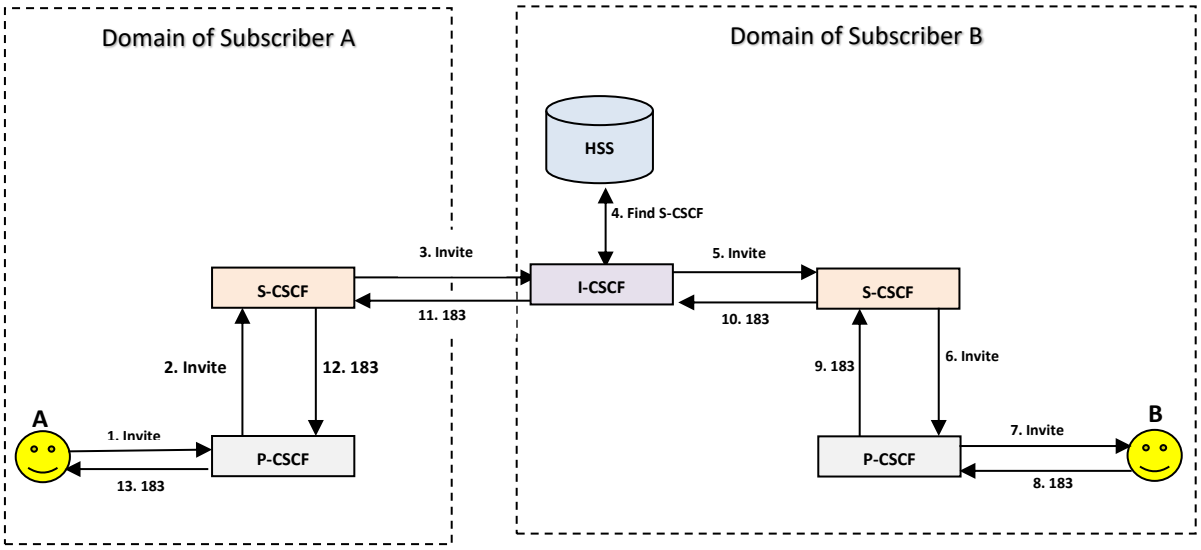


Fig. 2: Establishing a session between two users in two different IMS domains.

In HOC, a loss-based mechanism is implemented to control overload, using multi-agent with holonic organization to implement a user perspective of fairness if possible. HOC uses a greedy method on the network graph to obtain constant holarchy. When a proxy server is overloaded, the sending load is adjusted from the source servers, causing fewer network resources to be involved in the overload. Load fitting is done based on received requests and used as predictor. Therefore, each holon uses the server capacity amount of its offered

load. In HOC, when a holon involved in overload is recognized with a request from the due holon, the overload is most likely solved because servers with high dependency are placed in a holon in terms of load exchange with each other. Complexity decreased by the use of holonic organization and cooperation between holons through the communication, agreement and knowledge that is exchanged among them [36].

In the rate-based policy, the rate of sending the upstream server to the downstream server is limited. Setting a threshold for CPU consumption and sending this value at regular intervals is one of these techniques controlling the rate to enter the requests into the downstream server [5], [24], [37]. In the on-off policy, the downstream server can stop or connect the received load for a while. Determining the time required to process messages within the downstream server queue and announcing the downtime to the upstream server is among the methods of this policy [24].

Finally, a hop by hop method can be developed based on new policies, especially the concept of software networks, Network functions virtualization (NFV), and cloud environments [38]-[40].

Designing Load Controller

In many IMS load control algorithms, there are parameters determining the efficiency of the algorithm. For example, in the proposed algorithms [5] a threshold value is considered for the amount of CPU occupation. Whenever the percentage of CPU occupation exceeds this threshold, the overload control algorithm will operate. Method [28] also determines a delay threshold in the upstream server.

It determines the window size in the upstream server by comparing the delay of the received responses with this threshold. The best and most accurate way to calculate the values of strategic parameters is to use mathematical relations. Nonetheless, mostly due to the complexity or heuristic of the method, it is not possible to calculate the desired values using mathematical equations [41]. Trial and error is another way to determine the values of the parameters and check the value producing the most optimal answer by simulating and placing different values. The problem with this method is that the obtained values are only suitable for that network situation, and by changing the network conditions, the answers are no longer acceptable and the calculations must be performed again. Learning-based methods help in such situations. According to Fig. 2, IMS has three tandem SIP servers (S-CSCF, I-CSCF, S-CSCF) in which the I-CSCS server is stateless not contributing to overload. However, the two tandem servers are stateful and play a role in organizing and managing the sessions. Since the servers connected through one hop, the proposed controller used the hop by hop overload control [42].

As seen in Fig. 3, for each proxy server queue, two warning (T_w) and constraint (T_c) thresholds and thus three regions are defined determining the decision and controller response. If the number of session requests exceeds T_c (constraint region), requests will be rejected locally by messages 503. If the number is between two thresholds (warning region), the negotiation process with the upstream server begins. Nevertheless, if the number of requests is less than T_w, no reaction occurs.

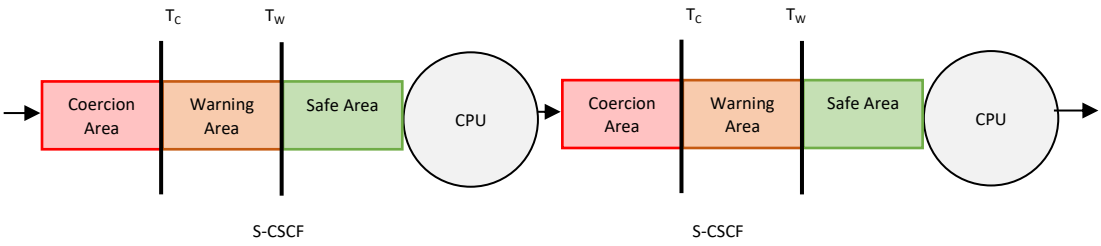


Fig. 3: Two tandem servers in IMS.

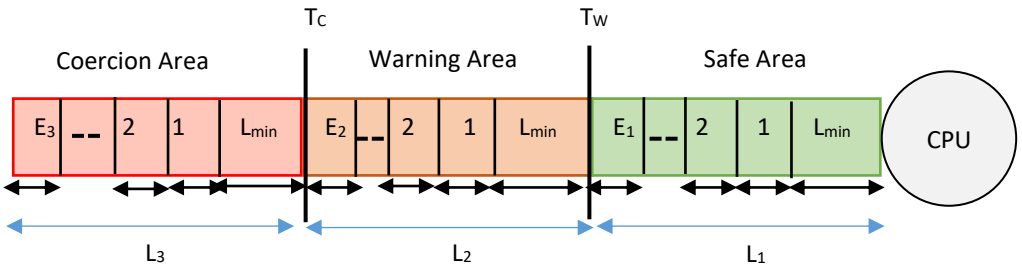


Fig. 4: Assigning states to the agent queue.

In this regard, each proxy server is defined as an intelligent agent with the ability to learn and negotiate. Agents learn the values of T_w and Tc by Q-learning through monitoring their resources and the knowledge from the environment. Q-learning allows the controller to be independent of expert knowledge and prior environmental information. Among the requisites for designing a Q-learning algorithm is the definition of states space, actions space, and appropriate reward function. In the designed controller, the control function implements the learning algorithm, the stimulator unit negotiates with the upstream agent, and the monitoring unit collects information about the queue length and the rate of incoming and outgoing requests. In the proposed controller, the monitoring unit, the control function, and the stimulator unit are located in the agent.

A. Defining Agent Q-Learning Requirements

The actions defined for each agent determine how the constant queue length is divided into each region. To define each action by the agent, a minimum constant length is assigned to each region. Moreover, a certain number of length increases are considered as an extension. The action space is defined by <N_a, L_{min}, N_{ex}, L_{ex}>. N_a indicates the number of regions, L_{min} represents the minimum length assigned to each region, N_{ex} denotes the number of extensions, and L_{ex} indicates the length of each extension. Server queue length (L) is obtained from (1).

$$L = N_a * L_{min} + N_{ex} * L_{ex} \quad (1)$$

The length assigned to each region is obtained from (2). In Fig. 4 the length of each region is equal to the sum of the minimum initial lengths and the number of allocated extensions

$$L_i = L_{min} + e_i * L_{ex} \quad (2)$$

where, e_i is the number of extensions of the i^{th} region. Since the sum of the extensions considered for all 3 regions is constant, the method of allocating the extensions to each region is calculated by (3) [14].

$$\sum_{i=1}^{N_a} e_i = N_{ex} \quad ; \quad e_i \in N \quad (3)$$

The answer of (3) specifies the number of actions of each agent and depends on the value of N_{ex}. As the N_{ex} increases, the number of responses, and consequently the number of actions of each agent increase. With (4), the number of extensions of each region can be controlled by the parameter Θ to reduce the action space to accelerate convergence [14].

$$\sum_{i=1}^{N_a} e_i = N_{ex} \; ; \quad e_i \in N; \quad e_i \leq \theta \quad ; \quad 1 \leq \theta \leq N_{ex} \quad (4)$$

The values of the parameters used to determine the actions space are given in Table 1. According to these values, the set of actions is 19 actions, which are given in Table 2. The actions are selected through an exploration.

To determine the states space, the number of transactions in each region is used. In this regard, the number of transactions in each region is arranged according to their order of entry and each arranged list corresponds to a state. The total number of states Table 3 is equal to the sum (ten states) of the inequality permutations of the number of areas transactions (six states) plus the possibility of equalizing the number of areas transactions (four states). For example, if T_i indicates the number of transactions in the i^{th} area, T_1> T_2> T_3 will represent the highest number of transactions in the safe region and the lowest number of transactions in the constraint region.

Table 1: The parameters used in determining the actions space

Parameters	Volumes
L	110
L_{min}	20
L_{ex}	10
N_{ex}	5
N_a	3
Θ	5
|Actions|	21
|States|	13

Table 2: The values of the length assigned to each region for possible actions

Actions					
Normal area	Warning area	Coercion area	Normal area	Warning area	Coercion area
20	70	20	30	60	20
20	60	30	30	50	30
20	50	40	30	40	40
20	40	50	30	30	50
20	30	60	30	20	60
20	20	70	50	40	20
40	50	20	50	30	30
40	40	30	50	20	40
40	30	40	60	30	20
40	20	50	60	20	30
70	20	20			

Table 3: The different states defined for each agent

States	
Inequality States	Equality States
T_1>T_2>T_3	T_1=T_2>T_3
T_1>T_3>T_2	T_1=T_3>T_2
T_2>T_1>T_3	T_2=T_3>T_1
T_2>T_3>T_1	T_1=T_2=T_3
T_3>T_1>T_2	
T_3>T_2>T_1	

The directing of the agent in the states space and actions to achieve an optimal policy is done by the reward function. The reward function determines how much closer an agent is to a goal through a state. In the proposed learning method, the concept of Goodput is used as a reward function.

The amount of reward received by an agent for acting is proportional to the amount of increase given by Goodput compare with the previous states. Hence, the values in Q-Table are updated with a delay step [35], the reward function for the ith agent is predicted by applying the Normalized Least Mean Square method (NLMS), Table 4.

Table 4: The agents' reward value prediction by NLMS

Number	Equations
1	$R_{n-2} = G_{n-1} - \bar{G}_{n-1}$
2	$\underline{W_n} = \underline{W}_{n-1} + \omega * \frac{R_{n-2} * \underline{G_{n-2}}}{\|\underline{E}_{n-2}\|2}$
3	$\underline{G}_{n-1} = combine G_{n-1} and\ \underline{G}_{n-2}$
4	$\overline{G}_n = \underline{W}_n^T * \underline{G}_{n-1}$

In Table 4, the underlined variables are vectors and the over-lined ones are to hold the predicted values. The **W** is the vector of prediction coefficient filter at size *q* and **G** is a vector to hold the *q* value of process reward. The initial value of **W** is zero which can be updated per each new data. $\overline{G}_n$ is predicted reward of G_n.

Q-Table is used during the learning process to store and update Q-function values. This table is considered as a two-dimensional matrix in which rows and columns specify states and actions, respectively. The initial value of Q-Table is considered zero. After the learning stage, according to the values of Q-Table, the thresholds T_w and T_c of each agent are extracted based on the length of the regions [30], [36].

B. Negotiation Protocol for Implementing the Policy

In negotiation, a set of agents is involved along with a set of variables dependent on agents. Agents negotiate a set of possibilities (values). To reach an agreement, the possibilities are assigned to the variables through negotiation. In the controller, the set of agents participating in the negotiation are the agents in Fig. 3. The variables define the amount of load sent to the downstream agent. The possibilities are also the values suggested by the agents to obtain the amount of load sent or received.

The agents will participate in the negotiation based on a defined strategy presented in Table 5. A suggestion cycle includes the initiator suggestion and the response of other agents to it. The initiator is the downstream agent initiating the negotiation process by passing the load through T_w, while the respondent is the upstream agent.

Table 5: The negotiation protocol to reduce the load

1	*K=1*
2	The downstream agent (j) asks the upstream agent (i) to reduce the number of requests sent in the period Δt based on R_{ij} according to (5). $R_{ij} = \frac{(load_j - Tw_j)}{\Delta t} * \left(1 - CPU^j_{Occupancy}\right) * CPU^j_{Capacity}$ (5) Load is the number of sessions within the queue.
3	If agent i is in the safe region, it calculates the number of requests that it can process during Δt while not leaving the safe region according to (6) and notifies agent j. Otherwise, agent, i asks the upstream agent (P-CSCS) to reject the request as $R_{ij} + R_{pi}$ randomly. $D_{ij} = \frac{(TW_i - load_i)}{\Delta t} * \left(-CPU^i_{Occupancy}\right) * CPU^i_{Capacity}$ (6)
4	Agent j receiving the answer of agent i, rejects the request locally and randomly as R_{ij}-D_{ij}. If it enters the safe region, it sends the value $R_{ij} = 0$ to agent i and the negotiation ends. Otherwise, agent i recalculates the load reduction rate according to (7) and sends it to j. $R_{ij} = \left(\left(\sum_{k=1}^{I} R^k_{ij} - D^k_{ij}\right)\right) * \beta * I$ (7) Where, *β* is the reduction coefficient and *K* is the number of negotiations.
5	*K=K+1*
6	The above process is repeated until the end of the negotiation.

Results and Discussion

The proposed method is implemented based on RFCs 3261 and 6026 in NS-2 (2.34). NS-2 is run on the same software and hardware platform to compare the studied mechanisms (Fedora Linux 20, Intel(R) Core(TM) i7-6700K CPU @ 4.00GHz 4.00 GHz, Cache Size 8.0 MB, Install RAM 16.0 GB). UDP is considered as the transmission layer protocol. The servers have a processing capacity of 300 sessions per second (SPS). The users have unlimited capacity and can send or receive multiple session requests at the same time. The priority of processing agents is to negotiate messages because failure to process these messages on time causes an overload on the network. If the queue is full when receiving a request, the request will be deleted.

Goodput, sessions delay, number of rejected sessions, stability, and rapid response are the main criteria for evaluating the performance of Reinforcement Learning Overload Controller (RLC).

Goodput is the number of successful sessions that the agent handles per unit of time. A session is successful, which is created in less than 10 secs. A session delay is a time for creating a session. Session response time is

exponential with average of 30 secs. Stability means that the overload controller should not cause throughput fluctuations on the proxy servers and prevent the Goodput from being zero. On the other hand, by the sudden reduction in the traffic imposed on the server, all the applied controls should be removed quickly and the situation should return to the normal state [25], [43].

The offered load was entered into the network as a Poisson distribution by the acceptance and rejection method. The performance evaluation criteria were extracted with a confidence interval of 95%. In the diagrams and tables, the offered loads were divided by the network capacity and normalized. The parameters used in the Q-learning algorithm were shown in Table 6. To obtain the values of T_w and T_c, different input modes with consecutive performances are considered and then the values of these two thresholds are used in RLC.

Table 6: The values of parameters in Q-learning

Parameters	Values
α	0.7
Learning rate	0.9 Decreasingly
Discount factor (γ)	0.4
Exploration rate	0.7

A. Evaluating the Accuracy of RLC Performance

Table 7 shows the average and variance of Goodput, sessions delay, and the number of rejected sessions of RLC, and Overload Controller (OC) when the average number of different input sessions is more than the network capacity; otherwise T_w and T_c have no role in the normal operation of the network. The optimal value of T_c is 37 and the optimal values of T_w are 13, 17 and 25, obtaining as trial and error for OC. In Table 7, the columns of improvement show the percentage which RLC improve performance relative to selected thresholds. In OC, thresholds have better performance, providing better average and less variance. Goodput at T_w= 17 has better performance, T_w = 17 performs better at delaying sessions, and T_w = 25 rejects sessions more efficiently. Since the rejects sessions are approximately equal for T_w=17 and T_w=25, the values T_w = 17 and T_c = 37 are chosen to compare the performance of RLC with OC. Therefore, Goodput is improved through 1.25%, sessions delay is decreased through 3% and number of rejected sessions is reduced through 1.07% by RLC compare with selected OC.

B. Performance Evaluation of RLC

RLC performance is compare with the known overload control methods of CPU occupancy end to end (EOCC), CPU occupancy hop by hop (HOCC) and Holonic overload control (HOC) [5], [36]. The reason for choosing these methods for comparison is 1) They are well-known and include standard codes 2) They have been used in many studies to compare performance, and 3) The end to end methods such as EOCC and HOC have better performance than hop by hop methods. Therefore, comparison with these methods can be a good benchmark to test RLC. Since there is no overload, when the offered load is less than the network capacity, comparisons are only made for offered load more than network capacity. Table 8 shows the improvement of the RLC over the compared methods.

Goodput is shown in Figs. 5. In HOCC, when the downstream S-CSCF is overloaded, it notifies its upstream S-CSCF. Upstream S-CSCF receives this message to reduce the load on the downstream S-CSCF, however, it continues to send since P-CSCF has no knowledge causing overload in the S-CSCF upstream and eventually the entire network.

Table 7: Checking validity and accuracy operation of the RLC

		RLC	T_w=13	Improvement	T_w=17	Improvement	T_w=25	Improvement
Goodput	Average	0.963	0.942	2.2%	0.951	1.25%	0.944	1.97%
	variance	1.571	1.611	2.48%	1.600	0.68%	1.710	8.13%
Sessions delay	Average	0.291	0.301	3.32%	0.300	3.00%	0.311	6.43%
	variance	0.022	0.043	48.8%	0.032	31.3%	0.033	33.3%
Number of rejections	Average	1.022	1.051	2.76%	1.033	1.07%	1.031	0.87%
	variance	298.4	317.1	5.89%	301.7	1.09%	300.6	0.73%

Table 8: RLC performance compare with studied mechanisms

RLC Compare with	HOC	EOCC	HOCC
Goodput	0.6%	13.5%	40.4%
Session delay	-6.7%	49.17%	62.35%
Number of rejections	9.4%	17.3%	31.04%

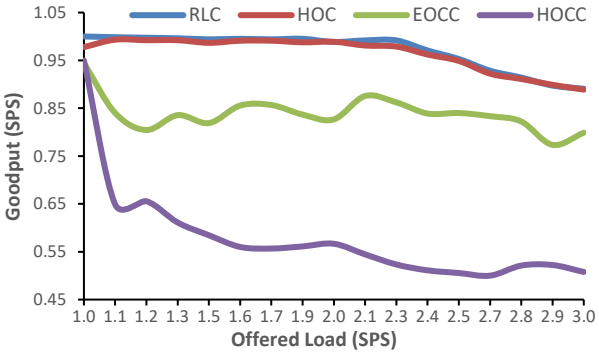


Fig. 5: Goodput for the studied mechanisms.

In EOCC, the probability of acceptance of sessions is adjusted so that the amount of CPU consumption is less than 90% so under heavy load, 10% of CPU capacity is wasted. Hence, the servers do not process at full capacity and Goodput is never exactly equal to *C*. In HOC, CPU capacity is fully utilized. In addition, overload occurrence is prevented in the proxy servers by holons, and the sending load is adjusted from the edge servers. Thereafter, the proxy servers are not overloaded. In RLC, whenever it detects that one of the agents leaves the safe region, it reacts quickly and tries to prevent overload from entering the network by preventing the additional load from entering the source. In RLC, the average of Goodput is almost 0.6% more than HOC because of the hierarchical structure of the HOC, it reacts more slowly than RLC. Goodput of RLC is 13.5% more than EOCC also 40.4% more than HOCC.

The results of sessions delay are shown in Fig. 6. Due to the local view of the overload, the HOCC spends some of the server capacity on processing requests that are eventually deleted, causing delays the processing of other requests. In EOCC, the probability of accepting load from the destination server to the source servers is reported. Upon receiving the restrictions, each server changes its information and sends the updated values to the source servers. As a result, updating the parameters and sending the appropriate amount of load to the destination take time, inadvertently causing the additional load to enter the network and delay the establishment of sessions. HOC has smaller sessions delay (average is 0.13 secs). In HOC, the retransmission mechanism is rarely activated due to keeping the servers in the safe section and not permitting the extra load to enter the network. The due delay exists because of corresponding holon calculation. RLC starts negotiations with the upstream server as soon as the load passes through the secure region. However, in RLC, negotiation process delay is added to RLC delay therefore, sessions delay is 6.7% more than HOC. But RLC sessions delay is 49.17% less than EOCC and 62.35% less than HOCC. Fig. 7 shows the total number of the rejected sessions. By increasing the number of rejected sessions, the network resources are spent to process the requests with no results.

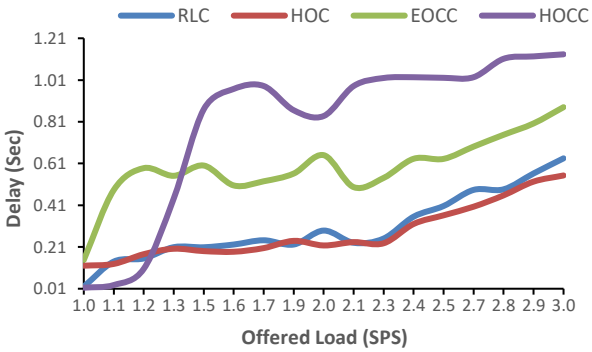


Fig. 6: The sessions delay for the studied mechanisms.

When the number of input sessions is equal to the network capacity, OCC methods reject the sessions due to the CPU consumption threshold of 90%. HOCC reject 31.04% and EOCC reject 17.3% of the session more than RLC. However, in RLC and HOC there are no restrictions. Thus, no session is rejected until the load passes through the network capacity, after which RLC has fewer rejected sessions. Because the rejection of the sessions is based on an intelligent process in accordance with the existing conditions, the negotiation of the agents makes the rejection of the sessions more logical, rather than selfish behavior in HOC. Therefore, RLC reject 9.4% of session less than HOC.

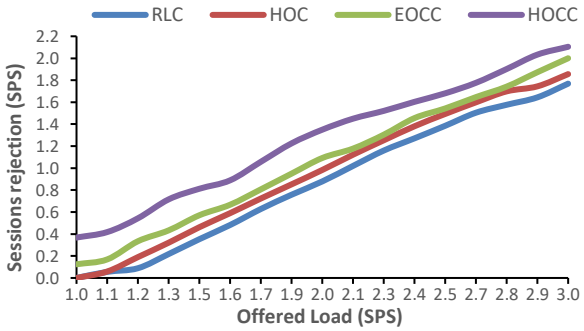


Fig. 7: The number of rejected sessions for the studied mechanisms.

To test the stability and rapid response of the mechanisms, the number of input sessions is initially considered equal to 200 SPS below the network capacity, which lasts up to 100 secs. During this period, Goodput corresponds to the offered load. At 100 secs, the number of offered load suddenly increases to three times the network capacity (900 SPS). With this technique, it can be seen how quickly the mechanisms under study react in the face of sudden load changes and maintain their stability or not. The offered load is returned to 200 SPS at 200 secs. The Goodput and sessions delay of tests are shown in Figs. 8 and 9. In Fig. 8, all mechanisms respond quickly to the offered load sharp changes. Moreover, HOC achieves better Goodput because the offered load is predicted. In HOC, the holons

react quickly to a sudden increase in the offered load and control it from the sources. However, its Goodput is more than RLC due to consecutive switching between the holons. RLC responds steadily to sudden load changes due to the negotiation. This is because the agents act quickly by increasing the offered load and entering the warning area. Moreover, they prevent overload from occurring through negotiation. In HOCC, the Goodput is initially decreased and then increased. Because the control parameters are updated based on 200 SPS and until the next update, sessions are accepted. EOCC methods require time to propagate overload information from destination to the sources, causing temporary instability. At 200 secs, Goodput of RLC returns to its previous value without any fluctuation. Therefore, RLC completely satisfies stability.

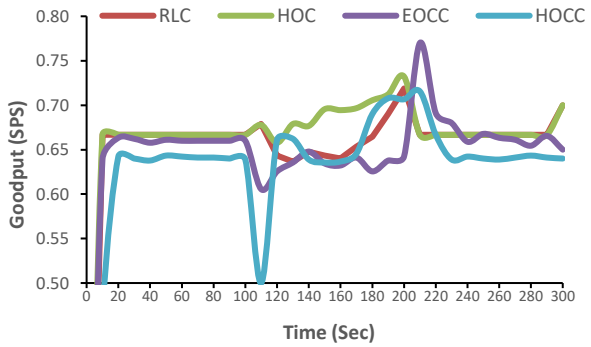


Fig. 8: Goodput of the studied mechanisms when the offered load changes suddenly.

In Fig. 9, RLC has less delay than other methods. When the load returns to its previous value, the RLC delay will return to the value before the change. As the offered load increases due to the lack of up to date parameters in OCC methods, a large amount of load enters the network, and the delay increases. By removing the overload, the delay is slightly reduced. Parameter values in EOCC are updated with more delay due to being end to end and passing of control values over the entire network. The HOC makes a temporary error because it predicts load based on previous observations and the current load is very different from the previous.

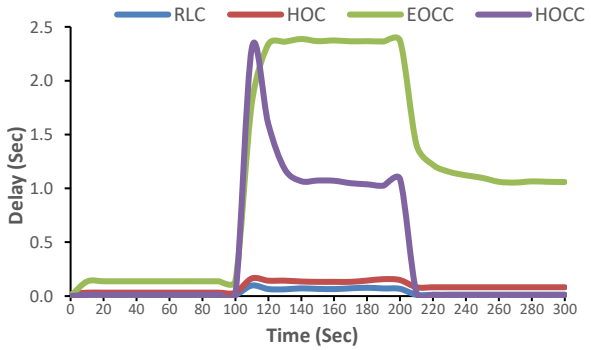


Fig. 9: Sessions delay of the studied mechanisms when the offered load changes suddenly.

To evaluate the performance of RLC in a real VOIP environment, the Goodput and sessions delay for variable offered load with Poisson distribution are shown in Figs. 10 and 11. According to Fig. 10, RLC follows the changes in the number of incoming sessions well and adapts to it without network failure. When the offered load and Goodput are not distinguishable, they have been overlapped because Goodput is equal to offered load. In addition, the delay is controlled and fluctuated with the load changes and the system becomes stable.

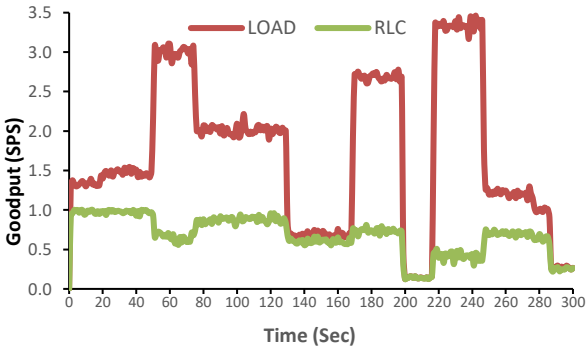


Fig. 10: Goodput of RLC under real offered load.

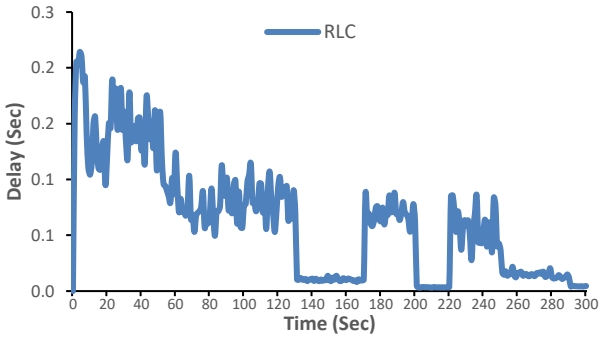


Fig. 11: The sessions delay of RLC under real offered load.

Conclusion

IMS will become the most important platform for multimedia applications. By increasing the number of users, the throughput of the IMS servers is decreased. On the other hand, the issue of overload control in IMS is a complex system, for which multi-agent systems are good alternatives to classical solving methods. In multi-agent systems, a large task can be divided into a set of smaller tasks so that each agent performs a task partially. In this study, IMS servers are considered as a learner and negotiator agents. Agents learn the values of thresholds by Q-learning and they implement a hop by hop control method through negotiates strategy with the upstream agent. To prove the performance of the proposed method, it was compared to similar methods. In Table 8, we reported the efficiency, the mean sessions delay, the average of Goodput and number of rejected sessions for different methods. As shown, RLC has better performance on all three measured parameters; while only, its delay is more than HOC. Because holonic

communication of HOC is faster than negotiation process of RLC. In the proposed method, the learning process is done independently by each agent. Although this type of learning is suitable for obtaining the parameters related to each agent, to show the optimal reactions by all agents, it is better to perform learning in the whole network to implement end to end methods through clustering. On the other hand, in IMS, there are HSS and DNS that are not based on SIP and contribute to the overload. These heterogeneous creatures can be inserted into the problem. Nowadays, by moving the process to cloud environments with NFV, the cost of IMS structure and platform has decreased. By quickly developing this method in scalable form, the proposed method can be implemented in cloud environments using NFV.

Author Contributions

M. Khazaei wrote the manuscript, designed the experiments, analysis the data, interpreted the results and revised the manuscript.

Acknowledgment

The author would like to thank the editor and reviewers for their helpful comments.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

IMS	IP Multimedia Subsystems
NGN	Next-Generation Network
SIP	Session Initiation Protocol
MAS	Multi Agent Systems
QOS	Quality of Services
HSS	Home Subscriber Server
SLF	Subscription Locator Function
CSCSF	Call Session Control Functions
S-CSCF	Serving CSCF
P-CSCF	Proxy CSCF
I-CSCF	Interrogating CSCF
NFV	Network Functions Virtualization
NLMS	Normalized Least Mean Square
SPS	Sessions per Second
C	CPU Capacity
RLC	Reinforcement Learning Overload Controller
OC	Overload Controller
EOCC	End to end Occupancy
HOCC	Hop by hop Occupancy

References

[1] P. Agrawal, Y. Jui-Hung, C. Jyh-Cheng, Z. Tao, "IP multimedia subsystems in 3GPP and 3GPP2: overview and scalability issues," IEEE Commun. Mag., 46: 138-145, 2008.

[2] K. K. Guduru, U. Jayadevappa, "Overload control in SIP signalling networks with redirect servers," Int. J. Wireless Mobile Comput., 19: 124-132, 2020.

[3] V. S. Vaishnavi, Y. M. Roopa, P. L. Srinivasa Murthy, "A survey on next generation networks," in Proc. ICCNCT 2019: 162-171, 2020.

[4] C. Shen, H. Schulzrinne, E. Nahum, "Session Initiation Protocol (SIP) server overload control: Design and evaluation," in Principles, Systems and Applications of IP Telecommunications. Services and Security for Next Generation Networks, S. Henning, S. Radu, and N. Saverio, Eds., ed: Springer-Verlag: 149-173, 2008.

[5] V. Hilt, I. Widjaja, "Controlling overload in networks of SIP servers," in Proc. IEEE International Conference in Network Protocols: 83-93, 2008.

[6] J. Davin, P. Riley, M. Veloso, "CommLang: Communication for coachable agents," in Proc. RoboCup 2004: Robot Soccer World Cup VIII. vol. 3276, D. Nardi, M. Riedmiller, C. Sammut, and J. Santos-Victor, Eds., ed: Springer Berlin Heidelberg: 46-59, 2005.

[7] M. Wooldridge, An Introduction to MultiAgent Systems, Newyork: Wiley, 2009.

[8] B. Jang, M. Kim, G. Harerimana, J. W. Kim, "Q-Learning algorithms: A comprehensive classification and applications," IEEE Access, 7: 133653-133667, 2019.

[9] M. Abdoos, N. Mozayani, A. C. Bazzan, "Hierarchical control of traffic signals using Q-learning with tile coding," Appl. Intell., 40: 201-213, 2014.

[10] H. C. Hsieh, J. L. Chen, "Distributed multi-agent scheme support for service continuity in IMS-4G-Cloud networks," Comput. Electr. Eng., 42: 49-59, 2015.

[11] D. Pereira, R. Oliveira, H. S. Kim, "A machine learning approach for prediction of signaling sip dialogs," IEEE Access, 9: 44094-44106, 2021.

[12] F. B. Mismar, J. Hoydis, "Unsupervised learning in next-generation networks: Real-time performance self-diagnosis," IEEE Commun. Lett., 25: 3330-3334, 2021.

[13] S. Chen, Z. Yao, X. Jiang, J. Yang, L. Hanzo, "Multi-agent deep reinforcement learning-based cooperative edge caching for ultra-dense next-generation networks," IEEE Trans. Commun., 69: 2441-2456, 2021.

[14] M. Abdoos, N. Mozayani, A. L. C. Bazzan, "Holonic multi-agent system for traffic signals control," Eng. Appl. Artif. Intell., 26: 1575-1587, 2013.

[15] M. Abdoos, N. Mozayani, A. L. C. Bazzan, "Traffic light control in non-stationary environments based on multi agent Q-learning," in Proc. 14th International IEEE Conference in Intelligent Transportation Systems (ITSC): 1580-1585, 2011.

[16] Y. Yao, V. Hilaire, A. Koukam, W. Cai, "A holonic model in wireless sensor networks," in Proc. International Conference in Intelligent Information Hiding and Multimedia Signal Processing: 491-495, 2008.

[17] S. M. Hosseini, N. Mozayeni, "An intelligent method for resource management in wireless networks," in Proc. 5th Conference in Information and Knowledge Technology (IKT): 371-376, 2013.

[18] E. Pei, L. Zhou, B. Deng, X. Lu, Y. Li, Z. Zhang, "A Q-Learning based energy threshold optimization algorithm in LAA networks," IEEE Trans. Veh. Technol., 70: 7037-7049, 2021.

[19] M. Cossentino, N. Gaud, V. Hilaire, S. Galland, A. Koukam, "ASPECS: an agent-oriented software process for engineering complex systems," Auton. Agents Multi-Agent Syst., 20: 260-304, 2010.

[20] M. Poikselkä, The IMS: IP multimedia concepts and services: Newyoek, J. Wiley & Sons, 2006.

[21] N. M. Ahmed, N. E. Rikli, "QoS-Based data aggregation and resource allocation algorithm for machine type communication devices in next-generation networks," IEEE Access, 9: 119735-119754, 2021.

[22] V. K. Gurbani, R. Jain, "Transport protocol considerations for session initiation protocol networks," Bell Labs Tech. J., 9: 83-97, 2004.

[23] M. Ohta, "Overload control in a SIP signaling network," Int. J. Electr. Electron. Eng., 3: 87-92, 2009.

[24] E. N. V. Hilt, C. Shen, A. Abdelal, "Design considerations for session initiation protocol (SIP) overload control," Internet Engineering Task Force (IETF), Request for Comments: RFC6357, 2011.

[25] J. Liao, J. Wang, T. Li, J. Wang, J. Wang, X. Zhu, "A distributed end-to-end overload control mechanism for networks of SIP servers," Comput. Networks, 56: 2847-2868, 2012.

[26] H. Dong-Yeop, P. Ji Hong, Y. Seung-wha, K. Ki-Hyung, "A window-based overload control considering the number of confirmation massages for SIP server," in Proc. Fourth International Conference in Ubiquitous and Future Networks (ICUFN): 180-185, 2012.

[27] M. Homayouni, H. Nemati, V. Azhari, A. Akbari, "Controlling Overload in SIP Proxies: An Adaptive Window Based Approach Using No Explicit Feedback," in Proc. IEEE Global Telecommunications Conference: 1-5, 2010.

[28] S. V. Azhari, M. Homayouni, H. Nemati, J. Enayatizadeh, A. Akbari, "Overload control in SIP networks using no explicit feedback: A window based approach," Comput. Commun., 35: 1472-1483, 2012.

[29] A. Montazerolghaem, M. H. Yaghmaee Moghadam, "Improving efficiency of SIP protocol using window-based overload conditions," Soft Comput. J., 2: 16-25, 2021.

[30] M. Khazaei, N. Mozayani, "A dynamic distributed overload control mechanism in SIP networks with holonic multi-agent systems," Telecommun. Syst., 63: 437-455, 2016.

[31] Y. Hong, C. Huang, J. Yan, "Applying control theoretic approach to mitigate SIP overload," Telecommun. Systems, 54: 387-404, 2013.

[32] A. Abdelal, W. Matragi, "Signal-Based Overload Control for SIP Servers," in Proc. 7th IEEE Consumer Communications and Networking Conference: 1-7, 2010.

[33] R. G. Garroppo, S. Giordano, S. Niccolini, S. Spagna, "A Prediction-Based Overload Control Algorithm for SIP Servers," Network and Service Management, IEEE Transactions, 8: 39-51, 2011.

[34] S. Jing, T. Ruixiong, H. Jinfeng, Y. Bo, "Rate-based SIP flow management for SLA satisfaction," in IFIP/IEEE International Symposium on Integrated Network Management: 125-128, 2009.

[35] M. Khazaei, "Occupancy overload control by Q-learning," in Fundamental Research in Electrical Engineering, Singapore: 765-776, 2019.

[36] M. Khazaei, N. Mozayani, "Overload management with regard to fairness in session initiation protocol networks by holonic multiagent systems," Int. J. Network Manage., 27: e1969, 2017.

[37] A. Akbar, S. M. Basha, S. A. Sattar, "A cooperative overload control method for SIP servers," International Conference in Proc. Communications and Signal Processing (ICCSP): 1296-1300, 2015.

[38] A. Montazerolghaem, S. K. Shekofteh, M. H. Yaghmaee, M. Naghibzadeh, "A load scheduler for SIP proxy servers: design, implementation and evaluation of a history weighted window approach," Int. J. Commun. Sys., 2015.

[39] A. Montazerolghaem, M. H. Y. Moghaddam, A. Leon-Garcia, "OpenSIP: Toward software-defined SIP networking," IEEE Trans. Netw. Serv. Manage., 15: 184-199, 2018.

[40] R. Gandotra, L. Perigo, "SDVoIP—A software-defined VoIP framework for SIP and dynamic QoS," Comput. J., 64: 254-263, 2019.

[41] L. D. Cicco, G. Cofano, S. Mascolo, "Local SIP overload control: controller design and optimization by extremum seeking," IEEE Trans. Control Network Syst., 2: 267-277, 2015.

[42] Y. Hong, C. Huang, J. Yan, "Modelling chaotic behaviour of SIP retransmission mechanism," Int. J. Parallel Emerg. Distrib. Syst., 28: 95-122, 2013.

[43] J. Wang, J. Liao, T. Li, J. Wang, J. Wang, Q. Qi, "Probe-based end-to-end overload control for networks of SIP servers," J. Network Comput. Appl., 41: 114-125, 2014.

Biographies

Mehdi Khazaei received a B.Sc. degree in computer Engineering (Computer Hardware) from Iran University of Science and Technology (Tehran, IRAN); M.Sc. and Ph.D. degree in computer systems Architecture from Iran University of Science and Technology (Tehran, IRAN) in 2017. He is currently assistant professor in the School of Information Technology at Kermanshah University of Technology (Kermanshah, Iran).

- Email: m.khazaei@kut.ac.ir
- ORCID: 0000-0002-4780-065X
- Web of Science Researcher ID: NA
- Scopus Author ID: 1027784
- Homepage: https://kut.ac.ir/en/profile/6-mehdi-khazaei

How to cite this paper:
M. Khazaei, "Reinforcement learning-based load controller in IP multimedia subsystems," J. Electr. Comput. Eng. Innovations, 11(1): 21-32, 2023.

DOI: 10.22061/JECEI.2022.8723.546

URL: https://jecei.sru.ac.ir/article_1713.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Feasibility of Digital Circuit Design Based on Nanoscale Field-Effect Bipolar Junction Transistor

A. Shokri, M. Amirmazlaghani*

Nanoelectronics Lab (NEL), Shahid Rajaee Teacher Training University, Tehran, Iran.

Article Info

Article History:
Received 09 January 2022
Reviewed 12 March 2022
Revised 04 May 2022
Accepted 14 May 2022

Keywords:
BJT
FET
Inverter
logic gate
Nanoscale

*Corresponding Author's Email Address:
m.mazlaghani@sru.ac.ir

Abstract

Background and Objectives: The Field-effect Bipolar Junction Transistor (FEBJT) is a device with a bipolar junction transistor (BJT) characteristics except that it is designed with standard CMOS technology. Therefore, it can be implemented in nanometer dimensions without the usual restrictions in fabricating the nanoscale BJTs. In addition to the advantages that FEBJT has as a bipolar junction transistor in analog circuits, it can also be used to design digital circuits. Here, we have investigated the capability of FEBJT as the base of a new digital family in nanometer scales.

Methods: To do this, we have designed and simulated an inverter logic gate based on FEBJT. We have presented this logic gate's static and dynamic assessment criteria and compared these characteristics with other technologies. Also, a three-stage ring oscillator circuit based on FEBJT is designed and presented. A three-dimensional TCAD Mixed-Mode simulator has been used for the simulations.

Results: The value of maximum frequency, PDP, dynamic power, and ring frequency are calculated 0.25THz, 38×10^{-17} J, 94uW, and 85GHz, respectively.

Conclusion: The excellent function of the FEBJT-based inverter gate and oscillator demonstrates that this device can be used as the base of new digital circuits and can open a doorway to the nanoscale CMOS digital family.

Introduction

The first semiconductor transistor was built in 1947. This element was called the bipolar junction transistor (BJT) developed by Schokley, Barden, and Brattain at Bell Lab [1]. The BJT has high current gain, high speed, and low input capacitance. With the advent of technology and the shrinking of electronic circuits, the fabrication of BJTs on the nanometer scale faced some restrictions. Reducing the width of the base area to nanometer dimensions was a significant obstacle to making this transistor smaller. MOS Field-Effect Transistors (MOSFETs) were another type of transistor that evolved rapidly, and CMOS technology replaced BJT in many circuits. One of the reasons for the development of CMOS technology was the possibility of manufacturing these elements in nanometer dimensions. Despite the recent advances in field-effect transistors, the specific features of bipolar transistors such as high current gain have led researchers to continue to look for solutions to utilize BJT transistors alongside MOSFETs. To this end, several papers have been presented that attempt to use both transistors at the same time or provide a solution to fabricate a smaller bipolar junction transistor to utilize the powerful features of the BJT transistor in today's small-scale digital industries [2]-[20].

By integrating BJT and CMOS transistors, BiCMOS technology simultaneously uses each of these transistors' special features [2]-[8]. The BJT transistor has been implemented horizontally, allowing for a smaller base area width. Plasma charging and polarity control of

electrodes have been proposed as other methods to avoid the need for silicon doping to create bases, collectors, and emitters in the BJT structure [9]-[17]. Another proposed device is the Field-Effect bipolar junction transistor (FEBJT), which is a BJT that is designed based on the idea of changing the doping level of the semiconductor by the electric field of the gate electrodes [18]-[20]. In other words, this device enables the implementation and fabrication of a BJT transistor with CMOS technology. In FEBJT, a BJT transistor's base, collector, and emitter regions are created using three-gate electric fields, called the gate-base, gate-collector, and gate-emitter, respectively. This structure can reduce the width of the BJT transistor to 7 nm. In addition to the analog advantages of FEBJT, this device can turn on and off with the base, emitter, and collector gates [18]-[20]. Therefore, digital electronics can be designed based on FEBJT without the current shrinking of the base electrode in BJT-based digital circuits. What this paper aims to present is to show the feasibility of digital applications for FEBJT.

In this paper, an inverter logic gate, as the base block of the digital family, is simulated and presented. The circuit design and transition characteristic diagrams are examined. The transient state responses of the segment and the times of the ups and downs have also been measured and calculated. Noise calculation at the primary logic gate is also performed and presented.

Field-effect Bipolar Junction Transistor (FEBJT)

Fig. 1 shows the schematic of FEBJT. This structure is simulated by the TCAD-3D simulator. Fig. 1(a) is the side view of this element. The design parameters are detailed in Table 1. The structure has three electrodes on the insulator: gate-collector, gate-base, and gate-emitter. It also has three common BJT electrodes: the collector, the emitter, and the base. The base electrode can be on either side of the structure. Fig. 1(b) shows the top view of the structure and the location of the base electrode. By applying a positive voltage to each gate, the n-type region can be created, and by using a negative voltage, the p-type region can be formmade on the silicon surface under the gate. Thus, it can be said that the structure is similar to MOSFET, except that the three gates are placed on the oxide instead of one gate [18], [20].

Due to the positive or negative voltage applied to the gates, there are eight modes for the silicon channel under the gates. The eight modes are shown in Table 2. Among the eight possible modes for the transistor channel, mode 3, which creates the npn structure within the channel, which is in accordance with the npn-BJT, is considered as on mode. Mode 6 of this table is considered as OFF mode [18]. Fig. 2 illustrates the circuit schematic of this device. Using this segment, two types of n-channel and p-channel can be designed according to npn and pnp BJTs. The device is designed based on SOI (Silicon on Insulator) technology. The base width of the proposed FEBJT is considered 20nm. The feature that determines FEBJT's speed and current gain is the size of the base gate, which is 20nm. The mixed-mode module of the ATLAS simulator is used for circuit model extraction. Using this module, the devices can be simulated numerically. A SPICE-like circuit description is provided in the MixedMode module. In other words, after the definition of a new concept device (like FEBJT) in the ATLAS-TCAD 3D simulator, the MixedMode module can extract the library of the new device. Different analog or digital circuits can then be defined in a SPICE-like description.

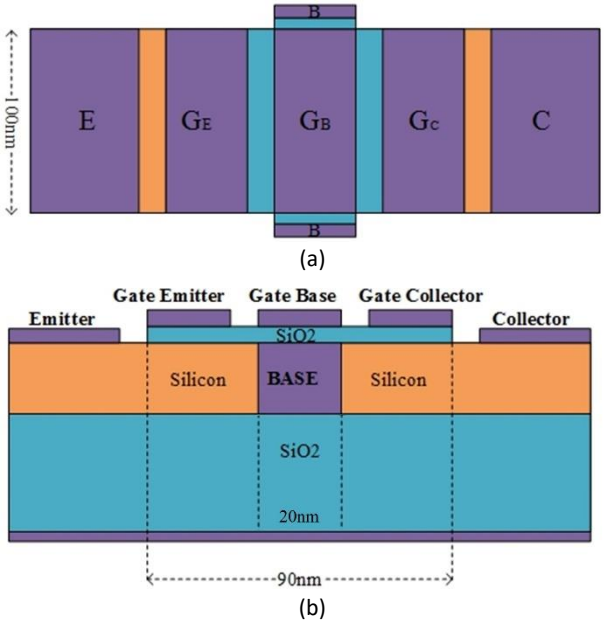


Fig. 1: (a) Side view of FEBJT. Three gates over a silicon channel induce the channel's emitter, base, and collector area. (b) Top view of the FEBJT structure. GE, GB, and GC are gate-emitter, gate-base, and gate-collector, respectively.

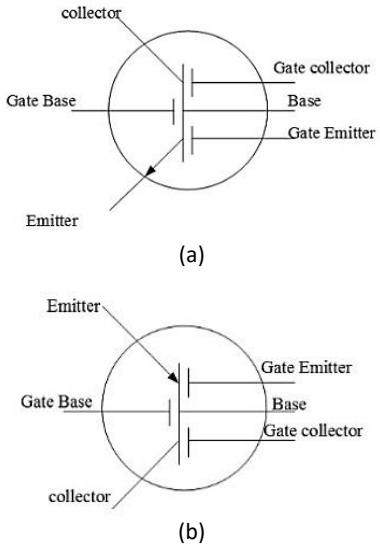


Fig. 2: (a) Circuit schematic model for npn FEBJT, (b) Circuit schematic model for pnp FEBJT.

Table 1: The design parameters of FEBJT

Symbol	Value
Thickness of Silicon film	100nm
Thickness of buried oxide	400nm
Thickness of gate oxide	5nm
Width of Structure	100nm
Length of G_E and G_C	30nm
Length of G_B	20nm
Emitter Length	40nm
Collector Length	40nm
Emitter Doping Concentration	$1\times10^{19}cm^{-3}$
Collector Doping Concentration	$1\times10^{18}\ cm^{-3}$
Channel Doping	$3\times10^{17}\ cm^{-3}$

Table 2: Possible modes for silicon channels in the FEBJT structure

STATE	Type of Structure	V_{GE}	V_{GB}	V_{GC}
1	n-nnn-n	+	+	+
2	n-nnp-n	+	+	-
3	n-npn-n	+	-	+
4	n-pnn-n	-	+	+
5	n-ppn-n	-	-	+
6	n-pnp-n	-	+	-
7	n-ppn-n	-	-	+
8	n-ppp-n	-	-	-

Inverter Logic Gate Based on FEBJT

A. Static Characteristics

Fig. 3 shows the designed inverter logic gate using FEBJT. Like CMOS inverter gate, which is designed using two complement n-channel and p-channel MOSFETs, the FEBJT inverter gate is also designed by Table 1: The structural design parameters for FEBJT. Combining two n-channel and p-channel transistors. The voltage applied to the electrodes of this structure is illustrated in Fig. 3.

Table 3 shows how to apply voltage to the device gates to generate the *low* and *high* logic outputs. Important parameters must be considered to check the quality of an inverter gate. The most important of these factors are voltage transient characteristic, output transient mode characteristic, output capacitance, output resistance, Power Delay Product (PDP), and speed [21]-[30]. The voltage transfer characteristic is a graph showing the output voltage changes relative to the input voltage. Fig. 4 shows the transient voltage characteristic of the inverter logic gate circuit with the FEBJT device. The input voltage is swept from -1 to 1 volt to obtain this characteristic. As specified in Table 3, the input voltage is applied to the gate-emitter (GE) and gate-collector (GC). Gate-base (GB) is also biased with VCC (1V).

The values are shown in Fig. 4 are used to calculate the noise margin. V_{OH} refers to the maximum output value known as logic 1. V_{OL} refers to the minimum output value known as logic 0. V_{IH} refers to the maximum input value known as logical input 1. V_{IL} refers to the minimum input value known as logical input 0. Table 4 shows the values of the parameters extracted from the voltage transient characteristic. The Noise Margin Low (NML) and Noise Margin High (NMH) values can be calculated from (1) and (2) [31]:

$$NM_L = V_{IL} - V_{OL} \quad (1)$$

$$NM_H = V_{OH} - V_{IH} \quad (2)$$

By placing the values extracted from Fig. 4 in the above equations, the NML and NMH are calculated 0.43 and 0.79 volts, respectively. It is worth noting that depending on whether the input signal is applied to the gate or base electrodes, different circuits can be designed as NOT-gate based on FEBJT. Fig. 4, the manuscript shows one possible configuration for the NOT-gate circuit in which the input is applied to the side gates of both transistors. Each designed inverter circuit would have specific voltage transfer characteristics curves (VTC), which the carrier concentrations and dopant densities can change through the channel. The reason for this change is the resistivity of the channel, which is controllable by different doping concentrations.

Fig. 5 shows the VTC for four different doping levels of the channel. To explain the reason for changing the curves shown in Fig. 4, look back at the schematic of the inverter gate in Fig. 3 inside the manuscript. As shown, there are two complementary FEBJT in a NOT-gate to create inverting behavior. Increasing the doping level of each transistor results in decreasing the channel resistivity in that transistor. By increasing the hole concentrations in pnp FEBJT, the resistivity of the above transistor decrease, and consequently, the VTC shifts to higher voltages in the right side of VTC. When the electron densities increase in npn FEBJT, the resistivity of the bottom transistor decrease and the VTC shifts to the smaller voltages and right side of the curve.

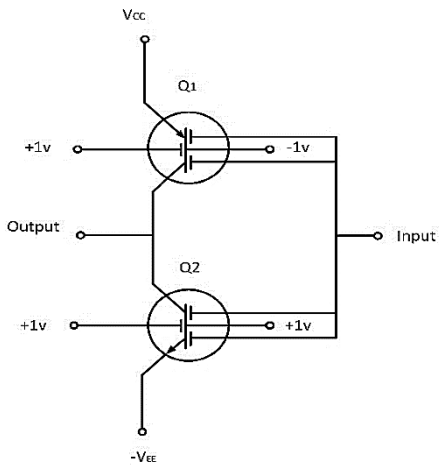


Fig. 3: The designed inverter logic circuit based on FEBJT.

Table 3: Truth table for FEBJT-based inverter logic gate

Input	Q_1	Q_2	G_E	G_C	G_B	Output
Low	on	off	input	input	Bias(+1v)	High
High	off	on	input	input	Bias(+1v)	Low

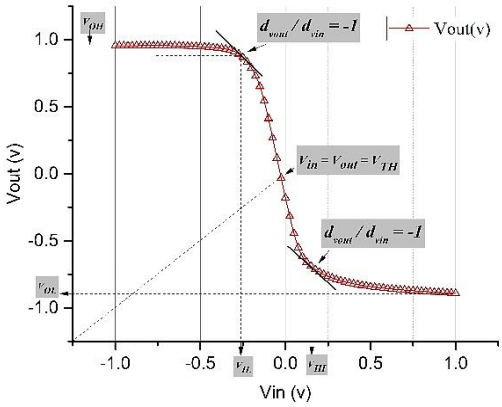


Fig. 4: The voltage transfer characteristic of the inverter logic gate circuit or not gate with the FEBJT device.

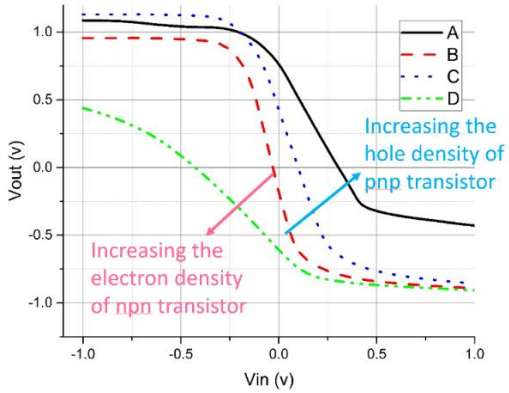


Fig. 5: The VTC for four different doping levels of the channel. By increasing the hole concentrations in pnp FEBJT in NOT-gate, the resistivity of the above transistor decreases. Consequently, the VTC shifts to higher voltages in the right side of VTC. When the electron densities increase in npn FEBJT, the resistivity of the bottom transistor decrease and the VTC shifts to the smaller voltages and right side of the curve.

B. Dynamic Characteristics

Fig. 6 shows the transient response of the inverter circuit based on FEBJT. As is demonstrated in Fig. 6(a), a voltage pulse in the range of (-1V to 1V) with a period of 2 microsecond is applied as the input signal. The output signal is also demonstrated in this figure. As can be seen, the output signal exactly follows the input in reverse. To characterize the dynamic performance of a logic-circuit family, the propagation delay of the basic inverter gate is usually examined [31].

2 microsecond is applied as the input signal. The output signal is also demonstrated in this figure. As can be seen, the output signal exactly follows the input in reverse. To characterize the dynamic performance of a logic-circuit family, the propagation delay of the basic inverter gate is usually examined [24].

Table 4: The value of voltage transfer characteristic parameters

Voltage	Amount (v)
V_{OH}	0.95
V_{OL}	-0.88
V_{IL}	-0.27
V_{IH}	0.16
V_{TH}	-0.04

The inverter propagation delay time (τ_p) is defined as the average of the high to low and low to high propagation delays as follows [31]:

$$\tau_p = \frac{\tau_{phl}+\tau_{plh}}{2} \quad (3)$$

In this equation, τ_{phl} and τ_{plh} are defined as the required time for the output to reach 50% of the rail-to-rail voltage. Two other parameters to examine the dynamic performance of digital circuits are the rise and the fall time of the output pulse. These two parameters (τ_r) and (τ_f) are defined as the required time for the output to change from 10% to 90% and from 90% to 10% of the rail-to-rail voltage, respectively. The transient responses are zoomed-in Figs. 6(c) and 6(d). The calculated dynamic parameters for the FEBJT inverter are listed in Table 5.

We have measured the output of the proposed inverter gate for four different load capacitors. The propagation delay times decrease when the output capacitor decrease. The rise, fall, and propagation delay times are shown in Fig. 7 as a function of the load capacitor. This curve shows that the parasitic capacitor at the output node is around 500fF. One of a digital circuit's most important design parameters is the Power Delay Product (PDP). This parameter indicates the amount of energy needed to change the output from the maximum value to its minimum value and vice versa. PDP can be calculated from the following equation [31]:

$$PDP = C_{load} \times {V_{dd}}^2 \quad (4)$$

C_{load} is the load capacitance that indicates the output node capacitance, and V_{dd} is the power supply voltage. There are different solutions for estimating the output node capacitance [31]. Here, we first put an external 1pf capacitor at the output node to calculate the output resistance from the time constant of the output curve (see Fig. 6(b)). After determining the output resistance, we looked back to the output curve of the inverter when no external capacitance was connected (see Fig. 6(a)). By knowing the value of the output resistance (25kΩ), the output node capacitance (C_{load}) was calculated 0.06 fF from the time constant of the output curve. Using (4), PDP was calculated about 38×10^{-17} J. Dynamic power, and the maximum frequency of the inverter logic gate are the

other important dynamic parameters that estimate the performance of a new digital family [24]. Using (5) and (6) [31], the maximum frequency and the dynamic power of the FEBJT inverter gate were calculated 0.25THz and 94uW, respectively.

$$F_{max} = \frac{1}{(\tau_{phl} + \tau_{plh})} \quad (5)$$

$$P_D = C_{load} \times {V_{dd}}^2 \times F_{max} \quad (6)$$

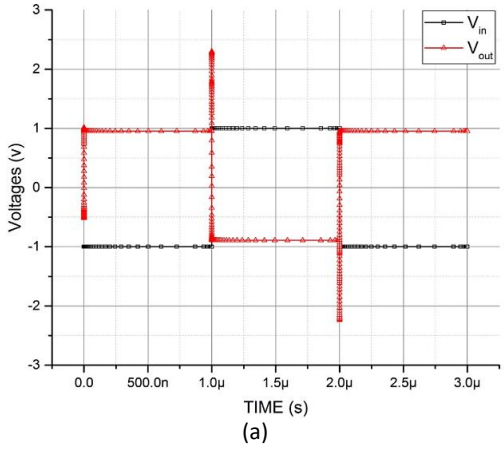
The important digital factors of the FEBJT inverter are compared with other field-effect transistors with an almost similar gate length in Table 6.

C. Ring Oscillator

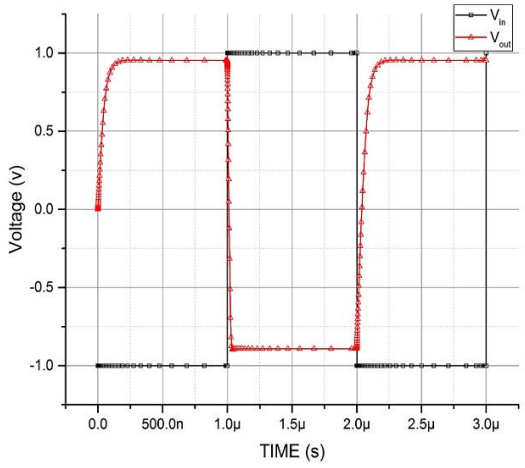
In this part, a three-stage ring oscillator based on FEBJT is designed and simulated, and its performance is evaluated and compared with other technologies. Fig. 8(a) demonstrates the designed ring circuit based on FEBJT. This oscillator is a combination of an odd number of inverters, and the output of each stage is given as input to the next stage. The output of the last stage is connected to the first stage, thus forming a ring. Each inverter stage provides a specific delay time; thus, the three-stage circuit starts to oscillate at a particular frequency [30]. The oscillation frequency is the function of the delay time of each stage and the number of stages used in the ring circuit [30]:

$$F_{osc} = \frac{1}{2n\tau_p} \quad (7)$$

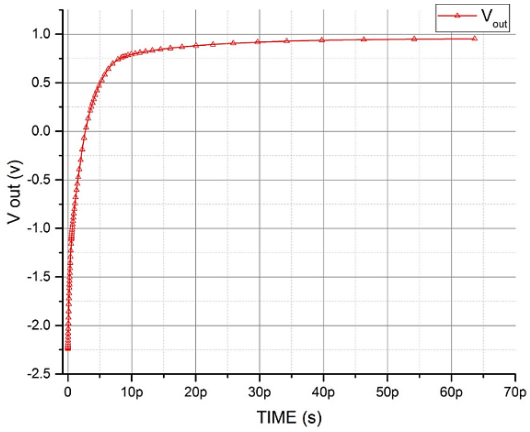
N is the number of the stages, and τ_p is the delay time of a single inverter stage. Considering n=3 and τ_p=2.01p, the oscillation frequency of the designed ring is calculated 85GHz. The output voltage of the ring oscillator based on FEBJT is demonstrated in Fig. 8(b). The performance of the designed ring oscillator is compared with other technologies in the almost similar gate length in Table 7. As can be seen in this table, the operation frequency of the ring oscillator based on FEBJT is larger than the other comparable proposed devices in many published works, and that's while the dynamic power of the inverter stage is smaller than the others.



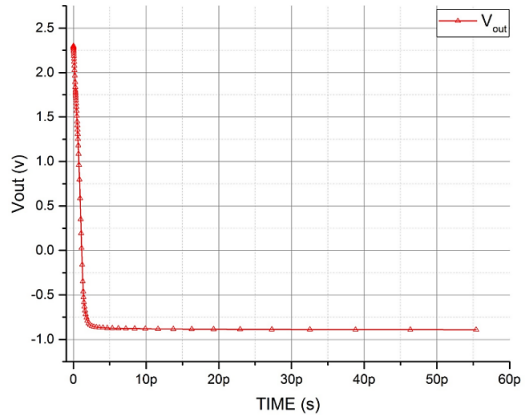
(a)



(b)



(c)



(d)

Fig. 6: The transient response of the inverter circuit based on FEBJT. (a) without the external capacitor, (b) with 1pf external capacitor. (c) The rise up of the output voltage is zoomed, τ_r is measured 7.79 ps. (d) The fall down of the output voltage is zoomed, τ_f is measured 0.84 ps.

Results and Discussion

The static and dynamic simulation results for the inverter logic gate based on FEBJT and the calculated performance of the ring oscillator built with FEBJT demonstrate the ability to use FEBJT in the design of digital circuits.

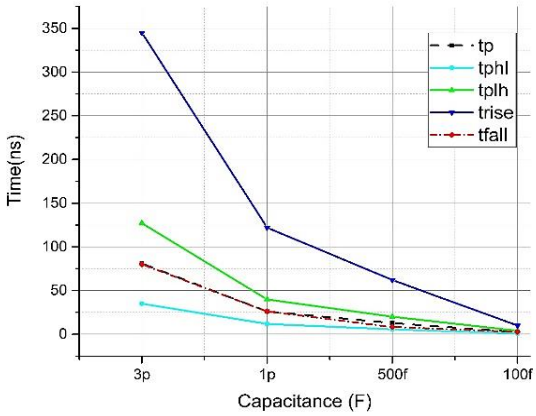


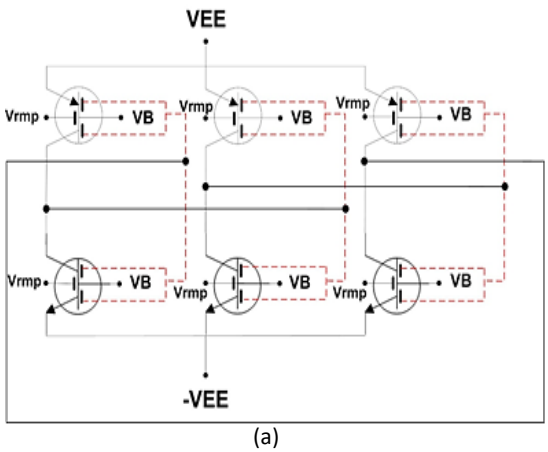
Fig. 7: The output of the proposed inverter gate for four different load capacitors. The rise, fall, and propagation delay time as capacitor function. As can be seen, the propagation delay times decrease when the output capacitor decrease.

Table 5: The value of transient state parameters

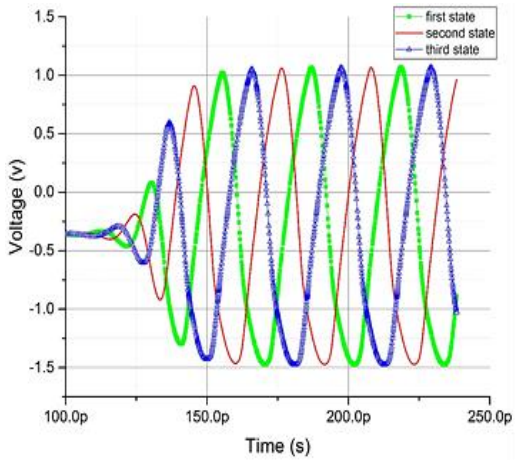
Time	Amount (s)
t_{phl}	1.11p
t_{plh}	2.89p
t_{fall}	0.84p
t_{rise}	7.79p
t_p	2.01p

Table 6: Comparison of the Field-effectBJT Reverse Logic Gate with other devices [29], [32]-[36]

Device	Structure	T_p (ps)	PDP (j)	Gate Length (nm)
This work	Field-effectBJT	2.01	38×10^{-17}	20
[29]	S-FED	1.04	6.2×10^{-18}	25
[32]	FINFET	8	7.0×10^{-18}	20
[33]	S-bulk Finfet	53000	3.0×10^{-13}	17
[34]	HTFET	710	3.6×10^{-20}	20
[35]	FINFET	2.3	0.01×10^{-21}	16-32
[36]	CNTFET	24	0.04×10^{-18}	20



(a)



(b)

Fig. 8: (a) The designed three-stage ring oscillator circuit based on FEBJT. (b) The oscillation of the three-stage circuit shown in Fig. 6(a). The oscillation frequency is measured 84GHz.

Table 7: The performance of the designed ring oscillator is compared with other technologies [37]-[47]

Device	Structure	State of ring	Frequency (GHz)	$Power_{dyn}$ (w)	Gate length(nm)
This work	FEBJJT	3	82	94×10^{-6}	20
[37]	DJ-JNT	3	4	-	20
[38]	FDSOI	3	2.45	-	32
[39]	Dg-JNT	3	52	-	20
[40]	CMOS	4	8.33	435×10^{-6}	65
[41]	FDSOI	3	49	3.77×10^{-3}	28
[42]	CMOS	4	16	46.2×10^{-3}	20
[43]	FINFET	3	40	-	20
[44]	V-TFET	11	0.6	86×10^{-9}	20
[45]	DG-FET	3	4.14	12×10^{-6}	20
[46]	TFET	9	-	1.5×10^{-15}	14
[47]	NCFET	17	2.9	-	18

Conclusion

A digital circuit was designed using the FEBJT element. This circuit is an inverter logic gate. Important features and values calculated, such as voltage transient characteristic, transient state, output capacitance, output resistance, and PDP, indicate that FEBJT is applicable for digital applications. Further, it is recommended to research other digital basic circuits with the help of this device with unique features.

Author Contributions

M. Amirmazlaghani designed the experiments. A. Shokri performed the simulations. M. Amirmazlaghani and A. Shokri interpreted the results and wrote the manuscript.

Acknowledgement

The authors would like to thank Dr. Farshid Raissi for his scientific comments.

Conflict of Interest

The author declares that there is no conflict of interests regarding the publication of this manuscript.

Abbreviations

FEBJT	Field-effect Bipolar junction transistor
BJT	bipolar junction transistor
PDP	Power Delay Product
CMOS	Complementary Metal Oxide Semiconductor
FET	Field-effectTransistor
MOSFET	Metal Oxide Semiconductor Field-effectTransistor
NML	Noise Margin Low
NMH	Noise Margin High

References

[1] J. N. Bureghartz, "Guide to State-of-the-art Electron Devices," John Wiley & Sons, 2013.

[2] K. Washio, "SiGe HBT and BiCMOS technologies for optical transmission and wireless communication systems," IEEE Trans. Electron devices, 50(3): 656-668, 2003.

[3] S. Athanasiou, C. A. Legrand, S. Cristoloveanu, P. Galy, "Novel ultrathin FD-SOI BIMOS device with reconfigurable operation," IEEE Trans. Electron Devices, 64(3): 916-922, 2017.

[4] J. H. Seo, K. Zhang, M. Kim, W. Zhou, Z. Ma, "High-performance flexible BiCMOS electronics based on single-crystal Si nanomembrane," NPJ Flexible Electron., 1(1): 1-7, 2017.

[5] J. Cai, T. H. Ning, "Bipolar transistors on thin SOI: concept, status and prospect," in Proc. 7th International Conference on Solid-State and Integrated Circuits Technology, 3: 2102-2107, 2004.

[6] K. Nadda, M. J. Kumar, "Vertical bipolar charge plasma transistor with buried metal layer," Sci. Rep., 5, 7860, 2015.

[7] S. Raman, P. Sharma, T. G. NeoGi, M. R. Leroy, R. Clark, J. F. McDonald, "On the performance of lateral SiGe heterojunction bipolar transistors with partially depleted base," IEEE Trans. Electron Devices, 62(8): 2377-2383, 2015.

[8] F. Bashir, S. A. Loan, M. Nizamuddin, H. Shabir, A. M. Murshid, M. Rafat, A. R. Alamoud, S. A. Abbasi, "A novel high performance nanoscaled dopingless lateral PNP transistor on silicon on insulator," in proc. IMECS, 2014.

[9] A. Sahu, L. K. Bramhane, J. Singh, "Symmetric lateral doping-free BJT: a novel design for mixed signal applications," IEEE Trans. Electron Devices, 63(7): 2684-2690, 2016.

[10] P. Agnihotri, P. Dhakras, J. U. Lee, "Bipolar junction transistors in two-dimensional WSe2 with large current and photocurrent gains," Nano lett., 16(7): 4355-4360, 2016.

[11] P. Chevalier, F. Gianesello, A. Pallotta, J. A. Goncalves, G. Bertrand, J. Borrel, L. Boissonnet, E. Brezza, M. Buczko, E. Canderle, D. Celi, "PD-SOI CMOS and SiGe BiCMOS technologies for 5G and 6G communications," in Proc. 2020 IEEE International Electron Devices Meeting (IEDM): 34-4, 2020.

[12] E. Preisler, "A commercial foundry perspective of SiGe BiCMOS process technologies," in Prpc. 2020 IEEE BiCMOS and Compound Semiconductor Integrated Circuits and Technology Symposium (BCICTS): 1-5, 2020.

[13] P. Chevalier, F. Gianesello, A. Pallotta, J. A. Goncalves, G. Bertrand, J. Borrel, L. Boissonnet, E. Brezza, M. Buczko, E. Canderle, D. Celi, "PD-SOI CMOS and SiGe BiCMOS Technologies for 5G and 6G communications," in Proc. 2020 IEEE International Electron Devices Meeting (IEDM): 34-4, 2020.

[14] L. Bramhane, S. Salankar, M. Gaikwad, M. Panchore, "Impact of work function engineering in charge plasma based bipolar devices," Silicon, 14: 3993-3997, 2022.

[15] A. Sahu, A. Kumar, S. P. Tiwari, "Exploration of logic gates and multiplexer using doping-free bipolar junction transistor," Solid-State Electron., 180, 107994, 2021.

[16] S. Zafar, M. A. Raushan, S. Ahmad, M. J. Siddiqui, "Reducing off-state leakage current in dopingless transistor employing dual metal drain," Semicond. Sci. Technol., 35(1): 015016, 2019.

[17] F. Raissi, M. Amirmazlaghani and A. Rajabi, "Field-effect BJT: an adaptive and multifunctional nanoscale transistor," Appl. Nanosci., 1-13, 2022.

[18] M. Amirmazlaghani, A. Rajabi, "New design of bipolar transistor with field-effect doping," in Proc. KBEI2017, Science and Technology university, 20DEC, 2017.

[19] F. Raissi, "A brief analysis of the field-effectdiode and breakdown transistor," IEEE Trans. Electron Devices, 43(2): 362-365, 1996.

[20] F. Raissi, M. Amirmazlaghani, A. Rajabi, "A multifunctional sub-10nm transistor," arXiv preprint arXiv:2108.00297, 2021.

[21] Q. Zhao, W. Sun, J. Zhao, L. Feng, X. Xu, W. Liu, X. Guo, Y. Liu, H. Yang, "Noise margin, delay, and power model for pseudo-CMOS TFT logic circuits," IEEE Trans. Electron Devices, 64(6): 2635-2642, 2017.

[22] S. H. C. Baek, K. W. Park, D. S. Kil, Y. Jang, J. Park, K. J. Lee, B. G. Park, "Complementary logic operation based on electric-field controlled spin–orbit torques," Nat. Electron., 1(7): 398-403, 2018.

[23] H. Elgabra, A. Siddiqui, S. Singh, "Design and simulation of a novel bipolar digital logic technology for a balanced performance in 4H-SiC," IEEE Electron Device Lett., 37(3): 257-260, 2016.

[24] A. Siddiqui, H. Elgabra, S. Singh, "Design considerations for 4H-SiC lateral BJTs for high temperature logic applications," IEEE J. Electron Devices Soc., 6: 126-134, 2017.

[25] X. Gao, C. Sui, S. Hemmady, J. Rivera, S. J. Yakura, D. Pommerenke, A. Patnaik, D. G. Beetner, "Modeling static delay variations in push–pull CMOS digital logic circuits due to electrical disturbances in the power supply," IEEE Trans. Electromagn. Compat., 57(5): 1179-1187, 2015.

[26] K. Nayak, M. Bajaj, A. Konar, P. J. Oldiges, K. Natori, H. Iwai, K. V. Murali, V. R. Rao, "CMOS logic device and circuit performance of Si gate all around nanowire MOSFET," IEEE Trans. Electron Devices, 61(9): 3066-3074, 2014.

[27] S. Guin, M. Sil, A. Mallik, "Comparison of logic performance of CMOS circuits implemented with junctionless and inversion-mode FinFETs," IEEE Trans. Electron Devices, 64(3): 953-959, 2017.

[28] T. K. Chiang, "A new device-physics-based noise margin/logic swing model of surrounding-gate MOSFET working on subthreshold logic gate," IEEE Trans. Electron Devices, 64(1): 306-311, 2016.

[29] B. J. Touchaei, N. Manavizadeh, "Design and simulation of low-power logic gates based on nanoscale side-contacted FED," IEEE Trans. Electron Devices, 64(1): 306-311, 2016.

[30] M. A. A. Hafiz, L. Kosuru, M.I. Younis, "Microelectromechanical reprogrammable logic device," Nat. commun., 7(1): 1-9, 2016.

[31] Y. Leblebici, S. M. Kang, "CMOS digital integrated circuits: analysis and design," McGraw-Hill, 1996.

[32] A. L. Zimpeck, C. Meinhardt, R. A. L Reis, "Impact of PVT variability on 20 nm FinFET standard cells," Microelectron. Reliab., 55(9-10): 1379-1383, 2015.

[33] S. Dubey, P. N. Kondekar, "Performance comparison of conventional and strained FinFET inverters," Microelectron. J, 55: 108-115, 2016.

[34] H. Vallabhaneni, A. Japa, S. Shaik, K. S. R. Krishna, R. Vaddi, "Designing energy efficient logic gates with Hetero junction Tunnel fets at 20nm," in Proc. 2014 2nd International Conference on Devices, Circuits and Systems (ICDCS): 1-5, 2014.

[35] U. Mushtaq, V. K. Sharma, "Performance analysis for reliable nanoscaled FinFET logic circuits," Analog Integr. Circuits Signal Process., 107(3): 671-682, 2021.

[36] M. K. Q. Jooq, A. Mir, S. Mirzakuchaki, A. Farmani, "Semi-analytical modeling of high performance nanoscale complementary logic gates utilizing ballistic carbon nanotube transistors," Physica E, 104: 286-296, 2018.

[37] N. Grag, Y. Pratap, M. Gupta, s. Kabra, "Impact of different localized trap charge profiles on the short channel double gate junctionless nanowire transistor based inverter and Ring Oscillator circuit," AEU Int. J. Electron. Commun., 108: 251-261, 2019.

[38] S. K. Shelke, V. N. Nitnaware, S. Rode, "Design of ring oscillator with better temperature, supply voltage & process stability with 32nm FDSOI transistor for ISM band application," in Proc. 2017 11th International Conference on Intelligent Systems and Control (ISCO): 311-313, 2017.

[39] A. Baidya, T. R. Lenka, S. Baishya, "Application of 3D double gate Junctionless transistor for ring oscillator," in Proc. 2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE): 1-4, 2016.

[40] S. Salem, M. Tajabadi, M. Saneei, "The design and analysis of dual control voltages delay cell for low power and wide tuning range ring oscillators in 65 nm CMOS technology for CDR applications," AEU Int. J. Electron. Commun., 82: 406-412, 2017.

[41] M. Abou Chahine, H. Bazzi, A. Mohsen, A. Harb, A. Kassem, "A low-noise voltage-controlled ring oscillator in 28-nm FDSOI technology for UWB applications," AEU Int. J. Electron. Commun., 97: 94-101, 2018.

[42] J. C. Chien, P. Upadhyaya, H. Jung, S. Chen, W. Fang, A. M. Niknejad, J. Savoj, K. Chang, "2.8 A pulse-position-modulation phase-noise-reduction technique for a 2-to-16GHz injection-locked ring oscillator in 20nm CMOS," in Proc. 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC): 52-53, 2014.

[43] P. K. Pal, B. K. Kaushik, S. Dasgupta, "Investigation of symmetric dual-k spacer trigate FinFETs from delay perspective," IEEE Trans. Electron Devices, 61(11): 3579-3585, 2014.

[44] M. R. Tripathy, A. K. Singh, A. Samad, S. Chander, K. Baral, P. K. Singh, S. Jit, "Device and circuit-level assessment of GaSb/Si heterojunction vertical tunnel-FET for low-power applications," IEEE Trans. Electron Devices, 67(3): 1285-1292, 2020.

[45] S. Seifollahi, S. A. S. Ziabari, A. Kiani-sarkaleh, "A design of nanoscale double-gate FET based ring oscillator with improved oscillation frequency using device engineering," AEU Int. J. Electron. Commun., 134, 153701, 2021.

[46] H. Lu, P. Paletti, W. Li, P. Fay, T. Ytterdal, A. Seabaugh, "Tunnel FET analog benchmarking and circuit design," IEEE J. Explor. Solid-State Comput. Devices Circuits, 4(1): 19-25, 2018.

[47] M. Y. Kao, G. Pahwa, A. Dasgupta, S. Salahuddin, C. Hu, "Analysis and modeling of polarization gradient effect on negative capacitance FET," IEEE Trans. Electron Devices, 67(10): 4521-4525, 2020.

Biographies

Alireza Shokri received M.S. degree from Shahid Rajaee Teacher Training University, Iran in 2020, micro and nano electronics engineering. He is currently a Ph.D. student in Department of Electrical engineering, Shahid Rajaee Teacher Training University, Tehran, Iran, where he is doing research on betavoltaic batteries. His research interest includes nano scale transistors and betavoltaic cell.

- Email: arshokri@sru.ac.ir
- ORCID: 0000-0001-8766-6486
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Mina Amirmazlaghani received her Ph.D. in the field of Nanoelectronics from K.N.Toosi University, Tehran, Iran, in 2014. Her PhD thesis was about "Design and fabrication of Graphene-based IR and THz detectors". During 2012, she was a visiting researcher at TML (Terahertz and Millimeter wave laboratory), MC2, Chalmers University. She was with AIST, Tsukuba, Japan, early in 2014. From 2014, she has been an assistant professor at electronics department of Shahid Rajaee University in Tehran, Iran where she has established Nanoelectronics Lab#1 and #2, for simulation and fabrication process, respectively.
Her current research interests include Graphene-Based Electronics, Design and Modeling of Nano-Scale Semiconductor Devices, Design and Fabrication of IR and THz Detectors, Beta-cell Batteries based on semiconductors and High Frequency Electronics.

- Email: m.mazlaghani@sru.ac.ir
- ORCID: 0000-0003-4235-3245
- Web of Science Researcher ID: AHC-9391-2022
- Scopus Author ID: NA
- Homepage: https://www.sru.ac.ir/en/faculty/school-of-electrical-engineering/mina-amir-mazlaghani/

How to cite this paper:
A. Shokri, M. Amirmazlaghani, "Feasibility of Digital Circuit Design Based on Nanoscale Field-Effect Bipolar Junction Transistor," J. Electr. Comput. Eng. Innovations, 11(1): 33-40, 2023.

DOI: 10.22061/JECEI.2022.8287.503

URL: https://jecei.sru.ac.ir/article_1712.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Fast and Power Efficient Signed/Unsigned RNS Comparator & Sign Detector

***Z. Torabi*[1,*], *A. Belghadr*[2]**

[1]*Faculty of Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.*

[2]*Department of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran.*

Article Info

Article History:
Received 15 January 2022
Reviewed 15 March 2022
Revised 18 April 2022
Accepted 25 may 2022

Keywords:
Computer arithmetic
Residue number system
Complicated operations
Signed number comparison
Dynamic range partitioning

*Corresponding Author's Email Address: z.torabi@sru.ac.ir

Abstract

Background and Objectives: Residue number system (RNS) is considered as a prominent candidate for high-speed arithmetic applications due to its limited carry propagation, fault tolerance and parallelism in "Addition", "Subtraction", and "Multiplication" operations. Whereas, "Comparison", "Division", "Scaling", "Overflow Detection" and "Sign Detection" are considered as complicated operations in residue number systems, which have also received a surge of attention in a multitude of publications.
Efficient realization of Comparators facilitates other hard-to-implement operations and extends the spectrum of RNS applications. Such comparators can substitute the straightforward method (i.e. converting the comparison operands to binary and comparing them with wide word binary comparators) to compare RNS numbers.
Methods: Dynamic Range Partitioning (DRP) method has shown advantages for comparing unsigned RNS numbers in the 3-moduli sets $\{2^n, 2^n \pm 1\}$ and $\{2^n, 2^n - 1, 2^{n+1} - 1\}$, in comparison with other methods. In this paper, we employed DRP components and designed a unified unit that detects the sign of operands and also compares numbers, for the 5-moduli set $\gamma = \{2^{2n}, 2^n \pm 1, 2^n \pm 3\}$. This unit can be used for comparison of signed and also unsigned RNS numbers in the moduli set γ.
Results: Synthesized comparison results reveal 47% (54%) speed-up, 35% (32%) less area consumption, 25% (24%) lower power dissipation, and 60% (65%) less energy for $n = 8$ (16) in comparison to the straightforward signed comparator.
Conclusion: According to the results of this study, DRP method for sign detection and comparison operations outperforms other methods in different moduli sets including 5-moduli set $\gamma = \{2^{2n}, 2^n \pm 1, 2^n \pm 3\}$.

Introduction

Nowadays, with the increased versatility of electronic products, high-performance computations with low-power consumption are of vital importance. Residue number system has offered the advantage of high-speed and low-power addition, subtraction, and multiplication operations, and thus it has received much attention for high-throughput computations, particularly in digital signal processing [1], data transmission [2], cryptography [3], steganography [4], and image processing [5].

Residue Number System (RNS) is a number system with k integer modulus $\{m_1, m_2, \dots, m_k\}$. A number X is represented as $(x_1, x_2, \dots, x_k)$, where $x_i = |X|_{m_i}$ (i.e., the remainder of integer division $\frac{X}{m_i}$). Cardinality of the residue number system is maximized (i.e., $M = m_1 \times \dots \times m_k$), where the moduli are pair-wise prime. In RNS,

which is an unweighted number system, some arithmetic operations such as division, scaling, comparison and sign/overflow detection are difficult to implement. Whereas, these complicated operations are fundamental to develop processors with practical interest. For example, comparison, sign and overflow detection are essential for some nonlinear procedures, such as median and rank-order filtering [6].

Sign detection is needed in applications dealing with positive and negative numbers. In such cases, dynamic range (i.e., M) is partitioned into two parts of $[0, \lfloor M/2 \rfloor)$ and $[\lfloor M/2 \rfloor, M)$ in order to represent positive and negative numbers, respectively. The straightforward sign detection method in RNS is based on converting the operand to binary format and then comparing it with $\frac{M}{2}$.

Comparison plays a crucial role in the development of division and overflow/sign detection units in RNS, therefore an efficient comparison method would be cost-effective to implement other complicated operations [7]-[9]. Contrary to the parallelism that residue number system offers to the addition and multiplication, no parallel RNS comparison scheme can be envisaged via independent modular comparator in concurrent residue channels. For example, in the moduli set {64,7,9,11,5}, 44352=(0,0,0,0,2) is greater than 6=(6,6,6,6,1), which is not clearly apprehended from their modular representations.

The RNS comparison schemes proposed so far [6], [10]-[16] can be categorized into a conversion-based method [6], [10], [11], [15], parity checking technique [12], [14], and mapping function [13], [16]-[23] that will be described in second Section. For RNS unsigned number comparison in 3-moduli sets, Dynamic Range Partitioning (DRP) method [17] yields the best performance [17], [18]. However, we have not encountered any DRP-based RNS comparator for moduli sets with more than three moduli.

In many RNS applications, domain of numbers is expanded. Utilizing wider moduli and increasing the number of moduli, are two different ways to fulfill the need for expanded range of numbers, while both of them make reverse conversion and complex operations more complicated. However, since the conversion process is not frequent, the burden of a lengthier reverse conversion for moduli sets with more than three moduli is bearable [18]. Several moduli sets with four to eight moduli have been reported in the literature. For example moduli set $\gamma = \{2^{2n}, 2^n \pm 1, 2^n \pm 3\}$ with $6n$-bit dynamic range whereas its signed/unsigned reverse converter have been introduced in [24], [25].

In this paper, we focus on the realization of a DRP-based sign detector and comparator for the moduli set γ. To this aim, we convert the 5-residue operands of γ to an equivalent 3-moduli set $\{2^{2n}, 2^{2n}-1, 2^{2n}-9\}$, where the DRP can be applied. For evaluation of the proposed comparator, we have not found any hardware realization of a comparator for γ, thus, we compare our method with the straightforward comparators [24], [25] and one of the recent previous general comparators [24].

We also compare the proposed sign detection method with the γ-sign-detection unit of [25]. The proposed work has considerable merits on the reference works [24], [25], in terms of latency, area, and power, presented through analytical and synthesized evaluations.

The rest of the paper is organized as follows. The second section reviews briefly different RNS comparison and sign detection methods. In the third section, the new sign detector and signed/unsigned comparator for γ are proposed, while its implementation scrutiny is discussed in the fourth section.

Evaluations are found in the fifth section and finally in the last section we draw our conclusions.

Background Materials and Related Works

In this section, we describe the representation of signed numbers in RNS, sign identification methods, and then review a number of comparison methods briefly.

In RNS, numbers are defined as positive integers in the range between [0, M−1], but in applications with signed numbers, as shown in Fig. 1, dynamic range is divided into two parts, positive and negative numbers. The sign of an RNS number X can be detected by (1).

$$\text{Sign}(X) = \begin{cases} 0 & if \quad 0 \le X < \lfloor M/2 \rfloor \\ 1 & if \quad \lfloor M/2 \rfloor \le X < M \end{cases} \tag{1}$$

Sign(X) usually indicates by the most significant bit (MSB) of X, therefore in fast and low power sign detection methods, before complete conversion of the operand to binary format via mixed radix representation (MRC) [26] or Chinese remainder theorem (CRT) [26], MSB of the operand is extracted.

In [27] and [28], with the usage of last MRC digit, MSB of the operand and consequently sign bit extracted. In [25] a sign detection unit and signed reverse converter is proposed for γ, based on CRT.

A wide variety of techniques have been proposed for RNS comparison in the literature [6], [9]-[23], some of which are summarized in Table 1. Most of the comparison methods compare two unsigned numbers and cannot be easily extended to compare signed RNS numbers due to the complexity of sign detection process.

In conversion-based methods [6], [9]-[11], [15], before full reverse conversion, comparison takes place. Comparing the corresponding MRC digits [26] or New CRT coefficients [29] fall into this category.

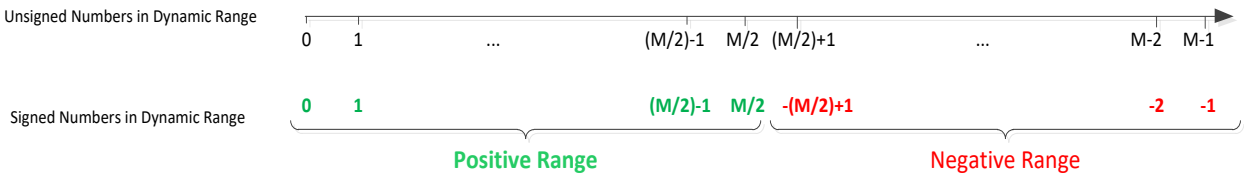


Fig. 1: Distribution of positive and negative numbers in dynamic range.

In parity checking technique [12], comparison is based on the parity of the operands and their difference. One of the major drawbacks of this method is that it is applicable only on moduli sets which do not have even moduli, while in practice numerous moduli sets comprise at least a power-of-two modulo, owing to an efficient arithmetic channel realizations.

In the mapping technique [13], a number is assigned to each RNS number in the dynamic range. For comparing two numbers X and Y, $D(X)$ and $D(Y)$ are compared, such that $D(X) > D(Y)$ leads to $X > Y$. This method, similar to the CRT, is based on a large modulo SQ operation, where $SQ = \sum_{i=1}^{n}(M/m_i)$. Since direct implementation of diagonal function is not efficient for comparing two RNS numbers, some modifications for diagonal function computation were proposed [23], [30]. In [23], $D(X)$ is computed in modulo 2^u, where $u = \log(m_n - 1)SQ$ and m_n is the largest modulo in the moduli set. Although 2^u is smaller than SQ, in comparison to other methods, [23] still needs computation in the large module 2^u.

Efficient computations of diagonal function results in introducing new moduli sets that allow for efficient hardware implementation of $D(X)$. Some algorithms were introduced in [30] to generate 3- and 4- moduli sets in such a way that $SQ = 2^v$ and $SQ = 2^v - 1$, respectively, for some v. In [31], similar to [30], several methods proposed to design moduli sets with SQ forms 2^n, $2^n - 1$, and 2^n+1.

In [16], [19], for implementing non-modular operations including comparison, sign detection, division, and scaling, the authors proposed a method to compute the interval evaluation of $X = (x_1, x_2, \dots, x_k)$. Such computations are performed in limited precision of fractional representation of X.

Ambiguity cases arise when X is very small or big, in such cases MRC digits were used for non-modular operations in this method which leads to sequential computations.

In [20], [21], dynamic range $[0, M)$ is divided into $M_k = m_1 \times m_2 .. \times m_{k-1}$ intervals. With a large amount of computations, the numerical intervals which contain X and Y are determined and after that, comparison can be done by comparing numerical intervals of X and Y.

Minimum-range monotonic core function is proposed in [22] which is a modification of core function [32].

In this solution, comparison of every two number is carried out through comparing their core functions. In [22], core function is monotonic and computed in module M_k. They also show that diagonal function is a special case of core function.

DRP [17], divides the dynamic range of any 3-moduli set into m_1 partitions of size $m_2 \times m_3$, where each partition is divided into m_2 sections of size m_3. For any moduli set $\{m_1, m_2, m_3\}$, DRP components (i.e. $p_1(X)$ and $p_2(X)$), are defined in (2), where $x_{23} = |X|_{m_2 m_3}$, $x_2 = |X|_{m_2}$, $x_3 = |X|_{m_3}$ and $M_1 = m_2 \times m_3$. $p_1(X)$ and $p_2(X)$ are the number of partition and section that are computed for an RNS number X, respectively.

$$\begin{cases} p_1(X) = \left|\left|{M_1}^{-1}\right|_{m_1}(x_1 - x_{23})\right|_{m_1} \\ p_2(X) = \left|\left|{m_3}^{-1}\right|_{m_2}(x_2 - x_3)\right|_{m_2} \end{cases} \quad (2)$$

Comparison of two numbers $X = (x_1, x_2, x_3)$ and $Y = (y_1, y_2, y_3)$ can be reduced to the comparison of $[p_1(X), p_1(Y)]$, $[p_2(X), p_2(Y)]$, $[x_3, y_3]$ in three different comparators.

Sign detection and signed number comparison of [6] for the moduli set $\{2^n - 1, 2^{n+x}, 2^n + 1\}$ are based on an optimized version of the MRC. It performs the comparison through utilizing the sign bits of comparison operands and their difference. In this method, the sign of RNS numbers can be identified by comparing the third MRC digit with 2^{n+k-1}.

Proposed Sign Detector and Comparator

In this section, a new DRP-based method is derived for sign detection and comparing two RNS numbers X and Y. As mentioned earlier, DRP has been utilized in unsigned numbers comparison methods [17], [18]. However, in this paper, DRP is applied to sign identification (Theorem 1) and comparison for the 5-moduli set γ.

The above DRP scheme (2) for 3-moduli RNS comparison can be extended to 5-moduli cases. In fact, the aforementioned 5-moduli set γ, can be reduced to the 3-moduli set $\tau = \{2^{2n}, 2^{2n} - 1, 2^{2n} - 9\}$, where the conjugate moduli $2^n \pm 1$ and $2^n \pm 3$, are combined to moduli $2^{2n} - 1$ and $2^{2n} - 9$ through two simple reverse conversion operations.

Table 1: Comparison of 10 previous RNS comparators

Ref.	Category	Moduli set	Signed / Unsigned numbers	Method
[9]	Reverse conversion	$\{2^n \pm 1, 2^n, m\}$, $m \in \{2^{n+1} \pm 1, 2^{n-1}-1\}$	Unsigned	New CRT
[17]	Mapping function	$\{2^n \pm 1, 2^n\}$	Unsigned	DRP
[18]	Mapping function	$\{2^n-1, 2^n, 2^{n+1}-1\}$	Unsigned	DRP
[10]	Reverse conversion	$\{2^n \pm 1, 2^n\}$	Unsigned	MRC-CRT
[11]	Reverse conversion	Arbitrary moduli set	Unsigned	New CRT
[12]	Parity checking	Odd moduli set	Unsigned	Parity checking
[13], [23]	Mapping function	Arbitrary moduli set	Unsigned	Diagonal mapping
[14]	Parity checking	$\{2^n \pm 1, 2^{n+1} \pm 1\}$	Unsigned	parity checking
[16], [19]	Mapping function	Arbitrary moduli set	Unsigned	floating-point interval evaluation, MRC
[20], [21]	Mapping function	Arbitrary moduli set	Unsigned	interval evaluation
[22]	Mapping function	Arbitrary moduli set	Unsigned	Core function
[6]	Reverse conversion	$\{2^n-1, 2^{n+x}, 2^n+1\}$	Signed	MRC
[15]	Reverse conversion	$\{2^{n+k}, 2^n \pm 1, 2^{n\pm1}-1\}$	Signed	MRC

Therefore the 3-moduli DRP method can be applied to the new 3-moduli set. Here we compute DRP components for the new 3-moduli set τ. Prior to that, the required multiplicative inverses are described as β_1, β_2 and β_3.

Property 1: $\beta_1 = |(2^n+3)^{-1}|_{2^n-3}$

$$= \begin{cases} \frac{2^{n-1}-1}{3} & n = 2k+1 \\ -\frac{2^{n-1}-2}{3} & n = 2k \end{cases}$$

Property 2: $\beta_2 = |(2^{2n}-9)^{-1}|_{2^{2n}-1} = -2^{2n-3}$

Property 3: $\beta_3 = |((2^{2n}-9)(2^{2n}-1))^{-1}|_{2^{2n}}$

$$= \begin{cases} \frac{2^{2n+3}+1}{9} & n = 3p \\ \frac{2^{2n+1}+1}{9} & n = 3p+1 \\ 2^{2n-1} + \frac{2^{2n-1}+1}{9} & n = 3p+2 \end{cases}$$

Let $m_1 = 2^{2n}, m_2 = 2^n-1, m_3 = 2^n+1, m_4 = 2^n-3, m_5 = 2^n+3$ and the corresponding residues of an operand X for the new moduli set τ based on CRT and New CRT be denoted as (x_1, x_{23}, x_{45}) where $x_1 = |X|_{2^{2n}}$, $x_{23} = |X|_{2^{2n}-1} = |x_3 + (2^n+1)2^{n-1}(x_2 - x_3)|_{2^{2n}-1}$ and $x_{45} = |X|_{2^{2n}-9} = x_5 + (2^n+3)\,|\beta_1(x_4 - x_5)|_{2^n-3}$.

In the following Eqns. 3 and 4, we derive $p_2(X)$ and $p_1(X)$ as DRP components in moduli set τ, based on Eqn. set 2, where $x_{2345} = |X|_{(2^{2n}-9)(2^{2n}-1)} = x_{45} + (2^{2n}-9)|(2^{2n}-9)^{-1}(x_{23} - x_{45})|_{2^{2n}-1}$.

$$p_2(X) = |\beta_2(x_{23} - x_{45})|_{2^{2n}-1}$$
$$= |2^{2n-3}(-x_{23} + x_{45})|_{2^{2n}-1} \quad (3)$$

$$p_1(X) = |((2^{2n}-9)(2^{2n}-1))^{-1}(x_1 - x_{2345})|_{2^{2n}}$$
$$= \left|\beta_3\left(x_1 - x_{45} + 9p_2(X)\right)\right|_{2^{2n}} \quad (4)$$

Theorem 1: X in the moduli set γ is negative if and only if $\mathrm{MSB}(p_1(X)) = 1$.

Proof: Based on the DRP method [8], in the moduli set τ we have $X = p_1(X)M_1 + x_{23} = p_1(X)(2^{2n}-1)(2^{2n}-9) + x_{23}$ and $p_1(X) < 2^{2n}$. With consideration of $\frac{M}{2} = 2^{2n-1}(2^{2n}-1)(2^{2n}-9)$, our proof consists of two parts as follows:

a. $(\mathrm{MSB}(p_1(X)) = 1) \Longrightarrow X \geq \frac{M}{2}$

If $MSB(p_1(X)) = 1 \Longrightarrow p_1(X) \geq 2^{2n-1} \Longrightarrow X \geq 2^{2n-1}(2^{2n}-1)(2^{2n}-9) \Longrightarrow X$ is negative.

b. $X \geq \frac{M}{2} \Longrightarrow (\mathrm{MSB}(p_1(X)) = 1)$

let $x_{23} = 2^{2n}-2$ to find the minimum value of $p_1(X)$, where X is negative. The following condition must hold:

$$p_1(X)(2^{2n}-1)(2^{2n}-9) + 2^{2n} - 2 \geq 2^{2n-1}(2^{2n}-1)(2^{2n}-9)$$

which leads to $p_1(X) \geq 2^{2n-1}$ and $MSB\big(p_1(X)\big) = 1$. ■

Therefore by implementing one of the DRP components (i.e., $p_1(X)$), the sign of an RNS number (i.e., $sign(X)$) in the moduli set γ is identified. For comparing two signed RNS numbers, which belong to the same range and both have the same sign (positive or negative), comparing them without considering their signs determines the result. Therefore, for comparing two RNS numbers X and Y, first the signs of operands are identified. If only one of them is positive, the result of comparison is clear, whereas both of them are positive or negative, comparison is undertaken via DRP components (i.e. $p_1(X), p_1(Y), p_2(X)$ and $p_2(Y)$). Comparison can be reduced to the comparison of $p_1(X)$ and $p_1(Y)$. In the case of $p_1(X) = p_1(Y), p_2(X)$ and $p_2(Y)$ are compared. If $p_1(X) = p_1(Y)$ and $p_2(X) = p_2(Y)$, comparison of x_{45} and y_{45} yields the final result. Flowchart of the proposed comparator is illustrated in Fig. 2.

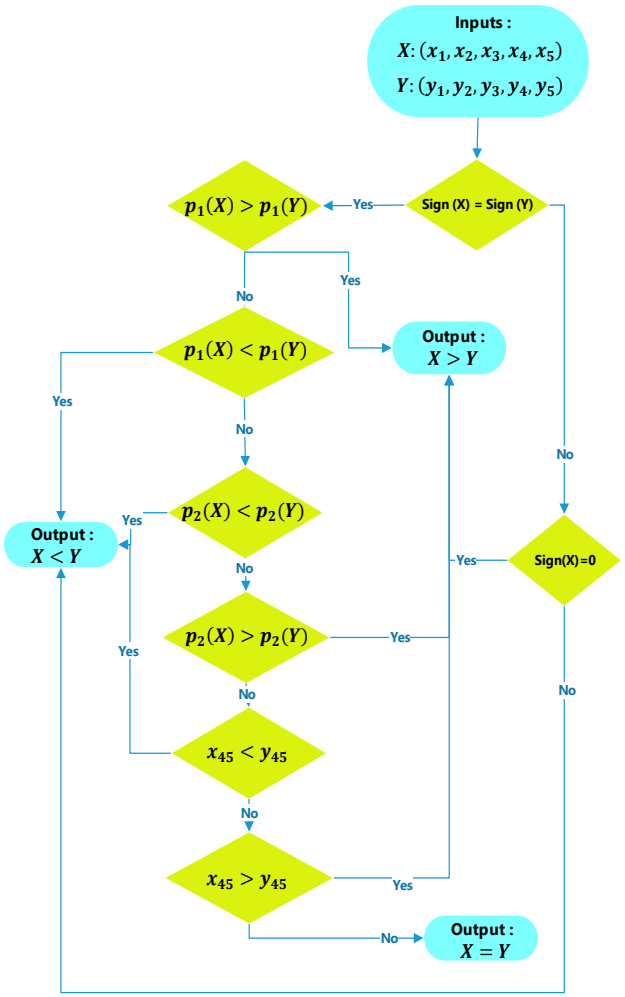


Fig. 2: Algorithm of the proposed comparator.

Since sign detection is performed with $p_1(X)$, and it is also required for comparison, with eliminating first step of Fig. 2 (comparing sign(X) and sign (Y)), it can be used for unsigned comparison. The overall architecture for signed/unsigned comparator is visualized by Fig. 3 where E and C show that $X = Y$ and $X > Y$ respectively.

Example 1. Consider $\gamma = \{256, 15, 17, 13, 19\}$ with $n = 4$. Let $X = 1 = (1,1,1,1)$ and $Y = 1000000 = (64,10,9,1,11)$ be two RNS numbers to be compared. The equivalents of X and Y in the corresponding moduli set $\tau = \{256, 255, 247\}$ are $(1,1,1)$ and $(64,145,144)$ respectively. Based on Eqns. 3 and 4, $p_1(X) = p_2(X) = 0$, $p_1(Y) = 15$ and $p_2(Y) = 223$. According to the Theorem 1 and Fig. 2, both X and Y are positive and $p_1(Y) > p_1(X)$ so $Y > X$.

Implementation

Sign detection and comparator units in the proposed work are based on DRP components, therefore, in this section, we provide the implementation details of $p_1(X)$ and $p_2(X)$ generators. Here with the assumption of $n = 3p + 1$ and usage of the properties 1-3, we investigate implementation-friendly equations for $p_1(X)$ and $p_2(X)$. Computation and implementation of DRP components with $n \neq 3p + 1$ are quite similar.

$$p_1(X) = \left| \frac{2^{2n+1}+1}{9} \left(\begin{matrix} -(2^n+3) \left| \frac{2^{n-1}-1}{3}(x_4 - x_5) \right|_{2^n-3} \\ +x_1 - x_5 + 9\, p_2(X) \end{matrix} \right) \right|_{2^{2n}} \quad (5)$$

$$p_2(X) = \left| 2^{2n-3} \left(\begin{matrix} (2^n+3) \left| \frac{2^{n-1}-1}{3}(x_4 - x_5) \right|_{2^n-3} \\ x_5 - x_3 - (2^n+1)2^{n-1}(x_2 - x_3) \end{matrix} \right) \right|_{2^{2n}-1} \quad (6)$$

Replacing $-x_3 = \overline{x_3} - 2^{n+1} + 1$, $-x_2 = \overline{x_2} - 2^n + 1$, $-x_5 = \overline{x_5} - 2^{n+1} + 1$, $U = \left| \frac{2^{n-1}-1}{3}(x_4 - x_5) \right|_{2^n-3} = \left| \sum_{i=0}^{i=\frac{n-3}{2}} 2^{2i}\,(x_4 + \overline{x_5}) - 5 \times \frac{2^{n-1}-1}{3} \right|_{2^n-3}$ and $-U = \overline{U} - 2^n + 1$ in (5) and (6) result (7) and (8), respectively.

$$p_1(X) = \left| \frac{2^{2n+1}+1}{9}(x_1 + \overline{x_5} + (2^n+3)\overline{U} \right. \left. \begin{matrix} \\ -2^{n+2} + 4) + p_2(X) \end{matrix} \right|_{2^{2n}} \quad (7)$$

$$p_2(\mathrm{X}) = \left| \begin{matrix} 2^{2n-3}x_5 + (2^{n-3} + 3 \times 2^{2n-3})\, U + 2^{2n-3}\overline{x_3} + \\ 2^{2n-3} - 2^{n-2} + (2^{2n-4} + 2^{n-4})(x_3 + \overline{x_2}) \end{matrix} \right|_{2^{2n}-1} \quad (8)$$

One $(n-1,\ 2^n - 3)$ multi operand modular adder (MOMA) [33] followed by an n-bit modular adder is required to generate U expression. Based on (8), after computation of U, $p_2(X)$ is obtained with a two-level CSA followed by a $2n$-bit modular adder. In parallel with $p_2(X)$, $\frac{2^{2n+1}+1}{9}(x_1 + \overline{x_5} + (2^n+3)\overline{U} - 3 \times 2^n + 4)$ is being obtained through a $(2n-4\ ,2^{2n})$ MOMA. The required architecture for generation of $p_1(X)$ and $p_2(X)$ is depicted in Fig. 4.

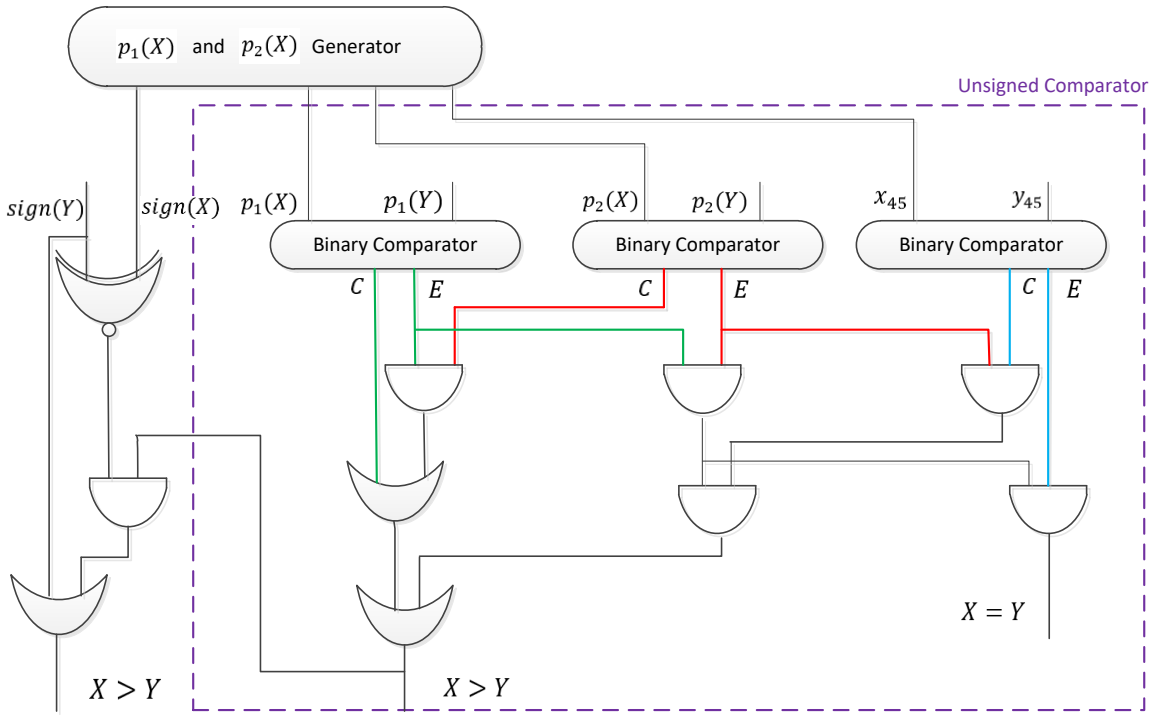


Fig. 3: The proposed signed/unsigned γ −comparator.

Evaluation

In this section, we present the performance evaluation of the proposed design and compare it with previous related works.

The proposed design consists of sign detection module and comparator. In the literature, two reverse converters [24], [25] and one sign identifier [25] designed for moduli set γ.

In [25], a $\gamma -$ signed reverse converter proposed wherein the sign of operand is extracted in the middle of conversion. In the case of negative sign, the output of reverse converter should be added to 2's complement of M.

In [24] a γ −reverse converter proposed which is based on New CRT [27] and the output is positive number in the range $[0\ M)$.

We evaluate the proposed comparator against a straightforward comparator which consists of two reverse converters for converting the operands to binary format and a binary comparator for comparing two operands. Moreover, we evaluate unsigned general comparators of [16], [19]-[23], which are based on mapping function that has recently received attentions in literature.

In addition, we evaluate the proposed sign detection method with sign detection module of [25] and straightforward sign detection method of [24] (i.e., Conversion of operand to binary format and comparing it with $\frac{M}{2}$).

The delay and cost measures of the proposed comparator and sign detector are compiled in Tables 2 and 3, based on the unit gate model [34].

In our analytical evaluations, the cost and delay of one simple 2-input logic gate (e.g., AND, OR, NAND, NOR) are considered as 1 unit of cost ($\#G$) and delay (ΔG). For example, delay and cost of an n-bit carry ripple adder is assumed to be $2n\Delta G$ and $7n\#G$ respectively. The comparators of [28], [29] have less delay in return of extra cost.

Between general comparators described in Tables 2 and 3 (i.e., [16], [19]-[23]), the comparators proposed in [22] and [27] have reasonable delay and cost.

So as to find better insight into merits of the proposed design, we have synthesized the proposed comparator and γ −comparators of [24], [25], and comparator of [22] in case of $n = 8$ and $n = 16$, with the TSMC 90nm CMOS standard logic cell library by Synopsys Design Compiler. Synthesized results are compiled in Table 4 which approve superiority of the proposed comparator in comparison with the reference designs, in terms of delay, area, power and energy.

Based on the results of Table 4, the ratios of delay and power ($n = 8$) of straightforward signed comparator are higher than other methods.

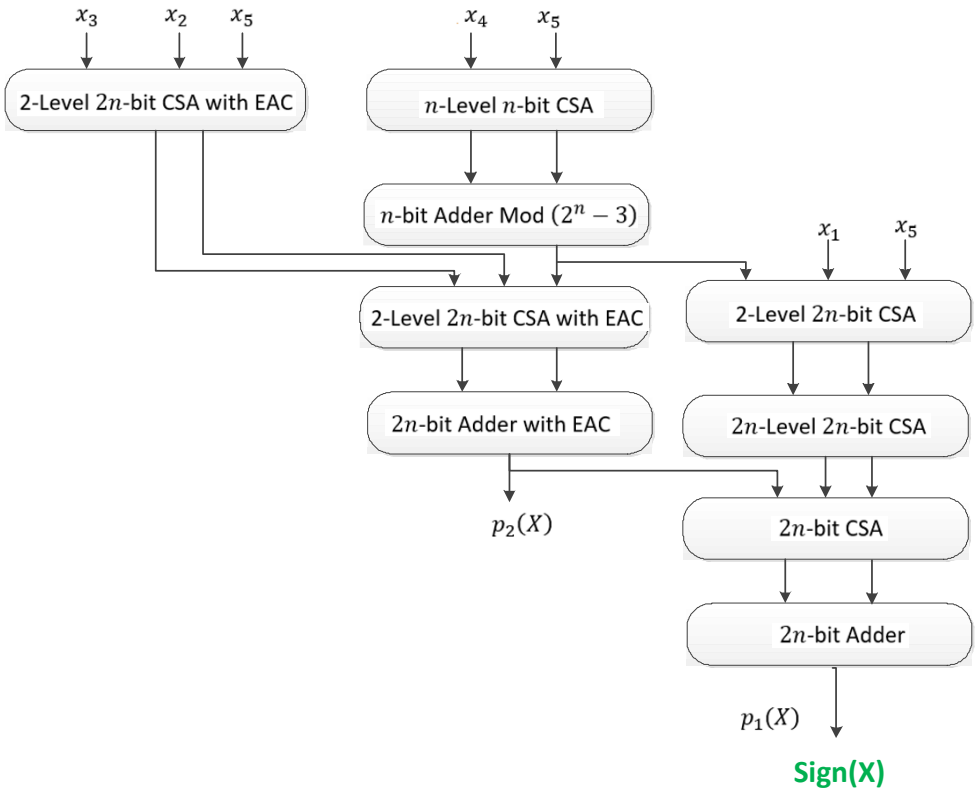


Fig. 4: $p_1(X)$ and $p_2(X)$ generator and sign detection circuit.

Table 2: Analytical delay comparison

Operation	Method	Adder					CSA	Comparator				Total Delay
		n-bit	$2n$-bit	$4n$-bit	$6n$-bit	$7n$-bit		n-bit	$2n$-bit	$6n$-bit	$7n$-bit	
Sign Detection	[24]		3	1			$3\log n + 1$			1		$(32n + 12\log n + 4)\Delta G$
	[25]	1	4	1	1		$\log n + 5$					$(38n + 4\log n + 20)\Delta G$
	proposed	1	2				$\log 2n + \log n + 3$					$(10n + 8\log n + 16)\Delta G$
Comparison	[24]		3	1			$3\log n + 1$			1		$(32n + 12\log n + 4)\Delta G$
	[25]	1	4	1	1		$\log n + 5$			1		$(50n + 4\log n + 20)\Delta G$
	proposed	1	2				$\log 2n + \log n + 3$		1			$(14n + 8\log n + 16)\Delta G$
	[23]					1	$\log 2n$				1	$(28n + 8\log n)\Delta G$
	[16], [19]	4	1				$12\log n$	1	1			$(18n + 48\log n)\Delta G$
	[22]			1			$\log n$		2			$(16n + 4\log n)\Delta G$
	[20], [21]		2				$\log n$		1			$(12n + 4\log n)\Delta G$

Table 3: Analytical cost comparison

Operation	Method	Adder					CSA	Comparator				Total Cost
		n -bit	$2n$- bit	$4n$- bit	$6n$- bit	$7n$ -bit	n-bit	n -bit	$2n$ - bit	$6n$- bit	$7n$ - bit	
Sign Detection	[24]	1	5	1			$8n+6$			1		$(56n^2+273n)\#G$
	[25]	1	4	1			$5n+42$					$(35n^2+141n)\#G$
	proposed	1	2				$5n+14$					$(35n^2+133n)\#G$
Comparison	[24]	1	5	1			$8n+6$			3		$(56n^2+357n)\#G$
	[25]	1	4	1	2		$5n+42$			1		$(35n^2+267n)\#G$
	proposed	1	2				$5n+14$		3			$(35n^2+175n)\#G$
	[23]					1	$10n$				1	$(70n^2+98n)\#G$
	[16], [19]	4	1	1			$10n$	4	3			$(70n^2+140n)\#G$
	[22]			1			$6n$		4			$(42n^2+84n)\#G$
	[20], [21]	$4(2^{2n}-1)(2^{2n}-9)+5$					$(2^{2n}-1)(2^{2n}-9)n$	$4(2^{2n}-1)(2^{2}$ $-9)$				$(2^{2n}-1)(2^{2n}-9)(14n)\#G$

Table 4: Synthesis based comparison results

Design	n	Delay (ns)	Ratio	Area (μm^2)	Ratio	Power (mW)	Ratio	Energy (pJ)	Ratio
[24]	8	10.80	1.56	84075.61	2.22	69.74	2.21	720.79	3.31
[25]	8	13.20	1.91	58972.18	1.56	42.15	1.33	556.38	2.55
[22]	8	12.70	1.84	212472.50	5.62	110.85	3.51	1407.79	6.46
proposed	8	6.90	1.00	37800.76	1.00	31.54	1.00	217.62	1.00
[24]	16	16.30	1.58	189451.71	1.32	191.22	1.25	3116.88	1.98
[25]	16	22.70	2.20	210462.79	1.47	200.03	1.31	4540.68	2.88
[22]	16	14.8	1.43	350311.09	2.45	178.43	1.16	2640.76	1.67
proposed	16	10.30	1.00	142929.73	1.00	152.81	1.00	1573.94	1.00

Conclusion

In residue number systems, one of the most complicated operations are sign detection and comparison which also play a prominent role in the development of division and overflow detection components in RNS. The 5-moduli set $\gamma = \{2^{2n}, 2^n \pm 1, 2^n \pm 3\}$, has been shown to have efficient RNS arithmetic circuits as well as efficient reverse converter. To extend applicability of this moduli set, we provided the first efficient signed/unsigned RNS comparator circuit in this work.

In the proposed comparator, with the advantage of dynamic range partitioning technique, sign of the operands are identified and then comparison performed effectively. Synthesis-based results confirmed analytical

evaluation and revealed 47% (54%), 35% (32%), 25% (24%) and 60% (65%) delay, area, power, and energy improvements, respectively, for the new signed RNS number comparator in comparison with the reference design.

As regards the relevant future work, we plan to apply DRP method to other 4- and 5-moduli sets, to improve comparison operation and so other complicated operations.

Author Contributions

Zeinab Torabi contributed to the idea, simulation, writing, review, and editing the paper. Armin Belghadr contributed for writing, review, and editing the paper.

Acknowledgment

The authors would like to thank the editor and anonymous reviewers.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

RNS	Residue Number System
MRC	Mixed Radix Representation
CRT	Chinese Remainder Theorem
DRP	Dynamic Range Partitioning
MSB	Most Significant Bit

References

[1] G. C. Cardarilli, L. D. Nunzio, R. Fazzolari, A. Nannarelli, M. Petricca, M. Re, "Design space exploration based methodology for residue number system digital filters implementation," IEEE Trans. Emerging Top. Comput., 10(1): 186-198, 2020.

[2] I. Z. Alhassan, E. D. Ansong, G. Abdul-Salaam, S. Alhassan, "Enhancing image security during transmission using residue number system and k-shuffle," Earthline J. Math. Sci., 4(2): 399-424, 2020.

[3] D. Schoinianakis, "Residue arithmetic systems in cryptography a survey on modern security applications," J. Cryptographic Eng., 10(3): 249-267, 2020.

[4] M. A. Belhamra, E. M. Souidi, "Steganography over Redundant Residue Number System Codes," J. Inf. Secur. Appl., 51: 102434, 2020.

[5] M. I. Youssef, A. E. Emam, M. Abd Elghany, "Image multiplexing using residue number system coding over MIMO-OFDM communication system," Int. J. Electr. Comput. Eng., 9(6): 4815-4825, 2019.

[6] L. Sousa, P. Martins, "Sign detection and number comparison on RNS 3-Moduli sets $\{2^n\text{-}1, 2^{n+x}, 2^n+1\}$," Circ. Syst. Signal Process., 36: 1224-1246, 2017.

[7] C. Y. Hung, B. Parhami, "An approximate sign detection method for residue numbers and its application to RNS division," Comput. Math. Appl., 27: 23–35, 1994.

[8] T. Tomczak, "Fast sign detection for RNS $\{2^n-1, 2^n, 2^n+1\}$," IEEE Trans. Circuits Syst. I Regul. Pap., 55(6): 1502–1511, 2008.

[9] Z. Torabi, G. Jaberipur, "Fast low energy RNS comparators for 4-moduli sets $\{2^n \pm 1, 2^n, m\}$ with $m \in \{2^{n+1} \pm 1, 2^{n-1}\text{-}1\}$", Integr. VLSI J., 55: 155-161, 2016.

[10] S. Bi, W. J. Gross, "The mixed-radix Chinese remainder theorem and its applications to residue comparison," IEEE Trans. Comput., 57(12): 1624-1632, 2008.

[11] Y. Wang, X. Song, M. Aboulhamid, "A new algorithm for RNS magnitude comparison based on new Chinese remainder theorem II," in Proc. Ninth Great Lakes Symposium on VLSI: 362-365, 1999.

[12] M. Lu, J. S. Chiang, "A novel division algorithm for the residue number system," IEEE Trans. Comput., 1: 1026–1032, 1992.

[13] G. Dimauro, S. Impedovo, G. Pirlo, A. Salzo, "RNS architectures for the implementation of the diagonal function," Inf. Process. Lett., 73: 189–198, 2000.

[14] L. Sousa, "Efficient method for magnitude comparison in RNS based on two pairs of conjugate moduli," in Proc. IEEE Symposium on Computer Arithmetic (ARITH): 240-250, 2007.

[15] S. Kumar, C. H. Chang, TF Tay, "New algorithm for signed integer comparison in $\{2^{n+k}, 2^n\text{-}1, 2^n+1, 2^{n\pm1}\text{-}1\}$ and its efficient hardware implementation," IEEE Trans. Circuits Syst. I: Reg. Pap., 64(6): 1481-1493, 2016.

[16] K. Isupov, "Using floating-point intervals for non-modular computations in residue number system," IEEE Access, 8: 58603-58619, 2020.

[17] Z. Torabi, G. Jaberipur, "Low-power/cost RNS comparison via partitioning the dynamic range," IEEE Trans. Very Large Scale Integr. VLSI Syst., 24(5): 1849-1857, 2016.

[18] Z. Torabi, A. Belghadr, "Efficient RNS comparator via dynamic range partitioning: The case of $\{2^n\text{-}1, 2^n, 2^{n+1}\text{-}1\}$," CSI J. Comput. Sci. Eng., 16(2): 38-43, 2019.

[19] K. Isupov, "High-performance computation in residue number system using floating-point arithmetic," Comput., 9(2): 9-24, 2021.

[20] V. A. Krasnobayev, A. S. Yanko, S. A. Koshman. "A method for arithmetic comparison of data represented in a residue number system," Cybern. Syst. Anal., 52(1): 145-150, 2016.

[21] V. Krasnobayev, S. Koshman, K. Myslyvtsev,, K. Kuznetsova, , T. Ivko, T. Katkova, "Method of arithmetic comparison of data in the residue numeral system," in Proc. IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S&T): 483-487, 2019.

[22] M. Babenko, S. J. Piestrak, N. Chervyakov, M. Deryabin, "The study of monotonic core functions and their use to build RNS number comparators," Electron., 10(9): 1041, 2021.

[23] M. Babenko, M. Deryabin, S. J. Piestrak, P. Patronik, N. Chervyakov, A. Tchernykh, A. Avetisyan, "RNS number comparator based on a modified diagonal function," Electron., 9(11): 1784, 2020.

[24] H. Ahmadifar, G. Jaberipur, "A new residue number system with 5-Moduli Set: $\{2^{2q}, 2^q \pm 3, 2^q \pm 1\}$," The Comput. J., 58: 1548-1565, 2014.

[25] M. Mojahed, A. S. Molahosseini, A. A. E. Zarandi, "A multifunctional unit for reverse conversion and sign detection based on the 5-moduli set," Comput. Sci., 22(1), 2021.

[26] N. S. Szabó, R. I. Tanaka, Residue Arithmetic and Its Applications to Computer Technology. New York, NY, SA: McGraw-Hill, 1967.

[27] L. Sousa, P. Martins, "Efficient sign identification engines for integers represented in RNS extended 3-moduli set {2n − 1, 2n + k , 2n + 1}," 50(16): 1138-1139, 2014.

[28] M. Xu, Z. Bian, R. Yao, "Fast sign detection algorithm for the RNS moduli set {2^{n+ 1}-1, 2^{n}-1, 2^{n}\}," IEEE Trans. Very Large Scale Integr. (VLSI) Syst., 23(2): 379-383, 2014.

[29] Y. Wang, "New Chinese remainder theorems," in Proc. IEEE Asilomar Conference on Signals, Systems and Computers: 165-171, 1998.

[30] P. Boyvalenkov, N. I. Chervyakov, P. Lyakhov, N. Semyonova, A. Nazarov, M. Valueva, G. Boyvalenkov, D. Bogaevskiy, D. Kaplun, "Classification of moduli sets for residue number system with special diagonal function," IEEE Access, 8: 156104-156116, 2020.

[31] M. Valueva, G. Valuev, N. Semyonova, P. Lyakhov, N. Chervyakov, D. Kaplun, D. Bogaevskiy, "Construction of residue number system using hardware efficient diagonal function," Electron., 8(6): 694, 2019.

[32] G. Dimauro, S. Impedovo, G. Pirlo, "A new technique for fast number comparison in the residue number system," IEEE Trans. Comput., 42(5): 608-612, 1993.

[33] B. Cao, C. H. Chang, T. Srikanthan, "Adder based residue to binary converters for a new balanced 4-moduli set," in Proc. IEEE International Symposium on Image and Signal Processing and Analysis: 820-825, 2003.

[34] A. Tyagi, "A reduced-area scheme for carry-select adders," IEEE Trans. on Comput., 42: 63–70, 1993.

Biographies

Zeinab Torabi received her Ph.D. degree in Computer Architecture from Shahid Beheshti University, Tehran, Iran, in 2016. She is currently an Assistant Professor in Faculty of Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran. Her research interests include computer arithmetic, residue number system, and algorithms.

- Email: z.torabi@sru.ac.ir
- ORCID: 0000-0002-2526-688X
- Web of Science Researcher ID: ABG-9144-2022
- Scopus Author ID : 56958405600
- Homepage: https://www.sru.ac.ir/en/school-of-computer/zeinab-torabi/

Armin Belghadr received the B.S. degree in computer hardware engineering and the M.S. degree in computer architecture from Shahid Beheshti University, Tehran, Iran, in 2011 and 2013, respectively. He has also received his Ph.D. degree in computer architecture with the Department of Computer Science and Engineering, Shahid Beheshti University in year 2019. His-research interests include computer arithmetic and particularly residue number systems.

- Email: a_belghadr@sbu.ac.ir
- ORCID: 0000-0003-4835-6607
- Web of Science Researcher ID: Q-7750-2019
- Scopus Author ID: 55865963000
- Homepage: http://facultymembers.sbu.ac.ir/a_belghadr/

How to cite this paper:
Z. Torabi, A. Belghadr, "Fast and power efficient signed/unsigned RNS comparator & sign detector," J. Electr. Comput. Eng. Innovations, 11(1): 41-50, 2023.

DOI: 10.22061/JECEI.2022.8321.505

URL: https://jecei.sru.ac.ir/article_1717.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Depth Estimation and Deblurring from a Single Image Using an Optimized-Throughput Coded Aperture

M. Masoudifar[1,], H. R. Pourreza[2]*

[1]*Department of Computer Engineering, Hakim Sabzevari University, Sabzevar, Iran.*

[2]*Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran.*

Article Info

Article History:
Received 22 January 2022
Reviewed 17 March 2022
Revised 28 April 2022
Accepted 28 May 2022

Keywords:
Coded apertures
Depth from defocus
Defocus deblurring

*Corresponding Author's Email Address: *masoudi@hsu.ac.ir*

Abstract

Background and Objectives: Depth from defocus and defocus deblurring from a single image are two challenging problems caused by the finite depth of field in conventional cameras. Coded aperture imaging is a branch of computational imaging, which is used to overcome these two problems. Up to now, different methods have been proposed for improving the results of either defocus deblurring or depth estimation. In this paper, an asymmetric coded aperture is proposed which improves results of depth estimation and defocus deblurring from a single input image.

Methods: To this aim, a multi-objective optimization function taking into consideration both deblurring results and depth discrimination ability is proposed. Since aperture throughput affects on image quality, our optimization function is defined based on illumination conditions and camera specifications which yields an optimized throughput aperture. Because the designed pattern is asymmetric, defocused objects on two sides of the focal plane can be distinguished. Depth estimation is performed using a new algorithm, which is based on perceptual image quality assessment criteria and can discern blurred objects lying in front or behind the focal plane. **Results:** Extensive simulations as well as experiments on a variety of real scenes are conducted to compare the performance of our aperture with previously proposed ones.

Conclusion: Our aperture has been designed for indoor illumination setting. However, the proposed method can be utilized for designing and evaluating appropriate aperture patterns for different imaging conditions.

Introduction

When a scene is captured by a limited depth of field camera, objects lying at different depths are registered with varying degree of defocus blur. Depth from defocus (DFD) is a method that recovers depth information by estimating the amount of blur in different areas of a captured image. The concept of DFD was first introduced in [1], [2] and then various techniques were proposed, which are briefly reviewed in Sec. 2.

Despite the desirable results of DFD techniques in conventional apertures, there are some drawbacks rooted in the inherent limitation of circular apertures. For example, single image DFD methods and even some of multiple image DFD methods are unable to distinguish between defocused objects placed before and after the focal plane. In addition, in single image DFD methods, the lower depth of field, which provides enhanced depth discrimination ability, is obtained at the cost of losing image quality. In larger blur scales, most of image frequencies are lost. It makes the estimation of depth and

deblurring more ambiguous and vulnerable to image-noise [3].

Coded aperture photography is a method used for modifying the defocus pattern generated by lens. The shape of PSF (Point Spread Function) can be changed by using a coded mask on lens. So far, a variety of mask patterns have been proposed for improving the results of depth estimation [3]-[5] or defocus deblurring [6]-[8]. However, there are a few number of techniques for extracting both depth and high quality deblurred image [9], [10]. These methods use multiple images captured by a single aperture [9] or multiple aperture patterns [10].

In this paper, we propose an asymmetric single pattern, which is used for capturing a single image. This image is processed to achieve a depth map and an all-focus high quality deblurred image.

To find the proposed optimal aperture pattern, a new multi-objective function containing two evaluation functions is defined. The first function determines the expected value of deblurring error using the correct PSF. The second function computes the expected value of deblurring error using the incorrect PSFs. Both functions are defined in the frequency domain. A non-dominated sorting-based multi-objective evolutionary algorithm [11] is used to find a Pareto-optimal solution. An optimal pattern is chosen in a way that it can also distinguish between defocused objects placed before and after the focal plane. As a result, an asymmetric pattern is proposed which is appropriate for depth estimation and deblurring in a single captured image.

According to [12], [13], illumination conditions and camera specification affect the performance of coded aperture cameras. Therefore, our objective functions are formulated by considering the imaging circumstances. In this way, the designed mask acquires a reasonable throughput that ensures the acceptable signal-to-noise ratio (SNR) of the captured image.

The proposed mask is compared with circular aperture, and a number of state-of-the-art coded aperture patterns. The performance comparison includes depth estimation accuracy and the quality of deblurring results.

In accordance with the proposed multi-objective function, a depth estimation algorithm is introduced in which a blurred image is deblurred by a set of PSF scales. Then, a PSF with the best quality deblurring result is considered as the correct blurring kernel. The quality of deblurred images is measured by an aggregate measure of no-reference image quality assessment criteria.

A. Key Contributions

1) A new multi-objective function is proposed for defining a single pattern, which yields the minimum deblurring error with correct PSF and the maximum deblurring error with incorrect PSFs.

2) The blurring problem is redefined with respect to the aperture throughput and imaging system conditions. Hence, in the design of coded aperture, the amount of additive noise and image brightness are taken into account.

3) An aggregate no-reference image quality assessment measure is used for depth estimation. The quality of images deblurred by different PSFs is measured and the PSF that yields a deblurred image of the highest quality is chosen as the true PSF.

4) The results of simulation on a dataset show less variance in correct depth/kernel estimation across the entire range of depths/kernel sizes compared to previous aperture patterns.

B. Scope and Limitations

1) The image formation model is assumed to be linear.

2) An affine noise model is used to describe the combined effects of signal-dependent and signal-independent noise. Signal dependent Poisson noise is approximated using a Gaussian noise model. Signal independent noise is assumed only read-noise.

3) The aperture pattern is designed based on the assumption that the exposure time and lighting condition is fixed.

4) The proposed aperture pattern and depth estimation algorithm can be used for both grey-level and color imaging systems.

The rest of this paper is organized as follows: In Sec. 2, related works are briefly reviewed. In Sec. 3, the blurring problem is formulated and pattern evaluation functions are introduced. Section 4 describes the optimization method used to find the optimal pattern. The proposed aperture is analyzed in accordance with spectral properties and depth sensitivity in Sec. 5. Our depth estimation algorithm is presented in Sec.6. Experimental results in both synthetic and real scenes are presented in Sec. 7. Finally, conclusions are drawn in Sec. 8.

Previous Works

The concept of DFD was first introduced in [1], [2] and then a variety of techniques were proposed that used a single image [14]-[19] or multiple images [20]-[24].

Single image DFDs usually estimate the blur scale either by assuming some prior information about PSF [14], [18], texture [16], color information [17] or by using learning methods [19]. However, multiple image DFDs are more variant and use various techniques to extract depth information. Some methods capture two or more images from a single viewpoint under different focus settings or various sizes of aperture [1], [2], [22], [24], [25]. Other methods use two or more images from different viewpoints such as stereo vision with identical focus setting [26] or different focal settings [9].

As mentioned in Sec.1, DFD with conventional apertures suffers from some drawbacks. In the past

decades, coded aperture photography has been used to resolve these problems. Here, some of the proposed apertures and DFD methods are briefly reviewed.

Hiura et al. [27] use multiple images taken by a single aperture pattern from a single viewpoint under different focus settings. Zhou et al. [10] propose a pair of aperture masks. Two blurred images are taken from a single viewpoint with a similar focus setting and two different asymmetric aperture patterns. In real applications, a programmable aperture is needed to ensure that the viewpoint of the two captured images remain unchanged. Otherwise, images should be first registered, and then depth estimation algorithm be applied. Takeda et al. [9] use stereo imaging with a single aperture pattern, yet different focal settings to improve the results of depth estimation presented in [10].

Levin et al. [4] design a single symmetric pattern with the aim of increasing the depth discrimination ability. Kullback-Leibler divergence between different sizes of blur is used to rank aperture patterns. The optimal symmetric pattern is achieved through a full-search of all binary masks. An efficient deblurring algorithm is also used to create high quality deblurred results. Since the proposed mask is symmetric, before and after focal plane cannot be differentiated.

Sellent et al. [5] define a function in the spatial domain for the aperture pattern evaluation. A parametric maximization problem is defined to find a pattern that produce the most possible difference among images blurred of different PSF scales. By solving this problem, non-binary patterns are obtained that can be pruned to binary forms. This technique is then used to find asymmetric patterns suitable to discriminate the front and back of the focal plane [3].

Aperture Evaluation

In this section, first the blurring problem is briefly reviewed and then our criteria for evaluating aperture patterns are introduced. Based on the proposed criteria, a multi-objective function is defined, which is capable to compare aperture patterns with varying throughputs.

A. Problem Formulation

Image degradation due to out of focus blurring and noise can be modeled by convolution of a PSF or kernel function (k_d) with the sharp image (f_s) and then adding noise (ω):

$$f = k_d \otimes f_s + \omega, \qquad \sum_i k_d^i = 1 \tag{1}$$

the subscript *d* indicates that kernel size is a function of depth of scene. The sum of kernel elements (i.e. k_d^i) equals 1, meaning that the image brightness does not change by blurring.

When we use a binary-coded aperture, the shape and throughput of the aperture are determined by this mask. As noted in [12], [13] an aperture pattern must be evaluated by consideration of both shape and throughput. Therefore, we redefine the well-known defocus problem in terms of these factors.

A binary coded mask with n open cells can be considered as a grid of size N×N, where n holes distributed over the grid are kept open [5], [12]. The pattern of open holes determines the shape of PSF, and their number specifies the mask throughput.

For a simple fronto-parallel object at depth *d*, defocusing is redefined as the convolution of a defocus kernel (k_d) with a sharp image(f_n) that generates spatial invariant blur:

$$f = k_d \otimes f_n + \omega_n, \qquad \omega_n \sim N(0, \sigma_n^2), \qquad \sum_i k_d^i = 1 \tag{2}$$

The subscript *n* shows that the brightness of sharp image (f_n) and the amount of added noise (ω_n) depend on the aperture throughput (*n*). Due to the additive properties of light, in a constant definite exposure time, the brightness of sharp image (f_n) is increased linearly with an increase in the number of open holes. The value of ω_n also changes with the number of holes. In this study, the growth of ω_n is investigated by considering the number of holes, imaging system's specifications and scene illumination. As mentioned earlier, the sum of kernel elements (i.e. k_d^i) equals 1, meaning that the image brightness does not change by blurring. As we see in Sec. 3. B, the added noise is modeled by normal distribution, which its variance depends of the aperture throughput.

Equivalently, if the Fourier transforms of each variable is shown by its corresponding capital letter, the spatially invariant blur in the frequency domain is defined as follows:

$$F = K_d \,.\, F_n + \Omega_n \tag{3}$$

where the convolution operation in the Fourier domain is changed to a simple point-by-point multiplication. The subscripts *d* and n indicate the depth of scene and aperture throughput, respectively.

B. Noise Model

The imaging noise can be modeled as the sum of two distinct factors: read noise and photon noise [12]. Read noise, which is independent of the measured signal, is commonly modeled by a zero mean Gaussian random variable *r* with variance σ_r^2. Photon noise is a signal dependent noise with Poisson distribution. When the mean value of photon noise is large enough, it can be approximated by a random Gaussian variable with variance $\sigma_p^2 = J_n$[12], [28]. J_n refers to the mean number of photons received by a single pixel in a camera with an

n open-hole aperture.

As noted in [12], the total noise variance is computed as follows:

$$\sigma_n^2 = \sigma_r^2 + \sigma_p^2 = \sigma_r^2 + J_n = \sigma_r^2 + n.J \quad (4)$$

In this study, the mean signal value in photoelectrons (*J*) of a single-hole aperture is computed by [12]:

$$J = 10^{15}.\frac{1}{F\#^2}.R.I.q.\Delta^2.t \quad (5)$$

where *F#*, *R*, and *I* refer to camera f-number, average scene reflectivity that varies in range 0 to 1, and amount of scene illumination (measured in *lux*), respectively. *q* is the quantum efficacy of the image sensor, which measures the effectiveness of an imaging device to convert incident photons into photoelectrons. *Δ* is the size of a pixel in an image sensor and *t* refers to the exposure time. In our experiments, the assumption about scene and imaging system parameters, which represent the typical settings in consumer photography, are as follows:

q = 0.5 (typically CMOS sensors)
R = 0.5, t = 10^{ms}, F# = 18
Δ^2 = 5.1×5.1$^{\mu m}$ (SLR camera, typically Canon 1100D)
I = 300^{lux} (typically office light level)

In the following section, first our criteria regarding the intensity level of images are proposed. Then, the proposed formula in terms of photoelectron are redefined so that masks with different throughputs can be compared.

C. Mask Search Criteria

Suppose an image F_n is blurred with an unknown Kernel K_1 (3). If it is deblurred with a typical kernel K_2 and Wiener filter is used for deconvolution, then the total error of deblurring (e_n) is computed as (6):

$$\begin{aligned} e_n &= F_n - \hat{F}_n = F_n - \frac{K_2^* F}{|K_2|^2 + |C_n|^2} \\ &= F_n - \frac{K_2^*(K_1 F_n + \Omega_n)}{|K_2|^2 + |C_n|^2} \\ &= \underbrace{\frac{F_n K_2^*(K_2 - K_1)}{|K_2|^2 + |C_n|^2}}_{e_n^{(1)}} + \underbrace{\frac{F_n|C_n|^2 - K_2^*\Omega_n}{|K_2|^2 + |C_n|^2}}_{e_n^{(2)}} \\ &= e_n^{(1)} + e_n^{(2)} \end{aligned} \quad (6)$$

where $|C_n|^2$ is defined as the matrix of expected value for noise to signal power ratios (NSR) of natural images. (i.e. $|C|^2 = \frac{\sigma^2}{A}$ where A is the expected power spectrum of natural images and σ^2 is the variance of additive noise [7].) According to (6), the total error consists of two parts:

$$e_n^{(1)} = \frac{F_n K_2^*(K_2 - K_1)}{|K_2|^2 + |C_n|^2} \quad \text{error of wrong kernel estimation} \quad (7)$$

$$e_n^{(2)} = \frac{F_n|C_n|^2 - K_2^*\Omega_n}{|K_2|^2 + |C_n|^2} \quad \text{deblurring error} \quad (8)$$

If an accurate PSF is used for deblurring (i.e. $K_1 = K_2$), then the only term that determines the total error of deblurring will be $e_n^{(2)}$ (i.e. $e_n^{(1)} = 0$). On the other hand, if a wrong kernel is used as PSF ($K_1 \neq K_2$), both $e_n^{(1)}$ and $e_n^{(2)}$will generate errors in the deblurring result. As will shown in sec. 4.A, the values of $e_n^{(1)}$ are much greater than $e_n^{(2)}$ (See Fig. 2). Therefore, when $K_1 \neq K_2$, the main determinant of the total error will be $e_n^{(1)}$. Hence, consistent with our objective, a suitable pattern is defined as a pattern that minimizes the norm of $e_n^{(2)}$ and maximizes the norm of $e_n^{(1)}$. The norm of $e_n^{(2)}$ is computed as follows:

$$\begin{aligned} \left\|e_n^{(1)}\right\|_2^2 &= \left(\frac{F_n K_2^*(K_2 - K_1)}{|K_2|^2 + |C_n|^2}\right)^* \left(\frac{F_n K_2^*(K_2 - K_1)}{|K_2|^2 + |C_n|^2}\right) \\ &= |F_n|^2|K_2|^2 \frac{|K_2 - K_1|^2}{||K_2|^2 + |C_n|^2|^2} \end{aligned} \quad (9)$$

Since the power spectra of all natural images follow a certain distribution, the expectation of $\left\|e_n^{(1)}\right\|_2^2$ can be computed with respect to F_n. According to *1/f* law of natural images [29], the expectation of $|F_n|^2$ is computed as $A_n(\xi) = \int_{F_n} |F_n(\xi)|^2 \, d\mu(F_n)$ where ξ is the frequency and $\mu(F_n)$ is the measure of sample F_n in the image space [7]. Accordingly, the expectation of $\left\|e_n^{(1)}\right\|_2^2$ is computed as (10):

$$\begin{aligned} D_n(K_2, K_1) &= \mathbb{E}_{F_n}\{\left\|e_n^{(1)}\right\|_2^2 \\ &= \sum_{\xi} \frac{A_{n_\xi}|K_2|_\xi^2}{(|K_2|_\xi^2 + |C_n|_\xi^2)^2}|K_2 - K_1|_\xi^2 \end{aligned} \quad (10)$$

This measure can be considered as a distance criterion between two kernels. It can also help distinguish between defocus points lying in front or back of the focal plane. It should be noted that the defocus PSF in front of the focal plane is the flipped version of the defocus PSF at the back of the focal plane (See Fig. 1. a), meaning that these PSFs have an identical spectral response but different phase properties. Equation (10) includes the term K2-K1, which can compute both spectral and phase differences of two kernels. Hence, by having an asymmetric aperture, the deblurring with the flipped version of a PSF generates $e_n^{(1)}$ error and helps distinguish sides of the focal plane (See Fig. 1. b)

The expected value of $\left\|e_n^{(2)}\right\|_2^2$ can be computed in a similar manner. (Details are found in [7]):

$$R_n(K_1) = \left\|e_n^{(2)}\right\|_2^2 = \sum_{\xi} \frac{\sigma_n^2}{|K_1|_\xi^2 + |C_n|_\xi^2} \quad (11)$$

This value has been used by Zhou et al. [7] as a criterion to find aperture patterns with least errors in deblurring results. However, it has been redefined here to allow

studying patterns with different throughputs. Additionally, we search for a pattern that is suitable for both depth estimation and deblurring.

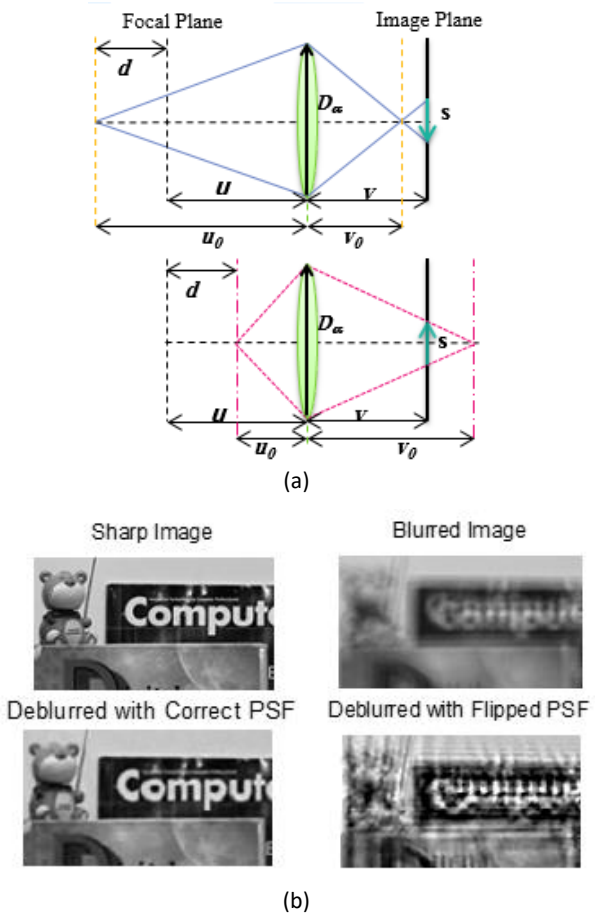


Fig. 1: (a) the defocus PSF in front of the focal plane is the flipped version of the defocus PSF at the back of focal plane. (b) If an asymmetric pattern is used for imaging, then deblurring with the flipped PSF yields more errors in the deblurred image (see (10)).

If the camera response function [30] is assumed linear, then relations (10) and (11) can be stated in terms of photon as follows:

$$\begin{cases} D_n(K_2,K_1) = \sum_{\xi} \dfrac{{J_n}^2 . A_{1_\xi} |\mathrm{K}_2|_\xi^2}{(|K_2|_\xi^2 + \dfrac{\sigma_r^2 + J_n}{{J_n}^2 . A_{1_\xi}})^2} |K_2 - K_1|_\xi^2 \\ R_n(K_1) = \sum_{\xi} \dfrac{\sigma_r^2 + J_n}{|K_1|_\xi^2 + \dfrac{\sigma_r^2 + J_n}{{J_n}^2 . A_{1_\xi}}} \end{cases} \quad (12)$$

where A_1 refers to the expected power spectra of natural images taken by a single hole aperture. Since we assume the aperture has n holes and the camera has a linear response function, the number of absorbed photoelectrons in an n hole aperture, is n times of a single hole aperture (i.e. $J_n = n.J$). We also assume $\sigma_n^2 = \sigma_r^2 + J_n$ based on what was described in Sec.3.B (See (4)).

The values of R_n and D_n grow with n. So, the range of these values is different for apertures with a different number of open holes. If we desire to study patterns with different throughputs, then D_n and R_n must be normalized. Hence, our multi-objective function is defined as follows:

$$\begin{cases} min \;\; R(K_{s_1}) = \dfrac{1}{n^2} . R_n(K_{s_1}) \\ \qquad\qquad\qquad\qquad , s_1, s_2 \in S \; and \; n \in [1..N^2] \\ max \;\; D(K_{s_1}, K_{s_2}) = \dfrac{1}{n^2} D_n(K_{s_1}, K_{s_2}), \;\; s_1 \neq s_2 \end{cases}$$
$$s.to: \qquad 0 \leq |K_s(\xi)| \leq 1, \quad s \in S \quad (13)$$

where S refers to a limited range of blur scales and N is the mask resolution.

Aperture Pattern Design

In this study, the mask resolution (N) is determined in a way that each single hole provides the least possible diffraction. According to the formula proposed in [31], a 7×7 mask is appropriate for an imaging system with an aperture-diameter of 20^{mm} and pixel-size of $5.1^{\mu m}$. Based on the camera specifications used in our experiments, this resolution is selected for our mask, and thus the number of open holes (n) will be in the range of [1-49].

A. Optimization

Multi-objective optimization is usually described in terms of minimizing a set of functions. Therefore, we rewrite our objective functions as follows:

$$\min \{ R(K_{s_1}), -D(K_{s_1}, K_{s_2}) \},$$
$$for \; s_1, s_2 \in [1..10] \; and \; s_1 \neq s_2 \quad (14)$$

These evaluation functions are clear and concise, but their solution in the frequency domain is challenging. Since we search for a binary pattern with specific resolution, the objective function must also be able to satisfy some other physical constraints in the spatial domain. It is difficult to derive an optimal solution that satisfies all constraints in both frequency and spatial domains. Therefore, a heuristic search method is used to solve the problem. In evaluating each pattern, R and D values are computed for 10 different scales of kernels (See (14)). Then, the maximum value of R and minimum value of D are used to evaluate the pattern.

The main goal of a multi-objective optimization problem is to find the best Pareto optimal set of solutions [11]. In this study, NSGA-II [32], which is an appropriate method for solving multi-objective optimization problems, is used to optimize our objective functions. A generation of binary patterns with a population size of 1500 is created. A pattern is defined by a vector of 49 binary elements. According to [33], this population size is sufficient to converge to a proper solution. Other parameters are set by default values adjusted in the prepared software. Fig. 2 shows the values of objective

functions in the Pareto-front. The values of proposed objective functions are also computed for some other apertures and then added to the figure.

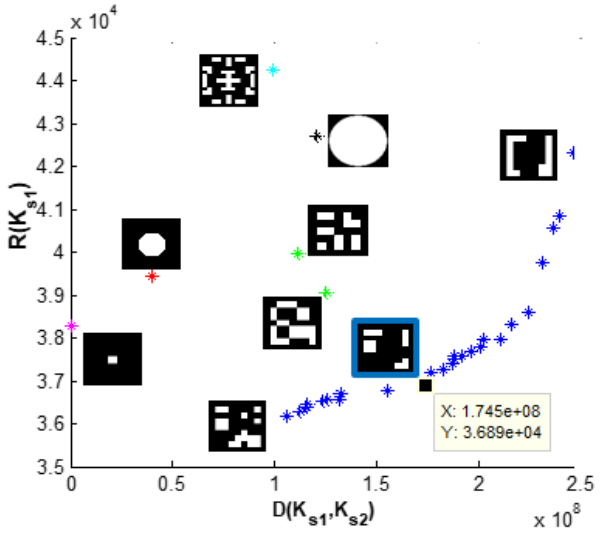


Fig. 2: D values vs. R values of final patterns in the Pareto optimal solution (blue), Open circular aperture (black), conventional aperture (red), pinhole aperture(magenta), patterns proposed in [3] (green) and [4] (cyan). Final selected pattern has been highlighted by the blue border.

According to Fig. 2, in the Pareto optimal solution, with an increase in the symmetry of patterns, the deblurring error (*R*) and the error of using a wrong scale kernel (*D*) rises. However, it does not mean that any symmetric pattern outperforms all other asymmetric patterns in terms of discrimination ability (D). For example, objective functions were also computed for the pinhole aperture, open circular aperture and circular aperture with a throughput equal to the selected coded pattern (highlighted by the blue border)[1] as well as the symmetric pattern proposed by Levin et al. [4]. Although these patterns are symmetric, the provided *D* values are not essentially greater than all asymmetric patterns. On the other hand, *R* values provided by asymmetric patterns are not essentially smaller than any symmetric ones. In fact, *R* and *D* values depend on several factors such as mask throughput and spectral properties.

As noted earlier, NSGA-II provides a set of solutions. Since just one pattern has to be selected, we compute $D_r = D(K_{s1}, rot(K_{s1}, 180))$ for all patterns derived from the Pareto optimal solution. In a similar manner, this value is computed for asymmetric patterns proposed in [3]. Fig. 3 shows the computed values.

As shown in Fig. 3, with an increase in symmetry, D_r declines. Given the significance of criterion D_r, the pattern highlighted by the blue border is selected as a sample of the derived patterns.

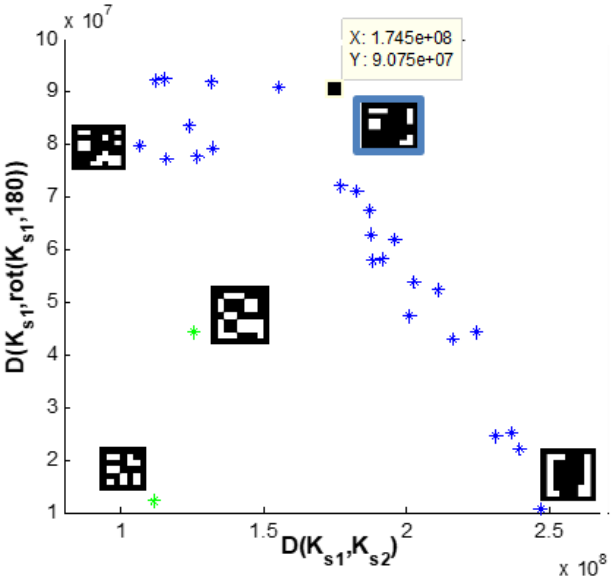


Fig. 3: *D* value of wrong scale kernels vs. D_r value of the flipped correct scale for the patterns obtained by NSGA-II (blue) and asymmetric patterns proposed in [3] (green).

It must be mentioned that the selected pattern is not the best option under all conditions. However, since it provides appropriate values of *D, R* and D_r, it is selected as the final pattern. Indeed, the final pattern should provide a minimum value for the weighted sum of all criteria, which each weight representing the importance of the associated criterion. This study adopts NSGA-II, which does not use the weighted sum for optimization.

Aperture Pattern Analysis

In this section, a brief analysis of the proposed pattern is presented. The transmission rate (compared to the open circular aperture) of our optimized aperture is 0.265, which is almost equal to the Levin's pattern [4]. Hence, the SNR of images captured by this aperture is about 14.4dB[2], which is in the range of [10..40], meaning that the captured images have an acceptable (not ideal) SNR [34]. In the following; the aperture pattern is examined with respect to its spectral properties and depth sensitivity.

A. Spectral analysis

At the first step, an analogy is drawn between the spectral properties of the selected pattern and the conventional aperture. It should be noted that both apertures have similar throughput so under different imaging conditions; the same amount of additive noise is added to the captured images. In this situation, the spectral properties of apertures determine the results. Fig. 4 shows 1D slices of spectral response for each aperture at five different blur scales. According to [4],

[1] In the rest of text, the circular aperture with the same throughput of selected coded pattern is called conventional aperture.

[2] $\mathrm{SNR}_{\mathrm{capture}} = 10\log_{10}(J_n/\sigma_n)$

when a pattern has various frequency responses in each scale, it is more convenient to distinguish blur scales. As shown in Fig. 4, in the conventional aperture, the zero amplitude obtained from different scales overlaps in some frequencies, making it difficult to distinguish between blur scales. However, the coded pattern has diverse spectral responses in different scales.

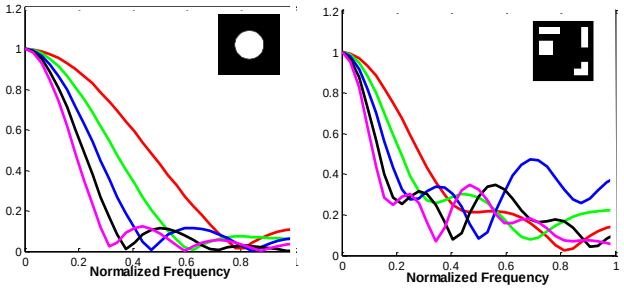


Fig. 4: The 1D slide of spectral response at 5 different blur scales for conventional and coded aperture.

The spectral response of these two apertures is also compared at 4 different scales. As shown in Fig. 5, the minimum spectral response of our pattern is greater than the conventional aperture, especially in larger blur scales. Therefore, in the proposed pattern, attenuation of frequencies in the captured image is reduced and thus deblurring results are improved.

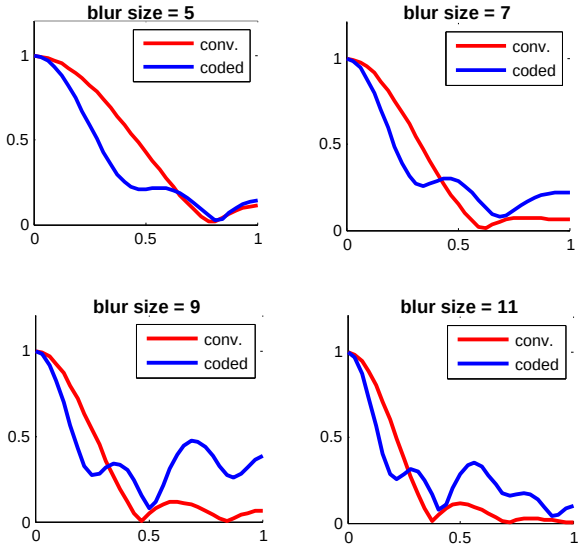


Fig. 5: 1D slices of Fourier transforms for conventional aperture (red) and the proposed pattern (blue) at 4 different scales.

B. Depth Sensitivity

Another advantage of the proposed pattern is its high sensitivity to the depth variation. It is known that DoF declines with an increase in the aperture diameter. In the proposed pattern, open holes are located in the margin of the mask. Hence, this aperture pattern is more sensitive to depth variations than the conventional aperture. To examine the depth sensitivity difference in these apertures, the blur size is computed in a limited range of depth (before and after the focal point) for a typical lens (EF 50mm f/1.8 II).

The blur size (s) is computed based on thin lens formula [35]:

$$s = \frac{D_a(v - v_0)}{v_0}, \quad v_0 = \frac{Fu_0}{u_0 - F}, \quad v = \frac{Fu}{u - F} \qquad (15)$$

The parameters used in (15) were introduced in Fig. 1(a). The aperture diameter (D_a) is assumed 20^{mm} and 8.21^{mm} for coded and conventional patterns respectively. As shown in Fig. 6, the proposed pattern is more sensitive to depth variation. Therefore, depth estimation is easier in images captured by the coded pattern. On the other hand, according to Fig. 5, coded mask gives a higher spectral response, and is thus expected to obtain better results in both deblurring and depth estimation in real imaging.

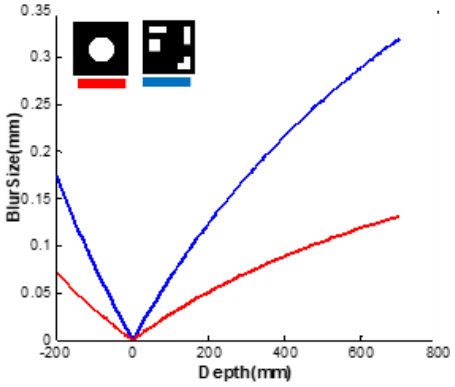


Fig. 6: Blur size vs. depth for conventional (red) and coded (blue) apertures (focus length (u) = 1200^{mm}, v = 50^{mm}). Code aperture is more sensitive to depth variation.

Depth Estimation

Depth estimation is performed using an algorithm described here. The method is based on the proposed objective function (13) and can be used for detecting both the scale and the orientation of PSF. The main idea is that deblurring with inaccurate kernels, whether in scale or direction, produces low-quality images while deblurring with correct kernel yields high quality images (See Fig. 7). For depth estimation, the blurred image is deblurred with a limited set of blurring kernels. The quality of each deblurred image is measured using an aggregate no-reference image quality measure. A PSF, which generates a deblurred image of the highest quality, is selected as the true kernel. As stated earlier, if the aperture pattern is asymmetric, this method can be used for detecting both size and direction of the PSF (Fig. 1).

Several no-reference image quality measures have been proposed in the literature. In one of the most comprehensive studies [36] a weighted sum of 8 different criteria is used for evaluating the image quality (Recent

studies show that using an aggregate measure of image quality assessment criteria is more precise [8], [36], [37] Although this measure can be used for depth estimation, it is more complicated than it is necessary. In our application, a comparison is drawn between the qualities of deblurred versions of the same image.

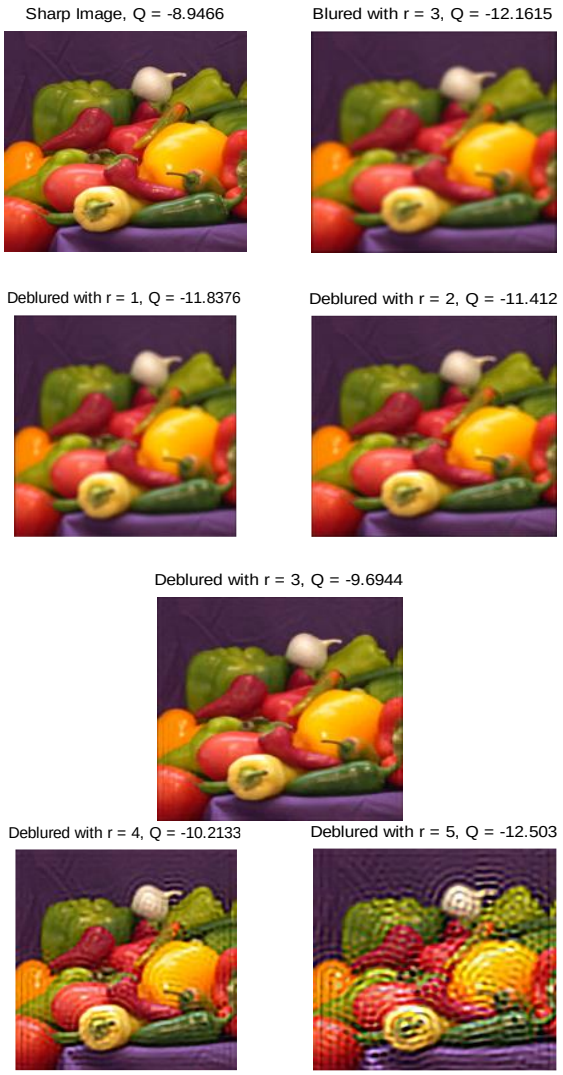


Fig. 7: Deblurring results with different radii of the kernel (r=1..5) in imaging with conventional aperture. Deblurring with smaller kernels results in blurry images and deblurring with larger PSFs yields images with artifacts. The quality of each image is evaluated by the no-reference quality assessment measure proposed in [36]. A larger Q-value indicates higher quality.

In fact, here the quality measure is more of a relative measure not a strict one. Therefore, measures of lower complexity can be applied for quality assessment. The speed of depth estimation algorithm is improved by reducing the number of criteria. In this study, the quality of deblurred images is evaluated by an aggregated measure containing four criteria: *Norm-Sparsity-Measure* [38], *Sparsity-Prior* [4], *Sharpness-Index* [39] *and Pyramid-Ring* [36], which are well-suited for our application. These criteria are sensitive to blur or artifact or both of them. The no-reference aggregate image quality measure is defined in (16), where higher values indicate greater quality. The process of computing this measure has been described in our previous work [40].

$$Quality = -12.65 * normSps + 0.073 * sharpIndex - 0.289 * sparsity - 9.86 * pyrRing \quad (16)$$

A similar measure has been used in [3] to find only the direction of PSF. Sellent et al. [3] use a depth estimation algorithm [35] to determine the scale of PSF. Then, a quality assessment measure is used to find the direction of PSF. Our proposed method is almost similar to [3], but no prepared database is used for PSF estimation here. We use the proposed measure to evaluate the quality of deblurred images (or patch of images) derived by different PSFs. A PSF, which yields a deblurred image with the best quality, is chosen as true PSF. This method is used for detecting both size and direction of PSF.

A. Handling Depth Variations

In real world scenes, there are depth variations. Therefore, each part of an image might be blurred with a different kernel. A common method of depth estimation in these images involves using fairly small patches in which the depth is assumed to be constant. The blur kernel is estimated for the patch, and this estimation is assigned to its central pixel. By repeating this stage for all pixels of the image, a raw depth map is obtained. Then, a coherent map labeling is performed using the raw depth map, image derivative information and some smoothness priors [4], [17].

In this study, first two blur scales that generate deblurred patches of the highest quality are considered as the possible true scales of the central pixel. The probability of each scale is computed based on its relative quality. Higher quality increases probability and the sum of two probability values are equal to 1. At the end of this stage, a three-dimensional matrix is obtained. In other words, for a H×W image and S possible depths, matrix $D_R \in \mathbb{R}^{H \times W \times S}$ includes the raw depth map in which $D_R(h,w,s)$ represents the probability of depth $s \in S$ in pixel (h,w).

There may be some errors in the depth estimation of the raw depth map, especially in depth discontinuities. Therefore, in the second step, a coherent blur map is obtained by minimizing an energy function defined as follows [17]:

$$Min\, E(D_c) = \sum_{p} D_p(s_p) + \sum_{(p,q) \in N} \lambda_{p,q} V(s_p, s_q) \quad (17)$$

where p and q refer to image pixels. The first term $D_p(s_p)$ indicates fidelity to the previous probability blur scale (s) estimation at position p. The second term $V(s_p, s_q)$ is a

smoothness term, which guarantees that neighbor pixels of similar gray levels have identical blur scales. D_c denotes a solution for coherent data map with minimum energy (E). A coherent map with min(D_c) is estimated by a method proposed in [17].

To assign a penalty to depth change in D_p, the early probabilities of blur scale ($p_p(s)$) are convolved with a Gaussian filter ($N(0,0.1)$) to reach the smoothed probabilities ($\hat{p}_p(s)$). Then $-\log(\hat{p}_p(s))$ is used as $D_p(s)$. (See [17]). This function could also be used for cases in which one or more probabilities are assigned to the initial blur scale.

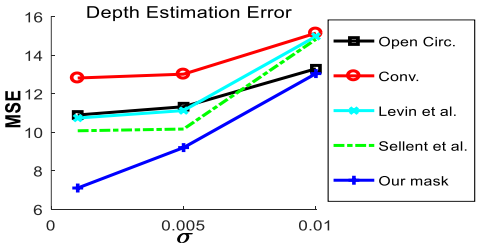
The smoothness term $V(s_p, s_q)$ examines depth discontinuity in neighboring pixels. For each pixel p, depth similarity is investigated with its eight surrounding pixels with $V(s_p, s_q) = |s_p - s_q|$.The relative significance of the difference between depths of two adjacent pixels is determined by the difference of their gray level (g_p and g_q). Hence, $\lambda_{p,q}$ is defined as $\lambda_{p,q} = \lambda_0 e^{-(\frac{\|g_p - g_q\|^2}{\sigma_\lambda^2})}$ [17].

In our experiment, parameters are set to λ_0=1000 and σ_λ=0.006. Finally, α-expansion is used to minimize the energy function [41].

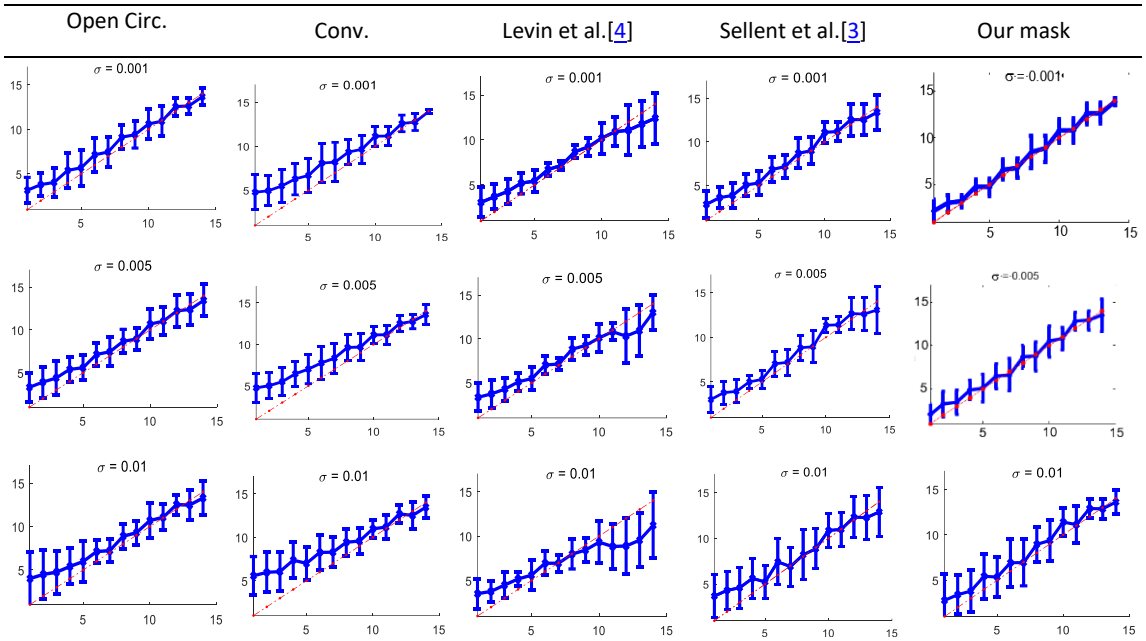
Experiment

The proposed mask and depth estimation method are validated in several experiments. The mask is compared with circular aperture, conventional aperture and two other masks designed for depth estimation [3], [4] (It must be mentioned that our study does not include aperture patterns proposed for deblurring, which assume to have sufficient information about blurring kernel and only focus on deblurring results). Among the masks proposed by Sellent et al. [3], we choose the 7×7 mask, which is the best according to our evaluating criteria (see Fig. 2 and 3). Our study contains synthetic and real experiments. It is expected that the designed mask increases the accuracy of PSF estimation and provides desirable deblurring results.

(a) A few number of patches used in the experiments.



(b) Average of depth estimation error of blur scales (s = 1:14).



(c) The average and variance of estimated blur scale (vertical axis) in comparison with ground truth scale (horizontal axis). Red diagonal represents the ideal estimation.

Fig. 8: Results of depth estimation for five apertures at 3 noise levels (σ=0.001, 0.005, 0.01) and 14 blur sizes (s = 1:14).

A. Synthetic Experiments

I) Depth Estimation Accuracy

In the first experiment, a number of various images are blurred uniformly with various blur scales (s=1:14). Then, 50 patches of these images are randomly selected and their depth is estimated by the method described in Sec.6. Fig. 8(a) shows some of the selected patches. In each scale, the mean and variance of estimated size of PSFs are computed over all patches.

This experiment is repeated for different aperture patterns at three levels of noise (σ = 0.001, 0.005, 0.01). Based on the results shown in Fig. 8(c), the depth estimation accuracy is reduced by increasing noise. However, results are satisfactory especially in our mask and the mask proposed by Sellent et al. [3]. It must be mentioned that since both symmetric and asymmetric patterns are studied in this experiment, only one side of the focal plane is considered.

For better comparison of studied aperture patterns, in each scale, the norm of difference between the ground truth blur scale (s_{gt}) and the estimated blur scale (s_{es}) is computed over all patches (i.e.$\sum_{p=1}^{50}\left(s_{gt}^{p}-s_{es}^{p}\right)^{2}$). Then, this value is averaged over all studied blur scales. Fig. 8(b) shows the mean square error (MSE) of depth estimation for different apertures at three noise levels. It shows that under equal circumstances, where all imaging conditions (including throughput) are the same, coded pattern has greater performance than its corresponding conventional aperture.

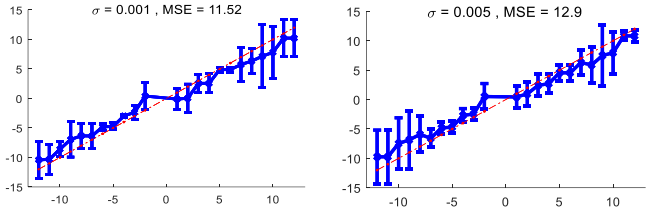
The depth estimation experiment is repeated for asymmetric patterns with blur sizes in the range of -12:12 pixel. Since a blur size of 0 is meaningless and ±1 indicates a sharp image, 23 different sizes of blur are indeed examined. According to Fig. 9, our method provides favorable results at σ = (0.001, 0.005) with the depth estimation error (MSE) of the proposed aperture being less than the pattern in [3].

II) Deblurring Results

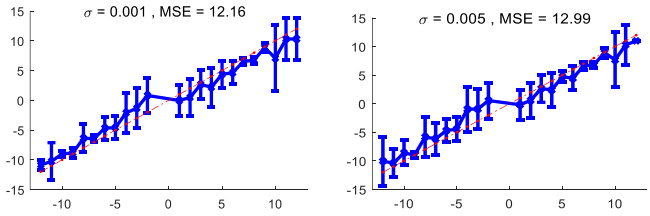
In the second experiment, deblurring results of aperture patterns are examined. For different scales of blur, each blurred patch is deblurred with a correct scale of PSF. Then, the Root Mean Square Error (RMSE) of the difference between original sharp image and its deblurred version is computed. The average of RMSE is calculated over all patches.

As shown in Fig. 10, our pattern provides the least error, especially in large blur scales, while the conventional aperture is the best aperture in lower blur scales.

A sample of deblurring result for Circular Zone Plate (CZP) chart is shown in Fig. 11. In all experiments, images are deblurred by the sparse deconvolution algorithm proposed by Levin et al. [4].



(a) Our mask



(b) Mask proposed by Sellent et al. [3].

Fig. 9: The average and variance of estimated blur scale (vertical axis) compared to ground truth scale (horizontal axis) at 2 noise levels (σ=0.001, 0.005) in the depth range of -12:12. Red diagonal represents the ideal estimation.

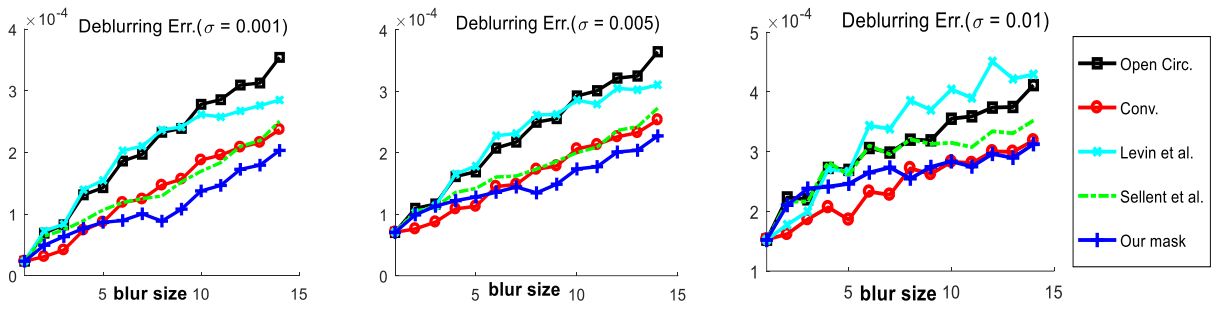


Fig. 10: Deblurring error of five apertures at 3 noise levels (σ=0.001, 0.005, and 0.01) for 14 blur sizes (s = 1:14).

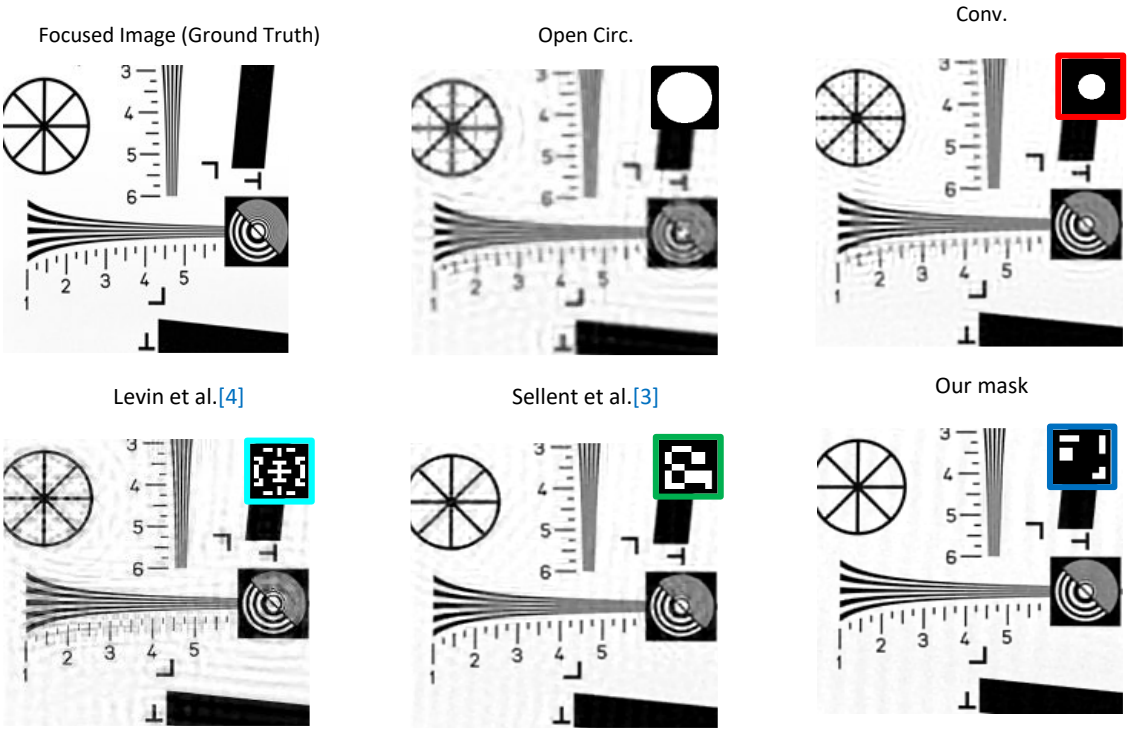


Fig. 11: Comparison of deblurring results derived from different aperture patterns (blur size = 13, σ=0.005).

B. Real Scene

For real experiments, the proposed pattern is printed on a single photomask sheet. It is cut out of the photomask sheet and inserted into a camera lens. In our experiment, a Canon EOS 1100D camera with an EF 50mm f/1.8 II lens is used. The disassembled lens and the one assembled with the proposed mask are shown in Fig. 12(a, b).

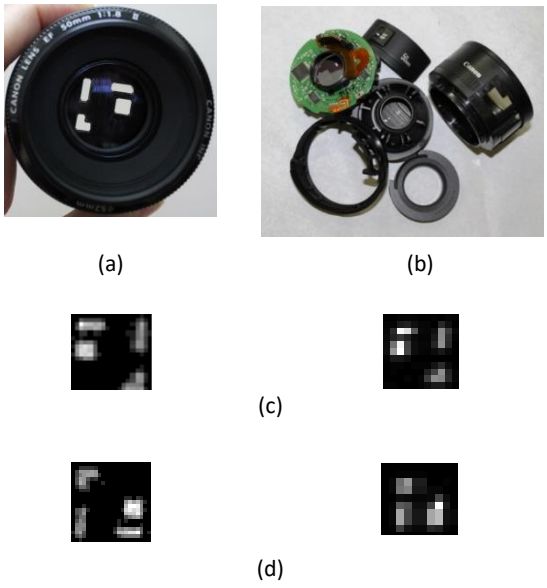


Fig.12: (a) lens assembled with the proposed mask, (b) disassembled lens. (c), (d) calibrated PSFs of evaluated pattern.

A very thin LED is used to calibrate the true PSF. The LED is mounted behind a pierced black cardboard to make a point light source. Since the position of focal point may be changed in each experiment, the camera focus is set to a sample point. Then, the camera is moved back and forth up to 60cm in 5cm increments and an image is captured at each depth. Each image is cropped according to the surface in which the point light spreads. Afterward, using some threshold values, the residual light is cleared and the result is normalized. In some rare cases, there is a jump in the PSF scale in consecutive measured PSFs. Under these conditions, other PSF scales are generated synthetically from the obtained PSFs. In this way, a bank of PSFs is generated that covers all possible sizes of PSF in the range [-19:+19]. The camera is set to F# = 2 and the illumination is set to office room lighting condition (i.e. 300 lux). Fig. 12(c, d) shows some calibrated PSFs in forward and backward points of focus.

In the first experiment, the focal point is set to the farthest point and all objects are placed in front of it. The captured images and results are shown in Fig. 13(a). The index number in the color-bar shows relative distance to the camera so that in each figure, the closer object is colored with smaller index.

Although the results are acceptable, there are some errors of depth estimation on the floor of the scene that should be corrected by the user or other segmentation techniques, which may not be so sensitive to intensity similarity.

In the second experiment, three objects are placed in the back of, over and in front of the focus point.

Fig. 13(b) shows the captured image along with the depth map. In the third experiment, the focal point is set to the nearest object with all other objects being placed behind that.

According to Fig. 13(c) our method can achieve acceptable results in this case.

Each depth-map is slightly corrected and then deblurring [4] is performed with the modified depth map. Fig. 13(c) shows all-focus images derived from deblurring.

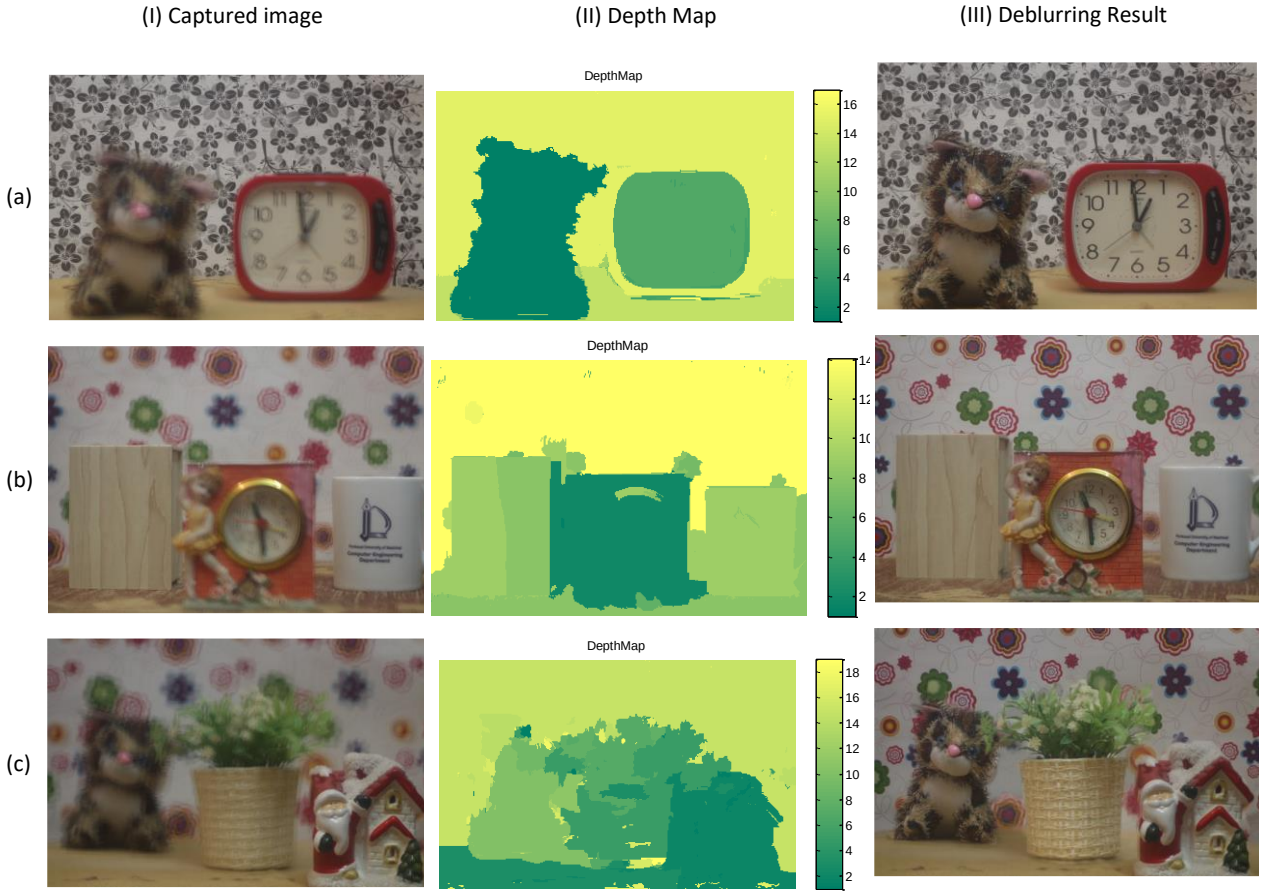


Fig. 13: Depth map estimation of depth varying scenes: (a) in front of the focal plane, (b) both sides of the focal plane, (c) at the back of the focal plane.

Conclusion

In this paper, a new method of aperture mask evaluation was proposed, which could reduce estimation error in both depth map and deblurring results. Asymmetric apertures make different PSFs in the back and front of the focal point. This feature could help discriminate blurred objects on two sides of the focal plane. The aperture pattern was designed for a specific imaging condition. Our future work will be concerned with defining an objective function in which the exposure time is also considered as an unknown variable of the problem and the SNR of captured images determines the lower bound of the mask throughput. Our proposed mask was intended for indoor illumination setting.

d

by considering the aperture throughput and imaging conditions, an exact evaluation of masks with different throughput could be done. Analytical and experimental results showed that our proposed mask could estimate an appropriate depth map of objects captured in one image regardless of the side of the focal plane. This was achieved with the help of a new depth estimation algorithm proposed in this article. According to the proposed algorithm, the deblurring result of correct PSF has the highest quality, which helps PSF estimation. Although the proposed no-reference quality measure yielded desirable results in depth estimation, more studies are required to obtain better measures which can reduce depth estimation error in both conventional and coded aperture imaging.

Author Contributions

M. Masoudifar designed and implemented the experiments, carried out the data analysis, and wrote the manuscript. H. Pourreza interpreted the results and revised the manuscript.

Acknowledgment

This work is completely self-supporting, thereby no financial agency's role is available. The authors gratefully thank the anonymous reviewers and the editor of JECEI.

Conflict of Interests

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and/or submission, and redundancy have been completely observed by the authors.

Abbreviations

DFD	Depth from defocus
PSF	Point Spread Function
NSGA	Non-dominated Sorting Genetic Algorithm
SNR	Signal to Noise Ratio
DoF	Depth of Field
MSE	Mean Square Error
RMSE	Root Mean Square Error

References

[1] A. P. Pentland, "A new sense for depth of field," IEEE Trans. Pattern Anal. Mach. Intell., 9(4): 523-531, 1987.

[2] M. Subbarao, N. Gurumoorthy, "Depth recovery from blurred edges," in Proc. Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'88: 498-503, 1988.

[3] A. Sellent, P. Favaro, "Which side of the focal plane are you on?," in Proc. 2014 IEEE International Conference on Computational Photography (ICCP): 1-8, 2014.

[4] A. Levin, R. Fergus, F. Durand, W. T. Freeman, "Image and depth from a conventional camera with a coded aperture," ACM Trans. Graphics, 26(3): 70, 2007.

[5] A. Sellent, P. Favaro, "Optimized aperture shapes for depth estimation," Pattern Recognit. Lett., 40: 96-103, 2014.

[6] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, J. Tumblin, "Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing," ACM Trans. Graphics, 26(3): 1-12, 2007.

[7] C. Zhou, S. Nayar, "What are good apertures for defocus deblurring?" in Proc. 2009 IEEE International Conference on Computational Photography (ICCP): 1-8, 2009.

[8] B. Masia, L. Presa, A. Corrales, D. Gutierrez, "Perceptually optimized coded apertures for defocus deblurring," Comput. Graphics Forum, 31(6): 1867-1879, 2012.

[9] Y. Takeda, S. Hiura, K. Sato, "Fusing depth from defocus and stereo with coded apertures," in Proc 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR): 209-216, 2013.

[10] C. Zhou, S. Lin, S. K. Nayar, "Coded aperture pairs for depth from defocus and defocus deblurring," Int. J. Comput. Vision, 93(1): 53-72, 2011.

[11] A. Konak, D. W. Coit, A. E. Smith, "Multi-objective optimization using genetic algorithms: A tutorial," Reliab. Eng. Syst. Saf., 91(9): 992-1007, 2006.

[12] K. Mitra, O. S. Cossairt, A. Veeraraghavan, "A framework for analysis of computational imaging systems: Role of signal prior, sensor noise and multiplexing," IEEE Trans. Pattern Anal. Mach. Intell., 36(10): 1909-1921, 2014.

[13] V. Paramonov, I. Panchenko, V. Bucha, A. Drogolyub, S. Zagoruyko, "Depth camera based on color-coded aperture," in Proc. the IEEE Conference on Computer Vision and Pattern Recognition Workshops: 1-9, 2016.

[14] V. Aslantas, "A depth estimation algorithm with a single image," Opt. express, 15(8): 5024-5029, 2007.

[15] S. Zhuo, T. Sim, "Defocus map estimation from a single image," Pattern Recognit., 44(9): 1852-1858, 2011.

[16] J. Lin, X. Ji, W. Xu, Q. Dai, "Absolute depth estimation from a single defocused image," IEEE Trans. Image Process., 22(11): 4545-4550, 2013.

[17] X. Zhu, S. Cohen, S. Schiller, P. Milanfar, "Estimating spatially varying defocus blur from a single image," IEEE Trans. Image Process., 22(12): 4879-4891, 2013.

[18] S. Gur, L. Wolf, "Single image depth estimation trained via depth from defocus cues," in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 7683-7692, 2019.

[19] P. Hambarde, S. Murala, "S2DNet: Depth estimation from single image and sparse samples," IEEE Trans. Comput. Imaging, 6: 806-817, 2020.

[20] A. N. Rajagopalan, S. Chaudhuri, "Optimal selection of camera parameters for recovery of depth from defocused images," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition: 219-224, 1997.

[21] M. Watanabe, S. K. Nayar, "Rational filters for passive depth from defocus," Int. J. Comput. Vision, 27(3): 203-225, 1998.

[22] P. Favaro, S. Soatto, "A geometric approach to shape from defocus," IEEE Trans. Pattern Anal. Mach. Intell., 27(3): 406-417, 2005.

[23] S. Matsui, H. Nagahara, R.I. Taniguchi, "Half-sweep imaging for depth from defocus," Image Vision Comput., 32(11): 954-964, 2014.

[24] M. Ye, X. Chen, Q. Li, J. Zeng, S. Yu, "Depth from defocus measurement method based on liquid crystal lens," Opt. Express, 26(2): 28413-28420, 2018.

[25] S. W. Hasinoff, K. N. Kutulakos, "Confocal stereo," Int. J. comput. vision, 81(1): 82-104, 2009.

[26] A. Rajagopalan, S. Chaudhuri, U. Mudenagudi, "Depth estimation and image restoration using defocused stereo pairs," IEEE Trans. Pattern Anal. Mach. Intell., 26(11): 1521-1525, 2004.

[27] S. Hiura, T. Matsuyama, "Depth measurement by the multi-focus camera," in Proc. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition: 953-959, 1998.

[28] O. Cossairt, M. Gupta, S. K. Nayar, "When does computational imaging improve performance?," IEEE Trans. Image Process., 22(2): 447-458, 2013.

[29] Y. Weiss, W. T. Freeman, "What makes a good model of natural images?," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07): 1-8, 2007.

[30] P. E. Debevec, J. Malik, "Recovering high dynamic range radiance maps from photographs," in ACM SIGGRAPH 2008 classes: 1-10, 2008.

[31] O. Cossairt, "Tradeoffs and limits in computational imaging," Columbia University, 2011.

[32] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," IEEE Trans. Evol. Comput., 6(2): 182-197, 2002.

[33] Y. Gao, "Population size and sampling complexity in genetic algorithms," in Proc. the Bird of a Feather Workshops: 178-181, 2003.

[34] S. Theodoridis, R. Chellappa, Academic Press Library in Signal Processing: Image, Video Processing and Analysis, Hardware, Audio, Acoustic and Speech Processing: Elsevier Science, 2013.

[35] M. Martinello, "Coded aperture imaging," Heriot-Watt University, Edinburgh, Scotland, 2012.

[36] Y. Liu, J. Wang, S. Cho, A. Finkelstein, S. Rusinkiewicz, "A no-reference metric for evaluating the quality of motion deblurring," ACM Trans. Graph., 32(6): 175, 2013.

[37] B. Hu, L. Li, J. Qian, "Perceptual quality evaluation for motion deblurring," IET Comput. Vision, 12(6): 796-805, 2018.

[38] D. Krishnan, T. Tay, R. Fergus, "Blind deconvolution using a normalized sparsity measure," in Proc. 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR): 233-240, 2011.

[39] G. Blanchet, L. Moisan, "An explicit sharpness index related to global phase coherence," in Proc. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): 1065-1068, 2012.

[40] M. Masoudifar, H. R. Pourreza, "Depth estimation from a single defocused image using no reference image quality assessment metrics," presented at the Fifth International Conference on Technology Development in Iranian Electrical Engineering, 2021.

[41] A. Delong, A. Osokin, H. N. Isack, Y. Boykov, "Fast approximate energy minimization with label costs," Int. J. comput. vision, 96(1): 1-27, 2012.

Biographies

Mina Masoudifar received the B.Sc. degree in computer engineering from Sharif University, Tehran, Iran, and the M.Sc. and Ph.D. degrees from Ferdowsi University of Mashhad, Iran, in 1996, 1999, and 2017, respectively. She is an Assistant Professor in computer engineering at the Department of Computer Engineering, Hakim Sabzevari University, Sabzevar, Iran. Her main research interests are machine vision, computational photography, image quality assessment, and deep learning.

- Email: masoudi@hsu.ac.ir
- ORCID ID: 0000-0002-9609-1853
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: http://staff.hsu.ac.ir/persons/?perid=100252

HamidReza Pourreza is currently Professor of Computer Science and Engineering at Ferdowsi University of Mashhad (FUM). He received his B.S. degree in Electrical Engineering from FUM in 1989, and received his M.S. and Ph.D. degrees in Electrical Engineering and Computer Engineering from Amirkabir University of Technology in 1993 and 2003, respectively. His research interests are in the areas of Deep Learning, Computer Vision, Hardware Design, and Intelligent Transportation Systems (ITS).

- Email: hpourreza@um.ac.ir
- ORCID ID: 0000-0002-3560-8070
- Web of Science Researcher ID: B-8754-2015
- Scopus Author ID: 23968187500
- Homepage: https://hpourreza.profcms.um.ac.ir/

How to cite this paper:
M. Masoudifar, H. R. Pourreza, "Depth estimation and deblurring from a single image using an optimized-throughput coded aperture," J. Electr. Comput. Eng. Innovations, 11(1): 51-64, 2023.

DOI: 10.22061/JECEI.2022.8630.537

URL: https://jecei.sru.ac.ir/article_1718.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Fabrication of Micro Glass Spherical Resonator by Chemical Foaming Process (CFP)

M. Kookhaee, A. Khooshehmehri, A. Eslami Majd*

Faculty of Electrical and Computer, Malek-Ashtar University of Technology, Tehran, Iran.

Article Info

Article History:
Received 05 January 2022
Reviewed 10 March 2022
Revised 26 April 2022
Accepted 07 May 2022

Keywords:
Hemispherical resonator gyroscope
The etched cavity in the silicon substrate
Anodic bonding
Glass blowing
Height to radius ratio of the shell

*Corresponding Author's Email Address: *a_eslamimajd@mut-es.ac.ir*

Abstract

Background and Objectives: The Hemispherical Resonator Gyroscope (HRG) is an inertial sensor which is a good choice for space missions and inertial navigation due to their low noise, low energy consumption, long life, and excellent accuracy and sensitivity. It consists of three main parts: the shell, the excitation and detection system, and the control circuits. In recent years, with using MEMS technology in the construction of HRG, vibrating shells with low volume and low price are made.

Methods: The hemispherical shell is the main part and the beating heart of hemispherical resonator gyroscopes and is responsible for sensing. An optimized shell is required to implement the excitation and detection system and operate the gyroscope properly. In this research, the structure of a spherical shell with an environmental base that does not need to release the shell from its environment for its excitation and detection system is selected and the relationships governing this type of shell to improve the parameters of the glass blowing method will be investigated. Also, all sub-processes of this type method of fabrication to optimize the glass-blown spherical shell are implemented.

Results: The process of making spherical shell by glass blowing using the chemical foaming process is used to obtain shells with height to radius ratio greater than 1, and finally, a glass shell with an etched cavity with a radius of 562 µm and depth of 524 µm created by the CNC process, with height to radius ratio of approximately 1.8 has been achieved. In this method, using direct transfer of calcium carbonate to the etched cavity, before anodic bonding, the glass shell volume has been increased from 0.602 nL to 1.04 nL.

Conclusion: The result is that to achieve a glass shell with a height to radius ratio of more than 1, in addition to improving the fabrication process, it is necessary to transfer the solid foaming agent to the etched cavity. Finally, in the fabrication of the glass-blown spherical shell, we have used the chemical foaming process (CFP) to obtain shells with a height to radius ratio greater than 1.

Introduction

Coriolis vibration gyroscopes are sensors in which the angle of rotation or angular velocity is measured using the Coriolis force applied to the vibrating mass [1]. One of the types of Coriolis vibration gyroscopes is the vibrating shell gyroscopes that vibrate with a hemispherical resonator or so-called "wine glass" [2], [3].

Hemispherical resonator gyroscopes (HRG) are considered a suitable choice for space missions and inertial navigation due to their low noise, low energy

consumption, long life, and excellent accuracy and sensitivity [4]. These types of gyroscopes can also be made microelectromechanically [5].

The structure of a hemispherical resonator gyroscope consists of three main parts: the shell, the excitation, and detection system, and the control system. The resonator shell is the main part and the beating heart of these gyroscopes and is responsible for sensing. The resonator shell needs a perfectly symmetrical structure for proper operation to show desirable characteristics in terms of equilibrium, natural frequency difference, vibration-induced deformation, and damping [6].

Due to the diversity of the structure of resonator shells, different methods are used to make the shell, which can be mainly referred to as two methods of silicon-based bulk and surface micromachining and surface tension processes [7]. In the construction of the shell by micromachining process using several stages of photolithography and etching, the resonator shell is created in the substrate [8].

The fabrication of the shell by surface tension processes is also based on the surface tension force at the melting point of the materials and is done in the two methods of blow torching [9] and glass blowing [10].

In 2014, Taheri-Tehrani et al. Reported the construction of a hemispherical shell with a diameter of 1 mm, depth of 250 µm, and thick of 1 µm by micromachining process. In this method, the diamond shell is fabricated by the chemical vapor deposition (CVD) process in cavities created by wet and isotropic etch [11].

In 2015, Khalil Najafi et al. Created shells using the blow torching process. In this fabrication method, the shell with its solid stem is separated from the mold and finally released using a type of wax and chemical mechanical polishing (CMP) [12]. In 2018, Dingbang Xiao et al. Also examined the process of making a shell by blowtorching and releasing it with a femtosecond laser to improve the cutting quality of the side walls [13], [14].

In 2015, Senkal et al. Reported the fabrication of a micro-fused silica shell using the glass blowing process [15].

In 2015, Senkal investigated the fabrication of pyrex shells by the glass blowing process. In the mentioned thesis, DRIE was used to etch cylindrical cavities with a central post to a depth of 250 µm on a 1mm silicon wafer. In this reference, to evaluate the structure of the hemispherical resonator shell, characteristics such as structure symmetry, shell surface roughness, and material composition before and after glass blowing have been analyzed [10].

In 2015, AM Shkel et al. Reported the fabrication of a spherical shell by glass blowing and surrounding electrodes, in which a pyrex micro-spherical resonator with a radius of 500 µm was made by glass blowing process and surrounded by four electrodes. In the process of fabricating this shell, cylindrical cavities with a radius of 265 nm and a depth of 800 µm are created by dry etching (DRIE) on a 1 mm silicon wafer [16].

AM Shkel et al. In 2011 and Binzhen Zhang et al. In 2016 proposed a method for glass-blown spherical shell fabrication with three-dimensional metal electrodes created at the same time as the shell [17], [18].

In 2018, Jianbing Xie et al. Used a method of transferring the solid foaming agent ($CaCO_3$) using a precipitation reaction to etched cavities by the DRIE process before anodic bonding to achieve larger sphericity. The tallest spherical glass shell, created by an etched cavity with a radius of 250 µm and a depth of 800 µm, has a height to radius ratio of 1.58 [19].

Jintang Shang et al. In 2011 and 2015 also used TiH_2 foaming agents to fabricate glass bubbles [20], [21].

An optimized shell is required to use the shell as a sensor and to measure the amount of input rotation to the resonator. In fact, to implement the excitation and detection system, one must first obtain a shell with the desired structural features to design and then implement the excitation and detection system according to its geometry. In this research, we have selected to implement a glass-blown spherical shell with an environmental base that does not need to release the shell from its surroundings for its excitation and detection system. The sub-processes of the glass blowing method include etching cavities, anodic banding, thinning and polishing, and blowing. We used the CNC process to create etched cavities in a silicon substrate. Using this method in creating cavities with great depth takes less time. After implementing all the sub-processes of this type of method of fabrication and improving their parameters, the glass shell with a height to radius ratio of more than 1, was still not obtained. To achieve a glass shell with a desired height to radius ratio (>1), in addition to improving the fabrication process, we used the direct transfer of the solid foaming agent (calcium carbonate) to the etched cavity before the anodic bonding process.

Finally, we have achieved a spherical shell with a height to radius ratio of approximately 1.8 by creating an etched cavity with a radius of 562 µm and a depth of 524 µm using the CNC process and direct transfer of calcium carbonate to the etched cavity. Although our cavity radius was greater, and our cavity depth was less than the values used in similar articles, we were able to achieve a glass shell very close to the sphere by the chemical foaming process.

Governing Relationships the Spherical Shell with Peripheral Base

To improve the glass blowing process, it is necessary to study the governing relationships of the spherical shell. The geometrical parameters of the glass-blown spherical

shell are shown in Fig. 1 where R_0 is the radius of the etched cavity, h is the depth of the etched cavity, R_g is the radius of the spherical shell and h_1 is the height of the spherical shell.

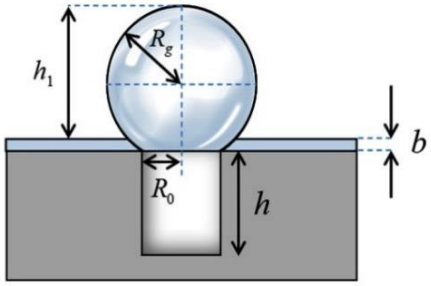


Fig. 1: Display of geometric parameters of a spherical shell with peripheral base.

The impact of gravity and the viscous force of the softened glass are neglected in the height model, and the thickness of the glass shell is considered to be uniform; hence, the volume of expanded gas confined obeys the ideal gas law.

Using the geometric parameters shown in Fig. 1, the volume of the etched cavity can be obtained from (1) [19]:

$$V_E = \pi R_0^2 h \tag{1}$$

According to Fig. 1, where R_0 is the radius of the etched cavity, h is the depth of the etched cavity. The volume of the glass shell can also be obtained using (2) [19]:

$$V_g = \left(\frac{T_f}{T_b} - 1\right) V_E \tag{2}$$

where V_g represents the volume of the glass shell, V_E represents the volume of the etched cylindrical cavity, T_f represents the furnace's heating temperature, and Tb represents the anodic bonding temperature. Thus, considering the volume of the etched cavity and the volume of the glass shell, the height of the glass shell as a function of the furnace temperature, bonding temperature, depth, and radius of the etched cavity can be obtained using (3) [19]:

$$h_1 = \frac{\left[\left(3V_g + \sqrt{R_0^6\pi^2 + 9V_g^2}\right)\pi^2\right]^{\frac{2}{3}} - R_0^2\pi^2}{\pi\left[\left(3V_g + \sqrt{R_0^6\pi^2 + 9V_g^2}\right)\pi^2\right]^{\frac{1}{3}}} \tag{3}$$

The mathematical equation describing the relationship between the height of the glass shell (h_1) and the radius of the glass shell (R_g) is (4) [19]:

$$R_g = \frac{R_0^2 + h_1^2}{2h_1} \tag{4}$$

Fabrication of the Glass-Blown Spherical Shell

The steps of fabricating a glass-blown spherical shell have five main sub-processes; include etching cavities, transfer the solid foaming agent to the etched cavity, anodic banding, thinning and polishing, and blowing. The schematic of this type of fabrication is shown in Fig. 2.

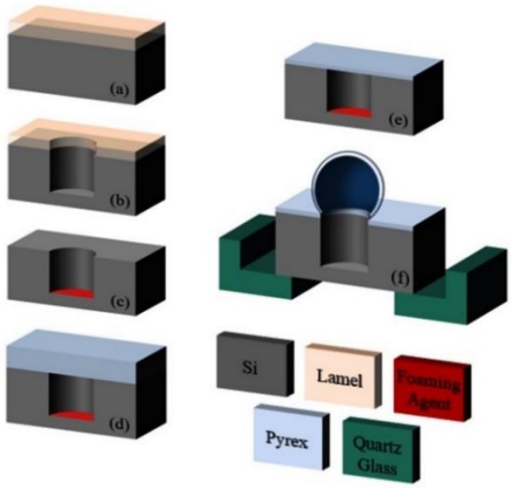


Fig. 2: Schematic of the process of fabricating a glass-blown spherical shell by chemical foaming process. (a) Paste the lamel to silicon and preparing it for the CNC process, (b) Creating a cylindrical cavity in a p-type silicon substrate using the CNC process, (c) Direct transfer of the solid foaming agent ($CaCO_3$) into the etched cavity, (d) Anodic bonding of pyrex layer to silicon substrate, (e) Decreasing the thickness of the pyrex layer to 200 µm using the thinning and polishing process, (f) Putting the sample on a quartz base and transfer it to a furnace to form a glass shell.

The young's modulus of the substrate has a greater effect on the quality factor (Q) of the resonator than its density, in other words, the substrate with a higher young's modulus leads to less energy loss. Silicon substrate with a higher young's modulus is a better choice than fused silica substrate [22]. Silicon substrate consists of two common types n and p [23], for the reason that the p-type silicon substrate has a better performance in the anodic bonding process than the n-type substrate, it is better to use this type of substrate [24].

In the cavities etching process, first, a p-type silicon substrate with a thickness of 740 µm is selected and a cavity with a radius and depth of 550 µm is created on it using the CNC process. In creating a cavity with a CNC machine, by measures such as changing the tool movement program, reducing the tool speed, changing the tool, and pasting the lamel on the silicon, it is possible to improve the lip and the dimensions of the cavity and bring it closer to the ideal (Fig. 2 (a), (b)). The crystalline orientation of the silicon substrate does not affect the creation of a cavity using CNC [30].

Then, to increase the height to radius ratio of the glass shell, the solid foaming agent is transferred into the etched cavity in the silicon (Fig. 2 (c)). Here, for the process of adding the foaming agent, calcium carbonate ($CaCO_3$) with the thermal decomposition temperature of 825 °C is used as a foaming agent, which its chemical

decomposition is in the furnace heating temperature in (5) [25]:

$$CaCO_3(S) \xrightarrow{900°C} CaO(S) + CO_2(g) \quad (5)$$

This material's thermal decomposition temperature is lower than that of the furnace (~900 °C) but greater than that of anodic bonding (~400 °C), thus it fits the requirements of the following procedure.

Calcium carbonate ($CaCO_3$) can be transferred into the etched cavity inside the silicon substrate in two ways, using direct transfer and the precipitation reaction of Na_2CO_3 solution and $CaCl_2$ solution. The chemical equation (6) is used to place the foaming agent into the etched cavity using a precipitation reaction:

$$Na_2CO_3(aq) + CaCl_2(aq) = CaCO_3(S) + 2NaCl \quad (6)$$

In the precipitation reaction method, two syringe pumps and two microliter syringes with a needle with an outer diameter of 190 µm are used to inject solutions into the cavity. The outer diameter of the syringe needle is smaller than the diameter of the etched cavity. The quantity of calcium carbonate in the etched cavity may be adjusted by varying the injected volume and concentration of the two reaction solutions. The sample is put on a hot plate after the micron-injection procedure, and the determined amount of calcium carbonate ($CaCO_3$) is left in the cavity [19].

In this paper, the direct transfer is used to transfer the solid foaming agent into the etched cavity in silicon. Directly transferring the solid foaming agent is very difficult. There are two primary causes for this difficulty: one, solid calcium carbonate (particularly powders) would cause difficult-to-remove bonding surface contamination, and the second, CaCO3 is insoluble in most solvents [19]. In the direct transfer of the solid foaming agent, calcium carbonate powder is first mixed with DI water to form a suspension. Using a microliter syringe with a needle with an outer diameter of 300 µm, the suspension is injected into the cavity. After DI water evaporates, calcium carbonate powder settles to the bottom of the cavity.

The cleanliness of the silicon surface is very important at this step because the powder particles remaining on the surface will not cause proper bonding of silicon and pyrex and will lead to problems in subsequent processes [24]. In this process, the amount of gas emitted by the foaming agent may be regulated by adjusting the number of moles of calcium carbonate ($CaCO_3$), which is equivalent to the number of moles of CO_2.

Next, the silicon wafer with the cavity is anodically bonded to a pyrex layer, at a temperature of 400 °C and a voltage of 1500 V, and the air and the solid foaming agent are trapped inside the etched cavity in the silicon (Fig. 2 (d)). In the anodic bonding process, the roughness of the two surfaces that are placed on top of each other is so important that two samples with a surface roughness of more than 50 nm in this process, at the higher the applied temperature and voltage, are not connected.

Bond strength is a critical element in the anodic bonding process since it is related to bond quality and dependability. A good bond is created when the bond strength is high. As the bonding temperature increases, the bond strength increases [26]. In the fabrication of a spherical shell by glass blowing, reducing the bonding temperature increases the air pressure trapped inside the etched cavity and improves the height to radius ratio of the shell. This decrease in bonding temperature may create unbonded points near the cavity, which will cause problems such as reduced air pressure inside the cavity, breaking pyrex during thinning, and asymmetry of the shell.

The use of mechanical tests, such as pressure, pull, shear, or bending tests to evaluate the anodic bonding process is not desirable because it destroys the specimen and makes it impossible to continue the fabrication process [24]. In this paper, an optical microscope, which is a non-destructive method, has been used to evaluate the anodic bonding process. In this method, using the color difference, the bonded spots and the unbonded spots are determined. A visual evaluation to observe unbonded spots is shown in Fig. 3.

Fig. 3: Examination of unbonded spots using color differences using an optical microscope.

The thickness of the pyrex layer bonded to the silicon substrate with an etched cavity should be something around a few hundred micrometers [27], [28]. In this paper, a thickness of 200 µm is considered as the desired thickness, which can be used to Achieve this desired thickness using the thinning and polishing process (Fig. 2 (e)).

In this step of the fabrication process, in order to form the glass shell, the bonded wafer with the pyrex layer facing upwards is placed onto a quartz glass stencil and then both together Are transferred into the quartz tube furnace (Fig. 2 (f)). The type of furnace used in this step

depends on the softening point of the bonded pyrex (820 °C). To prevent shock to the bonded wafer, it is transferred to the center of the furnace in 3.5 minutes.

While glass can be shaped at a broad range of temperatures, empirical tests show that if the furnace temperature is less than 800 °C, the glass spheres will take a long time to form. Also, because of the poor viscosity at higher temperatures, the spheres tend to break if the furnace temperature is higher than 950 °C. Therefore, the best temperature range for the furnace is between 850 °C and 900 °C [29].

By placing the bonded wafer at a temperature higher than the softening point of pyrex (870 °C) in a furnace, the trapped air in the cavity and the gas produced by the calcium carbonate expand and increase the pressure inside the cavity. However, this increase in pressure occurs via the uniform surface pressure distribution before driving the high-temperature molten glass membrane to reshape into a hollow shell. As the pyrex layer softens and the gas is released by the foaming agent at the furnace temperature, the thin pyrex layer on top of the cavity that has been etched in silicon substrate becomes a spherical glass shell. The surface tensile force in this method causes high symmetry and minimal roughness on the surface of the spherical shell [18].

After 20-60 seconds, the formed shells will be removed quickly to cool down in the air in order to avoid the collapse of the shells. The heating time should be chosen carefully because if the heating time is very long, the glass shell would deform even break. On the other hand, if the heating time is not long enough, the glass would not have enough time to become soften and blown into hemisphere shape [19].

Experimental Characteristics

In the governing relationships of the spherical shell, the radius of the etched cavity (R_0), the depth of the etched cavity (h), the temperature at which the etched cavity was bonded to a glass wafer (T_b), and the heating temperature in the furnace (T_f) are considered as inputs and the height of the spherical shell (h_1) and the radius of the spherical shell (R_g) is calculated as output.

The hollow glass shell that is more similar to a sphere could provide more favorable properties of the hemispherical resonator gyroscope. The larger the height to radius ratio of the glass shell, it is closer to being spherical and therefore more desirable.

The advantages of the high ratio of height to a radius of the glass shell can be referred to reducing four-node wineglass resonant frequency, which is useful for the excitation and detection of the HRG, and larger surface area for adjustment higher aspect surrounding capacitive electrodes, which can increase the sensitivity of HRG [19]. Therefore, two parameters of the height to radius ratio of the glass shell (h_1/R_g) and the height to diameter ratio of the glass shell ($h_1/2R_g$) are also considered as output parameters. The effect of changing the radius of the etched cavity on the height to radius ratio of the shell is shown in Fig. 4.

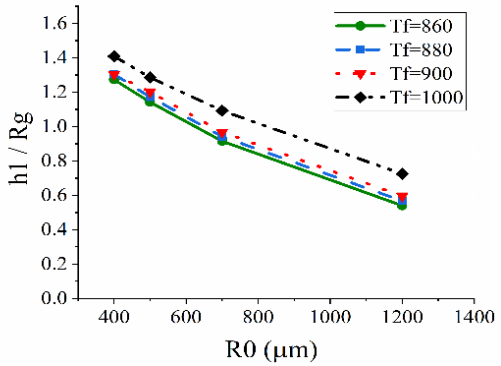


Fig. 4: The effect of changing the radius of the etched cavity on the height to radius ratio.

In the diagram in Fig. 4, the depth of the etched cavity (600 μm) and the anodic bonding temperature (400 °C) is considered constant, and the radius of the etched cavity and the heating temperature in the furnace has been changed. As can be seen in Fig. 4, as the radius of the etched cavity increases, the height to radius ratio of the glass shell decreases, and as the heating temperature in the furnace increases, the height to radius ratio of the glass shell increases. Also, the effect of changing the depth of the etched cavity on the height to radius ratio of the shell is shown in Fig. 5.

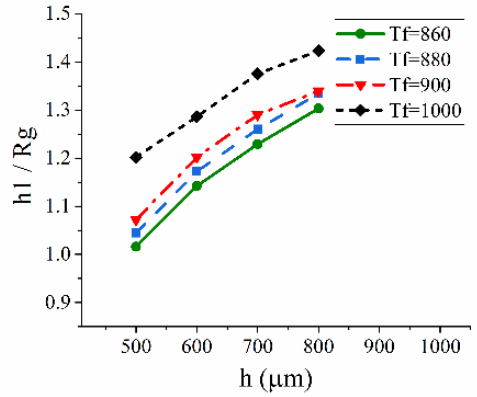


Fig. 5: The effect of changing the depth of the etched cavity on the height to radius ratio.

In the diagram in Fig. 5, the radius of the etched cavity (500 μm) and the anodic bonding temperature (400 °C) are considered constant, and the depth of the etched cavity and the heating temperature in the furnace has been changed. As can be seen in Fig. 5, as the depth of the etched cavity and the heating temperature in the furnace increase, the height to radius ratio of the glass shell increases. The effect of the temperature of anodic bonding change on the height to radius ratio of the shell is shown in Fig. 6.

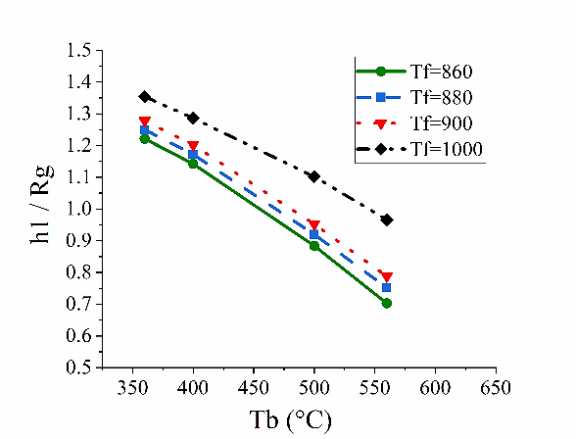


Fig. 6: The effect of the temperature of anodic bonding change on the height to radius ratio.

In the diagram in Fig. 6, the radius (500 µm) and depth (600 µm) of the etched cavity are considered constant, and the anodic bonding temperature and the heating temperature in the furnace have been changed. As can be seen in Fig. 6, as the anodic bonding temperature increases, the height to radius ratio of the glass shell decreases, and as the heating temperature in the furnace increases, the height to radius ratio of the glass shell increases. The effect of the heating temperature in the furnace change on the height to radius ratio of the shell is shown in Fig. 7.

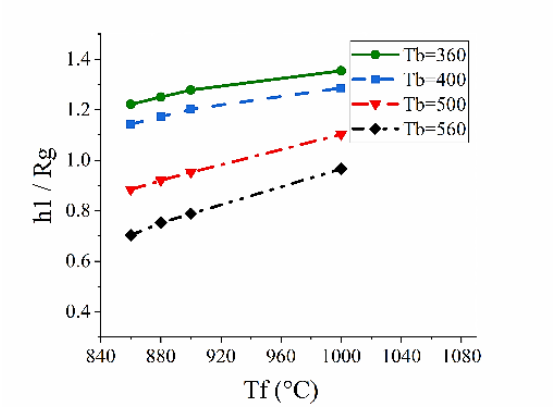


Fig. 7: The effect of the temperature in the furnace change on the height to radius ratio.

In the diagram in Fig. 7, Similar to the diagram in Fig. 6, the radius and depth of the etched cavity are considered constant, and the heating temperature in the furnace and the anodic bonding temperature have been changed. As it is known, in the diagram of Fig. 7, with increasing the heating temperature in the furnace, the height to radius ratio of the glass shell increases, and with increasing the anodic bonding temperature, the height to radius ratio of the glass shell decreases.

This result is obtained from the analysis of the above diagrams that to achieve the height to radius ratio of the glass shell of more than 1, the radius of the etched cavity and the anodic bonding temperature should be reduced and the depth of the etched cavity and the heating temperature in the furnace should be increased.

The depth and radius of the etched cavity, the thickness of the pyrex layer, the anodic bonding temperature, the temperature at which the glassblowing was executed, and even the cooling process of the softened glass shell all play an important role in the final shape of the glass shell [10], [19].

By performing the sub-processes of fabrication include etching cavities, anodic banding, thinning and polishing, and blowing, the glass shell is formed. It is difficult to obtain a glass shell with a height to radius ratio greater than 1 with these sub-processes. To obtain the glass shell with a height to radius ratio of more than 1, methods such as reducing the radius of the etched cavity and the anodic bonding temperature and increasing the depth of the etched cavity and the heating temperature in the furnace have been performed. These changes applied have been effective in increasing the height to radius ratio of the glass shell, but the result is far from a spherical shell, and the volume of the glass shell is still under restrictions from the process parameters. In addition, applying these changes has been faced challenges. For example, a thick silicon substrate is needed to increase the depth of the etched cavity (800-1000 µm). In this paper, to increase the depth of the etched cavity, a layer of pyrex with a thickness of 2 mm is anodically bonded to the double side polished silicon substrate from the back to provide the depth of the cavity to be increased. Then, to create an etched cavity using the CNC process, the sample is pasted on the lam in such a way that the silicon is facing up, and a lamel is glued on it to improve the lip. The image and schematic of this prepared sample are shown in Fig. 8.

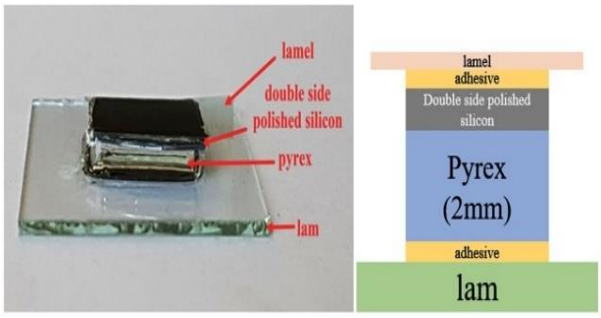


Fig. 8: Anodic bonding the pyrex wafer to the double side polished silicon substrate to increase the depth of the etched cavity.

Increasing the thickness of the sample causes prolongation of time in the fabrication process and increases the cost. In addition, it affects subsequent processes and makes the anodic bonding and dicing processes difficult [19], [29]. One of the problems with increasing the thickness is the need to apply a very high voltage (2600 V) to perform the anodic bonding process

of the upper pyrex layer at low temperatures (400 °C). Applying excessive voltage during the anodic bonding process increases the likelihood of the sample sparking and breaking. Demonstration sample 1 is shown in Fig. 9.

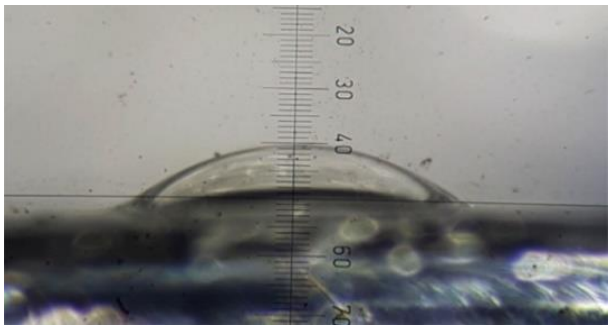


Fig. 9: Demonstration glass shell created by optimizing the four sub-processes of etching cavity, anodic banding, pyrex thinning and polishing, and blowing.

In Fig. 9, the glass shell is created by optimizing the four sub-processes of etching cavity, anodic banding, pyrex thinning and polishing, and blowing.

As shown in this figure, the glass shell created is spaced from the spherical shell and has a height to radius ratio of below 1.

Results and Discussion

In this work, to achieve a glass shell with a height to radius ratio of more than 1, in addition to improving the fabrication sub-processes, transferring the foaming agent to the etched cavity has also been used.

First, two cavities with the same conditions and dimensions are created using a CNC process inside a silicon substrate and some foaming agent ($CaCO_3$) is transferred into one of these cavities.

Fig. 10 (a) shows a spherical shell blown via an etched cavity in silicon substrate without a foaming agent and Fig. 10 (b) shows a spherical shell blown via an etched cavity in a silicon substrate with a quantified foaming agent.

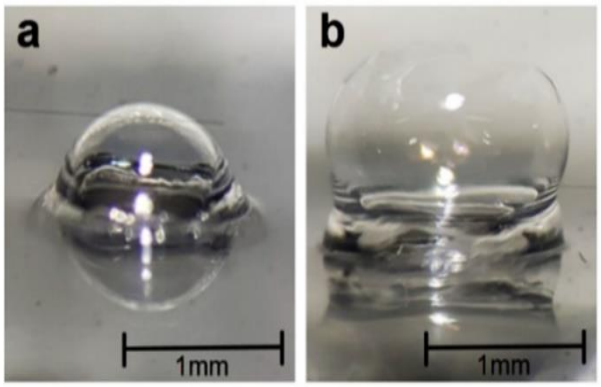
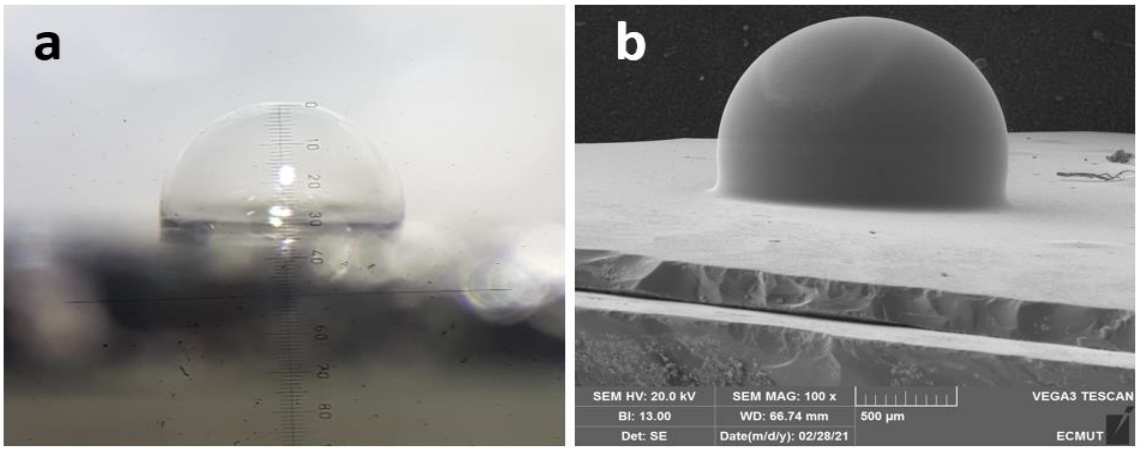


Fig. 10: (a) Spherical shell blew via an etched cavity in silicon substrate without a foaming agent, (b) Spherical shell blew via an etched cavity in a silicon substrate with a quantified foaming agent.

As shown in Fig. 10, the height to radius ratio of the glass shell that in its fabrication process used the foaming agent is greatly improved compared to the glass shell that there is only air in the etched cavity.

The height to radius ratio of the hemispherical shell resonators (HSRs), With the addition of foaming agent in the etched cavity to a depth of 200 µm may approach (even exceed) the shell has been blown by the etched cavity to a depth of 800 µm without no addition foaming agent [19].

Fig. 11 shows the optical microscope and scanning electron microscope (SEM) images of spherical shells created by the chemical foaming process (samples 2 and 3).



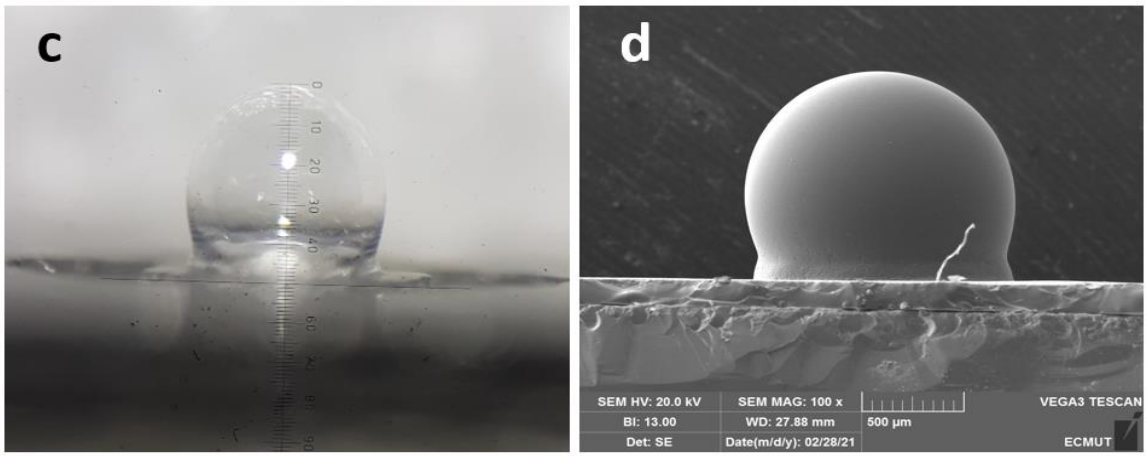


Fig. 11: Optical microscope and scanning electron microscope (SEM) images of spherical shells created by the chemical foaming process. (a, b) Sample 2. (c, d) Sample 3.

The experimental parameters and the experimental results of the samples in Fig. 11 are reported in Table 1. According to the governing relationships of the spherical shell, if the foaming agent not used in the fabrication of the glass shell of sample 2, ideally it should have reached a height of 472 µm and a radius of 550 µm, in which case the height to radius ratio of the shell is below 1. Sample 2, using the foaming agent ($CaCO_3$), the approximate height and radius of the glass shell 931 µm and 655 µm were achieved, respectively, and the height to radius ratio of the shell reached 1.42.

Also, if the foaming agent not used in the fabrication of the glass shell of sample 3, ideally it should have reached a height and radius of 553 µm and 560 µm, respectively, in which case the height to radius ratio of the shell is less than 1. Sample 3, using the foaming agent ($CaCO_3$), has reached the approximate height and radius of the glass shell of 1129 µm and 637 µm, respectively. As can be seen, the glass shell of sample 3 is very close to the sphere and its height to radius ratio is approximately 1.8.

Table 1: The experimental parameters and the experimental results of the glass shell samples

	Input							Output					
	R_0 (μm)	h (μm)	e (μm)	b (μm)	T_b ($^\circ C$)	T_f ($^\circ C$)	t (min)	h_1 ($predicted$) (μm)	h_1 (μm)	R_g ($predicted$) (mm)	R_g (μm)	$\frac{h_1}{R_g}$ ($predicted$)	$\frac{h_1}{R_g}$
Sample	the radius of the etched cavity	the depth of the etched cavity	edge of the etched cavity	the thickness of the glass wafer	the anodic banding temperature	the heating temperature in the furnace	the heating time in the furnace	the predicted height of the glass shell	the height of the glass shell	the predicted radius of the glass shell	the radius of the glass shell	the predicted height to radius ratio of the glass shell	the height to radius ratio of the glass shell
1	587.5	796	20	200	400	870	1.5	732	250	0.60	-	1.216	-
2	550	422	10	175	400	870	0.45	472	930.80	0.55	654.78	0.858	1.42
3	562.5	524	10	125	400	870	0.33	553	1128.71	0.56	636.25	0.987	1.77

Conclusion

In this study, to optimize the glass-blown spherical shell, all the sub-processes of this type of fabrication method, including etching cavity, anodic banding, pyrex thinning and polishing, and blowing, have been carefully investigated and implemented in practice. In spherical shell optimization, the result is that to achieve a glass shell with a height to radius ratio of more than 1, in addition to

improving the fabrication sub-processes, it is also necessary to transfer the foaming agent to the etched cavity. Finally, using the chemical foaming process (CFP) and direct transfer of calcium carbonate to the etched cavity by the CNC process with a radius of 562 µm and a depth of 524 µm, before the anodic bonding process, a glass shell with a height to radius ratio of approximately 1.8 has been obtained.

Author Contributions

All the authors participated in the conceptualization, implementation and M. Kookhaee wrote the manuscript.

Funding

The authors received no special funding for this effort.

Acknowledgment

The authors would like to acknowledge the Faculty of Electrical & Computer Engineering, Malek Ashtar University of Technology, and the Microelectronic laboratory, for their support and contribution to this study.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

MEMS	Micro-Electro-Mechanical Systems
HRG	Hemispherical Resonator Gyroscope
DRIE	Deep Reactive Ion Etching
CVD	Chemical Vapor Deposition
CNC	Computer Numerical Control
SEM	Scanning Electron Microscopy
CMP	Chemical Mechanical Polishing

References

[1] J. Lee, S. Yun, J. Rhim, "Design and verification of a digital controller for a 2-Piece hemispherical resonator gyroscope," Sensors, 16(4): 555, 2016.

[2] A. M. Shkel, "Type I and type II micromachined vibratory gyroscopes," in Proc. IEEE/ION PLANS 2006: 586-593, 2006.

[3] Z. Xu, G. Yi, M. Er, C. Huang, "Effect of uneven electrostatic forces on the dynamic characteristics of capacitive hemispherical resonator gyroscopes," Sensors, 19(6): 1291, 2019.

[4] A. A. Trusov, M. R. Phillips, G. H. Mccammon, D. M. Rozelle, A. D. Meyer, "Continuously self-calibrating CVG system using hemispherical resonator gyroscopes," in Proc. 2015 IEEE International Symposium on Inertial Sensors and Systems (ISISS): 1-4, 2015.

[5] A. Heidari, M. L. Chan, H. A. Yang, G. Jaramillo, P. Taheri-Tehrani, P. Fonda, H. Najar, K. Yamazaki, L. Lin, D. A. Horsley, "Hemispherical wineglass resonators fabricated from the microcrystalline diamond," J. Micromech. Microeng., 23(12): 125016, 2013.

[6] Z. Xu, G. Yi, H. Fang, Y. Cao, L. Hu, G. Zhang, "Influence of elasticity modulus on the natural frequency in hemispherical resonator," in Proc. 2019 IEEE DGON Inertial Sensors and Systems (ISS): 1-11, 2019.

[7] B. Luo, M. A. Zhang, C. Lu, J. Shang, "Wafer-level fabrication of silicon-in-glass electrodes for electrostatic transduction," in Proc. 2016 IEEE 66th Electronic Components and Technology Conference (ECTC): 1278-1283, 2016.

[8] J. J. Bernstein, M. G. Bancu, J. M. Bauer, E. H. Cook, P. Kumar, E. Newton, T. Nyinjee, G. E. Perlin, J. A. Ricker, W. A. Teynor, M. S. Weinberg, "High Q diamond hemispherical resonators: fabrication and energy loss mechanisms," J. Micromech. Microeng., 25(8): 085006, 2015.

[9] J. Cho, J. Yan, J. A. Gregory, H. Eberhart, R. L. Peterson, K. Najafi, "High-Q fused silica birdbath and hemispherical 3-D resonators made by blow torch molding," in Proc. 2013 IEEE 26th International Conference on Micro Electro Mechanical Systems (MEMS): 177-180, 2013.

[10] D. Senkal, "Micro-glassblowing Paradigm for Realization of Rate Integrating Gyroscopes," PhD Thesis University of California, Irvine, 2015.

[11] P. Taheri-Tehrani, T. Su, A. Heidari, G. Jaramillo, C. Yang, S. Akhbari, H. Najar, S. Nitzan, D. Saito, L. Lin, D.A. Horsley, "Micro-scale diamond hemispherical resonator gyroscope," in Proc. Hilton Head Workshop: 289-292, 2014.

[12] J. Y. Cho, K. Najafi, "A high-q all-fused silica solid-stem wineglass hemispherical resonator formed using micro blow torching and welding," in Proc. 2015 28th IEEE International Conference on Micro Electro Mechanical Systems (MEMS): 821-824, 2015.

[13] W. Li, Z. Hou, Y. Shi, K. Lu, X. Xi, Y. Wu, X. Wu, D. Xiao, "Application of micro-blowtorching process with whirling platform for enhancing frequency symmetry of microshell structure," J. Micromech. Microeng., 28(11): 115004, 2018.

[14] Y. Shi, X. Xi, Y. Wu, W. Li, K. Lu, Z. Hou, X. Wu, D. Xiao, "Wafer-level fabrication process for micro hemispherical resonators," in Proc. 2019 IEEE 20th International Conference on Solid-State Sensors, Actuators and Microsystems & Eurosensors XXXIII (TRANSDUCERS & EUROSENSORS XXXIII): 1670-1673, 2019.

[15] D. Senkal, M. J. Ahamed, M. H. Asadian Ardakani, S. Askari, A. M. Shkel, "Demonstration of 1 million q-factor on microglassblown wineglass resonators with out-of-plane electrostatic transduction," J. Microelectromech. Syst., 24(1): 29-37, 2015.

[16] J. Giner, A. M. Shkel, "The concept of "collapsed electrodes" for glassblown spherical resonators demonstrating 200: 1 aspect ratio gap definition," in Proc. 2015 IEEE International Symposium on Inertial Sensors and Systems (ISISS): 1-4, 2015.

[17] I. P. Prikhodko, S. A. Zotov, A. A. Trusov, A. M. Shkel, "Microscale glass-blown three-dimensional spherical shell resonators," J. Microelectromech. Syst., 20(3): 691-701, 2011.

[18] R. Wang, B. Bai, H. Feng, Z. Ren, H. Cao, C. Xue, B. Zhang, J. Liu, "Design and fabrication of micro hemispheric shell resonator with annular electrodes," Sensors, 16(12): 1991, 2016.

[19] J. Xie, L. Chen, H. Xie, J. Zhou, G. Liu, "The application of chemical foaming method in the fabrication of micro glass hemisphere resonator," Micromachines, 9(2): 42, 2018.

[20] J. Shang, B. Chen, W. Lin, C.P. Wong, D. Zhang, C. Xu, J. Liu, Q. A. Huang, "Preparation of wafer-level glass cavities by a low-cost chemical foaming process (CFP)," Lab on a Chip, 11(8): 1532-1540, 2011.

[21] B. Luo, J. Shang, Y. Zhang, "Hemipherical wineglass shells fabricated by a Chemical foaming process," in Proc. 2015 IEEE 16th International Conference on Electronic Packaging Technology (ICEPT): 951-954, 2015.

[22] A. Darvishian, B. Shiari, J. Y. Cho, T. Nagourney, K. Najafi, "Anchor loss in hemispherical shell resonators" J. Microelectromech. Syst., 26(1): 51-66, 2017.

[23] G. S. May, S. M. Sze, Fundamentals of Semiconductor Fabrication, USA: Hoboken, John Wiley & Sons, 2004.

[24] A. C. Lapadatu, H. Jakobsen, Handbook of Silicon Based MEMS Materials and Technologies (Second Edition), Chapter 30, Anodic Bonding (pp. 599-610) Elsevier BV, 2015.

[25] R. Barker, "The reversibility of the reaction CaCO3⇄ CaO+ CO2" J. Appl. Chem. Biotechnol., 23(10): 733-742, 1973.

[26] J. Wei, H. Xie, M. L. Nai, C. K. Wong, L. C. Lee, "Low temperature wafer anodic bonding" J. Micromech. Microeng., 13(2): 217-222, 2003.

[27] C. Zhang, A. Cocking, E. Freeman, Z. Liu, S. Tadigadapa, "On-Chip glass microspherical shell whispering gallery mode resonators," Sci. Rep., 7(1): 14965, 2017.

[28] J. Giner, L. Valdevit, A. M. Shkel, "Glass-blown pyrex resonator with compensating Ti coating for reduction of TCF," in Proc. 2014 IEEE International Symposium on Inertial Sensors and Systems (ISISS): 1-4, 2014.

[29] E. J. Eklund, A. M. Shkel, "Glass blowing on a wafer level," J. Microelectromech. Syst., 16(2): 232-239, 2007.

[30] A. Khooshehmehri, A. Eslami Majd, S. A. Hosseini, "Design and fabrication of hemispherical shell resonator by glass blowing method," J. Electr. Comput. Eng. Innovations, 10(1): 37-46, 2022.

Biographies

Maryam Kookhaee was born in tehran in iran, 1993. She received the B.Sc. degree in Bioelectrical Engineering from Sahand University of Technology, Tabriz, Iran, in 2016. Also, she received the M.Sc. degree in Electrical Engineering from Malek-Ashtar University of Technology in 2021. Her research interests include Coriolis Gyros, MEMS, and semiconductor field effect devices.

- Email: maryam_kookhaee@mut.ac.ir
- ORCID: 0000-0002-1991-6954
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Aylar Khooshehmehri received the B.Sc. degree in Electrical Engineering from K. N. Toosi University of Technology, Tehran, Iran, in 2007. Also, she received the M.Sc. and Ph.D. degree in Electrical Engineering from Malek-Ashtar University of Technology in 2011 and 2018, respectively. Her research interests include superconducting amplifying and detector devices, MEMS, and semiconductor field effect devices.

- Email: khooshehmehri@mut.ac.ir
- ORCID: NA
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Abdollah Eslami Majd was born in Hamadan, Iran, on March 23, 1976. He received the B.E. degree in applied physics from bu-ali Sina University, Hamadan, Iran in 1998. He received M.E. degree in atomic and molecular physics from Amir kabir University of Technology Tehran Polytechnic, Tehran, Iran in 2001. He received the Ph.D. Degree in photonics from laser and plasma Institute of Shahid Beheshti University, Tehran, Iran in 2011. Since joining electrical engineering and electronic department of Malek Ashtar University of Technology in 2012, he has engaged in research and development of stary light in the satellite camera, laser induced breakdown spectroscopy (LIBS) and hemispherical resonator gyroscope (HRG). He is co-author of more than 30 publications. Dr. Eslami is a member of Optics and Photonics Society of Iran and Physics Society of Iran.

- Email: a_eslamimajd@mut-es.ac.ir
- ORCID: 0000-0002-7538-3160
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:
M. Kookhaee, A. Khooshehmehri, A. Eslami Majd, "Fabrication of micro glass spherical resonator by Chemical Foaming Process (CFP)," J. Electr. Comput. Eng. Innovations, 11(1): 65-74, 2023.

DOI: 10.22061/JECEI.2022.8737.549

URL: https://jecei.sru.ac.ir/article_1719.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Design and Fabrication of Coaxial Plasma Waveguide Filter with the Ability to Reconfigure the Frequency Band

S. H. Mohseni Armaki*, M. Tohidlo, M. Kazerooni

Faculty of Electrical and Computer Engineering, Malek-Ashtar University of Technology, Iran.

Article Info

Article History:
Received 06 February 2022
Reviewed 17 March 2022
Revised 30 May 2022
Accepted 01 June 2022

Keywords:
Coaxial plasma waveguide
Plasma frequency
AC plasma excitation
Reconfiguration
Transverse electromagnetic
Cognitive radio

*Corresponding Author's Email Address: *mohseni@mut.ac.ir*

Abstract

Background and Objectives: This study aims to present a new structure based on coaxial waveguide, which can change the bandwidth, return losses, and input impedance by changing the plasma parameters of the coaxial waveguide. This structure consists of a metal body and a gas tube inside it, which uses a high voltage alternating current converter, can change the plasma parameters and, consequently the waveguide parameters. The input and output of the waveguide are also designed using the indirect capacitive coupling method.

Methods: In the Field of plasma research and related emerging technologies, recently, it has achieved a special place in various industries such as radar and Aerospace industries. The creation of telecommunication structures such as antennas and Waveguides with plasma, has given features such as adaptability, the ability to reconfigure the characteristics of the structure, and improve the sensitivity of this type of structure.

Results: By applying and changing the plasma excitation parameters, a change in the bandwidth was observed in the frequency band range of 0.5-4 GHz and a maximum of 1.38 GHz. Also, increasing the intensity of the excitation current improved the return losses in the resonance frequencies and, on the other hand, increased the band ripple.

Conclusion: According to the results, the change of Plasma parameters depends on the change of plasma excitation frequency, and the value of Excitation current applied. As the Value of excitation current increases, the matching to the resonance frequencies improves, but on the other hand, the passband ripple of the plasma waveguide filter increases. As the plasma excitation pulse frequency increases, the bandwidth and resonance frequencies change to higher frequencies, and the matching to the resonance frequencies improves. But on the other hand, the passband ripple increases. This new waveguide filter can be used in cognitive/ adaptive telecommunication systems due to the constant change of frequency band.

Introduction

With the idea of using Plasma as a substitute for metal in telecommunication structures [1], researchers have made great efforts make the most of this material with its special properties. In recent years, various researches have been conducted in the field of plasma structures in radio frequencies, including plasma waveguides [2], plasma antennas [3]-[4]. and, frequency selective Surfaces [5]-[6]. Plasma is a highly ionized gas whose number of free electrons is approximately equal to the number of its positive ions and is commonly referred to as the fourth state of matter [7]. The existence of plasma

was first proven by Sirviliam Crooks in 1879.

Plasma can be generated by a variety of methods, some of which include: AC and DC excitation [8]-[9], radio frequency (RF) excitation [10], high power pulsed laser [11], and high-energy (nuclear) methods [12]. Meanwhile, the high plasma ionization capability has made it possible to use it as a substitute for metal in microwave conducting structures [13]. In metal, free electrons move and radiate along the metal conductor, causing electromagnetic fields to pass through or radiate. in Plasma, electrons released from positive ions formed during the ionization process [14]. They pass or radiate electromagnetic fields. The difference between metal and plasma RF structures does not end here. For example, plasma waveguides, unlike metal waveguides, have a reconfiguration property [15] and can be changed and controlled by plasma parameters such as plasma frequency, collision frequency, and plasma density, waveguide parameters from Sentences change the frequency bandwidth, Reflection coefficient of the passing band, ripple of the passing band and input impedance, a new generation of controllable and flexible waveguides. Certainly, the nanosecond rate of change of plasma parameters is remarkable compared to the speed of mechanical change of metal structures and the many advantages of Plasma. Waveguides are used in telecommunication systems to transmit a wave from the generating part to the antenna and vice versa or to transmit a wave between different parts of a RF system [16]. Waveguides have different dimensions, shapes, and types depending on the application and transmission wave parameters [17].

Due to the dependence of their parameters on their physics, waveguides also have different filtering capabilities and, consequently, have their frequency bandwidth [18]. In some telecommunication systems, such as some meteorological and monitoring radars, depending on transmitter/receiver system, it is sometimes necessary to change the frequency of the wave transmitted from the generator and then transmit it to the antenna [19]. In this case, the use of broadband waveguides will be used. Broadband waveguides have their advantages and disadvantages as power limitation, ripple bandwidth, and fixed input impedance, and high manufacturing costs [20].

In the field of adaptive/ cognitive radars, continuous and instantaneous frequency band change is very important [21]-[22]. Therefore, various filters have been designed and manufactured for this purpose. Configurable filters are usually microstrip, which is pass frequency band controlled by MEMS devices or PIN Diode. The most important disadvantage of these structures is the low power and step change(Discontinuity in change) of bandwidth [23]-[24].

The system proposed in this paper is a reconfigurable coaxial plasma filter waveguide in terms of the frequency band, ripple bandwidth, and input impedance. The system follows a coaxial waveguide-based plasma structure consisting of a plasma tube, body, capacitive couplers, and alternating high voltage excitation circuit. The proposed system will be able to change the plasma parameters such as plasma frequency and collision frequency by changing the output frequency or input current of the excitation circuit. By changing the plasma parameters, the waveguide parameters can be configured and controlled. Weakpoints of the proposed structure are sensitivity to temperature stresses, need for independent high voltage excitation circuit, and more passband ripple than conventional filters. Section 2 deals with the theory and parameters of Plasma. Section 3 deals with the results of coaxial plasma waveguide simulation, and Section 4 describes the laboratory method of coaxial plasma waveguide test with the proposed Excitation. Section 5 deals with the results of applying Excitation current waveforms at variable frequencies to coaxial waveguide parameters. Finally, the conclusion will be made.

Theory and Parameters of Plasma

The fourth state of matter is called Plasma. Plasma is a quasi-neutralized ionized gas that has lost all or a significant portion of its atoms to one or more electrons and become positive ions. This highly ionized gas equals the number of free ions in its positive electrons. The degree of ionization can vary from 100% (fully ionized gases) to low degrees (partially ionized) [25]. Plasma can be created by electric and magnetic fields, radiated heating, and laser excitation. The electric method itself is divided into two sections: alternating current and direct current. In the field of plasma antennas, it should always be noted that the plasma frequency (ω_p) is quite different from the frequency of the RF Structure (ω) and must be distinguished. The plasma frequency is the measure of plasma ionization, while the frequency of the plasma antenna is the frequency at which the plasma antenna transmits and receives. The plasma frequency of a metal antenna in the X-ray range of the stabilized electromagnetic spectrum means that it has a plasma frequency equal to 30 PetaHertz(3×10^{16}) up to 30 ExaHertz(3×10^{19}). Still, the plasma frequency of the plasma antenna can vary. Plasma, an environment that contains free charge, generates natural oscillations due to thermal and electrical disturbances. Because of these coordinate oscillations, the density of electrons can oscillate around the angular frequency (ω_p) [26].

Because the Plasma is a dispersive material, it has its own electrical and magnetic properties, which occur at

different excitations, each depending on the type of Excitation. As mentioned, the plasma environment is homogeneous, nonlinear, and dispersive in terms of electromagnetic properties. Therefore, its electrical and magnetic parameters can vary depending on the frequency and other factors, and consequently the Plasma is an environment with special properties. Thus, the Plasma behaves differently against the electromagnetic waves emitted at each specific frequency and different degrees of ionization. Electromagnetic waves are transmitted, scattered, or transmitted by radiation to the Plasma [27]-[28].

The relation between the electrons and the electric field in the excitation state with alternating current is as follows [10]:

$$F = eE = eE_0 e^{-j\omega t} = \frac{d}{dt}(mv) \quad (1)$$

$$\frac{d}{dt}(mv) = m\frac{dv}{dt} + mv\upsilon_c \quad (2)$$

$$v = \left(\frac{e}{m}\right)\frac{1}{\upsilon_c - j\omega}E \quad (3)$$

where F is the electric force, υ_c is the plasma collision frequency, v is the velocity of the electron under the field E, e is the electron charge, and m is the mass of the electron. These interpretations, plasma inner surface current, are defined by (4):

$$J = n_e ev = \left(\frac{n_e e^2}{m}\right)\frac{1}{\upsilon_c - j\omega}E \quad (4)$$

where n_e is the density of free electrons per cubic meter. Plasma discharge power can be written as (5), and Plasma electrical permeability is also described by (6):

$$P = J.E = \left(\frac{n_e e^2}{m}\right)\frac{E^2 e^{-2j\omega t}}{\upsilon_c - j\omega} \quad (5)$$

$$\varepsilon_r = 1 - \frac{\omega_p^2}{\omega(\omega - j\Upsilon)} = 1 - \frac{\omega_p^2}{\omega^2 + \Upsilon^2} - \frac{j\Upsilon}{\omega}\frac{\omega_p^2}{\omega^2 + \Upsilon^2} \quad (6)$$

$$\omega_p = \left(\frac{n_e e^2}{m_e \varepsilon_0}\right)^{\frac{1}{2}} \quad (7)$$

$$f_p = \frac{\omega_p}{2\pi} \approx 9000\sqrt{n_e} \quad (Hz) \quad (8)$$

$$\gamma = \alpha + j\beta = jk_0\sqrt{\mu_r \varepsilon_r} \quad (9)$$

where ω_p is the plasma frequency, and Y is the natural collision frequency of the electron. The Plasma in this experiment has low temperature and is unbalanced. In other words, the temperature of the electrons is higher than the temperature of the ions [29]. As a result, the plasma frequency is calculated according to (7). By placing the values of the charge and mass of the electron on (7), (8) will result. According to (6), if the wave frequency delivered to the plasma surface is greater than the plasma frequency, that is, $\omega > \omega_p$, in this case, the propagation constant (γ) is imaginary (9), and the Plasma is found for the wave as a transparent medium, and the wave from It passes. However, if the frequency of the wave delivered to the plasma is lower than the plasma frequency, that is $< \omega_p$, in this case, the wave propagation constant is real, the wave does not pass through the Plasma, and the Plasma acts as a metal [27].

Simulation of Coaxial Plasma Waveguide Filter

Initially, to evaluate the proper functioning of the coaxial plasma waveguide filter, it was simulated with CST software. This software, which is one of the most powerful software in the field of antenna and microwave simulation, can simulate Dispersive environments with special properties such as plasma. The model used in this software for plasma simulation is called the Drude model. to create the structure of Plasma, this model requires two main parameters of Plasma, namely collision frequency, and plasma frequency. Fig. 1 and Fig. 2 show the simulation scheme of a coaxial plasma waveguide filter.

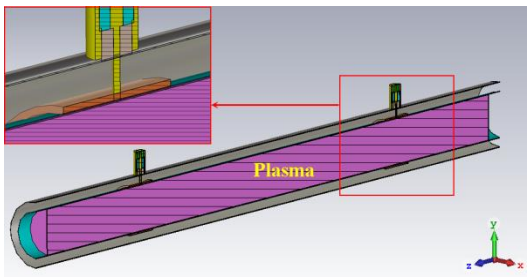


Fig. 1: Capacitive coupler view with connector and cut view of coaxial plasma waveguide filter.

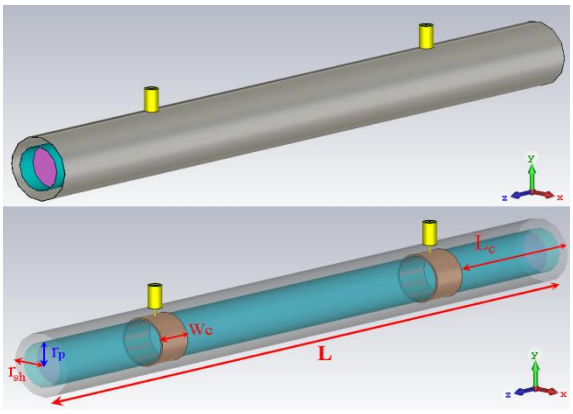


Fig. 2: Full and transparent view from inside the structure of coaxial plasma waveguide filter with details and parameters.

physical parameters of waveguide filter, including waveguide length, tube glass radius, Shield radius,

coupler width, and capacitive coupler distance from the end of the waveguide.It is presented in Table 1. It should be noted that the values of these parameters are defined based on the actual values of the 9-watt linear fluorescent lamp.

Table 1: Physical parameters of simulated plasma waveguide filter

Size (mm)	parameter of the waveguide
200	waveguide length(L)
7	tube glass radius(r_{p})
20	Shield radius (r_{sh})
10	coupler width(W_c)
30	capacitive coupler distance from the end of the waveguide(L_c)

Then, by calculating these two parameters using the above relations and in different excitation frequency values, the plasma frequency value (ω_p) is 3.6×10^{11}, 2.6×10^{11}, and 1.6×10^{11}, and the collision frequency (υ_c) is defined in all three cases 4×10^{8}. Fig. 3 shows the S-parameter for different plasma frequencies.

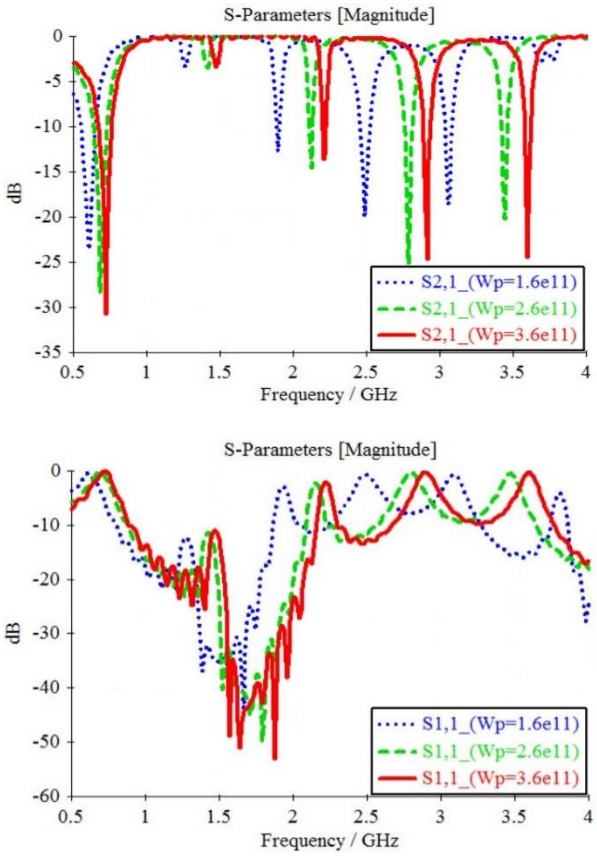


Fig. 3: S-parameter of simulated coaxial plasma waveguide filter at different plasma frequencies.

As you can see, the waveguide with plasma frequency transmits the wave in the bands 1-2, 2.2-2.8, and 3-3.5 GHz, which with the change of the plasma frequency (ω_p), the values of the resonance frequencies and the bandwidths decrease. On the other hand, by reducing the plasma frequency, the matching in the resonance frequencies decreases, which can be compensated by increasing the collision frequency (υ_c).

The simulation results of the matching reduction compensation at low plasma frequencies with increasing collision frequency (υ_c) are shown in Fig. 4. In plasma frequency 1.6×10^{11}, the value of collision frequency is increased from 4×10^{8} to the value of 1.4×10^{9}, and the results are recorded. As can be seen, the matching to the resonance frequencies is improved, but on the other hand, the ripple bandwidth is increased, which is not desirable. For Coaxial Plasma Waveguide Filter operation at higher pass frequencies, It is necessary that Increase the plasma frequency($\omega<\omega_p$). On the other, the plasma frequency will not increase to a certain extent, because it depends on the density and material of the gas used, the value of excitation voltage and current, and the size of the gas tube. By changing the mentioned parameters (in order to increase the plasma frequency), the dimensions of the structure will increase and the pass frequency band of Plasma Waveguide Filter will decrease. Also, in this structure, due to its special design (coaxial waveguide) at higher frequencies, high-order modes are excited and the structure will not be a TEM transmission line [30].

In the following, the laboratory equipment and how to change the plasma parameters to change and improve the plasma coaxial waveguide parameters are discussed.

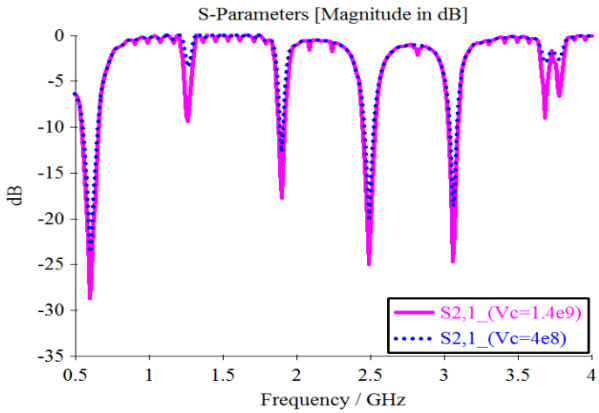


Fig. 4: S-parameter of coaxial plasma waveguide filter with constant plasma frequency and increase of collision frequency to compensate for the reduction of matching.

Implementation of Coaxial Plasma Waveguide Filter and Excitation Circuits

In this section, According to the simulation results, the construction and design the coaxial plasma waveguide filter and its most important part, the excitation circuit, are discussed. To implement the structure of the plasma tube, a 9-watt linear fluorescent lamp is used, which has

an effective length of 20 cm. Since the designed waveguide is of coaxial type and needs a metal body (shield), an aluminum tube with a thickness of 0.5 mm and a length of 20 cm has been used as a metal body. The whole set of excitation circuits is embedded in one box. Fig. 5 shows a general schematic of the laboratory equipment.

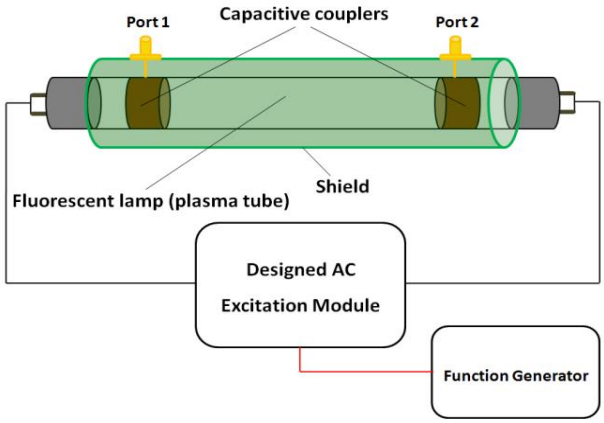


Fig. 5: Schematic of laboratory equipment to test plasma coaxial waveguide filter.

Fig. 6 shows the view of the plasma coaxial waveguide filter. It should be noted, however, that it is not possible to measure the parameters of the Drude model, plasma frequency, and collision frequency without access to the Plasma inside the tube (by Langmuir probe), so the exact Value of these parameters varies in different stimuli. And so far, there is no clear method for measuring or calculating them. Therefore, in the measurements section of this article, the exact Value can not be calculated for them, and only the waveguide parameters are considered.

Fig. 6: Real view of plasma coaxial waveguide filter.

Plasma Excitation Circuits

Since the change of plasma parameters to change and improve the waveguide parameters depends on the excitation circuits, the excitation circuit was designed with several capabilities [31]. The circuit in Fig. 7 can generate pulses from 500 Hz to 40 kHz using an internal stable multi-vibrator circuit. But since we want to test the excitation current with different waveforms, we have a wider frequency and, at the same time, change the excitation voltage and current by the function generator. Using the key embedded in the surge box, we will be able to change the excitation circuit state by applying an internal excitation waveform (stable multi-vibrator) to an external excitation waveform (function generator). The transistor used acts as a buffer to provide more excitation current.

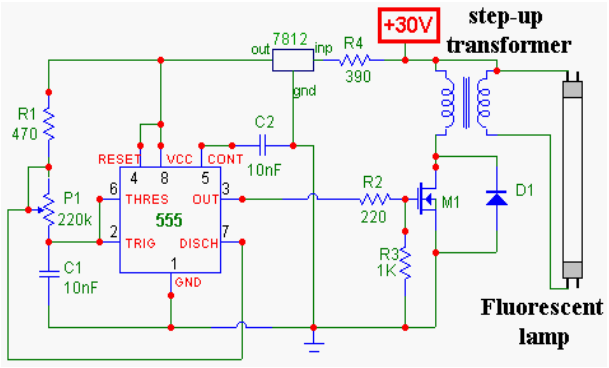


Fig. 7: Plasma AC excitation circuit with the ability to apply adjustable internal and external excitation waveforms.

The function generator device is used to apply the external excitation waveforms and adjust the frequency and amplitude of the wave. Depending on the supply voltage and the amplitude of the applied waveform, this circuit can generate an output voltage of 500 to 20,000 volts. The output current is also 0.1 to 1 amp, depending on the input values.

Signal Coupler

Because there is no direct access to the Plasma inside the tube to send or receive radio signals, as with conventional metal waveguides, we have to use a signal coupler. This coupler consists of a conductive copper strip with a thickness of 0.1 mm and a width of 1 cm, which is wrapped around the two ends of the lamp at a distance of 4 cm from the ends of the fluorescent lamp. The SMA connector core is connected to this coupler, and its body is connected to a metal body (Shield). Fig. 8 shows how the capacitive couplers of the signal are positioned around the two ends of the fluorescent tube and a view of the coaxial waveguide.

Fig. 8: Connecting the capacitive signal coupler to the fluorescent tube.

Plasma Coaxial Waveguide Test

To test the plasma coaxial waveguide with different excitation frequencies and its effect on plasma parameters and compare it with the simulation results of a function generator device manufactured by EZ-Digital with model FG-7005C and to Plasma waveguide, S-parameter was measured using an Agilent network vector analyzer (VNA) model E5071C. After applying the Excitation to the fluorescent lamp and keeping the input current to the module constant, as well as the amplitude of the external Excitation waveform, the measurement parameter and its results were recorded Fig. 9.

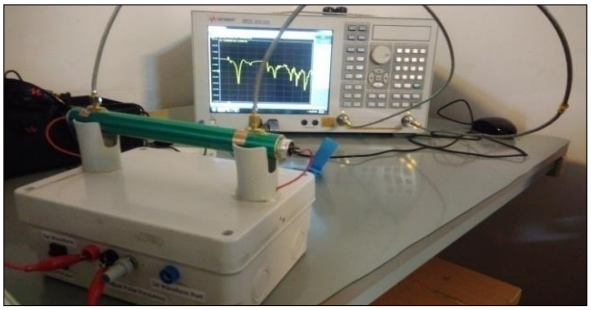


Fig. 9: Measurement of waveguide filter parameter with VNA.

Fig. 10 shows an example of parameter measurement using a VNA device. In the following, the results obtained from measuring and changing the input current and excitation frequency are discussed.

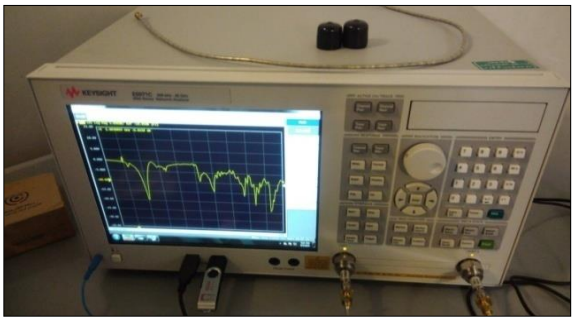


Fig. 10: Example of S_{21} parameter measurement with VNA.

Experimental Results and Discussion

First, AC excitation is investigated. After the connections and calibration of the VNA device, by keeping the applied wave amplitude constant, the pulse frequency was changed, and the effect of the excitation pulse frequency on the plasma waveguide parameters was measured step by step. The general results can be seen in Table 2.

Due to the ample space of the results in the table, some intermediate values have been removed. Then, without changing the excitation connections and their parameters and by keeping the excitation frequency constant at 30 kHz, the excitation input current applied to the Plasma Waveguide from 0.45(A) to 1.05(A) with steps 0.1(A) increased, and the effect of increasing the excitation input current on the ripple passband and improving the matching at resonance frequencies was observed. The general results can be seen in Table 3.

Table 2: Results of square wave excitation method (pulse)

Max, Min Pass Band Ripple (dB)	Waveguide Pass Bandwidth (GHz)	Excitation Current (A)	Excitation Frequency (KHz)
(0), (-1.1)	0.76 – 2.01 = 1.25 2.19 – 2.71 = 0.52 2.84 – 3.14 = 0.3	0.55	5.00
(0), (-1.2)	0.8 – 2.1 = 1.3 2.2 – 2.75 = 0.55 2.9 – 3.18 = 0.28	0.52	10.00
(0), (-2.1)	0.9 – 2.22 = 1.32 2.36 – 2.94 = 0.58 3.13 – 3.41 =0.28	0.50	15.00
(0), (-3.1)	0.95 – 2.29 = 1.34 2.4 – 3 = 0.6 3.2 – 3.46 =0.26	0.48	20.00
(0), (-4.4)	0.97 – 2.32 = 1.35 2.44 – 3.05 = 0.61 3.25 – 3.50 = 0.25	0.45	25.00
(0), (-5.9)	1 – 2.38 = 1.38 2.5 – 3.11 = 0.61 3.3 – 3.55 = 0.25	0.44	30.00
The results were the same with the 30 kHz excitation frequency.		0.43	35.00

Table 3: Results of increasing the excitation current on the passband ripple and improving the Matching at the resonance frequencies

Max, Min Reflection Losses (dB)	Min, Max Pass Band Ripple (dB)	Excitation Current (A)	Excitation Frequency (KHz)
(-2.98), (-34.9)	(0), (-5.75)	0.45	30.00
(-3.0), (-35.5)	(-0.21), (-6.10)	0.55	30.00
(3.40), (-36.8)	(-0.43), (-6.50)	0.65	30.00
(-3.98), (-37.1)	(-0.71), (-7.10)	0.75	30.00
(-4.30), (-38.3)	(-1.20), (-7.69)	0.85	30.00
(-4.60), (-39.1)	(-1.51), (-8.11)	0.95	30.00
(-4.80), (-39.8)	(-1.94), (-8.65)	1.05	30.00

In excitation alternating current with a square wave

(pulse), excitation frequencies below 5 kHz cause the transistor temperature to rise and heat loss to be high. According to the measured values in Table 2, the maximum Value of Excitation current frequency was measured to be about 30 kHz. From this frequency onwards, no change was observed in the passband ripple and the bandwidth frequency range. Increasing the excitation input current, as recorded in Table 3, improved the matching of the resonance frequencies and, on the other hand, increased the passband ripple, which is not desirable. In the diagram of Fig. 11, we can see the trend of changes in the passband ripple in opposition to the excitation input current.

It should be noted that the excitation current and frequency are actually the plasma excitation current and frequency and is different from the Excitation of the input and output signal ports of the waveguide.

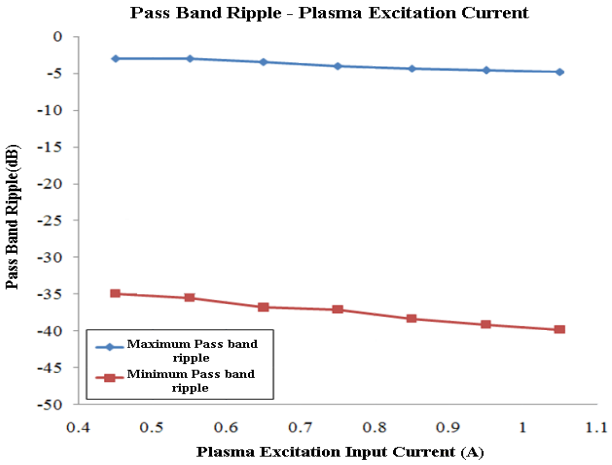


Fig. 11: Pass Band Ripple - Plasma Excitation Current.

In the diagram of Fig. 12, you can see an example of the parameters measured at different excitation frequencies. In Fig. 13, the excitation frequency is stabilized at 30 kHz, and the plasma excitation input current is increased.

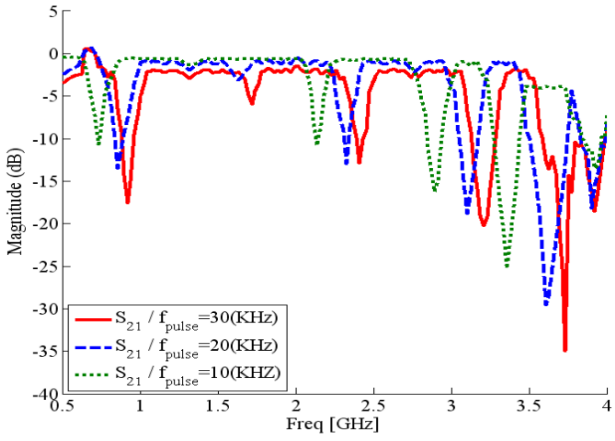


Fig. 12: S_{21} Parameter measured with different excitation frequencies

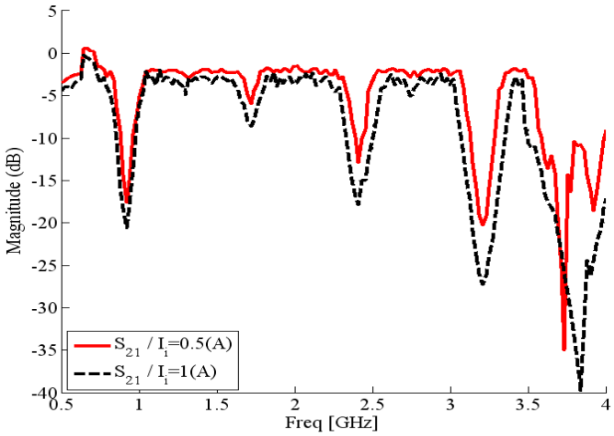


Fig. 13: S_{21} Parameter measurement at 30 kHz with increasing plasma excitation input current.

Fig. 14 shows the degree of conformity of the S_{21} parameter in both simulated and measured modes. In the simulated mode the rate ω_p is equal to its maximum rate of 3.6×10^{11} Rad/s, and in the measured mode, the plasma excitation frequency is assumed to be equal to its maximum Value of 30 kHz. The maximum value is when the value has not changed significantly in the simulated or measured results.

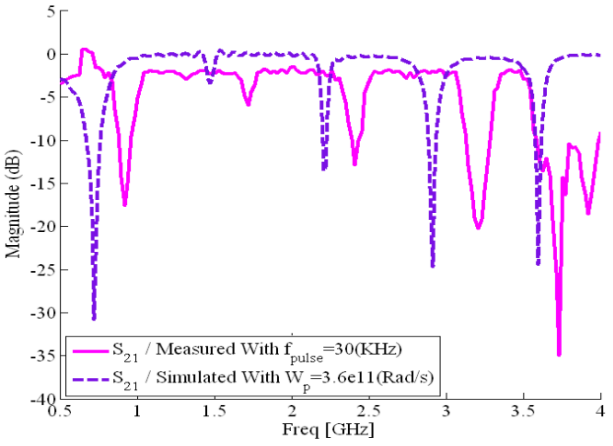


Fig. 14: The degree of conformity of the S_{21} parameter in both simulated and measured modes.

As can be seen from the parameter diagram of Fig. 12, with increasing the frequency of the plasma excitation pulse, the plasma frequency increases. Consequently, the frequency range of the Passband and its resonance frequencies also increase. Also, by increasing the excitation Frequency and transmission Passband of the waveguide filter, the passband ripple and matching at the resonance frequencies also increase. In other words, at lower frequency ranges, the waveguide filter passband will have fewer ripples. It should be noted, however, that low-frequency plasma excitation pulses cause several resonance frequencies to be lost. According to the S_{21} parameter diagram of Fig. 13, with increasing the plasma

excitation input current, the Matching to the resonance frequencies is improved. Still, on the other hand, the passband ripple is also increased, which will not be desirable.

As shown in Fig. 14, the simulation results slightly disagree with the practical results. The reasons for this can be 1- The new type of excitation used and the different frequencies of plasma Excitation, 2- Inability to measure plasma parameters in natural state, and 3- an ideal the simulation environment. On the other hand, as mentioned earlier, Plasma is a complex environment (nonlinear and Dispersive), and it isn't easy to match the simulation and Experimental results. For example, the ripples in the S21 parameter measured in Fig. 12 and Fig. 13 are due to the alternation of the plasma excitation signal, which is practically impossible to create such a thing using the Drude model in the simulation software.

Conclusion

During the article, the complete steps of designing a coaxial plasma waveguide filter with the ability to adjust the frequency range of the passband were followed, which are: simulation with CST software, implementation of plasma waveguide, excitation circuits, coupling design, and finally plasma coaxial waveguide test, the results of which were presented. The main focus of this paper was to reconfigure waveguide characteristics using altering and controlling plasma parameters through AC excitation. As mentioned, Plasma has two main parameters called collision frequency and plasma frequency. By changing them, the properties of plasma material and, consequently, without changing the physical structure, the parameters of plasma waveguide change.

According to the results, the change of these two parameters depends on the change of plasma excitation current frequency, and the value of excitation current applied. As the value of Excitation current increases, the matching to the resonance frequencies improves, but on the other hand, the passband ripple of the plasma waveguide filter increases. As the plasma excitation pulse frequency increases, the bandwidth and resonance frequencies change to higher frequencies, and the Matching to the resonance frequencies improves. But on the other hand, the passband ripple increases. This new type of waveguide filter can be used in cognitive/ adaptive telecommunication systems due to the constant change of frequency band.

Author Contributions

This paper is the result of M. Tohidlo's Research project which is supervised and advised by S. H. Mohseni armaki and M. Kazerooni, respectively. M. Tohidlo did the Simulations, Design and fabrication of Coaxial Plasma Waveguide Filter and wrote the manuscript. M. Tohidlo and S. H. Mohseni Armaki presented a new Reconfigurable Coaxial Plasma Waveguide Filter. S. H. Mohseni armaki interpreted the results and edited the manuscript. M. Kazerooni reviewed the manuscript.

Acknowledgment

The authors would like to thank the editor and anonymous reviewers.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

ω_p	Plasma frequency
υ_c	Plasma collision frequency
ω	Waveguide frequency
F	Electric force
v	Electron velocity
e	electron charge
m	Mass of the electron
n_e	density of free electrons
J	plasma inner surface current
P	Plasma discharge power
Υ	Electron natural collision frequency
γ	Propagation constant
L	Waveguide length
r_p	Tube glass radius
r_{sh}	Shield radius
W_c	Coupler width
L_c	Capacitive coupler distance from the end of the waveguide

References

[1] T. J. Dwyer, J. R. Greig, D. P. Murphy, J. M. Perin, R. E. Pechacek, M. Raileigh, "On the feasibility of using an atmospheric discharge plasma as an RF antenna," IEEE Trans. Antennas Propag., 32(2): 141–146, 1948.

[2] T. Anderson, Plasma Antennas. Artech House-1 edition: 203, 2011.

[3] H. Ja'afar, M. T. Ali, H. M. Zali, N. A Halili, A. N. Dagang, "Analysis and design between plasma antenna and monopole antenna," in Proc. IEEE International Symposium on Telecommunication Technologies (ISTT 2012): 47-51, 2012.

[4] C. Wang, B. Yuan, W. Shi, J. Mao, "Low-profile broadband plasma antenna for naval communications in VHF and UHF bands," IEEE Trans. Antennas Propag., 68(6): 4271-4282, 2020.

[5] T. Anderson, "Plasma frequency selective surfaces," in Proc. IEEE Antennas and Propagation Society International Symposium (APSURSI): 2096-2097, 2014.

[6] S. H. Zainud-Deen, H. A. E. A. Malhat, N. A. Shabayek, “Reconfigurable RCS reduction from curved structures using plasma based FSS,” Plasmonics., 15(2): 341-350, 2020.

[7] M. A. Lieberman, A. J. Lichtenberg, Principles of Plasma Discharges and Materials Processing, New York: Wiley: 757, 1994.

[8] J. Zhao, Y. Chen, Y. Sun, H. Wu, Y. Liu, Q. Yuan, “Plasma antennas driven by 5–20 kHz AC power supply,” AIP Adv., 5(12): 127114, 2015.

[9] J. P. Rayner, A. P. Whichello, A. D. Cheetham, “Physical characteristics of a plasma antenna,” in Proc. AIP Conference Proceedings: 392-395, 2003.

[10] F. Sadeghikia, M. T. Noghani, M. R. Simard, "Experimental study on the surface wave driven plasma antenna," AEU Int. J. Electron. Commun., 70(5): 652-656, 2016.

[11] A. V. Mitrofanov, D. A. Sidorov-Biryukov, M. V. Rozhko, N. V. Erukhimova, A. A. Voronin, M. M. Nazarov, A. B. Fedotov, A. M. Zheltikov, “Broadband ultrawide-angle laser-plasma microwave antennas,” Phys. Rev. A., 105(5): 053503, 2022.

[12] M. R. Harston, J. F. Chemin, “Mechanisms of nuclear excitation in plasmas,” Phys. Rev. C., 59(5): 2462, 1999.

[13] H. Q. Ye, M. Gao, C. J. Tang, “Radiation theory of the plasma antenna,” IEEE Trans. Antennas Propag., 59(5): 1497-1502, 2011.

[14] N. A. Dyatko, I. V. Kochetov, V. N. Ochkin, "Influence of the ionization process on characteristics of spatial relaxation of the average electron energy in inert gases in a uniform electric field," Phys. Rev. E., 104(6): 065204, 2021.

[15] A. Rezagholi, F. Mohajeri, “On the application of neon discharge plasmas in construction of plasma waveguide attenuators,” Iran. J. Sci. Technol. Trans. Electr. Eng., 44(1): 77-87, 2020.

[16] B. D. McVey, M. A. Basten, J. H. Booske, J. Joe, J. E. Scharer, “Analysis of rectangular waveguide-gratings for amplifier applications,” IEEE Trans. Microwave Theory Tech., 42(6): 995-1003, 1994.

[17] A. Abdoli-Arani, “Dispersion relation of TM mode electromagnetic waves in the rippled-wall elliptical plasma and dielectric waveguide in presence of elliptical annular electron beam,” IEEE Trans. Plasma Sci., 41(9): 2480-2488, 2013.

[18] Y. Herhil, S. Piltyay, A. Bulashenko, O. Bulashenko, “Characteristic impedances of rectangular and circular waveguides for fundamental modes,” in Proc. 2021 IEEE 3rd Ukraine Conference on Electrical and Computer Engineering (UKRCON): 46-51, 2021.

[19] J. Utkarsh, R. K. Raj, A. K. Lall, D. K. Upadhyay, G. K. Mishra, “Reconfigurable Bandpass Filter for use of 2.7–3.1 GHz radar spectrum,” in Proc. 2016 International Conference on Emerging Trends in Communication Technologies (ETCT): 1-4, 2016.

[20] K. C. Hwang, “Design and optimization of a broadband waveguide magic-T using a stepped conducting cone,” IEEE Microwave Wireless Compon. Lett., 19(9): 539-541, 2009.

[21] S. Z. Gurbuz, H. D. Griffiths, A. Charlish, M. Rangaswamy, M. S. Greco, K. Bell, “An overview of cognitive radar: Past, present, and future,” IEEE Aerosp. Electron. Syst. Mag., 34(12): 6-18, 2019.

[22] L. E. Brennan, L. S. Reed, “Theory of adaptive radar,” IEEE trans. Aerosp. Electron. Syst., (2): 237-252, 1973.

[23] H. Islam, S. Das, T. Bose, T. Ali, “Diode based reconfigurable microwave filters for cognitive radio applications,” a review IEEE Access., (8): 185429-185444, 2020.

[24] F. Gentili, F. Cacciamani, V. Nocella, R. Sorrentino, L. Pelliccia, “RF MEMS hairpin filter with three reconfigurable bandwidth states,” in Proc. European Microwave Conference: 802-805, 2013.

[25] H. Conrads, M. Schmidt, "Plasma generation and plasma sources," Plasma Sources Sci. Technol., 9(4): 441, 2000.

[26] C. D. Lorrain, P. Brityei, Electromagnetic Fields and Waves, USA 2nd edition: John Wiley& Sons: 656, 1976.

[27] D. H. Froula,"Plasma scattering of electromagnetic radiation: theory and measurement techniques," Academic Press: 497, 2011.

[28] W. Xiao-Po, S. Jia-Ming, "Scattering by two parallel plasma cylinders," in Proc. IEEE International Conference on Microwave and Millimeter Wave Technology (ICMMT): 1-4, 2012.

[29] A. Zhu, “Characteristics of AC-biased plasma antenna and plasma antenna excited by surface wave,” J. Electromagn. Anal. Appl., 4(7): 279–284, 2012.

[30] A. Chittora, S. Singh, A. Sharma, J. Mukherjee, “Design of wideband coaxial-TEM to circular waveguide TM 01 mode transducer,” in Proc. 2016 10th European Conference on Antennas and Propagation (EuCAP): 1-4, 2016.

[31] M. Tohidlo, S. M. Hashemi, F. Sadeghikia, “The effect of frequency and waveform of AC excitation on U-Shaped monopole plasma antenna,” Radar., 7(2): 89-95, 2020.

Biographies

Seyyed Hossein Mohseni Armaki was born in Kashan, Iran. He received his B.Sc. degree in Communication engineering from the KNTU, Tehran, Iran, in 1991, the M.Sc. degree in Communications engineering from the KNTU, in 1995 and the Ph.D. degree from Iran University of Science & Technology, Tehran, Iran, in 2011. He is an Associate Professor at Malek Ashtar University of Technology, Tehran, Iran. His research interests include antenna analysis, antenna measurements, and electromagnetic propagation.

- Email: Mohseni@mut.ac.ir
- ORCID: 0000-0002-6777-5658
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Majid Tohidlo was born in Tehran, Iran, in 1996. He received the B.Sc. degree in 2020 from Shahid Rajaee University (SRU), Tehran, Iran in Electrical Engineering, and M.Sc degree in 2022 from Malek Ashtar University of Technology(MUT), Tehran, Iran in Communications Engineering. His areas of research interests Reconfigurable Antenna, Plasma and Dispersive Environment, Dual Polarization Antenna, Phased Array System and Antenna, and Microwave Circuits.

- Email: M.tohidlo@mut.ac.ir
- ORCID: 0000-0002-8360-7099
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Morteza Kazerooni received the B.S. degree from the Department of Electronic Engineering, Shiraz University, Shiraz, Iran, in 1998, the M.S. degree from the Malek Ashtar University of Technology in 2001, and the Ph.D. degree from the Iran University of Science and Technology (IUST), in 2010, Tehran, Iran. He is currently an Associate Professor with the Malek Ashtar University of Technology. His research interests include design and analysis of Phased Array Systems, Synthetic Aperture Radar (SAR) and Microwave Passive Systems.

- Email: Kazerooni@mut.ac.ir
- ORCID: NA
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: Na

How to cite this paper:
S. H. Mohseni Armaki, M. Tohidlo, M. Kazerooni, "Design and fabrication of coaxial plasma waveguide filter with the ability to reconfigure the frequency band," J. Electr. Comput. Eng. Innovations, 11(1): 75-84, 2023.

DOI: 10.22061/JECEI.2022.8668.542

URL: https://jecei.sru.ac.ir/article_1723.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

A Novel Architecture Based on Business Intelligence Approach to Exploit Big Data

M. R. Behbahani Nejad[1,*], H. Rashidi[2]

[1]*Department of Computer & IT, Qazvin Branch, Islamic Azad University, Qazvin, Iran.*

[2]*Faculty of Statistics, Mathematics and Computer Science, Allameh Tabataba'i University, Tehran, Iran.*

Article Info

Article History:
Received 13 January 2022
Reviewed 15 March 2022
Revised 02 June 2022
Accepted 06 June 2022

Keywords:
Architecture
Big data
Business intelligence
Hadoop
CPN

*Corresponding Author's Email Address:
Reza2005nejad@gmail.com

Abstract

Background and Objectives: Big data is a combination of structured, semi-structured and unstructured data collected by organizations that must be stored and used for decision-making. Businesses that deal with the business intelligence system, as well as their data sources, have a major challenge in exploiting Big Data. The current architecture of business intelligence systems is not capable of incorporating and exploiting Big Data. In this paper, an architecture is developed to respond to this challenge.

Methods: This paper focuses on the promotion of business intelligence to create an ability to exploit Big Data in business intelligence. In this regard, a new architecture is proposed to integrate both Business Intelligence and Big Data architectures. To evaluate the proposed architecture, we investigated business intelligence architecture and Big Data architecture. Then, we developed a Unified Modeling Language diagram for the proposed architecture. In addition, using the Colored Petri-Net, the proposed architecture is evaluated in a case study.

Results: The results show that our architectural system has a higher efficiency in performing all steps, average time, and maximum time compared to business intelligence architecture.

Conclusion: The proposed architecture can help companies and organizations gain more value from their data sources and better support managers and organizations in their decision-making.

Introduction

Today, the use of Business Intelligence (BI) systems to support decision-makers at different levels of decision-making in companies and organizations is an essential requirement ([1], [2]). As the inputs to these systems are becoming more and more valuable, the results of these systems will improve the support of managers in decision-making. With the advent of a new phenomenon called Big Data, all IT issues such as intelligent business systems have been affected. Companies and organizations with business intelligence systems that also have data sources of big data type found themselves in trouble [3]. In this paper, the business intelligence architecture will be developed to solve this problem. The most important issue is the problems caused by the entry of Big Data into the business intelligence system because BI systems with their current architecture are not able to use Big Data. As a result, this paper focuses on the promotion and development of business intelligence to create the ability to exploit Big Data. This paper begins with highlighting the concepts of software architecture. We then focus on the way Big Data architecture is examined in terms of how to store, load, manage, and analyze data. Thereafter, concepts related to the BI are explored. Our new model derives from the architecture of intelligent architecture

and business intelligence architecture. Based on the evaluation indicators that are extracted from the review studies [4], the business intelligence architecture and the new model are compared and weaknesses are identified to determine which architecture has the most valuable value for a company or organization. The proposed architecture is evaluated with the assistance of Unified modeling language and Colored Petri-Nets (CPNs).

This paper focuses on the lack of support for the architecture of business intelligence systems from Big Data as one of the most important information resources of companies and organizations. The collision of two categories of business intelligence systems and Big Data together causes some problems in business intelligence systems. One of these problems is the impossibility of analytic, storage, and loading Big Data in the business intelligence system. This will make IT systems unable to value Big Data, and so much of the information resources of companies and organizations that have Big Data cannot be exploited by business intelligence systems. In this paper, we have tried to provide a solution to this issue by presenting a new model for BI business architecture.

The purpose of this research is to solve the problem of business intelligence systems in dealing with Big Data. For this purpose, a new model is presented in which the architecture of business intelligence and Big Data has become integrated. The rest of this paper is organized as follows. Section 2 presents background and related works. Section 3 presents the proposed architecture. Section 4 provides an evaluation of the proposed architecture with a case study. Section 5 is considered for the summary and conclusion.

Background and Related Works

In this section, the software architecture, business intelligence architecture, and Big Data architecture are briefly examined.

Software Architecture

Garlan and Shaw [3] define the software architecture as "a collection of computational components or simply components together with a description of the interactions between these components the connectors". In 2000, IEEE defined architecture as the fundamental organization of a system embodied in its components, their relationships to each other, the environment, and the principles guiding its design and evolution [5]. Software architecture is any system where software contributes essential influences to the design, construction, deployment, and evolution of the system as a whole ([6], [7]).

Software Architecture Style

According to a definition by Clements [8], "An architectural style is a dedication of components and communications among them with each other along with a set of rules and limitations about how to use them".In another definition, according to Taylor [9] "An architectural style is a named collection of architectural design decisions that (a) are applicable in a given development context, (b) constrain architectural design decisions that are specific to a particular system within that context, and (c) elicit beneficial qualities in each resulting system."

If the category of Garlan and Shaw's [3] is considered with architectural styles, then there are generally five types of architectural style: Dataflow style, Data-driven style, styles based on the promotion, Independent-component styles, and Virtual machine styles. These architectural styles are described more in detail below.

The first architectural style is Dataflow [7]. In this style, the architecture of the system determines how its data is exchanged between different components. In other words, the way data flows in the system plays a decisive role in the behavior of the system. The flow of data in these systems is very similar to the implementation of the logic of programming languages. Usually, data-flow systems can be a good option for modeling any kind of workflow. In these systems, the presence of at least two elements that flows between them is required. The processing is mainly performed in them, meaning that the output of an input element(s) will be in the data flow direction. The main subcategories of Dataflow styles include Pipe and Filter styles. The Filter includes several elements that are responsible for processing input data and converting them to output data. The Pipe establishes communication between filters, transfer data, and information. There are some rules and restrictions. For example, the type of pipes, their capacity, how they combine filters, and so on is considered as the rules and constraints.

The second architectural style is Data centered [7]. Today, most organizations around the world have a strong dependence on their data. Maintaining a company's data is vital to the extent that large companies are willing to spend millions of dollars to secure and maintain their data. In an environment that is so important to data preservation, the emergence of software-based software architectures based on Persistent Data is not surprising. With this strong motivation, a lightweight architecture emerges as a repository that provides the basis for sharing information among the components, individuals, and organs of data sharing. The main components of the tank style are: (a) The central data repository, which is, in fact, a large data structure that is shared between processes and various departments; (b) processing elements that are potentially independent of each other. This means that with the help of the central repository, data can satisfy all their communication needs and do not need to communicate

directly with each other. The third architectural style is Call/Return [7]. It is a style of software architecture that includes a variety of styles as well as the layered architecture. Systems that follow the layered style are inherently hierarchical. In a layered system, different layers provide transparency for the users.

The fourth architectural style is Service-oriented architecture [7]. It has been developed in recent years, known as service-oriented architecture (SOA). This style can somehow be expanded into a layer style or component-based style. The SOA is a model for developing software systems. Given the growth of information systems, organizations need to respond quickly to new business needs. While existing software architectures have provided some relief, evolutionary service architecture is a step-by-step service that helps organizations manage complex challenges [10].

The SOA is a matured component-based architecture, object-oriented design, and distributed systems. The SOA is a style of software design in which services are provided to the other components by application components, through a communication protocol over a network. The basic principles of the SOA are independent of vendors, products, and technologies [11]. The SOA enables application functionality to be provided as a set of services, as well as the creation of applications that make use of software services. The services are loosely coupled because they use standards-based interfaces that can be invoked, published, and discovered. In the SOA, the services are focused on providing a schema and message-based interaction with an application through interfaces that have application scope, rather than component-based or object-based. The SOA service should not be treated as a component-based service provider.

The SOA style can package business processes into interoperable services, using a range of protocols and data formats to communicate information. In the SOA, the clients and other services can access local services running on the same tier, or access remote services over a connecting network [11]. The looseness of the connection between the components of the software leads to their reusability, and the software is based on the service. In the SOA, services are divided into three categories: Service Request, Service Provider, and Registry Service.

Architectural Evaluation

For many years, analysts, engineers, and scientists have developed and used models to deal with complex systems. A model approximates the features of a real system it can also be used to evaluate systems that are not feasible in terms of method and economy before designing a system, regardless of design or initial output. As a result, existing information mismatches between different models can be overcome by reducing the high semantic distance between high-level needs and low-level architectural products.

Modeling outputs with reflecting some characteristics of the quality attributes give architects and designers of complex systems the power to visualize the entire system. Architecture is the first step in software development that can be traced to quality requirements. Qualitative attributes are considered at all stages of design and implementation and, if supported by architecture, be more easily detectable.

Models demonstrate runtime behavior by displaying architectural characteristics that can be used to evaluate many of the quality attributes, including performance and reliability. An executable model of software architecture is an implementation of the system, in which features are displayed, that includes non-narrative needs.

Applied architecture is created in the early stages of software development to reduce the risks associated with performance, operational capability, reliability, and so on. Having operational models in the initial phases provides the ability to evaluate the dynamic behavior of the system in different situations and to solve the existing problems. The real-time requirements with systematic needs and acquisition of the proper conditions for optimizing a system are the main issues that can be achieved by using executive models. Typically, modeling tools are used to create an implementation model.

There are various modeling tools available to display application architecture. The most important of these modeling tools are Petri Networks, queuing networks, simulation models, and process algebra. In addition, some of the languages that describe the architecture can also display a running architecture.

Architecture has a very important role in the software production process. Because the architect is the one who deals with all the stock-owners, they become a very influential person in the process. Designing an appropriate architecture due to the vagueness of the architectural specification is a very difficult task. With an architectural execution model, many architectural steps can be completed with high accuracy because the execution model at this level lowers the errors, recognizes Easier needs, better analyzing and evaluating of the system, and simplifies the presentation of the architecture.

Unified Modeling Language

The Unified Modeling Language is a Semi-formal and standard language for easy description of software architecture that is used to address the requirements of software engineering expertise. The main purpose of UML is to use its high descriptive power to model software architecture. The methods used in UML can handle only certain issues. Evaluation of software systems is not possible because UML is not a convenient approach for

evaluation. Therefore, to evaluate software systems, it is necessary to convert the actual model to the formal model.

The main problem with UML is in determining how to evaluate and analyze the system architecture using the documentation before the production of software. Presenting an effective method to evaluate and analyze the efficiency based on the software architecture may contribute to driving a software project successfully forward ([12], [13]). Since the UML-based system is not applicable, the system's behavior verification is delayed until its implementation; hence, the Colored Petri-Net (CPN) is used as an applicable model of software architecture. It is in particular well-suited for modeling systems in which communication, synchronization, and resource sharing are important. By transforming the actual model into a formal model, the possibility of evaluating the software architecture's performance on the official model is provided.

Qualitative features are the same non-obligatory system requirements that are largely determined by an architectural style. Performance, reliability, security, availability, usability, modifiability, portability, and testing capabilities are the most qualitative features to evaluate any software architecture.

Performance is the main quality attribute of software that demonstrates how well the software works concerning time-dependent issues [14]. Software performance is the process of predicting and evaluating whether the software satisfies performance goals defined by the users. The early identification of unsatisfactory performance of Software Architecture (SA) can greatly reduce the cost of design change. This is because correcting a design flaw is more expensive the later the change is applied during the software development process [14].

Because the performance is around timing, events (interrupts, messages, requests from users, or the passage of time) occur and the system must respond to them. There are a variety of characterizations of event arrival and the response, but the performance is essentially concerned with how long it takes the system to respond when an event occurs [7]. Performance appraisal at the early levels of software development reduces costs, risks development, and so on. Therefore, performance has a very important role to play in the success of software systems and an evaluation of the efficiency of the entire software development process should be considered.

Template-based software architecture is described with three different diagrams, including class diagrams, use case diagrams, and sequence diagrams. These diagrams should describe the behavior and architecture of software architecture. CRC (Class, Responsibilities, and Collaborators) graphs show the static structure of software architecture as a software component and the relationship between them. The use case diagram also specifies the services provided by the software system and describes the order of the behavior of the system.

The Object Management Group (OMG) introduced several UML extensions [15]. This group defined the SPT profile for scheduling, performance, and time specification. The performance sub-key, similar to other profiles, uses stereotypes and labeled values to support the expansion process. Each profile contains several stereotypes. Labeled values are attributes of stereotypes in the profile and are linked to the model's key element as explanations.

Evaluation of the performance of the software architecture using the characteristic sub-heading of efficiency and time is an approach proposed to build software systems so that the qualitative objectives are visible. The proposed approach builds on the UML Performance Sub-Platform for the proper modeling and evaluation of software performance throughout the software development process.

The use case diagram models the user's use of the system and is suitable for evaluating the system's performance. Because a set of use cases is used for the performance evaluation process, it is necessary to create a performance model for each use case of the software.

The sequence diagram describes the communication pattern established by the samples. The role of the samples in the order of the diagram is to accomplish a specific objective, namely interaction. From the point of view of efficiency, certain elements and structures are used to model the system load in the sequence diagram. Any message in the graph can be attached to a condition that expresses the probability that the message will be sent. In contrast, a component diagram maintains the role of service centers and their characteristics in terms of efficiency goals.

Petri Networks

The theory of Petri Networks was introduced by Carl Petrie [16]. The use of Petri's network to evaluate the software architecture and create an executive model due to its simplicity and high availability is very much considered. Petri Networks are displayed graphically, and they provide a mathematical framework for Analysis, Validation, and Performance evaluation. The focus of Petri's networks is on synchronization, coherence, and asynchronous operation.

These networks are powered by the power of systems behavioral modeling and are used as a tool for describing architecture.

The Colored Petri-Net has been introduced as a model developed from Petri Networks [16]. The Color Petri-Net uses the capabilities of simple Petri Networks and

programming languages. As shown in Fig. 1, data values in these networks are carried by the beads. Using the timestamp for a nut, it's easy to calculate the time of activation and the transfer of the nut to the destination location. Suppose that the goal is to evaluate the performance that is running from time to time – in which case it would be enough to refer to the timestamps and calculate the architectural efficiency that is associated with it.

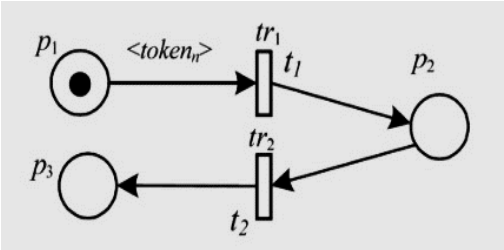


Fig. 1: A Colored Petri-Net with a timestamp for a bead [16].

Various tools support Colored Petri-Nets, including the CPN Tools software provided by the University of Aarhus, Denmark.

The first version of this software was released in October 2001 that is used for editing, simulating, and analyzing this kind of Petri-Nets.

Business Intelligence

Business Intelligence (BI) is the art of gaining a business advantage from data [17]. It is a technology-driven process for analyzing data and delivering actionable information that is used by managers, analysts, and executives to make informed business decisions. Fig. 2 shows a high-level business intelligence architecture that is used in practice.

This Fig. 3 shows that BI architecture consists of four components [18]. The first component is the data warehouse. It is a large repository of well-organized historical data. The second component is Business analytics, which are the tools that allow the transformation of data into information and knowledge. The third component is Business performance management (BPM) which allows monitoring, measuring, and comparing key performance indicators. The fourth component is the User interface (e.g., dashboards) that allows access and easy manipulation of other BI components [4].

The main business intelligence architecture is in the form of a service, as shown in Fig. 3. In this figure, there are three components for services, namely, Analytical and Reporting services, Data Management Services, and Integration Services.

The data sources can be relational Databases, Files sources, and other sources.

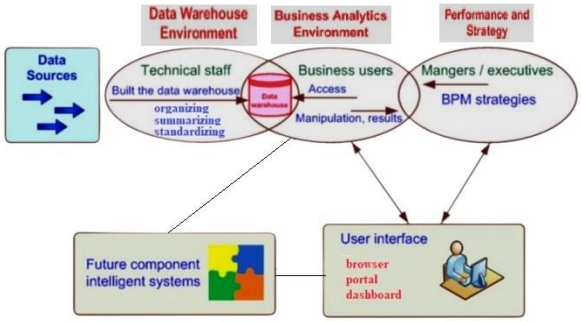


Fig. 2: A High-Level Architecture of BI [18].

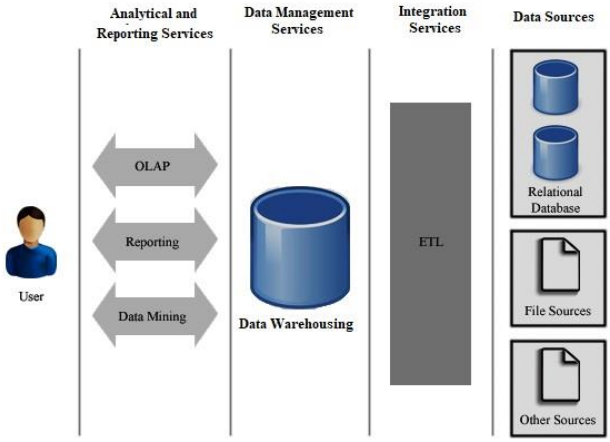


Fig. 3: Business Intelligence Architecture [4].

Big Data

Big Data is an abstract concept [19]. So far, there are many definitions of Big Data. In 2001, Doug Lenny of the Gartner Institute introduced a 3V model for defining Big Data [20]: "Data that is growing at a very high rate, has plenty of volumes so that it occupies a large amount of disk space…In addition, they are very diverse, that is, they consist of different structures of data". In other definitions, more features of metadata were provided, including data value, the complexity of data, the accuracy of data, and, based on these features, 4V, 5V, and even 7V models were also presented.

Apache most influential company in the field of Big Data. It introduced the main feature of a dataset that can be referred to as Big Data, which makes it impossible to store, manage, and process those data using common computational methods, and the rest of the Big Data Features Sub-attributes [19]. Apache defined Big Data as follows: “Big Data is referred to as a set of data that cannot be stored, managed, or processed by conventional computing methods.” And in the forthcoming article, this definition is the basis of work.

Apache's definition of Big Data seems to be closer to reality [20]. According to this definition, data that has 3V attributes is Big Data, and data that does not have 3V attributes, but is not commonly analyzed, is Big Data. If

Gartner's definition of Big Data is to be accepted, it is suggested to refer to data that does not have 3V attributes but cannot be stored and analyzed in the usual way, called "Semi-Big Data" or Semi Big Data (Table 1).

Table 1: Examining the types of data in terms of 3V features, and analyzing and storing in routine ways

Data types	Ability to analyze and store commonly used	3V features
Semi-Big Data	0	0
Structured data	1	0
Big Data	0	1
Such a situation is not possible	1	1

In Table 1, routine methods are referred to as methods that can be used to analyze and store structured data. The term semi-Big Data can be defined as follows: A dataset that has one or more attributes of Big Data features (V3 or V4 or V5 or ...) but can be stored, analyzed, managed, and controlled. They do not exist in the usual way at the expected time.

It is also recommended that the Semi-Big Data and Big Data be called "NODATA". "NODATA" stands for "NOT ONLY DATA", or it could be an abbreviation for "NOT ONLY STRUCTURE DATA". "NODATA" data type is usually stored in NOSQL databases. The words NODATA and NOSQL are similar in appearance. The term "NODATA" can be defined as a set of data that cannot be stored, analyzed, managed, and controlled by commonly used computing methods. However, "NODATA Technology" refers to technologies that can store and provide analysis, management, and control of NODATA.NODATA is a term that includes Big Data and Semi-Big Data.

Database Big Data

Since over 80% of the world's data is unstructured, relational databases are unable to store and manage such data [22]. To store this data, unmatched databases should be used. In general, there are four types of non-marketable databases classified according to columns, documents, key values, and graphs.

Hadoop

Apache is the most widely used company in the field of Big Data [19]. The company has been supporting and supporting the largest Big Data project called Hadoop, which has been published in an open-source. The extension of Hadoop has made a library of Big Data-related projects that includes a large number of sub-projects. The most important projects in the library are the distributed header file system project, which is responsible for data storage, loading, and management, and the mapping/reduction project, which is responsible for data super-data analysis, as well as the company's creation and development of a number of the non-relational databases that are used to store data, are very useful. Spark is another undergraduate sub-project on the top layer of Map-Reduce that extends the Map Reduce model.

The Hadoop can be likened to an operating system designed to handle and manage a large amount of data on different machines [23]. The best example for understanding the function of Hadoop is the difference between his software, Amway and Hadoop. He transforms a physical server into multiple virtual servers and converts a remote server into a virtual server.

Advantages of using Hadoop for business analytics include scalability, inexpensiveness, swift-paced, versatile, and no failures. Its use cases include advertisements, financial services, healthcare, gaming, and the web. Many companies all over the world use Hadoop for business analysis. Some of the largest corporations include Amazon web services, Cloudera, IBM, MapR technologies, and Microsoft [24]. The most important Big Data-related technologies are cloud computing, data centers, internet objects, and Hadoops [25].

Many national governments such as the U.S. also paid great attention to big data. In March 2012, the Obama Administration announced a 200-million-dollar investment to launch the “Big Data Research and Development Plan,” which was the second major scientific and technological development initiative after the “Information Highway” initiative [26].

Big Data Architecture

A picture of the high-level Big Data architecture is shown in Fig. 4.

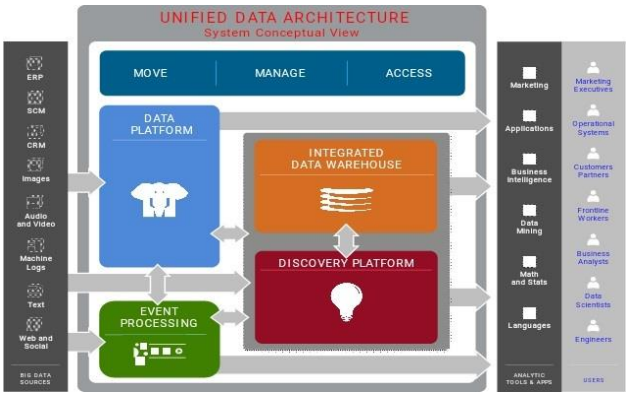


Fig. 4: High-level Big Data architecture [18].

The history of Big Data architecture projects is well represented in Table 2. In this table, the companies that worked on Business Intelligence architecture and the major activity of each company are highlighted in its right column.

Table 2: History of Projects Related to the Big Data Architecture ([27], [28])

Company (initial year)	Major Activity
Seisint (2000)	Developed a C++-based distributed file-sharing framework for data storage and query. The system stores and distributes structured, semi-structured, and unstructured data across multiple servers
Google (2003)	The genesis of Hadoop was the "Google File System" paper that was published in October 2003. This research spawned another one from Google – "Map Reduce: Simplified Data Processing on Large Clusters" in December 2004.
Yahoo (2006)	Development started on the Apache Nutch search engine project but was moved to the new Hadoop subproject in January 2006. Hadoop is born in Nutch. Hadoop 0.1.0 was released in April 2006.
Choice Point (2008)	Created parallel processing platforms
LexisNexis (2011)	Acquired Seisint Inc. in 2004. Acquired Inc. and their high-speed parallel processing platform in 2008. The two platforms were merged into HPCC (or High-Performance Computing Cluster) Systems in 2011.

Among the architectures provided for Big Data, the header has become more successful due to open source and its support by experts.

In Fig. 5, the architecture of the Big Data architecture is briefly outlined.

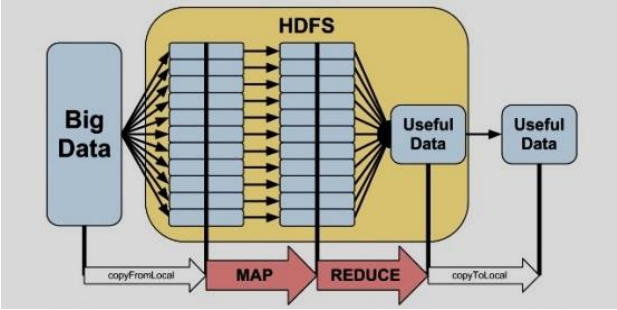


Fig. 5: General Framework for Big Data Architecture [29].

The left side of Fig. 6 depicts the mapping/reduction processes symbolically, while its right side depicts the mapping/reduction function to another type.

[1]TDWI (The Data Warehousing Institute) Research provides research and advice for business intelligence and data warehousing professionals worldwide.

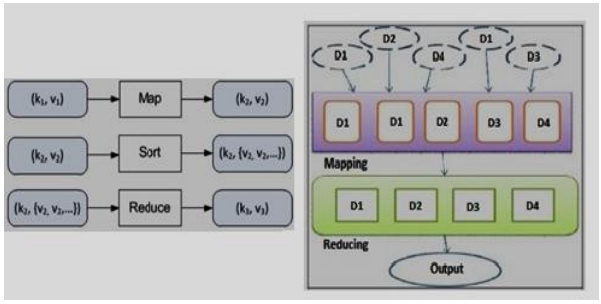


Fig. 6: Mapping and Reduction Processes [24] (Left-Side) and Mapping / Decrement Task View [28] (right-Side).

Integration of Architectures

In this part, the aim of integration, requirement of integration with examples of architectures is presented.

The Purpose of the Integration

TDWI[1] first proposed the integration of business intelligence and Big Data architectures in 2013. TDWI assumes that Hadoop usage will become mainstream in the coming years [29]. Hadoop has proved its usefulness with the toughest challenges in BI today, namely big data, advanced analytics, and multi-structured data.

The current business intelligence systems cannot support Big Data, and they have shortcomings, including the lack of performance and high costs. In a survey, the desire of companies and organizations to use the various features of the distributed file system of Hadoop was investigated. The results show that 78% of companies and organizations participating in this survey tend to use the distributed file system of Hadoop as a complement to the database management system. The data is especially useful for advanced analysis [29].

Big Data architecture often uses exploratory analysis, data mining, statistical analysis, sophisticated SQL queries, and more to analyze Big Data. The main advantage of Big Data architecture is that of intelligent business intelligence, scalability, and data diversity.

To improve the decisions process, every organization needs to use an active and integrated intelligence structure. For achieving such structure, collecting, storing for preparing data, analyzing, converting the result to useful information have a highly important role. For doing the analytical process, we should use a suitable environment that includes a warehouse, intellectual process, and a link. In the data storage, internal and external sources along with plenty of data must be stored.

This data storage can be separated based on marts and then saved in a warehouse, according to the activities of the company or organization. Recently, some big data should be added ([1], [30], [31]). An organization

integrates a distributed system of header files and a database management system that can use all the data to increase business value and reduce the cost of data management. In other words, the purpose of integrating business intelligence systems and Big Data architecture is to increase the value of the business in the organization or company, to create competitive advantage, reduce costs and productivity, and also use the value of all data, not just the use of value from part of the data.

The Needs for Integration

Utilizing shared architecture can lead to more accurate decisions in the organization and can prevent future failures in the system. The integration of Big Data architecture and the BI architecture allows us to exploit more data because it creates the opportunity for more data to be analyzed and ends with better results than when it comes to Big Data architecture. The important results from the integration of Big Data and business intelligence systems are as follows [29]:

- Improving business processes and procedures as well as achieving business goals in the target organization (the target organization could be a company, industry, education, a financial system, or a global system).
- Reducing scattered data and using smaller but more valuable data.
- Enabling more accurate decisions.
- Forecasting the future to prevent system failure.

Although there are many tools for big data architecture for implementation, when we want to integrate or merge data architecture and business intelligence systems, we must select the best tools for doing so and then combine them with architecture layers of business intelligence systems. This merging can lead to improving business plans which help to meet the goals of organizations.

We can make a comparison of HDFS with DBMS.HDFS system has a distributed file system without database management, but they have several capabilities of DBMS too. These capabilities such as titling and accidental access to intelligence support of SQL language improve optimizing and searching. Of course, the performance of several capabilities of HDFS is better than DBMS capabilities such as management of a large amount of data according to file and management of unstructured data. In BI systems, we can use DBMS for administrating storing of data and big data for the distributed file system HDFS, which accompany other tools like HBase, Impala operates like DBMS.

One of the weaknesses of the database management system is the inability to store Big Data. This includes the weaknesses of the distribution system because the file header is less accurate than the database management system. Thus, the best tool for storing the Big Data of the distributed system with a header file as well as structured data is the database management system, specifically for storing data types, which are not replaced for each other. If we use the tools in the Big Data architecture as a complement to the database management system, the performance of the database management system will also be better. It is due to the database management system working hard on some parts, especially if the data is unstructured or semi-structured, but the system with the distributed header file also works well on such parts.

Obviously, if the source of database management is focused on structured data and all of the mentioned resources exempt, the performance and efficiency will be improved. Many organizations already have data that the database management system in the organization's business intelligence systems cannot process, including public relations recordings, organization XML supply documents, sensor log files, machinery, and other unstructured data that may exist in the organization. The Big Data architecture can easily store and process these data.

The integration can lead to improved business processes and improved business plans to meet the goals of the organization. Because of the cost of the source in database management, capabilities of HDFS system, and free sources present with open code and cheap, we can reduce the costs utilizing the sources of HDFS and free sources of database management. In other words, since HDFS can be used as supplementary work for the database, we can take advantage of HDFS in computing and database management. With these advantages, data in HDFS is not processed. Instead, data in database management is processed and prepared for suitable usage. HDFS system supports comparability and multi-structured database. Processing data in database management makes high precision and it can increase the organizational facilities. HDFS system can store many kinds of data, including unstructured data and structured data with ETL operations. Storing and managing structured data in database management have better performance. The management and storing unstructured data in HDFS are better and has lower costs.

HDFS with supporting tools such as Hbase, can help to store and manage the data efficiently. In order to the monitoring of intelligence and reporting, we can apply usable linkages in business intelligence systems or the other designed tools in big database architecture. The differences between visualizing information or intelligence in big data and business intelligence systems are that tools directly have relations with data, but in business intelligence systems is not so. With the integration of HDFS and database management, we can create new and joint opportunities and it leads to increase

the capability of business intelligence systems, such as rating the data, Archive data source, Management unstructured data, Management of file-based, Increasing the power of processors, Management, store of data.

Integration of Big Data Architecture with other Architectures

In recent years, cloud computing has been integrated with other architectures such as BI architecture, DSS architecture, and SCM architecture. In 2013, the integrated Chan architecture was presented. Chan's integrated architecture is due to the integration of business intelligence architecture and Big Data architecture (See Fig. 6). In 2014, Samson and colleagues integrated the Big Data architecture with the BI architecture[32]. In Fig. 6, both architectures are presented without evaluation and have structural bugs. Table 3 depicts how Chan architecture and Samson architecture are compared in terms of their strengths and weaknesses.

Table 3: A comparison of architectures derived from the integration of business intelligence architecture and Big Data architecture, to provide a conceptual model

Researcher (year)	Strengths	Weaknesses
Chan (2013) [33]	- Suitable for Real-Time Systems	Failure to Structural, Lack of evaluation
Samson et al. (2014) [32]	- Scalability -Parallel processing of data warehouse with HDFS - Analysis beyond the map / reduce	Failure to Structural, Lack of evaluation
This research (2022)	-Presentation of a conceptual model at different levels and in different styles - Suitable for real-time systems and other systems - High scalability - Parallel processing of data warehouse and big data warehouse - Provide analyzes beyond Map / Reduce - Without structural forms - Architectural valuation of different methods	Evaluation is time-consuming

Comparing Architectures

We compared the architecture of Chan and Samson architecture with the proposed architecture. Table 3 shows the results of this comparison in terms of strengths and weaknesses.

Building a new architecture uses a systematic approach, that is, the architecture has an input and output. The first layer can be considered for the data source, the second layer for data storage and management, the third layer for the data analysis and intelligence, and the fourth layer for visualization tools and applications. It can be said that the first and second layers are the infrastructure layer, the third layer is the computational layer and the fourth layer is the application layer. Fig. 8 shows the proposed architecture. The performance of each layer is based on the name chosen for it.

In this architecture, each layer has several sub-layers, and each substrate can be included components. The multiplicity of components of each sub layer allows for the creation of multiple and different models based on the proposed architecture. In each model, one or more components of the sub-layer were used. Selecting one or more components of each sub-layer component is based on the need specified in the model.

A kind of architecture is proposed by Chan that applies to the business intelligence system [33]. The typical characteristics of this architecture are how to use them for real-time systems. The most problem of Chan architecture is the lack of assessment so there are some difficulties in the structure. One of these difficulties is in using tools or analyzing way (Map/Reduce) that portioned into two parts, one in the analytical sections and the other beside the HDFS. It is not structurally correct that use two parts with a name (Map / Reduce) with a fixed concept. Instead, we could name it in one place and refer it elsewhere. Samson architecture focuses on storing data through a system file with big database architecture. So it leads to facilitate comparability, whereas the traditional ways have not this capability. The most important capability mentioned for this architecture is a parallel processing warehouse with a big database so that it can provide more analyzes than those in the Map/Reduce of the Hadoop. It is made through direct accessing to data along with indirect access to data that is expected in this architecture. Samson and Chan have not presented any valuation on their architectures. From the points of view in structure or form of storing source, analyzing data, visualizing intelligence clearly has done. Big data architecture integrates with other ones in the article [34]. Moreover, big data architecture with a support system of decisions can be integrated. In this architecture, by identifying steps of the Simon model, integration of big database architecture is completed with the support of the system of design. Integration of the architecture of big databases with the architecture of the supply chain is done and is presented in the article [35].

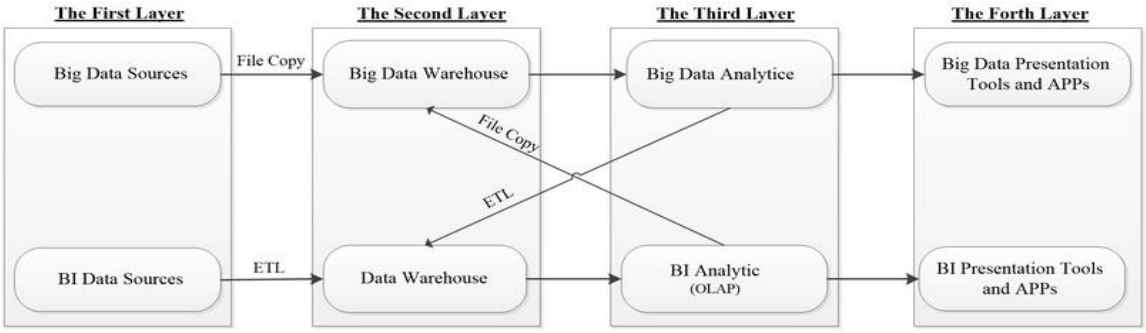


Fig. 7: The overall four-layered architecture of the proposed architecture with the system approach.

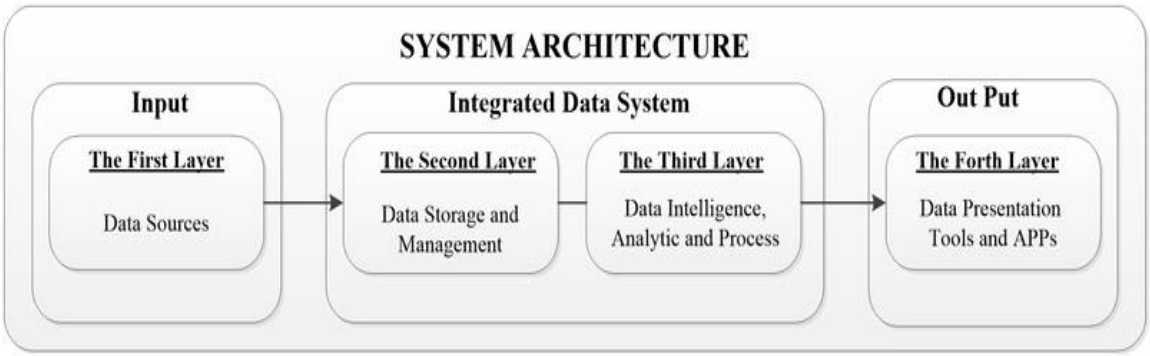


Fig. 8: A new four-layer architecture, derived from the integration of Big Data architecture and business intelligence architecture.

Proposed Architecture

The proposed architecture is a four-layered architecture that integrates Big Data and business intelligence architectures. This integration is illustrated in Fig. 7 of the template.

The first layer is a data source layer that has two substrates: The Big Data Sources sub-layer and the BI Sources sub-layer. Each of these sub-layers can be included in many data sources. The Big Data Sources sub-layer can consist of raw data streams, data flow, unstructured data, semi-structured data, multi-structured data, streaming events, images and videos, audio files, social media, text files, XML documents, webpage data, blogs, emails, Docs and PDFs, log server, marketing events, sensor data, GPS data, scientific research, machine logs, graph files, NOSQL files (Base, Mongo DB, Couch DB and sec). The BI Sources sub-layer contains structured data, operational systems data, RDBMS files (SQLServer, Oracle, Microsoft Access, MySQL and sec), software's structured data (CRM, SCM, ERP, and sec).

The second layer is the data storage and management layer**,** which can be considered as an infrastructure layer. This layer consists of two sub-layers: one is the Big Data warehouse sub-layer and the other is the Data warehouse sub-layer. Each of these substrates can include the components shown below. The Big Data Warehouse sub-layer includes HDFS, NOSQL (HBase, Mongo DB, and sec), Hive, Impala, Kudu, and RDD. The Data Warehouse sub-layer includes Data Marts, RDBMS, MPP (Massively Parallel Processing). It should be noted that the Big Data warehouse is a new concept similar to the concept of storage, with the difference that the storage warehouse is a reservoir of data in which the data are simply copied and there are no categories for the data. In some scientific sources, such a concept has been named "data lake" [36], which is referred to here as the storage bin.

The third layer is the Data Intelligence, Analytics, and Process Layer, which can be considered as the computation layer. This layer contains two sub-layers, one underlying Big Data Analytics sub-layer, and another BI Analytics sub-layer. Each of these substrates can include the following components: The Big Data Analytics sub-layer includes MAP Reduce, Real-Time Analytics, Pig, and Spark. The BI Analytics sub-layer includes OLAP.

The fourth layer is the Data Presentation Tools and Applications Layer. This layer also has two sub-layers, one underlying the Big Data Presentation Tools and Applications sub-layer, and the other underneath the BI Presentation Tools and Applications sub-layer. Each of these sub-layers has the components listed below:

The Big Data Presentation Tools and Applications sub-layer includes On-line APPs, Big Data tools (visualization, reporting, predictive, data mining, queries, machine learning), real-time APPs, and near real-time APPs. The BI Presentation Tools and Applications sub-layer includes dashboards, off-line APPs, developer environments, BI tools (visualization, reporting, predictive, data mining, queries, machine learning), custom APPs, and enterprise APPs.

Proposed Architectural Style

The proposed architectural style of different views is similar to some architectural styles but does not follow a particular style. It is the combination of several different styles and can be considered as an independent style.

The proposed architecture can be compared to data stream styles in similarity. One of a variety of data stream styles is pipe styles and filters. In the proposed architecture, the BI Analytic and Data Warehouse sections act as filters, because data is entered there, and after processing and modification, it becomes a series of information and then exits. Data changes in BI Analytic can be done using OLAP, and in the Data Warehouse, this change is done by the ETL process. On the other hand, the connection between these two parts can be considered as a pipe, which is responsible for transmitting data and information from one part to another.

In the similarity of proposed architecture with data-driven styles, one of a variety of data-driven styles is the repository style. The proposed architecture in this regard is similar to the architecture in the Data Warehouse and Big Data Warehouse sections of the data container.

As a Big Data repository, Big Data Warehouse shares its data between BI and Big Data Analytics. One of the components of data-driven styles is the processing elements that are potentially independent of each other. In the proposed architecture, BI Analytics and Big Data Analytics are process elements that are potentially independent and interconnected with the Big Data Warehouse database.

The proposed architecture can be compared to the layered style. One of a variety of styles based on overlays is a layered style. In layered style, the layers are hierarchically placed together and the lower layer provides a higher level of service, in other words, the upper layer of the client and the lower layer of the server. To display layered layout architecture, layers with headings (infrastructure layer, computing layer, and application layer) can be utilized. The proposed architecture in the form of a layered style can be shown in Fig. 9. As we can see, the first layer is an infrastructure layer that includes data sources and data management and storage. The second layer is the computation layer, which includes intelligence, data analysis, and data processing. In the third layer, the application layer includes visualization tools and application software.

The proposed architecture can also be compared with the service-oriented architectures. The proposed architecture is process-oriented, but it can be developed as a service-oriented architecture. In this case, the various components of the architecture will be linked together with a loose connection. This kind of connection will lead to the creation of a vibrant and dynamic system that can be considered as a competitive advantage for real-world businesses in today's marketplace.

In the proposed architecture, if data sources are in a cloud environment, the Data Sources layer, as well as the Data Storage and Management layers, can be provided as a service, called Infrastructure as a Service.

If we imagine the proposed four-layer architectures in the style of service-oriented architectures, it can be considered as three-layer architecture according to Fig. 10. As we can see, the infrastructure service includes data storage and management.

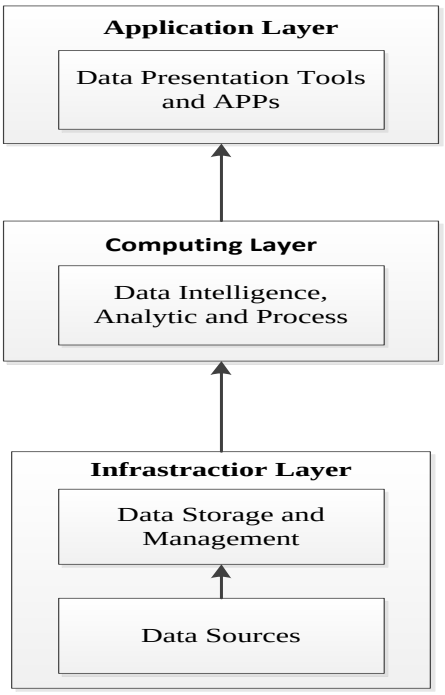


Fig. 9: Layered style architectural design.

In this figure, the computing service provides intelligence, analysis, and data processing. The application service provides visualization tools and applications. These three layers are explained below:

- The Infrastructure as a Service (Iaas) includes Big Data Warehouse as a Service and Data Warehouse as a Service. It extracts data from data sources that are commonly found in the cloud. It also has the task of storing and managing extracted data.

- The Computing as a Service (CaaS) includes Big Data Analytics as a Service and BI Analytics as a Service. It does the task of computing, analyzing data, intelligence, and processing data.
- The Application as a Service (AaaS) includes Big Data Presentation Tools and Applications as a Service and BI Presentation Tools and Applications as a Service. It does the task of visualizing and displaying information.

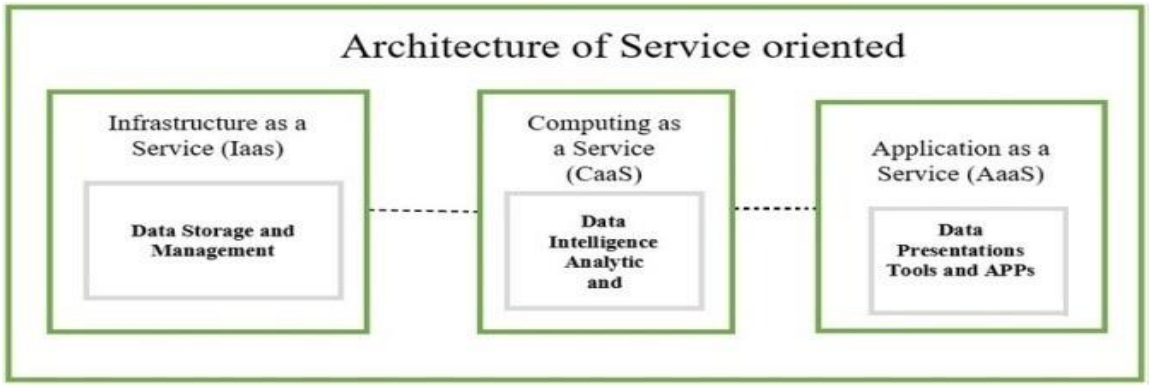


Fig. 10: Service-oriented architectural style architecture.

In the proposed architecture, the sub-layer subsystem, various components, and tools are named, all of those components and tools are developed in a service-oriented architecture, can be provided as a service. For example, in the Big Data Presentation Tools and APPS sub-layer subdirectory, the Data Mining tool is provided, this tool can be provided in the service architecture offered as Data Mining as a Service.

Evaluation with a Case Study

Before the implementation of the architecture, the colored Petri-Nets and the tools written for it could be used to evaluate the architecture and convert the actual architecture into a formal architecture, as well as provide an executable version of the architecture. Although this method has so far been less widely used in evaluating information systems, due to their high capabilities, it can be a reliable method for evaluating information systems and their architecture.

In this paper, we will focus on the proposed architecture. First, we will examine its strengths and weaknesses. Afterward, using an architecture simulation with the UML, it is then evaluated in a case study by Petri Networks.

Evaluation by Comparison of Weaknesses and Strengths

From different perspectives, one can study the weaknesses and strengths of an information system from an organizational, managerial, and Technology (technical) standpoint. Here, the strengths and weaknesses of Business intelligence, Big Data, and proposed architectures will be examined from a technological perspective.

The strengths and weaknesses of BI from the perspective of technology are as follows:

- Strengths of BI architecture include: improving decision making [37], standards support [38], generalization and adaptation [4], personalization [38], reduced costs [39], faster reporting, and more accuracy [37].
- Weaknesses of the BI Architecture include failure to support Big Data, failure to provide advanced analysis [41], lack of support for multi-structured data [29], and strong support for semi-structured and unstructured data [22].

The strengths and weaknesses of Big Data from the perspective of technology are as follow:

- Strengths of Big Data architecture include: analysis of Big Data ([29], [40]), high scalability ([24], [29]) , support for exploratory analysis [29] reducing miscellaneous data to reduce data volumes [40], providing more detailed for decisions-making by analyzing the Data, prevention of future system failure [40], acts as a good complement to the data warehouse, support for multicast data and low-cost hardware [29], low-cost Software ([21], [29]) and the storing and processing of data types (structured, semi-structured, and unstructured) ([24], [29]).
- Weaknesses of Big Data architecture include Lack of full SQL support, Low ability to query and access information in Real-Time, Evolving management tools ([29], [42]).

By integrating business intelligence and Big Data architectures into the proposed architecture, all of the weaknesses of the BI architecture are covered by the strengths of Big Data architecture, as well as the weaknesses of Big Data architecture being covered by the strengths of the business intelligence architecture. Thus, the outcomes from the integration of both business intelligence and Big Data architectures are revealed.

The strengths and weaknesses of the Proposed Architecture are as follows:

- Strengths of the proposed architecture include strengths of the Big Data architecture and strengths of the business intelligence architecture.
- Weaknesses of the proposed architecture include related to a possible problem in implementation, which can be measured after evaluation and experience collected.

It can be concluded that these two architectures can be a good complement to each other. Thus, if these two architectures can be integrated together, they will remain protected and the interaction of the two architectures would not give rise to new flaws.

In this case, with regard to the strengths of the proposed architecture, we can deduce that the strengths of the proposed architecture are almost equal to the sum of the strengths of business intelligence and Big Data architectures.

Concerning the weaknesses of the proposed architecture, we can also deduce that the proposed architectural weaknesses are equal in addressing the weaknesses of business intelligence architecture and the weaknesses of Big Data architecture. Because the weaknesses of one are the other strengths, the resulting weaknesses of the proposed architecture will be zero and there will be no weaknesses for the proposed architecture.

A weak point is that it cannot be resolved by integration, including the lack of implement the proposed architecture or the weakness created by the integration.

Evaluation by Simulation with Unified Modeling Language

A similar approach is integrated into the language of modeling.

Although this language is not able to provide an executable model of architecture with various charts, it can provide a quasi-official model of software architecture. Here, the system is supposed to be based on the architecture of the proposed system, and then work like this system will be implemented using the Unified modeling language. The schema of the simulated system is shown in Fig. 11.

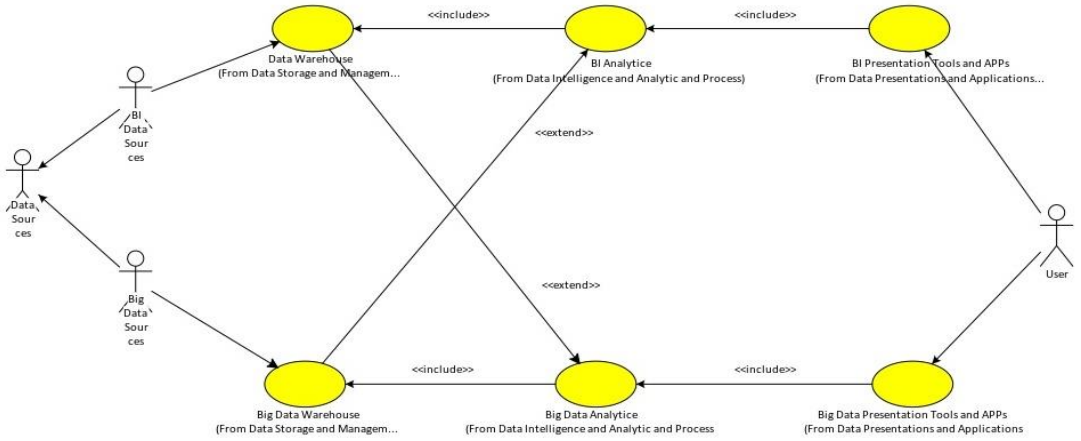


Fig. 11: The proposed architectural diagram of the proposed architecture.

Now, with the help of the use case diagrams, components, and sequences, we can the colored Petri-Net drew the architecture-based system and perform a simulation. In the simulation, we use the Colored Petri-Net in which a case study is required.

The Case Study

To evaluate the proposed architecture, we use a case study. The main goal, here, is to compare the proposed architecture with the traditional architecture used for business intelligence. With the traditional architecture of business intelligence in terms of quality performance attributes, performance is considered to be the most important quality attribute of software architecture.

The performance measurement means measuring the time needed to perform the processes of a software system. The less time this system uses, the more efficient the system. Comparing two architectures, the most important component of the analogy that shows the superiority of one architecture to another architecture is the quality attribute of efficiency. Of course, in a complete comparison of the two architectures, all the quality attributes and requirements for both architectures must be compared to fully demonstrate the superiority of the architecture to the other architecture. But one of the most important of these attributes is the efficiency attribute; the other attributes and requirements are

either not measurable at this stage or their significance is less effective than the quality attribute.

In this scenario-based case study, the traditional architecture of business intelligence with the proposed architecture is compared. First, the colored Petri-net of the simulated system is drawn that based on the business intelligence architecture. Also, the colored Petri-Net of the simulated system is drawn based on the proposed architecture. Thus, hundreds of simulated simulations are executed using hundreds of simulated petty networks, and then the average runtimes of the two systems are compared to determine which system is more efficient.

Fig. 12 shows the Petri-Nets of the business intelligence system based on the typical BI architecture after implementation.

The time values shown in the Finish section are the sum of the times from the beginning of the data entered into the system to the display of the final report to the user.

Fig. 13 shows the Petri-Net system based on the proposed architecture after implementation. The time values shown in the Finish section are the sum of the times from the beginning of the data entry to the end of the final report to the user.

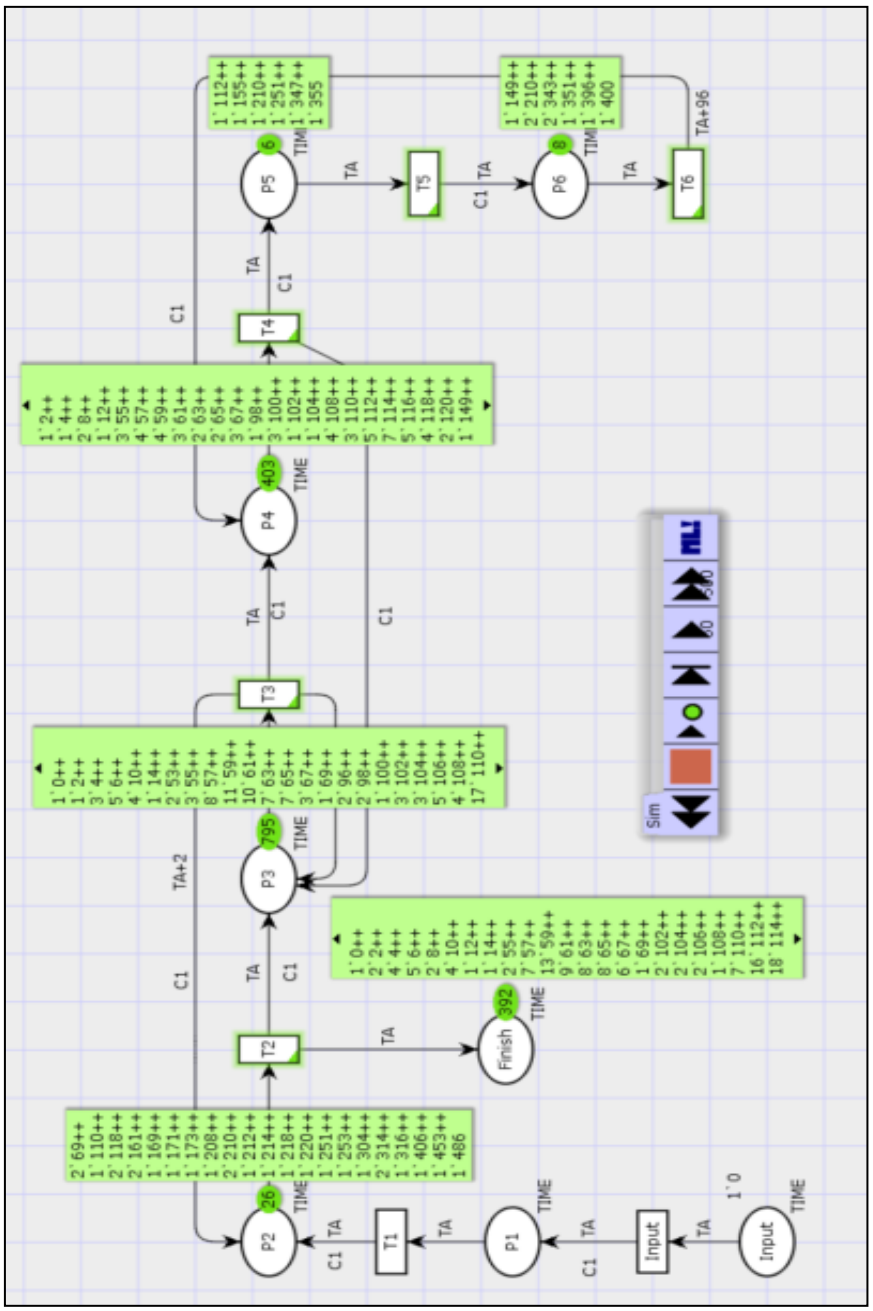


Fig. 12: Pure Color Networks of the BI System based on the architecture of the post-implementation.

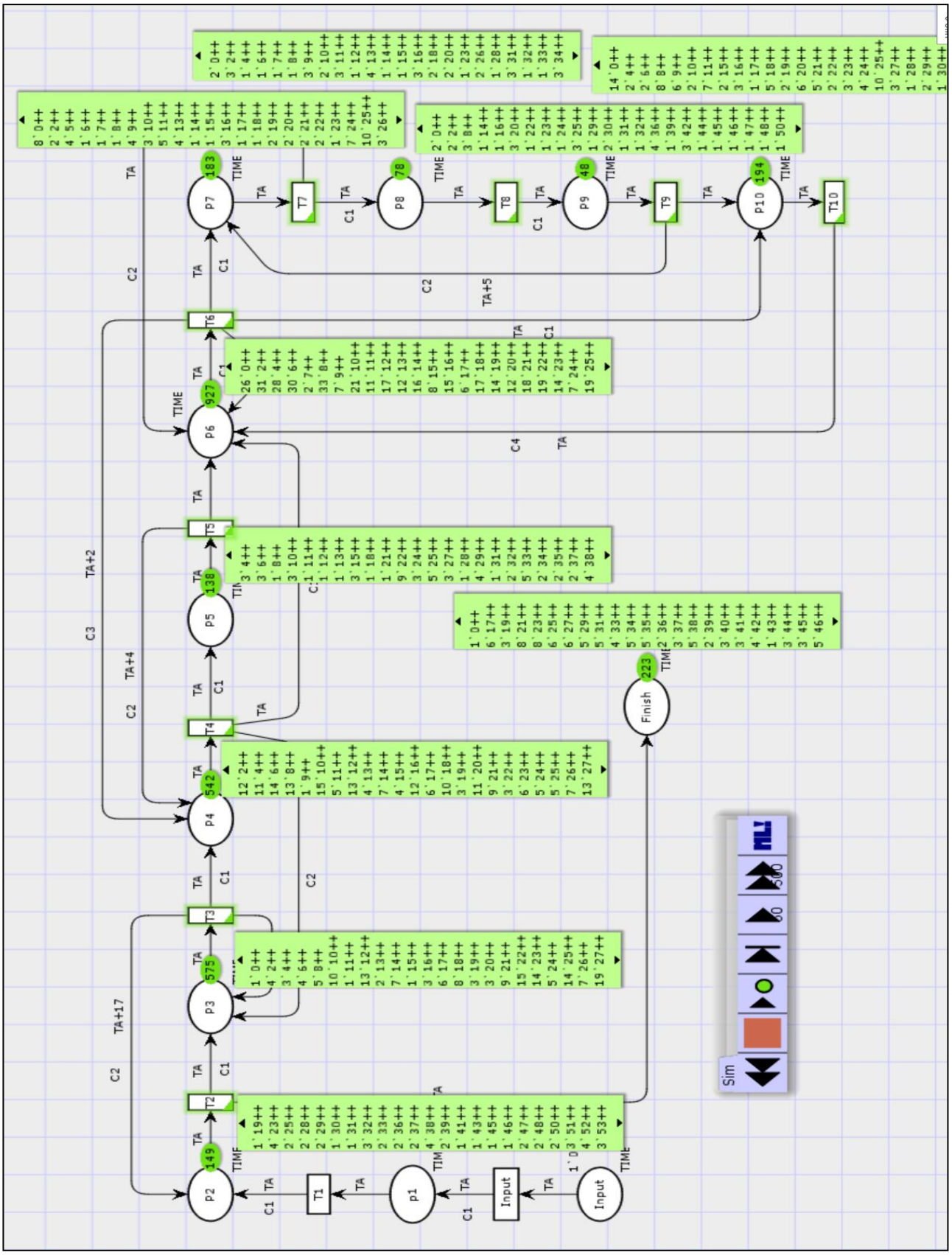


Fig. 13: Colored Petri-Net dish the proposed architecture-based architecture after implementation.

Table 4 illustrates the results obtained from the implementation of the color scheme of the Pure Color Network of the Business Intelligence Based Architecture and the proposed architecture. To compare the performance, usually, the metric of the count steps and time are used ([43], [44]). As we can see, the average runtime scenario in the business intelligence system is based on the common architecture of 378 and the average runtime scenario in the proposed architecture is 53. Also, other values, including the maximum time and total time of the proposed architecture are lower than that ofBI Architecture. Because of these lower values, the proposed system has more performance compared with the system-based system, based on the common architecture of business intelligence.

Table 4: Comparison of the results from the implementation of the Colored Petri-Nets (Time in Second)

Name	BI Architecture	Proposed rchitecture
All Count Steps	2000	1380
Count Finish Steps	268	268
Max Time	50845	138
Min Time	0	0
Total Times	101690	14331
Average Times	378.03	53.47

It is concluded that the system based on the proposed architecture is more efficient and can produce more valuable output data at a lower cost.

Results and Discussion

In this paper, we integrated the business intelligence architecture and Big Data architecture to address the problems of business intelligence systems in dealing with Big Data, resulting in a new architecture. By comparing the proposed architecture and the BI-architecture, the weaknesses of the business intelligence architecture were identified. Moreover, these weaknesses were resolved by integrating with the Big Data architecture. Also, in this paper, the evaluation of the business intelligence architecture as one of the complex architectures of information systems was carried out using the Colored Petri-Nets method. This proposed architecture allows companies and obtain more value from their data sources and receives stronger support from corporate executives and organizations in making executive decisions.

The most important result that can be drawn from this research is that we have an integrated the perspective. We promoted business intelligence systems and Big Data technologies to help managers create new opportunities in solving specific problems. This integration creates new opportunities for solving the problems of business intelligence and Big Data systems, which will greatly help to the managers and stakeholders of these systems.

Since the proposed architecture may have some side effects on the efficiency resources required, further research must be done. In addition, it is suggested that this proposed architecture be used in designing complex intelligence systems such as Business Intelligence Systems Decision support, resource management systems, supply chain management systems, and customer relationship management systems.

Author Contributions

This paper is the result of M. R. Behbahani Nejad M.Sc. project which is supervised by Mohammad Jafar Tarokh and advised by H. Rashidi and Participated by Bahman Nouriani. M. R. Behbahani Nejad proposed the main idea of the innovation and Integration of architectures, performed the simulations, carried out the data analysis, interpreted the results and wrote the first manuscript. H. Rashidi corrected the proofing the article and wrote the final version of the article. B. Nouriani Suggested petri nets for architectural evaluation.

Acknowledgment

This work is completely self-supporting, thereby no any financial agency's role is available.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

AaaS	Application as a Service
BI	Business Intelligence
BPM	Business Performance Management
CaaS	Computing as a Service
CPN	Colored Petri Net
CRC	Class, Responsibilities, Collaborators
Iaas	Infrastructure as a Service
NOSQL	Not Only SQL
OMG	Object Management Group
SA	Software Architecture
SOA	Service-based architecture
UML	Unified Modeling Language

References

[1] K. Batko, A. Ślęzak, "The use of big data analytics in healthcare," J. Big Data, 9(3): 1-24, 2022.

[2] A. Heydarian, H. Rashidi, “Storage, and access system of self-adapting Health big data based on smart IoT (in Persian),” presented at the Fourth International Conference on Electrical, Computer and Mechanical Engineering, 2021.

[3] C. Dobre, F. Xhafa, “Intelligent services for big data science,” Future. Gener. Comput. Syst. 37: 267–281, 2014.

[4] G. Muhammad, J. Ibrahim, Z. Bhatti, A. Waqas, “Business intelligence as a knowledge management tool in providing financial consultancy services,” Am. J. Inf. Syst., 2(2): 26-32, 2014.

[5] "IEEE recommended practice for architectural description for software-intensive systems," IEEE Std 1471-2000: 1-30, 9 Oct. 2000.

[6] W. Pedryz, Architecture in Big Data, Information Granularity, Big Data, and Computational Intelligence, Studies in Big Data, 8: 275-295, 2015.

[7] L. Bass, P. Clements, R. Kazman, Software Architecture in Practice, 4th Edition, SEI Series in Software Architecture, Addison-Wesley Professional, 2021.

[8] L. Clements, B. Paul, Rick Kazman, Software Architecture in Practice, 2nd Edition, SEI Series in Software Architecture, Addison-Wesley Professional, 2003.

[9] R. N. Taylor, N. Medvidović, E. M. Dashofy, Software architecture: Foundations, Theory and Practice. Wiley, 2009.

[10] M. M. Patil. Challenges and Issues in Handling Big Data, Int. J. Innovations Adv. Comput. Sci. (IJIACS), 4(Special Issue): 620 – 625, 2015.

[11] D. S. Linthicum, Service-Oriented Architecture (SOA)", Retrieved 09-21, 2016.

[12] P. Clements, K. Klein, Evaluating Software Architecture Methods and Case Studies, Addison Wesely, 2002.

[13] B. Kumar, J. Jaspernete, “UML profiles for modeling real-time communication protocols,” J. Object Tech., 9(2): 178-198, 2012.

[14] S. Balsamo, M. Marzolla, “A simulation-based approach to software performance modeling, ACM SIGSOFT software engineering notes,” in Proc. the 9th European Software Engineering Conference Held Jointly with 11th ACM SIGSOFT International Symposium on Foundations of Software Engineering, 28(5): 363-366, 2003.

[15] J. Dong, “UML extensions for design pattern compositions,” J. O. Tech. 1(5): 151-163, 2002.

[16] F. Liangbing, M. Obayashi, T. Kuremoto, K. Kobayashi Construction and application of learning Petri net, Manufacturing and Computer Science, 143-176, 2012.

[17] M. Muntean, Collaborative Business Environment Based on Federated Portals, Annals of T. Popoviciu Seminar, 4: 218-224, 2006.

[18] R. Sharda, D. Dursun, T. Efraim, Analytics, Data Science, & Artificial Intelligence: Systems for Decision Support, 11th edition, Pearson, Prentice-Hall, 2020.

[19] T. Erl, W. Khattak, P. Buhler, Big Data Fundamentals: Concepts, Drivers & Techniques, Pearson Service Technology Series from Thomas Erl, 1st Edition, 2017.

[20] D. Laney, “Application Delivery Strategies”, Gartner (META Group), White Paper, 2001.

[21] A. N. Nandakumar, N. Yambem, “A survey on data mining algorithms on apache hadoop platform,” Int. J. Emerging Technol. Adv. Eng., 4(1): 263-265, 2014.

[22] M. R. BehbahaniNejad, H. Rashidi, M. J. Tarokh, “Survey big data technologies architecture and the need to exploit it in DSS systems," presented at the 8th Iranian & 2th International Knowledge Management Conference, Tehran, 23-24 February 2016.

[23] S. R. Pakize, “A comprehensive view of Hadoop MapReduce Scheduling algorithms,” Int. J. Comput. Networks Commun. Secur., 2(9): 308–317, 2014.

[24] K. Kashyap, C. Deka, S. Rakshit, “A review on big data, Hadoop and Its impact on business,” Int. J. Innovative Res. Dev., 3(12): 78-82, 2014.

[25] A. Labrinidis, HV. Jagadish, Challenges, and opportunities with big data. Proc VLDB Endowment, 5(12): 2032–2033, 2012.

[26] M. Chen, S. Mao, Y. Liu, “Big data: A survey,” Mob. Netw. Appl. 19: 171–209, 2014.

[27] A. F. Mohammad, H. Mcheick, E. S. Grant, “Big data architecture evolution,” The fourth ACM International Symposium, 2014.

[28] Wikipedia. “Apache Hadoop” .Retrieve from https://en.wikipedia.org/wiki/Apache_Hadoop , 2015.

[29] Lockwood, Conceptual Overview of Map-Reduce and Hadoop, 2015.

[30] M. R. Behbahani Nejad, H. Rashidi, “A data architecture architecture with business intelligence approach,” presented at the first Seminar on Data Science and applications: 14-15, Tehran, 2021.

[31] H. Rashidi, “Corporate planning using object-oriented rules and considering big data (in Persian),” Allameh Tabataba’i Press, 2018.

[32] S. Oluwaseun Fadiy, S. Saydamb, V. Vanyduhe, “Advancing big data for humanitarian needs,” Procedia Eng., 78: 88-95, 2014.

[33] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, “Bigtable: A distributed storage system for structured data,” J. ACM Trans. Comput. Syst. (TOCS), 26(2): 1-14, 2008.

[34] T. Poleto, V. DioghoHeuer de Carvalho, A. P. Cabral Seixas Costa, “the roles of big data in the decision-support process: An empirical investigation,” Decision Support Systems V – Big Data Analytics for Decision Making. ICDSST 2015. Lecture Notes in Business Information Processing, 216: 10–21, 2015.

[35] B. Sanjib, S. Jaydip, “A proposed framework of next generation supply chain management using big data analytics,” presented at the National Conference on Emerging Trends in Business and Management: Issues and Challenges, , Kolkata, INDIA, March 17-18, 2016.

[36] R. Rossi, K. Hirama, “Characterizing big data management,” Issues Informing Scie. Inform. Technol. 12: 165-180, 2015.

[37] H. Borut, J. Jaklič, “Assessing benefits of business intelligence systems–a case study,” Manage. J. Contemp. Manage. Issues, 15(1): 87-119, 2010.

[38] S. Eybers S, J. H Strydom, Towards a classification framework of business intelligence value research. Italian Chapter of AIS (itAIS 2013), 2013.

[39] A. Carver, M. Ritacco, The Business Value of Business Intelligence. A Framework for Measuring the Benefits of Business Intelligence. Business Objects, 2006.

[40] M. Bahrami, M. Singhal, “The role of cloud computing rchitectour in big data, Chapter: Information Granularity, Big Data, and Computational Intelligence, Volume 8 of the series Studies in Big Data: 275-295, 2015.

[41] P. Russom, Integrating Hadoop into business intelligence and data warehouse, TDWI research. S. Madden, MIT, 2013.

[42] A. Abroshan, A. Harounabadi, J. Mirabedini, “Evaluation of software architecture using fuzzy colored Petri nets,” Manage. Scie. Lett., 3: 665-682, 2013.

[43] K. Dwivedi, S. K. Dubey, “Analytical review on Hadoop distributed file system”, presented at the 5th International Conference on Confluence the Next Generation Information Technology Summit: 25-26, Noida, India, 2014.

[44] J. Wang, X. He, Y. Deng, "Introducing software architecture specification and analyzing in SAM through an Example," Inform. Software Technol. J., 41(7): 451-467, 1999.

Biographies

Mohammad Reza Behbahani Nejad received the B.Sc. degree in Software Engineering from the Urmia University, Urmia, Iran and the M.Sc. degree in Software Engineering from the Islamic Azad University, Qazvin, Iran. He works on Business Intelligence and Big Data.

- Email: reza2005nejad@gmail.com
- ORCID: 0000-0001-6660-0699
- Web of Science Researcher ID: ADM-1867-2022
- Scopus Author ID: NA
- Homepage: NA

Hassan Rashidi is a Professor in Department of Mathematics and Computer Science of Allameh Tabataba'i University. He received the B.Sc. degree in Computer Engineering and M.Sc. degree in Systems Engineering and Planning, both from the Isfahan University of Technology, Iran. He obtained Ph.D. from Computer Science and Electronic System Engineering department of University of Essex, UK. His research interests include software engineering, software testing, and scheduling algorithms. He has published many research papers in International conferences and Journals.

- Email: Hrashi@gmail.com
- ORCID: 0000-0002-6588-5378
- Web of Science Researcher ID: AAE-2124-2022
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:
R. Behbahani Nejad, H. Rashidi, "A novel architecture based on business intelligence approach to exploit big data," J. Electr. Comput. Eng. Innovations, 11(1): 85-102, 2023.

DOI: 10.22061/JECEI.2022.8565.529

URL: https://jecei.sru.ac.ir/article_1727.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Pattern Measurement of Large Antenna by Sequential Sampling Method in Cylindrical Near-Field Test

M. Karimipour*

Department of Electrical Engineering, Arak University of Technology, Arak, Iran.

Article Info

Article History:
Received 18 February 2022
Reviewed 15 April 2022
Revised 31 May 2022
Accepted 25 June 2022

Keywords:
Cylindrical scanning
Fast Fourier technique
Near-field measurement

*Corresponding Author's Email Address:
m.karimipour@arakut.ac.ir

Abstract

Background and Objectives: Cylindrical scanning technique is a well-established indirect measurement method to characterize a wide range of antenna patterns such as fan-beam antennas and phased array antennas with versatile radiation patterns.
Methods: Cylindrical scanning technique which is based on the nearfield-to-far-field transformation based on cylindrical mode coefficients (CMCs), cannot predict the antenna radiation pattern with a very narrow beamwidth in the azimuth plane accurately, because a remarkable error occurs during the calculation of the derivative of high-order Hankel functions in the CMCs extraction. We aim to address this issue and introduce a simple yet rigorous technique namely the sequential sampling method (SSM) in conjunction with the two-dimensional Fast Fourier Transform (2D-FFT) to efficiently calculate the far-field radiation pattern of a super-directive antenna with a very narrow beamwidth in the azimuth plane. Briefly, the SSM offers several sequences of progressive azimuth angles and the corresponding order of Hankel functions in such a way that CMCs fully span 360 degrees of azimuth angles (φ) in the cylindrical coordinate system in each sequence. Afterward, by putting the far-field obtained by these sequences together, the final radiation pattern will have a high angular resolution. This technique can also be applied to determine the necessary criteria in the data acquisition step which should be satisfied to precisely measure the radiation pattern of super-directive antennas. These criteria are the maximum acceptable sampling resolution and the minimum value of the required azimuth angle (φ) in the data acquisition step if the far-field pattern is merely desired on the front side of the antenna.
Results: For verifications, the far-field radiation pattern of an electrically large slot array antenna including 81×15 slots is calculated at 8.75 GHz by the proposed technique and the results are compared with the array theory. The results show that the azimuth pattern can accurately be measured as small as 0.1° resolution by the SSM.
Conclusion: By comparing the results obtained by the proposed method and the traditional cylindrical scanning method, it can be inferred that the far-field pattern of an antenna with narrow beamwidth in the azimuth plane can easily be characterized by a cylindrical scanning system without any huge computational burden.

Introduction

Nowadays, there are several methods to measure antenna radiation patterns [1]. These methods are divided into two categories, namely direct and indirect. In the direct method, the far-field pattern is measured without any intermediate step. Depending on the

electrical size of antenna, the measurement can be performed either in an indoor anechoic chamber or in free space. For a wide range of antennas, it is not possible to measure the far-field radiation pattern directly. The main reason is the large electrical size of these antennas and the corresponding far zone region. For example, the size of antennas in phased array radars is approximately several wavelengths.

Therefore, the far-field region of these radiators will be more than several hundred meters. In this condition, setting up the far-field measurement system is a very complex and costly process. Therefore, measured results may not be valid due to several inevitable errors that may occur during the data acquisition. To overcome this drawback, indirect measurement methods have been introduced. In the indirect methods, some post-processing algorithms must be performed to determine the electromagnetic radiation pattern with high accuracy [2]. These post-processing algorithms are configured based on the kind of data acquisition. According to the geometric shape where the data are acquired, three well-established techniques, including planar, cylindrical, and spherical near-field measurements are introduced. Although, several other techniques such as data sampling over arbitrary geometries or curvatures are also reported in the literature [3]. These techniques are applied for special purposes.

Each of the aforementioned near-field methods is useful for testing a particular antenna. For example, the planar near-field method is useful for testing high directive antennas with a nontrivial front-to-back ratio of the pattern. When it is necessary to evaluate the radiation behavior of the antenna on the backside, the cylindrical near-field (CNF) system is a very useful solution. In most cases, the radiation pattern of array antennas with large size in the elevation plane and medium size in the azimuth plane can easily be measured by cylindrical scanning systems [2]. Conversely, the spherical near-field systems are useful to characterize medium and low-gain antennas with omnidirectional and isotropic-like patterns.

In the cylindrical scanning system, which is the case in this paper, the near-field data of the antenna under test (AUT), including amplitude and phase of electric fields, are acquired over a right circular cylinder or a cone-shape area via a simple probe. In practice, the acquisition process is performed by employing some measurement equipment and optical instruments (such as a laser tracker [1]) if high accuracy is needed. In the computational step as a post-processing task, the near-field to far-field transformation is performed to characterize the antenna pattern [4]-[8]. There are two main techniques including 2D-FFT and the matrix method to describe the far-field pattern from near-field information. The first one is fast, very efficient, and almost accurate [7], however, the near-field data should be acquired in a uniform grid. Meanwhile, the latter can be implemented for nonuniform sampling, and therefore it is a very suitable solution for considering probe position error during the data acquisition [8], [9]. The main disadvantage of the matrix method is the high computational cost of post-processing, which is unsuitable for pattern measurement of electrically large size antennas. Both techniques benefit from the description of electromagnetic fields outside the antenna by orthogonal basis functions namely cylindrical waves. If the sampling grid is regular, 2D-FFT can be efficiently applied to calculate the unknown coefficients of basis functions or CMCs. Afterward, the far-field pattern can easily be described by the 1D-FFT routine. Meanwhile, if the sampling grid is nonuniform, the derivation of the far-field pattern can be performed using the matrix operations [9] or interpolating the irregular NF data into regular [10], [11]. As pioneers in the field, AMETEK NSI-MI is the world leader in near-field measurements and produce a full range of standard and customized measurement system for certain applications [12]. NSI-MI says that the cylindrical near-field measurement system is useful to measure broad beams in the azimuth plane such as fan beams. In other words, the beamwidth of the azimuth pattern is a restricted factor for using the CNF system.

The basic theory behind the CNF systems expresses that the process of the far-field reconstruction pattern with an angular resolution of less than 0.5° in the azimuth plane involves describing near-field data with very high-order cylindrical basis functions [1]. In this fashion, the calculation of high-order CMCs corresponding to orthogonal basis functions may be associated with some errors. For example, the authors in [13] used a nearly huge grid including 512 samples on the φ-axis and 156 samples on the z-axis to accurately measure the far-field pattern of a very large L-band radar antenna. In this measurement, the maximum cylindrical mode function is in order of 256 which leads to the azimuth angular resolution of 0.7°. If the smaller resolution in the azimuth plane is necessary, the higher order of CMCs and Hankel function is required.

In this paper, a simple yet rigorous approach based on the SSM in conjunction with 2D-FFT is introduced to efficiently calculate the far-field radiation pattern with a very narrow beamwidth in the azimuth plane. The SSM offers several sequences of progressive azimuth angles along with the corresponding order of Hankel functions so that CMCs fully span 360 degrees for azimuth angles (φ) in each sequence. Afterward, by putting the far-field patterns obtained by these sequences together, the final radiation pattern will have a high angular resolution. We show that if these sampled sequences are arranged

together, the final resolution of the measured pattern in the azimuth plane would be as small as 0.1°. As an advantage of the proposed method, the monopulse antenna with a difference beam pattern in the azimuth plane can be rigorously measured with the accuracy of null position detection smaller than 0.1°.

To verify the concept, a 25×81 slot array antenna is considered as the AUT and the near-field data over a conceptual cylinder around the antenna is calculated via an ideal isotropic probe. The AUT is considered in the YZ plane such that the 81 elements are aligned on the y-axis. By modeling the radiation behavior of each slot by a finite-length magnetic dipole, a comprehensive mathematical framework is implemented to determine the radiation behavior of the entire slots in the array environment with every excitation distribution for slots provided that negligible mutual coupling exists between the slots. The proposed model enables the calculation of electromagnetic waves in both near-field and far-field regions. Therefore, the obtained results by the proposed model can also be employed to evaluate the near-field to far-field transformation algorithm developed by the SSM.

Analytical Model for Data Acquisition

The proposed model makes it possible to describe the radiated field of any slot or microstrip array in an arbitrary geometrical arrangement, flat or conformal, provided that the radiative behavior of the co-polarization (Co-Pol) and cross-polarization (Cross-Pol) components of the antenna can be modeled analytically. In addition, the model will be able to evaluate the radiated field in any desired area such as flat plane, cylinder, spherical, or even other conformal geometries such as cone. Therefore, one can employ it to verify the spherical scanning system as well. Note that this analytical method neglects the mutual coupling between the radiating elements in the array environment, hence, the model is valid while the coupling among the elements is negligible [14]. It is worth noting that the CADFEKO 2021 software is also able to calculate near-field waves around the antenna on a flat, cylindrical, spherical, and even cone areas. Therefore, it can be employed for data acquisition as well.

A. Theoretical Formulation

The radiation behavior of a single slot and microstrip patch can roughly be modeled by a magnetic and electric dipole, respectively. This model is also valid when these elements are arranged in an array configuration provided that the mutual coupling between elements is negligible. Therefore, the calculation process of electric and magnetic fields surrounding the dipole array with any arrangement leads to establishing a comprehensive analytical model for evaluating the radiation pattern of slot and microstrip array antennas. First, the mathematical formulation is introduced for the model, afterward, the model is verified by a case study simulated in the CST software.

According to the array theory, the electric field at any observation point in the free space can be described as the superposition of electric fields radiated from all elements. It is well-known that the electric field radiated by an electric and magnetic dipole which are aligned in the z-direction are as follows [15]:

$$Electric\ Dipole: E_\theta = G_\theta . I \tag{1}$$

$$Magnetic\ Dipole: E_\varphi = G_\varphi . I \tag{2}$$

where, G_θ is simply defined as below.

$$G_\theta = -j\eta \frac{e^{-jk_0 r_f}}{2\pi r_f}\left[\frac{cos\left(\frac{k_0 d_f}{2} cos\theta_f\right) - cos\left(\frac{k_0 d_f}{2}\right)}{sin\theta_f}\right] \tag{3}$$

Similarly, $G_\varphi = -G_\theta/\eta$. In (3), k_0 is the free space wavenumber, d_f is the dipole length which is greater than $\lambda/10$, (λ is the wavelength). The parameter η is the characteristic impedance in the free space and finally θ_f is the angular parameter defined in the spherical local coordinate system.

Equations (1) up to (3) are defined at the local coordinate system associated with each dipole element, i.e., (x_f, y_f, z_f) (See Fig. 1). To express the radiated field of the element in the global coordinate system, i.e., (x_g, y_g, z_g), one can employ the coordinate transformation technique presented in [16], and decompose the excitation current component of each dipole element into three components, I_x, I_y, I_z. Finally, regarding dipole directions, the electric field associated with each element is determined in every observation point in the global rectangular coordinate system. This scenario can be applied to every radiating element (including magnetic and electric dipole elements).

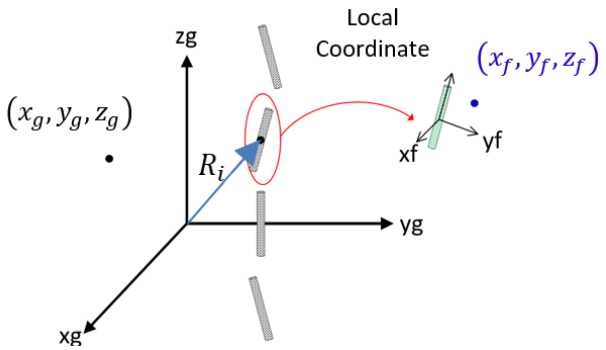


Fig. 1: General dipole array configuration along with the local and global coordinate system.

Using the superposition principle, the electric field radiated by all elements can be obtained at any arbitrary observation point.

Therefore, the general form of (1) can be represented in the matrix form as follows:

$$\begin{bmatrix}[E_x]_{N_s\times 1}\\ [E_y]_{N_s\times 1}\\ [E_z]_{N_s\times 1}\end{bmatrix}= \begin{bmatrix}[G_{xx}]_{N_s\times N_d} & [G_{xy}]_{N_s\times N_d} & [G_{xz}]_{N_s\times N_d}\\ [G_{yx}]_{N_s\times N_d} & [G_{yy}]_{N_s\times N_d} & [G_{yz}]_{N_s\times N_d}\\ [G_{zx}]_{N_s\times N_d} & [G_{zy}]_{N_s\times N_d} & [G_{zz}]_{N_s\times N_d}\end{bmatrix}\begin{bmatrix}[I_x]_{N_d\times 1}\\ [I_y]_{N_d\times 1}\\ [I_z]_{N_d\times 1}\end{bmatrix} \quad (4)$$

The conversion blocks, $[G_{ij}]_{N_s\times N_d}$, in (4) are defined to convert the dipole current in the '*j*' direction to the electric field component along the '*i*' direction. For example, if a specific dipole is aligned to the x-direction, all entries of conversion blocks, $[G_{ij}]_{N_s\times N_d}$, associated with that element are zero except the ones with $[G_{ix}]_{N_s\times N_d}$ indices. The parameters N_d and N_s are the number of dipoles and observation points, respectively. The derivation scenario of [G] blocks is summarized as follows:

- Determine $x_{g,NF}$, $y_{g,NF}$ *and* $z_{g,NF}$ according to the near-field sampling area. (For example, cylinder in this case)
- Determine r_g, θ_g and ϕ_g in the spherical coordinate system as follows*:*

$$\left(\theta_g, \phi_g\right)=\left(tan^{-1}\left(\frac{\sqrt{x_g^2+y_g^2}}{z_g}\right), tan^{-1}\left(\frac{y_g}{x_g}\right)\right)$$

$$r_g=\sqrt{x_g^2+y_g^2+y_g^2}$$

- Decompose each dipole to three dipoles along the x, y, and z-directions.
- Following [16] and determining the type of transfer matrix, namely $\Re$, for these three decomposed components which are defined for each dipole. For example: $\Re_{\mathrm{XZX}}, \Re_{ZXZ}$, $\Re_{ZYZ}$ or $\Re_{XYZ}$, where the $[\Re_{ijk}]$ denotes the rotation matrix around k, j, and i axes, respectively. This process should be done element by element in the array. It is worth noting that the dipole direction is the only determinative factor to use what form of $[\Re_{ijk}]$ needs to describe the electromagnetic behavior of dipole. The matrix, $[\Re_{ijk}]$, can be determined by multiplying three rotation matrices which are constructed from three decomposed components of the excitation current vector along the x-, y-, and z-directions. These three rotation matrices are defined based on the Euler angles α, β, γ. In the following, two forms of $[\Re_{ijk}]$ are described based on the Euler angles [16]:

$$\Re=\Re_{\mathrm{XZX}}(\alpha,\beta,\gamma)= \begin{bmatrix}\cos\gamma & \sin\gamma & 0\\ -\sin\gamma & \cos\gamma & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}1 & 0 & 0\\ 0 & \cos\beta & \sin\beta\\ 0 & -\sin\beta & \cos\beta\end{bmatrix}\begin{bmatrix}\cos\alpha & \sin\alpha & 0\\ -\sin\alpha & \cos\alpha & 0\\ 0 & 0 & 1\end{bmatrix}$$

$$\Re=\Re_{\mathrm{ZYZ}}(\alpha,\beta,\gamma)= \begin{bmatrix}\cos\gamma & \sin\gamma & 0\\ -\sin\gamma & \cos\gamma & 0\\ 0 & 0 & 1\end{bmatrix}\begin{bmatrix}\cos\beta & 0 & \sin\beta\\ 0 & 1 & 0\\ -\sin\beta & 0 & \cos\beta\end{bmatrix}\begin{bmatrix}\cos\alpha & \sin\alpha & 0\\ -\sin\alpha & \cos\alpha & 0\\ 0 & 0 & 1\end{bmatrix}$$

Accordingly, as shown in (4), three components of G are required to fully characterize the electromagnetic behavior of the dipole aligned in an arbitrary direction.

- Transfer the near-field points in the global coordinate system ($x_{g,p}$, $y_{g,p}$ *and* $z_{g,p}$) to the local coordinate system associated with the *mn*th element as follows:

$$\mathrm{x_{f,p}}=\Re_{11}(\mathrm{x_{g,p}}-\mathrm{x_{e,m}})+\Re_{12}(\mathrm{y_{g,p}}-\mathrm{y_{e,m}})+\Re_{13}(\mathrm{z_{g,p}}-\mathrm{z_{e,m}})$$
$$\mathrm{y_{f,p}}=\Re_{21}(\mathrm{x_{g,p}}-\mathrm{x_{e,m}})+\Re_{22}(\mathrm{y_{g,p}}-\mathrm{y_{e,m}})+\Re_{23}(\mathrm{z_{g,p}}-\mathrm{z_{e,m}})$$
$$\mathrm{z_{f,p}}=\Re_{31}(\mathrm{x_{g,p}}-\mathrm{x_{e,m}})+\Re_{32}(\mathrm{y_{g,p}}-\mathrm{y_{e,m}})+\Re_{33}(\mathrm{z_{g,p}}-\mathrm{z_{e,m}})$$

where p denotes the number of observation points in the near-field region and m is the element numbers in the array environment. In addition, the subscript g and f denote the local and global coordinate systems, respectively.

- Determine $r_{f,p}, \theta_{f,p}$ and $\phi_{f,p}$ in the spherical coordinate system from $\mathrm{x_{f,p}}$, $\mathrm{y_{f,p}}$ and $\mathrm{z_{f,p}}$.
- Calculate $G_{f,\theta}$ and $G_{f,p}^{\phi}$ for magnetic dipole are as follows [15]:

$$G_{f,p}^{\theta}=-j\eta\frac{e^{-jk_0r_{\mathrm{f,p}}}}{2\pi r_{\mathrm{f,p}}}\left[\frac{cos\left(\frac{k_0d_{\mathrm{f,p}}}{2}cos\theta_{\mathrm{f,p}}\right)-cos\left(\frac{k_0d_{\mathrm{f,p}}}{2}\right)}{sin\theta_{\mathrm{f,p}}}\right],\quad G_{f,p}^{\phi}=0$$

- Calculate $G_{f,p}^{x}$, $G_{f,p}^{y}$ and $G_{f,p}^{z}$ from $G_{f,p}^{\theta}$ and $G_{f,p}^{\phi}$ components determined in the previous step.
- Transform $G_{f,p}^{x}$, $G_{f,p}^{y}$ and $G_{f,p}^{z}$ to the global coordinate system $G_{g,p}^{x}$, $G_{g,p}^{y}$ and $G_{g,p}^{z}$, as follows:

$$G_{g,p}^{x}=\mathbb{Q}_{11}G_{f,p}^{x}+\mathbb{Q}_{12}G_{f,p}^{y}+\mathbb{Q}_{13}G_{f,p}^{z}$$
$$G_{g,p}^{y}=\mathbb{Q}_{21}G_{f,p}^{x}+\mathbb{Q}_{22}G_{f,p}^{y}+\mathbb{Q}_{23}G_{f,p}^{z}$$
$$G_{g,p}^{z}=\mathbb{Q}_{31}G_{f,p}^{x}+\mathbb{Q}_{32}G_{f,p}^{y}+\mathbb{Q}_{33}G_{f,p}^{z}$$

where $[\mathbb{Q}]=[\Re]^{-1}$.

Following the above guideline, all entries of coefficient matrix ($[\boldsymbol{G}]$) represented in (4) are determined whose dimension is $3N_s \times 3N_d$. The factor 3 observed in the matrix dimension returns to the three decomposed components of each dipole in the array antenna.

B. Verification of the Analytical Model for Data Acquisition

As an example, the radiation behavior of the array shown in Fig. 2(a) is characterized by the proposed model. 50 elements of magnetic dipoles are placed together in the YZ plane with θ angles shown in Fig. 2(b) The θ angles are defined relative to the z-axis. The working frequency and dipole spacing are considered to be 8.75GHz and

24mm, respectively. The current magnitudes of the dipoles are set in the form of Taylor distribution with a side-lode level (SLL) of -40dB and $\bar{n} = 4$ [15]. The progressive phases of the dipoles are set to zero, hence, the main beam of the pattern is aligned at $\theta = 90°$. Fig. 3 shows the [G] tensor defined in (4) as well as the far-field co-pol and cross-pol of the pattern at $\theta = 90°$. It is worth noting that an interesting application of this analytical model is to solve the inverse problem and extract the dipole currents from known E-field distribution around the array (which can be obtained by planar near-field measurement). As such, the *array calibration* can also be performed by this method. That is, by sampling the near-field around the antenna and finding the amplitude and phase of excitation current of the elements, it is possible to compare the real state of every element in the array configuration with their ideal states, and then, appropriate modifications can be adopted for faulty elements.

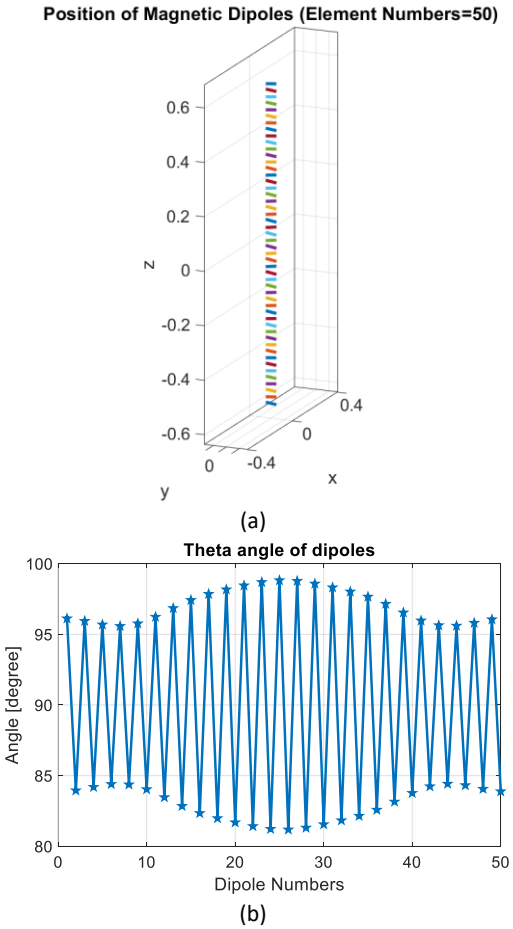


Fig. 2: (a) Dipole array configuration. (b) The values of θ angles relative to the z-axis for determining the element directions.

To further verify the proposed method, a 50-element slot array antenna with slot angles defined in Fig. 2(b) is simulated by the time-domain solver in the CST software, and the electric field data is extracted on a line at x=10 cm. In addition, the far-field pattern of the antenna is simulated and recorded in $\varphi = 0°$ plane for comparision with the results obtained by the theoretical model. The theoretical model is established by arranging 50 elements of magnetic dipoles together in such a way that the dipole directions are exactly aligned to the slot directions and dipole center positions coincide with slot centers. Dipole currents can be obtained by solving the inverse problem, $\bar{I} = \bar{\bar{G}}^{-1}\bar{E}_{NF}$, where $\bar{E}_{NF}$ is the near-field data extracted from the CST software. In the next step, the far-field pattern of the antenna is calculated by the theoretical model. This is simply accomplished by considering the observation points (r_g, θ_g and ϕ_g) in the far zone on the $\varphi = 0°$ plane.

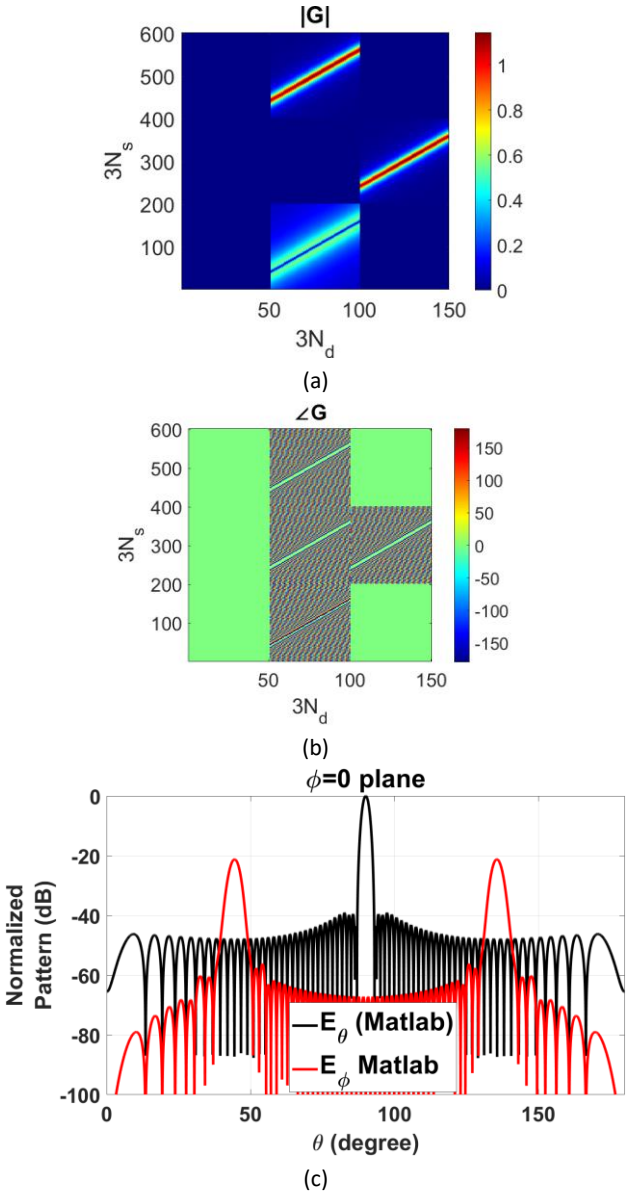


Fig. 3: (a) Amplitude and (b) phase of [G] tensor for antenna configuration shown in Fig. 2. (c) Calculated far-field radiation pattern on the $\phi = 0°$ plane.

Fig. 4 shows the simulation setup in the CST software to extract $\bar{E}_{NF}$ and the far-field pattern. The comparison

of the far-field pattern obtained by the CST software and the theoretical model is presented in Fig. 5. The results show that the proposed model enables to fully characterize the slot array antenna pattern. Therefore, the analytical model for data acquisition is prepared to extract the near-field data on a cylinder, which is a crucial step for setting up a CNF system.

Fundamentals of CNF Measurement Technique

The electromagnetic waves around an arbitrary antenna can be described by a set of CMCs as follows [1]:

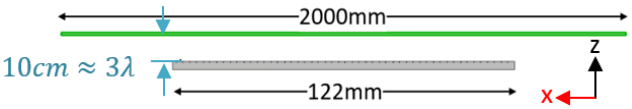


Fig. 4: Characterization of slot array radiation pattern by sampling nearfield data over the line at x=10 cm.

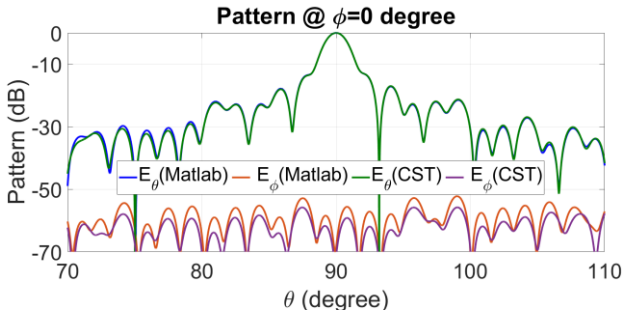


Fig. 5: The comparison between the far-field pattern of the slot array extracted from the theoretical model and the full-wave simulation.

$$E_z^{NF}(\varphi,z) = \sum_{n=-M}^{M}\int_{-\infty}^{+\infty} b_n(k_z)\frac{k_\rho^2}{k_0}H_n^{(2)}(k_\rho a)e^{jn\varphi}e^{-jk_z z}\,dk_z \quad (5)$$

$$E_\varphi^{NF}(\varphi,z) = \sum_{n=-M}^{M}\int_{-\infty}^{+\infty}\left\{b_n(k_z)\frac{nk_z}{ak_0}H_n^{(2)}(k_\rho a) - a_n(k_z)\frac{\partial H_n^{(2)}(k_\rho r)}{\partial r}\Big|_{r=a}\right\}e^{jn\varphi}e^{-jk_z z}\,dk_z \quad (6)$$

In (5) and (6), n is the mode index and M is the maximum number of modes that should be defined based on the accuracy of the calculations in the φ direction. The parameter a is the scanner cylinder radius. According to [1], $M = k_0 r_t + M_0$ where r_t is the minimum radius of a conceptual cylinder encompassed the AUT and M_0 is chosen larger than 10 corresponding to the required accuracy. For example, the pattern measurement of a large antenna with small beamwidth in the φ direction, requires a large number of CMCs for the far-field calculations.

As stated in [1], since the maximum separation among the sampling data in the φ direction is forced by $\Delta\varphi \le \frac{\pi}{M}$, we do not have freedom for choosing number of modes and data sampling steps in the φ direction, simultaneously. Therefore, to measure the patterns with narrow beamwidth in the azimuth plane, first, it should be selected the desired φ steps and, then determined the maximum mode number. The radial wavenumber, k_ρ, in (5) and (6) is equal to $\sqrt{k_0^2 - k_z^2}$, where $k_z = -\frac{\pi}{\Delta z_{scan}} + \frac{2m\pi}{n_{k_z}\,\Delta z_{scan}}$ and $m = [0, n_{k_z} - 1]$. The parameter n_{k_z} is the number of points in the spectral domain along the z-axis which is in the form of $2^q (q > 9)$. The coefficients $a_n(k_z)$ and $b_n(k_z)$ in (5) and(6) are CMCs which can easily be obtained by some mathematical manipulations [1]:

$$b_n(k_z) = \frac{k_0}{k_\rho^2 H_n^{(2)}(k_\rho a)}\mathcal{E}_v(n,k_z) \quad (7)$$

$$a_n(k_z) = \frac{1}{\frac{\partial H_n^{(2)}(k_\rho r)}{\partial r}\Big|_{r=a}}\left[b_n(k_z)\frac{nk_z}{ak_0}H_n^{(2)}(k_\rho a) - \mathcal{E}_H(n,k_z)\right] \quad (8)$$

Equations (7) and (8) show that CMCs are obtained in matrix form, that is $[a_n]_{n_{k_z}\times n}$ and $[b_n]_{n_{k_z}\times n}$. The quantities, $\mathcal{E}_v(n,k_z)$, and $\mathcal{E}_H(n,k_z)$ are the spectral components of the tangential near-field data on the sampling cylinder as follows [1]:

$$\mathcal{E}_{v|H}(n,k_z) = \frac{1}{4\pi^2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} E_{z|\varphi}^{NF}(\varphi,z)\,e^{-jn\varphi}e^{jk_z z}d\varphi dk_z = \frac{1}{4\pi^2} fftshift\left(FFT2\left(E_{z|\varphi}^{NF}(\varphi,z), n_{k_z}, n\right)\right) \quad (9)$$

Finally, the far-field components of electric fields can be described based on CMSs as follows [1]:

$$E_\theta^{FF}(\theta,\varphi) = -2jk_0 sin\theta\frac{e^{-jk_0R}}{R}\sum_{n=-N}^{N} j^n b_n(k_0 cos\theta)e^{jn\varphi} = -2jk_0\sqrt{1-\frac{k_z^2}{k_0^2}}\frac{e^{-jk_0R}}{R}\mathbb{C}[fft(b_n(k_0 cos\theta)e^{jn\varphi})] \quad (10a)$$

$$E_\varphi^{FF}(\theta,\varphi) = -2jk_0 sin\theta\frac{e^{-jk_0R}}{R}\sum_{n=-N}^{N} j^n a_n(k_0 cos\theta)e^{jn\varphi} = -2jk_0\sqrt{1-\frac{k_z^2}{k_0^2}}\frac{e^{-jk_0R}}{R}\mathbb{C}[fft(a_n(k_0 cos\theta)e^{jn\varphi})] \quad (10b)$$

In the above equations, the operator FFT2 is the two-dimensional Fast Fourier transform which can be efficiently evaluated by MATLAB software with a computational burden of $N(logN)$. Similarly, the operator "$fftshift$" is employed to shift the zero-frequency component to the center of the spectrum in MATLAB.

The operator $\mathbb{C}$ in (10b) returns the Fourier transform of each row of a_n matrix. As can be seen in (9), and (10) both near-field data extraction and the far-field pattern reconstruction processes can be performed by the Fast Fourier technique, (the two-dimensional form for the near-field data extraction and the one-dimensional form for the far-field calculation).

Sequential Sampling Method

A. Basic Concept

During the computation of CMCs in the CNF system, it is necessary to keep in mind that the choice of the sampling steps along the φ direction or, equivalently, the choice of mode numbers should be done in such a way that the cylindrical basis functions span a full cycle of 360°, without any overlap or gap among the samples [1]. The discussion presented in the previous section shows that when the sampling process in the φ direction is done with very small steps, e.g., 0.1°, it means that the parameter n in (5) and (6) should be selected in the interval $[-1800, 1800]$. This guarantees a full span of 360° in the azimuth plane by the steps of 0.1°. In other words, the derivative of Hankel-function type two should be calculated for mode numbers up to 1800. Our investigations show that the MATLAB software can return the values of the derivative of Hankel-function type two by a maximum mode number of 500 with negligible error. This mode number corresponds to the sampling step of 0.36° in the φ direction. This bottleneck restricts the capability of the CNF system. That is why the CNF system is almost applied to measure specific antennas with relatively wide beamwidth in the φ direction. If needed, the interpolation technique is used in classical approaches to obtain better results in the φ direction [1], [17]-[19]. However, in most cases, the interpolation is unable to accurately predict the antenna pattern at some angles where sharp variations occur in the beam such as the null position in the monopulse patterns.

Closer examination of (5) up to (10) makes clear the fact that the far-field pattern obtained by the CNF measurement system in the φ direction is in discrete summation form of electric fields, that is, the measured far-field pattern is calculated at those points that are sampled in the φ direction. For example, if the near-field data are recorded in the φ direction by 0.2° steps, the final measured pattern has a resolution of 0.2°. This is in contrast to the measured pattern in the z-direction (or elevation plane). In fact, the accuracy of the pattern in the z-direction can be enhanced by simply padding the near-field data in the z-direction. (See (9)). According to this fact, the far-field pattern in the φ direction can be reconstructed by a repetitive routine in which the high-resolution pattern is constructed part by part. To do this, we offer the sequences of sampling points in the φ direction to obtain the fine resolution yet keep the accuracy of the computation (by limiting the mode numbers to 500 or smaller). To better explain the proposed method, let us assume that the required accuracy of the measured pattern in the φ direction is 0.1°. (b)

Fig. 6 depicts the proper sequences of the φ angles where the near-field to-far-field transformation should be performed for each sequence to obtain P_i. By putting the P_i patterns together, the resulted pattern has a fine resolution as desired (i.e., 0.1°). As can be seen in Fig. 6, the data acquisition process should be accomplished with the desired resolution to provide the required data for each sequence. In the next step, the number of sequences, N_{seq}, should be determined based on $\Delta\varphi_{seq}$. In this regard, it can be expressed the number of sequences as follows:

$$N_{seq} = \frac{\Delta\varphi_{seq}}{\Delta\varphi_{Data\ acquisition}} = \frac{0.5}{0.1} = 5 \tag{11}$$

Acquisition data	0	0.1	0.2	0.3	0.4	0.5	....
	300	...	359.7	359.8	359.9		
M_Total	1800						

(a)

Sequsence Number	Phi angles to calculate i^{th} pattern relating to the i^{th} sequence						
Step 1	0	0.5	1	1.5	2	...	359.5
	Maximum Mode Number: $M_{seq1} = 360$ Calculated Far-field Patter= P_1						
Step 2	0.1	0.6	1.1	1.6	2.1	...	359.6
	Maximum Mode Number: $M_{seq2} = 360$ Calculated Far-field Patter= P_2						
Step 3	0.2	0.7	1.2	1.7	2.2	...	359.7
	Maximum Mode Number: $M_{seq3} = 360$ Calculated Far-field Patter= P_3						
Step 4	0.3	0.8	1.3	1.8	2.3	...	359.8
	Maximum Mode Number: $M_{seq4} = 360$ Calculated Far-field Patter= P_4						
Step 5	0.4	0.9	1.4	1.9	2.4	...	359.9
	Maximum Mode Number: $M_{seq5} = 360$ Calculated Far-field Patter= P_5						

(b)

Fig. 6: (a) The φ angle values needed for the data acquisition process by the SSM. (b) Proper φ angle sequences to limit the maximum mode numbers to 360 which leads to the calculation of P_i pattern with a high accuracy in MATLAB.

B. Verification of the SSM for Sum Pattern

To confirm the advantage of the SSM, a slot-array antenna including 25 rows and 81 columns is considered as the AUT (See Fig. 7). As shown in Fig. 7, the elements are located on the YZ plane. Since all 81 elements are in

the Y-axis direction, the antenna beamwidth in the azimuth plane will be narrow (about 1.4°), and hence, this is a good case to benchmark the SSM advantage. The working frequency is considered 8.75 GHz with the corresponding wavelength around 34 mm. The distance between each of the 81 elements in each row is 21.5mm. The width of each slot is equal to 2 mm. The separation among rows is 15 mm.

Therefore, the antenna dimension is 1.72 m × 0.36 m. The angle of slots with respect to the z-axis is determined by following the design method presented in [14]. Discussion about the slot array design method goes beyond the scope of the present paper.

Note that in a practical cylindrical scanner, the data acquisition process can be performed by rotating the AUT through a rotary positioner and moving a standard probe along the z-axis. The CNF configuration including the sampling points in the near-field region, where the tangential electric fields should be recorded, is shown in Fig. 7. The near-field components of the electric fields on the cylinder can be theoretically calculated by following the procedure outlined in the previous section. The radius of the scanning cylinder is considered to be 1.15m to avoid any physical contact between the antenna and probe during the data acquisition.

Fig. 8 shows the tangential electric field components of the slot array at 8.75 GHz which are depicted on the sampling cylinder. The height of the cylinder is considered as 2m to cover an acceptable angular range for the pattern in the elevation plane. As can be seen in Fig. 8, the peak values of the near-field components occur at the boresight direction, therefore, it is expected that the main beam of the far-field pattern is aligned in the boresight direction.

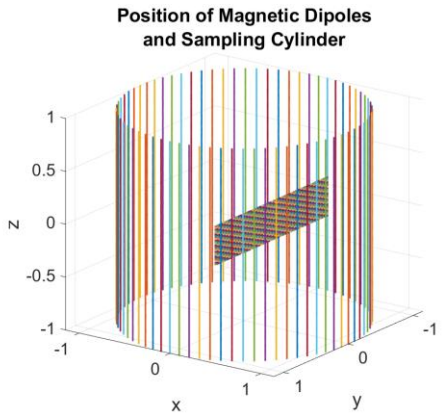
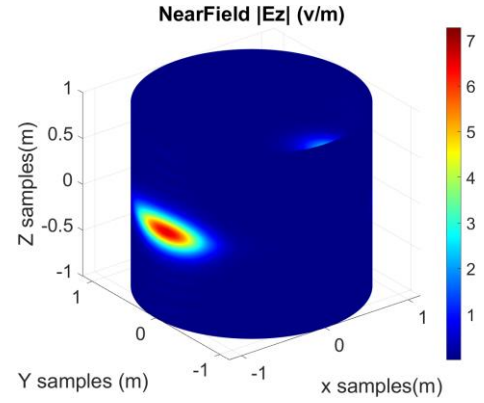


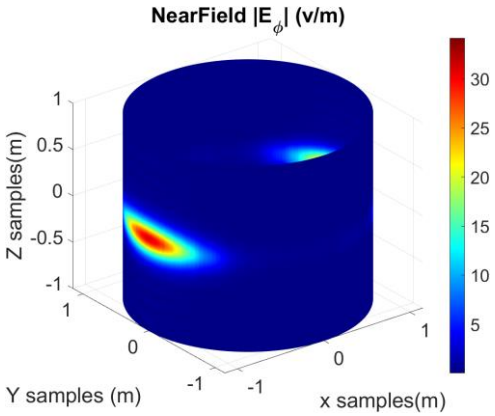
Fig. 7: Cylindrical sampling configuration developed in MATLAB.

According to the antenna dimensions, the maximum mode number is $M \approx 186$ assuming $M_0 = 10$, which leads to $\Delta\varphi < 0.96°$. Near-field-to-far-field transformation by the use of the sampling data with the step of $\Delta\varphi = 0.9°$ is accomplished in the azimuth plane. The results are shown in Fig. 9a. It can be observed that the measured pattern cannot follow the overall behavior of the desired pattern, especially in sidelobe regions and so the results are not satisfactory. If $\Delta\varphi = 0.5°, 0.1°$ corresponding to $M = 360, 1800$, the measured patterns obtained by the CNF system are shown in Fig. 9(b) and Fig. 9(c), respectively.

The reduction of the sampling step should naturally increase the accuracy of antenna pattern measurement, but this has not happened. This problem is solved by properly using the SSM.



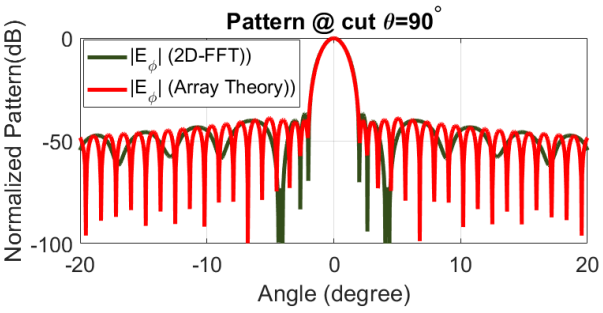
(a)



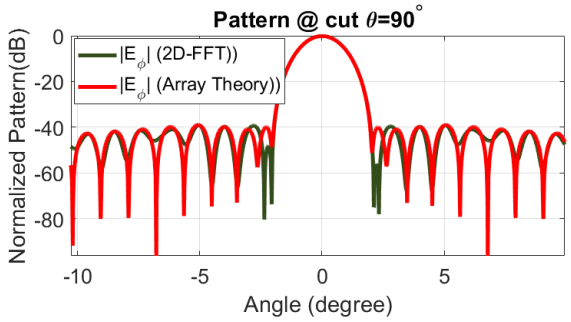
(b)

Fig. 8: Tangential electric field components of the slot array at 8.75 GHz. The sampling step in the φ and z directions are 0.1° and 1 mm, respectively. Note that the sampling step in the z-direction should satisfy the Nyquist theorem that is smaller than half of the wavelength.

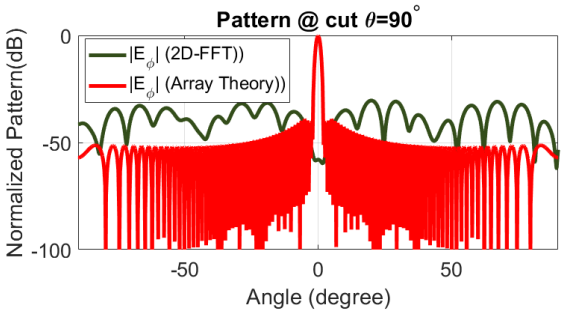
The results obtained by employing the SSM are shown in Fig. 10 which confirm the usefulness of applying the SSM in the CNF measurement system. Note that the far-field patterns represented in Fig. 9 and Fig. 10 with the label '*array theory*' are calculated by the same algorithm developed for near-field data acquisition. The difference is that (x_g, y_g, z_g) are defined in far-field region similar to that presented in Fig. 3.



(a)

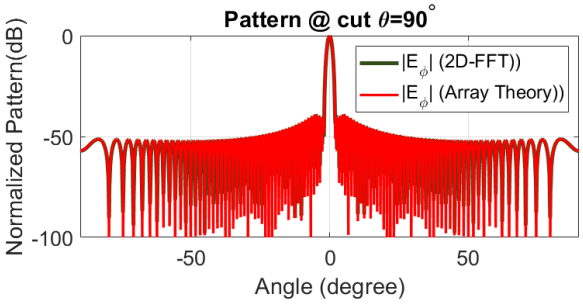


(b)

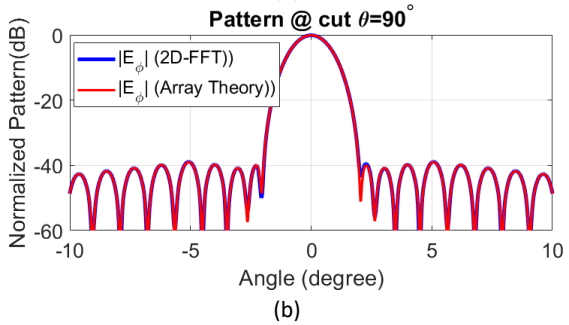


(c)

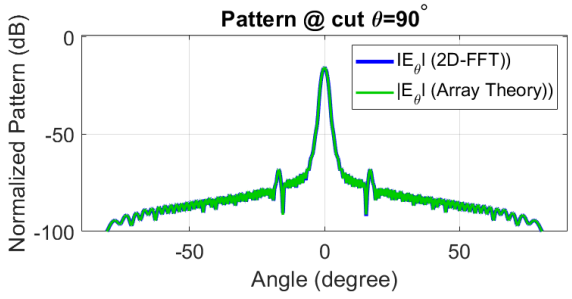
Fig. 9: The comparison between the far-field results obtained by the ideal array theory and the CNF system without using the SSM. The parameter $\Delta\varphi$ is considered as (a) 0.9° (b) 0.5° (c) 0.1°.



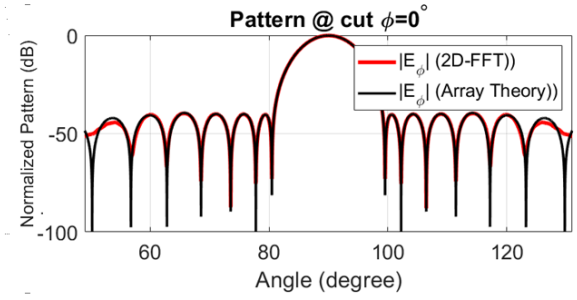
(a)



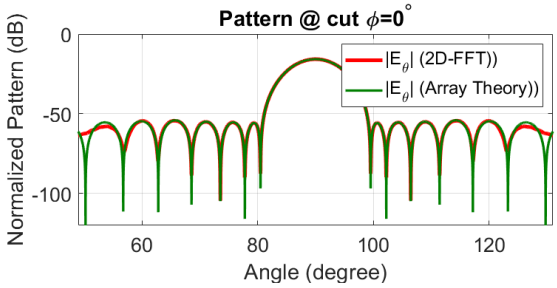
(b)



(c)



(d)



(e)

Fig. 10: The comparison between the far-field results obtained by the ideal array theory and the CNF system by using the SSM. $\Delta\varphi_{Data\ acquisition} = 0.1°$, $\Delta\varphi_{seq} = 0.5$, and $N_{seq} = 5$. (a) $|E_\varphi|$ at $\theta = 90°$. (b) closer view of $|E_\varphi|$ at $\theta = 90°$. (c) $|E_\theta|$ at $\theta = 90°$ (d) $|E_\varphi|$ at $\varphi = 0°$ (e) $|E_\theta|$ at $\varphi = 0°$.

C. Verification of the SSM by Measuring the Difference Pattern

It is well-known that monopulse antennas are widely used in radar systems for tracking purposes. Depending on the design goal, a deference pattern can be made in both azimuth and elevation planes. Detecting the null position in deference pattern and slope value of the beam around the null position, which all have a direct effect on the angle tracking error, are the main challenges in the pattern measurement of these antennas. For example, if the aperture size of a radar antenna is large in the azimuth plane (similar to that presented in Fig. 7) and the antenna is designed to radiate an azimuth difference pattern, the resulted monopulse pattern will be formed in the azimuth plane with a narrow beamwidth. In this fashion, antenna pattern measurement in the azimuth plane is difficult with the CNF system. This is why a large number of CMCs is required to accurately describe the far-field pattern. Therefore, as mentioned before, the measurement

results experience considerable errors if the conventional cylindrical scanning method is used.

In this subsection, an azimuth difference pattern with narrow beamwidth is generated by properly exciting the array elements described in section B. Afterward, the far-field radiation pattern is calculated with and without using the SSM in the CNF system to further highlight the advantage of the SSM to calculate the difference pattern with high accuracy, especially around the null position. In doing so, Taylor and Bayliss distributions in the elevation and azimuth planes are considered respectively as the desired excitation functions. For both Taylor and Bayliss functions, $\bar{n}$ and sidelobe levels are considered 4 and -40 dB, respectively. Fig. 11 shows the normalized excitation current distribution of the array to generate a difference pattern in the azimuth plane. Near-field components including E_z and E_ϕ are sampled on a cylinder with the radius and height equal to 1.15m and 2m, respectively as shown in Fig. 12. It can be observed that E_ϕ component has a null in the front side of the antenna which leads to the generation of a null in the far zone. Also, E_ϕ is a stronger field than E_z on the cylinder indicating that the major contribution of the far-field pattern is provided by E_ϕ.

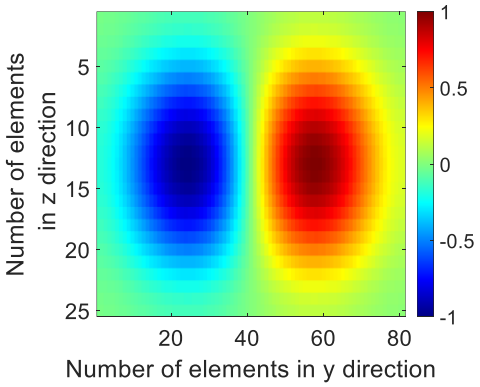
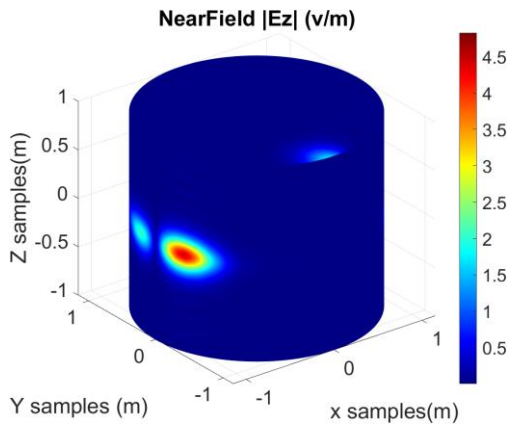
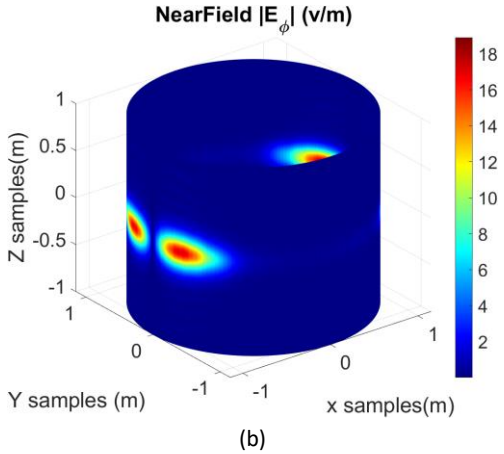


Fig. 11: Excitation current distribution of the array with the configuration shown in Fig. 7.



(a)



(b)

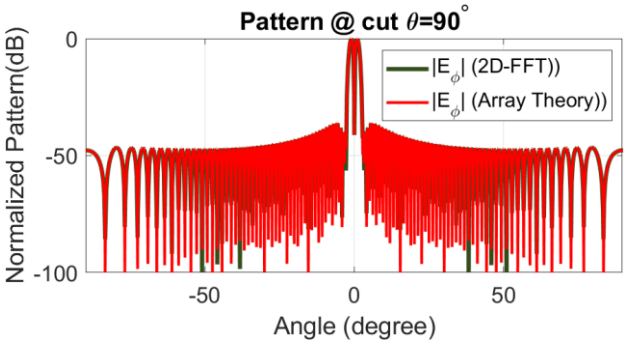
Fig. 12: Near-field components of the electric field on the scanning cylinder.

As the sampling step varies from $0.9°$ to $0.1°$, the observations are similar to those presented for the sum pattern in Fig. 9. To conserve space, the far-field results have not been presented in this subsection. A proper explanation that can be given in this regard is that when $\Delta\varphi$ is considered as $0.9°$, it can be inferred that the near-field-to-far-field transformation cannot predict the radiation pattern accurately. When the sampling step is reduced from $0.9°$ to $0.5°$, the resulted pattern will be more accurate; but as the sampling step decrease from $0.5°$ to $0.1°$, a considerable error is observed in the results. The reason behind this fact comes from the inaccuracy of calculating the CMCs during the transformation of the near-field data to the far-field pattern as the number of modes increases. Furthermore, the error due to the use of more CMCs in the far-field calculation process is much greater than the error due to the use of large step sizes in data acquisition.

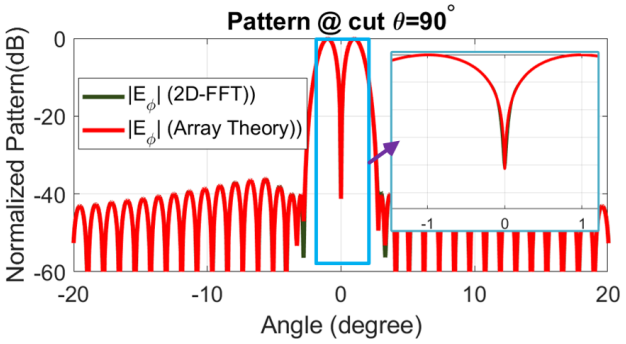
Note that this error affects the angle tracking error because the slope and position of the null in the difference pattern suffer some errors. Our investigations show that this error enhances when the beam is scanned. To prove this claim, a scanned beam is generated by the array and the far-field pattern of the antenna is calculated by two techniques, i.e., array theory and the classical CNF method. To do this, a given progressive phase is applied to the rows and columns of the array to scan the beam in the elevation and azimuth planes by 20° and 15°, respectively. It is found that the pattern in the side-lobe region along with the null position of the main beam and slope of the beam around the null include some errors which are more than the errors in broadside radiation.

In light of the above discussion, the need for an accurate method to calculate the difference pattern for all beam directions is inevitable. The classical CNF suffers from inaccuracy and the SSM technique can significantly enhance the measurement accuracy. Fig. 13 shows the far-field pattern of the antenna by the use of the SSM. In

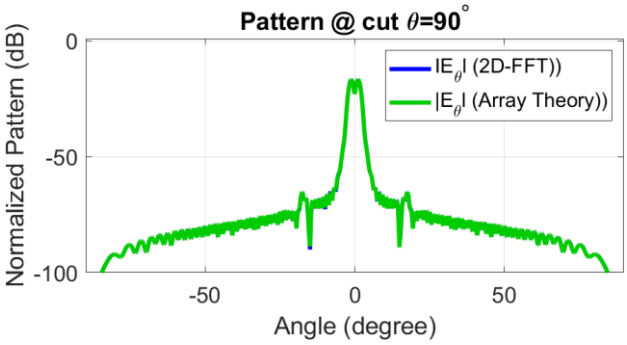
the SSM, $\Delta\varphi_{Data\ acquisition} = 0.1°$ and $\Delta\varphi_{seq} = 0.5$, $N_{seq} = 5$ are considered. In this design, the null position is considered at the broadside. It is observed that both co-pol and cross-pol patterns are calculated properly when compared with the ideal results.



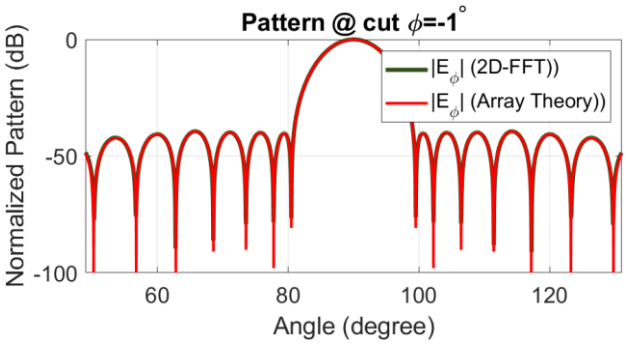
(a)



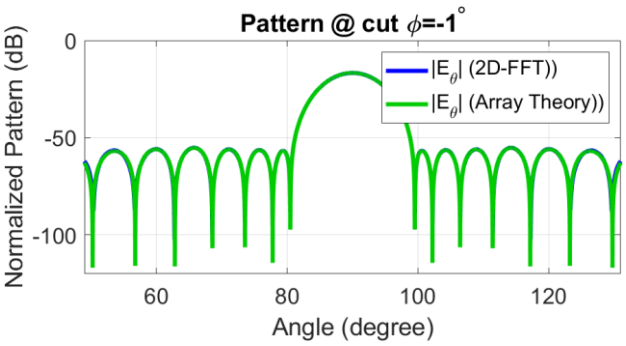
(b)



(c)



(d)



(e)

Fig. 13: The comparison between the far-field results of the difference pattern obtained by ideal array theory and the CNF system by using the SSM. $\Delta\varphi_{Data\ acquisition} = 0.1°$ and $\Delta\varphi_{seq} = 0.5$, $N_{seq} = 5$. (a) $|E_\varphi|$ at $\theta = 90°$. (b) closer view of $|E_\varphi|$ at $\theta = 90°$. (c) $|E_\theta|$ at $\theta = 90°$ (d) $|E_\varphi|$ at $\varphi = -1°$ (e) $|E_\theta|$ at $\varphi = -1°$.

Reducing Data Acquisition and Single-Cut Measurement Criteria in CNF system with the SSM

Reducing the data acquisition process is of interest in many mass production scenarios in which many identical antennas need to be measured. This work is performed by eliminating unnecessary data in the near-field region. For example, if the position and slope of the null in a monopulse pattern are required, there is no need to completely sample all of the near-field data in the three-dimensional space.

Equivalently, the shape of the main beam and first sidelobe levels are related to the near-field data on the front side of the antenna. This issue is also valid for necessary components of the electric field (E_z or E_φ or both of them) which should be sampled based on the required information about the pattern. (I) Reducing the sampling area to a sector with a given central angle, (II) single-cut measurement, (III) and choosing the necessary components of the electric field as desired, all significantly accelerate the pattern measurement process. Furthermore, in the factory calibration process of electrically large scanned-beam active phased array antennas, the fast and single-cut measurement would be helpful to find the possible errors in the antenna configuration very quickly. The CNF system in conjunction with the SSM is a good candidate to accomplish this task with acceptable accuracy. The CNF system in its simplest form can be employed as a single-cut measurement technique which is named the zero-height CNF system. The near-field-to-far-field transformation is done over a zero-height ring enclosing the AUT. In this section, several scenarios including sector sampling and single-cut sampling are investigated by the CNF system in conjunction with the SSM to determine the required criteria for reducing the unnecessary near-field data.

A. CNF and SSM based on Sector Sampling Area

The combination of the CNF and SSM as a powerful solution to characterize the antenna radiation pattern can be employed to determine the amount of near-field data required to fully describe the radiation field in the front of the antenna with minimal error. Fig. 14 shows the sector sampling scenario for pattern measurement of antenna described in the previous section. The results are presented in Fig. 15 up to Fig. 17 for various values of α angle (α=50° and α=120°).

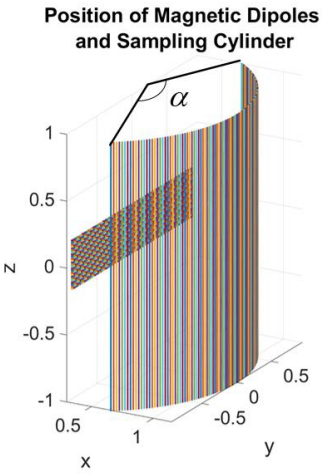
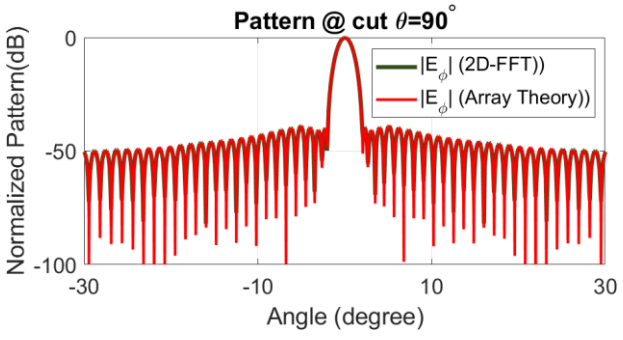
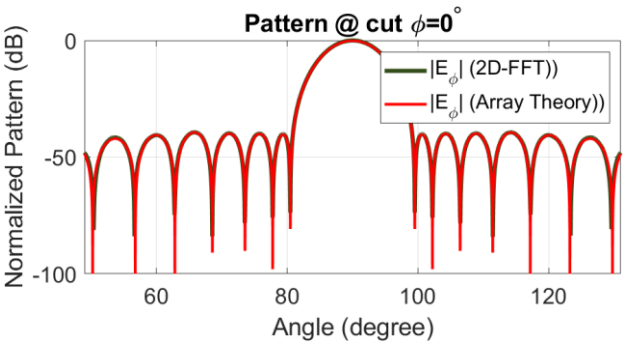


Fig. 14: Sector sampling area subtended by α angle.

It can be observed from Fig. 15 up to Fig. 17 that as far as the far-field on the front side of the antenna is concerned, one can extract the near-field data on a sector area with the subtended angle of α=120° around the peak value of the beam in the azimuth pattern. This fact is also valid for other types of antennas with narrow beamwidth in the azimuth plane and it is possible to find the minimum value of α angle to characterize the beam in front side of the antenna by the CNF system and SSM. Note that, if the SSM is not used to calculate the fields, the results will not be valid at all and we cannot determine the desired subtended angle, α, associated with the sampling sector area to accurately calculate the far-field patterns.

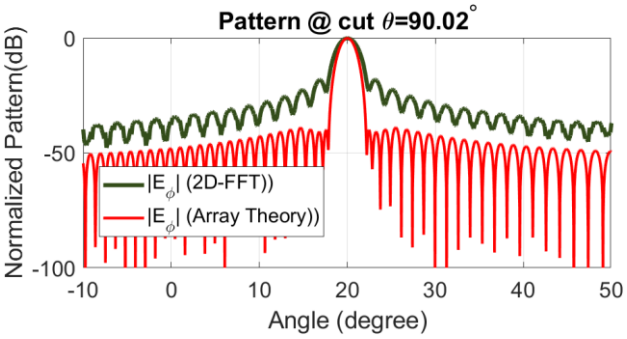


(a)

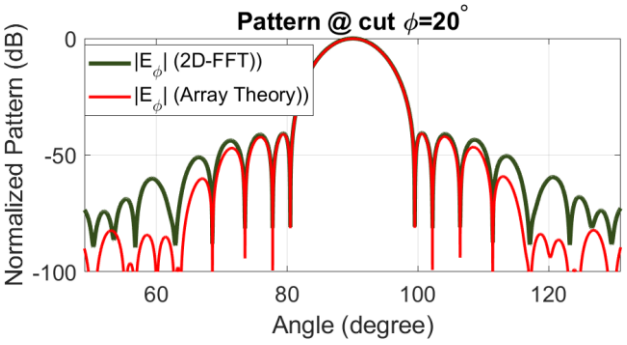


(b)

Fig. 15: Co-pol components of the pattern obtained by the sector sampling scenario with α=120° in (a) azimuth (b) elevation planes.

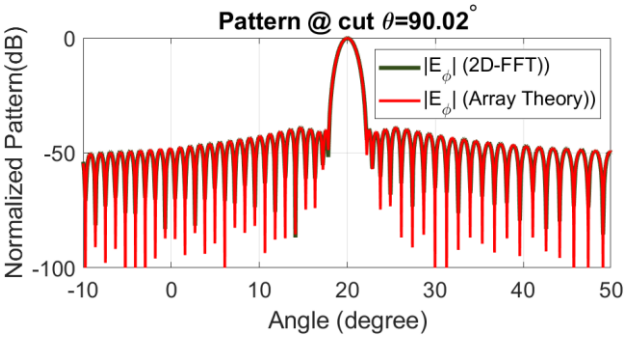


(a)

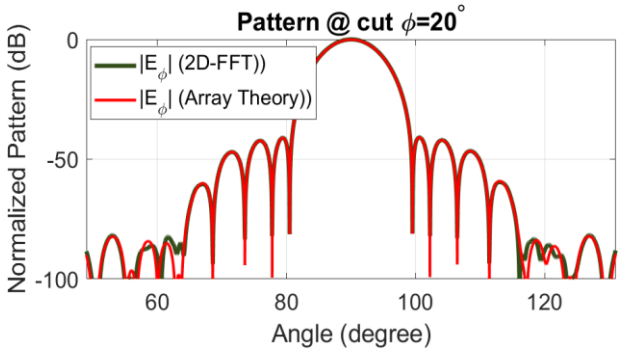


(b)

Fig. 16: Co-pol component results of a scanned beam pattern obtained by sector sampling scenario with α=50° in (a) azimuth (b) elevation planes. The beam is scanned to $\varphi = 20°$.



(a)



(b)

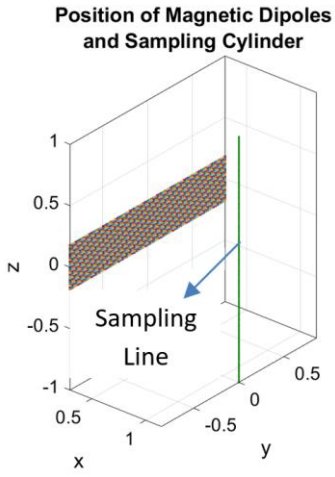
Fig. 17: Co-pol components of a scanned beam obtained by sector sampling scenario with α=120° in (a) azimuth (b) elevation planes. The beam is scanned to $\varphi = 20°$.

The results show that if we sampled a sector area around the main beam with the subtended angle equal to 120°, the calculated radiation pattern in the front side of the AUT has a good agreement with the ideal pattern. This argument can be made for other AUT with different beamwidth and beam directions and optimal subtended angle related to the sampling sector would be obtained based on the desired accuracy.

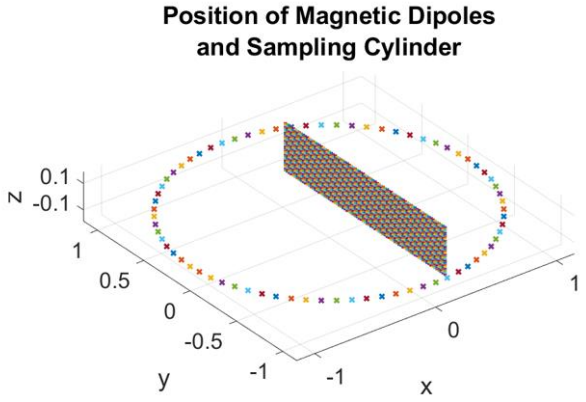
B. CNF and SSM based on Single-Cut Data Acquisition

As a special case of three-dimensional cylindrical scanning measurement, single-cut measurement (SCM) along the z and φ directions are an interesting solution to characterize the antenna patterns [20]. In general, the SCM drastically reduces the testing time; however, this technique is not useful for any type of antenna. For example, the authors in [23] exploited the SCM technique to characterize a large size one-dimensional array antenna (based station antenna as a practical case), in which the pattern is wide in the azimuth plane and directional in the elevation plane. In this fashion, the testing time can further be decreased by using multi-probe measurement, because the measurement is done only in one dimension [24]-[30].

Generally, the SCM can be implemented by the CNF system in two different states: (I) linear moving of the probe along the z-direction and (II) moving the probe in the φ- direction. The latter state is well-known as a zero-height CNF measurement. In classical CNF, it is difficult to apply the SCM scenario in the φ- direction due to the limitation of the CNF system in the measurement of the patterns with narrow beamwidth in the azimuth plane, hence, in this state the antenna should necessarily be rotated by 90° to accomplish the sampling process over a single zero-height ring enclosing the antenna. In the vast majority of cases, it is impossible to rotate the antenna because of mechanical limitations. The SSM can be effectively used to overcome this drawback. To implement the SCM, it is enough to apply a one-dimensional fast Fourier transform (1D-FFT) in the Near-field-to-far-field transformation process. Fig. 18 shows two possible SCM scenarios developed by the CNF system and the SSM. By sampling the near-field data along the z-axis, the pattern will be calculated in the elevation plane. Similarly, sampling over the ring enclosing the antenna leads to calculating the pattern in the azimuth plane.

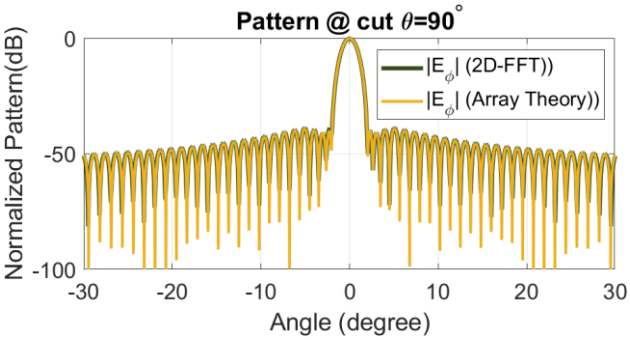


(a)

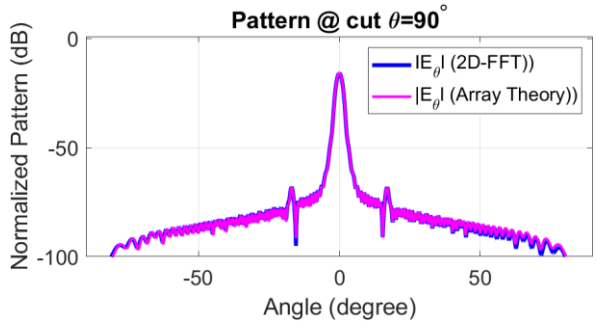


(b)

Fig. 18: Two scenarios of the SCM developed by CNF and the SSM.

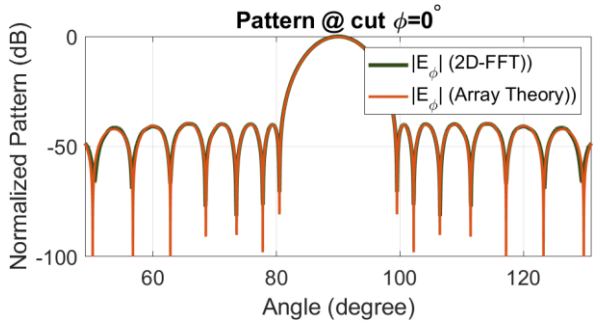


(a)

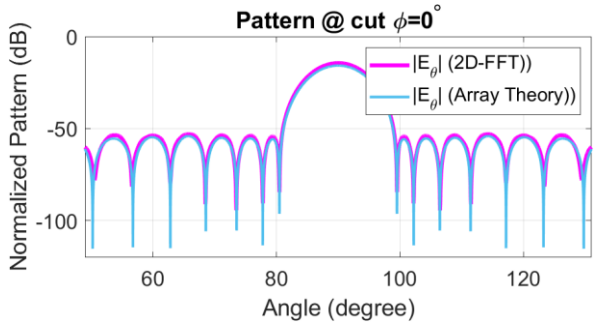


(b)

Fig. 19: The far-field pattern resulted from the SCM by scanning the near-field data over the straight line shown in Fig. 18(a).



(a)



(b)

Fig. 20: Far-field pattern resulted from the SSM by scanning the near-field data over the ring shown in Fig. 18 (b).

The far-field results are presented in Fig. 19 and Fig. 20. It can be observed that if the near-field data are recorded over the zero-height ring, the azimuth pattern can effectively be estimated by the CNF and SSM. In addition, if the near-field data are recorded along the z-axis, the elevation pattern will be estimated with acceptable accuracy.

Comparison and Discission

To complete the discussion, a comparison is accomplished between the CNF measurement system developed in this work and the previously reported works which is presented in Table 1. As can be seen, the authors in [13] tested a very large L-band radar antenna by a CNF system. The AUT in [13] has a sum pattern (with half-power beamwidth less than 2°) in the horizontal cut and a difference pattern in the elevation cut. To measure the sum pattern in the azimuth plane and characterize the main beam direction with the acceptable error, 512 samples are selected in the φ-direction. Apparently, if the deference pattern was in the azimuth plane, much more points would have to be taken. in the azimuth plane in [13]. Two other works listed in Table 1, that is [21] and [22], are related to the measurement of antennas with a wide-angle beam in the azimuth plane. As shown in Table 1 , the sampling steps in [21] and [22] are selected to be 6° and 2.5°, respectively which are much larger than the sampling step used in [13] because of wider beamwidth.

A closer examination of the achievements presented in [22] verifies the fact that higher-order Henkel functions are required for data acquisition if the azimuth beamwidth is small. This ultimately leads to smaller sampling steps. The authors in [22] measured a horn antenna wide wide-angle beam as the first practical case with CNF system by applying the Hankel function with the order up to 44 leading to $\Delta\phi$ = 4°. In the second case, a dish antenna was utilized as the AUT and the azimuth pattern was measured by applying the Hankel function with the order up to 72 leading to $\Delta\phi$ =2.5° as depicted in Table 1.

Table 1: The comparison between the results proposed in this work and the previously published works. H: height, W: width, D: depth. N.R stands for not reported.

Parameter	[13]	[21]	[22]	This work
AUT	L-Band Radar ANT	2- port ANT	Dish ANT	Slot array
Antenna dimension (mm)	large (N.R)	H:2254 W: 259 D: 99	Diameter = 0.34 cm	H: 360 W: 1720
Total Scan Length	7.5 m	2.7 m	990 mm	2 m
Distance probe to AUT	5m~7m	90 cm	20 cm, 40 cm	115 cm
Sampling step	0.7°	6°	2.5°	<0.2°
Sampling Length	10 cm	10 cm	15 mm	10 mm
HPBW in the azimuth	N. R	65°	~10°	1.4°
Sidelobe region estimation capability in the azimuth plane	Very good	weak	Relatively good up to ±20°	excellent
Estimation error of the beam direction in the azimuth plane	≈ 0.1°	Up to 1.19°	N. R	<0.02°

Another comparison is done for SCM which is given in Table 2. It can be seen that a pattern with a half-power beamwidth as small as 1.4° can be measured by the proposed method with excellent compatibility with the theoretical results. This is done by using the small sampling step, 0.2°, thanks to the use of the SSM. Meanwhile, the measurements developed in [20], [23] used the sampling steps equal to 1° to characterize the radiation patterns with broader beamwidths in the azimuth plane. Although, our investigations show that almost SCM systems are used to characterize sum patterns and therefore, proper evaluation and comparison (such as angle tracking error) cannot be done between the presented work and similar ones reported in the literature.

Table 2: The comparison between the results proposed in this work for SCM and the previously published works. H: height, W: width, D: depth. N.R stands for not reported.

Parameter	[20]	[23]	This work
AUT	LPDA	Sector ANT	Slot array
Antenna dimension (mm)	N. R	H: 834 mm	H: 360 W: 1720
Distance probe to AUT	91cm	10 cm	115 cm
Sampling step	1°	1°	<0.2°
HPBW in the azimuth	N. R	>50°	1.4°
Sidelobe region estimation capability in the azimuth plane	weak	Relatively good up to $\pm 60°$	excellent
Estimation error of the beam direction in the azimuth plane	N. R	N. R	<0.02°

Conclusion

In this paper, a novel method named the SSM is introduced to circumvent the serious limitation of the CNF measurement system and characterize the antenna pattern with narrow beamwidth in the azimuth plane. The presented method is based on artfully selecting the φ angle sequences and calculating the associated patterns in a repetitive routine. To verify the SSM, a comprehensive analytical model based on coordinate transformation with Euler angles is presented, which is useful to characterize the radiation pattern of every array antenna consisting of slot or microstrip patch elements.

Author Contributions

This paper is fully prepared by Majid Karimipour, assistant professor at the Arak University of Technology. The primitive idea, mathematical calculations, Matlab code, and simulations are all accomplished by Majid Karimipour.

Acknowledgment

The author would like to thank the editor and anonymous reviewers.

Conflict of Interest

The author declares that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancy have been completely observed by the authors.

Abbreviations

SSM	Sequential Sampling Method
Co-Pol	Co Polarization
Cross-Pol	Cross Polarization
CNF	Cylindrical Near-Field
FFT	Fast Fourier Transform
CMC	Cylindrical Mode Coefficient
2D	Two Dimensional
3D	Three Dimensional
AUT	Antenna Under Test
SCM	Single-Cut Measurement

Reference

[1] C. Parirni, S. Gregson, G. MacCormick, D. J. Van Rensburg, T. Eibert, Theory and Practice of Modern Antenna Range Measurements, v2, London: IET, 2020.

[2] A. D. Yaghjian, "An overview of near-field antenna measurements", IEEE Trans. Antennas Propag., 34(1): 30-45, 1986.

[3] A. Taaghol, T. K. Sarkar, "Near-field to near/far-field transformation for arbitrary near-field geometry, utilizing an equivalent magnetic current," IEEE Trans. Electromagn. Compat., 38(3): 536–542, 1996.

[4] M. Salucci, M. D. Migliore, P. Rocca, A. Polo, A. Massa, " Reliable antenna measurements in a near-field cylindrical setup with a sparsity promoting approach," IEEE Trans. Antennas Propag., 68(5): 4143-4148, 2020.

[5] J. S. Row, J. F. Tsai, "Cylindrical near-field antenna measurement using a polarization reconfigurable probe," Microwave Opt. Technol. Lett., 58(11): 2707–2711, 2016.

[6] F. R. Varela, M. J. L. Morales, R. T. Sanchez, A. T. M. Barrado, E. d. l. F. González, G. P. Quijano, C. Z. Torres, M. S. Pérez, M. S. Castañer, "Multi-probe measurement multi-probe measurement system based on single-cut transformation for fast testing of linear arrays," Sensors, 1744(21): 1-14, 2021.

[7] C. Apriono, Nofrizal, M. D. Firmansyah, F. Y. Zulkifli, E. T. Rahardjo, "Near-field to far-field transformation of cylindrical scanning antenna measurement using two dimension fast-fourier transform," Int. Conf. Qual. Res., 368-371, 2017.

[8] M. Farouq, M. Serhir, D. Picard, "Matrix method for far-field calculation using irregular near-field samples for cylindrical and spherical scanning surfaces," Prog. Electromagn. Res. Lett B., 63: 35–48, 2015.

[9] C. H, Schmidt, T. F. Eibert, "Assessment of irregular sampling near-field far-field transformation employing plane-wave field representation," IEEE Antennas Propag. Mag., 53(3): 213–219, 2011.

[10] O. M. Bucci, C. Gennarelli, C. Savarese, "Interpolation of electromagnetic radiated fields over a plane by nonuniform samples," IEEE Trans. Antennas Propag., 41(11): 1501–1508, 1993.

[11] A. Capozzoli, C. Curcio, A. Liseno, "Optimized near field antenna measurements in the cylindrical geometry," presented at the European Conf on Antennas. Propag, EUCAP., Lisbon, Portugal, 2015.

[12] https://www.nsi-mi.com/products/system-solutions/near-field-systems.

[13] S. Burgos, F. Martin, J. L. Besada," Cylindrical near-to-far-field transformation system for radar antennas: design, validation, and application," Microw. Opt. Technol. Lett., 50(10): 2527–2531, 2008.

[14] L. Jossefson, S. R. Rengarajan, Slotted Waveguide Array Antennas: Theory, analysis and design, vol. 1, London: Wiley, 2018.

[15] C. Balanis, Antenna theory: analysis and design, Hoboken, N.J.: Wiley-Interscience, 2016.

[16] Y. Rahmat-Samii, "Useful coordinate transformations for antenna applications." IEEE Trans. Antennas and Propag., (27)4, 571-574, 1979.

[17] B. Fuchs, L. L, Coq, M. D. Migliore, "On the interpolation of electromagnetic near field without prior knowledge of the radiating source," IEEE Trans. Antennas Propag., 65(7): 3568–3574, 2017.

[18] M. D. Migliore, "Near field antenna measurement sampling strategies: From linear to nonlinear interpolation," Electronics, 7(10): 1-21, 2018.

[19] Y. Hayashi, H. Arai "A reduction of measurement points in cylindrical near field measurement by complex interpolation," Int. Symp. on Antennas and Propag., Osaka, Japan, 2021.

[20] S. Omi, T. Uno, T. Arima, "Single-Cut near-field far-field transformation technique employing two-dimensional plane-wave expansion," IEEE Antennas Wirel. Propag. Lett., 17(8): 1538-1541, 2018.

[21] K. Phaebua, T. Lertwiriyaprapa, D. Torrungrueng, "Cylindrical near-field to far-field radiation pattern measurement system for a large mobile phone base station antenna," presented at the 2021 Int. Electr. Eng. Congr.,:10-12, Pattaya, THAILAND, 2021.

[22] J. Puskely, "Application of iterative fourier method in cylindrical phaseless antenna measurement technique," Radioengineering, 21(1): 422-429, 2012.

[23] Y. Sugimoto, H. Arai, T. Maruyama, M. Nasuno, M. Hirose, S. Kurokawa, "Fast far-field estimation method by compact single cut near-field measurements for electrically long antenna array," IEEE Trans. Antennas Propag., 66(11): 5859-5868, 2018.

[24] M. Sierra-Castañer, "Review of recent advances and future challenges in antenna measurement," Appl. Comput. Electromagn. Soc. J., 33(1): 99-102, 2018.

[25] R. Cornelius, T. Salmerón-Ruiz, F. Saccardi, L. Foged, D. Heberling, M. Sierra-Castañer, "A comparison of different methods for fast single-cut near-to-far-field transformation," IEEE Antennas Propag Mag., 56(2): 252-261, 2014.

[26] X. Li, G. Wei, L. Yang, B. Liao, "Fast determination of single-cut far-field pattern of base station antenna at a quasi-far-field distance," IEEE Trans. Antennas and Propag., 68(5): 3989-3996, 2020.

[27] F. R. Varela, R. T. Sánchez, M.J.L. Morales, M. Sierra-Castañer, "Near-field to far-field transformation for fast linear slide measurements" presented at the 14th Eur. Conf. antennas Propag. EuCAP Copenhagen, Denmark, 2020.

[28] L. J. Foged, G. Barone, F. Saccardi, "Antenna measurement systems using multi-probe technology," presented at the IEEE Conf. Antenna Meas. Appl. Chiang Mai, Thailand, 2015.

[29] F. Las-Heras, B. Galocha, J. L. Besada, "Far-field performance of linear antennas determined from near-field data," IEEE Trans. Antennas Propag., 50(3): 408-410, 2002.

[30] M. Orefice, M. A. Razzaq, G. Dassano, "Sidelobe level correction for parabolic antennas radiation pattern measurements in quasi-far-field conditions", Electron. Lett., 49(23): 1423-1425, 2013.

Biographies

Majid Karimipour received his B. Sc., M.Sc., and Ph.D. degrees all in Electrical Engineering in 2010, 2013, and 2018, respectively. The title of his Ph.D. thesis was " The application of the holographic technique in the synthesis of antenna radiation pattern" which was defened in the Iran University of Technology (IUST) with excellent grades. From 2014 to 2018 he worked with the communication satellite group, Iran Telecommunication Research Center, (ITRC) in Tehran, Iran, as a major researcher. He is currently an assistant professor with the Arak University of Technology. He has high experience in antenna near-field measurement and calibration systems, especially planar and cylindrical scanning systems. He has taught, communication systems, antenna theory, signal and system, and electromagnetic courses since 2015. His major research interests are the development and design of reflect array and transmit array antennas, phased array antennas, pattern synthesis, and near-field antenna pattern measurement systems.

- Email: m.karimipour@arakut.ac.ir
- ORCID: 0000-0001-7612-2512
- Web of Science Researcher ID: CAG-4398-2022
- Scopus Author ID: 55962138900
- Homepage: https://scholar.google.com/citations?user=3rZQKjAAAAAJ&hl=en

How to cite this paper:
M. Karimipour, "Pattern measurement of large antenna by sequential sampling method in cylindrical near-field test," J. Electr. Comput. Eng. Innovations, 11(1): 103-118, 2023.

DOI: 10.22061/JECEI.2022.8705.544

URL: https://jecei.sru.ac.ir/article_1733.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Innovative Paper

Indicators for Determining Salt Harvest Time Based on Salinity and Liquid Viscosity Using Microcontroller

A. Saleh*, A. S. Arifin

Department of Electrical Engineering, Politeknik Elektronika Negeri Surabaya, Surabaya, Indonesia.

Article Info

Article History:
Received 17 April 2022
Reviewed 25 April 2022
Revised 14 June 2022
Accepted 27 June 2022

Keywords:
Monitoring
Salinity
Conductivity
Database
Web

*Corresponding Author's Email Address: *akuwan@pens.ac.id*

Abstract

Background and Objectives: In general, traditional salt farmers determine the time to harvest salt by visiting and monitoring their salt ponds. Therefore, to assist salt farmers in determining the right time to harvest salt and determine the quality of the harvested salt, a wireless-based electronic device is needed that can monitor the salt content and viscosity of the brine.
Methods: An electronic device that is made to measure salt content (salinity) with a conductivity sensor and to measure fluid viscosity using a data processing method from sensor readings which is first converted to digital data with a program on the microcontroller. To find out whether the brine is ready to be harvested or not, the data obtained in the form of conductivity and stress are converted into percentages of NaCl and degrees of Baume. Then the data is sent to the ESP8266 Wifi module to be stored in a database and displayed on the Web.
Results: The results of the data obtained are based on testing in salt ponds for young water but it has been quite a long time the results have approached old water of around 64% and 14o Be. The results of the old water test that had just been moved to the last reservoir were close to harvest time of around 94% and 21oBe. If it has reached 25o Be then it is enough to be moved to the crystallization site. To determine the harvest period based on two parameters, namely the salt content and the viscosity of the liquid is 86-90% and the viscosity of the liquid is 20-24o Be. If you have reached both of these parameters, the salt can be harvested in about 7-10 days to make the water crystallize.
Conclusion: Equipment Indicators for determining salt harvest time based on salinity and liquid viscosity using a microcontroller that has been made have been successfully used to determine salt harvest time properly. The salt quality of this indicator tool is the salt content including the K-3 quality or the lowest quality of the 3 existing qualities.

Introduction

Advances in technology encourage humans to create equipment that can help humans simplify their work, making them more efficient and practical. In the industrial world, the process of making salt is produced using two methods of evaporation by sunlight in salt ponds and by boiling techniques. Salt is a very important commodity for people's lives, salt can not only be used as a consumption material but salt can also be categorized in industrial materials. This is stated in the Regulation of the Minister of Industry of the Republic of Indonesia Number: 88/M-IND/PER/10/2014 on the scope of salt it is stated that salt is the production of the Chlor Alkali chemical industry

group which consists of consumption salt and industrial salt [1]. Based on the quality of salt production requirements, there are several methods of purifying in order to fulfilled the requirements, included chemical and physical method. The purification process of salt is done using a washing technique and evaporation (recrystallization). Washing technique is done with a nearly saturated brine while evaporation (recrystallization) process [25]. Salt is widely used in various products and it is estimated that around 14,000 products use salt as an additive [2]. This need has not been followed by an adequate quantity and quality of national salt production. In general, people's salt is grouped into three types [4], namely:

1. K-1 is the best quality that meets the requirements for industrial and consumption materials with the following composition: NaCl : 97.46 %, CaCl2 : 0.723 %, CaSO4 : 0.409 %, MgSO4 : 0.04%, H2O : 0.63%, and Impurities : 0.65%.
2. K-2 is a quality below K-1, this type of salt must be reduced in levels of various substances in order to meet the standards as industrial raw materials. This salt content ranges from 90-94%.
3. K-3 is the lowest quality salt for people's production. Usually the levels are between 88-90%, sometimes mixed with soil, so the color is slightly brownish.

Currently, the controlling, monitoring, and reporting systems have developed rapidly and are wireless-based. Usually this wireless is used in an area or location where the user is always on the move, or at that location there is no wired network for data distribution [15]. In addition, wireless communication is experiencing a fairly rapid development and is widely used as an interface on electronic devices or computers as a means of remote control [17]. Based on these developments, the world of agriculture must also keep up with the latest technological developments. However, to have a monitoring system requires a large cost. In addition, salt farmers always have to go to the pond and heat up to see the condition of the salt water. The impact of the production system using the evaporation method with sunlight, many have experienced crop failures. Because the manufacture of salt by the method of evaporation of sea water by utilizing sunlight energy is influenced by several factors including the rate of evaporation is related to the amount of salt obtained and seawater concentration, related to the amount of dissolved salt [3]. In addition, the tools used to determine the salt harvest time are expensive and are still offline. Therefore, it is necessary to have an online (wireless) tool to determine the salt harvest time. Made.

Making a tool to determine the salt harvest time by monitoring the salt content and the thickness of the brine. There are two parameters that can be used to measure whether the salt is ready to be harvested or to the next process. Brine must have a NaCl content above 90% to 97% for iodized salt. As explained in the Regulation of the Minister of Industry of the Republic of Indonesia, household salt is iodized consumption salt with a minimum NaCl content of 94% [1]. Until December 2019 there were 39 SNIs related to salt [10], the Indonesian National Standard SNI 3556:2016 stipulates a minimum requirement of 94% sodium chloride (NaCl) and a minimum of 30 mg/kg of iodine as KIO_3. And the viscosity or concentration of salt water reaches 24° Be to 29° Be unit degrees Baume, if it is more then it tastes bitter (contains MgCl) and if it is less then it precipitates gypsum, calcium carbonate so that the salt becomes brittle and opaque. Therefore, it is necessary to conduct a study in order to identify the causes of crop failure, determine the amount of salt (NaCl) and the viscosity of the brine. NaCl easily obtained by evaporation of seawater. The making salt from seawater evaporation carried out the public, in general, is still conventionally to produce salt with low quality [22]. Salt production depends mainly on high sunlight intensity and temperature and low relative humidity, thus is favored mainly by summer months with high temperatures [30].

Material and Methods

In the design of this system, a monitoring system for salt levels and fluid viscosity is described in salt water. Seawater has an average salt content of 3.5%. Seawater also has varying salt content. To measure the salt content, a conductivity sensor is used, the sensors conductivity/TDS/salt levels have a compact design [24]. Input data in the form of sensors mounted on the microcontroller. In this circuit, the sensors used are the conductivity sensor and the baume meter. The sensor reading data will be processed and converted into digital data with the program in the microcontroller. The data obtained in the form of conductivity (µS) and voltage (V), will be converted into percentage of NaCl and degrees of Baume, respectively. Sensor data and output obtained will be forwarded by the ESP to the access point for real-time pond monitoring. Access point as a connecting device serves as a bridge between wired and wireless communication [16]. In this paper, the interface between the ESP8266 Wi-Fi module and the Arduino MCU is studied for system monitoring applications. As shown in Fig. 1. Fig. 1 shows that in the design of the system for making this indicator tool. In general, the description of the system is that several sensors function to capture sensor signals in the room. In this circuit, the sensors used are the conductivity sensor and the baume meter. The PC server will be connected to the Wifi Access Point and the ESP module to connect the internet with the microcontroller.

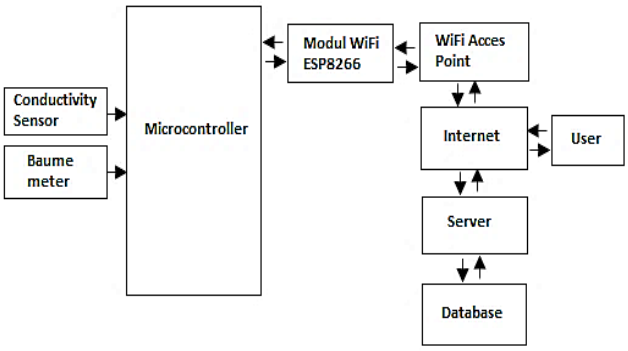


Fig. 1: System block diagram.

The ESP8266 Wi-Fi module is a self-contained system-on-chip (SOC) with integrated TCP/IP protocol stacks that can give any microcontroller access to a Wi-Fi network [13]. ESP8266 is a WiFi module that offers a complete and standalone Wi-Fi networking solution, enabling it to host applications or offload all Wi-Fi network functions from other application processors [21]. Parameter data obtained from the sensor will be stored in the database and can be accessed by the user. Systematics of making the system, adapted to the design that has been determined at the design stage.

Table 1: Salt condition parameters.

Condition	Parameters		
	Viscosity (°Be)	Salinity (%)	Description
Quality 1	24 - 29	<97	is the result of the crystallization process in solution
Quality 2	29 - 35	<94	is the remaining crystallization above in the solubility condition
Quality 3	>35	<90	is the remainder of the above concentrated solution at condition

The determination of the NaCl content was carried out using the Indonesian National Standard (SNI 01-3556-2016) method about consumption of iodized salt [5], the next calculation is as follows:

$$NaCl\ level = \frac{(V\ x\ N\ x\ fp\ x\ 58.5)}{W} x\ 100\% \quad (1)$$

where, V is the volume of $AgNO_3$ required in the titration (ml), N is the normality of $AgNO_3$, fp is the diluent factor, 58.5 is the molecular weight of NaCl and W is the weight of the test sample (mg).

In Fig. 2 is a system flowchart that contains an overview of the system starting from the microcontroller process until the data can be accessed via web hosting. The web is an information system technology that connects data from many sources and various services on the internet [18]. The working principle of the system from Fig. 2 is as follows: The program runs according to the sketch programmed into the microcontroller starting by setting the SSID, password and server address and setting the I/O port for sensors and libraries used by sensors. The microcontroller makes a connection between the ESP and the router according to the SSID that has been set in the sketch. In order for sensors to enter the hosting. The microcontroller reads data from the sensor via the I/O pins. The process of determining the parameters will process further I/O according to the program that has been sketched on the microcontroller.

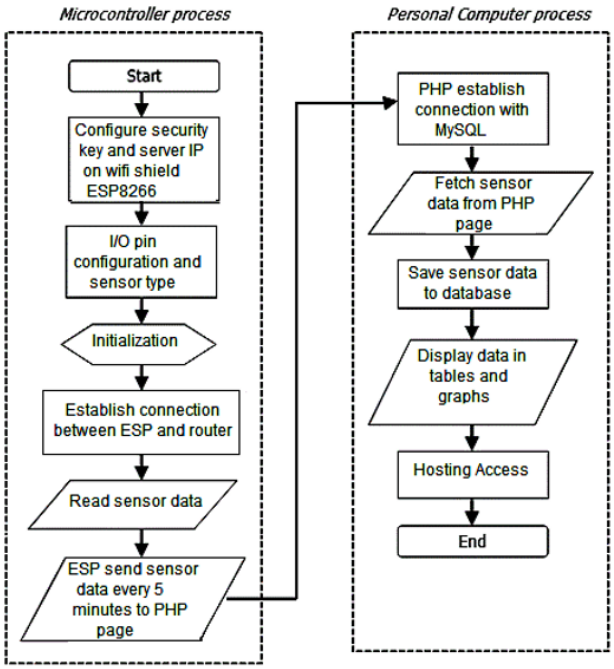


Fig. 2: System flowchart.

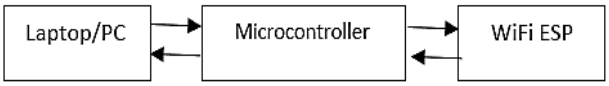
The ESP sends data from the microcontroller processing every few minutes to the server. PHP establishes a connection between Apache and MySQL. The dynamic web page captures the data sent by the ESP which is then stored in the database. Process average sensor on PHP data page. PHP retrieves sensor data from the database and then displays it in a line chart. Then accessing parameter results via hosting.

A. The Microcontroller Interface with The ESP8266 and The Conductivity Sensor and Baume

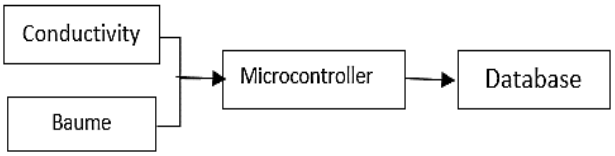
The equipment needed in the hardware design is an Arduino uno R3 microcontroller, ESP8266, a conductivity sensor, and a Baume meter. The block diagram of the microcontroller interface with the ESP8266 WiFi module, conductivity sensor and baume is shown in Fig. 3.

This sensor functions as a salt level or salinity reader. The sensor works by measuring the concentration of ions to conduct an electric current between two electrodes. As

shown in Fig. 3 above, by connecting 5V, output, and ground to the Arduino, you can then see the sensor readings directly on the serial monitor. The results of the data obtained by this conductivity sensor are the Conductivity and Total Dissolved Solids (TDS). Then converted into a concentration of salt content (NaCl%). The value of salt content will be greater when measured in old water when compared to young water and fresh water. Later the data will be sent to the database using the ESP8266 module.



(a) Microcontroller with ESP8266.



(b) Microcontroller with conductivity sensor and baume.

Fig. 3: Microcontroller interface with ESP8266 and sensor.

Salinity is the level of saltiness or the level of salt dissolved in water. These dissolved salts include sodium chloride, magnesium sulphates, potassium nitrates, and sodium bicarbonate [11]. Salinity is used also to trace seawater masses and to model ocean dynamics [29]. That is the number of grams of salt dissolved for each liter of solution. Usually expressed in units of ‰ (parts per thousand). Parts per thousand is g/kg for liquids and for solids is mL/L for gas mixtures. Therefore, a 1000 gram seawater sample containing 35 grams of dissolved compounds has a salinity of 35‰. The following equation is for a 35‰ salinity solution containing 35 grams of salt per 1000 grams of saltwater.

$$35‰ = \frac{35\ gram\ salt}{1000\ grams\ saltwater} \quad (2)$$

According to the classification of high and low salinity, salinity is divided into three parts, namely fresh water, brackish water and sea water. The higher the concentration of a solution, the higher the absorption capacity of the salt to absorb water. Salinity also affects the osmotic pressure of water. The higher the salinity in a water, the greater the osmotic pressure.

The Electrical Conductivity (EC) was calculated to TDS because the conductivity measurement is measured by using probe dipped into the water to measure the dissolved charge which corresponds to the analysis [8]. Measurement of salinity is related to chlorinity. This chlorine includes chloride, bromide and iodide.

$$Salinity = \left(\frac{TDS}{10}\right) + \left(\frac{Conductivity}{100}\right) + 27.1024 \quad (3)$$

The measurement of the total salt concentration of the aqueous extracts of soil samples can be done either directly through chemical analysis of the chemical constituents that constitute the soil salinity (or mass of the TDS) or indirectly through the measurement of the EC [9]. The correlations between TDS and EC depend on the type of ions and their concentrations in the aqueous solution. Therefore, a specific correlation should be generated for each type of brine. The use of equations developed for model brines may carry substantial error, particularly when analysing high salinity brines [6]. In general [8], the TDS – EC relationship is given by (4).

$$TDS = (0.55\ to\ 0.7)\ EC \quad (4)$$

The TDS in water samples is estimated by multiplying EC by an empirical factor. This factor may vary from 0.55 to 0.90 depending upon the nature of soluble ionic components, their concentration and the temperature of water. Conductivity or electrical conductivity (EC) and total dissolved solids (TDS) are frequently used as water quality parameters, especially in the coastal area [20].

B. The Baume Meter

Baume meter is a tool used to read viscosity. The trick is to modify it in such a way as to get the desired result. Measuring the Density of Liquids expressed in degrees Baume (Be). It is important to measure the density of liquids. Density is an important characteristic possessed by a substance [14]. There are 2 types that are used Bé Heavy for liquids that are heavier than water and Light Be for liquids that are lighter than water. For Bé a weight of 0 degrees is equal to a solution having a relative density of 1.842. As for light Bé, 0 degrees Bé is equal to the density of a 10% NaCl solution and 60 degrees Bé is the same as a solution having a relative density of 0.745.

The baume meter is mounted with copper wound around 25°Be. Then next to the paralon pipe a voltage reading indicator is installed. The blue wire is connected to the 5V Arduino and the green wire to the Arduino A1. So that the Arduino can detect the movement of the copper in the baume meter. Later the voltage will be converted to Baume degrees then the data is sent to the database. Fig. 4 shows the design of the baume meter that has been inserted into the paralon pipe.

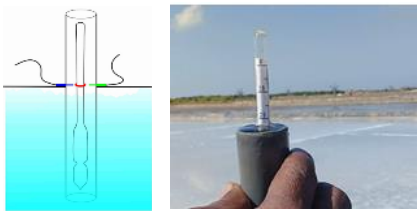


Fig. 4: Baume meter design.

Fig. 5 shows the indicator equipment for determining the salt harvest time that has been made.

Fig. 5: Salt harvest indicator equipment.

C. Web Server

Number Software realization by building a web server using XAMPP. In XAMPP there are Apache and MySQL which will be a local web server system. Web monitoring display as shown in Fig. 6.

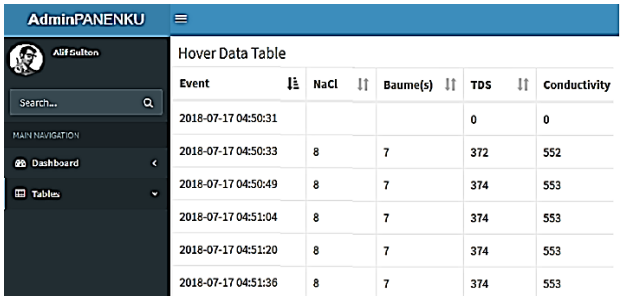


Fig. 6: Host's web view.

By using the internet or an online monitoring system, the monitoring process can be done anytime and anywhere and the data can be known more quickly and accurately [18]. Process monitoring on server is done at real time. Another monitoring service is to display the process database with immediate values [19].

D. Plans and Plots of Salt Ponds

In general, salt production models in Indonesia use evaporation, which is evaporation of seawater in shallow ponds by considering the thickness of seawater in these ponds [26].

Indonesia has a long coastline, potentially for salt pond. Salt Pond is an artificial shallow pond designed to produce salt from sea water or salt water [27]. The salt pond used in the research is located in Pandan Village, Galis District, Pamekasan Regency, Madura, East Java, Indonesia. Shown in Fig. 7.

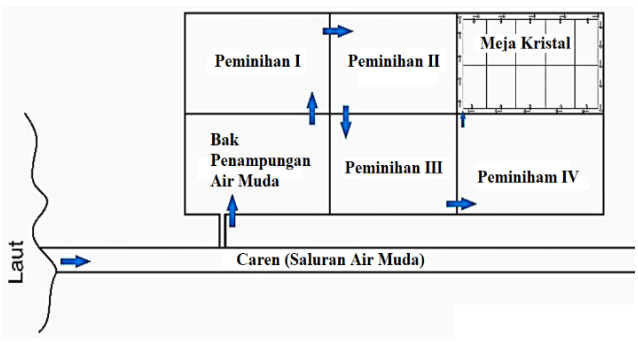


Fig. 7: Salt pond floor plan.

The process of making salt is carried out in the dry season, where the evaporation area (peminihan) is drained by seawater using a pump as shown in Fig. 7. In general, it consists of 6 plots of ponds, including Young water reservoir, Peminihan pool 1, Peminihan pool 2, Peminihan pool 3, Peminihan pool 4, Peminihan pool 5 and Crystal table or crystallization table.

While the map model and its size are shown in Fig. 8.

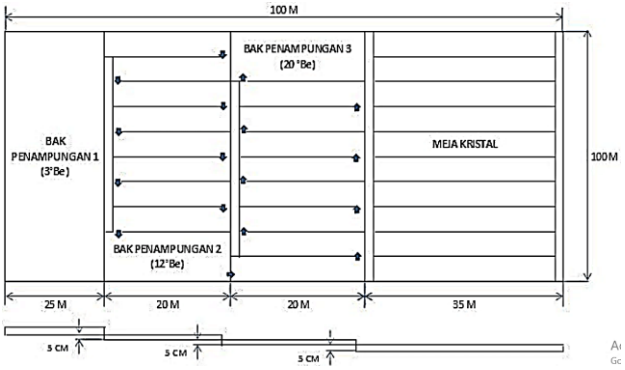


Fig. 8: Map and size the salt pond.

In this land seawater is evaporated so that it becomes old water. The old water flows to the crystallization table where later the salt will crystallize. In salt production process, besides producing salt also produced the liquid remaining crystallization called bittern [27]. The quality of the salt is controlled by removing or separating the bittern, which only crystallizes the salt at a concentration of 25^{O} to 30^{O} Be [7]. The harvested salt crystals are transported and taken to storage warehouses. The process can be continued by washing or can be directly sold as crushed salt. The resulting salt is in the form of white crystals which in addition to containing NaCl also contains other salts which are impurities.

Sea water (called young water) that flows into the salt pond will be accommodated in the young water reservoir, in the pool it is expected that as much sea water is accommodated depending on the area of land owned. In the holding pool the water depth is at least 1 meter, because the reservoir is a stock or supply of young water during the salt making process. In the pool the seawater that is accommodated is allowed to stand for a minimum of ten days, where as long as the young water is stored there will be deposition of impurities that are not needed

during the crystallization process. The problems found in this process are low quality of raw sea water, open transport of concentrated sea water into crystallization pond, high penetration of pre-crystallized water into the soil in crystallization pond and batch crystallization process [12]. In addition to this, there will also be a process of increasing the density of seawater, namely when seawater is flowed into a young water reservoir, the density is only 3° Be, after being stored for approximately ten days there will be an increase in the density of the seawater to 7° Be to 10° Be. This is because of the influence of the sun's heat, wind and geothermal heat.

After the young water circulates from the young water reservoir to the purification pond V, the brine solution will increase its density by approximately 20° Be to 22° Be (called old water), this happens because of the evaporation process during which the brine solution is transferred. and the concentration of the brine solution must really be reached in the purification pond V, and if the concentration is not sufficient, the brine solution should be held for a few days so that the concentration reaches 20° Be to 22° Be, after the concentration of the brine solution is sufficient, then old water was released onto the Crystal table [3].

Results and Discussion

Testing of the indicator equipment to determine the time of salt harvest is carried out by placing it on the salt pond to several points that are easily accessible. The types of water tested are young water and old water, where young water is water in the channel originating from sea water. While old water is water that has undergone several processes and has been moved from 1-4 purification which is then put into old water reservoirs. Old water is the last water before it is put into the crystal table and becomes salt.

A. Testing Process on Salt Pond

Steps for testing equipment on salt ponds. The first is the placement of the equipment in the planned place, to find out whether the sensor data is appropriate or not. In this test, the reading data from the sensor is displayed on the LCD. As shown in Fig. 9.

Fig. 9: Laying equipment on salt pond

Then from the sensor it is displayed in the database, by removing the LCD it is replaced with a configured ESP8266 module.

A. 1. Conductivity Sensor Test

In this test, water is added with iodized and non-iodized salt. Each was tested 5 times.

Table 2: Conductivity test data with iodized salt

No.	Iodine salt (g)	Water (ml)	ADC sensor	Conductivity (µS)
1	4	100	431	588
2	12	100	453	589
3	20	100	466	593
4	28	100	482	598
5	36	100	496	601

The conductivity value is obtained by the following equation:

$$y = 0.2142x + 494.93 \quad (5)$$

where : x = ADC value, and y = conductivity

Table 3: Conductivity test data non-iodized salt

No.	Iodine salt (g)	Water (ml)	ADC sensor	Conductivity (µS)
1	50	100	456	588
2	100	100	498	589
3	150	100	518	593
4	200	100	537	598
5	259	100	561	601

To get the percentage of salt content in the water, the results from the conductivity sensor are used, namely TDS and Conductivity. The equation is as follows:

$$y = 0.3417x + 281.08 \quad (6)$$

where : x = ADC value, and y = TDS

A. 2. Baume Sensor Testing

Tests were carried out on old water, young water, and fresh water.

To verify the long-term stability of the standard seawater composition, it was proposed to perform measurements of the standard seawater density. Since the density is sensitive to all salt components, a density measurement can detect any change in the composition [28].

As well as knowing what the density of each degree Baume.

Table 4: Baume relationship data with density

Baume (°Be)	Density
60	0.745
55	0.683
50	0.621
45	0.559
40	0.497
35	0.435
30	0.373
25	0.310
20	0.248
15	0.186
10	0.124
5	1.062

It is found that 0 degrees Bé is equal to the density of a 10% NaCl solution and 60 degrees Bé is the same as a solution that has a relative density of 0.745. So, it can be concluded that the relationship between °Be and density is x = 0.0124166667.

A. 3. Baume and Sensor Conductivity Testing

The results of the conductivity sensor and baume meter readings are as shown in Table 5. This test is carried out on young water that has been around for a long time. So, the results are close to old water around 86% and 20° Be. This water is taken and a sample is made to test the equipment for determining the salt harvest time.

Table 5: Young water test results data

Conductivity (µS)	Baume (°Be)	NaCl (%)	TDS
651.94	20	86.78	531.55
651.94	20	86.78	531.55
651.94	20	86.78	531.55
652.15	20	86.81	531.89

From the data table above, it shows that the salt is not ready to harvest because the NaCl value is less than 90% and the baume value 20° Be.

Table 6: Old water test results data

Conductivity (µS)	Baume (°Be)	NaCl (%)	TDS
656.40	24	90.52	537
656.08	24	90.48	529
656.30	24	90.52	520

From the data Table 6 above, it shows that the salt is ready to harvest because the NaCl value is more than 90% and the baume value 24° Be.

B. Test Results on Salt Ponds with Young and Old Water in Real Time

Data taken in the morning, afternoon and evening. In general, watering is done 2 times a day, namely in the afternoon and evening. Therefore, the test data was taken with 2 different water samples, namely young water and old water. For young water but it's been long enough. So, the results are close to old water around 64% and 14°Be. The test data with young water are as in Table 7. while the test data with old water are shown in Table 8

Table 7: The test data with young water

Id	Time	NaCl (%)	Baume (°Be)	TDS	Conductivity (µS)
142	2018-07-19 07:02:31	65	14	519	644
141	2018-07-19 07:02:15	64	15	521	645
140	2018-07-19 07:01:59	64	14	521	645
139	2018-07-19 07:01:44	64	14	520	645
138	2018-07-19 07:01:28	64	12	520	645
137	2018-07-19 07:01:12	64	14	520	645
136	2018-07-19 07:00:57	63	14	518	643
135	2018-07-19 07:00:41	66	14	520	645
134	2018-07-19 07:00:26	64	14	520	645
133	2018-07-19 07:00:10	62	13	520	645
132	2018-07-19 06:59:39	64	14	520	645
131	2018-07-19 06:59:23	61	14	518	643
130	2018-07-19 06:59:07	64	14	518	643
129	2018-07-19 06:58:52	62	15	518	643

This test is carried out for old water that has just been moved to the last reservoir. So, the results are close to harvest time of around 94% and 21° Be.

Table 8: The test data with old water

Id	Time	NaCl (%)	Baume (°Be)	TDS	Conductivity (µS)
127	2018-07-19 06:57:36	94	21	530	655
126	2018-07-19 06:57:21	92	19	530	655
125	2018-07-19 06:57:05	94	21	530	655
124	2018-07-19 06:56:49	92	24	537	653
123	2018-07-19 06:56:34	94	21	530	655
122	2018-07-19 06:56:18	93	20	530	655
121	2018-07-19 06:56:03	94	21	537	653
120	2018-07-19 06:55:47	95	22	537	653
119	2018-07-19 06:55:31	94	21	537	653

If it has reached 25° Be then it is enough to be moved to the crystallization site and salt can be harvested approximately 7-10 days to become crystals. As shown in Fig. 10. Salt harvest as shown in Fig. 11. Salt is a white crystalline solid which is the dominant group of compounds consisting of sodium chloride (> 80%) and other compounds such as magnesium chloride, magnesium sulfate, and calcium chloride [23].

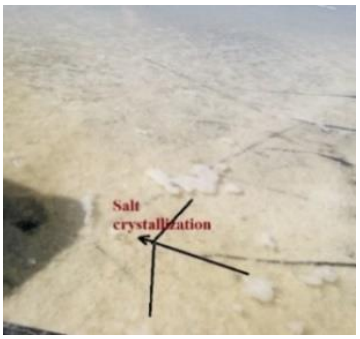


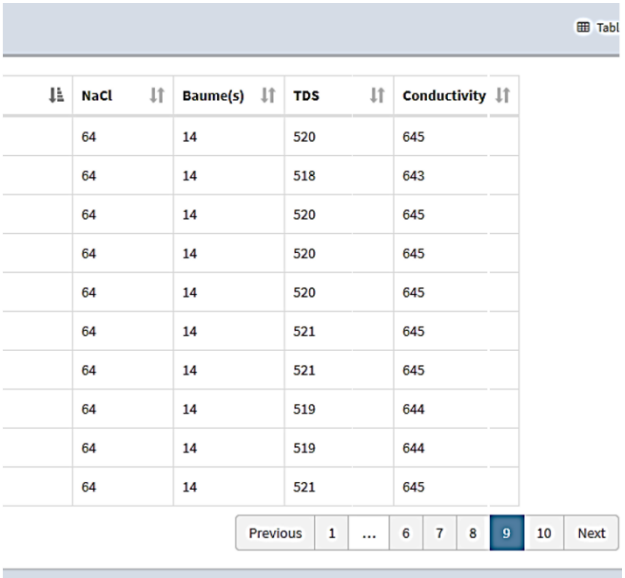
Fig. 10: Salt crystallization state.

Fig. 11: Salt harvest.

C. Test Display of Test Result Data on the Web

In this test, the microcontroller has been given an ESP8266 module that has been configured and connected to an access point, so it can be a server and connect to the localhost database. Then it has also been hosted on a website. The test results data on salt ponds are carried out in real time with a display on the web shown in Figs. 11 to 13 and a graphic display in Fig. 14.

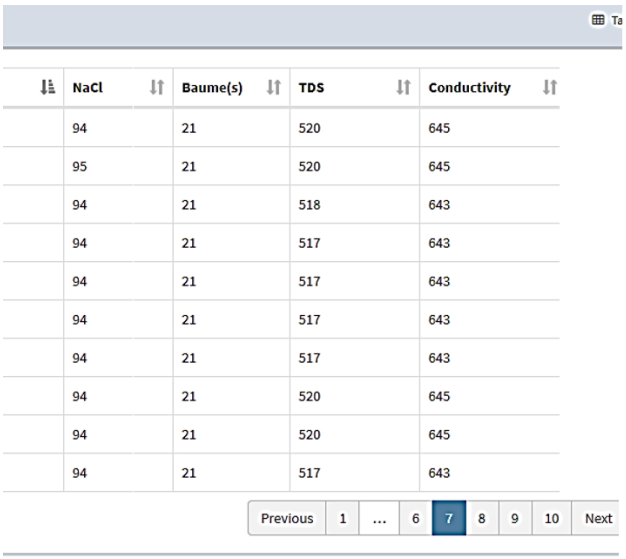
From this system, every data taken from the process is displayed on the screen as text, including date and time data, NaCl value, Baume value, TDS and conductivity value.



	NaCl	Baume(s)	TDS	Conductivity
	64	14	520	645
	64	14	518	643
	64	14	520	645
	64	14	520	645
	64	14	520	645
	64	14	521	645
	64	14	521	645
	64	14	519	644
	64	14	519	644
	64	14	521	645

Previous 1 ... 6 7 8 9 10 Next

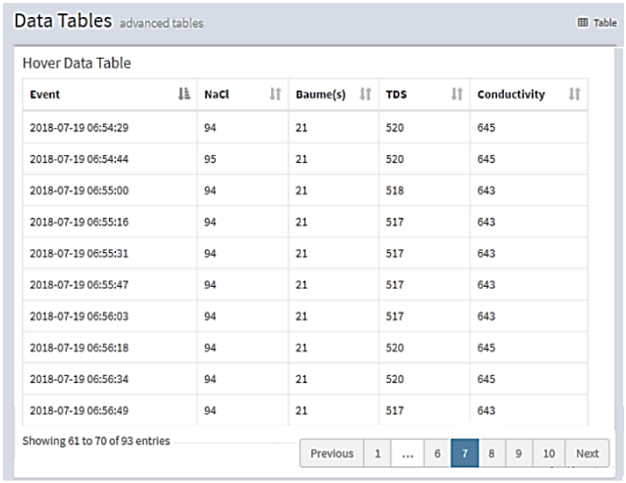
Fig. 11: Young water test results data on the web.



	NaCl	Baume(s)	TDS	Conductivity
	94	21	520	645
	95	21	520	645
	94	21	518	643
	94	21	517	643
	94	21	517	643
	94	21	517	643
	94	21	517	643
	94	21	520	645
	94	21	520	645
	94	21	517	643

Previous 1 ... 6 7 8 9 10 Next

Fig. 12: Old water test results data on the web.



Data Tables advanced tables

Hover Data Table

Event	NaCl	Baume(s)	TDS	Conductivity
2018-07-19 06:54:29	94	21	520	645
2018-07-19 06:54:44	95	21	520	645
2018-07-19 06:55:00	94	21	518	643
2018-07-19 06:55:16	94	21	517	643
2018-07-19 06:55:31	94	21	517	643
2018-07-19 06:55:47	94	21	517	643
2018-07-19 06:56:03	94	21	517	643
2018-07-19 06:56:18	94	21	520	645
2018-07-19 06:56:34	94	21	520	645
2018-07-19 06:56:49	94	21	517	643

Showing 61 to 70 of 93 entries

Previous 1 ... 6 7 8 9 10 Next

Fig. 13: Display the overall data of the test results on the web.

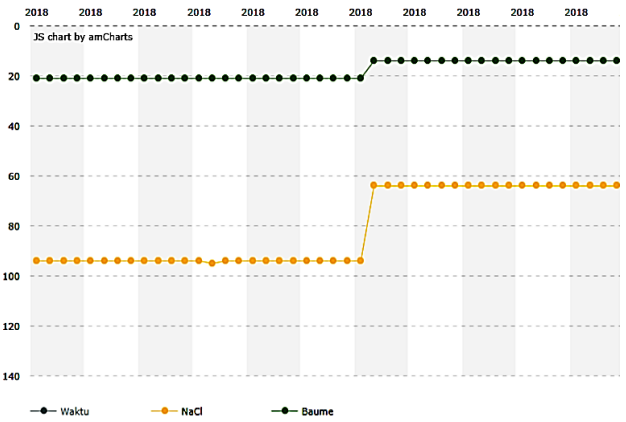


Fig. 14: Display of graphic data on the web.

In the graphic display there are 2 sensor parameters observed, namely NaCl and Baume. For a description of the parameter units, see the left of the graph and the time of data entry at the bottom of the graph.

Conclusion

From all the data obtained to determine the harvest period based on two parameters, namely salt content and liquid viscosity. Where the value of the salt content obtained is 86-90% and the fluid viscosity is 20-24° Be. If you have reached these two parameters, salt can be harvested in about 7-10 days until it becomes crystals. Parameter data access can be through localhost or hosting in real time.

Equipment The indicator for determining salt harvest time based on salinity and viscosity of the liquid using a microcontroller that has been made has been successfully used to determine the salt harvest time properly.

Salt quality is based on data that has been successfully retrieved by an indicator tool for determining the time of harvesting, its salt content includes K-3 quality or the lowest quality of the 3 existing qualities.

To further improve and perfect the equipment for determining the salt harvest time, it is necessary to use a mobile application to more easily monitor the condition of the salt pond water.

Author Contributions

The author's role in research participation is as follows:

Akuwan Saleh. Data analysis, interpreting the results and writing the manuscript, Alif Sultonul Arifin. designing experiments and collecting data.

Acknowledgment

The author would like to thank the salt farmers in Pandan Village, Galis District, Pamekasan Regency, Madura, East Java, Indonesia who have given permission to experiment and collect data in salt ponds. The research and writing of this journal manuscript were carried out independently without any financial assistance from any party.

Conflict of Interest

We declare that we do not have any conflict of interest. In this regard, I as the author fully disclose these interests to the Journal of Electrical and Computer Engineering Innovations (JECEI), and I have a plan to manage any potential conflicts that arise from the writing of the manuscript.

Abbreviations

Define abbreviations and acronyms in the text:

SNI	Indonesian National Standard
NaCl	Natrium Chloride / Sodium Chloride
CaCl	Calcium Chloride
MgCl	Magnesium Chloride
PHP	Hypertext Preprocessor
$AgNO_3$	Silver Nitrate
TDS	Total Dissolved Solids
EC	Electrical Conductivity
LCD	Liquid Cristal Display
SOC	Self-contained system-On-Chip
MCU	Microcontroller Units

References

[1] Regulation of the Minister of Industry of the Republic of Indonesia Number: 88/M-IND/PER/10/2014 concerning the Road Map for the Development of Salt Industry Clusters., Chapter-I, 1-3, 2014.

[2] B. Suwasono, A. Munazid., et al., "Strategic planning for capacity building production and salt farmer in region of surabaya city east java Indonesian," ASRJETS J., 12(1:) 53-65, 2015.

[3] A. Arwiyah, M. Zainuri, M. Efendi, "Studi kandungan nacl di dalam air baku dan garam yang dihasilkan serta produktivitas lahan garam menggunakan media meja garam yang berbeda," J. Kelautan, 8(1): 1-9, 2015.

[4] Y. U. Hoiriyah, "Peningkatan kualitas produksi garam menggunakan teknologi geomembran," Jurnal Studi Manajemen dan Bisnis, 6(2): 35-40, 2019.

[5] Badan Standardisasi Nasional, Standar Nasional Indonesia (Indonesian National Standard)., "Garam Konsumsi Beriodium," SNI 3556:2016, 2016.

[6] L. R. B. Rebello, T. Siepman, S. Drexler, "Correlations between TDS and electrical conductivity for high-salinity formation brines characteristic of south atlantic pre-salt basins," Water SA, 46(4): 602-609, 2020.

[7] A. L. Rositawati, et al., "Rekristalisasi garam rakyat dari daerah demak untuk mencapai SNI garam industri," Jurnal Teknologi Kimia dan Industri, 2(4): 217-225, 2013.

[8] S. Choo-in, "The relationship between the total dissolved solids and the conductivity value of drinking water, surface water and wastewater," in Proc. the 2019 International Academic Research Conference: 11-16, 2019.

[9] D. L. Corwin, K. Yemoto, "Salinity: Electrical conductivity and total dissolved solids," Methods Soil Anal., 2: 1-16, 2017.

[10] A. Wibowo, "Potential of developing Indonesian national standards (SNI) for yodium salt products to increase competitiveness," in Proc. PPIS 2020 – Tangerang Selatan, Indonesia: 79-88, 2020.

[11] L. N. Nthunya, et al., "Spectroscopic determination of water salinity in brackish surface water in nandoni dam, at vhembe district, Limpopo province, south Africa," Water, 10(8): 1-13, 2018.

[12] H. Susanto, N. Rokhati, G. W. Santosa, "Development of traditional salt production process for improving product quantity and quality in jepara district, central java, Indonesia," Procedia Environ. Scie. 23: 175-178, 2015.

[13] T. G. Oh, C. H. Yim, G. S. Kim, "Esp8266 Wi-Fi module for monitoring system application," Global J. Eng. Scie. Res., 4(1): 1-6, 2017.

[14] H. Putranta, et al., "A simple liquid density measuring instrument based on Hooke's law and hydrostatic pressure," Phys. Educ., 55(2): 1-9, 2020.

[15] M. Abadi, A. Saleh, "Rancang bangun alat pengukur langkah kaki dengan sensor accelerometer dan fasilitas komunikasi wireless 2, 4 GHz," EEPIS Final Project, 2013.

[16] A. Saleh, A. Haryadi Amran D., Suwito, "Kendali Gerak Robot Berdasarkan Isyarat Tangan Menggunakan Komunikasi Nirkabel," In Proc. SENTIA 2014-Politeknik Negeri Malang, 6: 73-78, 2014.

[17] T. A. S. Wibawa, Arifin., A. Saleh, "Rancang bangun robot soccer wireless berbasis mikrokontroller," EEPIS Final Project., 2012.

[18] A. Saleh, "Implementasi pengolahan citra pada sistem pemantau level cairan berbasis Web," in Proc. SENTIA 2014-Politeknik Negeri Malang, 6: 1-6, 2014.

[19] C. Bayrak, et al, "Web-Based System Monitoring and Control," Arkansas, [Online].

[20] A. F. Rusydi, "Correlation between conductivity and total dissolved solid in various type of water: A review," in Proc. IOP conference series: earth and environmental science, 118(1), 2018.

[21] T. Perets, "Investigation of wi-fi (esp8266) module and application to an audio signal transmission," University of Buea, Cameroon, [Online].

[22] T. Sulistyaningsih, D. Alighiri, "Quality monitoring of salt produced in Indonesia through seawater evaporation on HDPE geomembrane lined ponds," J.Phys. Conf. Ser., 983(1): IOP Publishing, 2018.

[23] P. Nugroho, A. Susandini, D. Islam, "Development of madura salt industrialization amid the covid-19 pandemic," PalArch's Journal of Archaeology of Egypt/Egyptology 17(9): 1621-1636, 2020.

[24] A. Ubaidillah, D. Rahmawati, R. Aiman, "Architecture tools measure the levels of salt and ph of seawater using a fuzzy logic-based android," JEEMECS (J. Electr. Eng. Mechatron. Comput. Scie.), 1(2): 46-51, 2018.

[25] T. Widjaja, et al, "Iodization of local salt based on purification technique using saturated brine washing method," in Proc. IOP Conference Series: Materials Science and Engineering, 543(1): IOP Publishing, 2019.

[26] M. Z. Mahasin, Y. Rochwulaningsih, S. Tri Sulistiyono, "Coastal ecosystem as salt production centre in Indonesia," E3S Web of Conferences, 202. EDP Sciences, 2020.

[27] S. M. Purnama, A. Cahyawati, D. Y. Hutapea, "Salt pond analysis using ALOS PALSAR case study Sampang, Madura-Indonesia," in Proc. IOP Conference Series: Earth and Environmental Science., 162(1): IOP Publishing, 2018.

[28] H. Schmidt, et al., "The density–salinity relation of standard seawater," Ocean Sci. Discuss., 14(1): 15-40, 2018.

[29] M. L. Menn, et al., "The absolute salinity of seawater and its measurands," Metrologia, 56(1): 015005, 2018.

[30] M. Śmiechowska, M. Ruszkowska, O. Olender, "Selected quality characteristics of sea salt important during transport and storage," TransNav, Int. J. Mar. Navig. Saf. Sea Transp., 15(3): 659-665, 2021.

Biographies

Akuwan Saleh was born in Surabaya, East Java, Indonesia, on November 23, 1967. He graduated from a Master of Engineering (S2) from the Sepuluh Nopember Institute of Technology Surabaya (ITS). Currently working as a lecturer at Electronic Engineering Polytechnic Institute of Surabaya (EEPIS). Field of interest in particular Multimedia Telecommunication Engineering.

• Email: akuwan@pens.ac.id
• ORCID: 0000-00002- 9082-1448
• Web of Science Researcher ID: GGI-5764-2022
• Scopus ID: 57202497538
• Homepage: https://sinta.kemdikbud.go.id/authors/profile/6197901#!

Alif Sultonul Arifin was born in Pamekasan, East Java, Indonesia, on January 11, 1997. He completed his Diploma III (D3) education from Electronic Engineering Polytechnic Institute of Surabaya (EEPIS). Currently working as a professional entrepreneur. Field of interest in particular Telecommunications Engineering.

• Email: alifsultonul@gmail.com
• ORCID: NA
• Web of Science Researcher ID: NA
• Scopus ID: NA
• Homepage: NA

How to cite this paper:
A. Saleh, A. S. Arifin, "Indicators for determining salt harvest time based on salinity and liquid viscosity using microcontroller," J. Electr. Comput. Eng. Innovations, 11(1): 119-128, 2023.

DOI: 10.22061/JECEI.2022.8900.559

URL: https://jecei.sru.ac.ir/article_1734.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Improved Bilinear Balanced Truncation for Order Reduction of the High-Order Bilinear System Based on Linear Matrix Inequalities

H. Nasiri Soloklo, N. Bigdeli*

Department of Control Engineering, Imam Khomeini International University, Qazvin, Iran.

Article Info

Article History:
Received 05 April 2022
Reviewed 18 May 2022
Revised 15 June 2022
Accepted 30 August 2022

Keywords:
Model order reduction
Bilinear system
Linear matrix inequality
Generalized lyapunov equations
Balanced truncation

*Corresponding Author's Email Address:
n.bigdeli@eng.ikiu.ac.ir

Abstract

Background and Objectives: This paper proposes a new Model Order Reduction (MOR) method based on the Bilinear Balanced Truncation (BBT) approach. In the BBT method, solving the generalized Lyapunov equations is necessary to determine the bilinear system's controllability and observability Gramians. Since the bilinear systems are generally of high order, the computation of the Gramians of controllability and observability have huge computational volumes. In addition, the accuracy of reduced-order model obtained by BT is relatively low. In fact, the balanced truncation method is only available for local energy bands due to the use of type I Gramians. In this paper, BBT based on type II controllability and observability Gramians would be considered to fix these drawbacks.

Methods: At first, a new iterative method is proposed for determining the proper order for the reduced-order bilinear model, which is related to the number of Hankel singular values of the bilinear system whose real parts are closest to origin and have the most significant amount of energy. Then, the problem of determining of type II controllability and observability Gramians of the high-order bilinear system have been formulated as a constrained optimization problem with some Linear Matrix Inequality (LMI) constraints for an intermediate middle-order system. Then, the achieved Gramians are applied to the BBT method to determine the reduced-order model of the bilinear system. Next, the steady state accuracy of the reduced model would be improved via employing a tuning factor.

Results: Using the concept of type II Gramians and via the proposed method, the accuracy of the proposed bilinear BT method is increased. For validation of the proposed method, three high-order bilinear models are approximated. The achieved results are compared with some well-known MOR approaches such as bilinear BT, bilinear Proper Orthogonal Decomposition (POD) and Bilinear Iterative Rational Krylov subspace Algorithm (BIRKA) methods.

Conclusion: According to the obtained results, the proposed MOR method is superior to classical bilinear MOR methods, but is almost equivalent to BIRKA. It is out-performance respecting to BIRKA is its guaranteed stability and convergence.

Introduction

Bilinear systems are a class of nonlinear systems that serve as a link between linear and nonlinear systems. States and inputs of these systems are linear, but they are jointly nonlinear.

Researchers have been interested in bilinear systems for many years due to several real-world examples exhibiting such behavior. They include power systems [1], heat transfer [2], and electrical circuits [3].

One of the main applications of bilinear systems is the approximation of weakly nonlinear systems with bilinear systems using the Carleman bilinearization [4], [5]. However, the approximation of nonlinear systems via bilinearization methods usually leads to a high-order bilinear model. Analysis, design, and implementation of the high-order bilinear systems are complicated and time-consuming. Therefore, researchers have considered Model Order Reduction (MOR) of bilinear systems for analyzing and control purposes in the literature. Indeed, a reduced-order approximation of the high-order bilinear model is determined to decrease the complexity. For this purpose, MOR methods created for linear systems [6]-[8] were extended to bilinear systems. These methods include proper orthogonal decomposition (POD) [9], moment matching techniques [10], [11], Krylov subspace methods [12]-[14], projection-based methods [15]-[17] and polynomial expansion series based methods [18]-[20].

H2-optimal MOR methods are the other approaches to approximate the high order systems [21]-[23]. In [23], a special class of linear parameter-varying systems were reformulated as bilinear dynamical systems. Then, a H_2 norm in the generalized frequency domain was minimized based on the gradient descent on the Grassmann manifold. In [24], H_2 optimal MOR problems were investigated for K-power systems as a special class of bilinear systems. Xu et al. shown that the H_2 optimal MOR problem of the bilinear system could be considered as unconstrained minimization problem on Grassmann manifold by using Gramians of controllability and observability [25] and cross Gramians of the bilinear systems [26]. In [27], the Riemannian trust-region method considered this minimization problem on the Stifel manifold. The time-limited and frequency-limited H_2 optimal MOR were other approaches to approximate the high-order bilinear systems in [28], [29].

As one of the most popular MOR methods for linear deterministic control systems, Moore introduced the balanced truncation method in [30]. Hsu et al. in [31] extended the BT method for order reduction of bilinear systems, called bilinear BT or BBT. Afterward, many researchers focused on model order reduction of bilinear systems based on the BT method and improved it [32], [33].

In the BBT method, the Gramians of controllability and observability are crucial. Solving the generalized Lyapunov equations is necessary to determine the bilinear system's controllability and observability Gramians. Since the bilinear systems are generally of high order, the computation of the Gramians of controllability and observability have huge computational volumes. In addition, the accuracy of reduced-order obtained by BT is not clear [34]. In fact, the balanced truncation method is only available for local energy bands due to the use of type I Gramians and no error bounds have been provided for BBT so far [34]. Therefore, high computational volume and relatively low accuracy especially in steady state are major drawbacks of the BBT method [35]-[37]. Despite these drawbacks, the BBT method ensures stability and convergence which makes it suitable for order reduction of the intermediate systems [38]. In order to improve BT for bilinear systems, in [34], BBT was extended based on type II Gramians, where the H_∞ error bounds were achieved for the reduced bilinear system in terms of the truncated Hankel singular values. In computing these type II Gramians, the equality constraints in the generalized Lyapunov equations are replaced with inequality. Therefore, optimal determination of the type II Gramians via these inequalities is essential. However, up to the knowledge of the authors, no solving method has provided for computing these Gramians in [34] and the related literature afterwards, leading to another challenge in this area. Therefore, employing optimal type II Gramians for improving the BBT method accuracy with lower computational volume and preserving BBT benefits would be exciting and essential.

This paper proposes a new method for MOR of bilinear systems based on the BBT method using the LMI approach to increase the accuracy of the BBT method. For this purpose, at first, a new iterative method is proposed for determining the proper order for the reduced-order bilinear model, which is related to the number of Hankel singular values of the bilinear system whose real parts are closest to origin and have the most significant amount of energy. Then, the problem of determining of type II controllability and observability Gramians of the high-order bilinear system have been formulated as a constrained optimization problem with some Linear Matrix Inequality (LMI) constraints for an intermediate middle-order system. Then, the achieved Gramians are applied to the BBT method to determine the reduced-order model of the bilinear system. Next, the steady state accuracy of the reduced model would be improved via employing a tuning factor. The proposed method has the advantage of increasing the accuracy of the BBT method, while its computational complexity is reduced. To evaluate the efficiency of the proposed method, three test systems have been then examined. The achieved results are compared with some well-known MOR methods such as BBT, bilinear Proper Orthogonal Decomposition (POD) and Bilinear Iterative Rational Krylov subspace Algorithm (BIRKA) methods [39]. The results show that the proposed method is superior to

classical bilinear MOR methods, but is almost equivalent to BIRKA. It is out-performance respecting BRIKA is its guaranteed stability and convergence. Therefore, the main contribution of the article can be summarized as follows:

- Development of a new iterative method to determine the proper order for the reduced model.
- Introduction of improved BBT method with increased accuracy and reduced computational complexity and steady-state error.
- Reformulation of the generalized Lyapunov equations as an LMI constrained optimization problem for the middle order approximation of system to determine the Gramians of controllability and observability as type II Gramians to reduce computational complexity and improve BBT accuracy.
- Reducing BBT steady-state error by tuning the feedforward gain of the reduced-bilinear model.

The rest of this paper is organized as follows: In section 2, the MOR of bilinear systems is introduced. In section 3, the basics of the bilinear BT method are presented. In section 4, the proposed method for MOR of the bilinear system is introduced.

In section 5, three high-order bilinear test systems are reduced using the proposed MOR method. The achieved results demonstrate that the proposed methods outperform.

Finally, the paper is concluded in section 6.

MOR of Bilinear Systems

Consider the single-input, single-output bilinear system as follows:

$$\zeta:\begin{cases}\dot{x}(t) = Ax(t) + Nx(t)u(t) + Bu(t)\\ y(t) = Cx(t)\end{cases} \quad (1)$$

where, $A, N \in R^{n\times n}, B, C^T \in R^n$ are the matrices of the bilinear system, $x(t) \in R^n$ is the state vector, and $u(t) \in R$ and $y(t) \in R$ are the input and output of the bilinear system, respectively. Also, n is the order of the bilinear system.

Suppose that the bilinear system of (1) is of high order. Model order reduction aims to create a system in which the original bilinear system's and reduced-order approximation's responses are almost identical, i.e., $y(t) \approx y_r(t)$ for all admissible inputs. Further, both (1) and the reduced-order system have the same structure. The reduced-order bilinear model can be represented as follows:

$$\zeta_r:\begin{cases}\dot{x}_r(t) = A_r x_r(t) + N_r x_r(t)u(t) + B_r u(t)\\ y_r(t) = C_r x_r(t)\end{cases} \quad (2)$$

where, $A_r, N_r \in R^{r\times r}, B_r, C_r^T \in R^r$ are the matrices of the reduced-order bilinear system, which are unknown and should be determined, $x_r(t) \in R^r$ is the state vector, and $u(t), y_r(t) \in R$ are the input and output of the reduced-order bilinear system, respectively. Also, $r \ll n$ is the order of the reduced model.

It should be noted that if the original bilinear system is stable, the reduced-order model should be stable, too.

Balanced Truncation for Bilinear Systems

The controllability Gramians of the bilinear system of (1) are defined as follows [35]:

$$P = \sum_{i=1}^{\infty}\int_0^{\infty}\cdots\int_0^{\infty} P_i P_i^T\, dt_1 \cdots dt_i \quad (3)$$

where

$$P_1(t_1) = e^{At_1}B$$
$$P_i(t_1,\cdots,t_i) = e^{At_i}NP_{i-1} \quad (4)$$

Also, the observability Gramians of the bilinear system of (1) is defined as follows:

$$Q = \sum_{i=1}^{\infty}\int_0^{\infty}\cdots\int_0^{\infty} Q_i^T Q_i\, dt_1 \cdots dt_i \quad (5)$$

where

$$Q_1(t_1) = Ce^{At_1}$$
$$Q_i(t_1,\cdots,t_i) = Q_{i-1}Ne^{At_i} \quad (6)$$

Type I Gramians

Theorem 1 [40]. Consider the bilinear system of (1) with a stable matrix *A*. The truncated type I controllability Gramian *P* of the system satisfies the generalized Lyapunov equation given by (7), as:

$$AP + PA^T + NPN^T + BB^T = 0 \quad (7)$$

As a result of extending theorem 1 for observing Gramians, it is concluded that truncated type I Gramians of observability can be obtained by solving the following generalized Lyapunov equation:

$$A^TQ + AQ + N^TQN + C^TC = 0 \quad (8)$$

The following iterative method has been used to solve the generalized Lyapunov equations of (7) and (8) [36].

Initially, the bilinear term of (7) is eliminated. Therefore, the generalized Lyapunov equation is transformed into the following Lyapunov equation:

$$A\hat{P}_1 + \hat{P}_1A^T + BB^T = 0 \quad (9)$$

An initial solution for the generalized Lyapunov equation is obtained by solving the Lyapunov equation of (9).

Then, in each iteration, the truncated type I controllability Gramians is derived by applying the following iterative formula:

$$A\hat{P}_i + \hat{P}_i A^T + N\hat{P}_{i-1}N^T + BB^T = 0, \quad i = 2,3,\cdots \tag{10}$$

Finally, the type I controllability Gramian is determined as follows:

$$P = \lim_{i\to\infty} \hat{P}_i \tag{11}$$

Similar to the controllability Gramian, the truncated type I observability Gramian can be computed.

Type II Gramians

The generalized Lyapunov equations of the bilinear system can be extended to the following inequality equations [41], which are called type II Gramians:

$$A^T P^{-1} + P^{-1}A + N^T P^{-1} N \le -P^{-1}BB^T P^{-1} \tag{12}$$

$$A^T Q + QA + N^T QN \le -C^T C \tag{13}$$

Although (12) is constructed based on inverse controllability Gramians, it can be rewritten in terms of controllability Gramians by multiplying *P* to left and right-sides of (12).

The type II Gramians have some advantages respect to type I Gramians. These advantages include additional information about the control, global energy bounds, and availability of an H_∞-error bound for the bilinear BT method [34]. Therefore, BT model order reduction for bilinear systems based on type II Gramians are superior to those based on type I Gramians. However, the main difficulty in employing type II Gramians is that finding the type II Gramians via solving the inequalities of (12) and (13), especially for large scale systems is not a straight forward task.

BBT Algorithm

Table 1: Algorithm of bilinear BT method

***Input**: The system matrices: A, N, B, C.*

1 Determine low-rank approximation of Gramians:

$$P \approx RR^T \text{ and } Q \approx SS^T;$$

2 Compute SVD of $S^T R$ as follows:

$$S^T R = U\Sigma V = [U_1 \quad U_2]\begin{bmatrix}\Sigma_1 & 0\\ 0 & \Sigma_2\end{bmatrix}[V_1 \quad V_2]^T$$

The Σ_1 contains the r largest singular values of $S^T R$

3 Construct the transformation matrices T_1 and T_2:

$$T_1 = SU_1\Sigma_1^{-\frac{1}{2}}$$

$$T_2 = RV_1\Sigma_1^{-\frac{1}{2}}$$

4 Determine the reduced-order bilinear model:

$$A_r = T_2^T A T_1\,, N_r = T_2^T N T_1\,, B_r = T_2^T B\,, C_r = CT$$

***Output**: A_r, N_r, B_r, C_r.*

Once the type I or type II Gramians are obtained for a bilinear system, they can be employed in the context of balancing for MOR of system. For this purpose, the bilinear BT algorithm can be used as presented in Table 1 [34].

Proposed MOR Method

Numerically, it is not easy to apply balanced truncation to a bilinear system because it requires a solution of two high-order generalized Lyapunov equations. On the other hand, determining the type II Gramians by solving the generalized Lyapunov inequalities is challengeable. Hence, in this paper a new method has been proposed to solve the generalized Lyapunov equations using the LMI approach. After solving the generalized Lyapunov equations by the proposed method, the bilinear BT method is applied to order reduction of bilinear systems. The proposed algorithm is implemented via four steps, as follows.

Step 1. Determining the order of the reduced system:

The first step of determining the reduced-bilinear model of (1) is specifying the desired order. It is necessary to determine which modes have the most significant amount of energy to accomplish this. Modes with higher energy can be evaluated using dominant poles or the Hankel singular value method.

The number of high-energy modes is equal to the order of the reduced model. The order of the reduced model is determined initially by the number of eigenvalues whose real parts are closest to the origin. It is recommended to choose a conservative order at the beginning. The initial order is decreased one by one using a bilinear MOR method such as BPOD [42] until the error index significantly increases.

Therefore, the lowest order with negligible error is the most appropriate.

Step 2. Finding the truncated type II Gramians via LMI:

As stated earlier, in implementing BBT based on type II Gramians, solving the generalized Lyapunov inequalities is complicated, especially for large-scale systems. To address this problem, in this paper, at first, the generalized Lyapunov inequalities would be represented as a LMI constrained optimal problem. Then, it will be solved via an intermediate approximation of the system.

Let us consider the generalized Lyapunov equation as represented by (12). By adding the unknown coefficient of λ to (12), the controllability Gramians equation is converted to the following inequality equation:

$$AP + PA^T + NPN^T + BB^T + \lambda I < 0 \tag{14}$$

On the other hand, the Gramians of controllability should be positive definite. Hence, the controllability Gramians equation can be converted to a constrained optimization problem as follows:

$$\begin{cases} Min\ \lambda \\ s.t.\ AP + PA^T + NPN^T + BB^T + \lambda I_n < 0 \\ P > 0 \end{cases} \quad (15a)$$

By solving the constrained optimization problem of (15), the controllability Gramians would be determined. Similar to controllability Gramians, observability Gramians can be determined, as:

$$\begin{cases} Min\ \mu \\ s.t.\ AQ + QA^T + NQN^T + C^TC + \mu I_n < 0 \\ Q > 0 \end{cases} \quad (15b)$$

In the next step, in order to facilitate solving the optimization problem of (15) and to decrease the computational volume, a middle-order approximation of the original system would be obtained by the conventional MOR methods such as BT or BPOD. It should be noted that both of these methods yields quite precise approximations in middle orders [43]. Then the LMI problem of (15) would be solved for this reduced system. In this case, the optimization problem can be considered as follows:

$$\begin{cases} Min\ \lambda \\ s.t.\ A_mP_m + P_mA_m^T + N_mP_mN_m^T + B_mB_m^T + \lambda I_m < 0 \\ P_m > 0 \end{cases} \quad (16a)$$

$$\begin{cases} Min\ \mu \\ s.t.\ A_mQ_m + Q_mA_m^T + N_mQ_mN_m^T + C_mC_m^T + \mu I_m < 0 \\ Q_m > 0 \end{cases} \quad (16b)$$

where, in (16) index *m* implies the system matrices and Gramians of the approximated middle-order system.

Step 3. Apply the bilinear BT method:

The achieved type II Gramians in the previous step are applied to the bilinear BT algorithm to obtain a reduced-order bilinear model.

Step 4. Adjust the gain of the reduced-order:

The obtained bilinear reduced-order model by the bilinear BT method usually suffers from steady-state error [44]. In other words, the final value of the achieved bilinear reduced-order model deviates from the final value of the original bilinear system. To address this problem, a feedforward tuning factor is added to the bilinear reduced-order model, which is determined in this step as follows.

Consider the reduced model of (17), in which the tuning factor of *K* has been added to the output equation as:

$$\begin{cases} \dot{x}_r(t) = A_rx_r(t) + N_rx_r(t)u(t) + B_ru(t) \\ y_r(t) = K\bar{C}_rx_r(t) \end{cases} \quad (17)$$

where $\bar{C}_r$ is the output vector determined by the bilinear BT method in previous step.

The steady-state error of the bilinear BT model is removed by properly adjusting *K*. To tune this parameter, two approaches can be considered.

In the first approach, the gain of the reduced-order model is tuned as an optimization problem by a swarm intelligence-based algorithm. The fitness function can be considered as a function of output error, such as the Integral Square of Error (ISE).

In the second approach, the tuning factor of *K* is set so that the steady-state output error for non-oscillating bounded inputs is removed. Let us define the steady-state output error as:

$$\lim_{t\to t_f} e(t) = \lim_{t\to t_f} |y(t) - Ky_r(t)| \quad (18)$$

where, t_f represents the large enough settling time of the response. Then, *K* is tuned to remove this steady state error as:

$$K = \frac{y(t_f)}{y_r(t_f)} \quad (19)$$

Therefore, the proposed approach can be implemented via the algorithm of Table 2 as follows:

Table 2: The proposed MOR algorithm

***Input**: The system matrices: A, N, B, C.*

1 Determine the suitable order of the reduced-bilinear model by an iterative method.

2 Find the middle order approximation of the system and solve the LMI constrained optimization problem of (16) to achieve type II controllability and observability Gramians.

3 Apply the obtained type II controllability and observability Gramians to the BT method to determine the reduced-order bilinear model.

4 Adjust the gain of the system by tuning the factor of K.

***Output**: A_r, N_r, B_r, C_r.*

Simulation Results

Here, three high-order bilinear systems are considered as test systems. These test systems are approximated by the proposed method. A bilinear system with an order of 200 is the first test system. Then, the Chaffee-Infante model would be approximated by the proposed method. The third test system is a nonlinear transmission line circuit converted to a bilinear system by Carleman bilinearization. To validate the proposed method, the obtained reduced-order models are compared with some well-known MOR methods such as BT, BPOD and BIRKA methods. The results show that the proposed method

matches the original systems more than other approaches.

Test system 1

In [24], a bilinear system of order 200 is presented with the following matrices:

$$\dot{x}(t) = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix} x(t) + \begin{bmatrix} 0 & 0 \\ N_1 & 0 \end{bmatrix} x(t)u(t) + \begin{bmatrix} B_1 \\ 0 \end{bmatrix} u(t) \quad (20)$$

$$y(t) = [0 \quad C_2]x(t)$$

where $A_1 \in R^{100\times100}$, $A_2 \in R^{100\times100}$, $B_1 \in R^{100\times1}$ and $C_2 \in R^{1\times100}$

$$A_1 = \begin{bmatrix} -10 & 2 & & \\ 7 & -10 & 2 & \\ & \ddots & \ddots & \ddots \\ & & 7 & -10 \end{bmatrix}, A_2 = \begin{bmatrix} -5 & 2 & & \\ 2 & -5 & 2 & \\ & \ddots & \ddots & \ddots \\ & & 2 & -5 \end{bmatrix} \quad (21)$$

$$N_1 = \begin{bmatrix} 2 & 1 & & \\ -1 & 2 & 1 & \\ & \ddots & \ddots & \ddots \\ & & -1 & 2 \end{bmatrix}, B_1 = [1 \quad \cdots \quad 1],$$

$$C_2 = [0 \quad \cdots \quad 0 \quad 1]^T$$

The procedure of order reduction by the proposed method is followed step by step as:

Step 1: In the first step, the order of the reduced bilinear model is determined. For this purpose, the initial order is determined based on the number of eigenvalues with the real part close to the origin. According to Fig. 1, this guess is 20. Then, the order of the reduced-model decreased one by one from 20 to 1, and for each order, the BPOD was applied. The H_2 norm of error for each order is calculated. The H_2 norm of the error for each order of the reduced test system 1 is shown in Fig. 2.

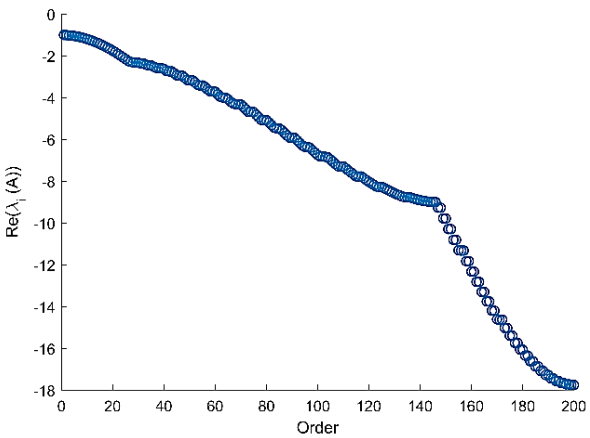


Fig. 1: Real part of eigenvalues of test system 1.

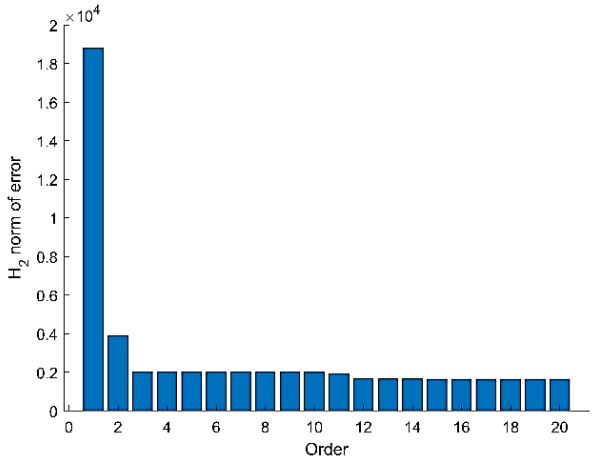


Fig. 2: H_2-norm of error versus order.

It can be shown that the proper order for the reduced-order bilinear model of test system 1 is 2.

Step 2: In this step, a new structure for computation of controllability and observability Gramians would be considered. By adding an unknown term of α to the controllability relation, this equation is converted to a LMI optimization problem. A similar approach can be applied to observability Gramians. Then, the optimization problem with inequality constraints of (15) is minimized to determine the alternative controllability Gramian. Following the same procedure, the observability Gramian is determined, as well.

Step 3: Using the bilinear BT method, the reduced-order bilinear model is derived.

Step 4: Adjust the gain of the reduced-order system to remove the steady-state error. Here $K = 1.016$ has selected as the ratio of the original and reduced order systems.

The achieved reduced-order bilinear model is presented as follows:

$$\dot{x}_r(t) = \begin{bmatrix} -1.0196 & -7.0814e-18 \\ -1.3391e-18 & -1.0092 \end{bmatrix} x_r(t) + \begin{bmatrix} 0 & 0 \\ 3.2332 & -1.1425e-18 \end{bmatrix} x_r(t)u(t) + \begin{bmatrix} 4.9068 \\ 2.5369e-17 \end{bmatrix} u(t) \quad (22)$$

$$y_r(t) = 1.016 \times [0 \quad 12.1142]x_r$$

Fig. 3 illustrates the response of the reduced bilinear model of Eq. (22) to input $u(t) = 0.05exp(-0.5t)$. Also, it is compared with some well-known MOR methods, including BT, BPOD and BIRKA methods. In addition, the absolute error has been also evaluated over time and depicted in Fig. 4. It can be shown that the proposed method matches the original system much better than other methods, and it has a smaller error compared to the other methods. Some important characteristics of the response are compared to provide a quantitative and numerical evaluation. These specifications include peak,

steady-state, and ISE index as an appropriate measures for evaluating the approximation error. The results of the comparison have been presented in Table 2.

Figs. 3, 4 and Table 3 show that the obtained reduced-order bilinear model via the proposed method provides the best approximation among the other indicated approaches.

Test System 2: Chaffee Infante

Another standard test system used to evaluate MOR methods is the one-dimensional Chaffee-Infante equation on $\Omega = (0, L) \times (0, T)$ [42]. In this equation, there is cubic nonlinearity:

$$v_t + v^3 = v_{xx} + v \quad (0, L) \times (0, T) \tag{23}$$

where ν is the viscosity parameter.

The initial and boundary conditions of the Chaffee-Infante equation have been considered in (24)-(26).

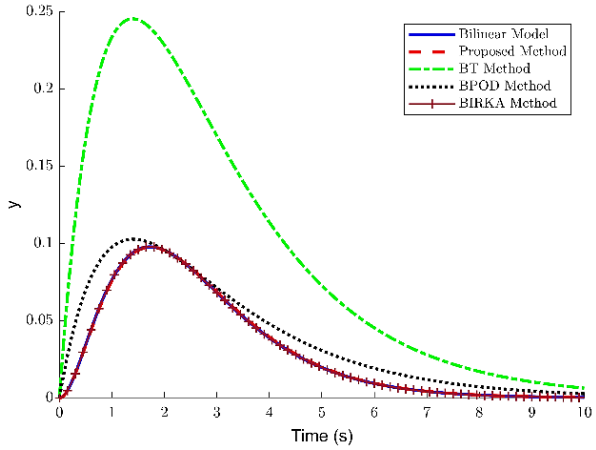


Fig. 3: Comparison of responses of the bilinear model of test system 1 and their reduced-order model approximations.

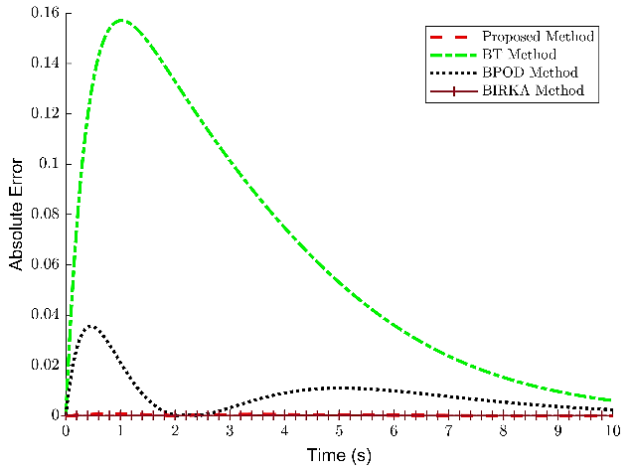


Fig. 4: Time evolution of absolute error of various methods for approximations of test system 1.

Table 3: Comparison of methods for test system 1

	Order	Final value	Peak	ISE
Original System	200	3.24e-04	0.0977	-
Proposed Method	2	3.31e-04	0.0975	1.02e-06
BT Method	2	0.0063	0.2453	0.0660
BPOD Method	2	0.0027	0.1027	0.0013
BIRKA Method	2	3.19e-04	0.0978	1.63e-08

$$\alpha v(0, t) + \beta v(0, t) = u(t) \tag{24}$$

$$v(L, t) = 0 \quad t \in (0, T) \tag{25}$$

$$v(x, 0) = v_0(x) \tag{26}$$

where α and β are constant parameters and $v_0(x)$ is initial condition of the system.

A finite-difference scheme was used to obtain the spatial discretization system. Then, Carleman bilinearization converts the nonlinear ODEs of the Chaffee-Infante equation to bilinear form. For this test system, $L = 0.1$ and $T = 5$. Also, the initial condition is considered zero, i.e., $v_0(x) = 0$. The discretization involved 31 points. Thus, the order of the bilinear model of the Chaffee-Infante is 992. The proposed method is applied to order reduction of the bilinear Chaffee-Infante model. The order of reduced approximation is 10. Similar to test system 1, the obtained reduced-order model is compared with some well-known model order reduction methods such as BPOD, BBT and BIRKA methods. In Fig. 5, responses of reduced-order bilinear models to input $u(t) = 0.5(1 + \cos(\pi t))$ are shown. Also, the absolute error versus time is presented in Fig. 6.

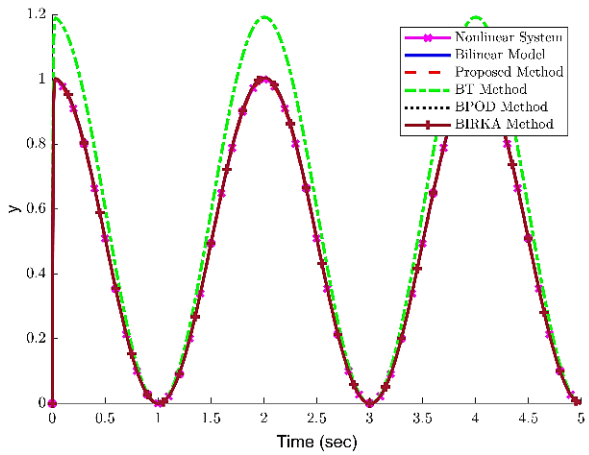


Fig. 5: Comparison of responses of the bilinear model of Chaffee Infante equation and their reduced-order model approximations.

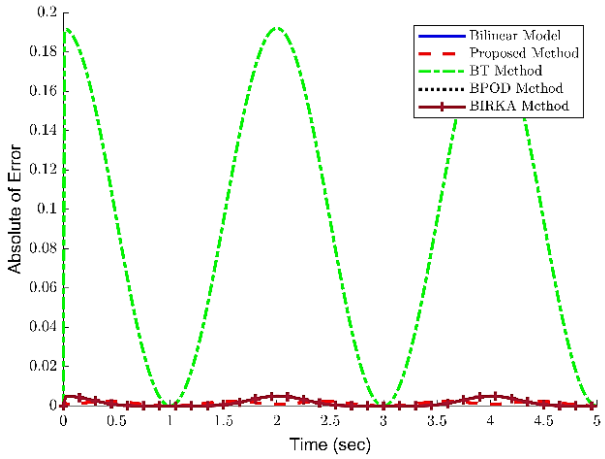


Fig. 6: Time evolution of absolute error of various methods for bilinear Chaffee Infante approximations.

According to Fig. 5 and Fig. 6, it is seen that the proposed method results as well as those of BIRKA are similar to the high-order bilinear model of the Chaffee Infante model.

Test system 3: Transmission Line Circuit

Fig. 7 depicts the transmission line circuit, including nonlinear resistors and an independent current source.

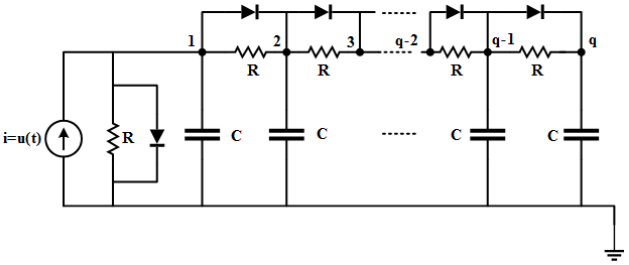


Fig. 7. Transmission line circuit.

Transmission lines can be modeled as a nonlinear state spaces form, as follows [6]:

$$\dot{x}(t) = \begin{bmatrix} -g(x_1) - g(x_1 - x_2) \\ -g(x_1 - x_2) - g(x_2 - x_3) \\ \vdots \\ -g(x_{k-1} - x_k) - g(x_k - x_{k+1}) \\ \vdots \\ -g(x_{q-1} - x_q) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} u(t) \tag{27}$$

$$y(t) = [1 \quad 0 \quad \cdots \quad 0]x(t)$$

where $x \in R^{q\times 1}$ is the state variables, $f \in R^{q}$ is nonlinear state evolution function, $b \in R^{q\times 1}$ and $c \in R^{1\times q}$ are input and output, respectively. Also, the relation between voltage and current of each resistor is modeled as $g(x) = exp(x) + x - 1$.

In this case, the number of resistors is chosen to be 20.

In order to approximate the nonlinear RC circuit system (27) using a bilinear model, the Carleman bilinearization is used [4], [5]. The order of the resulting bilinear model is $q^2 + q = 420$.

The obtained bilinear model of the transmission line is high-order and should be reduced. For this purpose, the proposed method is applied to approximate the reduced-order bilinear model of the transmission line model. The reduced-order bilinear model for the bilinear transmission line model is as follows:

$$\begin{aligned} &\dot{x}_r(t) \\ &= \begin{bmatrix} -68.20 & -9.87 & -3.11 \\ -6.42 & -92.79 & 9.69 \\ 2.96 & 18.77 & -61.27 \end{bmatrix} x_r \\ &+ \begin{bmatrix} 0.011 & 0.038 & -0.039 \\ 0.038 & 0.181 & -0.133 \\ -0.008 & -0.03 & 0.024 \end{bmatrix} x_r u \\ &+ \begin{bmatrix} -0.09 \\ -0.26 \\ 0.18 \end{bmatrix} u \end{aligned} \tag{28}$$

$$y_r(t) = 20 \times [-0.06 \quad -0.18 \quad 0.112]x_r$$

The responses of the nonlinear transmission line model and their approximations to input $u(t) = sin(10t)cos(t)exp(-1.5t)$ have been presented in Fig. 8. Also, the absolute error of the obtained reduced-order models has been shown in Fig. 9. In Table 4, some specifications of the achieved reduced-order bilinear model have been compared, as well. It is seen from Fig. 8 and Fig. 9 that the proposed method and the BIRKA have the most similarity and less error among the reduced-order bilinear systems. Besides, according to Table 4, it can be observed that the proposed method and the BIRKA method have the best approximation among the reduction methods.

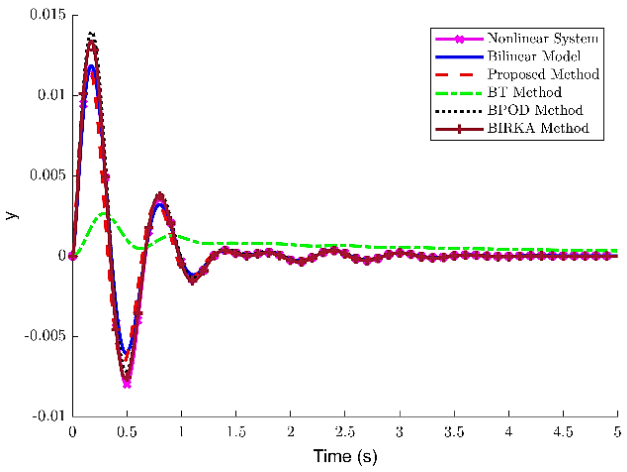


Fig. 8: Comparison of responses of the nonlinear transmission line system and their reduced-order model approximations.

However, although the BIRKA is as accurate as the

proposed method, its convergence is not guaranteed, generally [39].

In 50 simulations performed by the BIRKA to test System 3, it was observed that BIRKA diverges nine times, indicating an 18% failure rate in model order reduction. It is the main drawback of BIRKA, which is not observed in the BT family.

Indeed, as discussed earlier, convergence is guaranteed via employing BT. Therefore, the proposed method improves the bilinear BT method's performance and preserves the bilinear BT method's specifications, such as guaranteed stability and convergence.

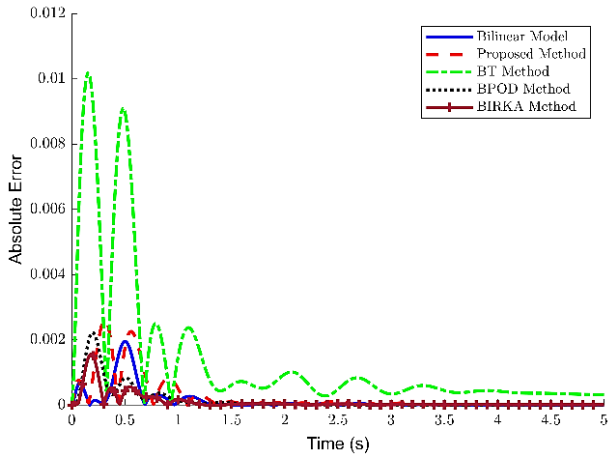


Fig. 9: Time evolution of absolute error of various methods for bilinear transmission line model approximations.

Table 4: Comparison of methods for test system 3

	Order	Final value	Peak	ISE
Nonlinear pendulum system	20	1.155e-05	0.0118	-
Bilinear Model	420	1.330e-05	0.0118	6.43e-07
Proposed Method	3	8.247e-07	0.0118	1.83e-06
BT Method	3	3.283e-04	0.0026	3.30e-05
BPOD Method	3	7.426e-07	0.0140	8.04e-07
BIRKA Method	3	-1.2903e-06	0.0134	3.51e-07

To further analysis, the simulation time of the MOR methods is compared. The time required for approximation of test system 3 by MOR methods is given in Table 5.

Table 5: Comparison of the simulation time of MOR methods for test system 3

	Simulation Time (sec)	Quality of Approximation
Proposed Method	6.24	Very good
BT Method	6.81	Not good
BPOD Method	32.69	Good
BIRKA Method	431.33	Very good

It can be seen that the proposed method and the bilinear BT method need less time to approximate the test system 3. However, the bilinear reduced-order model obtained by the bilinear BT method is not a good approximation. On the other hand, the proposed method and BIRKA have high quality to approximate test system 3, but the BIRKA needs about 70 times more simulation time.

It can be noted that the proposed method used to MOR of test system 3 is implemented by minimizing the optimization problem of (16). To do this, an initial bilinear reduced-order model with order 25 is approximated by a bilinear BT method. Then, the optimization problem of (16) is minimized to determine the bilinear reduced model. The required simulation time for this two-stage is 6.24 seconds.

Results and Discussion

This study investigates the MOR of the bilinear systems based on the improved bilinear BT method. The proposed method uses the concept of type II Gramians to determine controllability and observability Gramians. To determine these Gramians, a new LMI-based approach is applied. After determining the new Gramians, the bilinear BT method is used. Since type II Gramians have more advantages than type I Gramians, the accuracy of the obtained reduced-order bilinear model is higher than the bilinear BT method. The proposed method is not only more accurate than bilinear BT, but it also has the advantages of balanced truncation, including ensuring stability and convergence.

Furthermore, the steady-state error of the reduced-order bilinear model is removed by adjusting the tuning factor. Three test systems are considered and compared with some well-known MOR methods to evaluate the proposed method. The results show that the proposed method is more similar to high-order bilinear systems and outperforms other approaches.

Conclusions

This paper proposes a new MOR method based on the balance truncation approach with type II Gramians for order reduction of the bilinear systems. For this purpose, at first, a new iterative method is proposed for determining the proper order for the reduced-order bilinear model which is related to the number of eigenvalues of the bilinear system whose real parts are closest to origin and have the most significant amount of energy. Then, the generalized Lyapunov equations are constructed to determine the Gramians of controllability and Gramians of the observability. These generalized Lyapunov equations are transformed into a linear matrix inequality problem by adding an unknown coefficient. Next, the LMI problem is converted to a constrained optimization problem. New controllability and observability Gramians are determined by solving the constrained LMI optimization problem. The obtained Gramians are applied to the bilinear BT method to determine the reduced-order bilinear model.

These type II Gramians, determined by solving the constrained optimization problem as an LMI problem, contain additional information compared to type I Gramians. Also, type II Gramians lead to finding global energy bounds. Therefore, obtaining type II Gramians in the context of balancing leads to increasing the accuracy of the BT method. On the other hand, the order determined by the proposed method is more appropriate and accurate than other methods of determining the order of reduced systems. Three high-order bilinear test systems are approximated to show the efficiency and ability of the proposed method. The achieved results are compared with some classical order reduction methods such as the BT, BPOD and BIRKA. According to the obtained results, it can be concluded that the proposed MOR method is superior to classical bilinear MOR methods, but is almost equivalent to BIRKA. It is out-performance respecting BIRKA is its guaranteed stability and convergence.

Author Contributions

H. Nasiri Soloklo designed and simulated the proposed method and wrote the manuscript. N. Bigdeli chose strategies, analyzed the results, edited the manuscript, and managed the entire process.

Acknowledgment

A special thanks go out to the anonymous reviewers of this article and the editor of JECEI for their valuable suggestions.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

BBT	Bilinear Balanced Truncation
BIRKA	Bilinear Iterative Rational Krylov Subspace
BPOD	Bilinear Proper Orthogonal Decomposition
ISE	Integral square of Error
LMI	Linear Matrix Inequality
MOR	Model Order Reduction

References

[1] S. A. Al-Baiyat, M. Bettayeb, U. M. Al-Saggaf, "New model reduction scheme for bilinear systems," Int. J. Syst. Sci., 25: 1631–1642, 1994.

[2] P. Benner, T. Damm, "Lyapunov equations, energy functionals, and model order reduction of bilinear and stochastic systems," SIAM J. Control Optim·, 49(2): 686-711, 2011.

[3] P. Benner, P. Goyal, S. Gugercin, "H_2-quasi-optimal model order reduction for quadratic-bilinear control systems," SIAM J. Matr Anal. Appl., 39(2): 983–1032, 2019.

[4] W. J. Rugh, Nonlinear System Theory, Johns Hopkins University Press, Baltimore, 1981.

[5] S. V. Dushin, A. N. Abramenkov, E. J. Kutyakov, A. B. Iskakov, A. M. Salnikov, "Developing a weakly nonlinear power system model using the carleman bilinearization procedure," 2020 2nd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA): 963-967, 2020.

[6] U. Baur, P. Benner, L. Feng, "Model order reduction for linear and nonlinear systems: A system-theoretic perspective," Arch. Comput. Methods. Eng., 21: 331–358, 2014.

[7] H. Nasiri Soloklo, N. Bigdeli, "A PFC-based hybrid approach for control of industrial heating furnace," J. Electr. Comput. Eng. Innovations (JECEI), 7(1): 83-94, 2018.

[8] H. Hooshmand, M. M. Fateh, "Voltage control of flexible-joint robot manipulators using singular perturbation technique for model order reduction," J. Electr. Comput. Eng. Innovations (JECEI), 10(1): 123-142, 2022.

[9] V. B. Nguyen, M. Buffoni, K. Willcox, B. C. Khoo, "Model reduction for reacting flow applications," Int. J. Comput. Fluid Dyn., 28(3-4): 91-105, 2014.

[10] N. Stahl, B. Liljegren-Sailer, N. Marheineke, "Moment matching based model order reduction for quadratic-bilinear systems," in Proc. Faragó I., Izsák F., Simon P. (eds) Progress in Industrial Mathematics at ECMI 2018. Mathematics in Industry, (2019) 30, Springer, Cham.

[11] O. Agbaje, D. L. Shona, O. Haas, "Multimoment matching analysis of one-sided Krylov subspace model order reduction for nonlinear

and bilinear systems," in Proc. 2018 European Control Conference (ECC): 2599-2604, 2018.

[12] Y. Lin, L. Bao, Y. Wei, "A model-order reduction method based on Krylov subspace for MIMO bilinear dynamical systems," J. Appl. Math. Comput., 25: 293-304, 2007.

[13] Y. Lin, L. Bao, Y. Wei, "Order reduction of bilinear MIMO dynamical systems using new block Krylov subspace," Comput. Math. Appl., 58: 1093-1102, 2009.

[14] T. Breiten, T. Damm, "Krylov subspace methods for model order reduction of bilinear control systems," Syst. Control Lett., 59: 443-450, 2010.

[15] J. R. Philips, "Projection frameworks for model reduction of weakly nonlinear systems," in Proc. 37th Design Automation Conference, 184-189, 2000.

[16] Z. Bai, D. Skoogh, "A projection method for model reduction of bilinear dynamical systems," Linear Algebra Appl., 415: 406-425, 2006.

[17] M. Redmann, I. P. Duff, "Full state approximation by Galerkin projection reduced order models for stochastic and bilinear systems," Appl. Math. Comput., 420: 126561, 2022.

[18] L. Dai, Z. H. Xiao, R. Z. Zhang, "Laguerre-Gramian-based model order reduction of bilinear systems," in Proc. 2021 40th Chinese Control Conference (CCC): 1195-1200, 2021.

[19] X. Wang, Y. Jiang, "Model reduction of discrete-time bilinear systems by a Laguerre expansion technique," Appl. Math. Model., 40(13–14): 6650-6662, 2016.

[20] Y. Li, Y. Jiang, P. Yang, "Time domain model order reduction of discrete-time bilinear systems with Charlier polynomials," Math. Comput. Simul., 190: 905-920, 2021.

[21] L. Zhang, J. Lam, "On H_2 model reduction of bilinear systems," Automatica, 38(2): 205-216, 2002.

[22] P. Benner, T. Breiten, "Interpolation-based H_2 model reduction of bilinear control systems," SIAM J. Matr. Anal. Appl., 33(3): 859-885, 2012.

[23] P. Benner, X. Cao, W. Schilders, "A bilinear H_2 model order reduction approach to linear parameter-varying systems," Adv. Comput. Math., 45: 2241–2271, 2019.

[24] X. L. Wang, Y. L. Jiang, "On model reduction of K-power bilinear systems," Int. J. Syst. Sci., 45(9): 1978-1990, 2014.

[25] K. L. Xu, Y. L. Jiang, Z. X. Yang, "H_2 order-reduction for bilinear systems based on Grassmann manifold," J. Franklin Inst., 352: 4467–4479, 2015.

[26] K. L. Xu, Y. L. Jiang, Z. X. Yang, "H_2 optimal model order reduction by two-sided technique on Grassmann manifold via the cross-gramian of bilinear systems," Int. J. Control., 90(3): 616-626, 2017.

[27] P. Yang, Y. L. Jiang, K. L. Xu, "A trust-region method for H_2 model reduction of bilinear systems on the Stiefel manifold," J. Franklin Inst., 356(4): 2258-2273, 2019.

[28] K. L. Xu, Y. L. Jiang, "An approach to $H_{2,\omega}$ model reduction on finite interval for bilinear systems," J. Franklin Inst., 354: 7429-7443, 2017.

[29] U. Zulfiqar, V. Sreeram, M. I. Ahmad, X. Du, "Time and frequency-limited H_2-optimal model order reduction of bilinear control systems," Int. J. Syst. Sci., 52(10): 1953-1973, 2021.

[30] B. C. Moore, "Principal component analysis in linear systems: controllability, observability, and model reduction," IEEE Trans. Autom. Control., AC-26: 17-32, 1981.

[31] C. S. Hsu, U. B. Desai, C. A. Crawley, "Realization algorithms and approximation methods of bilinear systems," presented at the 22nd IEEE Conference on Decision and Control, San Antonio, Texas, 1983.

[32] I. P. Duff, P. Goyal, P. Benner, "Balanced truncation for a special class of bilinear descriptor systems," IEEE Control Syst. Lett., 3(3): 535-540, 2019.

[33] S. A. Al-Baiyat, A. S. Farag, M. Bettayeb, "Transient approximation of a bilinear two-area interconnected power system," Electr. Power Syst. Res., 26(1): 11–19, 1993.

[34] M. Redmann, "Type II balanced truncation for deterministic bilinear control systems,"arXiv: 1709.05655, 2017.

[35] P. Benner, P. Goyal, "Balanced truncation model order reduction for quadratic-bilinear control systems," Technical Report, 2017.

[36] P. Benner, P. Goyal, M. Redmann, Truncated Gramians for bilinear systems and their advantages in model order reduction, In Benner P, Ohlberger M, Patera T, Rozza G, Urban K, eds., Model reduction of parameterized systems. Springer International Publishing, 2017.

[37] S. K. Suman, A. Kumar, "Linear system of order reduction using a modified balanced truncation method," Circuits Syst. Signal Process, 40: 2741–2762, 2021.

[38] M. Redmann, "The missing link between the output and the H_2-norm of bilinear systems." arXiv, 1910.14427, 2019.

[39] R. Choudhary, K. Ahuja, "Inexact linear solves in model reduction of bilinear dynamical systems," IEEE Access, 7: 72297–72307, 2019.

[40] L. Q. Zhang, J. Lam, B. Huang, G. H. Yang, "On Gramians and balanced truncation of discrete-time bilinear systems," Int. J. Control., 76(4): 414–427, 2003.

[41] P. Benner, T. Damm, Y. R. Rodriguez Cruz, "Dual pairs of generalized lyapunov inequalities and balanced truncation of stochastic linear systems," IEEE Trans. Automat. Contr., 62(2): 782-791, 2017.

[42] K. Kerschen, J. Golinval, A. F. Vakakis, L. A. Bergman, "The method of proper orthogonal decomposition for dynamical characterization and order reduction of mechanical systems: An Overview., Nonlinear Dyn., 41: 147–169, 2005.

[43] A. C. Antoulas, D. C. Sorensen, "Approximation of Large-Scale Dynamical Systems: An Overview," Int. J. Appl. Math. Comput. Sci. 11 (5), 1093-1121, 2001.

[44] P. Benner, T. Breiten, "Two-sided projection methods for nonlinear model order reduction," SIAM J. Sci. Comput., 37(2): B239–B260, 2015.

Biographies

Hasan Nasiri Soloklo was born in Tehran in 1986. He received his M.Sc. degree in control engineering from Shahid Bahonar University of Kerman in 2012. Currently he is a Ph.D. candidate in Imam Khomeini International University of Qazvin. His research interests include model order reduction, bilinear systems, metaheuristic algorithms and evolutionary computation.

- Email: hasannasirisoloklo@edu.ikiu.ac.ir
- ORCID: 0000-0002-3712-9866
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Nooshin Bigdeli was born in 1978 in Iran, and completed her Ph.D. degree in Electrical Engineering majoring in Control at Sharif University of Technology, Tehran, Iran in 2007. She is currently professor of Control Engineering Department of Imam Khomeini International University, Qazvin, Iran. Her research interests include control systems, intelligent systems, chaos control, model predictive control as well as model order reduction in high order systems.

- Email: n.bigdeli@eng.ikiu.ac.ir
- ORCID: 0000-0001-5536-4491
- Web of Science Researcher ID: AAT-8622-2021
- Scopus Author ID: 8528681600
- Homepage: http://ikiu.ac.ir/members/?lang=1&id=23

How to cite this paper:
H. Nasiri Soloklo, N. Bigdeli, "Improved bilinear balanced truncation for order reduction of the high-order bilinear system based on linear matrix inequalities," J. Electr. Comput. Eng. Innovations, 11(1): 129-140, 2023.

DOI: 10.22061/JECEI.2022.8812.554

URL: https://jecei.sru.ac.ir/article_1766.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Improving the Classification of MPSK and MQAM Modulations by Using Optimized Nonlinear Preprocess in Flat Fading Channels

I. Kadoun, H. Khaleghi Bizaki*

Electrical and Computer Engineering Department, Malek Ashtar University of Technology, Tehran, Iran.

Article Info

Article History:
Received 11 April 2022
Reviewed 15 May 2022
Revised 15 July 2022
Accepted 30 August 2022

Keywords:
Automatic modulation classification
Linear discriminant analysis
Higher-Order cumulants
Mahalanobis distance

*Corresponding Author's Email Address: *bizaki@yahoo.com*

Abstract

Background and Objectives: Intelligent receivers, automatically detect the digital modulation type of the received signals for demodulation purposes where is well known as Automatic Modulation Classification (AMC) module. The performance of AMC algorithms depends on the channel conditions where for example, in fading channel its performance gets worse than the AWGN channel.
Methods: We propose a new algorithm for improving the AMC classification accuracy in flat fading channels. The proposed algorithm consists of an optimizable nonlinear preprocess followed by Linear Discriminant Analysis (LDA) technique. Two Lemmas have been found for extracting the optimization rule. And an optimization algorithm has been built based on the previous Lemmas.
Results: The simulation results show that the proposed algorithm improves the classification accuracy between 8-Phase Shift Keying (8PSK) and 16PSK (as an example of M-array PSK (MPSK) inter-class) for Signal-to-noise ratio (SNR) values greater than 13 dB, and between 16-quadrature amplitude shift modulation (16QAM) and 64QAM (as an example of M-array QAM (MQAM) inter-class) for SNR values greater than 4 dB. On the other hand, the classification accuracy of MPSK and MQAM is improved using the proposed algorithm compared with reference papers. Its improvement is up to 10.79% compared with the [1] and up to 38.552% compared with [2].
Conclusion: By using the proposed optimization algorithm, the AMC classification accuracy has been improved. Other classification problems can use this algorithm. And other nonlinear preprocess functions or optimization algorithms may be found in future work.

Introduction

With the significant development of modern communication technology, the AMC of the received signal is becoming more critical. Two primary AMC techniques are known: Likelihood-Based (LB) and Feature-Based (FB). LB techniques suffer from high computational complexity and need to estimate the unknown parameters [3], [4]. On the other hand, FB techniques have less complexity, don't need any parameter estimation [5], [6], and can work under different conditions like multipath fading channels [7].

Various studies were done to find good discriminative features for modulation classification like instantaneous time-domain features, Fourier and wavelet transform, higher-order moments, and cumulants [5]-[11]. Comparisons between the performances of these features were made in [12], [13]. According to their results, Higher-Order Cumulants (HOCs) are the best features under different conditions.

Different studies and simulations were done for AMC in fading channels using different HOCs. For example, in

[1], the author shows that the performance accuracy of MPSK and MQAM classification by using HOMs and HOCs is 84.37%. While in [2], the author shows that the performance accuracy of binary phase-shift keying (BPSK), Quadrature Phase-Shift Keying (QPSK), 8-phase-shift keying (8-PSK), 16-PSK, 16-quadrature amplitude modulation (16-QAM), 32-QAM, and 64-QAM classification by using cyclic cumulants is 89.8%, for SNR value of 15 dB.

The most critical inter-class modulation types are MPSK and MQAM [1], [2]. Most of the well-known intra-class modulation types are shown in Table 1.

Table 1: Chosen types of MPSK and MQAM digital modulations

Inter-class modulation	Intra-class modulation
MPSK	BPSK, QPSK, 8PSK, 16PSK
MQAM	8QAM, 16QAM, 32QAM, 64QAM

Our study improves the AMC performance by enhancing the discrimination between some intra-class digital modulation types in Table 1 (like 8PSK and 16PSK, and like 16QAM and 64QAM) for lower SNR values. This improvement is made by optimizing an added nonlinear preprocess function. Two primary optimizable nonlinear functions have been developed: regularized distance-based and nonlinear transformation. The simulation results show that these optimized functions could improve the discrimination between 8PSK and 16PSK for SNR values greater than 13 dB and between 16QAM and 64QAM for SNR values greater than 4 dB. On the other hand, the classification accuracy of MPSK and MQAM has been improved using the proposed algorithm compared with reference papers [1], [2]. The maximum improvement of our proposed algorithm compared with the reference paper [1] is 10.79%, and the maximum improvement of our proposed algorithm compared with the reference paper [2] is 38.552%.

System Model

Consider the received signal in flat fading channel as:

$$r_l(n) = \alpha w_l(n) + v(n) \tag{1}$$

where α is the complex channel fading coefficient which is considered $\alpha \in CN(0,1)$, $w_l(n)$ is the transmitted symbol which is considered an independent and identically distributed (i.i.d) process, and $v(n)$ is the additive white Gaussian noise (AWGN) and is considered $v(n) \in CN(0, \sigma_n^2)$.

The general mathematical form of the HOC is defined as [14], [15]:

$$C_{p,q} = Cum\left[\overbrace{r_1,...,r_{p-q}}^{p-q\,terms}, \overbrace{r^*_{p-q+1},...,r^*_p}^{q\,terms}\right] \tag{2}$$

where * denotes the complex conjugate, *p* is the order of the cumulant, *q* is the complex conjugate order of the cumulant, and *cum* function is defined as [14]:

$$Cum[r_1,...,r_n] = \sum_{\forall v}(-1)^{q-1}(q-1)!E\left[\prod_{j\in V_1} r_j\right]...E\left[\prod_{j\in V_q} r_j\right] \tag{3}$$

and the summation is being performed on all partitions $V = (V_1, V_2, ..., V_q)$ for the set of indexes $(1, 2, ..., n)$.

To cancel the effect of the power level of the received signal, the first type of normalization must be done [15], [16]:

$$C'_{pq} = \frac{C_{pq}}{(C_{21})^{p/2}} \tag{4}$$

The magnitude of the eighth, sixth, and fourth-order cumulants is greater than that of the second-order cumulants. As a result, we have different values for the other HOC orders. The second normalization can reduce the values range as [15], [17]:

$$\tilde{C}_{pq} = (C'_{pq})^{2/p} \tag{5}$$

According to our simulation results for the selected digital modulation types in Table 1, $\tilde{C}_{40}, \tilde{C}_{61}$, and $\tilde{C}_{80}$ (equations (6), (7), and (8)) have the most discrimination ability, so they have been chosen in our study [15], [18], [19]:

$$C_{40} = M_{40} - 3M_{20}^2 \tag{6}$$

$$C_{61} = M_{61} - 5M_{21}M_{40} - 10M_{20}M_{41} + 30M_{20}^2M_{21} \tag{7}$$

$$C_{80} = M_{80} - 35M_{40}^2 - 28M_{60}M_{20} + 420M_{20}^2M_{40} - 630M_{20}^4 \tag{8}$$

where [15], [18], [19]:

$$M_{pq} = E\left[r(k)^{p-q}\, r^*(k)^q\right] \tag{9}$$

is the moment of received signal *r*(*k*).

Mathematical Preliminary

A. Linear Discriminant Analysis (LDA)

This technique finds the optimum linear projection vector that maximizes the discrimination between digital modulation types [20]. We define the input features of the two classes for dataset *i* as:

$$\boldsymbol{x}_i = \left[\left(\tilde{C}_{40}(r_l)\right)_i \quad \left(\tilde{C}_{61}(r_l)\right)_i \quad \left(\tilde{C}_{80}(r_l)\right)_i\right]^T, r_l \in \boldsymbol{C}_x \ , \ i = 1..n_x \tag{10}$$

$$\boldsymbol{y}_i = \left[\left(\tilde{C}_{40}(r_l)\right)_i \quad \left(\tilde{C}_{61}(r_l)\right)_i \quad \left(\tilde{C}_{80}(r_l)\right)_i\right]^T, r_l \in \boldsymbol{C}_Y \ , \ i = 1,...,n_y \tag{11}$$

These input features can be written as:

$$\boldsymbol{x}_i := \left[\tilde{C}_{40,x_i} \quad \tilde{C}_{61,x_i} \quad \tilde{C}_{80,x_i}\right]^T \tag{12}$$

$$\boldsymbol{y}_i := \left[\tilde{C}_{40,y_i} \quad \tilde{C}_{61,y_i} \quad \tilde{C}_{80,y_i}\right]^T \tag{13}$$

The mean vectors of the input features can be calculated as [20]:

$$\boldsymbol{\mu}_x = \left[E\left(\tilde{C}_{40,x_i}\right) \quad E\left(\tilde{C}_{61,x_i}\right) \quad E\left(\tilde{C}_{80,x_i}\right)\right]^T := \left[\mu_{x,1} \quad \mu_{x,2} \quad \mu_{x,3}\right]^T \tag{14}$$

$$\boldsymbol{\mu}_y = \left[E\left(\tilde{C}_{40,y_i}\right) \quad E\left(\tilde{C}_{61,y_i}\right) \quad E\left(\tilde{C}_{80,y_i}\right)\right]^T := \left[\mu_{y,1} \quad \mu_{y,2} \quad \mu_{y,3}\right]^T \tag{15}$$

By defining the projection vector $\boldsymbol{u}$, the output features $x_i^{'}$, $y_i^{'}$ of the first and second classes respectively can be calculated as [20]:

$$x_i^{'} = \boldsymbol{x}_i^T \boldsymbol{u}, \ y_i^{'} = \boldsymbol{y}_i^T \boldsymbol{u} \tag{16}$$

Fisher criterion function represents the discrimination measurement between the output features of two classes as [20], [21]:

$$J(\boldsymbol{u}) := \frac{\left(\mu_y^{'}(\boldsymbol{u}) - \mu_x^{'}(\boldsymbol{u})\right)^2}{\sigma_x^{'2}(\boldsymbol{u}) + \sigma_y^{'2}(\boldsymbol{u})} = \frac{\boldsymbol{u}^T \mathbf{S}_\mathbf{B} \boldsymbol{u}}{\boldsymbol{u}^T \mathbf{S}_\mathbf{W} \boldsymbol{u}} \in \mathbb{R} \tag{17}$$

where $\boldsymbol{u}$ is the projection vector, $\mu_x^{'}(\boldsymbol{u}), \mu_y^{'}(\boldsymbol{u})$ are the means of the output features for the first and the second classes, respectively, $\sigma_x^{'2}(\boldsymbol{u}), \sigma_y^{'2}(\boldsymbol{u})$ are the variances of the output features for the first and second classes, respectively. $\mathbf{S}_\mathbf{B}$ And $\mathbf{S}_\mathbf{W}$ are defined as:

$$\mathbf{S}_\mathbf{B} = (\boldsymbol{\mu}_y - \boldsymbol{\mu}_x)(\boldsymbol{\mu}_y - \boldsymbol{\mu}_x)^T \in \mathbb{R}^{d \times d} \tag{18}$$

$$\mathbf{S}_\mathbf{W} := \frac{1}{n_x - 1} \sum_{i=1}^{n_x} (\boldsymbol{x}_i - \boldsymbol{\mu}_x)(\boldsymbol{x}_i - \boldsymbol{\mu}_x)^T + \frac{1}{n_y - 1} \sum_{i=1}^{n_y} (\boldsymbol{y}_i - \boldsymbol{\mu}_y)(\boldsymbol{y}_i - \boldsymbol{\mu}_y)^T \in \mathbb{R}^{d \times d} \tag{19}$$

where n_x, n_y are the numbers of samples for the two classes, respectively. The optimum projection vector $\boldsymbol{u}$ can be calculated by solving the maximization of the Fisher criterion function problem of (17) for $\boldsymbol{u}$. One of the solutions is using the Lagrange method as [21]:

$$L = \boldsymbol{u}^T \mathbf{S}_\mathbf{B} \boldsymbol{u} - \lambda(\boldsymbol{u}^T \mathbf{S}_\mathbf{W} \boldsymbol{u} - 1) \tag{20}$$

where λ is the Lagrange multiplier. Equating the derivative of L to zero gives [21]:

$$\frac{\partial L}{\partial \boldsymbol{u}} = 2\mathbf{S}_\mathbf{B} \boldsymbol{u} - 2\lambda \mathbf{S}_\mathbf{W} \boldsymbol{u} = 0 \Rightarrow \mathbf{S}_\mathbf{B} \boldsymbol{u} = \lambda \mathbf{S}_\mathbf{W} \boldsymbol{u} \tag{21}$$

which is a generalized eigenvalue problem. One possible solution to the above-generalized eigenvalue problem can be found as [21]:

$$\boldsymbol{u} = eig(\mathbf{S}_\mathbf{W}^{-1} \mathbf{S}_\mathbf{B}) \tag{22}$$

where *eig*(.) denotes the eigenvector of the matrix with the largest eigenvalue.

In the following Sections, the LDA algorithm is called the classical LDA.

B. Discrimination measurement

One of the well-known statistical distance measurements between two random variables is Mahalanobis Distance (MD). Suppose $\boldsymbol{v}_x$ and $\boldsymbol{v}_y$ are random variables. The MD distance between them can be calculated as [22]:

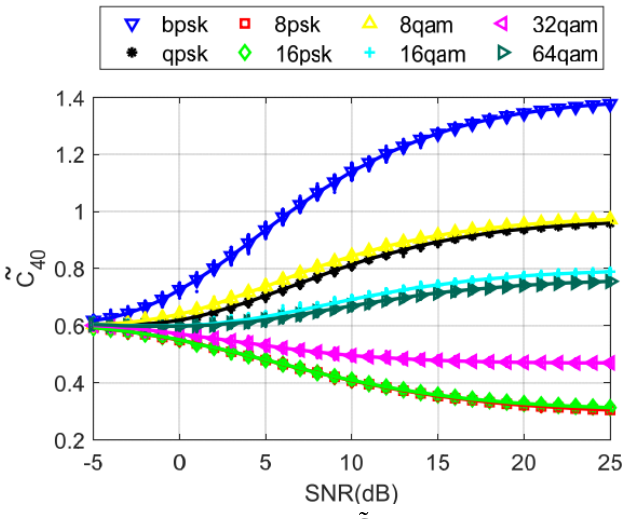
$$d(\boldsymbol{v}_x, \boldsymbol{v}_y) = \sqrt{(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y)^T (\mathbf{S}_\mathbf{x} + \mathbf{S}_\mathbf{y})^{-1} (\boldsymbol{\mu}_x - \boldsymbol{\mu}_y)} \in \mathbb{R} \tag{23}$$

where $\boldsymbol{\mu}_x, \boldsymbol{\mu}_y$ are mean vectors and $\mathbf{S}_\mathbf{x}, \mathbf{S}_\mathbf{y}$ are the covariance matrices of the random variables $\boldsymbol{v}_x, \boldsymbol{v}_y$, respectively.

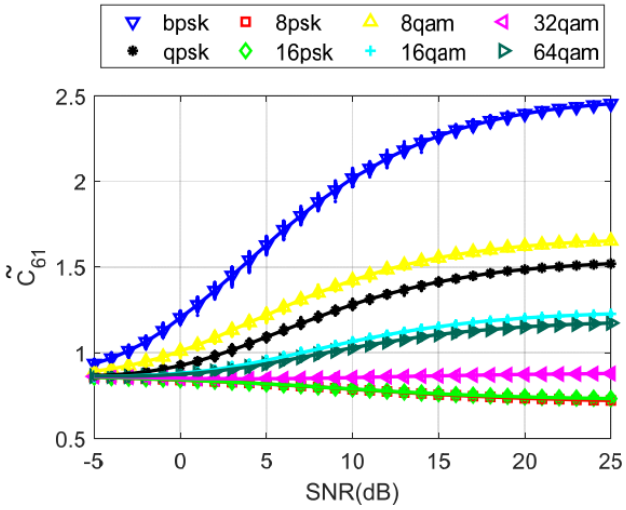
This study uses the MD as a discrimination measurement between two random variables.

Conventional Classical LDA-based AMC Problem

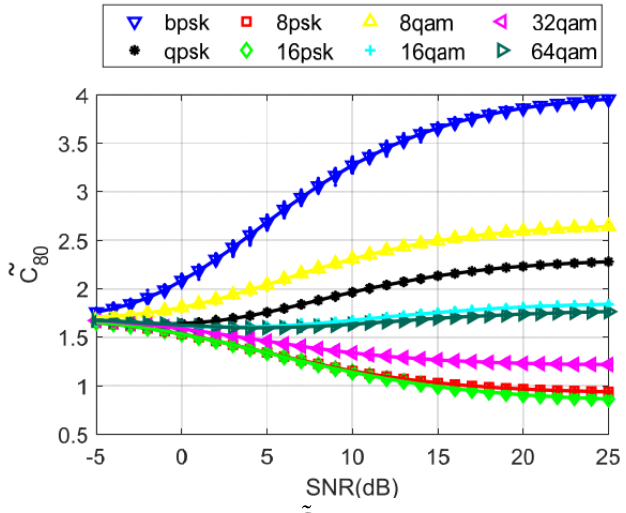
The values of the selected HOCs in Section 2, i.e. $\tilde{C}_{40}, \tilde{C}_{61}, \tilde{C}_{80}$, are shown in Fig. 1, for the selected digital modulations in Table 1, and SNR rang [-5:25] dB.



(a) $\tilde{C}_{40}$



(b) $\tilde{C}_{61}$



(c) $\tilde{C}_{80}$

Fig. 1: Values of $\tilde{C}_{40}$, $\tilde{C}_{61}$, and $\tilde{C}_{80}$ cumulants respectively.

From Fig. 1, some modulations can be classified easily (like BPSK, QPSK, 8QAM, and 32QAM). In contrast, the others are close to each other (like 8PSK and 16PSK, and like 16QAM and 64QAM). This situation would be worse for lower SNR values.

We define two different problems:

- 8PSK and 16PSK classification as problem p1.
- 16QAM and 64QAM classification as problem p2.

Our work aims to find a new algorithm that separately improves the classification accuracy for the two problems, p1 and p2.

Start with classical LDA to solve the mentioned problems p1 and p2. Calculation of the classification accuracy (ACC) for the problems p1 and p2 using the selected HOCs in Section 2 and the classical LDA algorithm have been done as shown in Fig. 2. As shown in Fig. 2, classical LDA doesn't improve the performance accuracy of 8PSK and 16PSK classification (problem p1) and 16QAM and 64QAM classification (problem p2). As shown in the next section, we propose modifying the classical LDA algorithm by adding an optimizable nonlinear preprocess.

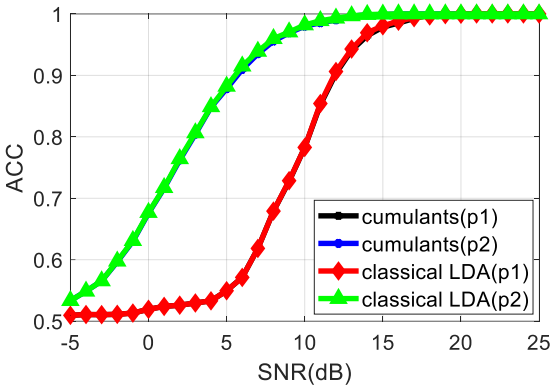


Fig. 2: Classification accuracy using the selected HOCs and the classical LDA for the problems p1 and p2.

Proposed Preprocess LDA Algorithm

The proposed preprocess LDA algorithm consists of an optimizable nonlinear preprocess, followed by the LDA algorithm.

A. The Proposed Mathematical Problem

The selected HOCs can be rewritten as: $C_{x_i,1} := \tilde{C}_{40,x_i}$, $C_{x_i,2} := \tilde{C}_{61,x_i}$, $C_{x_i,3} := \tilde{C}_{80,x_i}$ for the first class C_x , the input features vector becomes $\left[C_{x_i,1} \quad C_{x_i,2} \quad C_{x_i,3}\right]^T$, and $C_{y_i,1} := \tilde{C}_{40,y_i}$, $C_{y_i,2} := \tilde{C}_{61,y_i}$, $C_{y_i,3} := \tilde{C}_{80,y_i}$ for the second class C_y , the input features vector becomes $\left[C_{y_i,1} \quad C_{y_i,2} \quad C_{y_i,3}\right]^T$.

As shown in Section 4, the classical LDA algorithm needs adjustment to improve the discrimination between 8PSK and 16PSK, and between 16QAM and 64QAM modulations for low SNR values. An example of this adjustment is the addition of nonlinear function as follows: $f_1(C_{x_i,1} \text{ or } C_{y_i,1})$, $f_2(C_{x_i,2} \text{ or } C_{y_i,2})$, and $f_3(C_{x_i,3} \text{ or } C_{y_i,3})$ of the selected HOCs in Section 2 as shown in Fig. 3.

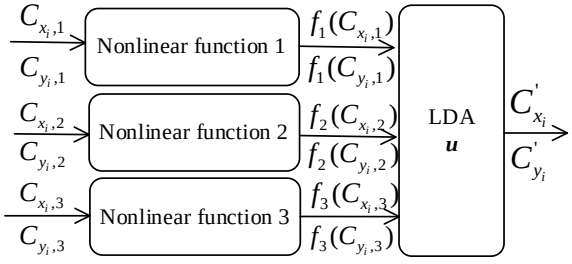


Fig. 3: Block diagram of the proposed nonlinear preprocess LDA algorithm.

where C'_{x_i}, C'_{y_i} are the output features of the first and second classes, respectively. These output features can be calculated as:

$$C_{x_i}^{'} = u_1 f_1(C_{x_i,1}) + u_2 f_2(C_{x_i,2}) + u_3 f_3(C_{x_i,3}) = \boldsymbol{u}^T \boldsymbol{f}_{x_i} \quad (24)$$

$$C_{y_i}^{'} = u_1 f_1(C_{y_i,1}) + u_2 f_2(C_{y_i,2}) + u_3 f_3(C_{y_i,3}) = \boldsymbol{u}^T \boldsymbol{f}_{y_i} \quad (25)$$

where $\boldsymbol{u} = [u_1, u_2, u_3]^T$ is the projection vector, $\boldsymbol{f}_{x_i} = \left[f_1(C_{x_i,1}) \quad f_2(C_{x_i,2}) \quad f_3(C_{x_i,3}) \right]^T := \left[f_{x_i,1} \quad f_{x_i,2} \quad f_{x_i,3} \right]^T$ is the vector of the values of the nonlinear functions for the first class, and: $\boldsymbol{f}_{y_i} = \left[f_1(C_{y_i,1}) \quad f_2(C_{y_i,2}) \quad f_3(C_{y_i,3}) \right]^T := \left[f_{y_i,1} \quad f_{y_i,2} \quad f_{y_i,3}) \right]^T$ is the vector of the values of the nonlinear functions for the second class.

The terms $\mu_x^{'}(\boldsymbol{u}), \mu_y^{'}(\boldsymbol{u})$ of the Fisher criterion function (17) are the means of the output features. They are calculated using (24) and (25) as:

$$\mu_x^{'} = \boldsymbol{u}^T E(\boldsymbol{f}_x) \quad (26)$$

$$\mu_y^{'} = \boldsymbol{u}^T E(\boldsymbol{f}_y) \quad (27)$$

The terms $\sigma_1^{'2}(\boldsymbol{u}), \sigma_2^{'2}(\boldsymbol{u})$ of the Fisher criterion function (17) are the variances of the output features. They are calculated as:

$$(\sigma_x^{'})^2 = \boldsymbol{u}^T cov(\boldsymbol{f}_x)\boldsymbol{u} \quad (28)$$

$$(\sigma_y^{'})^2 = \boldsymbol{u}^T cov(\boldsymbol{f}_y)\boldsymbol{u} \quad (29)$$

By using (26), (27), (28), and (29), the Fisher criterion (17) can be written as:

$$J(\boldsymbol{u}) = \frac{(\mu_y^{'}(\boldsymbol{u}) - \mu_x^{'}(\boldsymbol{u}))^2}{\sigma_y^{'2}(\boldsymbol{u}) + \sigma_x^{'2}(\boldsymbol{u})} = \frac{\boldsymbol{u}^T \left[E(\boldsymbol{f}_y) - E(\boldsymbol{f}_x) \right] \left[E(\boldsymbol{f}_y) - E(\boldsymbol{f}_x) \right]^T \boldsymbol{u}}{\boldsymbol{u}^T \left[cov(\boldsymbol{f}_x) + cov(\boldsymbol{f}_y) \right] \boldsymbol{u}} := \frac{\boldsymbol{u}^T \mathbf{S}_\mathbf{B} \boldsymbol{u}}{\boldsymbol{u}^T \mathbf{S}_\mathbf{W} \boldsymbol{u}} \quad (30)$$

where $\mathbf{S}_\mathbf{B} := \left[E(\boldsymbol{f}_y) - E(\boldsymbol{f}_x) \right] \left[E(\boldsymbol{f}_y) - E(\boldsymbol{f}_x) \right]^T$ and $\mathbf{S}_\mathbf{W} := \left[cov(\boldsymbol{f}_x) + cov(\boldsymbol{f}_y) \right]$.

The nonlinear preprocess function allows us to control $\mathbf{S}_\mathbf{B}, \mathbf{S}_\mathbf{W}$, which affects the discrimination performance.

The task is to find the rules that maximize the discrimination between two classes (Fisher criterion (30)) using the nonlinear preprocesses.

B. Necessary Lemmas

Lemma 1. For *J*, $\mathbf{S}_\mathbf{B}, \mathbf{S}_\mathbf{W}$ which are defined in (30) and (17), we find that:

$$\max(J) = \text{trace}(\mathbf{S}_\mathbf{w}^{-1} \mathbf{S}_\mathbf{B}) \quad (31)$$

Proof: Here, we mention some mathematical analyzes and results:

- The maximum value of the Fisher criterion function (17) is equal to the maximum value of eigenvalues of the matrix $\mathbf{S}_\mathbf{w}^{-1} \mathbf{S}_\mathbf{B}$ [23]:

$$\max(J) = \lambda_{\max}(\mathbf{S}_\mathbf{w}^{-1} \mathbf{S}_\mathbf{B}) \quad (32)$$

- The summation of eigenvalues of a matrix is equal to the trace of the matrix [24]:

$$\text{trace}(\mathbf{S}_\mathbf{w}^{-1} \mathbf{S}_\mathbf{B}) = \sum \lambda_i \quad (33)$$

- The production of eigenvalues of a matrix is equal to the determinant of the matrix [24]:

$$\det(\mathbf{S}_\mathbf{w}^{-1} \mathbf{S}_\mathbf{B}) = \prod \lambda_i \quad (34)$$

- By noticing $\mathbf{S}_\mathbf{B}$ calculation in (18), we find that $\det(\mathbf{S}_\mathbf{B})$ is equal to zero.
- $\det(\mathbf{S}_\mathbf{w}^{-1} \mathbf{S}_\mathbf{B}) = \det(\mathbf{S}_\mathbf{w}^{-1}) \det(\mathbf{S}_\mathbf{B}) = 0$, which means (34) is no longer helpful for calculating λ_i.
- According to [21], the rank of $\mathbf{S}_\mathbf{W}^{-1} \mathbf{S}_\mathbf{B}$ can be calculated as:

$$\text{rank}(\mathbf{S}_\mathbf{W}^{-1} \mathbf{S}_\mathbf{B}) = \min(T, n-1, L-1) \quad (35)$$

where *n* is the size of the dataset in each class, *L* is the number of classes, and *T* is the number of features. In our case, *L*=2, *T*=3, and *n*>>*L*, *T*. We find that the $\text{rank}\left(\mathbf{S}_\mathbf{W}^{-1} \mathbf{S}_\mathbf{B}\right)$ value is equal to 1. Which means we have *one nonzero eigenvalue*. By using (33) we find that:

$$\text{trace}(\mathbf{S}_\mathbf{w}^{-1} \mathbf{S}_\mathbf{B}) = \lambda \neq 0 \quad (36)$$

- Finally, by using (32) and (36), we find that $\max(J) = \text{trace}(\mathbf{S}_\mathbf{w}^{-1} \mathbf{S}_\mathbf{B})$.

Lemma 2: Maximization of the Fisher criterion (30) is equivalent to maximization of the Mahalanobis distance between the values of the nonlinear functions for each feature, i.e Fig. 3.

Proof: According to Lemma 1, maximization of the Fisher criterion function means maximization of $\text{trace}(\mathbf{S}_\mathbf{W}^{-1} \mathbf{S}_\mathbf{B})$. So, we have to study the effect of $\mathbf{S}_\mathbf{B}$ and $\mathbf{S}_\mathbf{W}$ elements on the Fisher criterion function. To simplify it, we study two-class cases where $\mathbf{S}_\mathbf{B}, \mathbf{S}_\mathbf{W} \in \mathbb{R}^{2\times 2}$:

$$\mathbf{S}_\mathbf{B} = \left[\boldsymbol{\mu}_y - \boldsymbol{\mu}_x \right] \left[\boldsymbol{\mu}_y - \boldsymbol{\mu}_x \right]^T = \begin{bmatrix} \Delta_1^2 & \Delta_1 \Delta_2 \\ \Delta_1 \Delta_2 & \Delta_2^2 \end{bmatrix} \in \mathbb{R}^{2\times 2} \quad (37)$$

where $\Delta_1 = \mu_{y,1} - \mu_{x,1}$ is the difference between the means of the two classes for the first feature and $\Delta_2 = \mu_{y,2} - \mu_{x,2}$ is the difference between the means of the two classes for the second feature.

and:

$$\mathbf{S_W} = \mathrm{cov}(\boldsymbol{x}) + \mathrm{cov}(\boldsymbol{y}) = \begin{bmatrix} (\sigma_{x,1})^2 & \rho_x \sigma_{x,1}\sigma_{x,2} \\ \rho_x \sigma_{x,1}\sigma_{x,2} & (\sigma_{x,2})^2 \end{bmatrix} + \begin{bmatrix} (\sigma_{y,1})^2 & \rho_y \sigma_{y,1}\sigma_{y,2} \\ \rho_y \sigma_{y,1}\sigma_{y,2} & (\sigma_{y,2})^2 \end{bmatrix} = \begin{bmatrix} g_1^2 + h_1^2 & \rho_x g_1 g_2 + \rho_y h_1 h_2 \\ \rho_x g_1 g_2 + \rho_y h_1 h_2 & g_2^2 + h_2^2 \end{bmatrix} \quad (38)$$

where $g_1 = \sigma_{x,1}$ and $g_2 = \sigma_{x,2}$ are the variances of the first and second features for the first class, $h_1 = \sigma_{y,1}$ and $h_2 = \sigma_{y,2}$ are the variances of the first and second features for the second class. ρ_x And ρ_y are the correlations between the features of the first class and second class, respectively. By using (37) and (38), the trace of $\mathbf{S_w^{-1}S_B}$ can be calculated as:

$$tr = trace(\mathbf{S_w^{-1}S_B}) = \frac{\Delta_1^2\left(g_2^2 + h_2^2\right) + \Delta_2^2\left(g_1^2 + h_1^2\right) - 2\Delta_1\Delta_2\left(\rho_x g_1 g_2 + \rho_y h_1 h_2\right)}{\left(g_1^2 + h_1^2\right)\left(g_2^2 + h_2^2\right) - \left(\rho_x g_1 g_2 + \rho_y h_1 h_2\right)^2} = \frac{\Delta_1^2\left(g_2^2 + h_2^2\right) + \Delta_2^2\left(g_1^2 + h_1^2\right) - 2\Delta_1\Delta_2\left(\rho_x g_1 g_2 + \rho_y h_1 h_2\right)}{\left(g_1^2 + h_1^2\right)\left(g_2^2 + h_2^2\right)\left[1 - \dfrac{\left(\rho_x g_1 g_2 + \rho_y h_1 h_2\right)^2}{\left(g_1^2 + h_1^2\right)\left(g_2^2 + h_2^2\right)}\right]} = \lambda \neq 0 \quad (39)$$

Discussion: Starting with $\mathbf{S_B}$, since Δ_1 and Δ_2 are the differences between the means of the first and second features for the two classes, respectively, we find from (39) that:

$$\lim_{\Delta_1 \to \mp\infty}(tr) = +\infty \,,\ \lim_{\Delta_2 \to \mp\infty}(tr) = +\infty \quad (39)$$

which means, by increasing the absolute values of the variables Δ_1 and Δ_2, the Fisher criterion function value (17) will increase and vice versa.

To study the effect of $\mathbf{S_W}$ elements on the Fisher criterion function, by noticing that g_1, h_1 are the variances of the first feature for the two classes, we find from (39) that:

$$\begin{aligned} &\text{Since}: \quad g_1 \geq 0 \,,\ h_1 \geq 0 \\ &\text{When}: \quad g_1 + h_1 \to 0 \Leftrightarrow g_1 \to 0 \ \textit{and}\ h_1 \to 0 \\ &\text{Thus}: \quad \lim_{\substack{g_1 \to 0 \\ h_1 \to 0}}(tr) = +\infty \end{aligned} \quad (40)$$

In the same way for g_2, h_2, we find from (39) that:

$$\begin{aligned} &\text{Since}: \quad g_2 \geq 0 \,,\ h_2 \geq 0 \\ &\text{When}: \quad g_2 + h_2 \to 0 \Leftrightarrow g_2 \to 0 \ \textit{and}\ h_2 \to 0 \\ &\text{Thus}: \quad \lim_{\substack{g_2 \to 0 \\ h_2 \to 0}}(tr) = +\infty \end{aligned} \quad (41)$$

which means, by decreasing the values of the variables $g_1 + h_1$ and $g_2 + h_2$, the Fisher criterion function value (17) will increase and vice versa.

From (40), (41), and (42), we find that to maximize the Fisher criterion function, we have to maximize Δ_1, Δ_2 of the matrix $\mathbf{S_B}$, and minimize $g_1 + h_1$, $g_2 + h_2$ of the matrix $\mathbf{S_W}$. The same thing must have been done for $\mathbf{S_B}$ and $\mathbf{S_W}$ in (30).

By defining the means and variances of functions values of features as follows: $\mu_{f,x,i} := E\left(f_{x,i}\right)$, $\mu_{f,y,i} := E\left(f_{y,i}\right)$, $\left(\sigma_{f,x,i}\right)^2 := \mathrm{var}\left(f_{x,i}\right)$, $\left(\sigma_{f,y,i}\right)^2 := \mathrm{var}\left(f_{y,i}\right)$. Maximizing the elements of the matrix $\mathbf{S_B}$ means finding the optimum nonlinear transformation which satisfies:

$$f_i = \underset{f_i}{\mathrm{argmax}}\left(E\left(f_{y,i}\right) - E\left(f_{x,i}\right)\right)^2 = \underset{f_i}{\mathrm{argmax}}\left(\mu_{f,y,i} - \mu_{f,x,i}\right)^2 \,,\ i = 1,2,3 \quad (42)$$

Minimizing the elements of the matrix $\mathbf{S_W}$ means finding the optimum nonlinear transformation which satisfies:

$$f_i = \underset{f_i}{\mathrm{argmin}}\left(\mathrm{var}\left(f_{x,i}\right) + \mathrm{var}\left(f_{y,i}\right)\right) = \underset{f_i}{\mathrm{argmin}}\left(\left(\sigma_{f,x,i}\right)^2 + \left(\sigma_{f,y,i}\right)^2\right),\ i = 1,2,3 \quad (43)$$

By combining (43) and (44) we find:

$$f_{\substack{i \\ i=1,2,3}} = \underset{f_i}{\mathrm{optimum}}\left\{\begin{matrix} \mathrm{argmax}\left(\mu_{f,y,i} - \mu_{f,x,i}\right)^2 \\ \mathrm{argmin}\left(\left(\sigma_{f,x,i}\right)^2 + \left(\sigma_{f,y,i}\right)^2\right) \end{matrix}\right\} \quad (44)$$

By noticing (23), for each feature *i*=1,2,3, we find that maximizing the MD is equivalent to the condition (45) as:

$$f_i = \underset{f_i}{\mathrm{argmax}}(MD(f_{x,i}, f_{y,i})) \,,\ i = 1,2,3 \quad (45)$$

which is the same as the condition denoted in lemma 2.

C. The Solution to The Proposed Mathematical Problem

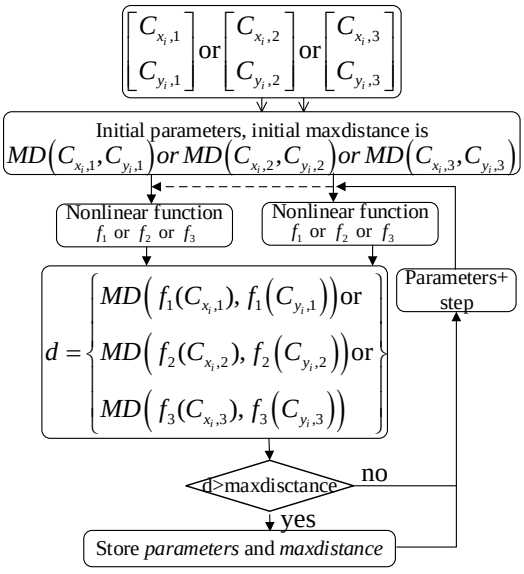


Fig. 4: Optimization of nonlinear preprocess.

Lemma 2 means, to maximize the Fisher criterion, we have to choose the optimum parameters that satisfy (46) for each feature. We propose a simple search algorithm to find these optimum parameters for each selected HOC in Section 2. The proposed algorithm is depicted in Fig. 4.

The proposed optimal nonlinear preprocess LDA algorithm of Fig. 4 consists of two stages where each stage consists of two steps (see Fig. 5):

a- Training stage:

Step 1:

In this step, we calculate the parameters of the nonlinear preprocess (47), (49), and (50):

- Calculate the parameters of the nonlinear preprocess (f_1) for feature1 as $C_{x_i,1}, C_{y_i,1}$.
- Calculate the parameters of the nonlinear preprocess (f_2) for feature2 as $C_{x_i,2}, C_{y_i,2}$.
- Calculate the parameters of the nonlinear preprocess (f_3) for feature3 as $C_{x_i,3}, C_{y_i,3}$.

Step 2:

- Calculate the linear projection vector $\boldsymbol{u}$ by solving the Eigenvalue problem (22) for $\boldsymbol{S_B}$ and $\boldsymbol{S_W}$ where is presented in (30).

b- Testing stage:

Step 1:

In this step, we apply the nonlinear preprocess (47), (49), and (50) using the calculated parameters in the previous stage as:

- Apply the nonlinear preprocess (f_1) for the feature1, i.e. $C_{x\ or\ y,1}$, using their calculated parameters.
- Apply the nonlinear preprocess (f_2) for the feature2, i.e. $C_{x\ or\ y,2}$, using their calculated parameters.
- Apply the optimized nonlinear preprocess (f_3) for the feature3, i.e. $C_{x\ or\ y,3}$, using their calculated parameters.

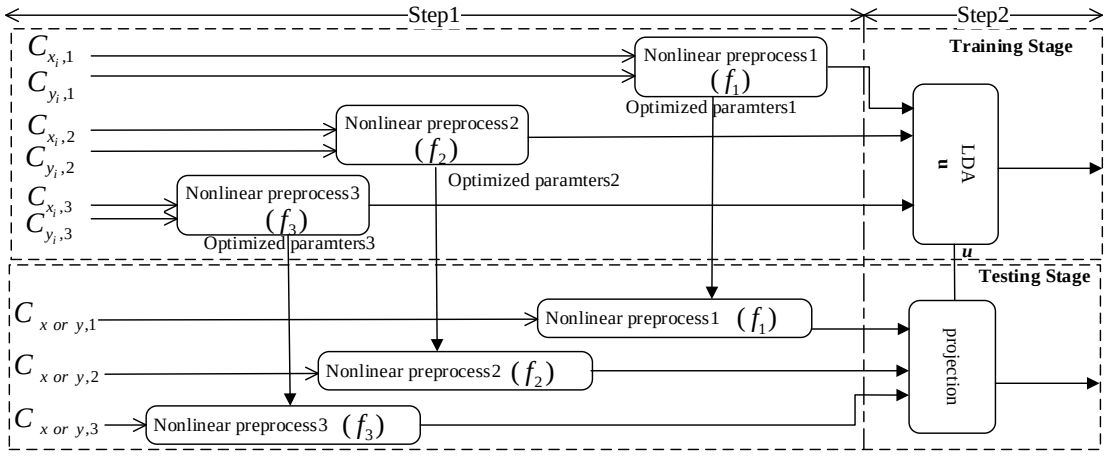


Fig. 5: Optimal nonlinear preprocess LDA-based algorithm.

Step 2:

- Apply the linear projection as (16) using the calculated projection vector in the previous stage.

Two general optimal nonlinear preprocesses have been studied here: regularized distance-based and optimized nonlinear transformation preprocesses.

D. Regularized Distance-Based Preprocess

To improve the discrimination between the classes, the distance between these features and the total mean is added to them as:

$$f_j\left(C_{x_i,j}\right) = C_{x_i,j} + (\Delta_{x_i,j} + \xi_j)^{-1}(C_{x_i,j} - \mu_j)$$
$$; i = 1...n, j = 1..d$$
$$f_j\left(C_{y_i,j}\right) = C_{y_i,j} + (\Delta_{y_i,j} + \xi_j)^{-1}(C_{y_i,j} - \mu_j) \quad (46)$$
$$; i = 1...n, j = 1..d$$

where

$$\Delta_{x_i,j} = (C_{x_i,j} - \mu_j)^2 \ , \ \Delta_{y_i,j} = (C_{y_i,j} - \mu_j)^2 \quad (47)$$

is the distance between the feature *j* (*j*=1, 2, or 3) and the total mean of the two classes $\mu_j = \frac{\mu_{x,j} + \mu_{y,j}}{2}$ of the feature *j*, $C_{x_i,j}$ is the input feature *j* of the first class, $C_{y_i,j}$ is the input feature *j* of the second class, $f_j(C_{x_i,j})$ is the regularized distance-based feature value of the first digital modulation type, $f_j(C_{y_i,j})$ is the regularized distance-based feature value of the second class, and ξ_j is the regularizer of the feature *j*. This regularizer aims to optimize this nonlinear transformation according to (46). We call it the proposed-dist LDA algorithm.

E. optimized nonlinear transformation

Another way to find an optimal nonlinear preprocess that satisfies (46), is to add some parameters (here we add two parameters like L_1, L_2) to some known nonlinear transformations. Two nonlinear transformations are used, Box-Cox [25] and tangent hyperbolic (tanh) transformations a [26]:

Box-Cox transformation is defined as [25]:

$$f_j(C_{x_i,j},L_1,L_2)=\begin{cases}\dfrac{(C_{x_i,j}+L_2)^{L_1}-1}{L_1} \; if \;\; L_1\neq 0\\ \log(C_{x_i,j}+L_2) \; if \; L_1=0\end{cases}$$
$$f_j(C_{y_i,j},L_1,L_2)=\begin{cases}\dfrac{(C_{y_i,j}+L_2)^{L_1}-1}{L_1} \; if \;\; L_1\neq 0\\ \log(C_{y_i,j}+L_2) \; if \; L_1=0\end{cases} \quad (48)$$

We call it the proposed-Box LDA algorithm.

Tangent hyperbolic (tanh) transformation can be defined as [26]:

$$f_j(C_{x_i,j},L_1,L_2)=\tanh(L_1(C_{x_i,j}+L_2))$$
$$f_j(C_{y_i,j},L_1,L_2)=\tanh(L_1(C_{y_i,j}+L_2)) \quad (49)$$

We call it the proposed-Tanh LDA algorithm.

The value of parameter L_2 for each feature is close to the total mean of feature $\mu_j=\frac{\mu_{x,j}+\mu_{y,j}}{2}$. This parameter can be determined quickly using the search algorithm in Fig. 4. While the value of L_1 depends on the feature values and the transformation function, it can be determined by using the search algorithm in Fig. 4. Still, it takes more time than itself for L_2 determination.

Time and Space Complexities

Here, we analyze the time and space complexities of the LDA and the proposed algorithms for the training and test stages. Then we compare them.

A. The Classical LDA Algorithm's Time and Space Complexities

To calculate time and space complexities, we suppose that the number of samples is $n_j=n$ for all classes *c* (*c*=2 in our case), and *d* is the number of features (*d*=3 in our case). Starting with training complexity, we find [20], [27]-[29]:

Table 2: Time and space complexities for training the classical LDA algorithm

Operation	Time complexity	Space complexity
$\{\boldsymbol{x}_i\}_{i=1}^n,\{\boldsymbol{y}_i\}_{i=1}^n$	*0*	*ncd=6n*
$\boldsymbol{\mu}_x,\boldsymbol{\mu}_y$	*cd(n+1)= 6(n+1)*	*cd=6*
$\boldsymbol{\mu}$	*ncd+d=6n+3*	*d=3*
$\mathbf{S_B}$	$cd^2+cd=24$	$2d^2=18$
$\mathbf{S}_W$	$ncd^2+ncd=24n$	$2d^2=18$
$\mathbf{S_W^{-1}}$	$O(d^3)=27$ [29]	$O(d^2)=9$ [28]
$\mathbf{S_W^{-1}S_B}$	$O(d^3)=27$ [29]	$d^2=9$
$\mathbf{u}=eig\left(\mathbf{S_W^{-1}S_B}\right)$	$O(d^3)=27$ [20]	$O(d^2)=9$ [30]
Final complexity	*36n+114*	*6n+72*
Our case $n\gg 100$	*36n*	*6n*

This result is similar to the result in [27] and for testing complexity, we find:

Table 3: Time and Space complexities for testing the classical LDA algorithm

Operation	Time complexity	Space complexity
$\{\boldsymbol{x}_i\}_{i=1}^n,\{\boldsymbol{y}_i\}_{i=1}^n$	*0*	*6n*
LDA projection	*12n*	*2n*
Final complexity	*12n*	*8n*

B. The Proposed Algorithm's Time and Space Complexities

Similar to the previous Section, we must calculate the training and testing complexities. According to the proposed nonlinear functions, the maximum number of optimizable variables is two. Suppose *a* and *b* are the number of loops for the first and second variables. Starting with training complexity, we find:

Table 4: Time and Space complexities for training the proposed algorithm

Operation	Time complexity	Space complexity
Apply nonlinear preprocess	*2n*	*2n*
$\boldsymbol{\mu}_x,\boldsymbol{\mu}_y$	*6(n+1)*	*6*
$\mathbf{S}_W$	*18n*	*6*
Calculate MD using $\boldsymbol{\mu}_x,\boldsymbol{\mu}_y,\mathbf{S}_W$	*12*	*3*
Total complexity of one-time preprocess	*10n+6*	*2n+15*
Repeat for the first variable *a* times	*a(26n+18)*	*2n+6*
Repeat for the second variable *b* times	*ab(26n+18)*	*2n+6*
Apply LDA	36*n*	6*n*
Final complexity	*ab(26n+18)+36n*	8*n*+6
Our case $n\gg 100$	*26abn+36n*	8*n*

and for testing complexity, we find:

Table 5: Time and Space complexities for testing the proposed algorithm

Operation	Time complexity	Space complexity
$\{\boldsymbol{x}_i\}_{i=1}^n,\{\boldsymbol{y}_i\}_{i=1}^n$	*0*	*2n*
Apply nonlinear preprocess	*2n*	*2n*
LDA projection	*12n*	*8n*
Final complexity	*14n*	*12n*

C. A Comparison Between the Complexities of the Classical LDA and the Proposed Algorithm

Calculating the ratio of the proposed algorithm's complexity over the classical LDA algorithm's complexity is done to compare their complexities, as shown in Table 6.

Table 6: The ratio of the complexity of the proposed algorithm over the complexity of the classical LDA algorithm

stage	Ratio of time complexity	Ration of space complexity
training	*≈ab+1*	*1.25*
testing	*1.17*	*1.5*

As shown in Table 6, the time complexity of training the proposed algorithm is higher than the time complexity of the classical LDA algorithm due to the optimization process. Otherwise, they are almost similar.

Simulation Results

A. Simulation of the Proposed Algorithm

Three steps for complete simulation:

I. Optimize the proposed-dist LDA, proposed-Box, and proposed-Tanh algorithms for each selected HOC in Section 2, as shown in Fig. 4.
II. Calculate of the linear projection vector ***u*** as shown in Fig. 5.
III. Calculate the Number of Misclassified Datasets (NoMD) for the two mentioned problems: problem p1 in Section 4, the classification between 8PSK and 16PSK, and problem p2, which is the classification between 16QAM and 64QAM.

B. The Proposed-Dist LDA Algorithm

Fig. 6 shows the simulation results of the normalized NoMD values of the classical LDA (16) and the proposed-dist LDA algorithms for the problems p1 and p2, and SNR values [-5: 20] dB.

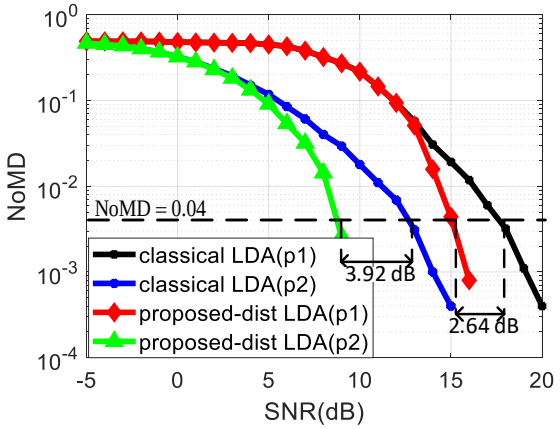


Fig. 6: Normalized NoMD values of the classical LDA and the proposed-dist LDA algorithms for the problems p1 and p2.

As shown in Fig. 6, the proposed-dist LDA algorithm could improve the discrimination between 8PSK and 16PSK for SNR values greater than 13 dB and between 16QAM and 64QAM for SNR values greater than 4 dB and for the normalized NoMD value of 0.04 (as an example), the improvement by using the proposed-dist LDA algorithm compared to the classical LDA algorithm is 2.64 dB for the problem p1 and 3.92 dB for the problem p2.

C. The Proposed-Box LDA Algorithm

Fig. 7 shows the simulation results of the normalized NoMD values of the classical LDA and the proposed-Box LDA algorithms for problems p1 and p2, and SNR values [-5: 20] dB.

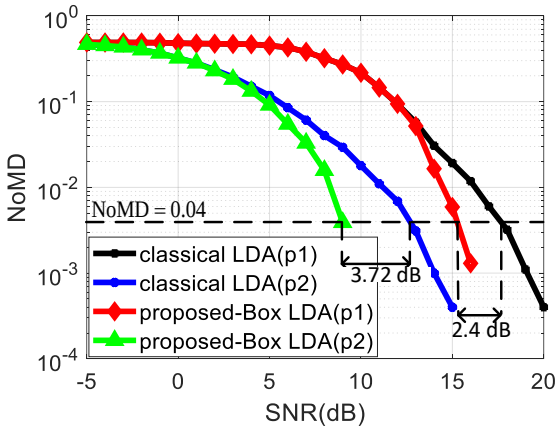


Fig. 7: Normalized NoMD values of the classical LDA and the proposed-Box LDA algorithms for the problems p1 and p2.

As shown in Fig. 7, the proposed-Box LDA algorithm could improve the discrimination between 8PSK and 16PSK for SNR values greater than 13 dB and between 16QAM and 64QAM for SNR values greater than 4 dB and for the normalized NoMD value of 0.04 (as an example), the improvement by using the proposed-Box LDA algorithm compared to the classical LDA algorithm is 2.4 dB for the problem p1 and 3.72 for the problem p2.

D. The Proposed-Tanh LDA Algorithm

Fig. 8 shows the simulation results of the normalized NoMD values of the classical LDA and the proposed-Tanh LDA algorithms for problems p1 and p2, and SNR values [-5: 20] dB.

As shown in Fig. 8, the proposed-Tanh LDA algorithm could improve the discrimination between 8PSK and 16PSK for SNR values greater than 13 dB and between 16QAM and 64QAM for SNR values greater than 4 dB and for the normalized NoMD value of 0.04 (as an example), the improvement by using the proposed-Tanh LDA algorithm compared to the classical LDA algorithm is 2.48 dB for the problem p1 and 4 dB for the problem p2.

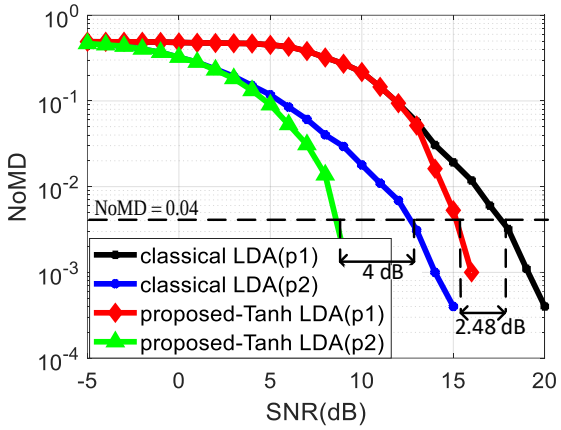


Fig. 8: Normalized NoMD values of the classical LDA and the proposed-Tanh LDA algorithms for the problems p1 and p2.

E. Classification Accuracy Improvement Compared with Reference Papers [1], [2]

MPSK and MQAM have been classified in reference papers [1], [2].

In [1], the author calculated the classification accuracy of MPSK and MQAM over a flat fading channel for SNR values of 10 dB and 0 dB. The classification accuracy of MPSK and MQAM is calculated using our optimized nonlinear LDA algorithm, i.e., the regularized distance-based LDA algorithm. The improvement of our proposed nonlinear LDA algorithm is calculated by subtracting the classification accuracy of the reference paper [1] from the classification accuracy of our proposed algorithm, as shown in Table 7.

Table 7: Comparison between the performance of the reference paper [1] and our proposed algorithm

SNR (dB)	0 dB	10 dB
Classification accuracy in Reference paper [1]	76.8%	84.37%
Classification accuracy of our proposed algorithm	78.25%	95.16%
The improvement of our proposed algorithm	1.45%	10.79%

As shown in Table 7, the classification accuracy of our proposed algorithm is improved compared with the reference paper [1]. The maximum improvement of our proposed algorithm compared with the reference paper [1] is 10.79%.

In [2], the author calculated the classification accuracy of MPSK and MQAM over a flat fading channel for SNR range [0: 20] dB. The improvement of our proposed nonlinear LDA algorithm is calculated by subtracting the classification accuracy of the reference paper [2] from the classification accuracy of our proposed algorithm, as shown in Fig. 9.

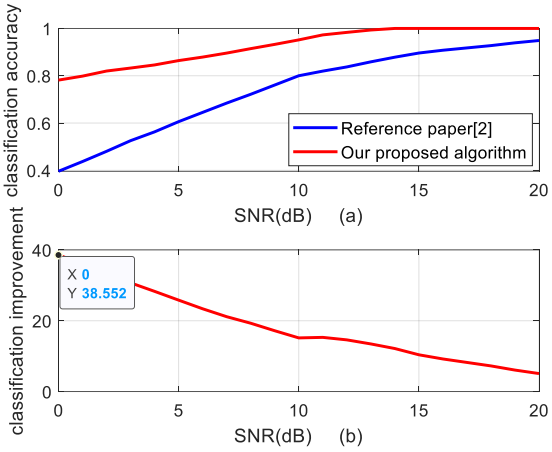


Fig. 9: Comparison between the performance of the reference aper [2] and our proposed algorithm.

As shown in Fig. 9, the classification accuracy of our proposed algorithm is improved compared with the reference paper [2]. The maximum improvement of our proposed algorithm compared with the reference paper [2] is 38.552%.

Conclusion

To improve the classification, an optimized nonlinear preprocess LDA algorithm has been developed. Three optimized functions have been used. These functions have similar performances.

According to Figs. 6, 7, and 8, the proposed preprocess LDA algorithms improve the classification between 8PSK and 16PSK for SNR values greater than 13 dB and between 16QAM and 64QAM for SNR values greater than 4 dB. The proposed-dist LDA algorithm has the best performance for classification between 8PSK and 16PSK. In contrast, the proposed-Tanh LDA algorithm has the best performance for classification between 16QAM and 64QAM. On the other hand, according to Table 7 and Fig. 9, the classification accuracy of our proposed algorithm is improved compared with the reference papers [1], [2]. The maximum improvement of our proposed algorithm compared with the reference paper [1] is 10.79%, and the maximum improvement of our proposed algorithm compared with the reference paper [2] is 38.552%.

By using the proposed optimization algorithm, the AMC classification accuracy has been improved. Other classification problems can use this algorithm. And other nonlinear preprocess functions or optimization algorithms may be found in future work.

Author Contributions

This paper is the result of I. Kadoun's Ph.d. thesis supervised by H. Khaleghi Bizaki. I. Kadoun and H. Khaleghi Bizaki proposed the main idea of the innovation of the paper. I. Kadoun performed the simulations, carried out the data analysis, interpreted the results and

wrote the manuscript. I. Kadoun and H. Khaleghi Bizaki corrected the proofing the article.

Acknowledgment

The authors would like to thank the editor and anonymous reviewers.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

AMC	Automatic Modulation Classification
SNR	Signal-to-Noise Ratio
LDA	Linear Discriminant Analysis
HOCs	Higher-Order Cumulants
MD	Mahalanobis Distance
MPSK	M-array Phase Shift Keying
MQAM	M-array Quadrature Amplitude shift Modulation
LB	Likelihood-Based
FB	Feature-Based
i.i.d	Independent and identically distributed
ACC	Classification accuracy

References

[1] S. A. Ghauri, I. M. Qureshi, A. Aziz, T. A. Cheema, "Classification of digital modulated signals using linear discriminant analysis on faded channel," World App. Sci. J., 29(10): 1220-1227, 2014.

[2] O. A. Dobre, A. Abdi, Y. Bar-Ness, W. J. W. P. C. Su, "Cyclostationarity-based modulation classification of linear digital modulations in flat fading channels," Wirel. Pers. Commun., 54(4): 699-717, 2010.

[3] O. A. Dobre, A. Abdi, Y. Bar-Ness, W. Su, "Survey of automatic modulation classification techniques: classical approaches and new trends," IET commun., 1(2): 137-156, 2007.

[4] O. A. Dobre, F. Hameed, "Likelihood-based algorithms for linear digital modulation classification in fading channels," in Proc. 2006 Canadian Conference on Electrical and Computer Engineering: 1347-1350, 2006.

[5] A. Hazza, M. Shoaib, S. A. Alshebeili, A. Fahad, "An overview of feature-based methods for digital modulation classification," in Proc. 2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA): 1-6, 2013.

[6] X.R. Jiang, H. Chen, Y.D. Zhao, W. Q. J. I. J. o. E. Wang, "Automatic modulation recognition based on mixed-type features," Int. J. Electron., 108(1): 105-114, 2021.

[7] D.-C. Chang and P.-K. J. I. C. Shih, "Cumulants-based modulation classification technique in multipath fading channels," IET Commun., 9(6): 828-835, 2015.

[8] P. S. Thakur, S. Madan, M. Madan, "Trends in automatic modulation classification for advanced data communication networks," Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET), 4(2): 496-507, 2015.

[9] Y. Wei, S. Fang, X. Wang, "Automatic modulation classification of digital communication signals using SVM based on hybrid features, cyclostationary, and information entropy," Entropy, 21(8): 745, 2019.

[10] Y. Kumar, M. Sheoran, G. Jajoo, S. K. Yadav, "Automatic modulation classification based on constellation density using deep learning," IEEE Commun. Lett., 24(6): 1275-1278, 2020.

[11] X. Zhang, J. Sun, X. Zhang, "Automatic modulation classification based on novel feature extraction algorithms," IEEE Access, 8: 16362-16371, 2020.

[12] D. H. Al-Nuaimi, I. A. Hashim, I. S. Zainal Abidin, L. B. Salman, N. A. Mat Isa, "Performance of feature-based techniques for automatic digital modulation recognition and classification—A review," Electronics, 8(12): 1407, 2019.

[13] S. Sobolewski, W. L. Adams, and R. Sankar, "Universal nonhierarchical automatic modulation recognition techniques for distinguishing bandpass modulated waveforms based on signal statistics, cumulant, cyclostationary, multifractal and Fourier-wavelet transforms features," in 2014 IEEE Military Communications Conference: 748-753, 2014.

[14] A. Smith, M. Evans, J. Downey, "Modulation classification of satellite communication signals using cumulants and neural networks," in Proc. 2017 Cognitive Communications for Aerospace Applications Workshop (CCAA): 1-8, 2017.

[15] I. Kadoun, H. K. Bizaki, "Advanced features generation algorithm for MPSK and MQAM classification in flat fading channel," Radioengineering, 31(1): 127, 2022.

[16] V. D. Orlic, M. L. Dukic, "Automatic modulation classification algorithm using higher-order cumulants under real-world channel conditions," IEEE Commun. Lett., 13(12): 917-919, 2009.

[17] A. E. Abdelmutalab, "Learning-based automatic modulation classification," 2015.

[18] S. A. Ghauri, I. M. Qureshi, A. N. Malik, T. A. Cheema, "Higher order cummulants based digital modulation recognition scheme," Res. J. Appl. Sci. Eng. Tech., 6(20): 3910-3915, 2013.

[19] A. Abdelmutalab, K. Assaleh, M. J. P. C. El-Tarhuni, "Automatic modulation classification based on high order cumulants and hierarchical polynomial classifiers," Phys. Commun., 21: 10-18, 2016.

[20] A. Tharwat, T. Gaber, A. Ibrahim, A. E. Hassanien, "Linear discriminant analysis: A detailed tutorial," AI Commun., 30(2): 169-190, 2017.

[21] B. Ghojogh, F. Karray, M. Crowley, "Fisher and kernel Fisher discriminant analysis: Tutorial," arXiv Prepr. arXiv1906.09436, 2019.

[22] A. Lahav, R. Talmon, Y. J. I. Kluger, I. A. J. o. t. IMA, "Mahalanobis distance informed by clustering," Inf. Inference A J. IMA., 8(2): 377-406, 2019.

[23] C. D. Meyer, Matrix analysis and applied linear algebra, 71. Siam, 2000.

[24] G. Strang, "Introduction to linear algebra Fifth," ed: Wellesley-Cambridge Press, 2016.

[25] A. C. Atkinson, M. Riani, A. J. S. S. Corbellini, "The Box–Cox transformation: Review and extensions," Stat. Sci., 36(2): 239-255, 2021.

[26] A. H. Namin, K. Leboeuf, R. Muscedere, H. Wu, M. Ahmadi, "Efficient hardware implementation of the hyperbolic tangent sigmoid function," in Proc. 2009 IEEE International Symposium on Circuits and Systems: 2117-2120, 2009.

[27] D. Cai, X. He, J. Han, "SRDA: An efficient algorithm for large-scale discriminant analysis," IEEE Trans. Knowl. Data Eng., 20(1): 1-12, 2007.

[28] A. Rupp, J. Pelzl, C. Paar, M. Mertens, A. Bogdanov, "A parallel hardware architecture for fast Gaussian elimination over GF (2)," in Proc. 2006 14th Annual IEEE Symposium on Field-Programmable Custom Computing Machines: 237-248, 2006.

[29] L. Zhao, W. Lin, Y. Wang, X. Li, "Recursive local summation of rx detection for hyperspectral image using sliding windows," Remote Sens., 10(1): 103, 2018.

[30] X. Li, S. Wang, Y. Cai, "Tutorial: Complexity analysis of Singular Value Decomposition and its variants," arXiv paper. arXiv1906.12085, 2019.

Biographies

Iyad Kadoun was born in Damascus, Syria, in 1979 and received his B.S. degree in communication engineering from HIAST in 2002, and his M.Sc. degree in communication engineering from Malek Ashtar University in 2012. His research interests include digital communications.

- Email: idiyad@yahoo.com
- ORCID: 0000-0003-1999-6066
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Hossein Khaleghi Bizaki received the Ph.D. degree in electrical engineering and communication systems from Iran University of Science and Technology, Tehran, Iran, in 2008. He is an author or coauthor of more than 100 publications. His research interests include information theory, coding theory, wireless communication, multiple-input–multiple-output systems, space–time processing, and other topics on communication system and signal processing.

- Email: bizaki@yahoo.com
- ORCID: 0000-0001-9458-8287
- Web of Science Researcher ID: NA
- Scopus Author ID: 23012350800
- Homepage: NA

How to cite this paper:
I. Kadoun, H. Khaleghi Bizaki, "Improving the classification of MPSK and MQAM modulations by using optimized nonlinear preprocess in flat fading channels," J. Electr. Comput. Eng. Innovations, 11(1): 141-152, 2023.

DOI: 10.22061/JECEI.2022.8743.550

URL: https://jecei.sru.ac.ir/article_1765.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Performance Analysis and Modeling of a Variable Reluctance Speed Sensor for Turbomachinery Applications

A. H. Nejadmalayeri*, P. Yousefi, M. Safaei

Technical and Engineering Faculty, Imam Hossein Comprehensive University (IHU), Tehran, Iran.

Article Info

Article History:
Received 27 May 2022
Reviewed 29 June 2022
Revised 19 July 2022
Accepted 30 August 2022

Keywords:
Variable reluctance speed sensor (VRS)
Electromagnetic sensors
Instantaneous angular speed (IAS)
Finite element (FEM)

*Corresponding Author's Email Address: *malayeri@ihu.ac.ir*

Abstract

Background and Objectives: The speed sensor is one of the main components of the control and monitoring systems of rotational machines which is widely used in the aviation industry, railway, and automotive applications. Variable Reluctance Speed sensor (VRS) is a kind of magnetic sensor that has been traditionally employed for many different industrial measurements because of several well-known advantages, such as passive nature, non-contact operations, robustness, low cost, low sensitivity to dirt, and large-signal output.
Methods: In this paper, a variable reluctance speed sensor is proposed. The design process of the proposed sensor is presented and both the magnetic and electrical models of this sensor are derived by assuming the effect of magnetomotive force caused by eddy current formed on the outer edge of the target gear at high frequencies. As a result, the proposed model can demonstrate the performance of the variable reluctance speed sensor at high frequencies very well.
Results: The proposed VRS is designed and simulated using MATLAB and Ansys Maxwell software to verify the theoretical results is constructed and tested.
Conclusion: In this paper, a variable reluctance speed sensor is proposed and studied. The magnetic and electrical models of the proposed sensor are derived and the output voltage equation has been calculated as a function of the air gap length. The proposed VR sensor is simulated using 2D Finite Element Analysis software to identify the main parameters that influence the sensor output and also to verify the accuracy of the model. According to the simulation results, the output waveform quality will be affected by parameters such as air gap length, target gear material, the self-inductance of the VR sensor, and the load component values. In terms of the electrical model, we were able to simulate the effect of load resistance and capacitance on the sensor output.

Introduction

The speed sensor is one of the main components of the control and monitoring systems of rotational machines which is widely used in the aviation industry, railway, and automotive applications [2]-[11]. For instance, for vibration-based damage detection of rotor blades in gas turbine engines, or for torsional vibration monitoring, it is necessary to measure instantaneous angular speed (IAS) as accurately as possible [12]-[19].

So far, several methods have been proposed for measuring IAS based on different kinds of sensing systems including Hall sensor-based, laser-based, optical encoders, capacitive and electromagnetic sensors, and potentiometric methods [20]-[25]. Variable Reluctance Speed sensor (VRS) is a kind of magnetic sensor that has been traditionally employed for many different industrial measurements because of several well-known advantages, such as passive nature, non-contact

operations, robustness, low cost, low sensitivity to dirt, and large-signal output. The main disadvantage of the VRS sensor is that the signal-to-noise ratio is very low at slow speeds and the output voltage depends on the target speed. In [26], a novel design and implementation of the VR sensors have been proposed which has resulted in an improvement in the speed measurement capabilities for turbomachinery. The proposed design provides an enhancement in output signal quality during the low and high-speed performance, and also during high-power operation. In addition, the new measuring system proposed in [26], has the capability for health monitoring of the engine bearings based on analysis of the differential output voltages from the two sensors. In [27], a numerical study has been performed on a VR sensor used for a coolant pump. In this research, a coupled circuit is introduced for calculating the induced voltage. The result of this study indicates that a sensor with a radially magnetized permanent magnet is more sensitive than C-shape. Also, a ferromagnetic yoke installation has the advantage of closing the magnetic circuit in order to increase the flux concentration within the circuit. All of these results will produce a higher output voltage. In [28], both the electrical and magnetic model of a VR sensor has been carried out and presented. In this paper, also the effect of load components on the output magnitude and resonant frequency of the output signal has been evaluated. In [29], [30], a speed measuring system is proposed based on the variable reluctance sensor in order to measure Instantaneous rotation speed and torsional vibration monitoring. In these papers, both the magnetic and electrical model of the system is proposed to evaluate the system quality. This model allows for simulating the behavior of the system, given the arbitrary shape and speed of the rotating target. All of the models presented in all of these articles have been simplified and the effect of magnetomotive force caused by eddy current formed on the outer edge of the target gear at high frequencies has not been considered. So, this model can't demonstrate the performance of the variable reluctance speed sensor at high frequencies very well. This paper is arranged as follows. In the first section, the basic theory of the variable reluctance speed sensor is described and both the magnetic and electrical equivalent circuits of the VR sensor are derived considering the effect of the eddy current that forms on the target gear. In the second section, 2D Finite Element Analysis (FEM) is used to model and simulate the output voltage generated by the sensor as a function of gear Instantaneous rotation speed. In the third section, an experimental setup is presented together with some experimental results confirming the results obtained from theory and simulation and finally, conclusions are drawn.

Modeling

As shown in Fig. 1, the structure of the VR sensor is made of a permanent magnet that is responsible for generating magnetic flux, and a sensing coil that is wrapped around an iron core that acts as a probe. When a ferromagnetic target passes through the probe, loading the permanent magnet occur and causes variation of the magnetic flux density, which consequently crosses the sensing coil and according to faraday law, induces voltage on the sensing winding.

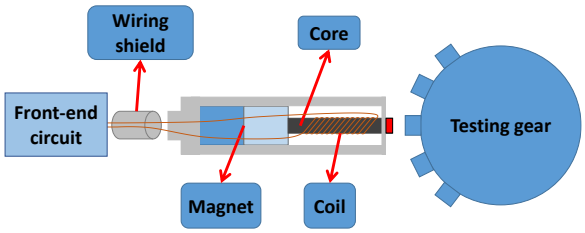


Fig. 1: Structure of variable reluctance speed sensor.

The system shown in Fig. 1 can be considered from two perspectives. In this section, both the magnetic and electrical model of the proposed VR sensor is derived. The magnetic part is composed of a magnet, ferrous iron core, and ferrous testing gear, and finally, the electrical section formed by the VRS coil, the wiring (usually a twin-axial cable), and the front-end electronics.

A. Magnetic Equivalent Circuit

The magnetic equivalent circuit of the system have shown in Fig. 1 can be derived using the usual approach exploiting the definition of the magnetic flux and applying the Ampere law to the identified flux line φ in Fig. 2(a), extending in the magnet for the section φ_m, and outside the magnet for the section φ_{lk}. Using the model shown in Fig. 2(a), the magnetic equivalent circuit can be derived as Fig. 2(b), in which R_m, R_{lk}, R_c, R_i and R_g are permanent magnet internal reluctance, leakage reluctance, iron core reluctance, sensor housing reluctance, and air gap reluctance respectively.

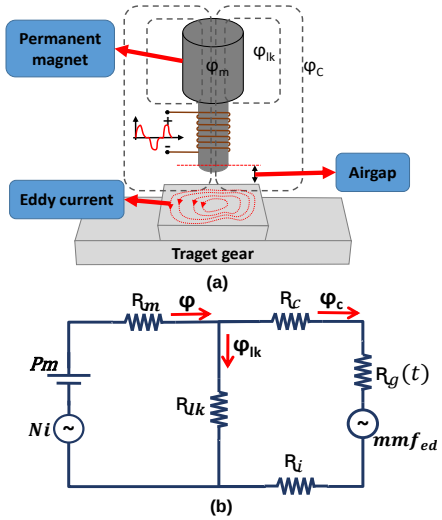


Fig. 2: Magnetic circuit derivation: (a) magnetic model, (b) magnetic circuit.

As shown in Fig. 2(a), air gap and sensor housing reluctances vary with the target gear motion. So, it can be considered as a function of time. Nevertheless, the value of $\mathcal{R}_{lk}$ is usually very large whereas $\mathcal{R}_i$ is so small, so they can be neglected in order to simplify the calculation. The flux in the magnetic circuit is generated by two magnetomotive forces (MMFs), a constant MMF relative to the permanent magnet, PM, in series with its internal R_m and the other given by MMF created by the sensing coil. By neglecting the current flowing in the sensing coil, this MMF can be ignored in calculations.

In Fig. 3, the result of the FEM analysis has been shown which indicates the accuracy of the magnetic equivalent circuit shown in Fig. 2(b).

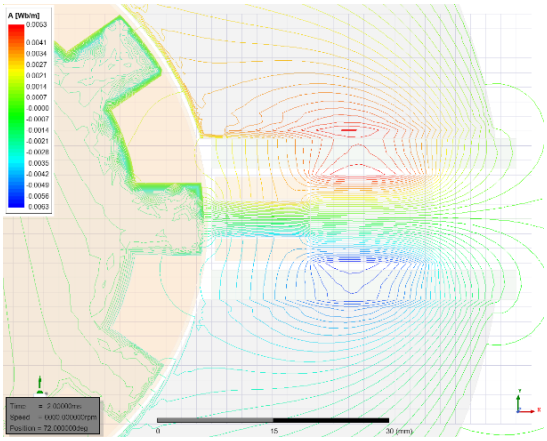


Fig. 3: FEM analysis of the target VR sensor.

The air gap reluctance can be defined by the following equation:

$$\mathcal{R}_g = \frac{l_g(t)}{\mu_0 A_g} \tag{1}$$

where A_g is the equivalent surface of the gap area, μ0 is the air magnetic permeability, whereas $l_g(t)$ is the equivalent gap length as a function of the time. Therefore, as the target gear rotates, the air gap reluctance will be a time-dependent variable, which in turn causes a variable flux in the air gap. The air gap variation when the target gear rotates and considering 1-D geometry is shown in Fig. 4.

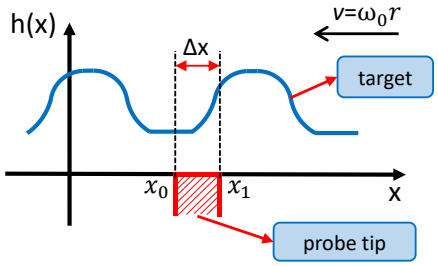


Fig. 4: Target gear and sensor tip profiles.

The target is supposed to move at the speed v, parallel to axis x. According to Fig. 4, the average effective gap length can be defined as follows:

$$l_g(t) = \frac{1}{\Delta x}\int_{x_0}^{x_1} h(x - vt)\,dt \tag{2}$$

As explained previously, the leakage reluctance of VR sensors can be ignored to simplify calculations. So according to the equivalent magnetic circuit shown in Fig. 2(b), the magnetic flux can be calculated as follow:

$$MMF_t = (\mathcal{R}_m + \mathcal{R}_c + \mathcal{R}_g(t) + \mathcal{R}_i)\varphi \tag{3}$$

The MMF_t in (3) is the total magneto motive force that can be defined as follow:

$$MMF_t = Ni(t) + Pm + mmf_{ed} \tag{4}$$

where N is the number of turns in the sensing coil. By using the VR sensor in an electrical circuit, a current flow throws the sensing coil. So, this current makes the magnetomotive force *Ni*(t) and also Pm is the MMF caused by the permanent magnet. MMF_{ed} in (4) is the magnetomotive force generated by eddy current formed on the outer edge of the target gear which is shown in Fig. 2(a). The magnetic flux equation can be written as follow:

$$\varphi(t) = \frac{Ni(t) + Pm + mmf_{ed}}{\mathcal{R}_m + \mathcal{R}_c + \dfrac{l_g(t)}{\mu_0 A_g} + \mathcal{R}_i} \tag{5}$$

B. Electrical Equivalent Circuit

The equation which defines the magnetic flux has been calculated in the previous section. For calculating the induced voltage in the sensing coil of the VR sensor, the Faraday law can be used as follow:

$$V = -N\frac{d\phi(t)}{dt} \tag{6}$$

As shown in (5), the magnetic flux has been composed of three terms. The flux generated by sensing coil current, the flux generated by permanent magnet and finally the flux generated by target gear eddy current which generates harmonics on the output voltage of the sensor at high frequencies.

$$V_o(t) = V_{coil}(t) + V_{Pm}(t) + V_{mmfed}(t) \tag{7}$$

So, using (5) and (6), the induced voltages can be defined as follow:

$$V_{coil}(t) = -\frac{N^2}{\mathcal{R}_t + \dfrac{l_g(t)}{\mu_0 A_g}}\frac{di(t)}{dt} + N^2 i(t)\frac{d}{dt}\frac{1}{\mathcal{R}_t + \dfrac{l_g(t)}{\mu_0 A_g}} \tag{8}$$

$$V_{coil}(t) = -\frac{N^2}{\mathcal{R}_t + \dfrac{l_g(t)}{\mu_0 A_g}}\frac{di(t)}{dt} + N^2 i(t)\frac{d}{dt}\frac{1}{\mathcal{R}_t + \dfrac{l_g(t)}{\mu_0 A_g}} \tag{9}$$

$$V_{mmf_{ed}}(t) = -N.mmf_{ed}\frac{d}{dt}\frac{1}{\mathcal{R}_t + \dfrac{l_g(t)}{\mu_0 A_g}} \tag{10}$$

where $\mathcal{R}_t$ is defined as follow:

$$\mathcal{R}_t = \mathcal{R}_m + \mathcal{R}_c + \mathcal{R}_i \tag{11}$$

The (8) that defines the voltage of the coil, is composed of two terms. According to the inductance voltage (12), the first term is the voltage that places on the self-inductance of the sensor. So, self-inductance can be defined as the (13).

$$V_{Lc} = L_c \frac{di(t)}{dt} \tag{12}$$

$$L_c(t) = \frac{N^2}{\mathrm{R}_t + \frac{l_g(t)}{\mu_0 A_g}} \tag{13}$$

As shown in (13), the inductance value is a time-dependent variable because it depends on the air gap length.

The second term of (8), is opposing electromotive force respect to the one generated by the magnet that is shown by (14).

$$V_{op(t)} = N^2 i(t) \frac{d}{dt} \frac{1}{\mathrm{R}_t + \frac{l_g(t)}{\mu_0 A_g}} \tag{14}$$

$$V_{coil}(t) = L_c \frac{di(t)}{dt} + V_{op(t)} \tag{15}$$

So, we have (16) for the output voltage:

$$V_o(t) = V_{Lc}(t) + V(t) \tag{16}$$

where:

$$V(t) = V_{op}(t) + V_{Pm}(t) + V_{mmfed}(t) \tag{17}$$

By assuming R_{coil} as the parasitic resistance of the sensing coil and according to the (16) and (17) that defines the output voltage and the voltage induced in the coil, the electrical model of the VR sensor can be derived as Fig. 5. The load component considered at the output of the sensor is used to adjust the amplitude of the output waveform for a range of frequencies. The variation of the output pulse for a different amount of the load impedance is shown in Fig. 6.

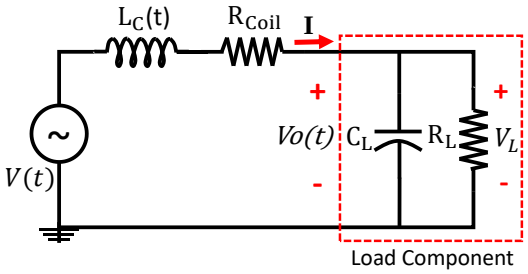


Fig. 5: Equivalent electrical circuit of the VR sensor.

The sensor current, can be calculated by solving the differential equation below:

$$L_C(t)\frac{dI(t)}{dt} + R_{Coil}I(t) + Z_L I(t) = V(t) \tag{18}$$

It is obvious that (18) is not an ordinary differential equation with constant parameters, so it has to be solved numerically.

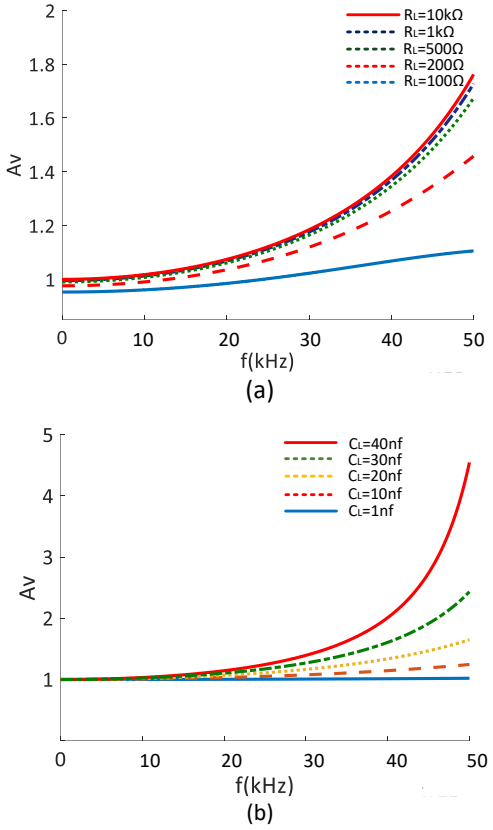


Fig. 6: The output voltage changes for different frequencies, (a) CL=1nf; (b) RL=10KΩ.

Design and Simulation of the proposed sensor

In this section, the design process of the proposed VR sensor is shown and then the simulations are performed according to the design results. As shown in Fig. 7, the design process of the VRS sensor is an iterative process. At the first step of the design process, the magnet and pole piece are selected as Table 1. By assuming *V(t)*=5v for the output voltage, the design process continue until the desired value for V(t) is reached.

The 2D Finite Element Analysis (FEM) is used for the simulation of the output voltage generated by the sensor. The distribution of the magnetic flux during target gear rotation is shown in Fig. 8 for two different states of assuming and regardless of the target gear eddy current. As previously explained, the magnetic flux produced by the target gear eddy current causes harmonics and makes distortion on the sensor output voltage.

Table 1: VR Sensor simulation parameters

	Material	Electromagnetic Properties	Dimensions (mm) length×diameter
Magnet	NeFeB	H_C = 1034507 A/m, μ_r = 1.05	8×4
Pole Piece	Ferrite	Linear B-H Curve	5×4

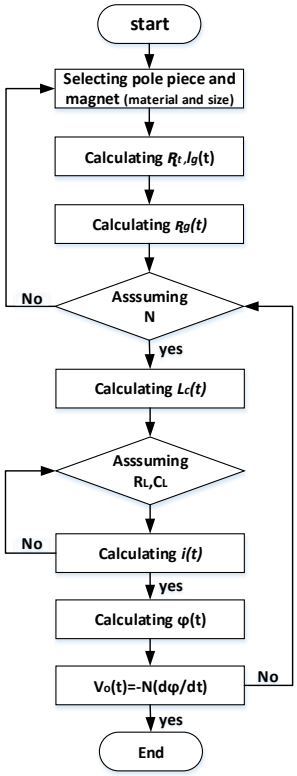


Fig. 7: The design process of the proposed VRS sensor.

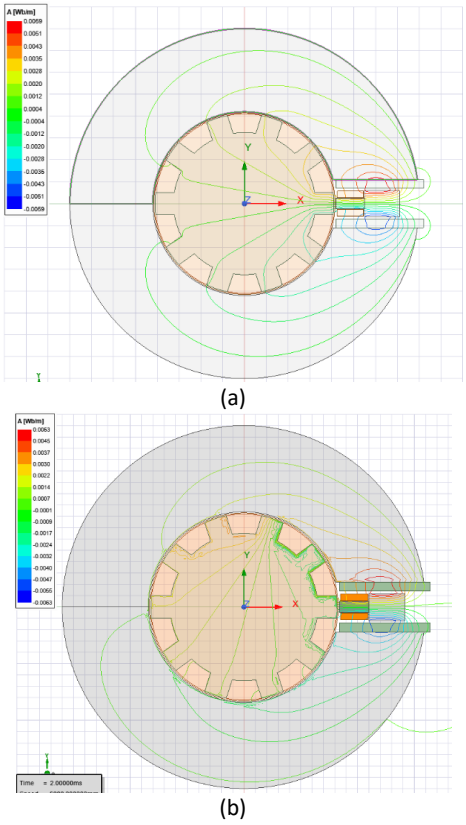


Fig. 8: Target simulation, (a) magnetic flux by Ignoring eddy current effect. (b) magnetic flux by considering eddy current effect.

So, in order to obtain the sensor coil voltage induced by the target gear eddy current, the simulation is performed in two ways, once by ignoring the effect of the flux generated by the target gear eddy current and then by considering the eddy current effect. The output voltage for both of the simulation states is shown in Fig. 9(a).

As shown in Fig. 9, the flux caused by the eddy current generated on the outer edge of the target gear makes distortion on the voltage induced in the sensor coil. The voltage caused by the eddy current is shown in Fig. 9(b).

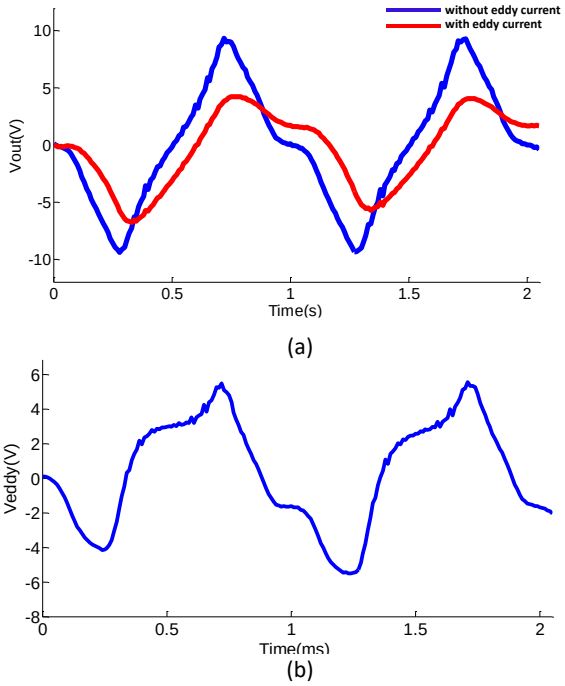


Fig. 9: Simulation results, (a) The output voltage of the VR speed sensor for 6000RPM (b) The voltage induced in the coil by the eddy current flux.

This voltage waveform is the result of the difference between the two waveforms shown in Fig. 9(a). A comparison of the two waveforms shown in Fig. 9(a)and Fig. 9(b) indicates that the flux generated by the target gear eddy current, acts in the opposite direction of the main flux and it weakens the main magnetic field. So, as shown in Fig. 2(b), it can be modeled as a magnetomotive force (MMF_{ed}) in the magnetic circuit. The value of the MMF_{ed} is proportional to the rotation frequency of the target gear. So, the higher instantaneous rotational speed, the greater distortion of the output waveform. The sensor output voltage for two different speeds is shown in Fig .10

The self-inductance variation is shown in Fig. 11 for the target gear. As shown in Fig. 11 and according to (13), the self-inductance varies with air gap length variations.

The amount of the induced voltage due to the MMF_{ed} is related directly to the magnetic permeability coefficient of the target gear material. As the magnetic permeability increases, the voltage induced by MMF_{ed} which is shown

in Fig. 9(b) decreases and the sensor output waveform will be more appropriate.

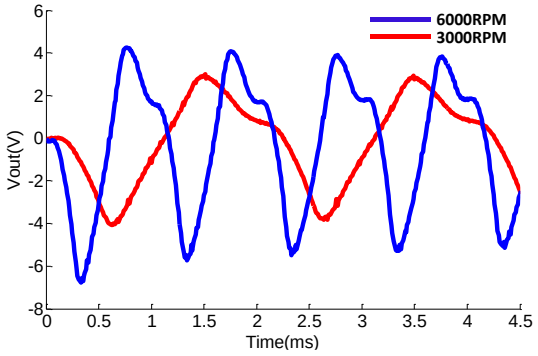


Fig. 10: The sensor output voltage for two different speed.

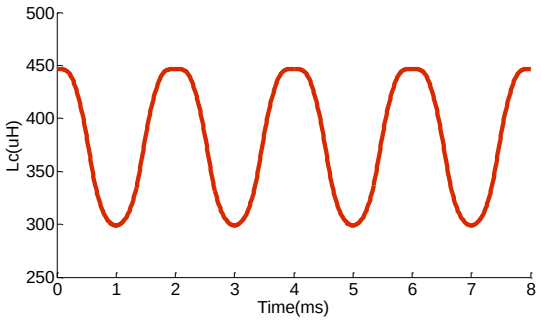


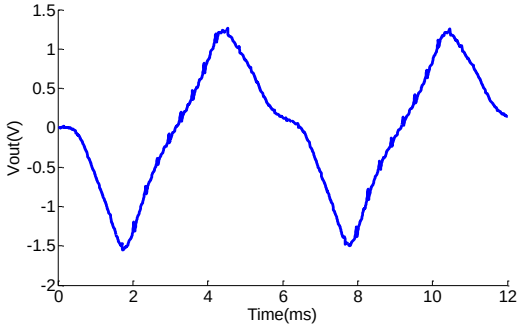
Fig. 11: VR sensor self-inductance variations in 3000RPM.

In order to show this issue, simulation has been performed for three different types of the target gear material which is shown in Table 2.

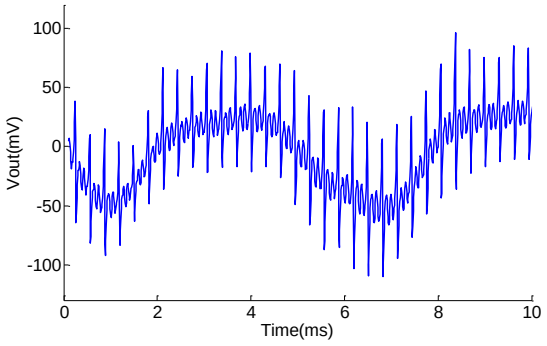
Table 2: Different target gear materials

Material	Relative Permeability	Metal Conductivity(S/m)
Iron	4000	10300000
Aluminum	1.000021	38000000
Steel	1	1100000

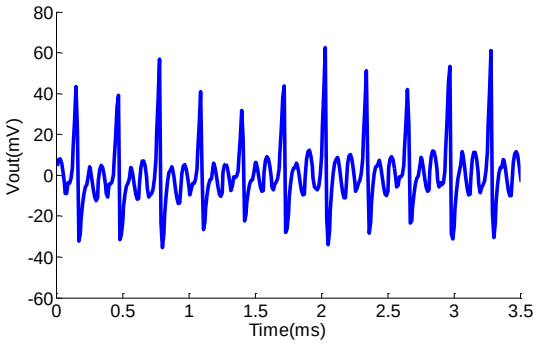
As shown in Fig. 12, the iron target gear with the highest magnetic permeability coefficient compared to the aluminum and steel materials has the best output result.



(a)



(b)



(c)

Fig. 12: The sensor output voltage for different target gear materials (a) Iron, (b) Aluminum, (c) Steel.

Experimental Results

An experimental test is used to validate the magnetic model derived and presented in previous sections. As shown in Fig. 13, the test system consists of a variable speed motor and, a target gear with 10 teeth and a VR sensor with the characteristics shown in Table 1.

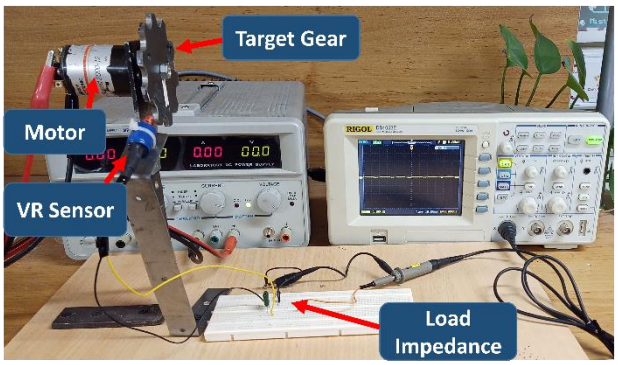


Fig. 13: Experimental set-up used for the tests.

The sensor output signal was acquired with a digital oscilloscope (RIGOL DS1052E) and The VR sensor is loaded by a large impedance as shown in figure5. The test is performed and the experimental result of the test is shown in Fig. 14.

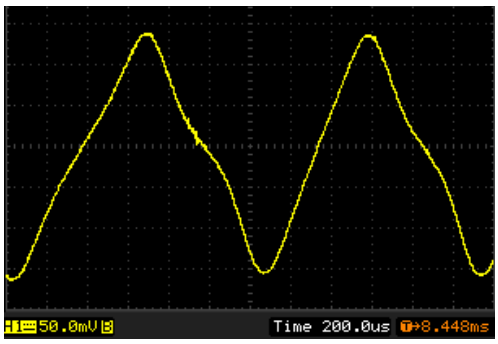
As explained at previous sections, the output waveform of the VR sensor is always a function of the target gear geometry, material and airgap length. The output voltage waveform of the VR sensor for an iron

target gear is shown in Fig. 14(a). According to the magnetic circuit shown in Fig. 2(b), and the high relative permeability of iron, most of the magnetomotive force (MMF_{ed}) generated in this mode is related to the sensor magnet and MMF_{ed} is so weaker than the total MMF. So, in this case, the voltage waveform has lower harmonics and is very similar to the target gear geometry.

The second test has been performed using an aluminum target gear as shown in Fig. 14(b). As shown in Fig. 2(b) and due to the low value of the relative permeability of aluminum, the voltage waveform is distorted. Therefore, to use this sensor for applications such as vibration monitoring or blade tip timing in turbomachinery, using complex electronic circuits to eliminate the sensor output waveform is necessary.

The amplitude of the output voltage waveform can be adjusted according to Fig. 6 by changing the output load impedance. So that the voltage amplitude changes at different frequencies do not deviate from the linear state.

The comparison of the experimental output waveform of the proposed VR sensor by the simulation results that shown in Fig. 12(a) and Fig. 12(b), ensures the accuracy of the modeling performed.



(a)

(b)

Fig. 14: Experimental test result of the proposed VR sensor (a)Iron target gear (b)Aluminum target gear.

Conclusions

In this paper, a variable reluctance speed sensor is proposed and studied. The magnetic and electrical models of the proposed sensor are derived and the output voltage equation has been calculated as a function of the air gap length. The proposed VR sensor is simulated using 2D Finite Element Analysis software to identify the main parameters that influence the sensor output and also to verify the accuracy of the model. According to the simulation results, the output waveform quality will be affected by parameters such as air gap length, target gear material, the self-inductance of the VR sensor, and the load component values. In terms of the electrical model, we were able to simulate the effect of load resistance and capacitance on the sensor output.

Author Contributions

Amir Hossein Nejadmalayeri in collaboration with Peyman Yousefi and Meysam Safaei, designed, simulated and carried out the data analysis. He collected the data and interpreted the results and wrote the manuscript.

Acknowledgment

The authors gratefully thank the anonymous reviewers and the editor of JECEI for their useful comments and suggestions.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

VRS	Variable Reluctance Sensor
IAS	Instantaneous Angular Speed
FEM	Finite Element
MMF	Magneto Motive Forces

References

[1] S. Merhav, Aerospace sensor systems and applications. Springer Science & Business Media, 1998.

[2] F. Chaaban, T. Birch, D. Howe, P. Mellor, "Topologies for a permanent magnet generator/speed sensor for the ABS on railway freight vehicles," in Proc. 1991 Fifth International Conference on Electrical Machines and Drives (Conf. Publ. No. 341), IET: 31-35, 1991.

[3] B. Wang, "Design of teaching platform for ABS wheel speed sensor," J. Phys. Conf. Ser., 2187(1): IOP Publishing: 012004, 2022.

[4] R. Przysowa, E. Rokicki, "Inductive sensors for blade tip-timing in gas turbines," J. KONBiN, 36(1): 147, 2015.

[5] D. Heller, I. Sever, C. Schwingshackl, "A method for multi-harmonic vibration analysis of turbomachinery blades using Blade Tip-Timing and clearance sensor waveforms and optimization techniques," Mech. Syst. Sig. Process., 142: 106741, 2020.

[6] A. Vercoutter, M. Berthillier, A. Talon, B. Burgardt, J. Lardies, "Estimation of turbomachinery blade vibrations from tip-timing data," in Proc. 10th International Conference on Vibrations in Rotating Machinery: 11-13, 2012.

[7] Y. S. Didosyan, H. Hauser, H. Wolfmayr, J. Nicolics, P. Fulmek, "Magneto-optical rotational speed sensor," Sens. Actuators, A, 106(1-3): 168-171, 2003.

[8] P. Procházka, F. Vaněk, "New methods of noncontact sensing of blade vibrations and deflections in turbomachinery," IEEE Trans. Instrum. Meas., 63(6): 1583-1592, 2013.

[9] T. Achour, M. Pietrzak-David, "Service continuity of an IM distributed railway traction with a speed sensor fault," in Proc. the

2011 14th European Conference on Power Electronics and Applications, : 1-8, 2011.

[10] Y. Maniwa, S. Kitamura, K. Aoyama, M. Matsuyama, "Turbomachinery control by CENTUM VP," Yokogawa Technical Report-English Edition-, 45: 47, 2008.

[11] M. Dowell, G. Sylvester, "Turbomachinery prognostics and health management via eddy current sensing: current developments," in Proc. 1999 IEEE Aerospace Conference (Cat. No. 99TH8403), 3: 1-9, 1999.

[12] Y. Li, F. Gu, G. Harris, A. Ball, N. Bennett, K. Travis, "The measurement of instantaneous angular speed," Mech. Syst. Sig. Process., 19(4): 786-805, 2005.

[13] S. Madhavan, R. Jain, C. Sujatha, A. Sekhar, "Vibration based damage detection of rotor blades in a gas turbine engine," Eng. Fail. Anal., 46: 26-39, 2014.

[14] A. Darpe, K. Gupta, A. Chawla, "Coupled bending, longitudinal and torsional vibrations of a cracked rotor," J. Sound Vib., 269(1-2): 33-60, 2004.

[15] S. W. Doebling, C. R. Farrar, M. B. Prime, "A summary review of vibration-based damage identification methods," Shock Vib. Digest, 30(2): 91-105, 1998.

[16] L. Doliński, M. Krawczuk, "Damage detection in turbine wind blades by vibration based methods," J. Phys. Conf. Ser., 181(1): IOP Publishing: 012086, 2009.

[17] C. Liu, D. Jiang, "Improved blade tip timing in blade vibration monitoring with torsional vibration of the rotor," J. Phys. Conf. Ser., 364(1): IOP Publishing: 012136, 2012.

[18] L. Naldi, M. Golebiowski, "New approach to torsional vibration monitoring," in Proc. the 40th Turbomachinery Symposium, Texas A&M University. Turbomachinery Laboratories, 2011.

[19] F. L. M. Dos Santos, B. Peeters, H. Van Der Auweraer, L. Góes, W. Desmet, "Vibration-based damage detection for a composite helicopter main rotor blade," Case Stud. Mech. Syst. Sig. Process., 3: 22-27, 2016.

[20] S. Kaul, R. Koul, C. Bhat, I. Kaul, A. Tickoo, "Use of alook-up'table improves the accuracy of a low-cost resolver-based absolute shaft encoder," Meas. Sci. Technol. 8(3): 329, 1997.

[21] X. Li, G. C. Meijer, "A novel low-cost noncontact resistive potentiometric sensor for the measurement of low speeds," IEEE Trans. Instrum. Meas., 47(3): 776-781, 1998.

[22] T. Fabian, G. Brasseur, "A robust capacitive angular speed sensor," IEEE Trans. Instrum. Meas., 47(1): 280-284, 1998.

[23] R. M. Kennel, "Encoders for simultaneous sensing of position and speed in electrical drives with digital control," IEEE Trans. Ind. Appl. 43(6): 1572-1577, 2007.

[24] M. Nandakumar, S. Ramalingam, S. Nallusamy, S. Srinivasarangan Rangarajan, "Hall-sensor-based position detection for quick reversal of speed control in a BLDC motor drive system for industrial applications," Electronics, 9(7): 1149, 2020.

[25] L. Avanesov Yuriy, N. Bukanova Ayna, S. Voronov Alexander, I. Evstifeev Michail, "Optimization of design parameters for depth electromagnetic speed sensor," J. Sci. Tech. Inf. Technol., Mech. Opt., 113(1): 140-146, 2018.

[26] J. J. Costello, A. C. Pickard, "A novel speed measurement system for turbomachinery," IEEE sens. Lett., 2(4): 1-4, 2018.

[27] H. Huh, J. S. Park, S. Choi, K. B. Park, S. Q. Zee, "Numerical research on new variable reluctance sensor with fixed permanent magnet for SMART main coolant pump," in Proc. the Korean Nuclear Society Conference, Korean Nuclear Society: 1045-1046, 2005.

[28] R. A. Croce Jr, I. Giterman, "Development of the Electrical and Magnetic Model of Variable Reluctance Speed Sensors."

[29] T. Addabbo et al., "Instantaneous rotation speed measurement system based on variable reluctance sensors for torsional vibration monitoring," IEEE Trans. Instrum. Meas., 68(7): 2363-2373, 2018.

[30] T. Addabbo et al., "Instantaneous rotation speed measurement system based on variable reluctance sensors: Model and analysis of performance," in Proc. 2018 IEEE Sensors Applications Symposium (SAS): 1-6, 2018.

Biographies

Amir Hossein Nejadmalayeri received his B.Sc. degree in Electrical Engineering from Shahid Bahonar University, Kerman, Iran, in 2018 and his M.Sc. degree in Electrical Engineering from Malek-Ashtar University of Technology, Tehran, Iran, in 2021. His research interests include design, modeling of power converters, and pulsed power systems.

- Email Address: malayeri@ihu.ac.ir
- ORCID: 0000-0002-8506-5959
- Web of Science Researcher ID: GLV-1497-2022
- Scopus Author ID: NA
- Homepage: NA

Meysam Safaei received his B.Sc. degree in electrical engineering from University of Eyvanekey, Teharn, Iran, in 2011 and his M.Sc. degree in artificial intelligence and robotics engineering from Malek-Ashtar University of Technology, Tehran, Iran, in 2022. His research interests include artificial intelligence and IOT.

- Email Address: safaei@ihu.ac.ir
- ORCID: 0000-0002-4544-4284
- Web of Science Researcher ID: GLV-2737-2022
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:
A. H. Nejadmalayeri, P. Yousefi, M. Safaei, "Performance analysis and modeling of a variable reluctance speed sensor for turbomachinery applications," J. Electr. Comput. Eng. Innovations, 11(1): 153-160, 2023.

DOI: 10.22061/JECEI.2022.8528.522

URL: https://jecei.sru.ac.ir/article_1767.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

A Novel Full-duplex Relay Selection and Resource Management in Cooperative SWIPT NOMA Networks

M. B. Noori Shirazi, M. R. Zahabi*

Faculty of Electrical and Computer Engineering, Babol Noshirvani University of Technology, Babol, Iran.

Article Info

Article History:
Received 23 April 2022
Reviewed 12 June 2022
Revised 26 July 2022
Accepted 30 August 2022

Keywords:
Energy efficiency
Energy harvesting
Full-duplex
NOMA
Relay selection
Sum rate

*Corresponding Author's Email zahabi@nit.ac.ir

Abstract

Background and Objectives: Non-orthogonal multiple access (NOMA) is a promising solution to meet a high data rate demand in the new generation of cellular networks. Moreover, simultaneous wireless information and power transfer (SWIPT) was introduced to enhance the performance in terms of energy efficiency. In this paper, a single-cell cooperative NOMA system with energy harvesting full-duplex (FD) relaying is proposed to improve the sum rate and energy efficiency.

Methods: A downlink model consisting of a base station (BS), two cell-center users (nearly located users), and two cell-edge users (far located users) are considered. In each signalling interval, the BS transmits a superposition signal of cell-center and cell-edge users based on the power domain (PD) NOMA strategy. Employing a relay selection criterion, a cell-center user is paired with a cell-edge user and acts as an FD decode and forward (DF) relay to improve the cell-edge user performance. An energy harvesting (EH) model is considered where a power splitting (PS) protocol is adopted at the relay node. The other cell-center user saves the harvested energy from the BS to exploit in the subsequent signalling intervals. Two problems of power allocation for sum rate and energy efficiency maximization in constraints of the minimum required data rate for each user and maximum transmit power at the BS are formulated for the proposed scheme. Due to the non-convexity, the optimization problems are transformed and approximated to the convex optimization problems and solved by iterative algorithms. Difference of convex (DC) programming is employed for solving the sum rate maximization problem where an effective combination of DC programming, bisection method, and Dinkelbach algorithm is utilized for dealing with the energy efficiency maximization problem.

Results: The sum rate and energy efficiency over maximum available power at the BS are presented. Also, the effects of the power splitting factor and the cell radius on the sum rate and energy efficiency are investigated. Moreover, a comparison with the OMA and NOMA schemes is studied for the different minimum required data rates.

Conclusion: Simulation results validate that the proposed scheme outperforms the OMA and NOMA schemes in terms of sum rate in all SNR regimes. Moreover, the energy efficiency of the proposed scheme achieves considerably better performance than OMA for all SNR values and obtains remarkable better performance than NOMA in most SNR values.

Introduction

With the rapid growth of wireless communications and the density of the cellular systems due to the appearance of the Internet of Things (IoT) and machine-type communications, demands for much greater data rates and efficient allocation of communication resources have

been the vital subjects. To address these aforementioned challenges, more efficient multiple access (MA) techniques have recently been proposed in 5G and beyond networks to achieve much higher system throughput and massive connectivity. Accordingly, non-orthogonal multiple access (NOMA) was introduced to make highly efficient use of the resources and serve multiple users at the same time, frequency, or code resources. In fact, the NOMA applies the superposition coding where a specific user performs the successive interference cancellation (SIC) to attain its information symbols [1]. It does this by initially decoding the signal(s) of the users with higher power levels, subtracting it from the superposed signal, and then decoding the difference as the user with a lower power level [2]. In contrast to conventional power allocation schemes, the NOMA users with weaker channel conditions are allocated more transmission power for successfully decoding their information symbols. It has been shown in [3] that the downlink NOMA with SIC can improve both the capacity and cell-edge user's throughput. In [4], the ergodic capacity maximization problem and then an optimal power allocation for multiple-input–multiple-output (MIMO) NOMA systems were proposed.

Recently, the user-cooperative relaying has been introduced into the NOMA transmission [5]-[7]. This scenario allows the users with stronger channel conditions to act as a relay for the users with weaker channel conditions. A cooperative NOMA scheme to further boost the performance of the system was proposed in [8]. Also, in [9], a downlink cooperative NOMA scenario is considered, where the base station communicates with multiple mobile users simultaneously with the help of a half-duplex amplify-and-forward (AF) relay.

It is worth noting that a half-duplex cooperative scheme might lead to spectral efficiency loss. As a result, a full-duplex (FD) relaying scheme is a promising solution to deal with this loss. In fact, the combining of cooperative NOMA and FD is a solution that is effective in achieving better spectral efficiency. A cooperative NOMA network with FD relaying was used in [10], in which the system outage probability and ergodic rate were derived. Also, in [11] a NOMA scheme with a near user as an DF relay was proposed where the resource allocation for maximizing the performance in terms of energy efficiency was achieved. In [12], the outage probability, user data rate, and energy efficiency were derived in a cooperative NOMA network with FD relaying. In addition, the performance of a full duplex relay (FDR) assisted cognitive radio (CR) network employing the NOMA scheme was investigated in [13]. In [14], the performance of an FD cooperative NOMA relaying system in the presence of imperfect successive interference cancellation (ISIC) was analysed and evaluated.

On the other hand, the FD relay node consumes more energy for the relay transmission. Hence, reducing the energy consumption of battery-assisted FD relaying users to improve the system performance in terms of energy efficiency has attracted great attention in modern communication systems. Accordingly, we need a solution to consume less energy in an efficient manner. Energy harvesting (EH) is a promising technique that leads to saving energy in a wireless network and permits an improvement in terms of energy efficiency [15]. In this regard, simultaneous wireless information and power transfer (SWIPT) was investigated first in [16]. Accordingly, the authors proposed EH on the basis of time switching (TS) and power splitting (PS) [17].

The efficient combining of SWIPT with the NOMA technique is an effective solution that both improves the system performance and saves energy. Therefore, in [18] SWIPT was applied to a cooperative NOMA system in which power allocation and PS coefficients were optimized by maximizing the energy efficiency. Also, studying a wireless-powered uplink communication system with NOMA and time-allocation method was proposed to maximize individual data rates and to improve the fairness of all users [19]. Moreover, SWIPT was applied to cooperative NOMA networks where the NOMA users near the source acted as EH relays to help far users. The authors in [20] analysed the outage probability and system throughput for a cooperative NOMA network with SWIPT and considered the impact of the PS factor on the performance of the users. The NOMA system's performance with decode-and-forward based multiple EH relays over Nakagami−m fading channels has been investigated in [21]. In [22], the authors presented a two-layered cooperative energy heterogeneous NOMA network, where each base station is powered by both the usual grid and alternative energy resources. Moreover, a joint power optimization, user association, carrier scheduling, and dynamic transmission control in dual-hop/multihop backhaul configurations of reliable NOMA HetNets with EH capability was investigated in [23].

Moreover, some new references employing FDR with SWIPT have been applied in the NOMA transmissions. In [24], the performance of a NOMA network with SWIPT based battery-assisted energy harvesting FDR in terms of outage probability has been investigated. Furthermore, the performance of a wireless powered cooperative spectrum sharing system based on NOMA transmission and a non-linear EH model with the secondary transmitter in the FD mode was analysed in [25]. The effects of beamforming on the energy efficiency in an FD user-assisted cooperative NOMA system were investigated in [26]. An FD TS-SWIPT cooperative NOMA-based IoT relay system with perfect SIC (PSIC) and ISIC was proposed in [27], where one master IoT node acts as an FD DF relay to enhance a cell-edge user's performance.

It should be mentioned that the implementation of machine learning techniques such as multi-agent deep reinforcement learning [28] and deep neural network (DNN) [29]-[31] are suggested for optimal resource management. A novel and effective deep reinforcement learning (DRL)-based approach to addressing joint resource management in a practical multi-carrier NOMA system with ISIC was presented in [32]. Also [33] has investigated a user selection and dynamic power allocation scheme in the SWIPT-NOMA relay system with DNN to optimize the user access and power allocation simultaneously to maximize the sum rate. A machine learning solution to improve harvesting energy based on clustering users was proposed in [34]. However, employing the machine learning based techniques is beyond the scope of our proposed scheme and can be considered as a candidate solution for future works.

Motivation and Contributions

To the best of our knowledge, the NOMA-FD-EH references have focused only on two users' cooperative relaying NOMA, where a cell-center NOMA user act as an FD relay node for a cell-edge NOMA user or an FD relay station after decoding of the BS transmitted symbols; retransmits the information for two users based on NOMA protocol. In contrast to these references, we introduce a model with two cell-center relaying users and two cell-edge users, where all user terminals are served by a BS in a NOMA strategy. In this case, we need to solve an optimization problem with more than two parameters (four parameters) which leads to a more challenging and general problem. Moreover, a novel relaying user selection is employed, while in the previous works there is only one cell-center user, and the relay selection is not required. It is worth noting that the relay selection criterion is not only based on the channel condition but also depends on the harvested energy and hence has a novelty. The existence of more than one cell-center user allows the unselected cell-center user in each signalling interval saves the harvested energy and accordingly improves the energy efficiency. Also, the relay selection follows a user pairing which determines the edge-user that should be paired with the selected relaying user. It should be noted that this scenario can be extended to a model with multiple cell-center users and multiple cell-edge users by employing a new pairing strategy among cell-center and cell-edge users which can exhibit better the superiority of our proposed scheme and can be considered as an attractive scenario for future works.

In this paper, we present a cooperative power domain NOMA for a downlink cellular network that consists of a BS, two cell-center users, and two cell-edge users. The BS sends the information of all users based on the NOMA strategy and a selected cell-center node adopting the energy harvesting model acts as a full-duplex relay user for the cell-edge users. The main contributions of this paper are summarized as follows.

- A cooperative NOMA-FD-EH model is investigated for improving the sum rate and energy efficiency. To the best of our knowledge, it is the first time that multiple users at both cell-center and cell-edge are suggested. The cell-center users detect their own data in addition to the cell-edge users' data and become a candidate as relay nodes for the cell-edge users. As a result, the sum rate formulation and theoretical analysis of this system model are different and more complex than the previous researches that have not been studied yet.
- A novel criterion for relay selection is proposed based on both the channel conditions between the cell-center and cell-edge users and also harvested energy level of each cell-center user. Accordingly, a cell-edge user whose date should be retransmitted is paired with the cell-center relaying user. The other cell-center user saves the energy for subsequent transmissions.
- Due to the non-convexity of the optimization problems, a suboptimal approach is proposed to obtain the power allocation for the sum rate and energy efficiency maximization by iteratively solving the approximated convex problems. We use the difference of convex (DC) programming for solving the sum rate optimization problem while an effective combination of DC programming, bisection method, and Dinkelbach algorithm is employed to efficiently solve the energy efficiency optimization problem.
- The proposed scheme is compared with the OMA and NOMA schemes. The results show the superior performance of the proposed scheme over the OMA and NOMA strategies in terms of both the sum rate and energy efficiency.

The rest of the paper is organized as follows. The system model and problems' formulations are derived. Then, the optimization problems and suboptimal power allocation algorithms for the maximization of the system's sum rate and energy efficiency are developed, respectively. The proposed algorithms' performances are evaluated by simulations and finally, the conclusion of the paper is presented.

Notation: $\mathbb{E}[x]$ is the expectation value of x. Also, $|x|$ denotes the absolute value of the complex scalar x. Moreover, $\nabla f(x_0)$ indicates the gradient of $f(x)$ at point x_0.

System Model and Problem Formulation

Let us consider a wireless network consisting of one BS and four mobile users distributed in a cell. There are two users near the BS and two users at the far locations from the BS. The cell-center users can be candidate as relay nodes for the cell-edge users. The transmission power for

the cell-center users is prepared based on employing the PS energy harvesting protocol. Both the BS and users are equipped with single transmit antenna and single receive antenna [14]. We assume that the cell-center users are operating in the FD mode. There is a direct link between BS and all mobile users but the coverage capability of BS for the edge users is potentially weak.

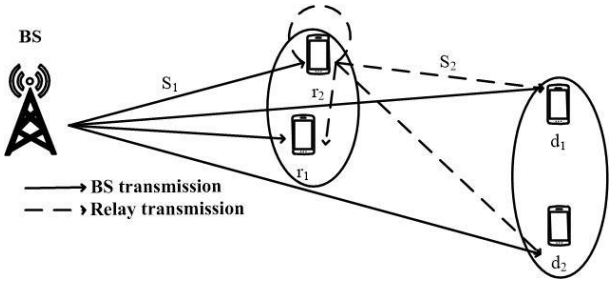


Fig. 1: System model.

The system model is presented in Fig. 1. In our system model, the BS transmits the superposition of all users' signals based on PD NOMA [35]. Since the NOMA users with strong channels can detect the messages of users with weak channels, the cell-center users (r_1 and r_2) are capable of detecting the cell-edge users' signals (d_1 and d_2) and also becoming a relay for retransmission of the cell-edge users' signals. It should be mentioned that the required power for retransmission at the relay is provided by EH in PS mode [17]. In each signalling interval, the selected cell-center user as relay node retransmits only one of the cell-edge users' signals to improve the system's performance. In other words, among the cell-center users, the one that has a maximum product of channel gain (to the cell-edge users) and harvested energy level is selected as a relay. Also, the cell-edge user whose data will be retransmitted is paired with the selected cell-center user based on the relay selection criterion.

Let h_{Bd_i} denote the channel coefficient between the BS and cell-edge user i $(i = 1,2)$ and h_{Br_i} exhibits the channel coefficient between the BS and the cell-center user $i(i = 1,2)$. Also, $h_{r_i d_j}|\, i,j\epsilon\{1,2\}$ denotes the channel coefficient between the cell-center users and the cell-edge users. The channel gains can be viewed as exponentially distributed random variables, providing that the channels are fading with Rayleigh distribution. Furthermore, $h_{r_1 r_2}$ represents the channel coefficient between two cell-center users. We assume the perfect channel state information where is obtained with negligible overhead before each signalling interval. In the proposed model, the superimposed signal transmitted by BS can be expressed as:

$$S_1(t) = \sum_{i=1}^{2} \sqrt{P_s \alpha_i} x_{r_i}(t) + \sum_{j=1}^{2} \sqrt{P_s \gamma_j} x_{d_j}(t) \quad (1)$$

where P_s denotes the BS transmit power and α_i and γ_j are the power allocation coefficients for ith cell-center user and jth cell-edge user, respectively. Also, $x_{r_i}$and x_{d_j} represent the signal of the ith cell-center user and jth cell-edge user, respectively. Without loss of generality, we consider that $|h_{Br_1}| \geq |h_{Br_2}| \geq |h_{Bd_1}| \geq |h_{Bd_2}|$, leading to power allocation coefficients in the descending order as $\gamma_2 \geq \gamma_1 \geq \alpha_2 \geq \alpha_1$ [35]. Moreover, the transmitted signal for users should be such that $\mathbb{E}\left\{|x_{r_i}|^2\right\} = 1$, $i\epsilon\{1,2\}$ and also $\mathbb{E}\left\{\left|x_{d_j}\right|^2\right\} = 1$, $j\epsilon\{1,2\}$.

Now, the criterion for the selection of the relaying user can be shown as follows:

$$r^* = \max_i \left\{ max\left(P_{r_i}|h_{r_i d_1}|^2, P_{r_i}|h_{r_i d_2}|^2\right)\right\}, i\epsilon\{1,2\} \quad (2)$$

where P_{r_i} denotes the harvested power at the ith cell-center user. It should be noted that criterion (2) in addition to the selection of relaying node, jointly determines the cell-edge user whose data will be retransmitted. Then, without loss of generality and assuming that r_2 is selected for relaying the signal of user d_1, the received signal at the relay user for information processing can be represented as:

$$y_{r_2}(t) = \sqrt{\beta_{r_2}} h_{Br_2} S_1(t) + \sqrt{\beta_{r_2}} h_{r_2} S_2(t-\tau) + n_{r_2}(t) \quad (3)$$

where $\beta_{r_2}$and $n_{r_2}(t)$ denote power splitting factor and additive white Gaussian noise (AWGN) at the relay node, respectively. Also, τ represents the delays which is caused by the processing and SIC implementation at the relay node [11]. Furthermore, $S_2(t)$ is the retransmitted signal by the relay node after detecting the cell-edge users' signals which is given by:

$$S_2(t) = \sqrt{P_{r_2}} x_{d_1}(t) \quad (4)$$

where P_{r_2} is the harvested power at the relay node before retransmission. Moreover, the total harvested power of r_2 in each signalling interval can be described as [1]:

$$P_{r_2}^{\mathrm{EH}} = \eta(1-\beta_{r_2})\mathbb{E}\left\{\left|\frac{y_{r_2}(t)}{\sqrt{\beta_{r_2}}}\right|^2\right\} \quad (5)$$

where $0 \leq \eta \leq 1$ is the energy conversion efficiency. After retransmission by the relay node and implementation of SIC and self-interference reduction, the Signal to Interference plus Noise Ratio (SINR) at r_2 is given by:

$$SINR_{r_2} = \frac{|h_{Br_2}|^2 P_s \alpha_2 \beta_{r_2}}{\frac{|h_{r_2}|^2}{\zeta} P_{r_2} \beta_{r_2} + |h_{Br_2}|^2 P_s \alpha_1 \beta_{r_2} + N_0} \quad (6)$$

where α_1 and α_2 denote the power, coefficients allocated to the r_1 and r_2, respectively and N_0 represents the noise power. Furthermore, ζ denotes the self-interference (SI) reduction factor defined as the ratio of the SI powers before and after SI suppression [11]. Also, h_{r_2} represents

the SI leakage channel of the relaying user [11] (where the SI cancellation is considered to be perfect).

Moreover, the received signal at the unselected relay (r_1) for information processing is given by:

$$y_{r_1}(t) = \sqrt{\beta_{r_1}} h_{Br_1} S_1(t) + \sqrt{\beta_{r_1}} h_{r_1 r_2} S_2(t-\tau) + n_{r_1}(t) \quad (7)$$

Considering β_{r_1} denotes the power splitting factor, the total harvested power at r_1 in each signalling interval can be expressed as [1]:

$$P_{r_1}^{\mathrm{EH}} = \eta\left(1-\beta_{r_1}\right)\mathbb{E}\left\{\left|\frac{y_{r_1}(t)}{\sqrt{\beta_{r_1}}}\right|^2\right\} \quad (8)$$

Also, the SINR at the unselected cell-center user (r_1) with the strongest channel gain after SIC implementation and cancellation of the signal from the relaying user (according to the awareness of the cell-edge users' signals) is presented by:

$$SINR_{r_1} = \frac{\left|h_{Br_1}\right|^2 P_s \alpha_1 \beta_{r_1}}{N_0} \quad (9)$$

Now, we represent the received signal at two cell-edge users as follows:

$$y_{d_i}(t) = h_{Bd_i} S_1(t) + h_{r_2 d_i} S_2(t-\tau) + n_{d_i}(t), \quad i\epsilon\{1,2\} \quad (10)$$

After SIC implementation and cancellation of the signal of user d_2, the SINR at the cell-edge user d_1 is given by (assuming the phase of the transmitted signal from the relay node is shifted to co-phase the received signals from the relay node and the BS at d_1):

$$SINR_{d_1} = \frac{\left|h_{Bd_1}\right|^2 P_s \gamma_1 + \left|h_{r_2 d_1}\right|^2 P_{r_2}}{\left|h_{Bd_1}\right|^2 P_s(1-\gamma_1-\gamma_2) + N_0} \quad (11)$$

Furthermore, the SINR equation for detection of the other cell-edge user (d_2) signal is exhibited as follows:

$$SINR_{d_2} = \frac{\left|h_{Bd_2}\right|^2 P_s \gamma_2}{\left|h_{Bd_2}\right|^2 P_s(1-\gamma_2) + N_0} \quad (12)$$

In the following, we present the theoretical analyses for the sum rate and energy efficiency, respectively.

A. Sum Rate Analysis

In this section, we will discuss the performance of the proposed scheme and analyze the sum rate maximization with the constraints on the total consumption power and minimum rate requirement for each user. In the other word, the optimal power allocation will be calculated such that the proposed scheme achieves the best performance in terms of sum rate on the given constraints. Accordingly, the problem formulation in terms of sum rate optimization can be represented as follows:

$$\max_{p_1,\dots,p_4} R_{sum} = \sum_{i=1}^{4} R_i \quad (13)$$

$$s.t.\, C1: \sum_{i=1}^{4} p_i \le P_s^{max}, \quad p_1 \le p_2 \le p_3 \le p_4$$
$$C2-C5: \quad R_i \ge R_i^{min}, \quad i = 1,\dots,4$$

where:

$$\begin{aligned} R_1 &= \log\left(1+SINR_{r_1}\right) \\ R_2 &= \log\left(1+SINR_{r_2}\right) \\ R_3 &= \log\left(1+SINR_{d_1}\right) \\ R_4 &= \log\left(1+SINR_{d_2}\right) \\ p_1 &= \alpha_1 P_s \\ p_2 &= \alpha_2 P_s \\ p_3 &= \gamma_1 P_s \\ p_4 &= \gamma_2 P_s \end{aligned} \quad (14)$$

where $\{R_i, p_i, i = 1,2\}$ and $\{R_i, p_i, i = 3,4\}$ represent the achievable rate equations and power assignments to the cell-center ($r_1$and r_2) and cell-edge ($d_1$and d_2) users, respectively. Also, P_s^{max} is the maximum power at the BS and R_i^{min} indicates the minimum required data rate for R_i. Due to the existence of the power allocation parameters at both the numerator and denominator of the achievable rate equations, the cost function is not convex. As a result, the problem (13) in its original form is neither a convex nor quasi-convex problem. Nevertheless, we show that it can be transformed into a convex problem via a linear transformation of the optimization variables [36].

Now we introduce the following variable transformation: $q_i = \sum_{j=1}^{i} p_j$ for $i = 1,2,\dots,4$ or conversely $p_i = q_i - q_{i-1}$ for $i = 2,\dots,4$ and $p_1 = q_1$.

Assuming that $I_1 = N_0$, $I_2 = \frac{\left|h_{r_2}\right|^2}{\zeta} P_{r_2}\beta_{r_2} + N_0$, $I_3 = \left|h_{r_2 d_1}\right|^2 P_{r_2} + N_0$, $I_4 = N_0$ and $I_5 = N_0$ we will have:

$$\begin{aligned} R_1 &= \log\left(\frac{I_1 + \left|h_{Br_1}\right|^2 q_1 \beta_{r_1}}{I_1}\right) \\ R_2 &= \log\left(\frac{I_2 + \left|h_{Br_2}\right|^2 q_2 \beta_{r_2}}{I_2 + \left|h_{Br_2}\right|^2 q_1 \beta_{r_2}}\right) \\ R_3 &= \log\left(\frac{I_3 + \left|h_{Bd_1}\right|^2 q_3}{I_4 + \left|h_{Bd_1}\right|^2 q_2}\right) \\ R_4 &= \log\left(\frac{I_5 + \left|h_{Bd_2}\right|^2 q_4}{I_5 + \left|h_{Bd_2}\right|^2 q_3}\right) \end{aligned} \quad (15)$$

These functions are still non-convex and also non-concave, but with the help of the following presentation and conversion of the maximization problem to a minimization problem, it is possible to define the cost function as a difference of two convex functions and consequently, we will have a DC programming with the convex constraints [37]. In our model, the two convex functions are presented as follows:

$$F(\boldsymbol{q}) = -\Big[\log\Big(I_1 + \big|h_{Br_1}\big|^2 q_1\beta_{r_1}\Big) + \log\Big(I_2 + \big|h_{Br_2}\big|^2 q_2\beta_{r_2}\Big) + \log\Big(I_3 + \big|h_{Bd_1}\big|^2 q_3\Big) + \log\Big(I_5 + \big|h_{Bd_2}\big|^2 q_4\Big)\Big]$$
$$G(\boldsymbol{q}) = -\Big[\log\Big(I_2 + \big|h_{Br_2}\big|^2 q_1\beta_{r_2}\Big) + \log\Big(I_4 + \big|h_{Bd_1}\big|^2 q_2\Big) + \log\Big(I_5 + \big|h_{Bd_2}\big|^2 q_3\Big)\Big] \quad (16)$$

It is obvious that both functions are the sum of the several convex functions and as a result, will be convex. Based on variable transformation, the constraint $C1$ in (13) changes from $\sum_{i=1}^{4} p_i \le P_s^{max}$ to $\sum_{i=1}^{4} p_i = q_4 \le P_s^{max}$. In addition, $p_1 \le p_2$, $p_2 \le p_3$, and $p_3 \le p_4$convert to $q_1 \le q_2 - q_1$, $q_2 - q_1 \le q_3 - q_2$, and $q_3 - q_2 \le q_4 - q_3$, respectively. For the constraints $C2 - C5$, the logarithmic functions substitute with their corresponding linear forms. For example, the constraint $R_1 = \log\left(\frac{I_1+|h_{Br_1}|^2 q_1\beta_{r_1}}{I_1}\right) = \log\left(1 + \frac{|h_{Br_1}|^2 q_1\beta_{r_1}}{I_1}\right) \ge R_1^{min}$ easily converts to $q_1 \ge \frac{I_1\left(2^{R_1^{min}}-1\right)}{|h_{Br_1}|^2\beta_{r_1}}$ and so on for the constraints $C3 - C5$. Finally, based on (16) and the aforementioned substitutions, the problem (13) will be transformed to:

$$\begin{aligned}
&\min_{q_1,\dots,q_4} Q(\boldsymbol{q}) = F(\boldsymbol{q}) - G(\boldsymbol{q})\\
&s.t.\ C1\colon q_4 \le P_s^{max},\ q_1 \le q_2 - q_1 \le q_3 - q_2 \le q_4 - q_3\\
&C2\colon q_1 \ge \frac{I_1\left(2^{R_1^{min}} - 1\right)}{\left|h_{Br_1}\right|^2\beta_{r_1}}\\
&C3\colon q_1 \le 2^{-R_2^{min}} q_2 + \frac{\left(2^{-R_2^{min}} - 1\right) I_2}{\left|h_{Br_2}\right|^2\beta_{r_2}}\\
&C4\colon q_2 \le 2^{-R_3^{min}} q_3 + \frac{\left(2^{-R_3^{min}} I_3 - I_4\right)}{\left|h_{Bd_1}\right|^2}\\
&C5\colon q_3 \le 2^{-R_4^{min}} q_4 + \frac{\left(2^{-R_4^{min}} - 1\right) I_5}{\left|h_{Bd_2}\right|^2}
\end{aligned} \quad (17)$$

For analyzing and solving a DC programming problem, for function $G(\boldsymbol{q})$, we must have an approximation by its linear form as $G^n(\boldsymbol{q}) = G\left(\boldsymbol{q}^{(n)}\right) + \nabla G^T\left(\boldsymbol{q}^{(n)}\right)\left(\boldsymbol{q} - \boldsymbol{q}^{(n)}\right)$ where $G\left(\boldsymbol{q}^{(n)}\right)$ and $\nabla G^T\left(\boldsymbol{q}^{(n)}\right)$ are the value and gradient of the $G(\boldsymbol{q})$ at the point $\boldsymbol{q}^{(n)}$, respectively. Now, with the convexity of $F(\boldsymbol{q})$ and the linearity of $G^n(\boldsymbol{q})$, the cost function is convex. Therefore, due to the linear constraints ($C1 - C5$), the problem (17) is convex and can be efficiently solved via convex optimization methods. Algorithm 1 illustrates the process of sum rate optimization.

Algorithm 1 suboptimal power allocation in sum rate maximization problem

1. Set iteration number $n = 0$
2. **initialize** $\boldsymbol{q}^{(0)} = \mathbf{0}$
3. **Repeat** Steps (5) to (7) until $\left|Q\left(\boldsymbol{q}^{(n+1)}\right) - Q\left(\boldsymbol{q}^{(n)}\right)\right| \le \epsilon$
4. Set $q_4{}^{(n)} = P_s^{max}$ for any n
5. Define convex approximation of $Q^n(\boldsymbol{q})$ at $\boldsymbol{q}^{(n)}$ as $Q^n(\boldsymbol{q}) = F(\boldsymbol{q}) - G^n(\boldsymbol{q}) = F(\boldsymbol{q}) - G(\boldsymbol{q}^{(n)}) - \nabla G^T(\boldsymbol{q}^{(n)})\left(\boldsymbol{q} - \boldsymbol{q}^{(n)}\right)$
6. Solve the convex problem
$$\boldsymbol{q}^{(n+1)} = \underset{\boldsymbol{q}}{argmin}\, Q^n(\boldsymbol{q})$$
$s.t.$ $C1 - C5$ of (17)
7. $n \leftarrow n + 1$

It should be mentioned that the value for q_4 will always be P_s^{max} in the minimization problem of (17), because the cost function is a decreasing function based on q_4. So, we assume in algorithm 1 that $q_4 = P_s^{max}$.

B. *Energy Efficiency Analysis*

In the following, the energy efficiency maximization can be defined and then solved. First, it is worth noting that we define energy efficiency as the achievable sum rate over total power consumption. The energy efficiency optimization problem is derived in (18) in which P_c indicates the circuit power consumption.

$$\begin{aligned}
&\max_{p_1,\dots,p_4} \sum_{i=1}^{4} R_i \Big/ \left(\sum_{i=1}^{4} p_i + P_c\right)\\
&s.t.\ \sum_{i=1}^{4} p_i \le P_s^{max},\ p_1 \le p_2 \le p_3 \le p_4\\
&R_i \ge R_i^{min},\ i = 1,\dots,4 \quad (R_i \text{ based on (14)})
\end{aligned} \quad (18)$$

The energy efficiency problem is neither a convex nor quasi-convex problem. But, with the help of transformation similar to the sum rate analysis, it can be transformed into a problem such as the following form:

$$\begin{aligned}
&\min_{q_1,\dots,q_4} \big(M(\boldsymbol{q}) - N(\boldsymbol{q})\big)/(q_4 + P_c)\\
&s.t.\ C1\colon q_4 \le P_s^{max},\ q_1 \le q_2 - q_1 \le q_3 - q_2 \le q_4 - q_3\\
&C2\colon q_1 \ge \frac{I_1\left(2^{R_1^{min}} - 1\right)}{\left|h_{Br_1}\right|^2\beta_{r_1}}\\
&C3\colon q_1 \le 2^{-R_2^{min}} q_2 + \frac{\left(2^{-R_2^{min}} - 1\right) I_2}{\left|h_{Br_2}\right|^2\beta_{r_2}}\\
&C4\colon q_2 \le 2^{-R_3^{min}} q_3 + \frac{\left(2^{-R_3^{min}} I_3 - I_4\right)}{\left|h_{Bd_1}\right|^2}\\
&C5\colon q_3 \le 2^{-R_4^{min}} q_4 + \frac{\left(2^{-R_4^{min}} - 1\right) I_5}{\left|h_{Bd_2}\right|^2}
\end{aligned} \quad (19)$$

where,

$$M(\boldsymbol{q}) = -\Big[\log\Big(I_1 + \big|h_{Br_1}\big|^2 q_1\beta_{r_1}\Big) + \log\Big(I_2 + \big|h_{Br_2}\big|^2 q_2\beta_{r_2}\Big) + \log\Big(I_3 + \big|h_{Bd_1}\big|^2 q_3\Big) + \log\Big(I_5 + \big|h_{Bd_2}\big|^2 q_4\Big)\Big] \quad (20)$$

$$N(\boldsymbol{q}) = -\Big[\log\Big(I_2 + \big|h_{Br_2}\big|^2 q_1\beta_{r_2}\Big) + \log\Big(I_4 + \big|h_{Bd_1}\big|^2 q_2\Big) + \log\Big(I_5 + \big|h_{Bd_2}\big|^2 q_3\Big)\Big]$$

As can be observed, both functions $M(\boldsymbol{q})$ and $N(\boldsymbol{q})$ are convex functions resulting in a difference of the two convex functions at the numerator of the fractional cost function.

Similar to the sum rate optimization problem, we exploit the linear approximation of $N(\boldsymbol{q})$ such that $N^k(\boldsymbol{q}) = N(\boldsymbol{q}^{(k)}) + \nabla N^T(\boldsymbol{q}^{(k)})(\boldsymbol{q} - \boldsymbol{q}^{(k)})$ to achieve a convex function at the numerator. For exploiting the Dinkelbach [38], it is necessary to have a concave function at the numerator and a convex function at the denominator of the fractional cost function. The denominator is a linear function of the problem parameters and hence is a convex function. On the other hand, with the conversion of minimization problem to a maximization one, the numerator will be concave. Now, it is possible to solve the problem by using the Dinkelbach algorithm. On the basis of the Dinkelbach algorithm, the following objective function should be introduced:

$$H(\boldsymbol{q}, \lambda) = \big(M(\boldsymbol{q}) - N(\boldsymbol{q})\big) - \lambda(q_4 + P_c) \quad (21)$$

where λ is a positive parameter. The optimal solution can be found by solving the problem parameterized by λ such that $H(\boldsymbol{q}, \lambda) = 0$ [36]. Consequently, the optimization problem is transformed and given by:

$$\max_{q_1,\dots,q_4} H(\boldsymbol{q}, \lambda) = \big(N(\boldsymbol{q}) - M(\boldsymbol{q})\big) - \lambda(q_4 + P_c) \quad s.t. \quad C1, -C5 \text{ of (19)} \quad (22)$$

Algorithm 2 illustrates the optimal power allocation in the energy efficiency maximization problem based on Dinkelbach approach. It should be noticed that considering $q_4 = P_s^{max}$ is not optimal anymore similar to sum rate optimization problem, because the numerator is a logarithmic function of q_4 while the denominator of the cost function is a linear function of q_4. Hence, the energy efficiency will not improve with the increasing the q_4 and hence we cannot employ the maximum value for q_4. But, knowing the fact that inequality $0 \leq q_4 \leq P_s^{max}$ is always established, it is possible to adopt the bisection algorithm to achieve q_4 and subsequently the other parameters based on algorithm 2. As a result, we employ the bisection with the combination of Dinkelbach algorithm to solve the optimization problem in (22).

Algorithm 2 suboptimal power allocation in energy efficiency maximization problem based on Dinkelbach algorithm

1. Set iteration number $k = 0$
2. **Initialize** $\boldsymbol{q}^{(0)} = \mathbf{0}$ **and** $\lambda^{(0)} = 0$.
3. Set $q_{4_{LB}}^{(0)} = 0$ and $q_{4_{UB}}^{(0)} = P_s^{max}$
4. **Repeat** Steps (5) to (12) until $\big|H(\boldsymbol{q}^{(k+1)})\big| \leq \epsilon_1$
5. **While** $q_{4_{UB}}^{(k)} - q_{4_{LB}}^{(k)} \geq \epsilon_2$ **do**
6. Set $q_4{}^{(k)} = \big(q_{4_{LB}}^{(k)} + q_{4_{UB}}^{(k)}\big)/2$
7. Define convex approximation of $H^{(k)}(\boldsymbol{q})$ at $\boldsymbol{q}^{(k)}$ as $H^k(\boldsymbol{q}) = M(\boldsymbol{q}) - N^k(\boldsymbol{q}) - \lambda^{(k)}\big(q_4{}^{(k)} + P_c\big) = M(\boldsymbol{q}) - N(\boldsymbol{q}^{(k)}) - \nabla N^T(\boldsymbol{q}^{(k)})(\boldsymbol{q} - \boldsymbol{q}^{(k)}) - \lambda^{(k)}(q_4{}^{(k)} + P_c)$
8. Solve the convex problem
   $\boldsymbol{q}^{(k+1)} = \underset{\boldsymbol{q}}{argmax}\, H^k(\boldsymbol{q})$
   $s.t. \quad C1 - C5$ of (19)
9. **if** $R_1^{(k)} \leq R_1^{min}$ **then**
   set $q_{4_{LB}}^{(k)} = \big(q_{4_{LB}}^{(k)} + q_{4_{UB}}^{(k)}\big)/2$
   else
   set $q_{4_{UB}}^{(k)} = \big(q_{4_{LB}}^{(k)} + q_{4_{UB}}^{(k)}\big)/2$
10. $\lambda^{(k+1)} = \big(M(\boldsymbol{q}) - N^{(k)}(\boldsymbol{q})\big)/\big(q_4{}^{(k)} + P_c\big)$
11. **if** $R_1^{(k)} \leq R_1^{min}$ **then**
    set $q_{4_{LB}}^{(k+1)} = q_{4_{LB}}^{(k)}$ and $q_{4_{UB}}^{(k+1)} = P_s^{max}$
    else
    set $q_{4_{LB}}^{(k+1)} = 0$ and $q_{4_{UB}}^{(k+1)} = q_{4_{UB}}^{(k)}$
12. $k \leftarrow k + 1$

In algorithm 2, LB and UB indices address the lower bound and upper bound, respectively.

Feasibility and Computational Complexity

As we know, a feasible solution for an optimization problem is a solution that satisfies all constraints that the program is subjected to, where it does not violate even a single constraint. In the sum rate optimization problem, the constraint on the maximum power of BS ($q_4 = P_s^{max}$) is always satisfied as equality. But it should be noted that for the satisfaction of the other constraints, inequalities $q_{i+1} \geq q_i$, $i = 1,2,3$ must be established. Hence, the feasibility occurs when these inequalities are satisfied that are dependent on the values of $\{R_i^{min}, i = 1, \dots, 4\}$, $\{I_i, i = 1, \dots, 5\}$, and also channels' gains. Moreover, another condition for the satisfaction of the constraints C_2 and C_3 is that $2^{-R_2^{min}} q_2 + \frac{\left(2^{-R_2^{min}} - 1\right) I_2}{|h_{Br_2}|^2 \beta_{r_2}} \geq \frac{I_1\left(2^{R_1^{min}} - 1\right)}{|h_{Br_1}|^2 \beta_{r_1}}$ to guarantee the accuracy of the obtained value for q_1. In the case of energy efficiency, conditions for the constraints are similar to the sum rate optimization, except that the constraint on the P_s^{max} must be satisfied in the form of inequality.

The computational complexity of the optimization problems depends on the utilized algorithms for solving the problems. Hence, we describe the complexity order

for each employed algorithm in the proposed optimization problems. From the general convergence properties of the DC algorithm, it has a linear convergence and only relies on a few basic operations, which leads to a low computational cost [39]. In fact, the DC programming optimization problem can be solved by using standard algorithms from convex optimization theory such as the interior point method and sequential quadratic programming [37]. Therefore, the complexity of the DC algorithm within some tolerance measured by ϵ is in the order of $\mathcal{O}(\sqrt{N}\log(N/\epsilon))$ [40], where N is the number of optimization problem parameters. On the other hand, the bisection line search is known to find an ϵ-accurate solution within the number of iterations bounded by $\mathcal{O}(\log(\epsilon_0/\epsilon))$, where $\epsilon_0 = |b-a|$ is the initial difference between the upper and lower bounds in the bisection method. Therefore, the overall complexity is bounded by $\mathcal{O}(N\log(\epsilon_0/\epsilon))$, which is linear in N [41]. It is worth noting that in our problem, $N = 1$ for the bisection method; because only one of the parameters is obtained by bisection.

Moreover, the convergence rate of Dinkelbach algorithm is super linear [42]. Assuming upper and lower bounds of the maximum energy efficiency value are available as U and L, we could find the optimal values updating λ according to the bisection method, instead of using Dinkelbach's update criterion. Although bisection converges typically slower than Dinkelbach method [42], it provides an estimate of Dinkelbach algorithm. By using the bisection method, the overall asymptotic complexity can be found within a tolerance ϵ with $\mathcal{O}(N\log(\lceil(U-L)/\epsilon\rceil))$ iterations [42]. As can be seen, the employed algorithms have appropriate conditions from the complexity point of view.

Results and Discussion

Simulation results are illustrated in this section to validate our theoretical works in terms of sum rate and energy efficiency. We consider a single-cell scenario employing NOMA-FD-EH with the distributed users in the cell. It is assumed that the cell-center and cell-edge users are randomly distributed within a disk with radius 5 meters and a ring with radii 5 and 10 meters, respectively [43].

The Path loss exponent is considered as 2 [26] and the noise power N_0 is 0.25. Energy harvesting efficiency η is set to 0.5 [44], and the minimum target data rate (R_i^{min}) is equal for all users where is considered to be $R_i^{min} = 0.5\ bps/Hz$ and $R_i^{min} = 1\ bps/Hz$ [26] in the simulations. As we mentioned, energy harvesting is based on the PS method for which the power splitting factor is set to $\beta_{r_1} = \beta_{r_2}$. In addition, circuit power consumption is set to $P_c = 0.1$ Watt. The OMA and optimal NOMA (optimal four users NOMA scheme with sum power and minimum achievable rate constraints) are two techniques that are considered for comparison with the proposed scheme. It is worth noting that the simulations are generated from 100000 independent realizations of different channel conditions. Also, the iteration error tolerances ϵ_1 and ϵ_2 are 0.001 [26];

Fig. 2 illustrates the sum rate performance of the proposed scheme versus the splitting factor for different SNRs (the maximum available power at the BS to the noise ratio) and various minimum target data rates. It can be seen that the optimum splitting factors for both $R_i^{min} = 0.5\ bps/Hz$ and $R_i^{min} = 1\ bps/Hz$ are achieved at the low values of the splitting factor. It means that more harvested power leads to improving the sum rate performance, especially in low SNR regimes, where the BS power is low and hence the requirement to relaying is more sensible. On the other hand, with the reduction of the minimum target data rate, the sum rate increases. Moreover, the performance is approximately the same for the different splitting factors in the high SNR values. It should be noted that, when the splitting factor is very small and approximately all of the BS power is employed for relaying, the performance degrades a little; where the cell-center users' rates approach zero. These results are used in the following simulations.

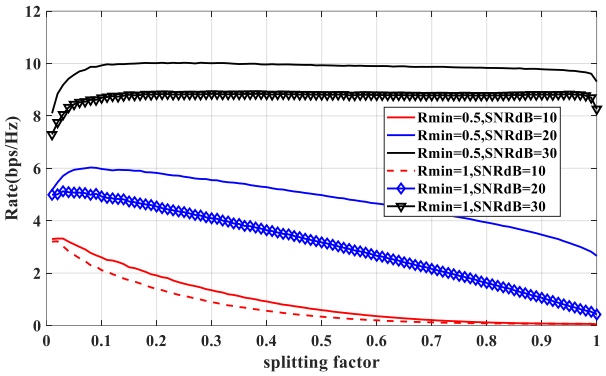


Fig. 2: Sum rate of the proposed scheme versus splitting factor for different SNRs and minimum target data rates ($R_i^{min} = 0.5\ ,1\ bps/Hz$).

In Fig. 3, the sum rate performance of the proposed scheme over SNR is compared with the OMA and optimal NOMA with the splitting factor of 0.15 in two cases as $R_i^{min} = 0.5\ bps/Hz$ and $R_i^{min} = 1\ bps/Hz$. As can be seen from this figure, the proposed scheme considerably outperforms the OMA and optimal NOMA schemes, especially in low SNR values for both minimum target data rates.

Also, the performance of the sum rate improves with the increase of SNR. On the other hand, it is obvious that the only proposed scheme has non-zero values in the SNR values less than 10dB, while the other schemes have zero values in this SNR regime.

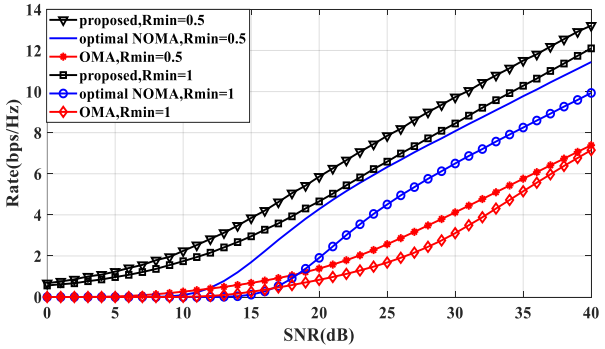


Fig. 3: Comparison of the proposed scheme with the OMA and optimal NOMA schemes in terms of sum rate over SNR for two minimum target data rates ($R_i^{min} = 0.5, 1\ bps/Hz$).

Fig. 4 illustrates the performance in terms of energy efficiency for the proposed scheme based on splitting factor for different SNRs and minimum target data rates where $R_i^{min} = 0.5\ bps/Hz$ and $R_i^{min} = 1\ bps/Hz$. For both minimum target data rates and with the increasing of the SNR, energy efficiency achieves higher values at the larger splitting factors. It means that in the low SNR regimes we need a smaller splitting factor and vice versa for achieving better performance. On the other hand, when the splitting factor has a high value and the harvesting is ignored, energy efficiency approaches zero in all cases. Moreover, the maximum energy efficiency in the case of $R_i^{min} = 0.5\ bps/Hz$ achieves with a larger splitting factor in comparison with the case $R_i^{min} = 1\ bps/Hz$.

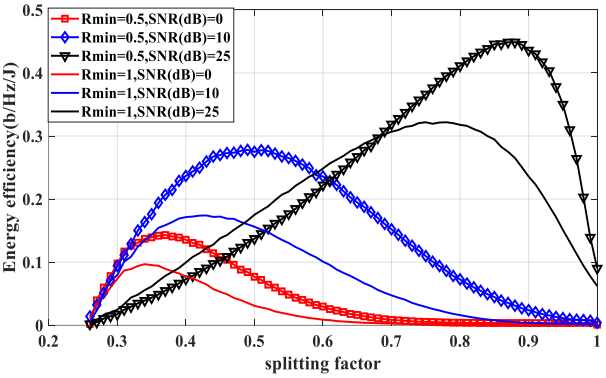


Fig. 4: Energy efficiency of the proposed scheme versus splitting factor for different SNRs and minimum target data rates ($R_i^{min} = 0.5, 1\ bps/Hz$).

As can be seen, there is no same optimum value for the splitting factor in all SNR regimes. As a result, we adopt a moderate value for the splitting factor to be employed in the energy efficiency.

In Fig. 5, the energy efficiency performance of the proposed scheme versus SNR is compared with the OMA and optimal NOMA in two cases as $R_i^{min} = 0.5\ bps/Hz$ and $R_i^{min} = 1\ bps/Hz$ with a splitting factor of 0.6 and 0.5, respectively. It can be seen from the figure that in both minimum target data rates the proposed scheme considerably outperforms the OMA in all SNR regimes and achieves higher values in comparison to the optimal NOMA in almost all ranges of SNR. However, in the very high SNR regimes, optimal NOMA approaches a constant value, while the proposed scheme has small values. This probably stems from that in the proposed scheme we don't subtract the harvested power of the unselected cell-center user in each signaling from the total BS transmit power. Moreover, based on the results in the Fig. 4 if we employed the smaller splitting factor, the performance would improve in the low SNR regimes and by utilizing the larger value for the splitting factor, energy efficiency would increase in the high SNR values.

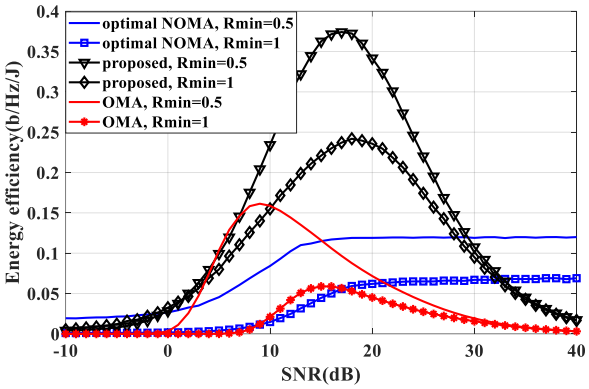


Fig. 5: Comparison of the proposed scheme energy efficiency with the OMA and optimal NOMA over SNR for $R_i^{min} = 0.5\ bps/Hz$ and $R_i^{min} = 1\ bps/Hz$.

Fig. 6 and Fig. 7 show the performances of the proposed, OMA, and NOMA schemes versus the the cell radius. It should be noted that all distances among nodes scale with the increase of the cell radius from 10 meters to 40 meters. Also, the minimum target data rates are set to $R_i^{min} = 0.5\ bps/Hz$ and $R_i^{min} = 1\ bps/Hz$ and SNR is considered as 20 dB for the sum rate and as 30 dB for the energy efficiency. Moreover, for the energy efficiency analysis in the Fig. 7, the splitting factors are considered as 0.6 and 0.5 for $R_i^{min} = 0.5\ bps/Hz$ and $R_i^{min} = 1\ bps/Hz$, respectively. As can be seen from the Fig. 6 and Fig. 7, increasing the cell radius leads to a decrease in the performance for both the sum rate and energy efficiency. However, the proposed scheme achieves better performance in terms of sum rate and energy efficiency for all cell radius values in both cases where $R_i^{min} = 0.5\ bps/Hz$ and $R_i^{min} = 1\ bps/Hz$. However increasing the cell radius up to 40 meters results in energy efficiency degradation for all schemes, the speed of this reduction is very slower in the proposed scheme. In addition, when the radius of the cell increases from 10 up to 15 meters, the performance of the proposed scheme improves. This likely stems from the less optimal transmitted power from the BS, because the sum rate decreases with the increase of the cell dimension.

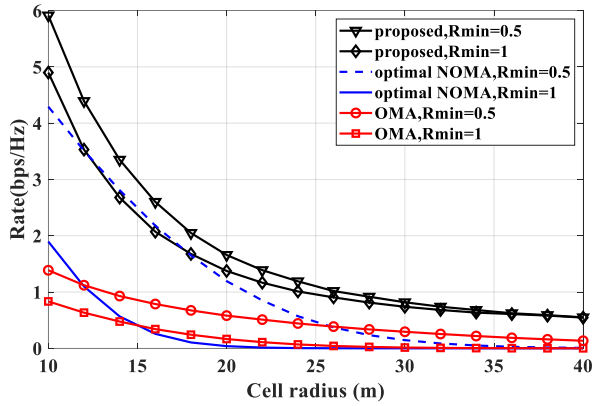


Fig. 6: Sum rate versus the cell radius for the proposed, OMA and optimal NOMA schemes with $R_i^{min} = 0.5\ bps/Hz$ and $R_i^{min} = 1\ bps/Hz$ and SNR of 20 dB.

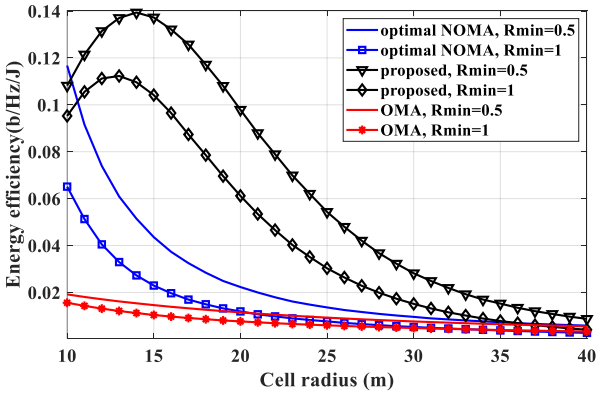


Fig. 7: Energy efficiency versus the cell radius for the proposed, OMA and optimal NOMA schemes with $R_i^{min} = 0.5\ bps/Hz$ and $R_i^{min} = 1\ bps/Hz$ and SNR of 30 dB.

Fig. 8 depicts the sum rate versus energy conversion efficiency for different SNRs and minimum target data rates as $R_i^{min} = 0.5\ bps/Hz$ and $R_i^{min} = 1\ bps/Hz$. It can be concluded from the figure that obviously approaching the η to the value of 1 improves the performance. It is worth noting that the same result is achieved for the energy efficiency over the value of η.

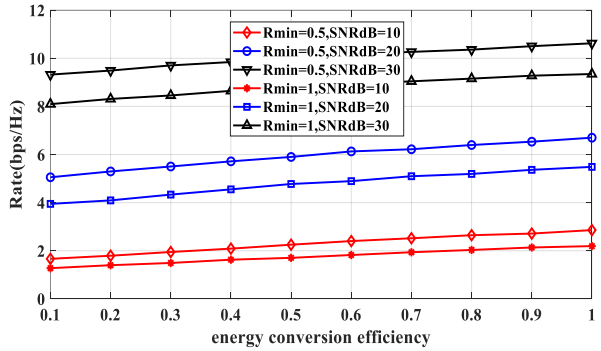


Fig. 8: Sum rate versus energy conversion efficiency for different SNRs and minimum target data rates as $R_i^{min} = 0.5\ ,1\ bps/Hz$.

Conclusion

This paper investigated a communication scheme, which efficiently combines cooperative NOMA, FD relaying, and EH techniques. Employing cooperative NOMA, the BS aims to broadcast the information to the mobile users that are distributed in a cell. Two users deployed in the near of the BS while two other users located at the far locations. In each signalling interval, one of the cell-center users is paired with a cell-edge user where the cell-center user employing PS protocol retransmits the cell-edge user's data. The DC programming and an efficient combination of DC, bisection method, and Dinkelbach algorithm were proposed to assign the suboptimal power allocations for maximizing the sum rate and energy efficiency performances, respectively. The numerical results demonstrated that when the SNR is 30dB, the square cell dimension is 10 meters, and $R_i^{min} = 1\ bps/Hz$, the proposed scheme using the appropriate values splitting factor significantly enlarges the system sum rate by 33% over optimal NOMA. This superiority approaches to 120% at SNR equal to 20dB. Moreover, the energy efficiency enhancement of the proposed scheme over optimal NOMA for $R_i^{min} = 1\ bps/Hz$, cell dimension equal to 10 meters, and SNR as 30dB and 20dB approach 200% and 40%, respectively. However, MIMO NOMA can be considered in the proposed schemes. Also, achieving the unequal optimal splitting factors is a suggestion for future works.

Author Contributions

This paper has been exploited from the M. B. Noori Shirazi's Ph.d. thesis supervised by M. R. Zahabi. M. B. Noori Shirazi and M. R. Zahabi proposed the main idea of the paper. M. B. Noori Shirazi performed the analyses and simulations and wrote the manuscript. M. R. Zahabi interpreted the results and corrected the manuscript.

Acknowledgment

Beforehand, the authors would like to thank the honorable editor and reviewers.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

NOMA	Non-Orthogonal Multiple Access
BS	Base Station
PD	Power Domain
FD	Full-Duplex
DF	Decode and Forward

EH	Energy Harvesting
PS	Power Splitting
IoT	Internet of Things
SIC	Successive Interference Cancellation
MIMO	Multiple-Input–Multiple-Output
AF	Amplify-and-Forward
SWIPT	Simultaneous Wireless Information and Power Transfer
TS	Time Switching
FDR	Full Duplex Relay
CR	Cognitive Radio
ISIC	Imperfect Successive Interference Cancellation
PSIC	Perfect Successive Interference Cancellation
DC	Difference of Convex
SI	Self-Interference
SNR	Signal to Noise Ratio
SINR	Signal to Interference plus Noise Ratio

Reference

[1] S. L. Wang, T. M. Wu, "Stochastic geometric performance analyses for the cooperative NOMA with the full-duplex energy harvesting relaying," IEEE Trans. Veh. Technol., 68(5): 4894-4905, 2019.

[2] Z. Elsaraf, F. A. Khan, Q. Z. Ahmed, "Deep learning based power allocation schemes in NOMA systems: A review," in Proc. 26th International Conference on Automation and Computing (ICAC): 1-6, 2021.

[3] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in Proc. IEEE 77th Vehicular Technology Conference: 1-5, 2013.

[4] Q. Sun, S. Han, I. Chin-Lin, Z. Pan,"On the ergodic capacity of MIMO NOMA systems," IEEE Wireless Commun. Lett., 4(4): 405–408, 2015.

[5] L. Lv, J. Chen, Q. Ni, ''Cooperative non-orthogonal multiple access in cognitive radio,'' IEEE Commun. Lett., 20(10): 2059-2062, 2016.

[6] L. Lv, J. Chen, Q. Ni, Z. Ding, ''Design of cooperative nonorthogonal multicast cognitive multiple access for 5G systems: User scheduling and performance analysis,'' IEEE Trans. Commun., 65(6): 2641-2656, 2017.

[7] Y. Yuan, P. Xu, Z. Yang, Z. Ding, Q. Chen, "Joint robust beamforming and power-splitting ratio design in SWIPT-based cooperative NOMA systems with CSI uncertainty," IEEE Trans. Veh. Technol., 68(3): 2386-2400, 2019.

[8] Z. Ding, M. Peng, H. V. Poor, "Cooperative non-orthogonal multiple access in 5G systems," IEEE Commun. Lett., 19(8): 1462–1465, 2015.

[9] J. Men, J. Ge, "Performance analysis of non-orthogonal multiple access in downlink cooperative network," IET Commun., 9(18): 2267–2273, 2015.

[10] L. Zhang, J. Liu, M. Xiao, G. Wu, Y. Liang, S. Li, "Performance analysis and optimization in downlink NOMA systems with cooperative full-duplex relaying," IEEE J. Sel. Areas Commun., 35(10): 2398–2412, 2017.

[11] Z. Wei, X. Zhu, S. Sun, J. Wang, L. Hanzo, "Energy-efficient full-duplex cooperative non-orthogonal multiple access," IEEE Trans. Veh. Technol., 67(10): 10123-10128, 2018.

[12] X.Yue, Y. Liu, S. Kang, A. Nallanathan, Z. Ding, "Exploiting full/half-duplex user relaying in NOMA systems," IEEE Trans. Commun., 66(2): 560–575, 2018.

[13] V. S. Babu, N. Deepan, B. Rebekka. "Performance analysis of cooperative full duplex NOMA system in cognitive radio networks," in Proc. IEEE International Conference on Wireless Communications Signal Processing and Networking: 84-87, 2020.

[14] X. Li, M. Liu, C. Deng, P. T. Mathiopoulos, Z. Ding, Y. Liu, "Full-duplex cooperative NOMA relaying systems with I/Q imbalance and imperfect SIC," IEEE Wireless Commun. Lett., 9(1): 17-20, 2019.

[15] S. Guo, Y. Shi, Y. Yang, B. Xiao, "Energy efficiency maximization in mobile wireless energy harvesting sensor networks," IEEE Trans. Mobile Comput., 17(7): 1524-1537, 2017.

[16] L. R. Varshney, "Transporting information and energy simultaneously," in Proc. IEEE international symposium on information theory: 1612- 1616, 2008.

[17] X. Zhou, R. Zhang, C. K. Ho, "Wireless information and power transfer: Architecture design and rate-energy tradeoff," IEEE Trans. Commun., 61(11): 4754–4767, 2013.

[18] Y. Zhang, J. He, S. Guo, F. Wang, "Energy efficiency maximization in wireless powered networks with cooperative non-orthogonal multiple access," IET Commun., 12(18): 2374–2383, 2018.

[19] P. D. Diamantoulakis, K. N. Pappi, Z. G. Ding, G. K. Karagiannidis, "Wireless powered communications with non-orthogonal multiple access," IEEE Trans. Wireless Commun., 15(12): 8422-8436, 2016.

[20] Y. Ye, Y. Li, D. Wang, G. Lu, "Power splitting protocol design for the cooperative NOMA with SWIPT," in Proc. IEEE International Conference on Communications (ICC): 1–5. 2017.

[21] S. Bisen, P. Shaik, V. Bhatia, "On performance of energy harvested cooperative NOMA under imperfect CSI and imperfect SIC," IEEE Trans. Veh. Technol. 70(9): 8993 - 9005, 2021.

[22] F. Nikjoo, A. Mirzaei, A. Mohajer, "A novel approach to efficient resource allocation in NOMA heterogeneous networks: Multi-criteria green resource management," Appl. Artif. Intell., 32(7-8): 583-612, 2018.

[23] A. Mohajer, F. Sorouri, A. Mirzaei, A. Ziaeddini, K. J. Rad, M. Bavaghar, "Energy-aware hierarchical resource management and Backhaul traffic optimization in heterogeneous cellular networks," IEEE Sys· J., 1-12, 2022.

[24] K. Agrawal, M. F. Flanagan, S. Prakriya, "NOMA with battery-assisted energy harvesting full-duplex relay," IEEE Trans. Veh. Technol., 69(11): 13952-13957, 2020.

[25] A. Hakimi, M. Mohammadi, Z. Mobini, Z. Ding, "Full-duplex non-orthogonal multiple access cooperative spectrum-sharing networks with non-linear energy harvesting," IEEE Trans. Veh. Technol, 69(10): 10925-10936, 2020.

[26] Y. Yuan, Y. Xu, Z. Yang, Z. Ding, "Energy efficiency optimization in full-duplex user-aided cooperative SWIPT NOMA systems," IEEE Trans. Commun., 67(8): 5753-5767, 2019.

[27] T. T. Nguyen, S.Q. Nguyen, P. X. Nguyen, Y. H. Kim, "Evaluation of Full-Duplex SWIPT cooperative NOMA-Based IoT relay networks over Nakagami-m fading channels," Sensors, 22(5): 1974, 2022.

[28] A. Mohajer, M. Bavaghar, H. Farrokhi, "Mobility-aware load balancing for reliable self-organization networks: Multi-agent deep reinforcement learning," Reliab. Eng. Syst. Saf., 202: 107056, 2020.

[29] H. Zhang, H. Zhang, K. Long, G. K. Karagiannidis, "Deep learning based radio resource management in NOMA networks: User association, subchannel and power allocation." IEEE Trans. Network Sci. Eng., 7(4): 2406-2415, 2020.

[30] S. P. Kumaresan, C. K. Tan, Y. H. Ng, "Deep Neural Network (DNN) for efficient user clustering and power allocation in downlink Non-Orthogonal Multiple Access (NOMA) 5G networks." Symmetry, 13(8): 1507, 2021.

[31] N. Yang, H. Zhang, K. Long, H. K. Hsieh, J. Liu, "Deep neural network for resource management in NOMA networks," IEEE Trans. Veh. Technol., 69(1): 876-886, 2019.

[32] S. Wang, T. Lv, W. Ni, N. C. Beaulieu, Y. J. Guo, "Joint resource management for MC-NOMA: A deep reinforcement learning approach," IEEE Trans. Wireless Commun., 20(9): 5672-5688, 2021.

[33] X. Xie, M. Li, Z. Shi, H. Tang, Q. Huang, "User selection and dynamic power allocation in the SWIPT-NOMA relay system," EURASIP J. Wireless Commun. Networking, (1): 1-19, 2021.

[34] J. Cui, M. B. Khan, Y. Deng, Z. Ding, A. Nallanathan, "Unsupervised learning approaches for user clustering in noma enabled aerial swipt networks," in Proc. IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC): 1-5, 2019.

[35] Z. Ding et al., "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," IEEE J. selected Areas Commun., 35.10 (2017): 2181-2195.

[36] M. Vaezi, Z. Ding, H. V. Poor, Multiple access techniques for 5G wireless networks and beyond, vol. 159, Berlin: Springer:2019.

[37] F. Fang, H. Zhang, J. Cheng, V. C. Leung "Energy-efficient resource allocation for downlink non-orthogonal multiple access networks," IEEE Trans. Commun., 64(9): 3722-3732, 2016.

[38] G. D. S. Peron, G. Brante, R. D. Souza, "Energy-efficient distributed power allocation with multiple relays and antenna selection," IEEE Trans. Commun., 63(12): 4797-4808, 2015.

[39] T. P. Dinh, H. M. Le, H. A. L. Thi, F. Lauer, "A difference of convex functions algorithm for switched linear regression," IEEE Trans. Autom. Control, 59(8): 2277-2282, 2014.

[40] A. Benhadid, F. Merahi, "Complexity analysis of an interior-point algorithm for linear optimization based on a new parametric kernel function with a double barrier term," Numer. Algebra Control Optim., doi: 10.3934/naco.2022003, 2022.

[41] D. Lee, D. W. Kim, "Multi-Objective LQG design with Primal-Dual method," arXiv preprint arXiv:2105.14760, 2021.

[42] D. Salvatore, A Zappone, S. Palazzo, M. Lops, "A learning approach for low-complexity optimization of energy efficiency in multicarrier wireless networks," IEEE Trans. Wireless Commun., 17(5): 3226-3241, 2018.

[43] Y. Cheng, K. H. Li, K. C. Teh, S. Luo, B. Li, "Two-Tier NOMA-Based wireless powered communication networks," IEEE Sys. J., 1-10, 2021.

[44] Y. Wang, Y. Wu, F. Zhou, Z. Chu, Y. Wu, F. Yuan, "Multi-objective resource allocation in a NOMA cognitive radio network with a practical non-linear energy harvesting model," IEEE Access, 6: 12973-12982, 2017.

Biographies

Mohammad Bagher Noori Shirazi received the B.S. degree in electrical engineering from University of Guilan, Rasht, Iran, in 2012, and M.S. degree in communications engineering from Khajeh Nasir Toosi University of technology, Tehran, Iran, in 2015. He is currently a Ph.D. student in the Department of Electrical and Computer Engineering, Babol Noshirvani University of Technology, Babol, Iran. His current research interests include wireless communications, cooperative based wireless networks, NOMA strategies and, energy harvesting protocols.

- Email: moh.noori.sh@gmail.com
- ORCID: 0000-0002-2900-8538
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

M. Reza Zahabi received his B.Sc. and M.Sc. degrees in electrical engineering from K.N. TOOSI University of Technology and Amir Kabir University of Technology, Tehran, Iran, respectively. He received his Ph.D. degree in 2008 in electrical engineering from Université de Limoges, France. Currently, he is a faculty member in Babol Noshirvani University of Technology. His current areas of interest are wireless communications, MIMO systems, 5G networks, coding protocols, and analog decoders.

- Email: zahabi@nit.ac.ir
- ORCID: 0000-0003-2811-8783
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://ostad.nit.ac.ir/home.php?sp=370385

How to cite this paper:
M. B. Noori Shirazi, M. R. Zahabi, "A novel full-duplex relay selection and resource management in cooperative SWIPT NOMA networks," J. Electr. Comput. Eng. Innovations, 11(1): 161-172, 2023.

DOI: 10.22061/JECEI.2022.8862.558

URL: https://jecei.sru.ac.ir/article_1768.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Object Detection by a Hybrid of Feature Pyramid and Deep Neural Networks

S. M. Notghimoghadam, H. Farsi*, S. Mohamadzadeh

Department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran.

Article Info

Article History:
Received 05 May 2022
Reviewed 12 June 2022
Revised 22 July 2022
Accepted 30 August 2022

Keywords:
Object recognition
Deep learning
Convolutional neural networks
Object classification

*Corresponding Author's Email Address: *hfarsi@birjand.ac.ir*

Abstract

Background and Objectives: Object detection has been a fundamental issue in computer vision. Research findings indicate that object detection aided by convolutional neural networks (CNNs) is still in its infancy despite having outpaced other methods.

Methods: This study proposes a straightforward, easily implementable, and high-precision object detection method that can detect objects with minimum least error. Object detectors generally fall into one-stage and two-stage detectors. Unlike one-stage detectors, two-stage detectors are often more precise, despite performing at a lower speed. In this study, a one-stage detector is proposed, and the results indicated its sufficient precision. The proposed method uses a feature pyramid network (FPN) to detect objects on multiple scales. This network is combined with the ResNet 50 deep neural network.

Results: The proposed method is trained and tested on Pascal VOC 2007 and COCO datasets. It yields a mean average precision (mAP) of 41.91 in Pascal Voc2007 and 60.07% in MS COCO. The proposed method is tested under additive noise. The test images of the datasets are combined with the salt and pepper noise to obtain the value of mAP for different noise levels up to 50% for Pascal VOC and MS COCO datasets. The investigations show that the proposed method provides acceptable results.

Conclusion: It can be concluded that using deep learning algorithms and CNNs and combining them with a feature network can significantly enhance object detection precision.

Introduction

Object detection is a term to describe a subset of computer vision and machine learning techniques highly influenced by deep learning and CNNs. Many studies have been conducted in this area. Object detection aims to identify and locate the types of objects in an image and categorize them into humans, animals, or vehicles [1]. As humans easily identify objects in an environment, an object detection system is designed to be trained like humans to be able to identify objects of different categories in fed images. A lower semantic distance between the machine and man evidently improves system performance. The conventional object detection methods are developed on the basis of handcrafted and shallow trainable architecture features. The design of intricate sets combining the features of many low-level images can easily impede their performance. The swift advancement in deep learning robust has led to the introduction of robust tools with semantic learning capabilities and deeper features for problem-solving in conventional architectures [2]. The emergence of neural networks influenced object detection methods. A neural network is a computational model formed by a large

number of interconnected nodes or neurons, each indicating an output function called the activation function [3]. Such networks have wide-ranging AI applications, such as in signal processing and automatic control [3], as well as various image processing areas [4]-[6], radar images processing [7], [8] and mobile telecommunications [9]. It should be stated that despite the wide application of neural networks, they generally fail to provide high precision in computer vision. Therefore, attempts were directed toward increasing the number of layers and the depth of these networks. However, there was a problem: increasing the number of layers would lead to the vanishing gradient problem, thus drastically slowing down the training process. Further, in more severe cases, it would halt the training process. Accordingly, the attempts to study deep learning increased. Deep learning, still under development, is a subset of machine learning that aims to learn from data using hierarchical architectures. It is extensively employed in artificial intelligence and machine vision [10]. Deep learning algorithms generally fall into four classifications: convolutional neural networks, restricted Boltzmann machine (RBM), autoencoders, and sparse encoders. These four categories are compared in this study. The computations of the RBM method are lengthier and time-taking. Autoencoders are not resistant to changes in the image, such as image rotation.

The training of features is impossible in sparse coding. CNNs are applicable in two-dimensional data. They are resistant to image changes, which makes them excellent options for object detection, allowing them to perform optimally. A critical factor that has recently directed the attention toward deep learning is the success of these algorithms in the ILSVRC challenge held by ImageNet every year [10]. In deep-learning-based methods, the object features should be extracted from the image. In other words, the data should be processed to identify the explicit and determining features of the objects. The higher the number of the extracted features results in the higher the precision of object detection. Research shows that using deep learning algorithms can significantly improve the precision of object detection systems and help achieve close-to-human precision. Deep learning also facilitates feature extraction from images without human intervention, which is an outstanding contribution and a critical advantage. As mentioned, the convolutional neural networks are perfect for object detection. The proposed method is a hybrid method with a core founded on the ResNet [11] deep neural network. This network has been designed and implemented in various depths. The details can be found in [11].

ResNet50, which is 50 layers deep, was selected among all ResNet layers. To improve the precision and quality of feature extraction, an FPN was designed based on [12], which was a hybrid of ResNet50 and a set of convolutional layers.

Related Work

Many studies address object detection, more specifically, to improve the precision of the existing systems and approximate neural networks to the human recognition system. Modern object detectors rarely can reach high-speed and precise inference with a short training. The TTFNet network has been proposed to balance precision, speed of inference, and training time [13]. Detectors with high-speed inference operators directly need less training time. Highly accurate detectors fall into detectors with low inference speeds and detectors requiring long training time. Authors in [13] stress the importance of shortening training time while maintaining the performance of well-known detectors. Since the time required for feature encoding and loss calculation is insignificant compared to feature extraction time. They adopt the training sample encoding approach to help enhance the learning rate and accelerate the training process. Authors in [14] propose a simple yet effective method called progressive self-knowledge distillation (PS-KD). This method employs its predictions as a model of trainer knowledge to strengthen the generalization performance of deep neural networks. In this method, the system plays the role of a learner that turns into a trainer over time. In this regard, the objectives are adjusted by combining the main background and the previous model predictions. The extensive research on image classification, object detection, and machine translation indicate improved performance by using this method. CNNs often encode an input image into a set of intermediate features. Although this structure is suitable for classification tasks, it fails to perform optimally in tasks requiring simultaneous detection and localization, such as object detection. The encoder-decoder architectures have been proposed as a solution. They are used by applying a decoder network on a backbone model. It has been noted that due to the decrease in the scale of the backbone, encoder-decoder architectures do not influence the establishment of robust multi-scale features. Accordingly, SpineNet, a backbone with scale-permuted features, is introduced [15]. The authors seek the answer to this question: Is the scale-permuted model a suitable backbone architectural design for simultaneous detection and localization? Intuitively, scale-permuted backbone architecture exterminates spatial information with down-sampling, challenging the retrieval of a decoder network. Object detection [16] is defined by estimating a very large but extremely sparse bounding box dependent probability distribution. This article introduces two new concepts: a corner-based region-of-interest estimator and a deconvolution-based CNN model. Most object detectors

are based on the anchor mechanism. To evaluate the alignment between the anchors and objects, they depend on calculating intersection over union parameters between the predefined anchor and real bounding boxes of objects. Authors in [17] question this type of using the intersection of union (IOU) and propose a new anchor alignment criterion. Anchors are a set of predefined reference boxes of a certain height and width. They are tiled across the image and help the network manages scale and object form changes by converting the object detection problem to a regression problem and classifying the anchor boundary box. This article proposes a mutual guidance mechanism that establishes an adaptive alignment between anchors and objects. The most modern object detection convolutional architectures are designed manually. In [18], the aim is to achieve a better architecture than an FPN for object detection. This article explores neural architectures and finds a new FPN in a new scalable searching space. This architecture is known as NAS-FPN, which is a combination of bottom-up and top-down connections. One-stage detectors simultaneously predict the classification time of objects and changes in the regression of the predefined boxes. This structure suffers from several flaws despite its good performance: The result of the predefined classification is inappropriately assigned to the regression during reasoning. Also, only one-time regression does not suffice for precise object detection. The present study first proposes a new module known as Reg-Offset-Cls (ROC) to solve the problem. The proposed module consists of three stages: bounding box regression, predicting the feature sampling location, and bounding box regression classification. Also, the hierarchical shot detector (HSD) detector was proposed to solve the second problem. This detector consists of two ROC modules and a feature-enhanced module [19]. Another study presents a systematic investigation of neural networks architecture designed for object detection and proposes several major optimizations to improve system performance. This article first proposes a bi-directional feature pyramid network (BiFPN) that allows the convenient and fast combination of multi-scale features. Next, a hybrid scaling method is proposed. The authors use the EfficientNet backbone to develop a new family of object detectors known as EfficientDet [20] . Humans perceive the world through sight, hearing, touch, and past experiences. Human experiences are taught by normal or unconscious learning. Authors in [21] propose an integrated network called YOLOR to encode implicit and explicit knowledge. Similar to the human brain, which is capable of conscious and unconscious knowledge acquisition, this integrated network can set up a display of simultaneous performance of tasks. This article summarizes how to design an integrated network that interpolates implicit knowledge into explicit knowledge. A systematic investigation of copy-paste indicates the sufficiency and acceptable performance of the simple mechanism of - randomly pasting the objects [22]. In [23], a large set of untagged images have been utilized for object detection. Authors study only several tagged images in each class. This method is known as multi-sample object detection. This procedure is iterated between the training model and the highly reliable sampling. In the training process, convenient samples are created first. Then, the initial weak model can improve. Over time, more reliable samples are selected, followed by another round of model improvement. The introduced framework is referred to as multi-modal self-paced learning for detection (MSPLD) [23]. An accurate, flexible, and completely anchor-free framework for object detection has been introduced [24].

Table 1: Related works

Method	Brief description of the method
TTFNet [13]	The TTFNet network has been proposed to balance precision, speed of inference, and training time.
PS-KD [14]	This method employs its predictions as a model of trainer knowledge to strengthen the generalization performance of deep neural networks.
SpineNet [15]	In this method SpineNet, a backbone with scale-permuted features, is introduced.
DeNet 101 [16]	In this method Object detection is defined by estimating a very large but extremely sparse bounding box dependent probability distribution.
Localize [17]	This article proposes a mutual guidance mechanism that establishes an adaptive alignment between anchors and objects
NAS-FPN AmoebaNet [18]	This article explores neural architectures and finds a new FPN in a new scalable searching space.
HSD [19]	Since the one-time regression does not suffice for precise object detection, HSD detector was proposed to solve this problem.
EfficientDet-D7x [20]	This article proposes a bi-directional feature pyramid network and a hybrid scaling method.
YOLOR-D6 [21]	Authors in this paper propose an integrated network called YOLOR to encode implicit and explicit knowledge.
Cascade Eff-B7 NAS-FPN [22]	A systematic investigation of copy-paste indicates the sufficiency and acceptable performance of the simple mechanism of - randomly pasting the objects.
MSPLD [23]	In this article a large set of untagged images have been utilized for object detection.
FoveaBox [24]	In this article an accurate, flexible, and completely anchor-free framework for object detection has been introduced.

Although almost all the highly advanced object detectors use predefined anchors, their proposed method directly learns the probability of the existence of an object and the bounding box coordinates without the anchor. This procedure involves two stages: the - prediction of class-sensitive semantic maps to measure the mentioned probability and the production of class-bounding boxes for each location with a potential object. In Table 1, related works are briefly stated.

The Proposed Method

Object classification and location are two critical steps in an object detection system design. The proposed method in this study is CNN-based and one-stage, with a ResNet50 core that can localize and classify the objects in the image. To achieve higher efficiency and better feature extraction, the FPN used here was combined with ResNet50. With the usage of a one-stage detector in this study, this question may arise: Is a simple one-stage detector able to achieve the same precision as a two-stage detector, as they have been known for their good performance? It should be noted that one-stage - detectors are applied to the regular and dense samplings of objects' locations scales. Recent research on one-stage detectors indicates that new one-stage detector designs are ten to forty percent more accurate than advanced two-stage methods [25]. Accordingly, the proposed method is a one-stage system with a ResNet50 core [11]. Moreover, the method involves an FPN to improve the feature extraction process and enhance the precision of the object detection system.

In general, the proposed method consists of two stages:

A) The training stage: Before the proposed system begins object detection, a convolutional neural network should be trained on a dataset.

B) The object detection stage: at this stage, the desired images (test images) are fed to the system. The system then detects the objects based on the training received in the previous stage. Below is the pseudocode of the proposed method.

Network ResNet

It is one of the deepest available architectures. It was introduced by Microsoft in 2015. With a depth of 152 convolutional layers and one fully-connected layer, this architecture has been recognized as the superior architecture in many competitions. It has also been able to reduce the ISLVRC challenge error by 3.57% [11]. This architecture has various layer depths discussed in [11]. The proposed method uses the Resnet50 structure. The number (50) indicates the layer depth of this architecture. (See Table 2). The ResNet architecture is one of the deepest architectures, which is presented in different depths (18, 34, 50, etc.) [11].

Algorithm 1: The pseudocode for proposed method

```
START
Downloading the dataset.
Implementing utility functions.
    coordinates of the corners
    coordinates of the center and the box
Computing pairwise Intersection Over Union (IOU.
 Implementing Anchor generator.
Resizing The Input IMAGE (1333*800)
Encoding labels.
Building the ResNet50 backbone.
Building Feature Pyramid Network(FPN).
Building the classification and box regression head.
Implementing decode predictions.
    confidence_threshold=0.05,
    nms_iou_threshold=0.5,
Implementing losses.
Setting up training parameters.
Initializing and compiling model.
Setting up callbacks.
Load the dataset.
Training the model And Loading weights.
Building inference model.
Generating Object detections.
    END.
```

Table 2: The layer details of ResNet50 [11]

Layer name	Output size	50-layer
Conv1	112×112	7×7, 64, stride 2
		3×3 max pool, stride 2
Conv2_x	56×56	$\begin{bmatrix} 1\times 1,64 \\ 3\times 3,64 \\ 1\times 1,256 \end{bmatrix}\times 3$
Conv3_x	28×28	$\begin{bmatrix} 1\times 1,128 \\ 3\times 3,128 \\ 1\times 1,512 \end{bmatrix}\times 4$
Conv4_x	14×14	$\begin{bmatrix} 1\times 1,256 \\ 3\times 3,256 \\ 1\times 1,1024 \end{bmatrix}\times 6$
Conv5_x	7×7	$\begin{bmatrix} 1\times 1,512 \\ 3\times 3,512 \\ 1\times 1,2048 \end{bmatrix}\times 3$
	1×1	Average pool, 1000-d fc, softmax

Before this architecture was introduced, other architectures had shown that increasing network depth has a direct effect on network efficiency and increases network efficiency but the problem was that in those architectures due to the problem of vanishing gradient. The depth of the network could be increased to a certain extent. In order to increase the depth of a shallow network, we need the identity layer. ResNet is a residual deep learning framework whose architecture is such that the identification of the layers is easily performed.

Therefore, by using this architecture, the depth of the network can be increased. In addition, ResNet's architectural structure is designed in such a way that there is an input from the previous step in each step, and this causes that better feature maps to be produced. Since in the proposed method we combine a feature pyramid network with the ResNet architecture and this increases the depth of the network, therefore, we choose the depth of 50 (ResNet50) among the different structures of the ResNet architecture so that the training of the network does not take too long and object detection is faster performed.

Feature Pyramid Network

FPNs can be applied to extract feature maps of images more efficiently. The proposed FPN used in [12] has a top-down architecture with lateral connections to create feature maps. This network significantly improves the feature extraction process. Multi-scale object detection is a major challenge in computer vision. This pyramid can enable a given model to detect objects in a wide range of scales by scanning locations and pyramid surfaces, hence the use of feature pyramids in the proposed method. The reported method in [12] purposes using the pyramidal and hierarchical structure of a convolution network. To achieve this goal, a structure has been relied on that combines low-resolution and semantically strong features with high-resolution and semantically weak features in a top-down manner. This pyramid can enable a given model to detect objects in a wide range of scales by scanning locations and pyramid surfaces, Therefore, following the ResNet network, we implemented a feature pyramid network to extract features more accurately.

Eight two-dimensional convolutional layers have been incorporated into the proposed method to develop this network (See Table 3).

Table 3: The convolutional layer details used in creating the FPN used

Padding	Stride	Dimensions	Number of filters	Layer name
Same	1	1×1	256	Conv 1
Same	1	1×1	256	Conv 2
Same	1	1×1	256	Conv 3
Same	1	3×3	256	Conv 4
Same	1	3×3	256	Conv 5
Same	1	3×3	256	Conv 6
Same	2	3×3	256	Conv 7
Same	2	3×3	256	Conv 8

Fig. 1 shows the structure of the core of the proposed system, and as it is known, the core of the system in our proposed method is created by combining a deep convolutional neural network called ResNet [11] with a feature pyramid network. This structure makes the extraction of features better and ultimately improves the accuracy of the object recognition system.

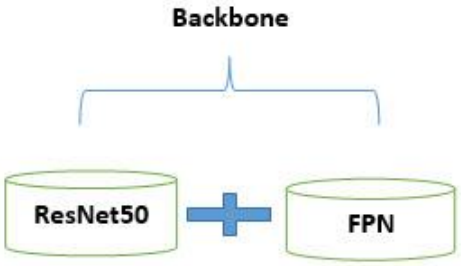


Fig. 1: The proposed system's core.

Image Resizing

The dimensions of the image change in a way that the length of the shorter side of the image equals 800 pixels. If the longer side of the image equals 1333 pixels after image resizing, the image size changes so that the longer side length equals 1333 pixels. In other words, in this stage, a minimum image side of 800 pixels and a maximum side length of 1333 pixels are defined.

Raw labels

Raw labels, including bounding boxes and class attributes, are applied in network training. This operation consists of the following stages:

- Creating anchor boxes in proportion to the database image dimensions.
- Assigning the ground truth box of objects (the main box refers to the actual box of every object in the image) to anchor boxes.

Here, the ground truth box of objects is assigned to anchor boxes based on overlapping. To that end, the IOU should be calculated among all anchor boxes during the training time and the ground truth box of the objects. The comparison based on IOU as a block diagram depicted in Fig. 2.

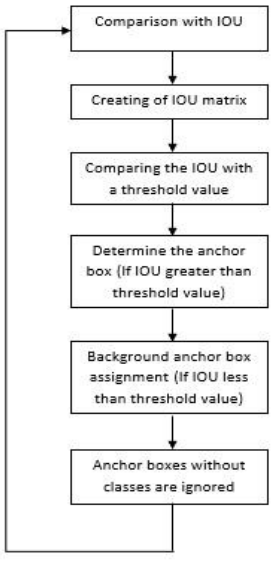


Fig. 2: The IOU block diagram.

System Training and Feature Extraction

In the training process, the system seeks to identify the best unknown parameters, such as the weight of convolutional filters and coefficients of the fully-connected layers, to achieve the least classification error rate. The proposed method uses the back propagation of errors and stochastic gradient descent (SGD) methods to update the weights in each iteration.

In stochastic gradient descent, the weights in each iteration are updated according to (1) . Equation (2) is the simpler re-expression of (1).

$$\theta = \theta - \eta \cdot \nabla_\theta J\left(\theta; x^{(i)}; y^{(i)}\right) \quad (1)$$

$$x = x - (\text{Learning\_rate}) * dx \quad (2)$$

where x is the order of parameters and weight of the filters and Learning_ rate indicates the rate at which the network is trained, determined at the beginning of the training process . It should be noted that the training rate parameter is modifiable. After a sufficient number of iterations, the network is trained to classify the database images.

After the training, the features of the dataset images should be extracted, which is a critical step as classification is performed based on these features and the rectangular boxes of objects (called windows, showing the location coordinates of objects). Finally, when an image is fed to the convolutional network, it crosses the convolutional layers, leading to a vector output. This output is known as the feature vector. By feeding all the images in the dataset to the network, there will be a set of feature vectors. Each vector is provided to the fully-connected layer for classification. The fully-connected layer in the proposed method here consists of 1000 neurons. Therefore, there will ultimately be 1000 specific classifications of database images. Generally, to summarize the first stage, the designed system is trained based on a set of datasets in the first stage. The optimum weights of filters and network parameters of the convolutional network, which are set to the optimal state to minimize classification errors, are then calculated to finalize the classification of the dataset images. This data is used to detect the existing objects in a new image. Next, for object detection, the test images are fed to the system. With the passing of images from the trained system, the feature vector of each image is extracted. Based on the classification set in the previous stage, the fully-connected layer determines the class of each object.

Activation Functions

Activation functions are essential in neural networks, playing a vital role in the network. The role of these functions is to make the optimization problem non-linear. More specifically, the activation functions decide whether or not a specific neuron in neural networks should be activated, determining to which category or class the output of the neural network belongs. Activation functions fall into various types, such as sigmoid, Tanh hyperbolic, and ReLU. The proposed method uses the ReLU and sigmoid functions as given by (3) and (4), respectively.

$$f(x)=\max(0,x) \quad (3)$$

$$\sigma(x) = 1/(1+e^{-x}) \quad (4)$$

Results and Discussions

This section discusses the evaluation criteria, datasets, and implementation results. Also, it compares the numerical results of the proposed method with some recent methods.

Evaluation Criteria

Here, the mAP criterion was used, as it is one of the most common and important evaluation criteria in object detection.

Intersection Over Union (IOU)

IOU is an evaluation criterion used to examine the precision of the predicted bounding box according to the ground truth box of the objects. Its value indicates the overlapping area between the predicted box and is a number that is the ground truth box of the object. In the case that this value is greater than a specific threshold, the system recognizes that anchor box as a predicted object. Therefore, this criterion determines the location precision by comparing the overlap between the ground truth box and the predicted bounding box. The determined value is between zero and one. The detection of an object is regarded as true when its IOU value exceeds a predefined threshold value. The usual threshold value is 0.5. When the predicted box overlaps with a ground truth box of more than 50%, the diagnosis made is considered valid. IOU is expressed as the follows:

$$\text{IOU} = (\text{area of overlap})/(\text{area of union}) \quad (5)$$

where the area of overlap is the intersection between the two boxes that also indicates the level of overlap. Area of union indicates an area with zero overlap between the two boxes [26].

The Precision and Recall Criteria

Precision (P) and recall (R) are used to evaluate the -classification ability of a method in question. In that regard, the values of the following parameters should be determined: true positive (TP) diagnosis, false positive (FP) diagnosis, and false negative (FN) diagnosis. These parameters are estimated based on location precision, -which is calculated based on IOU. Based on the IOU threshold values, then the values of TP and FP are calculated for the detected objects. The values of P and R are determined per each detection (true or false) based on TP and FP values using the Equations below [26]:

$$P = \frac{TP}{TP + FP} = \frac{TP}{\text{all detections}} \quad (6)$$

$$R = \frac{TP}{\text{all ground truths}} = \frac{TP}{TP + FN} \quad (7)$$

Average Precision (AP) and Mean Average Precision (Map)

The average precision criterion is calculated according to the values of P and R. This parameter is the average precision obtained per recall. The value of mAP is then - calculated by determining the mean of the AP value in different classifications [26], [27].

Datasets

To assess the performance of every object detection system, they should be run on suitable databases, and the standard criteria should be calculated accordingly. In this study, the Pascal VOC and COCO datasets were used for this purpose.

The Pascal Visual Object Classes (VOC) Dataset

This dataset was introduced in 2005. It comprises two sections: the first section consists of various image classes. The second section includes the annual tournaments. It consists of five challenges: classification, diagnosis, segmentation, action classification, and person layout. The various solution methods used in different competitions are discussed and compared in a workshop. The tournament was held from 2005 to 2012. Fig. 3 shows several sample images of this dataset [28].

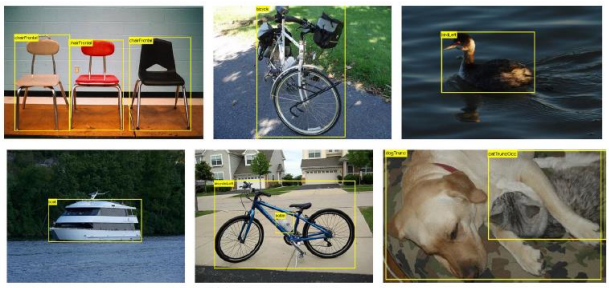


Fig. 3: Sample images of the Pascal VOC dataset [28].

The Microsoft Common Objects in Context (MS COCO) Dataset

Microsoft researchers introduced this database in 2014. It consists of 91 classes of objects and many samples and tagged images. Compared to Pascal VOC, it has fewer feature categories and samples in each category.

This dataset addresses three core problems: detection of non-iconic view of objects, contextual reasoning between objects, and precise two-dimensional location of objects.

Fig. 4 presents a sample image of images in this dataset [29].

Fig. 4: Sample images of the MS COCO dataset [29].

The Implementation Results

This section is a discussion of simulation results performed on two important object detection datasets. First, the average precision criterion is separately applied as a sample to each class of the existing objects in both datasets. The mAP of the proposed method is then compared with some recent methods.

Table 4 shows that the proposed method achieved a better average precision in 10 sample classes, such as Bird, Dog, Cat, and Person. Then, by calculating the precision of all classes and obtaining their means, the mAP value of the proposed method is calculated. This criterion indicates the performance of the object detection system.

Table 4: The average precision of the proposed method in ten different classifications in the Pascal VOC 2007 and MS COCO datasets

Class	Average precision (%)	
	Voc 2007 dataset	MS COCO dataset
Person	97.43	85.08
Cat	96.91	71.42
Dog	94.43	73.80
Bird	98.40	87.43
Train	91.23	67.42
Car	98.29	90.06
Motor	93.55	79.32
Bus	95.14	76.83
Cow	92.43	72.18
Sheep	94.27	85.1

Table 5 compares the proposed method with several recent methods based on the values of the mAP criterion.

As shown, the proposed method displayed a good object detection performance. The method proposed yielded a mean average precision (mAP) of 41.91 in the Pascal Voc2007 and 60.07% in COCO. Fig. 5 and Fig. 6 display samples of the detected objects by the proposed method in Pascal VOC and MS COCO datasets.

Table 5: The results of the mAP criterion of the proposed method and several recent methods in the Pascal VOC 2007 and MS COCO datasets

Method Name	Pascal VOC mAP%	Method Name	MS COCO mAP%
PS-KD [14]	79.7	TTFNet [13]	35.1
DeNet-101 [16]	77.1	SpineNet-49 [15]	45.3
Localize [17]	81.50	UniverseNet-20.08s [30]	47.4
HSD [19]	83.00	NAS-FPN AmoebaNet [18]	48.3
ReCoR [31]	83.90	EfficientDet-D7x [20]	55.1
Cascade Eff-B7 NAS-FPN [22]	88.6	YOLOR-D6 [21]	57.3
FoveaBox [24]	76.60	MSPLD [23]	56.6
proposed method	91.41	proposed method	60.07

Fig. 5: Samples of objects detected in Pascal VOC.

Fig. 6: Samples of objects detected in MS COCO.

The Proposed System Under Noise

To better evaluate the proposed model, some noise was introduced to the proposed system to examine its performance under noise. The test images of the datasets were combined with the salt and pepper noise [32] to obtain the value of mAP for different noise levels. Fig. 7 shows the mAP value against noise levels.

As shown, even at 0.3 noise level (for instance, when 30% of pixels of the image have been ruined by noise), the detection precision is above 50%. In real imaging situations, usually, the noise generated on the images is less than 0.05., hence the acceptable performance of the proposed model even in dealing with this level of noise. To obtain an intuitive understanding of the noise level, Fig. 8 display the output of the proposed model under different noise levels.

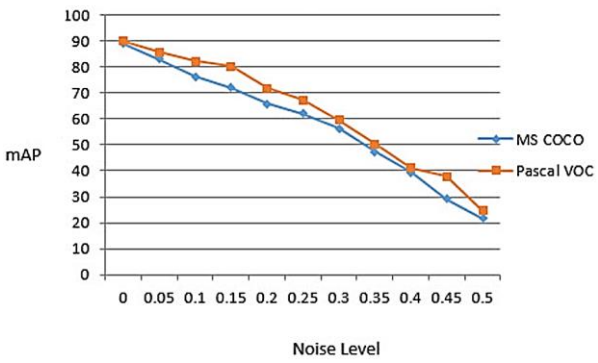


Fig. 7: The mAP diagram based on the noise level.

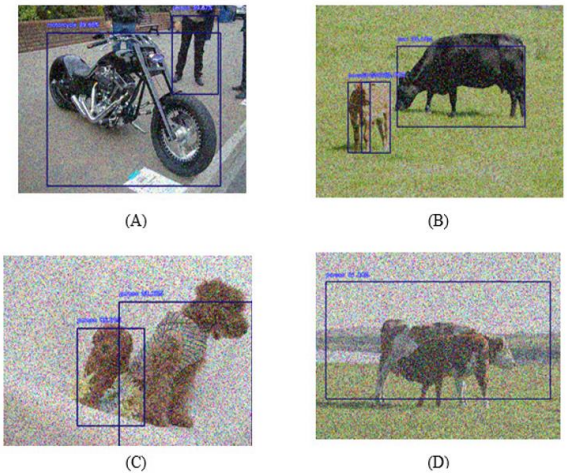


Fig. 8: Object detection with different noise levels: (A) 0.1, (B) 0.25, (C) 0.35 and (D) 0.5.

Conclusion

Since object detection is widely used in various fields such as medicine, self-driving cars, radar images processing and etc, the aim of this article is to implement a method for object detection with high accuracy. In this research, we studied the new articles which were presented in the topic of object recognition in the last few years and we found this reality that the object detection by convolutional neural networks is superior to other methods, but still the existing methods can be improved in terms of accuracy and speed. Therefore, our most important goal in this article was to present a method based on deep learning, using convolutional neural networks to recognize objects so that the proposed method detects objects with the least error in the shortest possible time. In general, object detection methods are classified into single-stage and two-stage detectors, and two-stage detectors are often more accurate but slower than single-stage detectors. Here, a one-stage object detection method was implemented using CNNs with a ResNet50 core. An FPN was used to improve the quality of the feature extraction process. The proposed system was then trained and evaluated with MS COCO and Pascal VOC2007 datasets. In the end, the value of mAP, one of the most notable criteria for performance evaluation of object detection systems, was calculated.

The proposed method was evaluated against seven other methods based on the mAP value obtained in each dataset. Moreover, a specific noise level was added to further investigate the system's performance in different noise level scenarios and mAP values.

The results indicated the better performance of the proposed method. Therefore, since the proposed system was trained using a deep-learning algorithm, the features of images were extracted hierarchically from the input images. The extracted data were then combined with an FPN. This combination improved the proposed method's detection precision. Accordingly, it can be concluded that using deep learning algorithms and CNNs and combining them with a feature network can significantly enhance object detection precision. And finally, considering the importance and many applications of object recognition a lot of research can be done in this field. Some examples of future work are as follows:

- Implementation of the proposed method presented in the article with other architectures and comparison with the results of our method.
- Investigating the use of noise reduction methods in object recognition.
- Implementation of the proposed method in order to object detection in the video.
- Designing a system by combining current object detection methods which may lead to improved performance.

Author Contributions

Prof. Hassan Farsi and Dr. were the supervisor and co-supervisor of the current research plan. They sketched the research framework and the roadmap. Also, they analyzed the results and tabulated the outcome derived from excerpted literatures. In this line, Seyed Mojtaba Notghimoghadam searched in authentic journals to gather all relevant papers. In addition to, he prepared the blueprint of the research plan.

Acknowledgment

This work is completely self-supporting, thereby no any financial agency's role is available.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviation

CNN	Convolutional Neural Network
FPN	Feature Pyramid Network
MAP	Mean Average Precision
ResNet	Residual Network
COCO	Common Objects in Context
MS COCO	MicrosoftCommon Objects in Context
VOC	Visual Object Classes
RBM	Restricted Boltzmann Machine
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
TTFNet	Training Time Friendly Network
PS-KD	Progressive Self-Knowledge distillation
IOU	Intersection Of Union
HSD	Hierarchical Shot Detector
BIFPN	Bi-directional Feature Pyramid Network
YOLOR	You Only Learn One Representation
MSPLD	Multi-modal Self-Paced Learning for Detection
DeNet	Danish Ethernet Network
AR	Average Precision

References

[1] Z. Zou, Z. Shi, Y. Guo, J. Ye, "Object Detection In 20Years: A Survey," arXiv preprint arXiv: 1905.05055, 2019.

[2] Z. Zhao, P. Zheng, S. Xu, X. Wu, "Object detection with deep learning: A Review," IEEE Trans. Neural Networks Learn. Syst., 30(11): 3212-3232, 2019.

[3] Y. Wu, J. Feng, "Development and application of artificial neural network," Wireless Pers. Commun., 102(2): 1645-1656, 2018.

[4] R. Nasiripour, H. Farsi, S. Mohamadzadeh, "Visual saliency object detection using sparse learning," IET Image Proc., 13(13): 2436-2447,2019.

[5] S. Pasban, S. Mohamadzadeh, J. Zeraatkar-Moghaddam, A. Shafiei, "Infant brain segmentation based on a combination of VGG-16 and U-Net deep neural networks," IET Image Proc., 14(17): 4756-4765, 2021.

[6] Z. Dorrani, H. Farsi, S. Mohamadzadeh, "Image edge detection with fuzzy ant colony optimization algorithm," Int. J. Eng., 33(12): 2464-2470, 2020.

[7] C. Seale, T. Redfern, P. Chatfield, C. Luo, k. Dempsey, "Coastline detection in satellite imagery: A deep learning approach on new benchmark data," Remote Sens. Environ., 278: 113044, 2022.

[8] K. Zeng, Y. Wang," A deep convolutional neural network for oil spill detection from spaceborne SAR images," Remote Sens., 12(6): 1015, 2020.

[9] H. Aliakbari, A. Abdipour, A. Costanzo, D. Masotti, R. Mirzavand, P. Mousavi, "ANN-Based design of a versatile millimetre-wave slotted patch multi-antenna configuration for 5G scenarios," IET Microwaves Antennas Propag, 11(9): 1288-1295, 2017.

[10] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, M. Lew, "Deep learning for visual understanding: A review," Neurocomputing, 187: 27-48, 2016.

[11] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in Proc. the IEEE conference on computer vision and pattern recognition: 770-778, 2016.

[12] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, "Feature pyramid networks for object detection," in Proc. the IEEE conference on computer vision and pattern recognition: 2117-2125, 2017.

[13] Z. Liu, T. Zheng, G. Xu, Z. Yang, H. Liu ,D. Cai, "Training-time-friendly network for real-time object detection," in Proc. the AAAI Conference on Artificial Intelligence, 34(07): 11685-11692, 2020.

[14] K. Kim, B. Ji, D. Yoon, S. Hwang, "Self-Knowledge distillation with progressive refinement of targets," in Proc. the IEEE/CVF International Conference on Computer Vision: 6567-6576, 2021.

[15] X. Du, T. Lin, P. Jin, G. Ghiasi, M. Tan, Y. Cui, Q. Le, X. Song, "SpineNet: Learning scale-permuted backbone for recognition and localization," in Proc. the IEEE/CVF Conference on Computer Vision And Pattern Recognition: 11592-11601, 2020.

[16] L. Tychsen-Smith, L. Petersson, "Denet: Scalable real-time object detection with directed sparse sampling," in Proc. the IEEE international Conference on Computer Vision: 428- 436, 2017.

[17] H. Zhang, E. Fromont, S. Lefèvre, B. Avignon, "Localize to classify and classify to localize: Mutual guidance in object detection," in Proc. the Asian Conference on Computer Vision, 2020.

[18] G. Ghiasi, T. Lin, Q. Le, "Nas-Fpn: Learning scalable feature pyramid architecture for object detection," in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 7036-7045, 2019.

[19] J. Cao, Y. Pang, J. Han, X. Li, "Hierarchical shot detector," in Proc. the IEEE/CVF International Conference on Computer Vision: 9705-9714, 2019.

[20] M. Tan, R. Pang, Q. Le, "Efficientdet: Scalable and efficient object detection," in Proc. the IEEE/CVF conference on computer vision and pattern recognition: 10781-10790, 2020.

[21] C. Wang, I. Yeh, H. Liao, "You only learn one representation: unified network for multiple tasks," arXiv preprint arXiv: 2105.04206, 2021.

[22] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T. Lin, E. Cubuk, Q. Le, B. Zoph., "Simple copy-paste is a strong data augmentation method for instance segmentation," in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 2918-2928, 2021.

[23] X. Dong, L. Zheng, F. Ma, Y. Yang, D. Meng, "Few-Example object detection with model communication," IEEE Trans. Pattern Anal. Mach. Intell., 41(7): 1641-1654, 2018.

[24] T. Kong, F. Sun, H. Liu, Y. Jiang, L. Li, J. Shi, "Foveabox: Beyound anchor-based object detection," IEEE Trans. Image Proc., 29: 7389-7398, 2020.

[25] T. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, "Focal loss for dense object detection," in Proc. the IEEE International Conference on Computer Vision: 2980-2988, 2017.

[26] P. Rafael, S. L. Netto, E. Silva, "A survey on performance metrics for object-detection algorithms," in Proc. 2020 International Conference on Systems, Signals and Image Processing (IWSSIP): 237-242, 2020.

[27] X. Wu, D. Sahoo, S. Hoi, "Recent advances in deep learning for object detection," Neurocomputing, 396(7): 39-64, 2020.

[28] M. Everingham, S. Eslami, L. Gool, C. Williams, J. Winn ,A. Zisserman, "The pascal visual object classes challenge: A retrospective," Int. J. Comput. Vision, 111(1): 98-136, 2015.

[29] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, C. Zitnick., "Microsoft COCO: common objects in context," in Proc. European Conference on Computer Vision: Springer, Cham: 740-755, 2014.

[30] Y. Shinya, "USB: Universal-scale object detection benchmark," arXiv preprint arXiv: 2103.14027, 2021.

[31] Z. Chen, J. Zhang, D. Tao, "Recursive context routing for object detection," Int. J. Comput. Vision, 129(1): 142-160, 2021.

[32] J. Azzeh, B. Zahran, Z. Alqadi, "Salt and pepper noise: Effects and removal," JOIV: Int. J. Inf. Visualization, 2(4): 252-256, 2018.

Biographies

Seyed Mojtaba Notghimoghadam received the B.Sc. degree in electrical engineering from the Islamic Azad University of Birjand, Birjand, Iran, 2012. He received the M.Sc. degree in telecommunication engineering from University of Birjand, Birjand, Iran, in 2022. He is currently Ph.D. student in university of Birjand, Birjand, Iran. His research interests in digital image processing, visual signal processing, deep learning and artificial intelligence.

- Email: m.notghimoghadam@birjand.ac.ir
- ORCID: 0000-0002-7320-929X
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Hassan Farsi received the B.Sc. and M.Sc degrees from Sharif University of Technology, Tehran, Iran, in 1992 and 1995, respectively. Since 2000, he started his Ph.D in the Centre of Communications Systems Research (CCSR), University of Surrey, Guildford, UK, and received the Ph.D degree in 2004. He is interested in speech, image and video processing on wireless communications. Now, he works as professor in communication engineering in department of Electrical and Computer Eng., University of Birjand, Birjand, IRAN.

- Email: hfarsi@birjand.ac.ir
- ORCID: 0000-0001-6038-9757
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://cv.birjand.ac.ir/hasanfarsi/en

Sajad Mohamadzadeh received the B.Sc. degree in electrical engineering from Sistan & Baloochestan, University of Zahedan, Iran, in 2010. He received the M.Sc. degree in telecommunication engineering from university of Birjand, Birjand, Iran, in 2012. Now, he works as associate professor in department of Electrical and Computer Eng., University of Birjand, Birjand, IRAN. His area research includes image processing, deep learning, pattern recognition, digital signal processing and sparse representation.

- Email: s.mohamadzadeh@birjand.ac.ir
- ORCID: 0000-0002-9096-8626
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://cv.birjand.ac.ir/mohamadzadeh/en

How to cite this paper:
S. M. Notghimoghadam, H. Farsi, S. Mohamadzadeh, "Instructions and formatting rules for authors of journal of electrical and computer engineering innovations, JECEI," J. Electr. Comput. Eng. Innovations, 11(1): 173-182, 2022.

DOI: 10.22061/JECEI.2022.9012.567

URL: https://jecei.sru.ac.ir/article_1769.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Displacement Effects on the Electrical Characteristics of a Single-Molecule Device

E. Rahimi[1,*], S. Dorouki[2]

[1]*Faculty of Electrical Engineering, Shahrood University of Technology, Shahrood, Iran.*

[2]*Kharazmi International Campus, Shahrood University of Technology, Shahrood, Iran.*

Article Info

Article History:
Received 19 June 2022
Reviewed 27 July 2022
Revised 20 August 2022
Accepted 25 August 2022

Keywords:
Molecular electronics
Single-molecule device

*Corresponding Author's Email Address:
erahimi@shahroodut.ac.ir

Abstract

Background and Objectives: The displacement of molecules is one of the major fabrication faults in manufacturing molecular electronic devices. In this paper, we profoundly study the effect of displacement on the current-voltage, and conductance-voltage characteristics of the Au-Benzenedithiol-Au single-molecule device.

Methods: The *ab-initio* calculations on the isolated molecules were performed to obtain the basic single-level quantum-dot model parameters. These parameters were then used within the self-consistent field algorithm to calculate the electrical characteristics of the device.

Results: The maximum conductance occurs when the molecule is placed exactly in the midpoint of the distance between the two electrodes, where the electrostatic capacitance reaches its minimum. When the molecule deviates from this point, and approaches one electrode, the conductance is decreased, and asymmetric behavior emerges. A molecular rectifier can be manufactured by placing the molecule close to one electrode.

Conclusion: Although modern software packages may employ advanced and complicated models including the combination of the density functional theory (DFT) and non-equilibrium Green's function (NEGF) methods to obtain accurate results, they are demanding in computer memory and time. Moreover, understanding the physical quantities of the systems from large-scale matrices is often difficult. The single-level model is a computationally light method, which provides a profound understanding of the device characteristics since all quantities are presented by numbers.

Introduction

Molecules, due to their advantages such as small size, chemical tunability, and self-assembly properties, are considered the main components of future electronic devices and systems [1]. The first molecular electronic device was proposed by Aviram and Ratner in 1974. During their theoretical calculations, they predicted that an organic molecule can act like a diode when attached to the electrodes. The proposed rectifier structure was similar to a P-N junction and consisted of electron donor and acceptor parts that were separated from each other by a tunneling bridge [2]. But the main challenge was measuring the current and conductivity of individual molecules or a small group of them. This challenge was first overcome in 1997 by an experimental group at Yale University, led by Mark Reed. By using the mechanically controllable break junction (MCBJ) method, they succeeded in measuring the current in single molecules [3]. After this pioneering experiment, the measurement of molecular conformation, current and conductivity in molecular devices continued by many experimental research groups around the world [4]-[17]. Therefore, by

using various experimental procedures, including the spectroscopic, thermoelectricity, and break-junction methods, the researchers succeeded in fabricating single-molecule devices, and as a result investigated the effect of various factors such as the types of molecule-electrode contact (bonding and anti-bonding), the length of the molecular bonds, different molecular functional groups and environmental factors (solvent and temperature) that affect their conductivity [7]-[13]. In recent years, many researchers have measured charge transfer in single organic molecules or a small group of them [9]-[13]. In addition, theoretical models at the semi-empirical and quantum chemical levels were utilized to understand the conductivity in such molecules. These models include the non-equilibrium Green's function (NEGF) multi-level models, tight-binding (TB) model, density functional theory model (DFT), local density approximation (LDA) and first-principles approaches [18]-[32]. Conductance of molecular devices is a challenging concept in which not only the chemical nature of the molecule plays a role, but also external factors such as the geometry of the metal-molecule chemical bonds and the electrostatics of the environment are important [8]-[10] . Therefore, although the results of calculations and models generally determine the current-voltage behavior of the device, the values obtained from the *ab initio* methods are usually several tens of times higher than the measurements and practical observations [3], [21]. One reason is that the exact geometry of the electrodes made in practice is unknown. Additionally, modeling of the molecule-electrode interface is challenging. In the last few decades, the molecular structure of benzene has been considered a small molecule in electronics. Benzene, with the molecular formula C_6H_6, is a planar molecule with a regular hexagonal structure in which each C-C-C bond angle is 120 degrees. Since each carbon atom is bonded to three other atoms, the orbitals are SP^2 type, and each carbon has an orbital perpendicular to the hexagonal plane; So, each orbital overlap with two neighboring orbitals to the same extent. Therefore, all six electrons are completely delocalized in the ring path, resulting in two donut-like electron clouds, one at the top and one at the bottom of the ring. This molecule can establish strong chemical bonds with the surface of gold electrodes by two thiol end groups [6]. In the manufacturing process of molecular devices, there is a possibility of deposition of a molecule in an inappropriate position between the electrodes. This paper aims to investigate the effect of the displacement of the molecule between two electrodes on the conductance of Au-1,4-benzenedithiol (BDT)-Au two-terminal device. Fig. 1 shows the BDT molecule between two gold electrodes. Changing the distance of the molecule to the source or drain changes its coupling to the electrodes (Fig. 2). Therefore, it is expected that the current and current-voltage characteristics of the device will be affected.

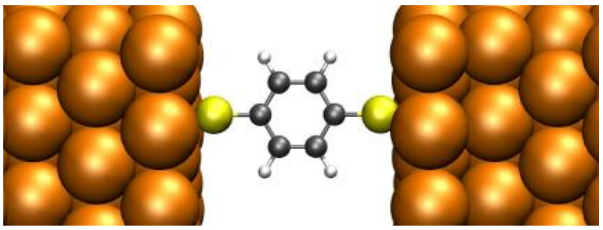


Fig. 1: BDT molecule between the source and drain electrodes.

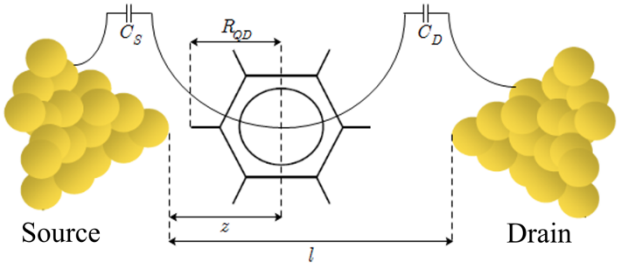


Fig. 2: Structure of the Au-BDT-Au molecular device. The source and drain contacts are modeled as the two plates of two planar capacitors.

In this research, we model the BDT molecule with a single-level quantum dot and investigate its displacement effect on the conductance of the Au-BDT-Au molecular device. Parameters in multi-level models are in the form of matrices, while they are in the form of numbers in single-level models. The single-level model is less complicated and accurate compared to other multi-level models, but because its parameters are numerical, it provides a better understanding of the physical behavior of the device in general. This paper is organized as follows. In the models and methods section, we briefly describe the single-level theoretical model, and in the discussion section, we analyze the effect of changing the position of the molecule on the device conductance, electrostatic capacitance, and molecule potential. Subsequently, we conclude in the last section.

Models and Methods

Although several mathematical and physical models and methods have been proposed for studying molecular devices, the results obtained using these models are often deviates from the measurements and experimental observations. Fig. 3 compares the practically measured current-voltage and conductance characteristics of the Au-BDT-Au device with the theoretical calculation results obtained from the multi-level DFT and NEGF model. This figure depicts that there is an impressive difference between theoretical and practical current values. To describe the electron transfer, we use the quantum dot (QD) model, which includes only one energy level. This is the simplest model to describe charge transfer in molecular devices, which includes the discrete nature of the energy spectrum of the molecule and ignores the effects related to the electrode-molecule contacts. In this model, the broadening of an energy level is approximated by the Lorentzian function as [20]:

$$g(\mathcal{E}) = \frac{2}{\pi}\frac{(\Gamma_S + \Gamma_D)/2}{(\mathcal{E} - \mathcal{E}_0)^2 + ((\Gamma_S + \Gamma_D)/2)^2} \quad (1)$$

where, Γ_S and Γ_D are the coupling energies of the source and drain electrodes to the molecule, respectively. The coupling energies are inversely related to the electron transfer rate, $\tau = \Gamma/\hbar$. In this paper we consider the coupling energies as follows:

$$\Gamma_S(z) = \Gamma_m \exp\left(-\frac{z - R_{QD}}{l/4}\right) \quad (2)$$

$$\Gamma_S(z) = \Gamma_m \exp\left(\frac{l - z - R_{QD}}{l/4}\right) \quad (3)$$

where, Γ_m denotes the maximum coupling energy, R_{QD} is the radius of the molecule, and z and l are the distance of the molecule from the source electrode, and the source-drain distance, respectively. The energy of the highest occupied molecular orbital (HOMO) of the molecule, $\mathcal{E}_0$, the energy of the lowest unoccupied molecular orbital, $\mathcal{E}_1$, and R_{QD}, and l have been calculated using the DFT at the B3LYP level of theory with 6-31G* basis set. These calculations for the isolated BDT molecule have been performed using the Gaussian-09 software. The electrochemical potential of the gold electrodes is $\mu = -5\text{eV}$ and the maximum coupling energy is $\Gamma_m = 0.1eV$.

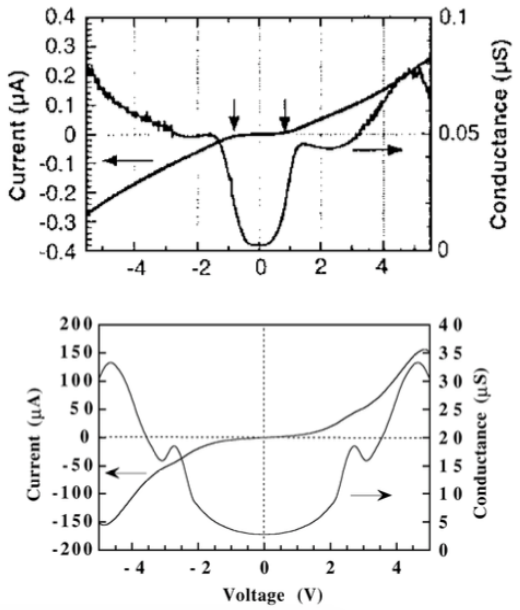


Fig. 3: The top figure depicts the measurement results [3], and the bottom figure shows the theoretical calculations [26] of the current-voltage and conductance of Au-BDT-Au device.

Table 1 lists the model parameters. According to the calculations, the electrochemical potential energy of the gold electrode is closer to the HOMO than to the LUMO, so it is expected that electron transfer will take place through the HOMO level.

Table 1: Model parameters used in this paper.

$\mathcal{E}_0$(e)	$\mathcal{E}_1$(eV)	l(nm)	R_{QD}(nm)	μ(eV)	Γ_m(eV)
-5.64	-0.36	5	1.5	-5	0.1

Moreover, in previous studies, it has been reported that the transfer in the Au-BDT-Au structure takes place through the frontier HOMO orbital [17]. Thus, it is reasonable to use one-level quantum-dot model.

After connecting the electrodes to the molecule, the Fermi energy level of the molecule is shifted by three mechanisms. The first mechanism is the state filling effect. Since initially the electrochemical potential of the electrodes is not equal to the Fermi energy of the molecule, some electric charge is transferred between the electrode and the molecule to equalize the Fermi energy level of the molecule to the electrochemical potential of the electrodes. Charge transfer in this case leads to the state filling effect. The second mechanism is the charging effect. Since the electron has an electric charge, the potential energy of the molecule changes due to electron transport. The third mechanism is the electrostatic effect of electrodes. Due to the fact that the molecule is located between two electrodes connected to the power source, the electric field created by the electrodes changes the energy level of the molecule. Therefore, the total change, U, in the energy level of the molecule includes the change due to charging effect, U_C, and the change caused by electrostatic effect, U_{ES}, which is $U = U_C + U_{ES}$. These energies can be approximated using the planar capacitor model. Considering the surface of gold electrodes and the molecule as parallel plates of a planar capacitor, the electrostatic capacitance is $C_{ES} = C_S + C_D$, as shown in Fig. 2. The charging energy of the molecule is given as [17]:

$$U_C = \frac{q^2}{C_{ES}}(N - N_0), \quad (4)$$

where, q is the elementary charge, N_0 denotes the number of electrons in the molecule before charge transfer, and N is the number of electrons after charge transfer, which can be calculated from the following equations.

$$N_0 = \int_{-\infty}^{+\infty} g(\varepsilon) f_m(\varepsilon) d\varepsilon \quad (5)$$

$$N = \int_{-\infty}^{+\infty} \frac{\Gamma_S f_S(\varepsilon, \mu_S) + \Gamma_D f_D(\varepsilon, \mu_D)}{\Gamma_S + \Gamma_D} g(\varepsilon - U) d\varepsilon \quad (6)$$

In (5) and (6), f_m, f_S and f_D are the Fermi-Dirac electron distribution functions of the molecule, source and drain electrodes, respectively. The current of the device is described by the self-consistent field (SCF) approach as follows [20],

$$I = \frac{q}{\hbar}\int_{-\infty}^{+\infty} g(\varepsilon - U)\frac{\Gamma_S \Gamma_D}{\Gamma_S + \Gamma_D}(f_S(\varepsilon, \mu_S) - f_D(\varepsilon, \mu_D))d\varepsilon \quad (7)$$

The applied voltage, V_{DS}, changes the electrochemical potential energy levels of the source and drain electrodes as, $-qV_{DS} = \mu_D - \mu_S$.

The Fermi level of the molecule changes when the molecule moves between the source and the drain since the source and drain capacitors are changed. Thus, using the simple capacitor voltage divider circuit, the distant

dependent Fermi energy of the molecule is,

$$E_F(z) = -\frac{z}{l}(\mu_D - \mu_S) + \mu_S. \quad (8)$$

The SCF algorithm, which provides the solution to the above equations is shown in Fig. 4.

SCF-Algorithm
Input V_{DS}, initial guess for U_{old} loop: calculate N based on (6) calculate U_{new} based on (4) if $\lvert U_{new} - U_{old}\rvert < \alpha$ then go to current else $U_{new} = U_{old} + \beta(U_{new} - U_{old})$ go to loop current: calculate current using (7) end

Fig. 4: The self-consistent field (SCF) algorithm for current calculation.

Results and Discussions

Considering the effects of the geometry of the metal-molecule chemical bonds and the electrostatic potential energy of the molecule, we investigate the effect of moving the BDT between the two electrodes on the three physical quantities; namely, the electrostatic capacitance, the potential energy of the molecule, and the device conductance.

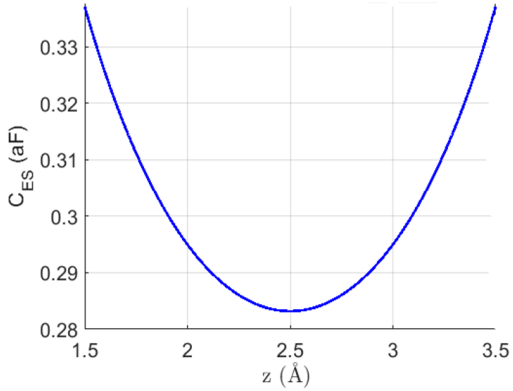


Fig. 5: The electrostatic capacitance vs. the distance of molecule from the source electrode, z.

A. The Potential Energy of the Molecule

Fig. 6 shows the change in the electrostatic potential of the molecule based on the change in the charging energy described by (4). when its distance from the source electrode changes from 1.5 Å to 3.5 Å. The results show that the electrostatic potential is not equal for the corresponding positive and negative voltages, when $z \neq l/2$. In the $R_{QD} < z < l/2$ range, the electrostatic potential in positive applied voltages has a greater value than in negative voltages. Because when the molecule is closer to the source, using (2) and (3) result in $\Gamma_S \gg \Gamma_D$. Consequently, for positive voltages, the charge transferred from the source to the molecule is greater than the charge transferred from the molecule to the drain per unit time, which ultimately causes the molecule to lose a small amount of charge. But in negative voltages, the charge transferred from the drain to the molecule is less than the charge transferred from the molecule to the source per unit of time, and thus, the molecule loses a large amount of charge.

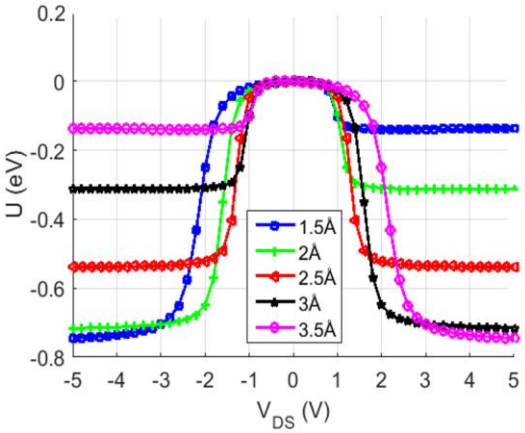


Fig. 6: The potential energy of the molecule vs. the distance of the molecule from the source electrode, z.

Based on the results depicted in Fig. 5 and using (8), it can be deduced that the electron transfer process in $V_{DS} > 0$ takes place in lower applied voltages. In the range $l/2 < z < l$ since $\Gamma_S \gg \Gamma_S$, the molecule loses more charge in $V_{DS} > 0$ than in $V_{DS} < 0$, and this process takes place in less negative voltages. Placing the molecule at $z = l/2$ results in $\Gamma_S = \Gamma_S$, and as a consequence, the amount of charge transferred from the molecule to the electrodes is equal in positive and negative voltages. The asymmetry properties suggest that by placing the molecule near one electrode, a molecular diode can be manufactured.

B. The Device Conductance

Fig. 7 and Fig. 8 depict the current-voltage and conductance-voltage characteristics of the Au-BDT-Au molecular device, when the distance of the molecule from the source electrode changes from 1.5Å to 3Å. It can be seen that when $z \neq l/2$, the electrical characteristics of the device are asymmetric and the asymmetry increases when the molecule deviates more from the middle point of the electrodes. As the molecule approaches the source, $R_{QD} < z < l/2$, the conductance is greater in the $V_{DS} > 0$ region than in $V_{DS} < 0$ region. Because, in $V_{DS} > 0$ region, the HOMO drops in the bias window in lower voltages and the conduction begins. The conduction starts when μ_D reaches the HOMO level at positive voltages, and when μ_S reaches the HOMO level at negative voltages. Fig. 8 shows that, in $R_{QD} < z < l/2$, the conductance peak is larger for $V_{DS} > 0$ than for $V_{DS} < 0$. Because, when the molecule moves in this region, $\Gamma_S \gg \Gamma_S$, and as a result, the electron transfer rate from the source to the molecule is higher than the electron transfer rate from the drain to the molecule, for $V_{DS} > 0$. Therefore, at positive voltages, the current reaches its maximum value with a greater slope. Instead, in the $l < z < l/2$ region, the conductance is higher at $V_{DS} > 0$ than at $V_{DS} > 0$. At $z = l/2$, the current-voltage (Fig. 7) and

the conductance-voltage (Fig. 8) characteristics are completely symmetric. The current-voltage and the conductance-voltage characteristics show that the maximum conductance occurs at $z = l/2$, and it is decreased when the molecule approaches the source or the drain electrode. Fig. 7 depicts the turn-on voltage is $V_{on} \approx 1V$ at $z = l/2$, which agrees with the measurement results [3] and first-principles calculations. Moreover, the conductance gap depicted in Fig. 8 is $V_{gap} \approx 2eV$, which is in agreement with the previous studies [3], [26].

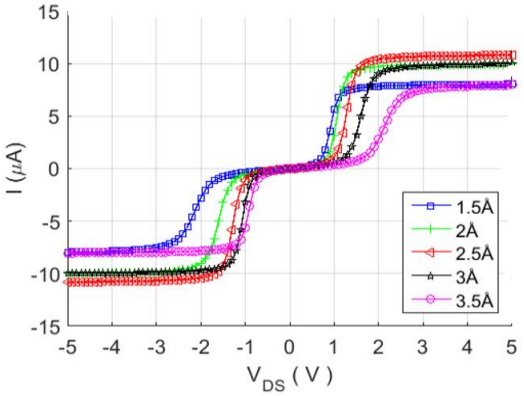


Fig. 7: The current-voltage characteristic of the device vs. the distance of the molecule from the source electrode, z.

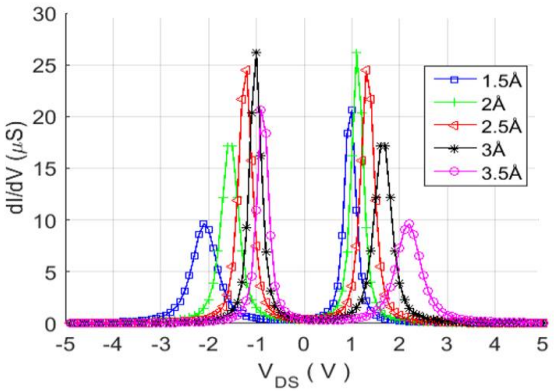


Fig. 8: The conductance-voltage characteristic of the device vs. the distance of the molecule from the source electrode, z.

Conclusions

In this paper, within the framework of the single-level quantum dot model, we have shown that the conductance of the Au-BDT-Au molecular device is maximized when it is placed at the midpoint of the distance between the source and the drain electrodes, where the electrostatic capacitance reaches is minimum. When the molecule approaches one electrode, the capacitance is increased, while the conductance is decreased. Moreover, the charge transferred through the HOMO is unequal in corresponding positive and negative voltages. In the symmetric distances from $z = l/2$, the charge transfer is equal for corresponding negative and positive applied voltages. The turn-on voltage, $V_{on} \approx 1V$ and the conductance gap, $V_{gap} \approx 2eV$ are in agreement with other theoretical studies [26] and practical measurements [3]. The results also indicate that displacing the molecule leads to rectification characteristics, which finds application in the manufacturing process of molecular diodes.

Author Contributions

The quantum chemical calculations have been done by E. Rahimi. Except that, all authors have contributed equally to developing the code, developing discussions, presenting the results and writing the article.

Acknowledgment

The authors are grateful to the editor and the reviewers for the careful and timely review process.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

DFT	Density Functional Theory
NEGF	Non-Equilibrium Green's Function
MCBJ	Mechanically Controllable Break Junction
BDT	Benzenedithiol
HOMO	Highest Occupied Molecular Orbital
LUMO	Lowest Unoccupied Molecular Orbital
SCF	Self-Consistent Field
TB	Tight-Binding
LDA	Local Density Approximation
QD	Quantum Dot

References

[1] P. Damle, T. Rakshit, M. Paulsson, S. Datta, "Current-voltage characteristics of molecular conductors: two versus three terminal," IEEE Trans. Nanotechnol., 1(6): 145-153, 2002.

[2] A. Aviram, M. A. Ratner, "Molecular rectifiers," Chem. Phys. Lett., 29(2): 277–283, 1974.

[3] M. A. Reed, C. Zhou, C. J. Muller et al., "Conductance of a molecular junction," Science, 278(2536): 252-254, 1997.

[4] X. D. Cui et al., "Reproducible measurement of single-molecule conductivity ", Science, 294(5336): 571-574, 2001.

[5] B. Xu, N. J. Tao, "Measurement of single-molecule resistance by repeated formation of molecular junction," Science, 301(5637): 1221-1223, 2003.

[6] X. Xiao, B. Xu, J. Tao, "Measurement of single molecule conductance: benzenedithiol and benzenedimethanedithiol," Nano Lett., 4(2): 267–271, 2004.

[7] F. Chen, X. Li, J. Hihath, Z. Huang, N. Tao, "Effect of anchoring groups on single-molecule conductance: comparative study of thiol-, amine-, and carboxylic-acid-terminated molecules," J. Am. Chem. Soc., 128(49): 15874–15881, 2006.

[8] X. Li, J. He, J. Hihath, B. Xu et al., "Conductance of single alkanedithiols: conduction mechanism and effect of molecule - electrode contacts" J. Am. Chem. Soc., 128(6): 2135–2141, 2006.

[9] Y. Jang, S. J. Kwon, J. Shin, H. Jeong, W. T. Hwang, J. Kim, J. Koo, T. Y. Ko, S. Ryu, G. Wang, and T. W. Lee, "Interface-engineered charge-transport properties in benzenedithiol molecular electronic junctions via chemically p-doped graphene electrodes," ACS Appl. Mater. Interfaces, 9(48): 42043-42049, 2017.

[10] D. Olson, N. Hopper, W. T. Tysoe, "Surface structure of 1, 4-benzenedithiol on Au (111)," Surf. Sci., 467: 21-25, 2016.
[11] E. Venkataraman, V. Amadi, T. S. Zaborniak, P. Zhang, C. Papadopoulos, "Negative differential resistance and hysteresis in self-assembled nanoscale networks with tunable molecule-to-nanoparticle ratios," Phys. Status Solidi (b), 257: 2000019, 2020.
[12] S. Kobayashi, S. Kaneko, M. Kiguchi, K. Tsukagoshi, T. Nishino, "Tolerance to stretching in thiol-terminated single-molecule junctions characterized by surface-enhanced raman scattering," J. Phys. Chem. Lett., 11(16): 6712-6717, 2020.
[13] Y. Naitoh, Y. Tani, E. Koyama, T. Nakamura, T. Sumiya, T. Ogawa, G. Misawa, H. Shima, K. Sugawara, H. Suga, H. Akinaga, "Single-molecular bridging in static metal nanogap electrodes using migrations of metal atoms," J. Phys. Chem. C, 124: 14007-1415, 2020.
[14] K. Horiguchi, M. Tsutsui, S. Kurokawa, A. Sakai, A., "Electron transmission characteristics of Au/1, 4-benzenedithiol/Au junctions," Nanotechnology, 20: 025204, 2008.
[15] L. L. Lin, C. K. Wang, Y. Luo, "Inelastic electron tunneling spectroscopy of gold–benzenedithiol–gold junctions: accurate determination of molecular conformation," ACS Nano, 5: 2257-2263, 2011.
[16] Y. Teramae, K. Horiguchi, S. Hashimoto, M. Tsutsui, S. Kurokawa, A. Sakai, "High-bias breakdown of Au/1, 4-benzenedithiol/Au junctions," Appl. Phys. Lett., 93: 083121, 2008.
[17] K. Baheti, J. L. Malen, P. Doak, P.Reddy, S.Y. Jang, T. D. Tilley, A. Majumdar, R. A. Segalman, "Probing the chemistry of molecular heterojunctions using thermoelectricity," Nano Lett., 8: 715-719, 2008.
[18] A. W. Ghosh, S. Datta, "Molecular conduction: paradigms and possibilities," J. Comput. Electron., 1: 515-525, 2002.
[19] M. Strange, C. Rostgaard, H. Hakinen, K. S. Thygesen, "Self-consistent GW calculations of electronic transport in thiol- and amine-linked molecular junctions," Phys. Rev. B, 83: 115108, 2011.
[20] S. Datta, "Electrical resistance: an atomic view," Nanotechnology, 15: S433, 2004.
[21] Z. H. Zhang, Z. Yang, J. H. Yuan, H. Zhang, X. Q. Ding, M. Qiu, "First-principles investigation on electronics characteristics of benzene derivatives with different side groups," J. Chem. Phys. 129: 094702, 2008.
[22] K. Stokbro, J. Taylor, M. Brandbyge, J. L. Mozos, P. Ordejon, "Theoretical study of the nonlinear conductance of di-thiol benzene coupled to Au (1 1 1) surfaces via thiol and thiolate bonds," Comput. Mater. Sci., 27: 151-160, 2003.
[23] E. G. Emberly, G. Kirczenow, "The smallest molecular switch," Physical Review Letters," 91: 188301, 2003.
[24] Y. Xue, M. A. Ratner, "Theoretical principles of single,Äêmolecule electronics: a chemical and mesoscopic view," Int. J. Quant. Chem., 102: 911-924, 2005.
[25] J. Tomfohr, O. F. Sankey, "Theoretical analysis of electron transport through organic molecules," J. Chem. Phys., 120: 1542-1554, 2004.
[26] M. D. Ventra, S. D. Pantelides, N. D. Lang, "First-principles calculation of transport properties of a molecular device," Phys. Rev. Lett., 84: 979-982, 2000.
[27] Y. Kim, T. Pietsch, A. Erbe, W. Belzig, E. Scheer, "Benzenedithiol: a broad-range single-channel molecular conductor," Nano Lett., 11: 3734-3738, 2011.
[28] R. B. Pontes, A. R. Rocha, S. Sanvito, A. Fazzio, A. J. R. Silva, "Ab initio calculations of structural evolution and conductance of benzene-1, 4-dithiol on gold leads," ACS Nano, 5: 795-804, 2011.
[29] W. C. Bauschlicher, A. Ricca, "Modelling of benzene-1, 4-dithiol on a Au (1 1 1) surface," Chem. Phys. Lett., 367: 90-94, 2003.
[30] S. V. Faleev, F. Léonard, D. A. Stewart, M. van Schilfgaarde, "Ab initio tight-binding LMTO method for nonequilibrium electron transport in nanosystems," Phys. Rev. B, 71: 195422, 2005.
[31] A. M. Scheer, G. A. Gallup, P. D. Burrow, "Unoccupied orbital energies of 1, 4-benzenedithiol and the HOMO–LUMO gap," Chem. Phys. Lett., 466: 131-135, 2008.
[32] T. Shimazaki, K. Yamashita, "A theoretical study of molecular conduction. V. NEGF-based MP2 approach," Int. J. Quant. Chem., 109: 1834-1840, 2009.

Biographies

Ehsan Rahimi received his B.Sc. from Ferdowsi University of Mashhad and his M.Sc. and Ph.D. from Iran University of Science and Technology in 2012, all in Electronic Engineering. He is currently an assistant professor at Shahrood University of Technology. His research interests include quantum and molecular electronic devices.

- Email: erahimi@shahroodut.ac.ir
- ORCID: 0000-0003-0774-8436
- Web of Science Researcher ID: GQR-0709-2022
- Scopus Author ID: 25122904100
- Homepage: https://shahroodut.ac.ir/fa/as/?id=S715

Saba Dorouki received her B.Sc. and M.Sc. in Electronic Engineering from Shahrood University of Technology in 2018. Her research interests include the design of single molecule devices.

- Email: dorouki.s@gmail.com
- ORCID: NA
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:
E. Rahimi, S. Dorouki, "Displacement effects on the electrical characteristics of a single-molecule device," J. Electr. Comput. Eng. Innovations, 11(1): 183-188, 2023.

DOI: 10.22061/JECEI.2022.9195.585

URL: https://jecei.sru.ac.ir/article_1771.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

An Ultra-Low Power Ternary Multi-Digit Adder Applies GDI Method for Binary Operations

N. Ahmadzadeh Khosroshahi[1], M. Dehyadegari[1,2,*], F. Razaghian[1]

[1]*Department of Electrical Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran.*

[2]*Department of Computer Engineering, k. N. Toosi University of Technology, Tehran, Iran.*

Article Info

Article History:
Received 23 June 2022
Reviewed 29 July 2022
Revised 23 August 2022
Accepted 06 September 2022

Keywords:
CNTFET
Low power ALU
Reversible logic
Ternary logic
Multi-valued logic

*Corresponding Author's Email Address:
Dehyadegari@kntu.ac.ir

Abstract

Background and Objectives: A novel low-power and low-delay multi-digit ternary adder is presented in this paper which is implemented in carbon nanotube field effect transistor (CNTFET) technology.
Methods: In the proposed design, CNTFET technology is used where reducing the power consumption is the main priority. A CNTFET's geometry directly determines the threshold voltage. In this architecture, at each stage, a half adder is applied to generate the intermediate binary signals which are called half-sum (HS) and half-carry (HC). To implement the binary operations of the design, the gate diffusion input (GDI) method is applied. A significant reduction of the power consumption is achieved while the PDP is improved.
Results: The proposed designs are simulated in synopsis HSPICE simulator. The Stanford 32 nm CNTFET technology is applied while the power supply is 0.9 v and the simulation is performed at room temperature. In this case, the pitch value of 20nm are chosen where the number of the tubes taken are 3. In this work a GDI based sum generator and a low-power encoder are used to calculate the final sum value of each stage. Furthermore, the proposed carry generation/propagation block results in a remarkable reduction of the overall propagation delay time. The simulation reveals a significant improvement in terms of the power consumption (up to 27%), the PDP (up to 41%) and the FO4 delay (up to 20%).
Conclusion: An efficient CNTFET based multi-digit ternary adder has been presented in this paper. The Synopsis HSPICE simulator is used where Stanford 32 nm CNTFET model are applied to simulate the design. According to the results, a significant saving in average power consumption is achieved where the power-delay product (PDP) is improved by 41% compared to the best existing design.

Introduction

Binary combinational logical circuits have an important concern in terms of the design complexity [1]. Generally, interconnects generate 70% of total on-chip capacitance [41]. Due to that, they are the most important sources of the power dissipation in a VLSI chip [2], [42], [43]. Recently, Multi Valued Logic (MVL) has been of interest.

It has brought many benefits in terms of hardware utilization, interconnects, number of digits required and average power consumption. Also reducing the scale size to the Nanometer limits the usage in the implementation of low-power digital designs. A review of the literates depicts that non-silicon devices of multi-valued logic systems are an offered circuit to implement high-

efficiency digital architectures. Since hardware complexity is directly proportional to the radix of a MVL system [3], calculations of base-3 (close to *e* = 2.718) is an optimum choice called ternary logic. New molecular devices have been widely developed to succeed conventional silicon-based CMOS technology. CNTFETs may be an alternative to silicon MOSFET devices due to their excellent operating characteristics with some similarity in their inherent characteristics [4].

The similar dimension of an N-type and a P-type CNTFET has the same mobility. It affects the sizing of the transistors in a complex circuit. CNTFETs are more suitable for implementing three-valued logic circuits because they can obtain multiple threshold voltages by changing the physical dimensions of the carbon nanotubes (CNTs). Chirality is a well-known characteristic of CNTFETs. It depends on the diameter of the CNTs. Any variation of the chirality affects the CNT's threshold voltage.

Multiple threshold voltages are required to implement a three-valued logic circuit. So, it can be obtained by CNTFETs of different chirality [5]. A wide different CNTFET-based ternary operations and logic circuits such as the logic gates, adders, and multipliers are discussed in [6]-[23].

Also, two CNTFET based Ternary ALUs (TALUs) which include a huge number of transistors have been presented in [21] and [22]. These power-consuming structures use a decoder-encoder based approach. A novel CMOS based ternary ALU is proposed by [26]. It employees the decoders, the ternary gates and the ternary buffers to implement various digital blocks like the adder, the subtractor and the multiplier. In this work [26], the depletion type MOS transistors and the large off-chip resistors significantly increase the power and the area consumption. A modified version of this TALU [26] is presented by [27] where the adder and the subtractor modules are in the same block and the outputs are multiplexed using the ALU select signals. There is a modified version of [27] which is proposed by [28], where the adder, the subtractor and the Ex-OR gate are combined in a single module. In this paper, a novel multi-digit ternary adder is introduced while it is more efficient in terms of the power and the PDP metrics. The proposed architecture applies the GDI method for implementing the binary operations.

Hence, an improvement is expected in terms of the power consumption. The rest of this paper is organized as the following: section two gives a concept of the ternary logic and the multi-digit ternary adders.

Our detailed proposed design is presented in Section three. In section four the simulation results and the comparisons have been fully discussed. Finally, section five is the conclusion.

Backgrounds of Research

A. CNTFET Based Ternary Logic

A three-valued logic (ternary logic) is an offered type of multi-valued logics. In this, the three truth values indicating *true*, *false* and an indeterminate 3rd value are included. In this case, 0, 1 and 2 directly correspond to a voltage value of 0, Vdd/2, and Vdd. In fact, f(y) is a ternary logic function of y= {y_1, y_2, …, y_n} maps $\{0.1.2\}^n$ to $\{0.1.2\}$. In this three-state logic, AND and OR functions are defined as min {y_i, y_j} and max {y_i, y_j} respectively. In three-state logic, the inverting gate, which is widely used in the design of other gates like NAND and NOR and most of the logic circuits, is defined as NTI, PTI and STI (which stand for negative, positive and standard ternary inverter respectively). Threshold voltage changes in exchange for CNT diameter changes have made them a viable option in MVL circuits. A graphite sheet is rolled up to create a CNT. In this case, C is the roll-up vector ($C = n\bar{a} + m\bar{b}$).

In this equation, $\bar{a}$ and $\bar{b}$ are the unit lattice vectors and the pair (n,m) is the chirality vector of the carbon nanotube. The integer values (n,m) determine the type of a CNT: metallic or semiconducting. If (n = m) and (n-m = 3i e.g. i is an integer), a CNT is a metal.

Otherwise, it treats as a semiconductor. The diameter of a CNT in Nano-meter is expressed as: $D_{CNT} \approx 0.0783\sqrt{n^2 + m^2 + n.m}$. Also, the threshold voltage of a FET in volt and a CNT's diameter in (nm) are inversely proportional as: $V_{th} = 0.43/D_{CNT}$. The most common chirality vectors that are used for implementation of MVL circuits are (19,0), (13,0) and (10,0) so that they provide threshold voltages of 0.289, 0.428 and 0.559 for an N-CNTFET and equivalent values with negative sign for a P-CNTFET respectively [29].

B. The GDI Transistor Level Implementation Technique

The gate diffusion input (GDI) method [30]-[34] is a well-known method which is developed based on a simple cell including two transistors.

It is shown in Fig. 1. G: the common gate input of the NCNT and the PCNT transistors (in CMOS technology NMOS and PMOS are used), P: the outer diffusion node of the PCNT transistor, N: the outer diffusion node of the NCNT transistor, D: the common diffusion node of the both transistors.

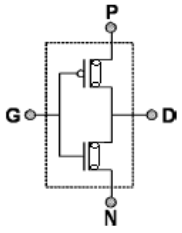


Fig.1: The basic cell of the GDI method [34].

Depending upon the circuit structure, P, N, and D would be either input or output ports. Table 1 includes several simple logic functions implemented by the GDI method.

Various Boolean functions could be simply implemented when a simple configuration changes occurred at the inputs. So, the GDI method gains simpler design, fewer transistor of the structure, and lower power dissipation [34].

Table 1: Several functions which are implemented by the basic GDI cell [34]

N	P	G	D	FUNCTION
$'\mathbf{0}'$	B	A	$\bar{A}B$	F1
$\boldsymbol{B}$	$'1'$	A	$\bar{A}+B$	F2
$'\mathbf{1}'$	B	A	$A+B$	OR
$\boldsymbol{B}$	$'0'$	A	AB	AND
$\boldsymbol{C}$	B	A	$\bar{A}B+AC$	MUX
$'\mathbf{0}'$	$'1'$	A	$\bar{A}$	NOT

C. Ternary Adders

Reviewing the literates introduces a simple architecture which implies several single-digit adders to implement a multi-digit adder [45]. This suffers from a long propagation delay time. A similar structure is presented by [46]. Another design including two half-adders are applied to generate the sum where a standalone circuit is utilized to generate the carry signal. The carry is ternary in nature and it is propagated to the next stage [47]. In some ternary adders provided in the literates, a ternary decoder is utilized to generate the binary version of the input. The three outputs of the decoder $(Y^0.Y^1.Y^2)$ are created according to the input (Y) as in (1). Since there are two possible states of the decoder outputs, the binary logic gates are utilized to generate the intermediate binary values. Then, these intermediate binary values produce the ternary sum/carry. The adder presented in [35] has a simple structure. A fast carry generation block is included in the design. Hence, an improved overall delay for multi-digit adders is expected spending large average power consumption. A multi-digit adder is introduced by [36] in which two half-adders are applied to produce the sum value. It applies a self-governing circuit to create the carry. The carry signal is ternary in nature while it needs to be transferred to the next stage. The most important drawback of the design is huge power consumption and a large amount of delay. Why so at any stage of the adder, a voltage divider is applied made the ternary carry. An energy efficient single-digit/multi-digit adder is illustrated in [36]. In the first stage of the architecture, positive and negative ternary complements of the inputs are generated. Then, the outputs of the first stage and the original inputs are fed to a network of transistors to estimate the intermediate output. This structure gains a moderate power consumption and PDP in comparison with the other existing designs.

$$\begin{aligned} y_i + y_j &= \max\{y_i.y_j\} \\ y_i.y_j &= \min\{y_i.y_j\} \\ \overline{y_i} &= 2 - y_i \end{aligned} \tag{1}$$

The Proposed Multi-Digit Ternary Adder

The proposed multi-digit ternary (multi-trit) adder applies some intermediate binary calculations to perform the operation. At each stage, two decoders are used to prepare the intermediate binary signals from the inputs. Then, the intermediate binary signals are applied to a half adder circuit. The outputs of the half adder are also binary signals which in turn drive the final sum generator circuit. The gate diffusion input method (GDI) method is utilized as a power efficient method to design the binary circuits. Finally, an encoder converts the binary sum to a ternary digit. The proposed structure is based on the model which is introduced in [37]. In the proposed architecture, a carry generator determines the carry-out of *i*th stage using the half-adder outputs of *i*th stage and the carry signals of the $(i-1)$ th stage. It leads to a reduction in carry propagation time while the complexity is slightly increased. Since the half adders work in parallel while a part of the carry-out is pre-calculated by the half adder, the carry propagation delay is decreased. To explain the design in detail, $(A_{N-1}.\ldots.A_1.A_0)$, $B(B_{N-1}.\ldots.B_1.B_0)$, $C_{in}$are considered as the ternary inputs. $Sum(Sum_{N-1}.\ldots.Sum_1.Sum_0)$, and *Cout* are the ternary outputs. The intermediate binary signals are expressed by Y_i^j which corresponds to *i*th digit-adder stage. There are two possible value of the intermediate binary signals: logic 2 (if $Y = j$) or logic 0 (if $Y \neq j$), where j {0, 1, 2}. For instance, A_0^1 corresponds to input of 0th digit-adder stage whose value is logic 2 only if ternary signal *A* is equal to logic 2. Fig. 2, shows the block diagram of the proposed multi-digit ternary adder for the ith stage. In this structure, ternary inputs are decoded to binary signals. A binary half adder computes the intermediate signals half-sum (HS) and half-carry (HC) which in turn are used to compute the final sum/carry signals in the binary sum/carry generator block. Finally, an encoder converts the binary signals to a ternary value. The binary operations, which are applied to compute the HS and HC signals, are shown in Fig. 3 [38]. A power and delay optimized decoder are designed to generate the mutually exclusive binary signals as in (1). In this design, an NTI is

used to calculate A_i^0, a pair of PTI-GDI inverter is used to compute A_i^2 and finally a GDI NOR is applied to calculate A_i^1.

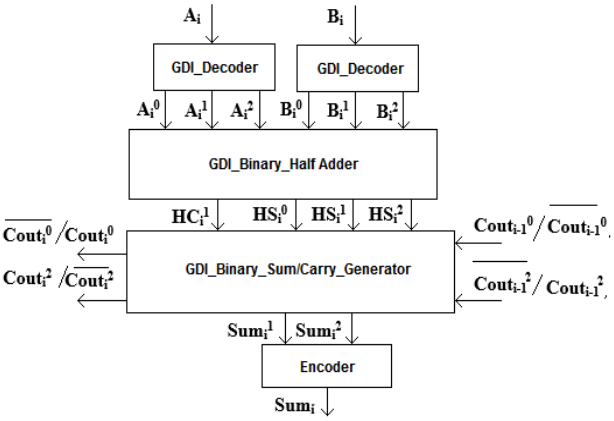


Fig. 2: The ith stage block diagram of the proposed ternary adder.

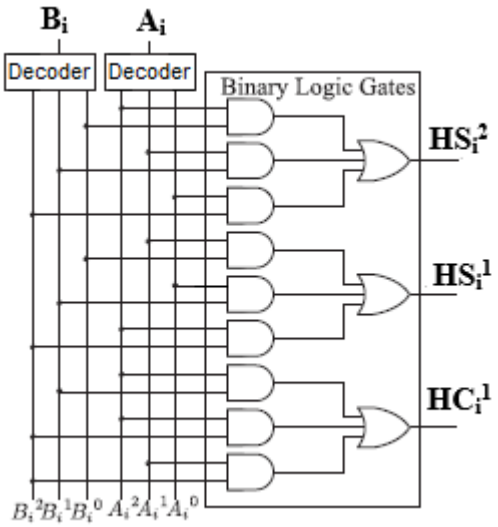


Fig. 3: The binary operations of the half adder to calculate the intermediate binary signals [38].

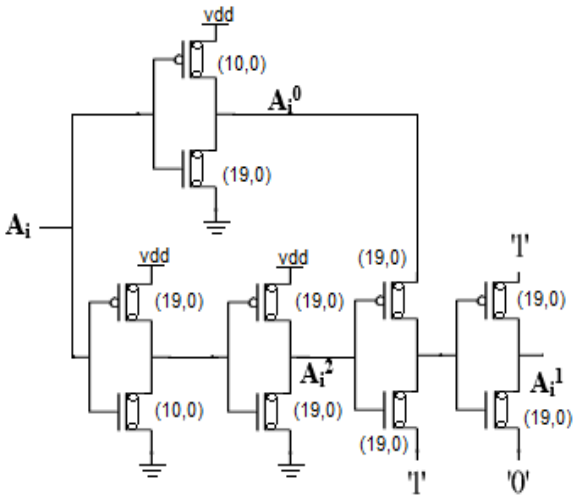


Fig. 4: The proposed GDI-based decoder.

The transistor level implementation of the proposed GDI-based decoder is shown by Fig. 4. These binary signals are fed to GDI-binary half adder to compute the intermediate binary signals. The proposed half adder generates mutually exclusive intermediate binary signals HS_i^0, HS_i^1 and HS_i^2. HC_i^1 represents the carry signal. Table 2 is the truth table of a ternary half adder which represent HS and HC for all of the inputs states. According to Table 2, (2)-(4) are obtained. Since HS_i^0, HS_i^1 and HS_i^2 are mutually exclusive signals, only two of these signals should be calculated. The third one could be implemented applying a NOR operation [37]. In this case, the GDI method is used to implement (2)-(4). Noting to the equations, "OR, AND, NOT" operations are implemented using the basic GDI cell.

Clearly, the transistor level implementation of the proposed GDI-binary half adder is presented by Fig. 5 where all the N-CNTs and the P-CNTs have the chirality of (19,0). The circuits in Fig. 5(a), (b), (c) and (d) are the CNTFET-based implementation of HS_i^0, HS_i^1, HC_i^1 and HS_i^2 respectively.

$$HS_i^2 = (A_i^1 + \bar{B}_i^1)\ (A_i^2 + \bar{B}_i^0)\ \ (A_i^0 + \bar{B}_i^2) \tag{2}$$

$$HS_i^1 = (A_i^1 + \bar{B}_i^0)\ (A_i^2 + \bar{B}_i^2)\ \ (A_i^0 + \bar{B}_i^1) \tag{3}$$

$$\overline{HC}_i^1 = (A_i^0 + \bar{B}_i^2)\ \ (\bar{A}_i^2 + \bar{B}_i^1) \tag{4}$$

Table 2: The truth table of a ternary half adder

A	B	HS	HC
0	0	0	0
0	1	1	0
0	2	2	0
1	0	1	0
1	1	2	0
1	2	0	1
2	0	2	0
2	1	0	1
2	2	1	1

Then, the intermediate binary signals are applied to the GDI-binary sum/carry generator block to calculate the sum and the carry.

There are two separated CNTFET implementation for the sum and the carry generators. The GDI method is also utilized for the binary operations. Table 3 is the truth table of the ith stage output carry ($Cout_i$) in order to the sum of

the inputs (A_i, B_i) and the previous stage carry (input carry of the ith stage: $Cout_{i-1}$).

stages while it generates $\overline{Cout_i^2}$ *and* $Cout_i^0$ for the odd stages.

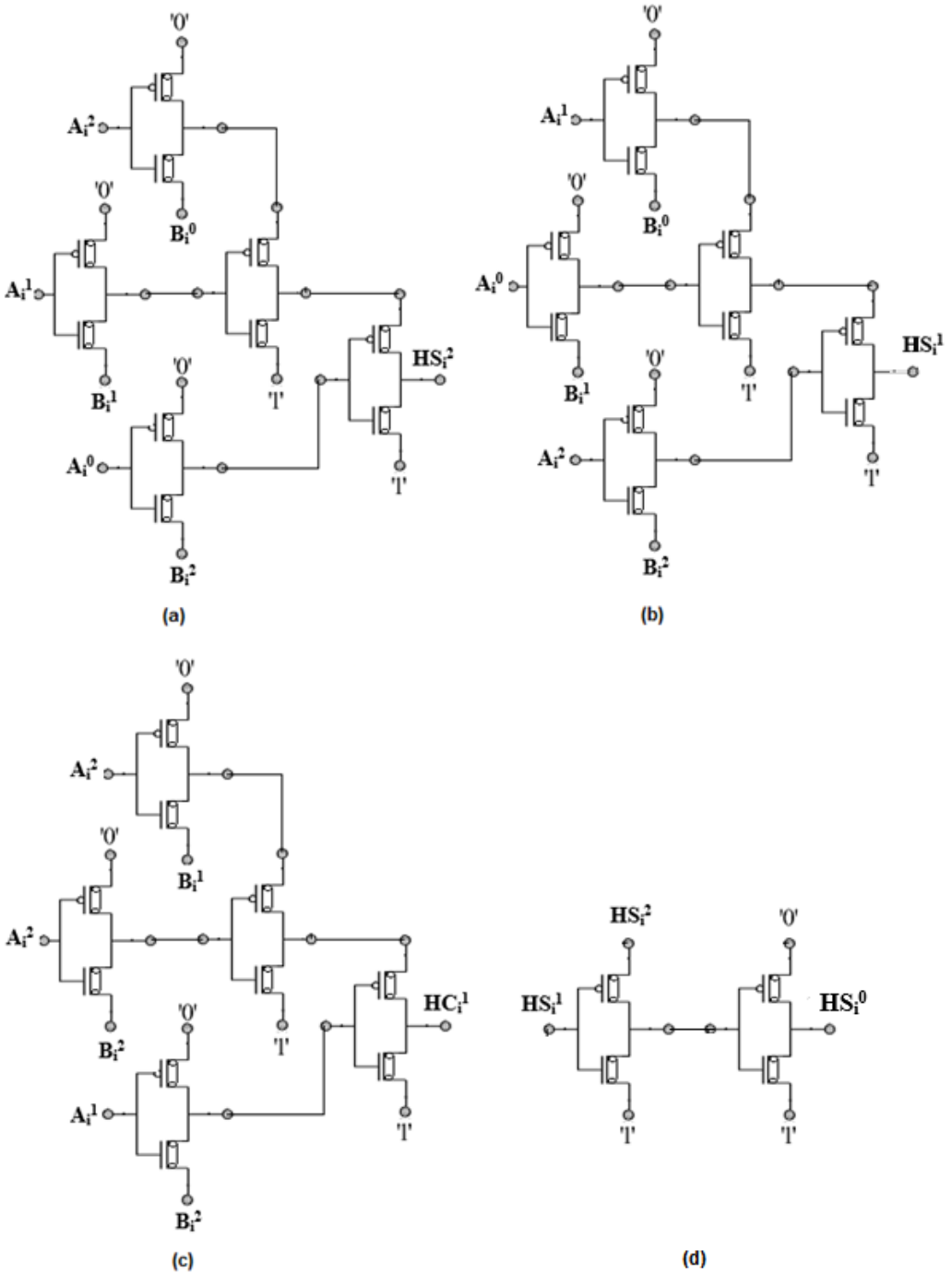


Fig. 5: The proposed GDI-binary half adder. All the CNTFETs have the chirality of (19,0).

The intermediate binary signals (HS_i, HC_i) are also included in the table. According to the truth table, the carry signal could be simply calculating by the following logical (5), (6). That is clear that only binary AND, OR and NOT are required to implement these equations. In this case, the basic CNTFET based GDI cell is used to implement the carry generator. The ith stage carry generator generates $\overline{Cout_i^0}$ *and* $Cout_i^2$ for the even

Fig. 6 shows the implementation of the carry generator in the 0th and 1th stage GDI-binary sum/carry generator block where the basic GDI cell is employed to implement the logical operations. The CNTFET based circuits in Fig. 6(a), (b), (c) and (d) are used to implement $Cout_0^0$, $Cout_1^0$, $Cout_0^2$ and $Cout_1^2$ respectively. In this design, all the CNTFETs have the chirality of (19,0).

$$Cout_i^0 = (\overline{HS_i^2} + Cout_{i-1}^0)\ \overline{HC_i^1}\ (HS_i^0 + \overline{Cout}_{i-1}^2) \quad (5)$$

$$Cout_i^2 = HS_i^1\ HC_i^1\ Cout_{i-1}^2 \quad (6)$$

2	0	2	0	1	1
3	1	0	1	1	1
4	1	1	1	1	2

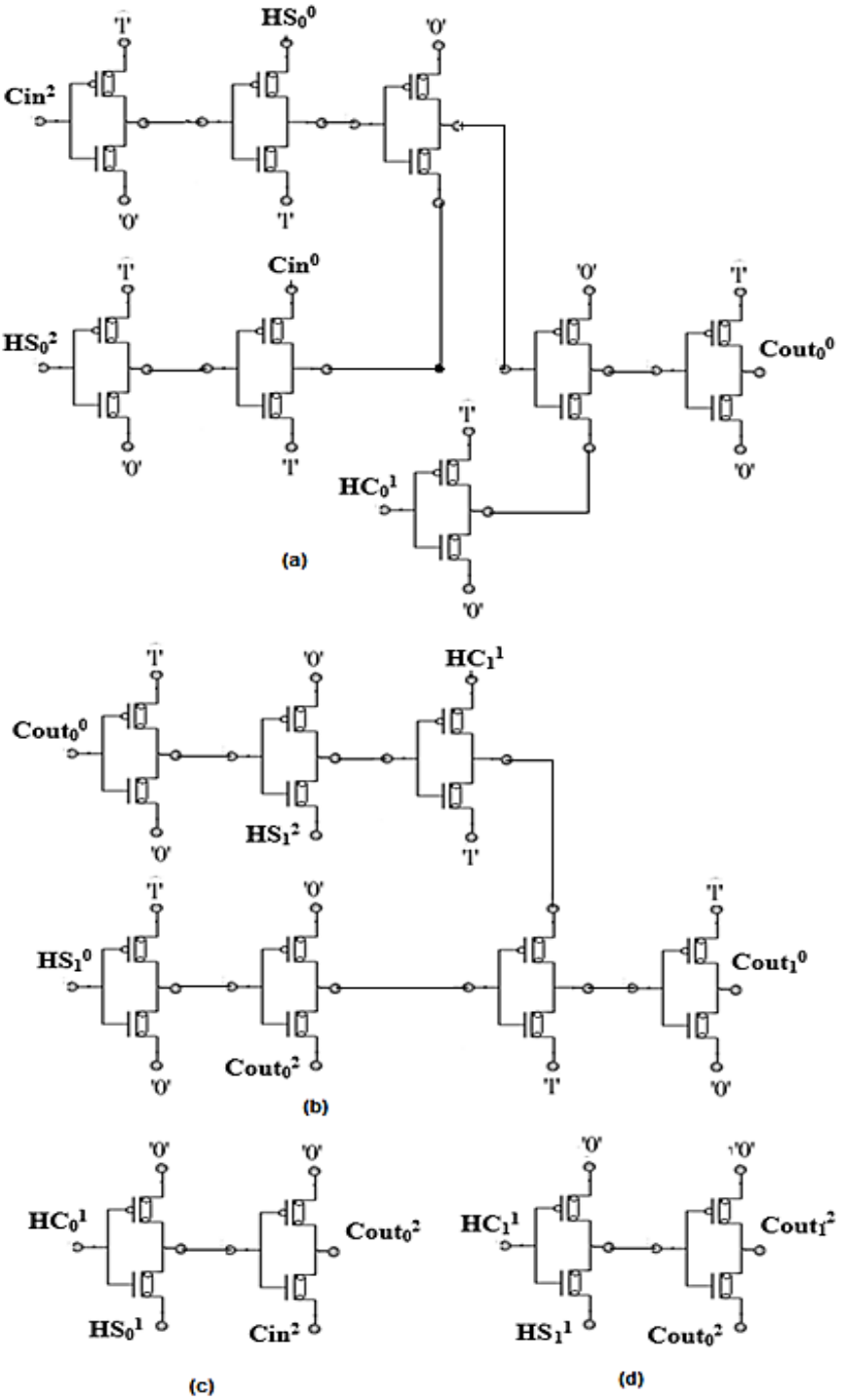


Fig.6. The proposed carry generator in the 0th and 1th GDI-binary sum/carry generator block. All the CNTFETs have the chirality of (19,0).

Table 3: The truth table of the ith stage output carry ($Cout_i$)

Sum (A_i,B_i)	HC_i	HS_i	$Cout_i$ | $Cout_{i-1=0}$	$Cout_i$ | $Cout_{i-1=1}$	$Cout_i$ | $Cout_{i-1=2}$
0	0	0	0	0	0
1	0	1	0	0	1

In the GDI-binary sum/carry generator block, a CNTFET based GDI circuit, which is similar to the HS calculation circuit in the half adder block, is used to generate the final ternary Sum (Sum_i^2, Sum_i^1).

In this case, the intermediate binary signals (HS) and the input carry of the stage is used to generate the final sum. Fig. 7(a), (b) represent the transistor level implementation to generate Sum_i^2 and Sum_i^2, while Fig. 7(c) is the implementation of $Cout_{i-1}^1$.

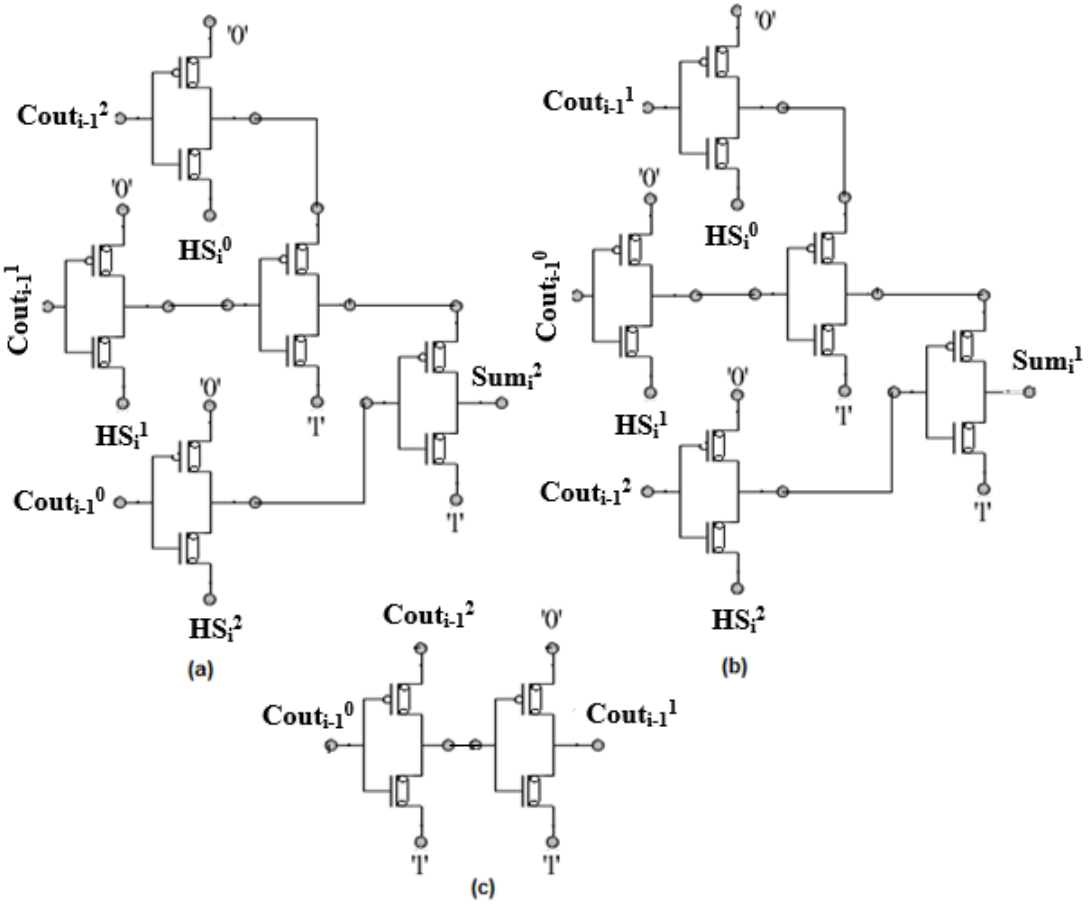


Fig. 7: The proposed final sum generator in the GDI-binary sum/carry generator block. All the CNTFETs have the chirality of (19,0).

The GDI basic cell, which is comprises of a P-CNTFET and an N-CNTFET with the chirality of (19,0), is used to implement the design. Encoder is an essential element in the architecture of a ternary adder that uses binary operations to estimate the binary intermediate signals. In this work, the power optimized encoders that are presented in [39] are applied to generate the ternary sum : stage. The implementation of the encoder which is used in the proposed design for the ternary sum generation at each stage (Sum_i) is presented by Fig. 8. Similarly, Fig. 8(b) represents the encoder which is used to implement the ternary final carry at the last stage ($Cout_{N-1}$).

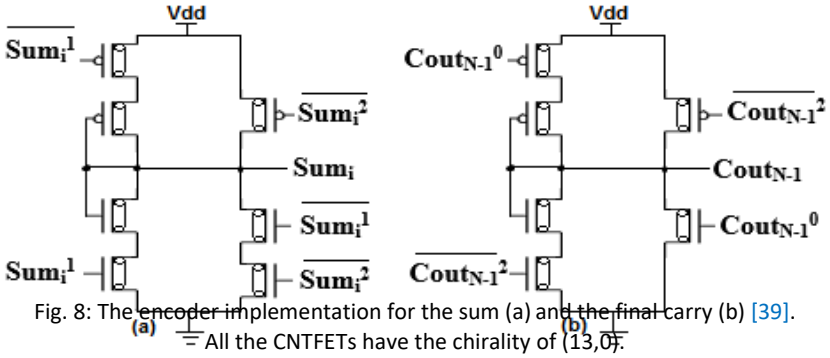


Fig. 8: The encoder implementation for the sum (a) and the final carry (b) [39].
All the CNTFETs have the chirality of (13,0).

All the CNTFETs for the both encoders have the chirality of (13,0). In the encoder circuit, which is applied in this work, when Sum_i^2 is equal to logic 2, logic 2 is generated at the output. In this case, if Sum_i^1 is equal to logic 2, a direct path between VDD and GND is created that forces the output to change to logic 1. If neither of Sum_i^2 and Sum_i^1 are logic 2 then the output is pulled down to logic 0. In this work, gate diffusion input is a method which is used to reduce the power dissipation, the propagation delay time, and the occupation area. The basic GDI cell is used to implement the GDI-decoder, GDI-binary half adder and GDI-binary sum/carry generator.

Due to this, the implementation of the proposed circuit is expected to be more efficient than previous works.

The Simulation Results and Discussion

The proposed designs are simulated in synopsis HSPICE simulator [49]. The Stanford 32 nm CNTFET technology is applied while the power supply is 0.9 v and the simulation is performed at room temperature. In this case, the pitch value of 20nm are chosen where the number of the tubes taken are 3. The main simulation parameters of the CNTs in Hspice simulator are included in Table 7. To have a meaningful comparison with the previous works, the proposed decoder, the proposed single digit adder and the proposed multi-digit adder are investigated in terms of the average power consumption, the power-delay product (PDP) and the fan out of 4 (FO4) delay. In this simulation the average power consumption is achieved by applying a random input pattern with the switching frequency of 500 MHz. The average power is composed of two types: dynamic and static. For instance, the static power consumption of the proposed single-digit adder is 0.51 uW while the average power is reported 0.88 uW. Table 4 is a comparison of the decoders.

Table 4: A comparison of the decoders

Decoder	Power (μW)	Delay-FO4 (ps)	PDP (10^{-18} J)
[38]	0.061	18.9	1.15
[37]	0.053	16.80	0.88
Proposed	0.042	15.28	0.64

Table 5: A comparison of the encoders

Encoder	Power (μW)	Delay-FO4 (ps)	PDP (10^{-18} J)
[38]	0.99	25.2	25.2
[12]	9.31	5.09	47.3
[6]	0.78	7.62	5.94
[39]	0.37	8.53	3.15

Table 6: A comparison of the single-digit adders

Single-digit adder	Power (μW)	Delay-FO4 (ps)	PDP (10^{-18} J)
[12]	32.9	37.27	1.225
[6]	2.47	26.85	0.066
[5]	1.51	134.7	0.204
[10]	7.43	34.28	0.255
[14]	2.11	29.23	0.062
[37]	1.14	37.78	0.043
Proposed	0.88	36.37	0.032

The proposed decoder design results in a reduction of 21% in power consumption, 9% in propagation delay and 22% in PDP compared with [37]. Furthermore, the fan-out requirement of the decoder in [37] is more than the proposed architecture. Hence, a circuit connected to the input of the decoder consumes lower energy in our work. The decoder which is used in the proposed design has lower power spending more delay time while the PDP is improved. Table 5 shows the simulation results of the several already provided encoders where 4 STI gate are connected as the load and a random pattern waveform as the input that results in the same output is applied. An implementation of a single-stage adder, which has no carry chain, could be done applying one stage of the proposed multi-digit adder. In some cases, it is called a full adder (FA). To have an exact consideration of FO4 delay, power and PDP, the output of the single-digit adder is connected to 4 STI as in [37] and the simulation waveform is the same as in [37]. [37] improves the results in terms of the average power and the PDP, but delay is 40% worse. The main reason for the defect is the high delay of the encoder. Although the same encoder is used in the proposed design, the overall delay is greatly reduced due to the GDI-encoder, GDI-binary half adder and GDI-binary sum/carry generator. The single-digit adder of [6] gains lower propagation delay time due to the simple encoder structure. However, the proposed single-digit adder has improved delay by 10% than [6]. In terms of the average power and the PDP, the proposed design could improve the results by 23% and 26% respectively compared with [37]. It's very important to study the effect of voltage variation on the behavior of the proposed CNT based design. Hence, the two graph of the propagation delay and the power consumption are plotted in term of the supply variation for the proposed design, [37] and [14]. Fig. 9 (a) shows that when the current of CNTFETs

increase due to the increase in voltage, the propagation delay decreases. On the other hand, as shown in Fig. 9(b), a positive change in supply voltage reduces the power consumption. Fig. 9(c) and Fig. 9(d) show how the temperature changes affect the design performance. Clearly, the temperature has a direct relation with the propagation delay and a reverse relation with the average power consumption. A metric describes the noise variation immunity

It is used to investigate the noise effect on the performance of a logic circuit. The NIC calculation is presented by [40].

The noise pulses are simply categorized by their width and amplitude [37].

A pulse with an adequate width and amplitude may cause a glitch (spurious switching).

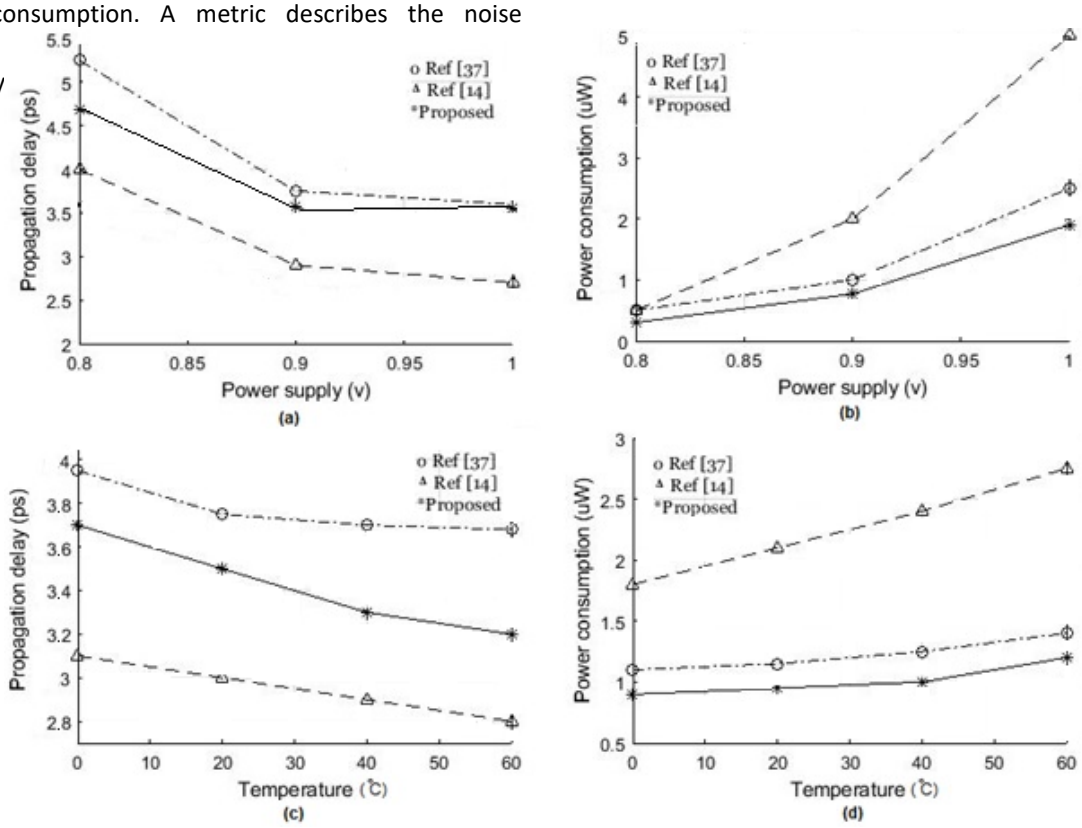


Fig. 9: The effect of the power supply (a, b) and temperature (c, d) variation on delay and power.

Fig. 10 is the NIC of the single-digit adders. Any point above the curve represents a glitch at the output. Fig. 10 clearly illustrates that the proposed single-digit adder has better noise immunity compared with [14] and [37].

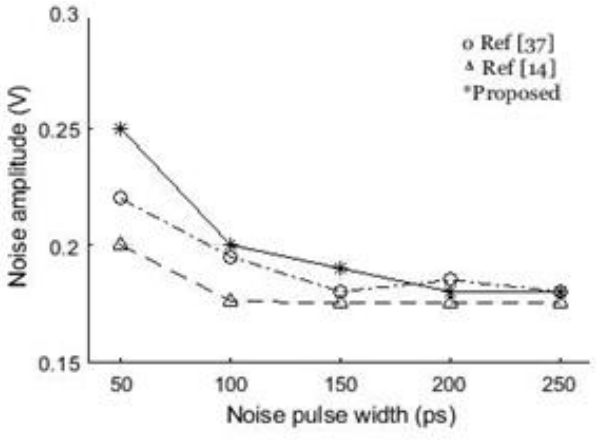


Fig. 10: NIC of single-digit adders.

To evaluate the improvement achieved in the proposed CNT-based multi-digit ternary adder, the main parameters including the power consumption, the propagation delay and the PDP are reported. In order to have a favorable comparison with previous works, two ripple carry multi-digit adder [6]-[12], two ripple carry adder which apply single-digit adder for multi-digit operations [5]-[10], and a conditional sum adder proposed by [14] are completely investigated by the same test patterns. Moreover, a lately published add/sub ternary circuit is investigated [48]. Table 8 reports the average power consumption of the several N-digit adders for N= {3,6,9,12}. According to the results, the proposed 3-digit, 6-digit, 9-digit and 12-digit adders improve the results by 10%, 25%, 27% and 19%, respectively, compared to the best design [37] in Table 8. In addition to the proposed power efficient GDI-CNT based circuits, the proposed power efficient decoder and the power efficient encoder [39], there is no encoder-decoder pairs in the carry propagation chain of the architecture [37]. There are the main reasons for the reduction of the power

consumption. A comparison of the multi-digit adders in term of the propagation delay are reported by Table 9.

Table 7: The simulation parameters of the CNTs [49]

Parameters	Descriptions	Value
Lch	Physical channel length	32nm
Lss	The length of doped CNT source-side extension region	32nm
Ldd	The length of doped CNT source-side extension region	32nm
Lgeff	The mean free path in the intrinsic CNT channel	100nm
Pitch	The distance between the centres of the two adjacent CNTs within the same device	20nm
Tox	The thickness of high-k top gate dielectric material	4nm
Csub	The coupling capacitance between the channel region and the substrate	40 uF/um
Efi	The Fermi level of the doped S/D tube	0.6 eV

Table 8: Power consumption of multi-digit adders

Power consumption (uW)				
(N)-digit adder	3	6	9	12
[12]	51.10	103.7	134.3	212.4
[6]	8.20	16.69	23.41	31.85
[5]	5.69	12.18	17.61	22.64
[10]	28.88	60.89	83.62	117.0
[14]	6.68	13.36	20.17	27.07
[37]	3.21	5.83	8.41	11.37
[48]	1.30	4.93	8.91	14.30
Proposed	2.89	4.32	6.11	9.21

Table 9: Propagation delay of multi-digit adders

Propagation delay (ps) _FO4				
(N)-digit	3	6	9	12
[12]	63.76	117.2	170.7	223.7
[6]	97.52	205.4	313.4	422.6
[5]	290.5	526.9	762.5	997.9
[10]	108.7	206.1	303.3	381.4
[14]	64.5	93.1	125.9	132.5
[37]	62.8	93.7	124.4	155.6
[48]	290.22	607.1	752.4	820.3
Proposed	53.40	74.21	101.3	122.7

In this case, the worst-case propagation delay is measured for all of the adders so that the signal changes propagate through the carry path and finally affect the $Cout_{N-1}$ and Sum_{N-1}. Unlike [5] and [6], [14] and [37] have lower complexity. Hence, an efficient carry generation/propagation results in lower propagation delay for each of them. The FO4 delay is calculated when four STI gates are connected at each output node of the critical path as in [37].

The simulation depicts that the proposed GDI based method improves the FO4-delay by 15% for N=3, 20% for N=6, 19% for N=6 and 8% for N=12 compared to the best design for each value of N. Although the overall structure of the proposed design is similar to [37], the proposed low-delay decoder and the low-complexity half adder has a great effect on reducing the overall propagation delay. The effect of the load capacitance on the propagation delay, is investigated. 1F, 2F and 3F capacitors are placed at the output node of the proposed multi-digit adder and two of the existing designs for N=6 and N=12. Fig. 11 shows the propagation delay against the load capacitance for 6-digit and 12-digit adders where the FO4- delays are also included.

The simulation reveals that there is lower dependency on the load capacitance for the both proposed 6-digit and 12-digit adders compared with the other designs. [14] has the worse result due to the limited driving capability. The product of the propagation delay and the power consumption (PDP) is the most widely used metric to validate the performance of a design. The PDP of the proposed multi-digit adder are compared with the other existing designs by Table 10. The proposed delay and power optimized decoder, half adder and carry/sum generator play an important role to reach the more efficient structure. So that the PDP is improved by 25% for N=3, 41% for N=6 and N=9 and 36% for N=12. Since the PDP is a well-known metric to evaluate the efficiency of a digital design, it is widely used in this work to compare the proposed architecture with the other existing designs. However, a circuit with low PDP (i.e. a very energy efficient design) may do the operations in a very slow manner. In this case, the energy-delay product (EDP) may be used as a much more preferable metric in some cases. To meet all aspects the EDP of the proposed and the other existing designs are reported by Table 11.

That is clear by the results that the proposed design is the most energy efficient design.

A CNTFET circuits consists of various diameter size CNTs. Due to that, it is so important to investigate the effect of the size variation. In this case Monte Carlo simulation is performed on the proposed single-digit adder and two other existing designs.

The simulation is performed with up to +-15% Gaussian distribution +- 3σ variations and 30 iterations for each

simulation. The simulation results are plotted in Fig. 12 where the power and delay variation are considered versus the diameter variation. That's clear that the proposed design depicts lower power variation compared with [14] and [37].

In term of the delay variation [14] gains the best performance.

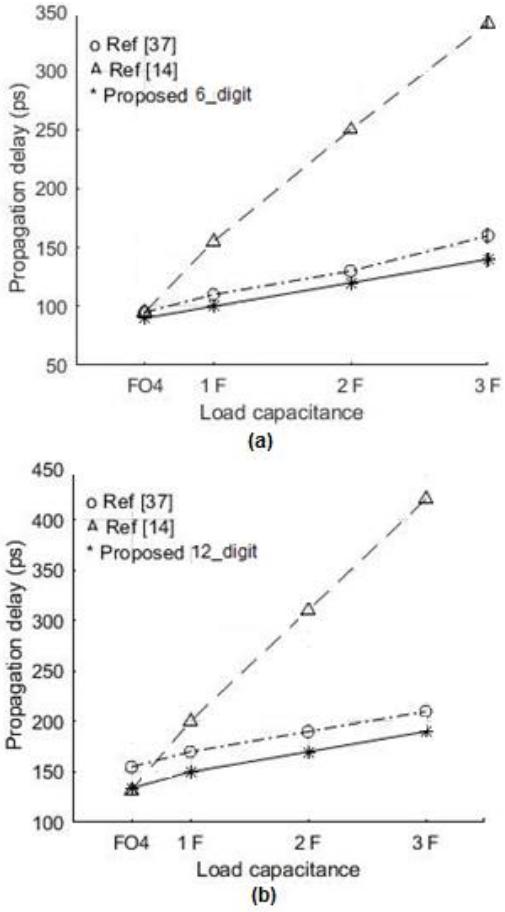


Fig. 11: Propagation delay vs load capacitance for 6-digit (a) and 12-digit (b) adders.

Table 10: PDP of multi-digit adders

Power-delay product (fJ)				
(N)-digit adder	3	6	9	12
[12]	3.26	12.2	22.9	47.5
[6]	0.80	3.43	7.34	13.4
[5]	1.66	6.42	13.4	22.6
[10]	3.14	12.6	25.4	44.6
[14]	0.43	1.24	2.54	3.59
[37]	0.20	0.55	1.05	1.77
[48]	0.38	2.99	6.70	11.73
Proposed	0.15	0.32	0.62	1.13

Table 11: EDP of multi-digit adders

Energy-delay product (x 10^-24 J.s)				
(N)-digit adder	3	6	9	12
[12]	0.21	1.43	3.9	10.6
[6]	0.08	0.7	2.3	5.6
[5]	0.48	6.42	3.4	22.5
[10]	0.34	2.6	7.7	17.01
[14]	0.03	0.12	0.32	0.48
[37]	0.012	0.051	0.13	0.27
[48]	0.11	1.1	5.04	9.6
Proposed	0.008	0.023	0.063	0.138

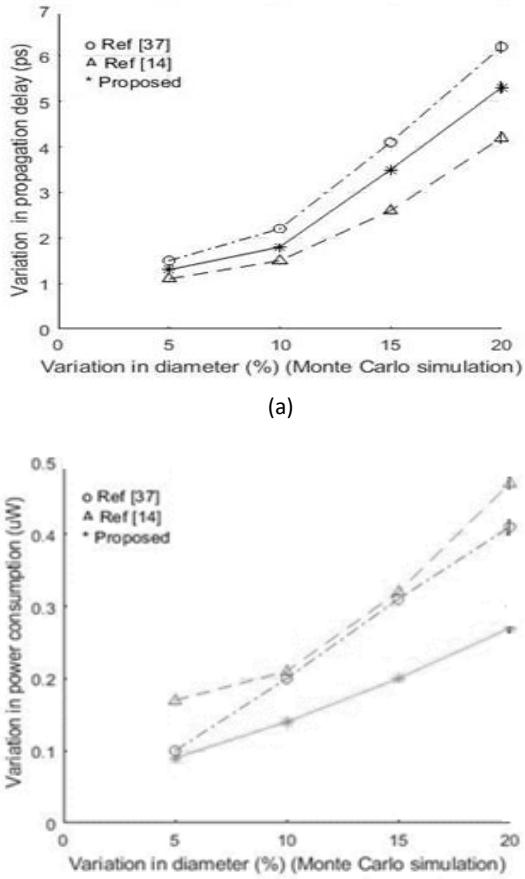


Fig. 12: Monte Carlo simulations for delay (a) and power (b).

Conclusion

A CNTFET based power and delay efficient multi-digit ternary adder has been presented in this paper. At each stage, a decoder converts the ternary inputs to the binary signals. The proposed structure includes a novel half adder to generate the intermediate binary signals. The binary signals have been used to calculate the final sum

and carry in the proposed sum/carry generator unit. The basic GDI cell has been widely used as an implementation method for the decoder, the half adder and the sum/carry generator designs.

The proposed design has been simulated in HSPICE with Stanford 32 nm CNTFET technology [49]. The simulation reveals a significant improvement in terms of power consumption (up to 27%), PDP (up to 41%) and FO4 delay (up to 20%).

Abbreviations

FA	Full Adder
MVL	Multi-Valued Logic
HS	Half Sum
HC	Half Carry
NIC	Noise Immunity Curve
GDI	Gate Diffusion Input

Author Contributions

The main idea of the paper is proposed by M.Dehyadegari and F.Razaghian and design and implementation and any of other simulation proposed by N.Ahmadzadeh Khosroshahi.

Acknowledgment

The authors would like to thank the editor and reviewers for their valuable comments on the manuscript.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

References

[1] M. Rezaei Khezeli, M. H. Moaiyeri, A. Jalali, "Analysis of crosstalk effects for multiwalled carbon nanotube bundle interconnects in ternary logic and comparison with cu interconnects," IEEE Trans. Nanotechnology, 16(1): 107-117, 2017

[2] A. Naeemi, R. Sarvari, J. D. Meindl, "On-Chip interconnect networks at the end of the roadmap: Limits and Nanotechnology opportunities," in Proc. 2006 International Interconnect Technology Conference, 2006.

[3] P. C. Balla, A. Antoniou, "Low power dissipation MOS ternary logic family," IEEE J. Solid-State Circuits, 19(5): 739-749, 1984.

[4] L. Yu-Ming, J. Appenzeller, J. Knoch, P. Avouris, "High-performance carbon nanotube field-effect transistor with tunable polarities," IEEE Trans. Nanotechnology, 4(5): 481-489, 2005.

[5] P. Keshavarzian, R. Sarikhani, "A Novel CNTFET-based ternary full adder," Circuits Syst. Signal Process., 33: 665–679, 2014.

[6] R. Faghih Mirzaee, K. Navi, N. Bagherzadeh, "High-Efficient circuits for ternary addition," VLSI Des., 2014: 534587, 2014.

[7] S. K. Sahoo, K. Dhoot, R. Sahoo, "High performance ternary multiplier using CNTFET," 2018 IEEE Computer Society Annual Symposium on VLSI (ISVLSI): 269, 2018.

[8] B. Srinivasu, K. Sridharan, "A synthesis methodology for ternary logic circuits in emerging device technologies," IEEE Trans. Circuits Syst I: Regul. Pap., 64(8): 2146-2159, 2017.

[9] G. Hills, C. Lau, A. Wright, et al. "Modern microprocessor built from complementary carbon nanotube transistors," Nature, 572: 595–602, 2019.

[10] S. L. Murotiya, A. Gupta, "Design of high speed ternary full adder and three-input XOR circuits using CNTFETs," in Proc. 2015 28th International Conference on VLSI Design, 2015.

[11] M. H. Moaiyeri, A. Doostaregan, K. Navi, "Design of energy efficient and robust ternary circuits for nanotechnology," IET Circuits Devices Syst., 5(4): 285-296, 2011.

[12] K. Sridharan, S. Gurindagunta, V. Pudi, "Efficient multiternary digit adder design in CNTFET technology," IEEE Trans. Nanotechnology, 12(3): 283-287, 2013.

[13] C. Vudadha, S. Katragadda, P. S. Phaneendra, "2:1 Multiplexer based design for ternary logic circuits," in Proc. 2013 IEEE Asia Pacific Conference on Postgraduate Research in Microelectronics and Electronics (PrimeAsia), 2013.

[14] B. Srinivasu, K. Sridharan, "Carbon nanotube FET-based low-delay andlow-power multi-digit adder designs," IET Circuits Devices Syst., 11(4): 352-364, 2017.

[15] S. L. Murotiya, A. Gupta, "A novel design of ternary full adder using CNTFETs," Arab J. Sci. Eng., 39: 7839–7846, 2014.

[16] C. Vudadha, S. Phaneendra Parlapalli, M. B. Srinivas, "Energy efficient design of CNFET-based multi-digit ternary adders," Microelectronics J., 75: 75-86, 2018.

[17] B. Srinivasu, K. Sridharan, "Low-Complexity multiternary digit multiplier design in CNTFET technology," IEEE Trans. Circuits Systems II: Express Briefs, 63(8): 753-757, 2016.

[18] C. Vudadha, M. B. Srinivas, "Design of high-speed and power-efficient ternary prefix adders using CNFETs," IEEE Trans. Nanotechnology, 17(4): 772-782, 2018.

[19] C. Vudadha, P. S. Phaneendra, G. Makkena, V. Sreehari, N. M. Muthukrishnan, M. B. Srinivas, "Design of CNFET based ternary comparator using grouping logic," in Proc. 2012 IEEE Faible Tension Faible Consommation, 2012.

[20] C. Vudadha, P. S. Phaneendra, M. B. Srinivas, "An efficient design methodology for CNFET Based ternary logic circuits," in Proc. 2016 IEEE International Symposium on Nanoelectronic and Information Systems (iNIS), 2016.

[21] N. Ahmadzadeh Khosroshahi, M. Dehyadegari, F. Razaghian, "A high-efficiency Wallace tree based multi-trit multiplier in CNTFET technology," Int. J. Electron., 1(19): 1244-1257, 2022.

[22] S. Lata Murotiya, A. Gupta, "Hardware-efficient low-power 2-bit ternary ALU design in CNTFET technology," Int. J. Electron., 103(5): 913-927, 2016.

[23] S. L. Murotiya, A. Gupta, S. Vasishth, "Novel design of ternary magnitude comparator using CNTFETs," in Proc. 2014 Annual IEEE India Conference (INDICON), 2014.

[24] G. Hills et al., "Understanding energy efficiency benefits of carbon nanotube field-effect transistors for digital VLSI," IEEE Trans. Nanotechnology, 17(6): 1259-1269, 2018.

[25] Y. Xie, Z. Zhang, D. Zhong et al., "Speeding up carbon nanotube integrated circuits through three-dimensional architecture," Nano Res., 12: 1810–1816, 2019.

[26] A. P. Dhande, V. T. Ingole, " Design and implementation of 2-bit ternary ALU slice," in Proc. Int. Conf. IEEE-Sci. Electron., Technol. Inf. Telecommun: 17-21, 2005.

[27] S. Lata Murotiya, A. Gupta, “Design of CNTFET-based 2-bit ternary ALU for nanoelectronics”, Int. J. Electron., 101(9): 1244-1257, 2014.

[28] S. Lata Murotiya, A. Gupta, “Hardware-efficient low-power 2-bit ternary ALU design in CNTFET technology”, Int. J. Electron., 103(5): 913- 927, 2016.

[29] T. Sharma, L. Kumre, “Energy-Efficient ternary arithmetic logic unit design in CNTFET technology,” Circuits Syst. Signal Process. 39: 3265–3288, 2020.

[30] A. Morgenshtein, A. Fish, A. Wagner, "Gate-diffusion input (GDI)-a novel power efficient method for digital circuits: a design methodology," in Proc. 14th Annual IEEE International ASIC/SOC Conference (IEEE Cat. No.01TH8558), 2001.

[31] M. Shaveisi, A. Rezaei, “Design and implementation of CNTFET-Based reversible combinational digital circuits using the GDI Technique for ultra-low power applications,” BioNanoSci. 10: 1063–1083, 2020.

[32] I. Sutherland, S. Fairbanks, "GasP: a minimal FIFO control," in Proc. Seventh International Symposium on Asynchronous Circuits and Systems. ASYNC 2001, 2001.

[33] R. O. Ozdag, P. A. Beerel, "High-speed QDI asynchronous pipelines," in Proc. Eighth International Symposium on Asynchronous Circuits and Systems, 2002.

[34] A. Morgenshtein, M. Moreinis, R. Ginosar, "Asynchronous gate-diffusion-input (GDI) circuits," IEEE Trans. Very Large Scale Integr. (VLSI) Syst., 12(8): 847-856, 2004.

[35] K. Sridharan, S. Gurindagunta, V. Pudi, "Efficient multiternary digit adder design in CNTFET technology," IEEE Trans. Nanotechnology, 12(3): 283-287, 2013.

[36] R. Faghih Mirzaee, K. Navi, N. Bagherzadeh, "High-efficient circuits for ternary addition," VLSI Design, 2014: 534587, 2014.

[37] C. Vudadha, S. Phaneendra Parlapalli, M. B. Srinivas, “Energy efficient design of CNFET-based multi-digit ternary adders,” Microelectronics J., 75: 75-86, 2018.

[38] S. Lin, Y. Kim, F. Lombardi, "CNTFET-Based design of ternary logic gates and arithmetic circuits," IEEE Trans. Nanotechnology, 10(2): 217-225, 2019.

[39] C. Vudadha, P. S. Phaneendra, M. B. Srinivas, "An efficient design methodology for CNFET based ternary logic circuits," in Proc. 2016 IEEE International Symposium on Nanoelectronic and Information Systems (iNIS), 2016.

[40] F. Sharifi, M. H. Moaiyeri, K. Navi, N. Bagherzadeh, “Robust and energy-efficient carbon nanotube FET-based MVL gates: A novel design approach”, Microelectronics J., .46(12): Part A, 2015.

[41] P. Soleimani Abhari, F. Razaghian, "A novel median based image impulse noise suppression system using spiking neurons on FPGA," Comput. Methods Biomech. Biomed. Eng.: Imaging Visualization, 8(6): 631-640, 2020.

[42] P. Soleimani Abhari, F. Razaghian, "Hardware implementation of LIF and HH spiking neuronal models," Signal Process. Renewable Energy, 3(1): 35-42, 2019.

[43] P. Soleimani ABHARI, M. Dosaranian Moghadam, "Design and Simulation of a modified 32-bit ROM-based direct digital frequency synthesizer on FPGA," AUT J. Electr. Eng., 47(1): 23-29, 2015.

[44] P. Soleimani Abhari, F. Razaghian, S. Talebi Toti, "Single spiking neuron as direct digital frequency synthesizer," Signal Process. Renewable Energy, 3(3): 73-81, 2019.

[45] P. Keshavarzian, R. Sarikhani, “A novel CNTFET-based ternary full adder”, Circ. Syst. Signal Process. 33 (3): 665-679, 2014.

[46] S. L. Murotiya, A. Gupta, “Design of high speed ternary full adder and three-input XOR circuits using CNTFETs, in Proc. 2015 28th International Conference on VLSI Design, 2015.

[47] R. F. Mirzaee, K. Navi, N. Bagherzadeh, “High-efficient circuits for ternary addition”, VLSI Des., 2014: 534587, 2014.

[48] S. Rani, Suman, B. Singh, "Cntfet based 4-trit hybrid ternary adder-subtractor for low power & high-speed applications," Silicon, 14: 689–702, 2022.

[49] J. Deng, H. S. P. Wong, “A compact SPICE model forcarbon-nanotube field-effect transistors including nonidealitiesand its application—Part II: Full device model and circuitperformance benchmarking,” IEEE Trans. Electron. Devices, 54(12): 3195–3205, 2007.

Biographies

Nader Ahmadzadeh Khosroshahi was born in Tabriz, Iran in 1985. He received his B.Sc. degree in Electronic Engineering in 2007 from Islamic Azad University, Urmia Branch, Urmia, Iran, M.Sc. degree in Electrical Engineering from Islamic Azad University, Tabriz Branch, Tabriz, Iran and he is now a. Ph.D. student in Electronic Engineering in Islamic Azad University, South Tehran Branch, Tehran, Iran. His current research interests are Nano-Electronic, Low-Power high-performance VLSI, Digital logic and circuit design.

- Email: st_n_ahmadzadeh@azad.ac.ir
- ORCID: 0000-0002-5291-2568
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Masoud Dehyadegari received his Ph.D. degree from University of Tehran, Tehran, IRAN, in 2013 in computer engineering. He is currently an Assistant Professor of school of computer engineering with the K. N. Toosi University of Technology, Tehran, IRAN. Form September 2011 until December 2012, he was a visiting scholar in University of Bologna, Italy. His research interests include Low-power system design, Network-on-chips, and Multi-Processor System-on-chip.

- Email: dehyadegari@kntu.ac.ir
- ORCID: 0000-0002-9473-5459
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Farhad Razaghian received the B.Sc. and M.Sc. degrees in Electronics Engineering from Islamic Azad university south Tehran branch, Tehran, Iran, and the Ph.D. degree in electronics engineering from the Islamic Azad university Science and Research branch, Tehran, Iran. His research interests are analog circuits, CMOS integrated circuits, power amplifiers and RFIC design as well as VLSI/FPGA implementation of algorithms.

- Email: razaghian@azad.ac.ir
- ORCID: 0000-0001-9005-4309
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:
N. Ahmadzadeh Khosroshahi, M. Dehyadegari, F. Razaghian, "An Ultra-low power ternary multi-digit adder applies GDI method for binary operations," J. Electr. Comput. Eng. Innovations, 11(1): 189-202, 2023.

DOI: 10.22061/JECEI.2022.9065.570

URL: https://jecei.sru.ac.ir/article_1772.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

An Approach for Evaluating Incentive Policy of Wind Resources Considering the Uncertainties in the Deregulated Power Market

M. Tolou Askari*

Department of Electrical and Electronic Engineering, Semnan Branch, Islamic Azad University, Semnan, Iran.

Article Info

Article History:
Received 24 June 2022
Reviewed 27 July 2022
Revised 15 August 2022
Accepted 26 September 2022

Keywords:
Restructured power market
Wind resources
Feed in tariff
Game theory

*Corresponding Author's Email Address:
m.askari@semnaniau.ac.ir

Abstract

Background and Objectives: Development of intermittent wind generation has necessitated the inclusion of several creative approaches in modeling the deregulated power market with presence of wind sources. The uncertain nature of wind resources will cause the private companies meet several risks in their medium and long term planning in a restructured power market. In addition to considering the uncertainties such as load, fuel costs, wind power generation and technical actions of rivals for modeling the restructured power market, the regulatory policies i.e. incentive policy for wind resources and Carbon tax should be assumed in this approach.

Methods: The first contribution of this article is to propose a developed mathematical model in order to evaluate the medium term deregulated power market by assuming the hybrid wind-thermal power plants. The second contribution is to evaluate the impact of different types of Feed in Tariff supports on Market Clearing Price, Expected Cost for Government, profits and contribution of each firm in electricity generation in the restructured power market. Also, the scenario based method has been used to generate the scenarios for wind uncertainties and then their reliability validate based on the statistical methods.

Results: The proposed mathematical model in the first contribution is calculated for each season and load levels based on the proposed wind scenarios. The functionality and feasibility of this model is demonstrated by simulation studies.

Conclusion: The proposed model in this article can be so useful for evaluating the different types of incentive policies for renewable energies. Moreover, this study confirms the previous researches that selected the Feed in Tariff as an efficient approach for developing the wind resources.

Introduction

Wind power plants, because of the various regulatory policies as well as their uncertainties, are among the risky investment in the deregulated power market. Thus, it is essential to provide a comprehensive model to investigate the effect of different types of incentive policies in the deregulated power market. In this model, all uncertainties in the electricity market, market regulatory policies and incentives of renewable power plant should be considered.

European Union (EU) countries have prepared several incentive policies to increase renewable resources [1]. Among renewable sources, the perfection rate of wind technology is higher than the others. The development has happened in consequence of several advantages of wind such as minimum environmental effects [2]-[5]. Furthermore , the wind has been developed very quickly

rather than the other renewable technologies because of the low R&D expenditure [6]. In the late 1990s, around 70% of wind generators had been installed in Europe, 19% in North America and just 9% in Asia [7]. Besides the expansion of wind technologies, the structure of the power market has reformed from the centralized to the decentralized power market [8], [9]. In this article, competitive power market has been mentioned as restructured power market. However, these changes have a serious effect on the goals of the various power markets. The main purpose of the players in restructured markets are not the same as the objectives explained by the government in regulated markets. The main aims of planning in the regulated power market is to respond the demand through the acceptable reliability [10], [11]. While, maximize the profit is the goal of the investors in the deregulated power market [12]-[15]. In the restructured power market the decision makers of wind resources are encounter with several challenges. These challenges related to the volatility of wind speed, uncertainties in the restructured power market, regulatory policies (such as incentive policies, CO2 taxation, etc.) and high capital investment of this technology. These challenges are barriers against the development of wind resources [16]. Moreover, the wind power plants could not compete with other conventional power plants in the restructured power market because of the intermittent nature of wind. Therefore, incentive policies need to expand the wind power plants [3], [17]. There are four main policies for developing the renewable energies which are including the auctions and fiscal incentives, tax credits, quotas and tradable green certificates and Feed-in tariffs (FIT). However, the FIT is the most effective incentive to develop the renewable resources [15], [21]. Feed-in tariffs are incentive mechanism suggested to speed up the investment in renewable energy by providing them reward (a "tariff") above the retail or wholesale rates of electricity. Spain is one of the first countries that have developed specific incentive mechanisms for implementing the renewable energies. They established the first FIT in 1994 through the fixed FIT. Then, Spain encouraged the renewable firms to sell their generation through the wholesale market and received premium in a restructured power market. Also, distribution utilities had been obliged to buy the whole renewable generation. Germany has supported the renewable energies by developing the technologies in order to decrease the generation cost of renewable energies. In addition to Research and development subsidies, the Feed in supports motivated the development of renewable in Germany [23], [24]. The other countries such as United States, Canada and Denmark are developing the renewable sections by considering the different types of supportive policies. Although FIT supports are effective policy to develop the renewable energies in developed countries, it exposes excessive cost to the government and costumers.

In addition, the wind resources' investors encounter more uncertainties rather than the conventional companies. These stochastic uncertainties are including the wind velocity, electricity demand and fuel price. Also, these stochastic parameters as well as the technical actions of rivals fluctuate the market clearing price. Therefore, the investors in the deregulated power market and governments should equip themselves with powerful and flexible decision making tools in order to investigate the effect of these uncertainties as well as the effect of various types of FIT on their, market clearing price (MCP), profits and contribution of each firm in electricity generation in the restructured power market. This decision tools should capable to model the uncertainties (such as demand, fuel price, wind and rational uncertainties) and the CO2 tax and bilateral contracts in addition of the FIT. Due to these uncertainties, modeling and planning in the restructured power market has encountered more risks. Different methods are available to evaluate the uncertainties. Usually, these methods are based on the probability and statistical methods. Accordingly, the decision problems such as planning and risk management should be solved by considering the stochastic uncertainties such as load and fuel price uncertainties and rational uncertainties that is the operational strategic behaviour of participants in the restructured environment. Moreover, the realities and the regulator policies should be considered in medium term planning.

A mathematical model has been suggested in [25] for evaluating the effect of fix FIT. In this model, uncertainties and also the realities of the deregulated power market has been neglected. In [26], a review paper has been given for assessing the policy in power market. It has been illustrated that policy assessment and Generation expansion planning are the most important issues [26]. furthermore, an approach has been given in [27]. Although this model provide useful data about the incentive mechanism for renewable sources in the deregulated power market, the CO2 tax has been dismissed [27].

In this article, a developed mathematical model is proposed to investigate the impact of FIT on the profit, MCP, Expected Cost for Government and contribution of each firm in the restructured power market. The main contribution of this paper is to propose a mathematical model in order to investigate the effect of different types of FIT by considering the hybrid wind-thermal power plants, rational and stochastic uncertainties and realities such as bilateral contracts and carbon tax for thermal units in a restructured power market. Also, the

uncertainty of output power of wind power plants has been considered in this study based on the scenario based method. The reliability of these scenarios validate based on the statistical methods. The developed mathematical model is examined with fixed FIT, variable FIT for different levels of demand and without FIT.

The rest of this paper is structured as follows: Section 2 illustrated the description of the proposed structure and describes the modeling of wind generation. Section 3 demonstrate the proposed mathematical model to simulate the deregulated power market. Section 4 implements the developed model on a power market and the results are discussed in Section 5. Finally, the last section is focused to the conclusion.

Description of the Proposed Structure

The flowchart of the proposed model is revealed in Fig. 1. It presents four main blocks, which are explained in the next paragraphs. In this study, a new method has been proposed to generate the wind scenarios according to the scenario-based method. The wind scenarios together with their probabilities have been generated based on the data mining algorithm. Then, their results are validated with the statistical method. The proposed method has been applied for a standard system in order to reveal the effectiveness of this approach. This section is presented in block 1. In the second block, a model has been developed to evaluate and analyze the medium term restructured power market with the presence of the hybrid wind-thermal firm. In this model, electricity demand and fuel costs as short term uncertainties are simulated through Monte-Carlo method annually. In addition, the wind generation's scenarios with probabilities of their occurrence have been calculated through the outcomes of block 1. These uncertainties are called the stochastic uncertainties which are considered in the proposed model to simulate the medium term power market. Besides the stochastic uncertainties, the strategic behavior of investors, as an effective parameter on the MCP, is analyzed based on the concept of Cournot game theory. This concept has been used in order to determine the MCP in the restructured power market. Also, regulatory interventions such as the CO2 tax, FIT, and bilateral contracts are assumed in this model as the exogenous parameters.

Then, the proposed model has been examined with various types of FIT which are represented in the third block. Finally, the output results of this model are given for each type of FIT. These results are represented in the fourth block. The details of these blocks are described in the next sections.

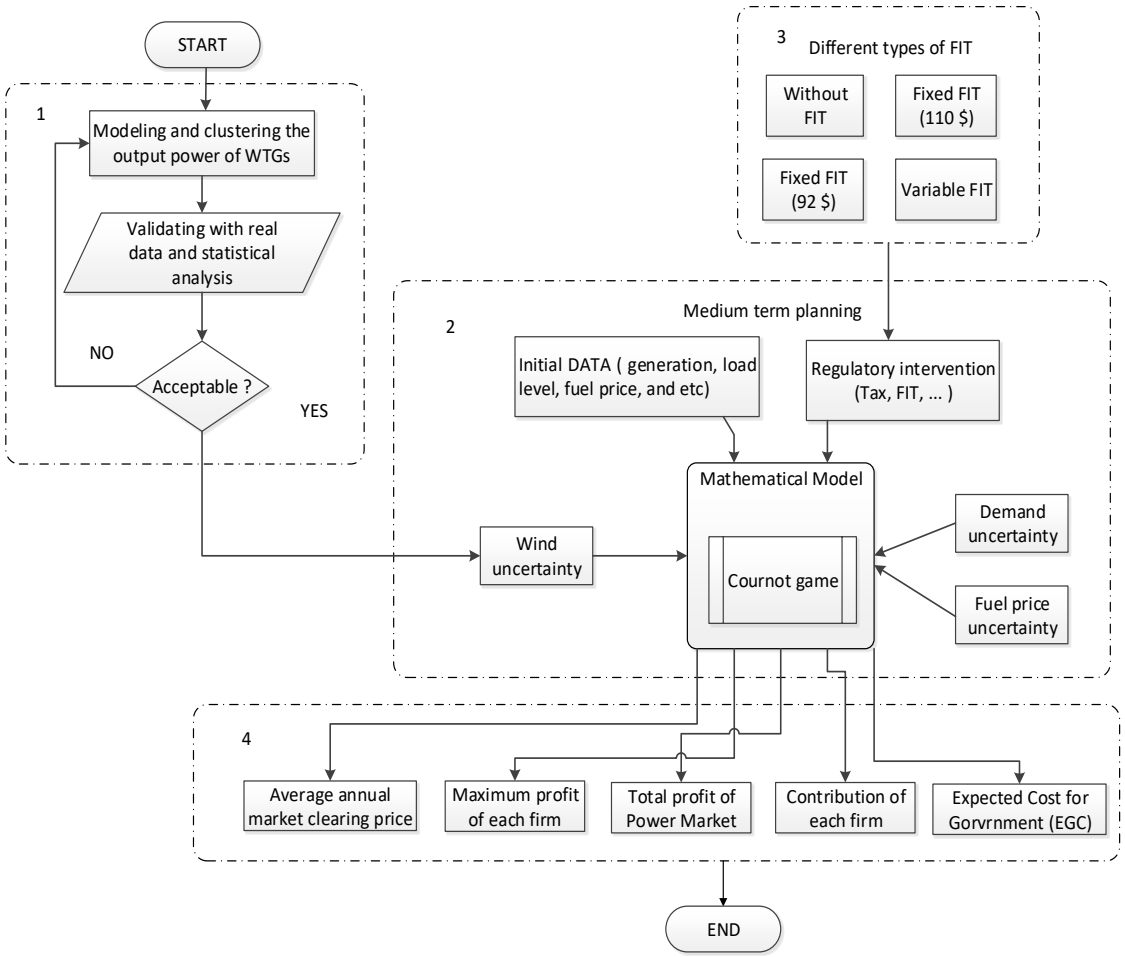


Fig. 1: Flowchart of the proposed structure.

Scenario Reduction Algorithm for Output Power of Wind Power Plants

Although, various methods have been implemented to decrease the generated scenarios for the velocity of wind, the precision of them were not justified through a scientific method [15], [28], [29]. In this article, the proposed method has been validated and justified based on the real data and Weibull probability plot test. In this article, the generated scenarios for wind power plants are selected based on data mining. The basis of this method is to classify the real wind speed data that have certain characteristics in a specific group. Data mining have various methods that in this article the K-means method is used to generate wind scenarios. The main purpose of scenario reduction algorithm is to downsize the data set whereas holding the stochastic data as intact as possible [30], [31]. In this paper, the clustering method is given to generate scenarios for the velocity of wind. In the proposed method, each year is divided into four seasons and then the acceptable number of scenarios determined for each season. Several scenarios can be generated for wind speed data in each season, but choosing the optimal number of scenarios can greatly help to increase the speed of the program. For this reason, a method should be proposed to select the appropriate number of scenarios in each season. In this article, after the wind scenarios have been generated, the parameters of the Weiball distribution function (scale and shape) are determined for these scenarios. Then, the scale and shape parameters compared for both Weibull distribution functions before and after the scenario generation. These steps are tested for the cases with different number of scenarios and reconciled with the Weibull parameters of real wind speed data. Finally, scenarios that are closer to the distribution function of real wind speed data in terms of scale and shape have been used for the next stage of this research. However, it is possible that the values of shape and scale parameters for each generated scenarios are very close to each other. Therefore, in order to avoid inaccurate selection, the authors of this article suggest a statistical method in addition to the previous method. In other words, in order to validate and justified the generated scenarios, an Anderson-Darling statistic test is applied according to Fig. 3. This test was done using Minitab software. Accordingly, if the scatter points are located between two references lines it means that the data set conform the Weibull distribution. Furthermore, if the P-value is higher than the significance level, for instance 0.05, then the Weibull probability plot test is meaningful and the data fit a Weibull distribution. Also, less AD values demonstrate a better fit. In this article, the Weibull distribution functions for real wind data and wind scenarios data for each season together with the Weibull probability test have been applied to validate the number of scenarios which is selected for each season. This is because the Weibull distribution gives the best fit for the wind speed data that have been used in this study [4], [32].

The k-means clustering algorithm which is proposed for this approach is as follows:

Step 1: To calculate the number of clusters based on the mean and standard deviation.

Step 2: To initialize the centroids of each cluster

Step 3: To optimize the following objective function:

$$\min_{\{C_1,C_2,\ldots,C_j\}} \sum_{j=1}^{K}\sum_{i=1}^{n}\left\| x_i^{(j)} - C_j^{\,2} \right\| \quad (1)$$

The k-means method is shown in Fig. 2.

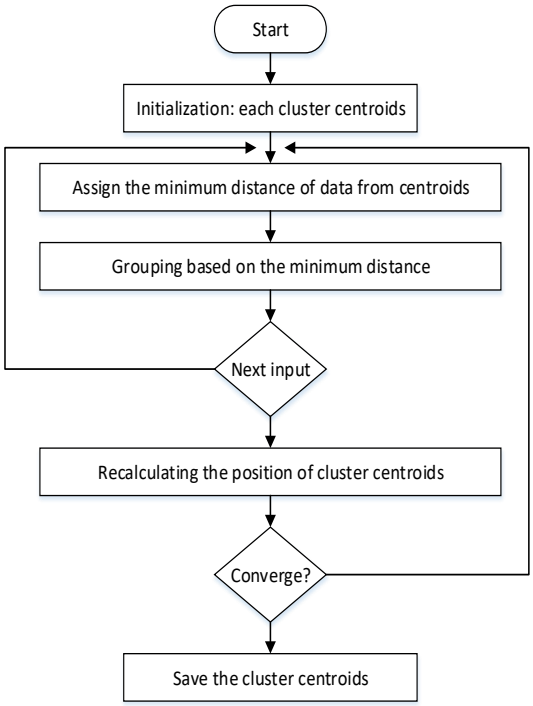


Fig. 2: Flow chart of K-means method.

The wind speed data were gathered from Swift Current in the Saskatchewan state-Canada [33]. Finally, the proposed scenarios for all seasons are shown in Table 1.

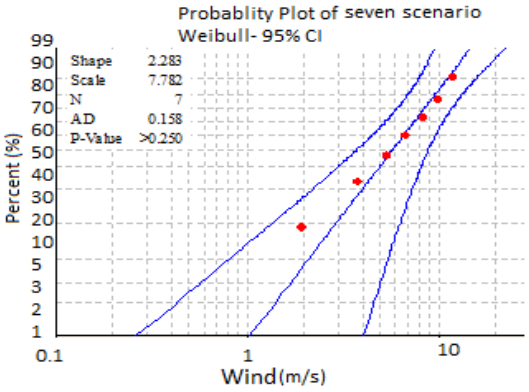


Fig. 3: Sample of Weibull probability plot test for first season.

Table 1: seasonal generated scenarios for wind farm

	Proposed Scenarios for electricity generation							Weibull parameters for each scenario		Weibull parameters from simulated data		Statistical results	
								Shape	Scale	Shape	Scale	P-value	AD
S1	Output [MW]	0	3.31	10.02	16.63	24.66	34.93	2.184	7.251	2.1522	6.5248	>0.25	0.149
	Prob [%]	14	23	24	19	14	6						
S2	Output [MW]	0	1.41	5.92	12.42	26.05		2.291	6.361	2.1695	5.1198	>0.25	0.166
	Prob [%]	24	22	17	26	11							
S3	Output [MW]	0	5.13	11.71	19.61	27.17	38.24	2.279	7.671	2.5788	6.9993	>0.25	0.138
	Prob [%]	20	29	30	11	8	2						
S4	Output [MW]	0	2.73	11.48	28.31			2.432	8.44	2.8548	8.0269	>0.25	0.195
	Prob [%]	20	23	35	22								

Mathematical Formulation for Deregulated Power Market

The main goal of the proposed objective function is to maximize the profit of each players in the deregulated power market. The objective function is presented in (2) to (4). Also, the constraints are shown in (5) to (9).

$$
\begin{aligned}
Max\ Benefit = & \sum_{e}^{Ne}\sum_{s}^{Ns}\sum_{l}^{Nl} d_{sl}\times\left(\left(\sum_{u}^{Nu} N_{e,u}\times P_{e,u,sl}\right)-Q_{e,sl}\right)\times SP_{sl} \\
& + \sum_{e'}^{Ne'}\sum_{s}^{Ns}\sum_{l}^{Nl} d_{sl}\times\left(\left(\sum_{u'}^{Nu'} N_{e',u'}\times P_{e',u',sl}\right)-Q'_{e',sl}\right)\times SP_{sl} \\
& + \sum_{e}^{Ne}\sum_{s}^{Ns}\sum_{l}^{Nl} d_{sl}\times Q_{e,sl}\times BP_{sl} \\
& + \sum_{e'}^{Ne'}\sum_{s}^{Ns}\sum_{l}^{Nl} d_{sl}\times Q'_{e',sl}\times BP_{sl} \\
& + \sum_{e'}^{Ne'}\sum_{s}^{Ns}\sum_{l}^{Nl} d_{sl}\times\left(\sum_{uw}^{Nuw} N_{e',uw}\times PW_{e',uw,sl,n}\right)\times(SP_{sl}+inc)'' \\
& - \sum_{e}^{Ne}\sum_{s}^{Ns}\sum_{l}^{Nl} d_{sl}\times\left(\sum_{u}^{Nu} N_{e,u}\times FP_{u}\times\left(a_{u}+b_{u}\times P_{e,u,sl}+c_{u}\times(P_{e,u,sl})^{2}\right)\right) \\
& - \sum_{e}^{Ne}\sum_{s}^{Ns}\sum_{l}^{Nl} d_{sl}\times\left(\sum_{u}^{Nu} N_{e,u}\times Tax\times\left(a_{u}+b_{u}\times P_{e,u,sl}+c_{u}\times(P_{e,u,sl})^{2}\right)\times EM_{u}\right) \\
& - \sum_{e'}^{Ne'}\sum_{s}^{Ns}\sum_{l}^{Nl} d_{sl}\times\left(\sum_{u}^{Nu} N_{e',u'}\times FP_{u'}\times\left(a_{u'}+b_{u'}\times P_{e',u',sl}+c_{u'}\times(P_{e',u',sl})^{2}\right)\right) \\
& - \sum_{e'}^{Ne'}\sum_{s}^{Ns}\sum_{l}^{Nl} d_{sl}\times\left(\sum_{u'}^{Nu'} N_{e',u'}\times Tax\times\left(a_{u'}+b_{u'}\times P_{e',u',sl}+c_{u'}\times(P_{e',u',sl})^{2}\right)\times EM_{u'}\right)
\end{aligned}
\tag{2}
$$

$$g_{e,sl} = \sum_{u}^{Nu} N_{e,u} \times P_{e,u,sl} \quad (3)$$

$$g_{e',sl} = \sum_{uw}^{Nuw} N_{e',uw} \times PW_{e',uw,sl} + \sum_{u'}^{Nu'} N_{e',u'} \times P_{e',u',sl} \quad (4)$$

Subject to:

$$BP_{sl} = SP_{sl,t-1}\left(1+GP\right) \quad (5)$$

$$g_{e,e',sl} \le D_{sl} \quad (6)$$

$$P_{e,u.min} \le P_{e,u} \le P_{e,u,max} \quad (7)$$

$$P_{e',u'min} \le P_{e',u'} \le P_{e',u'max} \quad (8)$$

$$PW_{e',uw,nmin} \le PW_{e',uw,n} \le PW_{e',uw,nmax} \quad (9)$$

The main objective function which is represented in (2) is made up of two main sections. The revenue of renewable and traditional power plants as well as the income for each firm in contractual markets is shown in the first section. Also, the incentive policy for wind firms is given in first section. Then the costs such as the CO2 tax and costs for heat power plants are shown in the second part of the proposed model. The amounts of generations of traditional and hybrid traditional-renewable private firms are shown in (3), and (4), respectively.

Electricity price for bilateral contract in each season and each load level is represented in (5). The demand constraints is shown in (6), because the private firms are not responsible to response the all request of the power market. Upper and lower limit of the generation capacity of each firms and units are represented in (7), and (8). Furthermore, the generation restriction in renewable units of hybrid firm is shown in (9).

The outputs of the proposed model are including the whole energy generated in market, electricity produced by each company, profit of power market, and profit of each company. For this situation, investors in the power market with lower expenses as opposed to different firms will amplify their benefits by incrementing their generations. On the other hand, different individuals with excessive generating costs decide on not to take part in this type of power market or they will take part with minimum generation. Therefore, the equilibrium price and equilibrium quantity will no longer be provided.

In the balance condition the amount of electricity load is equivalent with the amount of supply, which is the main characteristic to pursue choice in the deregulated market. In addition, the investors in the deregulated market has less data about the decision of rivals and they need a consistent model to make their operational decisions. Game theory is "the study of mathematical models of conflict and cooperation between intelligent rational decision-makers" [34].

There are three famous games to evaluate the competitive power market which are including Cournot, Bertrand, and van Stackelberg. In the first game theory model, each company picks a result amount to maximize benefit. Firms are accepted to deliver homogeneous merchandise that are non-storable. The equilibrium price is calculated based on an auction process. The mannequin additionally assumes that all corporations in the enterprise can be recognized at the beginning of the game, and decision-making by way of companies happens simultaneously [35]-[37].

Subsequently, on account of numerous likenesses between Cournot game model and the deregulated power market, the idea of Cournot game has been implemented in this paper to decide the Market Clearing Price (MCP). The concepts which is used in this paper in order to calculate the MCP by considering the uncertainties of demand and fuel costs are shown in Fig. 4 and Fig. 5.

In order to clarify the issue, the steps of the proposed algorithm are explained in 8 steps as follows:

Step 1: Calculate the power generation of each company using the proposed objective function based on the initial electricity price.

Step 2: Update the electricity price using the following linear demand function.

$$D_{sl}\left(SP_{sl}\right) = -A_{sl} \times SP_{sl} + B_{sl} \quad (10)$$

where D_{sl} is the total generation of power market in each season and load levels. Also, A and B are constant value which are determined based on (11) and (12) respectively.

$$A_{sl} = \frac{B_{sl}}{pc.\pi_{base,sl}} \quad (11)$$

$$B_{sl} = dc.D_{base,sl} \quad (12)$$

Step 3: The calculated price is compared with the initial price. If these 2 values are equal, the program is saved and the results are shown. Otherwise, steps 1 and 2 are repeated until the Nash equilibrium is reached. These three steps show in internal loop of Fig. 4.

Step 4: The above three steps are implemented for all the wind scenarios generated in the previous section of this article. This step represents in external loop of Fig. 5.

Step 5: Simulating the demand and fuel costs for each firm based on the Monte Carlo technique. In this paper, normal distribution function is considered to generate random data for these uncertainties. The proposed algorithm shows in Fig. 5.

Step 6: Save the output results which are including the average market clearing price (AMCP), annual average market clearing price (AAMCP), expected cost for government (ECG), average annual profit (AAP) for each firm and total profit of power market.

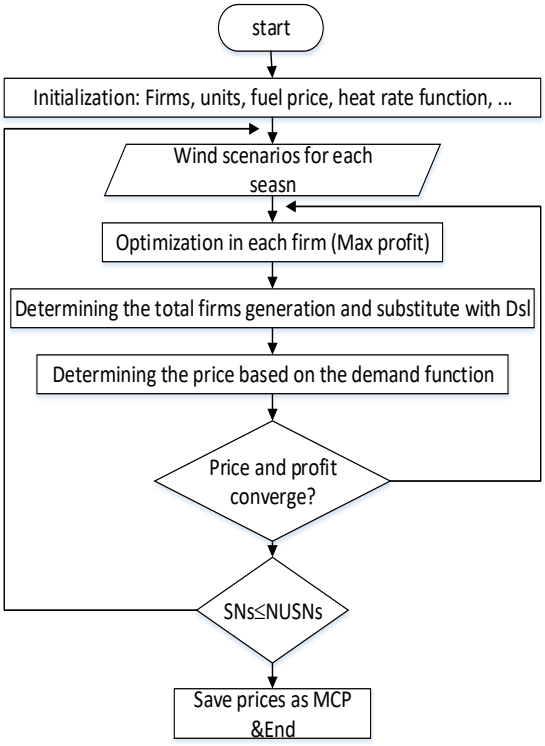


Fig. 4: Proposed algorithm to determine the MCP.

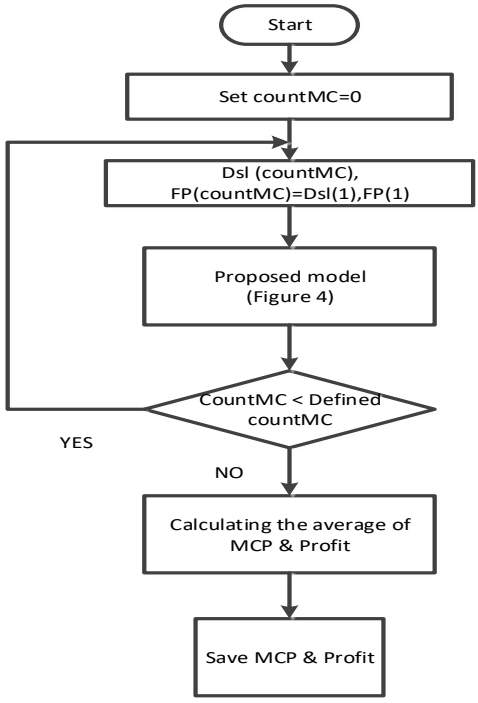


Fig. 5: Monte-Carlo for simulating the uncertainties (Demand and fuel costs).

Case Study

The proposed approach in this study is tested by applying IEEE RTS1 [38]. The whole installed capacity and the peak load of the study system are 2850 and 2500 MW, respectively. Characteristics regarding the firms in this system are revealed in Table 2. The study system includes five price maker investment companies.

The fuel costs and the information related with the producing units of these investment companies are taken from sources [39], [40] and are shown in Table 3. It is considered one year for the period of this study. Four windy seasons were considered for each year. Also, three load levels (off-peak, medium and peak levels) were assumed in this study for each season. The results of simulating wind scenarios are shown in Table 1.

The load duration data have been gathered from [41] and illustrated in Table 4. The amounts of CO2 emission for candidate units are illustrated in Table 5. The percentages of FIT for each load level are illustrated in Table 6.

Table 2: specifications of units

	Type of units			
	Oil/steam	Coal/steam	Wind	Nuclear
Firm 1	2	2	-	-
Firm 2	3	3	-	1
Firm 3	4	4	-	-
Firm 4	1	2	-	-
Firm 5	1	2	6	-

Table 3: Fuel cost and power limitation for each unit

Types of unit	Oil/steam	Coal/steam	Wind	Nuclear
Fuel[$/MBTU]	5.27	1.68	-	1.65
Max P [MW]	12-197	76-350	50	400
Min P [MW]	2.4-68.95	15.2-140	0	100

Table 4: Load duration data for each season and each load level

Season	Load level and duration [MW]&[hrs]		
	Off-peak (Duration)	Medium (Duration)	Peak (Duration)
1	950(876)	1800(985.5)	2300(328.5)
2	1200(876)	1650(985.5)	2370(328.5)
3	1300(766.5)	1900(876)	2500(547.5)
4	1000(876)	1550(985.5)	2250(328.5)

[1] Reliability test system

Table 5: CO2 emission [lb/MMBTU]

	Oil/steam	Coal/steam	Wind	Nuclear
CO2 EM	170	210	0	0

Table 6: Percentages of FIT for each load level

	Base (%)	Medium (%)	Peak (%)
SN1	50	60	70
SN2	70	80	90
SN3	110	120	130
SN4	120	130	140

Results and Discussions

In this section, the findings of this research are examined. These results will be analyzed in three parts. The first part related to the generated scenarios for wind power plants.

Then the robustness of the proposed mathematical model is examined in the second part based on the sensitivity analysis technique. Finally, the proposed model is investigated for different types of FIT in the last part.

A. Wind Scenario Gereration

In this section, the results of the Weibull distribution functions of wind speed data as well as the clustering wind data for each season are shown in Fig. 5 and Fig. 12.

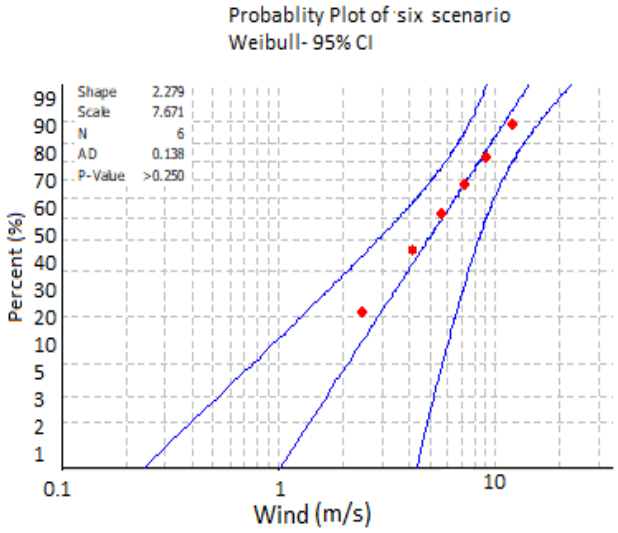


Fig. 5: Weibull probability plot test – first season.

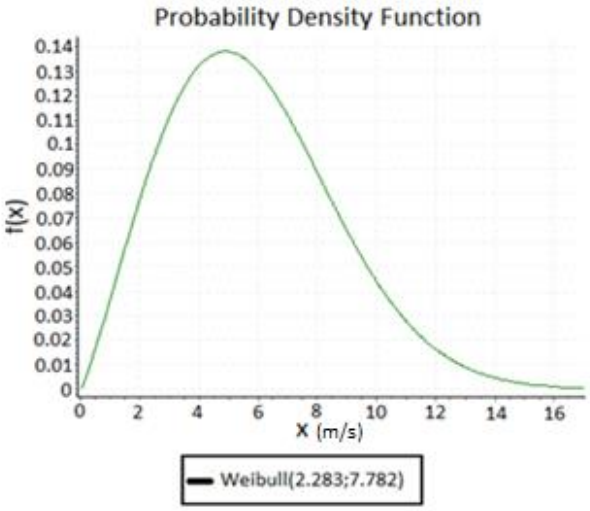


Fig. 6: Weibull distribution function - first season.

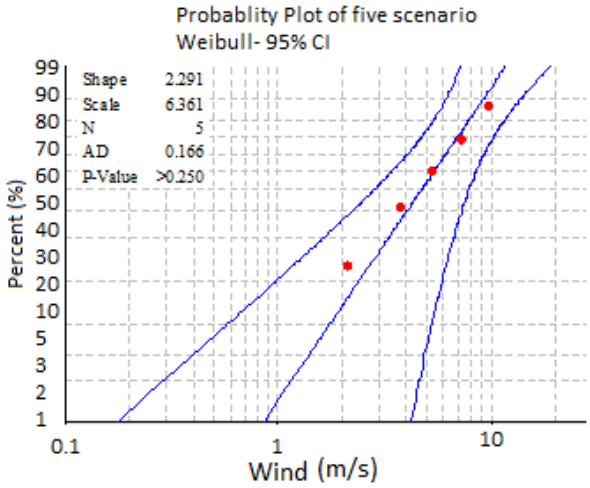


Fig. 7: Weibull probability plot test - second season.

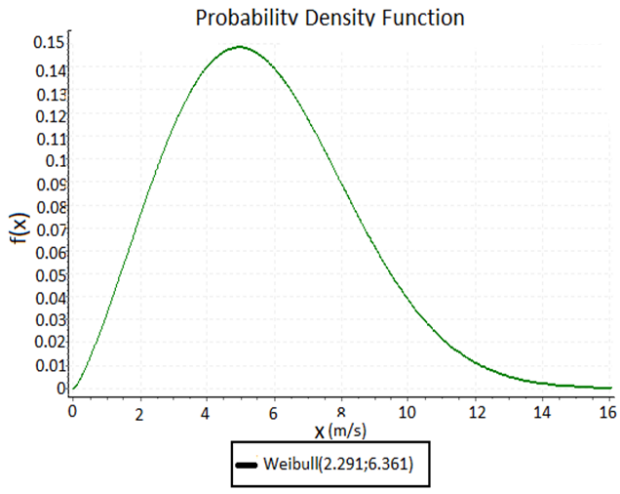


Fig. 8: Weibull distribution function for second season.

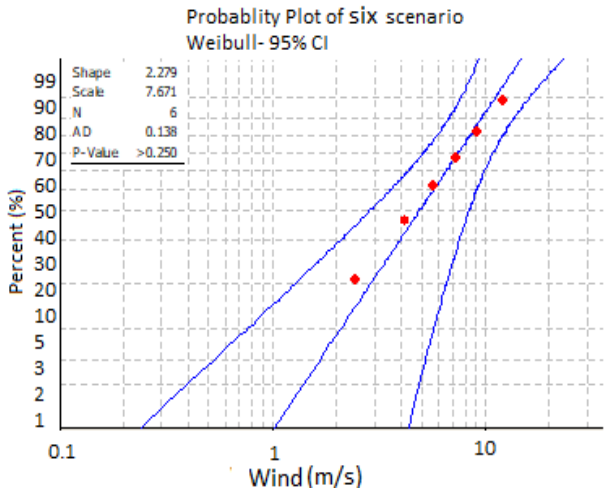


Fig. 9: Weibull probability plot test for third season.

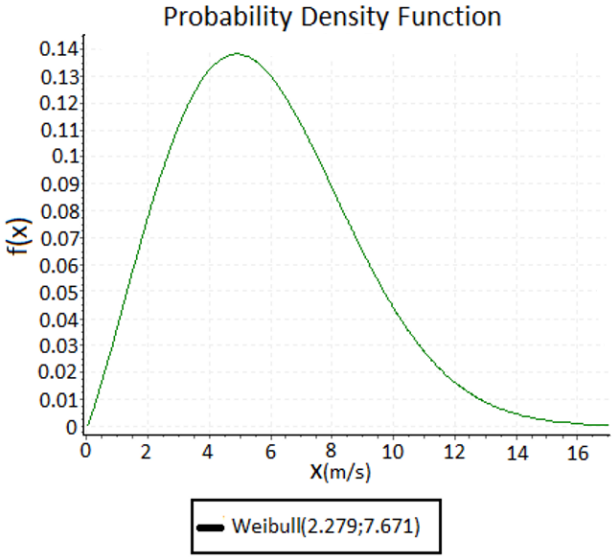


Fig. 10: Weibull distribution function for third season.

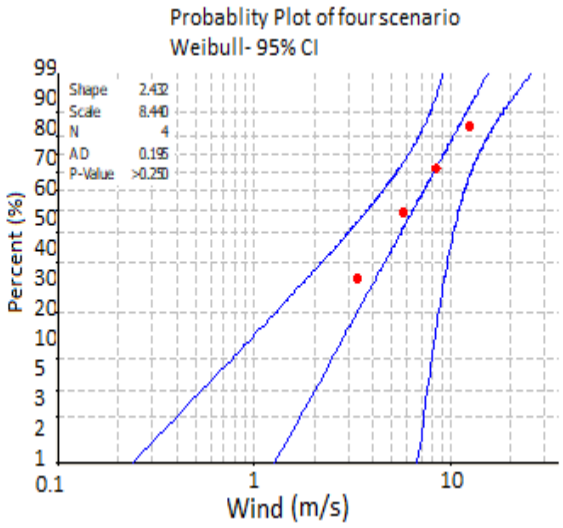


Fig. 11: Weibull probability plot test for forth season.

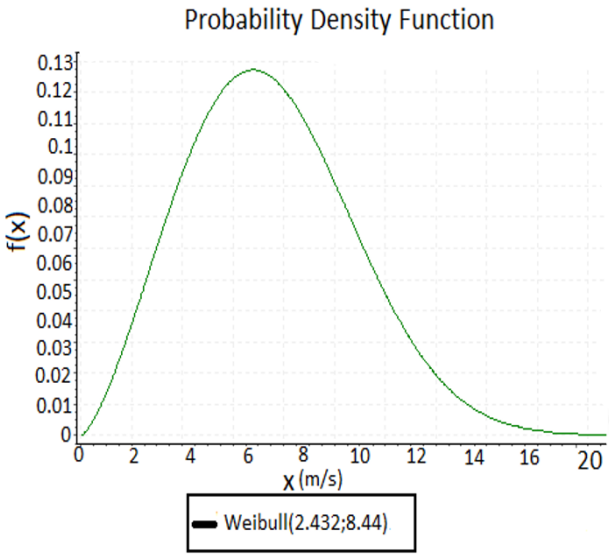


Fig. 12: Weibull distribution function for forth season.

B. *Sensitivity Analysis of Proposed Model*

The main purpose of this section is to validate and verify the proposed model. In fact, sensitivity analysis is used to examine the relationship between a specific dependent variable and independent variables. In this article, the effect of fuel price changes on the profit of each company and also the whole market is investigated. The results are given in Table 7. Accordingly, with the increase in the fuel price, the total profit of the power market decreases. These results show the correctness of the proposed model in which the profit of firms decrease as their costs increase.

Table 7: Sensitivity analysis of proposed model

Growth of fuel price	Profit [M$]					
	Firm1	Firm2	Firm3	Firm4	Firm5	Total
2%	25.56	78.99	51.11	62.95	29.66	248.3
4%	24.55	75.29	49.1	62.06	27.59	238.6
6%	24.18	73.96	48.36	61.33	25.35	233.2

C. *Investigating the Different Types of FIT*

The developed model is tested for four types of FIT and their results are illustrated in Table 7 and Table 8. Each case study is described as below:

Case study No. 1: In this case study, the FIT is not considered for wind resources.

Case study No. 2: The proposed model for simulating the restructured power market is solved with fixed FIT. Two different scenarios have been considered for this case study. These scenarios are including the FIT lower than the average fixed FITs in European countries (92

$/MWh), and FIT higher than the average fixed FITs in European countries (110 $/MWh).

Case study No. 3: In this case study, the FIT changes for different load levels. But these variations are considered fixed for each load level. Three different scenarios are imagined for solving the proposed model. In the first scenario, 80% of the MCP for each load levels (Low, Medium, and High) is paid to wind resources as FIT. In the second and third scenario 100% and 120% of the MCP for each load levels is considered for wind resources as FIT, respectively.

Case study No. 4: The MCP is increased by growing the electricity demand. Accordingly, the more FIT can be proposed in peak load level and it reduces to lower level. In case study No. 4, the proposed model solves for four different types of scenarios. The percentages of FIT for each load level are illustrated in Table 6. In scenario No.1, For instance, 50%, 60% and 70% of the MCP will be paid to wind resources as FIT for base, medium, and peak load levels, respectively. The average market clearing price (AMCP), annual average market clearing price (AAMCP) and expected cost for government (ECG) are illustrated in Table 8. ECG is mentioned as a factor in order to determine the cost that the government should pay as incentive. It calculates according to (13). According to Table 7, the AAMCP increases by growing the rate of fixed FIT from zero to 110 $/MWh. Also, The ECG increases through these variations.

$$ECG_{e'} = \sum_{s}^{Nn} \sum_{l}^{Nl} (g_{e',sl} \times inc) - (g_{e',sl} \times MCP_{sl}) \quad (13)$$

Table 8: Simulation results for investigating the effect of type of FIT on MCP and ECG

Cases	Types of FIT	AMCP [$/MWh]			AAMCP [$/MWh]	ECG[$]
		L	M	H		
No. 1	W/out FIT	22	42.924	90.748	51.891	0
No. 2	Fixed FIT (92)	22.028	42.96	90.743	51.91	93141.95
	Fixed FIT (110)	22.033	42.972	90.748	51.918	135734.6
No. 3	F/V SN1	22.02	42.915	90.547	51.827	99307.1
	F/V SN2	22.02	42.975	90.501	51.832	122898.8
	F/V SN3	22.036	42.913	90.788	51.912	148283.9
No. 4	V/V SN1	22.028	42.979	90.793	51.933	79625.51
	V/V SN2	22.013	42.982	91.074	52.023	105031.9
	V/V SN3	22.019	42.954	90.847	51.94	155961.7
	V/V SN4	22.008	42.99	90.481	51.826	167411.8

The results for investigating the effects of different case studies on the average annual profit (AAP) for each firm and total profit of power market have been represented in Table 9.

Accordingly, the firm No. 5 as a hybrid wind-thermal firm could not compete with conventional firms without FIT. Also, it will gain the maximum profit in case study No. 2. Since the demand and fuel price are considered as uncertain parameters in the proposed model, the standard deviation (SD) of each case study is shown in the following table.

In order to complete the discussion, the share of each private company in the market are compared for different case studies. These results are represented in Fig. 6.

This figure represents that the share of the generation of firm 5 as hybrid wind-thermal firm increases by considering the fixed FIT (110 $) rather than others. This analysis is referred to evaluate the incentive policy on the proposed model to simulate the deregulated power market.

Table 9: Simulation results for investigating the effect of type of FIT on AAP and total profit of power market

Cases	Types of FIT	AAP for each firm [M$]					Total profit [M$]	SD
		firm 1	firm 2	firm 3	firm 4	firm 5		
No. 1	W/out FIT	26.01	82.61	52.02	65.49	-4.36	221.79	3.94
No. 2	Fixed FIT (92)	26.43	83.45	52.87	64.40	22.56	249.72	5.50
	Fixed FIT (110)	26.67	82.85	53.35	64.11	32.16	259.16	4.48
No. 3	F/V SN1	26.08	82.75	52.16	64.76	14.89	240.65	6.49
	F/V SN2	26.08	81.83	52.17	64.01	20.28	244.38	8.78
	F/V SN3	26.20	83.04	52.41	64.78	25.28	251.74	7.54
No. 4	V/V SN1	26.48	82.71	52.96	63.92	11.09	237.17	6.79
	V/V SN2	26.52	82.92	53.05	63.39	16.35	242.26	7.23
	V/V SN3	26.27	83.09	52.54	64.10	25.57	251.59	8.78
	V/V SN4	26.29	83.15	52.59	63.90	27.86	253.82	7.48

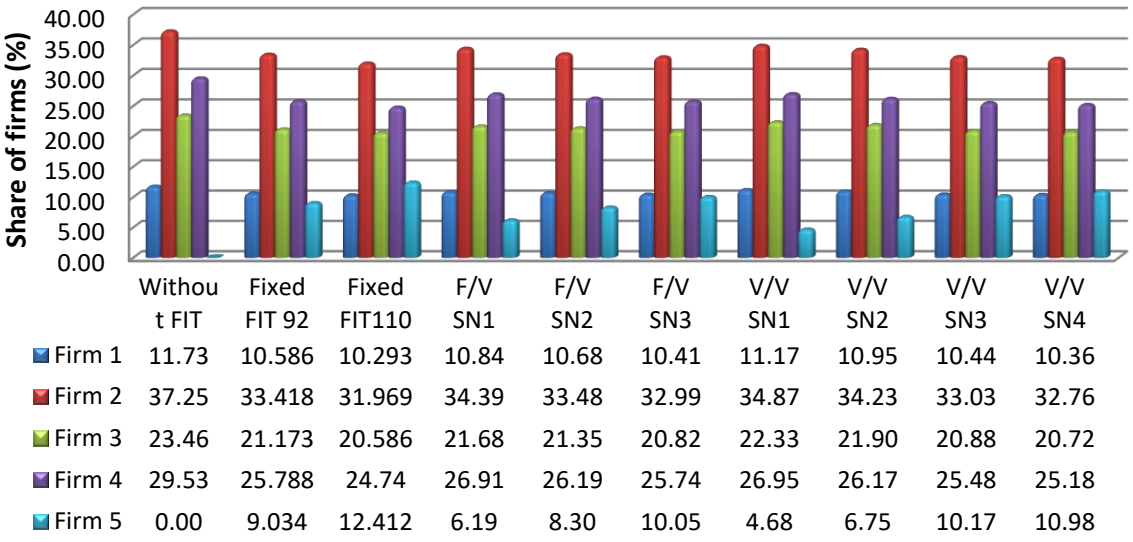


Fig. 13: Contribution of each firm in different types of FIT.

Conclusion

The first contribution of this article is to propose a developed model in order to simulate the medium term deregulated power market by assuming the hybrid wind-thermal firm. The second contribution is to evaluate the impact of different types of FIT on MCP, ECG, profits and contribution of each firm in electricity generation in the restructured market of energy. To this end, the wind power generation has been evaluated based on the scenario based method.

The wind scenarios are generated based on the data mining technique. Besides the wind uncertainty, the demand and fuel price uncertainties are assumed in this approach based on the Monte-Carlo technique. Also, the strategic behavior of other participants in each tactical period evaluates based on the Cournot game theory. The findings affirm that the wind resources could not compete with conventional firms. Furthermore, the proposed model in this article can be so useful for evaluating the

different types of incentive policies for renewable energies. Moreover, this study confirms the previous researches that selected the FIT as an efficient incentive policy for developing the wind resources. For future work this model can be examined with quota or other incentive policies in the restructured power market.

Author Contributions

Mohammad Tolou Askari: Programmer, Software, Validation, Conceptualization, Visualization, Investigation, Writing - Reviewing and Editing, Conceptualization, Methodology, Visualization, Investigation, Writing - Original draft preparation.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Acknowledgment

The authors gratefully thank the anonymous reviewers and the editor of JECEI for their useful comments and suggestions.

List of Variables and Parameters in the Model

i.Indices

l	Load level
s	Season
e	Traditional generation firm
$e^{'}$	Hybrid traditional and renewable generation firm
u	Thermal unit of traditional firm
u'	Thermal unit of hybrid traditional and renewable generation firm
uw	Wind unit of hybrid traditional and renewable generation firm

ii. parameters

d_{sl}	Duration of time [hours]
$N_{e,u}$	Number of thermal units in e
$N_{e',u'}$	Number of thermal units in $e^{'}$
$N_{e^{'},uw}$	Number of wind units in $e^{'}$
$Q_{e,sl}$	Total generation contracted by firm e in sl [MW]
$Q_{e',sl}$	Total generation contracted by firm $e^{'}$ in sl [MW]
SP_{sl}	Average electricity price of power market in sl [$/MW]
BP_{sl}	Contracted electricity price in sl [$/MW]
FP_u	Fuel price of unit u [$/MBtu]
a_u, b_u, c_u	Constant coefficients of heat rate function for unit u
$a_{u'}, b_{u'}, c_{u'}$	Constant coefficients of heat rate function for unit u’
Tax	CO2 tax rate [$/lbCO2]
EM_u	CO2 produced by unit u[lb/MMBTU]
GP	Percentage of electricity price
D_{sl}	Average demand in sl [MW]
$P_{e,u\min}$	Minimum generation of thermal unit u of firm e [MW]
$P_{e,u\max}$	Maximum generation of thermal unit u of firm e [MW]
$P_{e',u'\min}$	Minimum generation of thermal unit u’ of firm e' [MW]
$P_{e',u'\max}$	Maximum generation of thermal unit u’ of firm e' [MW]
$PW_{e',uw,n\min}$	Minimum generation of wind unit uw of firm e' for scenario n [MW]
$PW_{e',uw,n\max}$	Maximum generation of wind unit uw of firm e' for scenario n [MW]

iii. Decision variables

$P_{e,u,sl}$	Power generation by thermal unit u of firm e in sl [MW]
$P_{e',u',sl}$	Power generation by thermal unit u’ of firm e’ in sl [MW]
$PW_{e',uw,sl,n}$	Power generation by wind unit uw of firm e’ in sl for scenario n [MW]
$g_{e,e^{'},sl}$	Total power generation of firms e, e' [MW]
MCPsl	Market clearing price in sl [$/MW]

References

[1] P. Brown, "European Union Wind and Solar Electricity Policies: Overview and Considerations," 2013.

[2] C.-D. Yue, C.-M. Liu, E. M. Liou, "A transition toward a sustainable energy future: feasibility assessment and development strategies of wind power in Taiwan," Energy Policy, 29: 951-963, 2001.

[3] R. Saidur, M. Islam, N. Rahim, K. Solangi, "A review on global wind energy policy," Renewable Sustainable Energy Rev., 14: 1744-1762, 2010.

[4] M. Askari, M. Ab Kadir, H. Hizam, J. Jasni, "A new comprehensive model to simulate the restructured power market for seasonal price signals by considering on the wind resources," J. Renewable Sustainable Energy, 6: 023104, 2014.

[5] M. Askari et al., "Modeling optimal long-term investment strategies of hybrid wind-thermal companies in restructured power market," J. Mod. Power Syst. Clean Energy, 7: 1267–1279, 2019.

[6] P. Harborne, C. Hendry, "Pathways to commercial wind power in the US, Europe and Japan: The role of demonstration projects and field trials in the innovation process," Energy Policy, 37: 3580-3595, 2009.

[7] T. Ackermann, L. Söder, "An overview of wind energy-status 2002," Renewable Sustainable Energy Rev., 6: 67-127, 2002.

[8] T. Barforoushi, M. Parsa Moghaddam, M. Javidi, M. Sheik-El-Eslami, "A new model considering uncertainties for power market simulation in medium-term generation planning," Iran. J. Electr. Electron. Eng. (IJEEE), 2: 71-81, 2006.

[9] T. Barforoushi, M. P. Moghaddam, M. H. Javidi, M. K. Sheikh-El-Eslami, "Evaluation of regulatory impacts on dynamic behavior of investments in electricity markets: a new hybrid DP/GAME framework," IEEE Trans. Power Syst., 25: 1978-1986, 2010.

[10] D. Kothari, I. Nagrath, Modern power system analysis: McGraw-Hill Europe, 2003.

[11] A. S. Chuang, F. Wu, P. Varaiya, "A game-theoretic model for generation expansion planning: problem formulation and numerical comparisons," IEEE Trans. Power Syst., 16: 885-891, 2001.

[12] J. B. Park, J. H. Kim, K. Y. Lee, "Generation expansion planning in a competitive environment using a genetic algorithm," 3: 1169-1172, 2002.

[13] W. M. Lin, T. S. Zhan, M. T. Tsay, W. C. Hung, "The generation expansion planning of the utility in a deregulated environment," 2: 702-707, 2004.

[14] J. Zhu, M. Chow, "A review of emerging techniques on generation expansion planning," IEEE Trans. Power Syst., 12: 1722-1728, 1997.

[15] E. Alishahi, M. P. Moghaddam, M. Sheikh-El-Eslami, "A system dynamics approach for investigating impacts of incentive mechanisms on wind power investment," Renewable Energy, 37: 310-317, 2012.

[16] G. Richards, B. Noble, K. Belcher, "Barriers to renewable energy development: A case study of large-scale wind energy in Saskatchewan, Canada," Energy Policy, 42: 691-698, 2012.

[17] J. M. Kissel, S. C. Krauter, "Adaptations of renewable energy policies to unstable macroeconomic situations—Case study: Wind power in Brazil," Energy Policy, 34: 3591-3598, 2006.

[18] V. Di Dio, S. Favuzza, D. La Cascia, R. Miceli, "Economical Incentives and systems of certification for the production of electrical energy from renewable energy resources," in Proc. International Conference on Clean Electrical Power (ICCEP'07): 277-282, 2007.

[19] J. P. Painuly, "Barriers to renewable energy penetration; a framework for analysis," Renewable Energy, 24: 73-89, 2001.

[20] L. Ying, C. Yin, R. Yuan, H. Yong, "Economic incentive mechanism of renewable energy generation," in Proc. International Conference on Electrical Machines and Systems (ICEMS): 2689-2694, 2008.

[21] L. A. Barroso, H. Rudrick, F. Sensfuss, P. Linares, "The green effect," IEEE Power Energy Magazine, 8: 22-35, 2010.

[22] J. Sawin, "National policy instruments: Policy lessons for the advancement & diffusion of renewable energy technologies around the world," Renewable Energy, A Global Review of Technologies, Policies and Markets, 2006.

[23] J. Lipp, "Lessons for effective renewable electricity policy from Denmark, Germany and the United Kingdom," Energy Policy, 35: 5481-5495, 2007.

[24] V. Lauber, L. Mez, "Three decades of renewable electricity policies in Germany," Energy & Environment, 15: 599-623, 2004.

[25] M. Askari, M. Ab Kadir, E. Bolandifar, "Evaluation of FIT impacts on market clearing price in the restructured power market," in Proc. 2015 IEEE Student Conference on Research and Development (SCOReD): 224-227, 2015.

[26] S. Ahmad, R. M. Tahar, F. Muhammad-Sukki, A. B. Munir, R. A. Rahim, "Application of system dynamics approach in electricity sector modelling: A review," Renewable Sustainable Energy Rev., 56: 29-37, 2016.

[27] A. Ibanez-Lopez, J. Martinez-Val, B. Moratilla-Soria, "A dynamic simulation model for assessing the overall impact of incentive policies on power system reliability, costs and environment," Energy Policy, 102: 170-188, 2017.

[28] F. Zia, M. Nasir, A. A. Bhatti, "Optimization methods for constrained stochastic wind power economic dispatch," in Proc. 7th IEEE International Power Engineering and Optimization Conference (PEOCO): 129-133, 2013.

[29] X. Liu, W. Xu, "Economic load dispatch constrained by wind power availability: A here-and-now approach," IEEE Trans. Sustainable Energy, 1(1): 2-9, 2010.

[30] L. A. Barroso, A. J. Conejo, "Decision making under uncertainty in electricity markets," in Proc. IEEE Power Engineering Society General Meeting, 2006.

[31] A. J. Conejo, M. Carriâon, J. M. Morales, Decision making under uncertainty in electricity markets , 153: Springer, 2010.

[32] R. Karki, P. Hu, R. Billinton, "A simplified wind power generation model for reliability evaluation," IEEE Trans. Energy Convers., 21: 533-540, 2006.

[33] (2006). Canada's National Climate Archive [online].

[34] M. J. Osborne, An introduction to game theory vol. 3: Oxford University Press New York, 2004.

[35] A. S. Chuang, F. Wu, P. Varaiya, "A game-theoretic model for generation expansion planning: problem formulation and numerical comparisons," IEEE Trans. Power Syst., 16: 885-891, 2001.

[36] D. Kirschen, G. Strbac, Front Matter: Wiley Online Library, 2005.

[37] H. Singh, "Introduction to game theory and its application in electric power markets," IEEE Comput. Appl. Power, 12(4): 18-20, 1999.

[38] C. Grigg, P. Wong, P. Albrecht, R. Allan, M. Bhavaraju, R. Billinton, et al., "The IEEE reliability test system-1996. A report prepared by the reliability test system task force of the application of probability methods subcommittee," IEEE Trans. Power Syst., 14: 1010-1020, 1999.

[39] E. Gnansounou, J. Dong, S. Pierre, A. Quintero, "Market oriented planning of power generation expansion using agent-based model," in Proc. IEEE Power Systems Conference and Exposition: 1306-1311, 2004.

[40] M. Shahidehpour, H. Yamin, Z. Li, "Market operations in electric power systems, 2002," ed: John Wiley & Sons, Inc. New York.

[41] T. Barforoushi, M. P. Moghaddam, M. H. Javidi, M. K. Sheikh-El-Eslami, "Evaluation of regulatory impacts on dynamic behavior of investments in electricity markets: a new hybrid DP/GAME framework," IEEE Trans. Power Syst., 25: 1978-1986, 2010.

Biographies

Mohammad Tolou Askari enrolled at University of Applied Science and Technology of Mashhad and obtained his first degree in Bachelor of Power Electrical Engineering in 2005. He continued his education in Master of Power Electrical Engineering in 2008. He received PhD degree in Power electrical engineering (2014) from University Putra Malaysia. Currently, he is Assistant Professor with Islamic Azad University, Semnan, Iran. His research interests include electrical transformers, Smart grids, Micro grids, GEP & TEP, Distribution systems.

- Email: m.askari@semnaniau.ac.ir
- ORCID: 0000-0002-5473-7708
- Scopus Author ID: 36103897600, 57221941296
- Web of Science Researcher ID: AAO-2829-2021
- Homepage: https://scimet.iau.ir/Mohammad_TolouAskariSedehiEsfahani

How to cite this paper:
M. Tolou Askari, "An approach for evaluating incentive policy of wind resources considering the uncertainties in the deregulated power market," J. Electr. Comput. Eng. Innovations, 11(1): 203-216, 2023.

DOI: 10.22061/jecei.2022.8980.563

URL: https://jecei.sru.ac.ir/article_1784.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Novel Ultra-Low-Power Mirrored Folded-Cascade Transimpedance Amplifier

S. Sadeghi[1], M. Nayeri[1,], M. Dolatshahi[2], A. Moftakharzadeh[3]*

[1]*Department of Electrical Engineering, Yazd Branch, Islamic Azad University, Yazd, Iran.*

[2]*Department of Electrical Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Iran.*

[3]*Department of Electrical Engineering, Yazd University, Yazd, Iran.*

Article Info

Article History:

Received 06 June 2022
Reviewed 15 July 2022
Revised 18 August 2022
Accepted 26 September 2022

Keywords:

Trans-impedance amplifier
Folded-mirror
Optical receiver
Folded-cascade
Ultra-low-power

*Corresponding Author's Email Address: *nayeri@iauyazd.ac.ir*

Abstract

Background and Objectives: In this paper, a novel structure as a Folded-Mirror (FM) Trans-impedance Amplifier (TIA) is designed and introduced for the first time based on the combination of the current-mirror and the folded-cascade topologies. The trans-impedance amplifier stage is the most critical building block in a receiver system. This novel proposed topology is based on the combination of the current mirror topology and the folded-cascade topology, which is designed using active elements. The idea is to use a current mirror topology at the input node. In the proposed circuit, unlike many other reported designs, the signal current (and not the voltage) is being amplified till it reaches the output node. The proposed TIA benefits from a low input resistance, due to the use of a diode-connected transistor, as part of the current mirror topology, which helps to isolate the dominant input capacitance. So, as a result, the data rate of 5Gbps is obtained by consuming considerably low power. Also, the designed circuit employs only six active elements, which yields a small occupied chip area, while providing 40.6dBΩ of trans-impedance gain, 3.55GHz frequency bandwidth, and 664nArms input-referred noise by consuming only 315µW power using a 1V supply. Results justify the proper performance of the proposed circuit structure as a low-power TIA stage.

Methods: The proposed topology is based on the combination of the current mirror topology and the folded-cascade topology. The circuit performance of the proposed folded-mirror TIA is simulated using 90nm CMOS technology parameters in the Hspice software. Furthermore, the Monte-Carlo analysis over the size of widths and lengths of the transistors is performed for 200runs, to analyze the fabrication process.

Results: The proposed FM TIA circuit provides 40.6dBΩ trans-impedance gain and 3.55GHz frequency bandwidth, while, consuming only 315µW power using a 1V supply. Besides, as analyzing the quality of the output signal in the receiver circuits for communication applications is vital, the eye-diagram of the proposed FM TIA for a 50µA input signal is opened about 5mV, while, for a 100µA input signal the eye is opened vertically about 10mV. So, the vertical and horizontal opening of the eye is clearly shown. Furthermore, Monte-Carlo analysis over the trans-impedance gain represents a normal distribution with the mean value of 40.6dBΩ and standard deviation of 0.4dBΩ. Also, the value of the input resistance of the FM TIA is equal to 84.4Ω at low frequencies and reaches the value of 75Ω at -3dB frequency. The analysis of the effect of the feedback network on the value of the input resistance demonstrates the input resistance in the absence of the feedback network reaches up to 1.4MΩ, which yields the importance of the existence of the feedback network to obtain a broadband system.

Conclusion: In this paper, a trans-impedance amplifier based on a combination of the current-mirror topology and the folded-cascade topology is presented, which amplifies the current signal and converts it to the voltage at the output node. Due to the existence of a diode-connected transistor at the input node, the input resistance of the TIA is comparatively small. Furthermore, four out of six transistors are PMOS transistors, which represent less thermal noise in comparison with NMOS transistors. Also, the proposed Folded-Mirror topology occupies a relatively small area on-chip, due to the fact that no passive element is used in the feedforward network. Results using 90nm CMOS technology parameters show 40.6dBΩ trans-impedance gain, 3.55GHz frequency bandwidth, 664nArms input-referred noise, and only 315µW power dissipation using a 1volt supply, which indicates the proper performance of the proposed circuit as a low-power building block.

Introduction

The beam of light, as the fastest signal carrier, was always an attractive candidate for communication systems. Optical fibers, as a proper medium for transferring a beam of light, introduce better performance in terms of crosstalk, bandwidth, electromagnetic interference, and channel loss in comparison with conventional mediums [1].

Also, the rapid increase in the transit frequency of CMOS technologies made deep-submicron CMOS devices a proper candidate to provide an acceptable level of integration besides a proper level of speed and low cost.

In Fig. 1 the building block of the transmitter and the optical receiver system are demonstrated. At the receiver stage after the photodiode, the trans-impedance amplifier (TIA) stage, as the most critical building block in a receiver system, is shown in gray. The photodiode receives the optical signal and proportionally produces a weak signal current.

At the Far-end, a weak signal current in the range of microampere is detected [1], [2], which requires to be amplified with low noise and a proper bandwidth, to be detectable in the digital circuitry. Of course, nonlinearities and second-order effects besides the trade-offs among gain, bandwidth, speed, noise, power consumption, and voltage headroom are part of the challenges the designer must consider when using deep-submicron technologies. Furthermore, a large parasitic capacitance in the input node of the TIA limits the frequency bandwidth at the beginning [4]-[10].

Of course, many researchers have published many different structures such as regulated cascade (RGC) structures [3], [11]-[13] to compensate for the effect of this large parasitic capacitance. In [11]-[13] broadband circuits are introduced using passive inductors and resistors to enlarge the bandwidth, which of course requires a large occupied area on the chip.

Additionally, high voltage headroom is required for RGC structures at high-speed applications, which is not possible due to the occurrence of the quantum tunneling phenomena in nanometer CMOS technologies.

In [14] a method, which converts the transconductance of a transistor into a trans-impedance, is proposed.

In this method, no resistor is required to do the conversion, and a further degree of freedom is obtained in comparison with previously published circuit structures, but the usage of passive inductors in this structure yields a large occupied chip area. Moreover, a π-network as the TIA stage is proposed in [15] alongside a shunt amplifier based on folded-cascade structures, which benefits from a high gain and low-noise characteristic, while, suffering from high power consumption and a large occupied chip area due to the use of passive inductors.

Also inverter [5], [6]-[21] is another attractive structure used in designing TIA stages. In [5], a cascaded circuit structure is employed in a conventional inverter structure, which eliminates the Miller capacitance and enlarges the bandwidth, but limits the output swing. Furthermore, a three-stage cascaded push-pull conventional inverter, which uses a series inductive peaking technique to extend the bandwidth, is proposed in [16].

Of course, this technique also requires a largely occupied area on-chip. Additionally, an inverter employing a diode-connected NMOS and a cascaded PMOS is proposed in [17], which provides a wide dynamic range with 227MHz frequency bandwidth.

Moreover, a conventional inverter employing active feedback with an extra gain stage is proposed in [18], while, a similar circuit is proposed in [19], which uses an inverter structure in its input stage, followed by a 1.5KΩ feedback resistor.

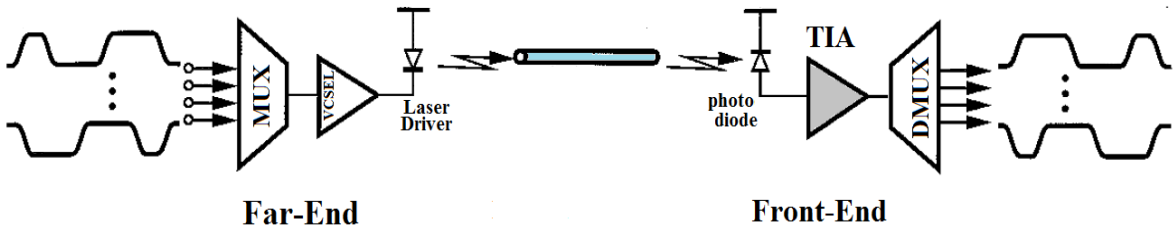


Fig. 1: Location of the TIA stage in an optical receiver system [1].

Clearly, a 1.5KΩ resistor occupies a considerable area on-chip. Also, a conventional inverter is employed as the booster amplifier in an RGC structure [19], which introduces low input resistance that isolates further the parasitic capacitance of the photodiode, while, suffering from the miller capacitance in the inverter stage.

In [7] a current-mirror-based TIA is proposed, which amplifies the current signal and covers it to the voltage at the output node, unlike many other reported TIAs, which convert the current signal to a voltage signal at the begging, and then try to amplify the voltage.

In this paper, a new trans-impedance amplifier namely "Folded-Mirror" (FM) is proposed, which benefits from a low input resistance (due to the use of a diode-connected transistor at the input node), and a relatively low noise behavior (due to the use of PMOS transistors instead of NMOS transistors). As the time constant of the dominant pole is reduced due to the small value of the input resistance, the circuit is capable of providing an extended bandwidth, without the requirement of consuming extra power.

The Proposed TIA

This novel proposed topology is based on the combination of the current mirror topology and the folded-cascade topology, which is designed using active elements.

The idea is to use a current mirror topology at the input node, as in Fig. 2-1 (a), which introduces a small value of $(gm)^{-1}$ as the input resistance. Then, the signal requires to be amplified further in a cascade stage. So, a cascade stage is added to the structure, and a current source is used as its load, as in Fig. 2-1 (b). As these two stages cannot provide proper trans-impedance gain, a folded cascade structure is used instead of the cascade structure, to fold the current signal toward M5, as it is shown in Fig. 2-1 (c).

The M5 transistor, as the diode-connected load, is used in a current mirror topology to further amplify the signal, as in Fig. 2-1 (d). So, the signal is now amplified in three steps, in which M2 and M5 are used commonly in the current mirror structure and in the folded cascade structure, simultaneously. Finally, Fig. 2-1 (e) demonstrates the active type of the proposed open-loop TIA.

Fig. 2 (b) demonstrates the final version of the proposed FM TIA. The produced signal of the photodiode amplifies in a current-mirror structure (consists of M1 and M2), a folded-cascade structure (consists of M2, M3, M4, and M5) and in the second current-mirror topology (consists of M5 and M6), respectively. Usage of a current-mirror stage at the input node introduces a low input resistance, which isolates the parasitic capacitance of the photodiode. Moreover, the proper usage of the voltage-current feedback decreases the output resistance and the input resistance even more. Also, the usage of four PMOS transistors out of six transistors yields less generated thermal noise, due to the less mobility of holes in the PMOS transistors.

Moreover, Fig. 3 shows the model of the photodiode [25], [26] and Fig. 4 demonstrates the equivalent circuit of the proposed FM TIA.

So, the open-loop trans-impedance gain (AV) of the FM TIA can be calculated as follows:

$$A_V = \frac{g_{m2}}{g_{m1}} \times \frac{g_{m6}}{g_{m5}} \times r_{o6} \tag{1}$$

which, g_m represents the transconductance, and r_o represents the drain-source resistance of the MOSFET.

Considering the fact that the gate-source voltage of M1 and M2 are equal ($V_{gs1} = V_{gs2}$) and also $V_{gs5} = V_{gs6}$, and M1, M2, M5 and M6 are PMOS transistors ($\mu_{p1} = \mu_{p2}$,$C_{ox5} = C_{ox6}$) with a same length at a specific technology ($L_1 = L_2$, $L_5 = L_6$), (1) can be simplified as follows:

$$A_V = \frac{w_2}{w_1} \times \frac{w_6}{w_5} \times r_{o6} \tag{2}$$

which $\left(\frac{w_2}{w_1} \times \frac{w_6}{w_5}\right)$ defines the current amplification.

The input resistance of the proposed TIA is comparatively small, due to the use of the diode-connected transistor M1, which yields the value of $\left(\frac{1}{g_{m1}}\right)$ as the input resistance, for the open-loop FM TIA. So, the input resistance of the closed-loop FM TIA ($R_{in,f}$) can be calculated as follows:

$$R_{in,f} = \frac{\frac{1}{g_{m1}}}{1 + A_V . \frac{1}{R_f}} = \frac{g_{m5}\, R_f}{g_{m1}\, g_{m5}\, R_f + g_{m2}\, g_{m6}\, r_{o6}} \tag{3}$$

where R_f represents the feedback resistance.

Besides, the output resistance of the open loop FM TIA is equal to (r_{o6}).

In order to calculate the closed-loop output resistance, it can be written as follows:

$$R_{out,f} = \frac{r_{o6}}{1 + A_V \frac{1}{R_f}} = \frac{g_{m1}\, g_{m5}\, R_f\, r_{o6}}{g_{m1}\, g_{m5}\, R_f + g_{m2}\, g_{m6}\, r_{o6}} \tag{4}$$

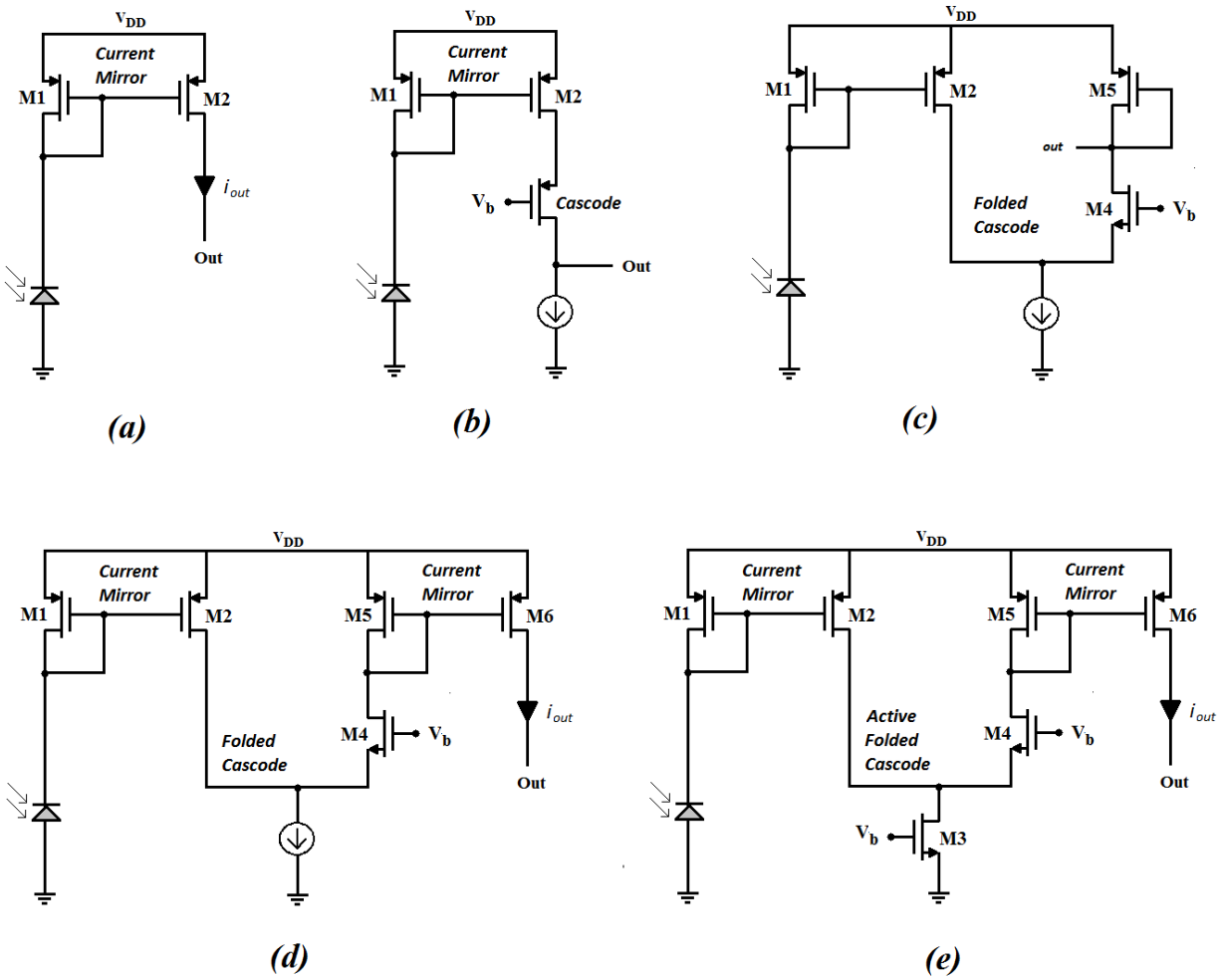


Fig. 2-1: Design Process, a) current mirror topology, b) a cascade stage is added to (a), c) a folded cascade structure instead of the cascade structure in (b), (d) a current mirror topology is added to (c), and e) the active type of the proposed open-loop TIA.

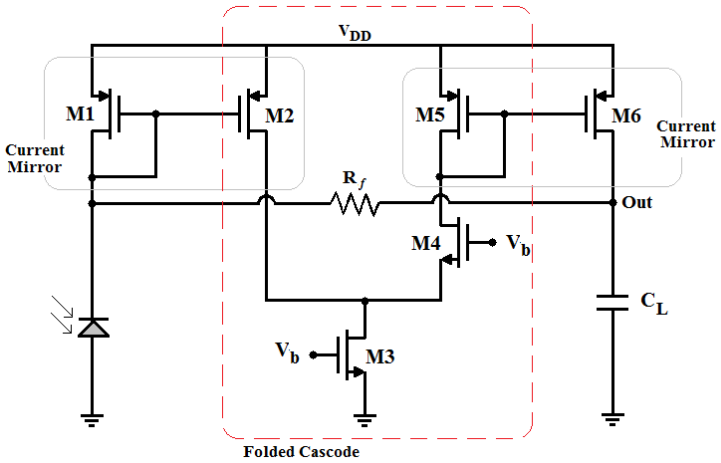


Fig. 2-2: The Final Model of the proposed TIA.

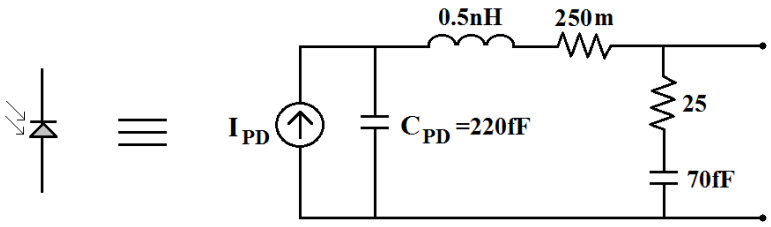


Fig. 3: Model of the Photodiode.

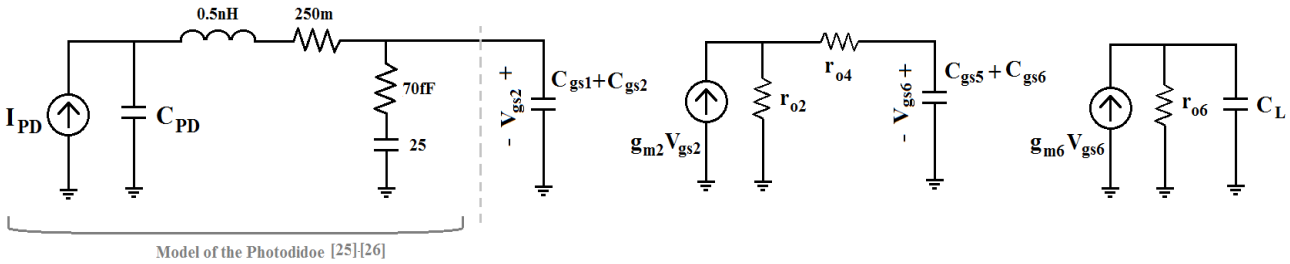


Fig. 4: Equivalent Circuit of the Proposed FM TIA.

Also, for the capacitance seen at the input and output nodes of the FM TIA, it can be written as follows:

$$C_{in} = C_{gs1} + C_{gs2} + C_{Pd} \approx C_{Pd} \tag{5}$$

$$C_{out} = C_L \tag{6}$$

which C_{gs} represents the gate-source parasitic capacitance of a MOSFET, C_{Pd} is the parasitic capacitance of the photodiode and C_L represents the load capacitance. As the parasitic capacitance of the photodiode contains comparatively a large value, it can be concluded that the input capacitance of the TIA is approximately equal to C_{Pd}.

So, in order to calculate the closed-loop trans-impedance gain of the proposed FM TIA at low frequencies, it can be written as follows:

$$A_{V,f} = \frac{A_V}{1 + A_V \frac{1}{R_f}} = \frac{g_{m2}\, g_{m6}\, R_f\, r_{o6}}{g_{m1}\, g_{m5}\, R_f + g_{m2}\, g_{m6}\, r_{o6}} \tag{7}$$

And hence, the transfer function of the proposed FM TIA can be achieved as follows:

$$A_V(S) = \frac{A_{V,f}}{\left(1 + S.C_{in}.R_{in,f}\right)\left(1 + S.C_{out}.R_{out,f}\right)} \tag{8}$$

By using (3) to (7) and considering the fact that $C_{in} \gg C_{out}$, (8) can be re-written as follows:

$$A_V(S) = \frac{g_{m2}\, g_{m6}\, R_f\, r_{o6}}{\left(g_{m1}\, g_{m5}\, R_f + g_{m2}\, g_{m6}\, r_{o6}\right)} \frac{1}{\left(1 + S.C_{Pd}\, \frac{g_{m5}\, R_f}{g_{m1}\, g_{m5}\, R_f + g_{m2}\, g_{m6}\, r_{o6}}\right)} \tag{9}$$

As (9) reveals, the proposed FM TIA is approximated as a single pole circuit, with its pole equal to $S \approx -\frac{g_{m1}.g_{m5}.R_f + g_{m2}.g_{m6}.r_{o6}}{C_{pd}.g_{m5}.R_f}$. So, the -3dB frequency can be written as follows:

$$f_{-3dB} \approx \frac{g_{m1}.g_{m5}.R_f + g_{m2}.g_{m6}.r_{o6}}{2\pi.C_{pd}.g_{m5}.R_f} \tag{10}$$

Results and Discussions

In the following, the circuit performance of the proposed folded-mirror TIA is simulated using 90nm CMOS technology parameters. The frequency response of the proposed TIA up to 10GHz is demonstrated in Fig. 5. As Fig. 5 presents, the proposed FM TIA circuit provides 40.6dBΩ trans-impedance gain and 3.55GHz frequency bandwidth, while, consuming only 315µW power using a 1V supply.

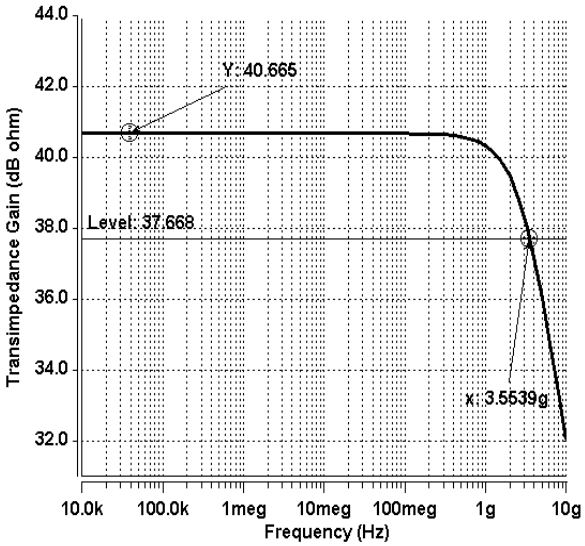
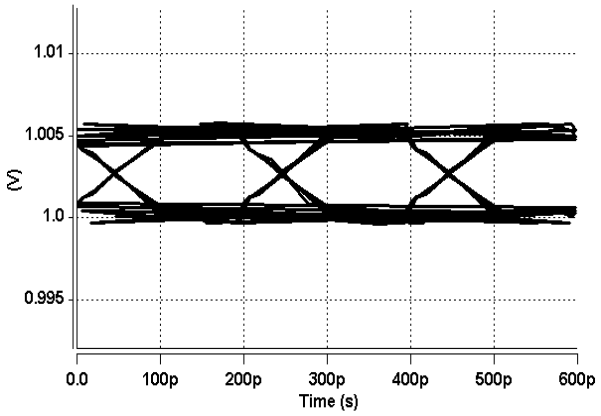


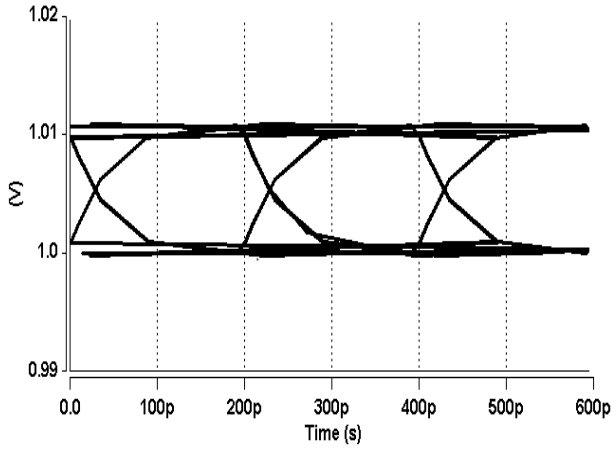
Fig. 5: Frequency Response of the proposed FM TIA.

As analyzing the quality of the output signal in the receiver circuits for communication applications is vital, the eye-diagram of the proposed FM TIA is demonstrated in Fig. 6, using Non-Return to Zero (NRZ) Pseudo-Random Bit Sequence (PRBS) 2^7-1 for two different values of 50µA and 100µA input signals, respectively. As Fig. 6 suggests for a 50µA input signal, the eye is opened about 5mV, while, for a 100µA input signal the eye is opened vertically about 10mV. So, the vertical and horizontal opening of the eye is clearly shown.

Furthermore, the Monte-Carlo analysis over the size of the widths and lengths of the transistors is performed for 200runs, to analyze the fabrication process. Fig. 7 demonstrates the results over frequency response, while, Fig. 8 demonstrates the results over the trans-impedance gain. Monte-Carlo analysis over the trans-impedance gain represents a normal distribution (the red line) with the mean value of 40.6dBΩ and standard deviation of 0.4dBΩ, as in Fig. 8.



(a)



(b)

Fig. 6: The eye-diagram of the FM TIA using NRZ PRBS for (a) 50µA and (b) 100µA input signal.

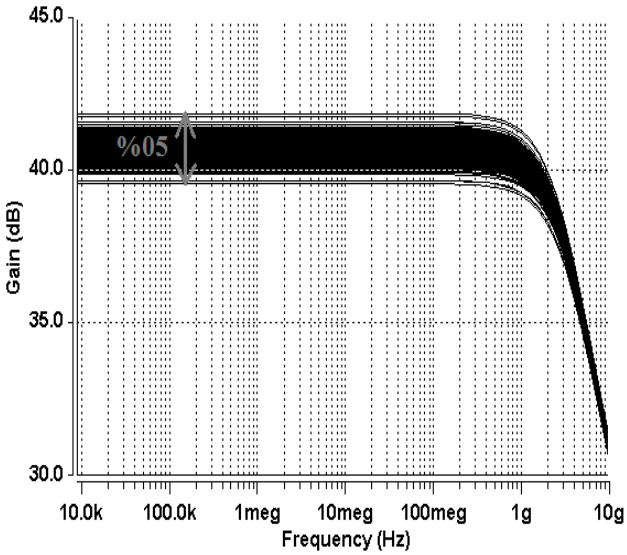


Fig. 7: Monte-Carlo Analysis over frequency response.

Also, the input resistance of the optical receivers (the TIA stage) is a challenging parameter, as discussed before. So, the input resistance of the proposed FM TIA versus frequency is shown in Fig. 9.

As it was theoretically discussed before, the input resistance of the FM TIA should be relatively small due to the existence of a diode-connected transistor at the input node, and the use of a voltage-current feedback. So, Fig. 9 displays the value of the input resistance of the FM TIA, which is equal to 84.4Ω at low frequencies and reaches the value of 75Ω at -3dB frequency.

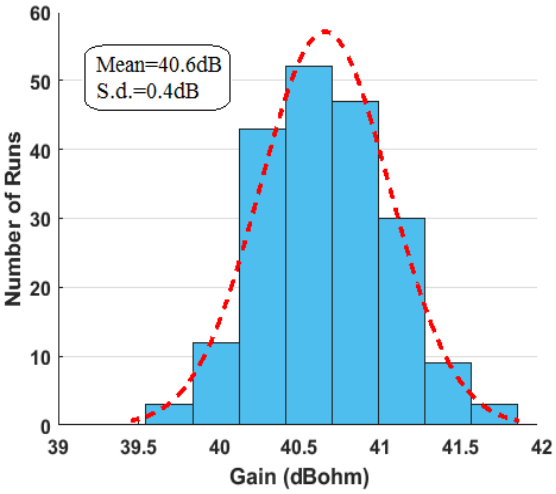


Fig. 8: Monte-Carlo Analysis over transimpedance gain.

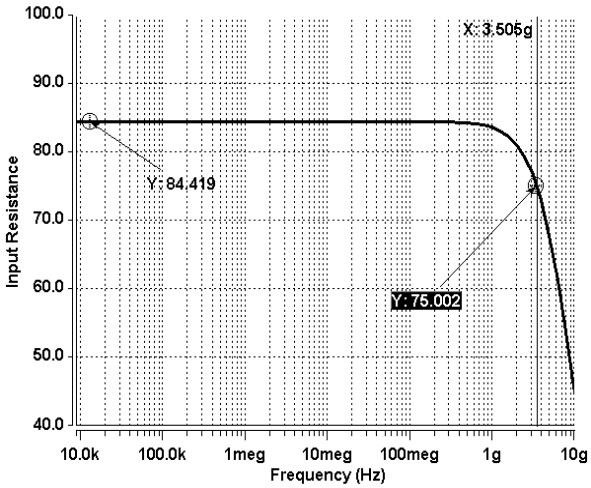


Fig. 9: Input resistance of the FM TIA.

Furthermore, the effect of the feedback network on the value of the input resistance is analyzed and summarized in Table 1. As it can be concluded from Table 1, the input resistance in the absence of the feedback network reaches up to 1.4MΩ, which yields the importance of the existence of the feedback network to obtain a broadband system.

Table 1: effect of the feedback network on the input resistance

	The Open-loop TIA	The Closed-loop TIA
Input resistance (@low freq.)	84.4Ω	1.4MΩ

As it is important that a broadband system can operate properly in a reasonable range of temperature, the effect of temperature variations on the frequency response of the proposed FM TIA is analyzed, and the results are given for three different values of -30ºC, +30ºC, and +90ºC in Fig. 10. As Fig. 10 suggests, increasing the temperature results in an increased gain while resulting in a decreased frequency bandwidth, which shows the trade-off between the trans-impedance gain and the frequency bandwidth. Table 2 numerically summarizes this analysis.

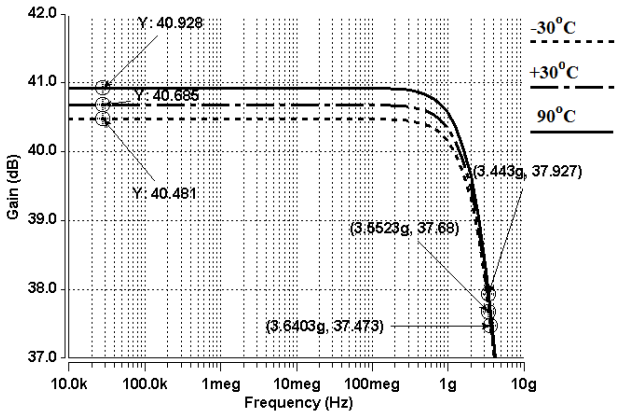


Fig. 10: Effect of temperature variations on frequency response.

Table 2: Effect of temperature variation on transimpedance gain, frequency bandwidth and power consumption

	-30ºC	+30ºC	+90ºC
Transimpedance Gain	40.4dBΩ	40.6dBΩ	40.9dBΩ
Frequency Bandwidth	3.64GHz	3.55GHz	3.44GHz
Power Consumption	274µW	318µW	355µW

Moreover, the sensitivity of the proposed FM TIA to VDD is analyzed and the results are given in the following. In Fig. 11, the result of %10 variations of the supply voltage (VDD) is shown over frequency response. According to Fig. 11, the trans-impedance gain varies from 40.34dBΩ to 41.02dBΩ (varies about 0.68dB), while, the frequency bandwidth varies from 3.425GHz to 3.695GHz (270MHz). Also, Table 3 summarizes the numerical analysis of supply voltage variations. As Table 3 reveals, a %10 reduction in the value of the supply voltage (from VDD to 0.9VDD), results in 0.06 less power dissipation, and 0.04 less bandwidth, while, 0.01 more gain value can be achieved.

Table 3: Effect of V_{DD} variation on Transimpedance gain, frequency bandwidth and power consumption

	$1.1V_{DD}$	V_{DD}	$0.9V_{DD}$
Transimpedance Gain	40.34dB	40.66dB	41.02dB
Frequency Bandwidth	3.69GHz	3.56GHz	3.42GHz
Power Consumption	413µW	315µW	298µW

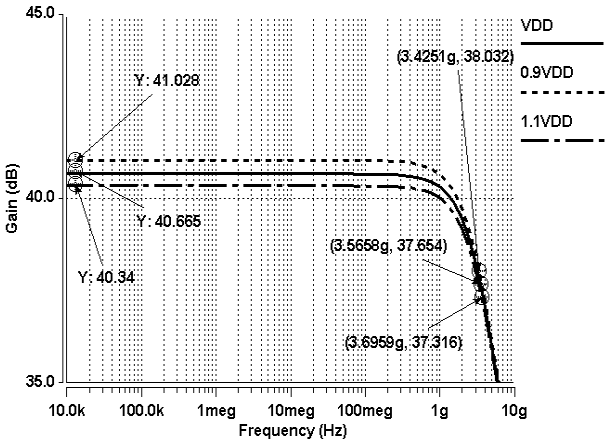


Fig. 11: Supply voltage variations Vs. frequency response.

Noise Analysis

The demonstration of noise sources in the block diagram of the proposed FM TIA as in Fig. 12, provides a better understanding of the noise performance in this circuit. According to (11), input-referred noise of the proposed FM TIA circuit structure can be calculated as the sum of noise in the core of the TIA, and the feedback network, as follows [1]:

$$\overline{I_{n,in}^2} = \overline{I_{n,Rf}^2} + \frac{\overline{V_{n,Core}^2}}{R_f^2} \tag{11}$$

which

$$\overline{I_{n,Rf}^2} = \frac{4KT}{R_f} \tag{12}$$

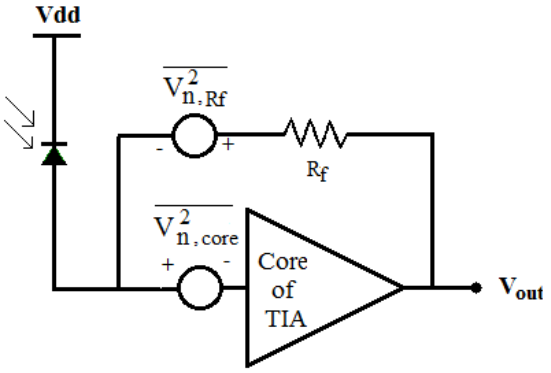


Fig. 12: Demonstration of noise sources in the block diagram of the FM TIA.

Now, according to (11) calculation of the noise of the TIA core is required. So, a current source is put in parallel with the drain-source terminals of transistors, to demonstrate the produced thermal noise in each transistor, as in Fig. 13. First of all, it should be noted that M1 is operating in the triode region, due to the fact that it is used as a diode-connected transistor. Hence, the generated noise of M1 is shunted to the ground [27]. So, according to the shunted parasitic capacitance of the photodiode, the produced thermal noise of M1 can be calculated as follows:

$$\overline{I_{n,M1}^2} = \frac{KT}{C_{Pd}} \tag{13}$$

which, K is the Boltzmann constant and T is the temperature.

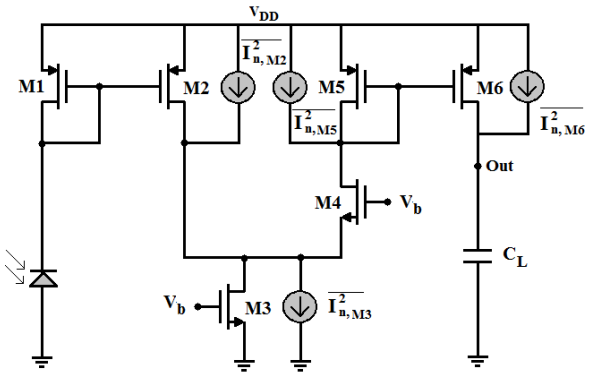


Fig. 13: representation of noise in the core of the FM TIA by current sources.

Besides, as M4 forms a cascade structure, the thermal noise generated by M4 is negligible [27]; due to the fact that if the channel length modulation of M4 is neglected, it can be said that $I_{n2} + I_{D2} = 0$, and hence M4 does not affect $V_{n,out}$.

So, considering (11) and Fig. 13, $\overline{V_{n,Core}^2}$ can be calculated as follows:

$$\overline{V_{n,Core}^2} = 4KT\gamma\left[\frac{g_{m2}+g_{m3}+g_{m5}}{|g_{m2}|^2} + \frac{g_{m6}}{\left|\frac{g_{m6}}{g_{m5}}.\frac{g_{m2}}{g_{m1}}\right|^2}\right] \tag{14}$$

where γ refers to the channel noise factor of a MOSFET.

So, in order to calculate the input-referred noise of the TIA, considering (11), (12), and (14), it can be written as follows:

$$\overline{I_{n,in}^2} = \frac{4KT}{R_f}\left[1 + \frac{\gamma}{R_f.|g_{m2}|^2}\left(g_{m2}+g_{m3}+g_{m5} + \frac{(g_{m5}.g_{m1})^2}{g_{m6}}\right)\right] \tag{15}$$

As (15) suggests, by increasing the transconductance of g_{m2}, it is possible to decrease the input referred noise of the proposed FM TIA. Additionally, the input referred noise and the output noise of the proposed TIA are shown in Fig. 14 and Fig. 15, respectively. As Fig. 14 shows, the input referred noise at low frequencies is equal to 10pA/√Hz, and reaches the value of 11.1pA/√Hz at -3dB

frequency. Also, the total input referred noise current of the proposed FM TIA is equal to 10.4pA/√Hz (664nA$_{rms}$).

As it was discussed before, the employed feedback resistor decreases the thermal noise of the FM TIA. Table 4 compares the value of the input referred noise of the proposed FM TIA with and without the feedback network. As Table 4 reveals, the feedback network considerably decreases the thermal noise current of the FM-TIA.

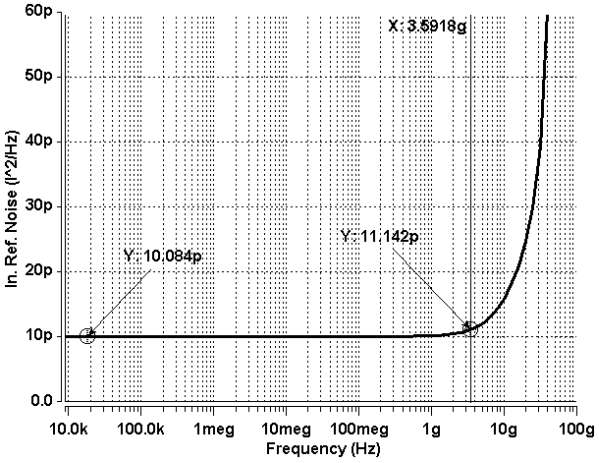


Fig. 14: Input referred noise of the FM TIA.

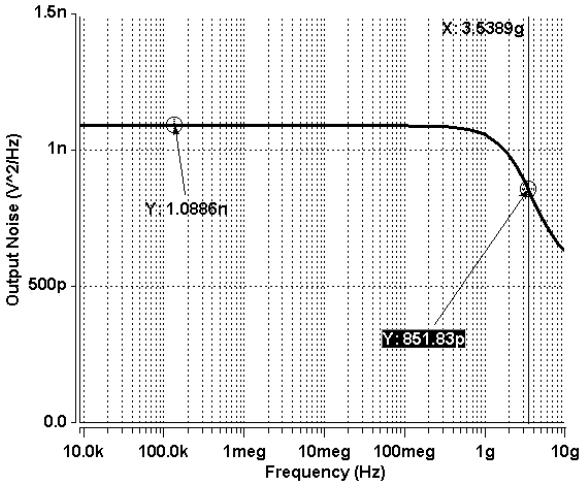


Fig. 15: Output noise of the FM TIA.

Table 4: Effect of the feedback network on the input referred noise current density

	The Open-loop TIA	The Closed-loop TIA
Input referred noise current density	664nA$_{rms}$	9.5mA$_{rms}$

Table 5: Performance comparison among the proposed TIA and other reported designs

	[14]	[15]	[22]	[23]	[24]	[28]	[29]	[30]	[31]	This Work
Year	2017	2016	2021	2016	2021	2015	2016	2017	2016	
Technology (CMOS)	0.18µm	0.18µm	90nm	0.13µm	90nm	0.13µm	0.13µm SiGe BiCMOS	0.13µm SiGe BiCMOS	0.18µm	90nm
Gain(dBΩ)	59	58	41	54	42.3	50.1	72	83.7	55-69	40.6
Bandwidth (GHz)	7.9	8.1	6.5	11.5	5	7	38.4	32.1	1	3.55
Power Consumption (W)	18m	34.8m	1.67m	45m	2.7m	7.5m	261m	150m	6m	315µ
Cpd (fF)	300	300	250	-	250	250	-	-	-	220
Supply Voltage (V)	1.8	1.8	1	1.5	1	1.5	3.3	3.3	1.8	1
Input referred noise(pA/√Hz)	23	15	33.4	6.8	32.5	31.3	14.8	-	9.33	10.4
No. of passive inductors	2	2	0	2	0	0	0	0	0	0
FoM1	425	184.8	436	128	167	299	585	3276	417	1206
FoM2	5.54	3.69	3.25	-	1.3	2.4	-	-	-	25.5
Area	0.11 mm^2	-	-	0.048 mm^2	312 µm^2	16200 µm^2	-	2.345 mm^2	7500 µm^2	98 µm^2
Work*	Sim	Sim	Sim	Exp	Sim	Exp	Exp	Exp	Sim	Sim

* Sim and Exp refer to experimental and simulation results, respectively.

Moreover, Fig. 16 demonstrates the layout of the FM TIA.

As the proposed circuit contains only six transistors in the feedforward network, and a small resistance equal to 50Ω as the feedback network, the occupied chip area of the proposed TIA is only 98µm^2, which is a small area.

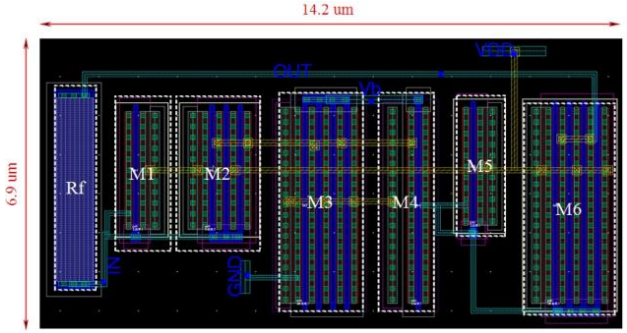


Fig. 16: The circuit layout of the proposed FM TIA.

Table 5 provides a summary of performance and compares the parameters of the proposed TIA circuit with other reported designs. The power consumption value of the proposed TIA is shown to be significantly less than other reported designs. However, in order to provide a fair comparison, two Fig.s of Merit (FOMs) are defined in Table 5, as follows:

$$FOM1 = \frac{Gain \times B.W.}{P_{DC}} \left(\frac{\Omega.GHz}{mW}\right) \quad (16)$$

$$FOM2 = \frac{Gain \times B.W. \times C_{in}}{P_{DC} \times In.Ref.Noise} \left(\frac{\Omega.GHz.pF}{mW.\left(\frac{pA}{\sqrt{Hz}}\right)}\right) \quad (17)$$

Conclusion

In this paper, a trans-impedance amplifier based on a combination of current-mirror topology and folded-cascade topology is presented, which amplifies the current signal and converts it to the voltage at the output node.

Due to the existence of a diode-connected transistor at the input node, the input resistance of the TIA is comparatively small.

Furthermore, four out of six transistors are PMOS transistors, which represent less thermal noise in comparison with NMOS transistors. Also, the proposed Folded-Mirror topology occupies a relatively small area on-chip, due to the fact that no passive element is used in the feedforward network.

Results using 90nm CMOS technology parameters show 40.6dBΩ trans-impedance gain, 3.55GHz frequency bandwidth, 664nArms input-referred noise and only 315µW power dissipation is using 1volt supply, which indicates the proper performance of the proposed circuit as a low-power building block.

Author Contributions

Authors have had an equal contribution in the problem and data analysis, interpreting the results and writing the manuscript.

Acknowledgement

The authors gratefully thank the anonymous reviewers and the editor of JECEI for their useful comments and suggestions.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

TIA	Trans-impedance Amplifier
FM	Folded-Mirror
RGC	Regulated Cascode
PMOS	Positive Metal-Oxide Semiconductor
NMOS	Negative Metal Oxide Semiconductor
CMOS	Complementary metal–oxide–semiconductor
NRZ	Non-Return to Zero
PRBS	Pseudo-Random Bit Sequence
MOSFET	metal-oxide semiconductor field-effect transistor

References

[1] B. Razavi, Integrated Circuit for Optical Communications, Second edition, New York John Wiley & Sons Inc, New Jersey, 2012.

[2] K. Schneider, H. Zimmermann, Highly Sensitive Optical Receivers, Springer Series in advanced Microelectronics, Netherland, 2006.

[3] S. Zohoori, M. Dolatshahi, "A Low-power, CMOS transimpedance amplifier in 90-nm technology for 5-Gbps optical communication applications," Int. J. Circuit Theory Appl., 46: 1-14, 2018.

[4] D. Li, G. Minoia, M. Repossi, D. Baldi & etc., "A Low-noise design technique for High-speed CMOS optical receivers," IEEE J. Solid-State Circuits, 49: 1437-1446, 2014.

[5] S. Zohoori, M. Dolatshahi, "A CMOS Low-power optical front-end for 5Gbps applications," Fiber Integr. Opt., 37: 37-56, 2018.

[6] M. Atef, H. Zimmermann, "Optical receiver using noise cancelling with an integrated photodiode in 40nm CMOS technology," IEEE Trans. Circuits Syst. I: Regul. Pap., 60: 1929-1936, 2013.

[7] S. Zohoori, M. Dolatshahi, M. Pourahmadi, M. Hajisafari, "A CMOS, Low-power current-mirror-based transimpedance amplifier for 10Gbps optical communications," Microelectronics J., 80: 18-27, 2018.

[8] Y. H. Chien, K. L. Fu, Sh. I. Liu, "A 3-25 Gb/s four-channel receiver with noise-cancelling TIA and power-scalable LA," IEEE Trans. Circuits Syst., 61: 845-850, 2014.

[9] B. Nakhkoob, M. Mostafa Hella, "A 5-Gb/s noise optimized receiver using a switched TIA for wireless optical communications," IEEE Trans. Circuits Syst. I: Regul. Pap., 61: 1255-1268, 2014.

[10] S. Zohoori, M. Dolatshahi, "An inductor-less, 10Gbps trans-impedance amplifier operating at low supply-voltage," in Proc. 25th Iranian conference on electrical Engineering (ICEE2017), Tehran, Iran, 2017.

[11] Zh. Lu, K. S. Yeo, W. M. Lim, M. A. Do, Ch. CH. Boon, "Design of a CMOS broadband transimpedance amplifier with active feedback," IEEE Trans. Very Large Scale Integr. VLSI Syst., 18: 461-472, 2010.

[12] X. Zhi-gang, CH. Ying-mei, W. Tao, Ch. Xue-hui, ZH. Li, "A 40 Gbit/s fully integrated optical receiver analog front-end in 90nm CMOS," J. China Univ. Posts Telecommun., 19: 124-128, 2012.

[13] M. Seifouri, P. Amiri, M. Rakide, "Design of broadband transimpedance amplifier for optical communication systems," Microelectronics J., 46: 679-684, 2015.

[14] M. Seifouri, P. Amiri, I. Dadras, "A transimpedance amplifier for optical communication network based on active voltage-current feedback," Microelectronics J., 67: 25-31, 2017.

[15] M. Rakide, M. Seifouri, P. Amiri, "A folded cascade-based broadband transimpedance amplifier for optical communication systems," Microelectronics J., 54: 1-8, 2016.

[16] L. Liu, J. Zou, N. Ma, Zh. Zhu, Y. Yang, "A CMOS transimpedance amplifier with high gain and wide dynamic range for optical sensing system," Optik, 126: 1389-1393, 2015.

[17] M. Atef, "Transimpedance amplifier with a compression stage for wide dynamic range optical applications," Microlelectronics J., 46: 593-597, 2015.

[18] F. Aznar, W. Gaberl, H. Zimmermann, "A 0.18um CMOS transimpedance amplifier with 26 dB dynamic range at 2.5Gb/s," Microelectronics J., 42: 1136-1142, 2011.

[19] M. Atef, F. Aznar, S. Schidl, A. Polzer, W. gaberl, H. Zimmermann, "8 Gbit/s inductorless transimpedance amplifier in 90 nm CMOS Technology," Analog Integr. Circuits Signal Process., 79: 27-36, 2014.

[20] M. Atef, H. Zimmermann, "Low-power 10Gb.s Inductorless inverter based common-drain active feedback transimpedance amplifier in 40nm CMOS," Analog Integr. Circuits Signal Process., 76: 367-376, 2013.

[21] S. Zohoori, M. Dolatshahi, M. Pourahmadi, M. Hajisafari, "An inverter-based, CMOS, Low power Optical Receiver Front-End," Fiber Integr. Opt, 38: 1-19, 2019.

[22] S. Honarmand, M. Pourahmadi, M. R. Shayesteh, K. Abbasi, "Design of an inverter-base, active-feedback, low-power transimpedance amplifier operating at 10 Gbps," J. Circuits Syst. Comput., 30(06): 2150110, 2021.

[23] P. Andre, S. Jacobus, "Design of a high gain and power efficient optical receiver front-end in 0.13µm RF CMOS technology for 10Gbps applications," Microw. Opt. Technol. Lett., 58: 1499–1504, 2016.

[24] S. Honarmand, M. Pourahmadi, M. R. Shayesteh, K. Abbasi, "A multi-stage TIA based on cascoded-inverter structures for low-power applications," J. Integr. Circuits Syst., 16(3): 1-12, 2021.

[25] C. Toumazou, S. M. Park, "Wideband low noise CMOS transimpedance amplifier for gigahertz operation," Electron. Lett, 32: 1194-1196, 1996.

[26] S. M. Rezaul Hasa, "Design of a low-power 3.5GHz broadband CMOS transimpedance amplifier for optical transceivers," IEEE Trans. Circuits Syst. I Regul. Pap., 52: 1061-1072, 2005.

[27] B. Razavi, Design of Analog CMOS Integrated Circuits, McGraw-Hill, Singapore, 2001.

[28] M. H. Taghavi, L. Belostotski, J.W. Haslett, P. Ahmadi, "10-Gb/s 0.13-µm CMOS inductor less modified-RGC transimpedance amplifier," IEEE Trans. Circuits Syst. I: Regul. Pap., 62:, 1971–1980, 2015.

[29] K. Honda, H. Katsurai, M. Nada, "A 56-Gb/s transimpedance amplifier in 0.13-µm SiGe BiCMOS for an optical receiver with −18.8dBm input sensitivity," in Proc. the IEEE Compound Semiconductor Integrated Circuit Symposium (CSICS) Austin, TX, USA, 2016.

[30] Y. Chen, J. Li, Z. Zhang, H. Wang, Y. Zhang, "12-Channel, 480 Gbit/s optical receiver analogue front-end in 0.13µm BiCMOS technology," Electron. Lett., 53: 492–494, 2017.

[31] R. Y. Chen, Z.Y. Yang, "CMOS transimpedance amplifier for gigabit-per-second optical wireless communications," IEEE Trans. Circuits Syst. II, 63: 418–422, 2016.

[32] G. Royo, C. Sanchez-Azqueta, A. D. Martinez Perez, C. Aldea, S. Celma, "Fully differential transimpedance amplifier for reliable wireless communications," Microelectron. Reliab., 83: 25-28, 2018.

[33] K. Monfared, Y. Belghisazar, "Improved low voltage low power recycling folded fully differential cascode amplifier," Tabriz J. Electr. Eng., 48: 327-334, 2018.

[34] P. Amiri, M. Seifouri, B. Afarin, A. Hedayati Pour, "Design of RGC preamplifier with bandwidth 20GHz and transimpedance 60 dBΩ for telecommunication systems," Tabriz J. Electr. Eng., 46: 15–23, 2016.

[35] M. Seifoui, P. Amiri, I. Dadras, "An electronic transimpedance amplifier for optical communications network based on active voltage-current feedback," Tabriz J. Electr. Eng., 48: 737-744, 2018.

Biographies

Sahar Sadeghi received she B.Sc and M.Sc degrees in Electrical Engineering from Islamic Azad University yazd branch, Yazd, Iran. Currently, She is a Ph.D Student in Electrical Engineering, Department of Electrical Engineering, Yazd Branch, Islamic Azad University, Yazd, Iran. She has presented numerous articles in the National and International confrencess and published an article in a reputed journal.

- Email: s. sadeghi@iauyazd.ac.ir
- ORCID: 0000-0002-5015-5348
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Maryam Nayeri received the Ph.D. degree in electronics from the Science and Research Branch, Islamic Azad University, Tehran, Iran, in 2016. She is currently with the Department of Electrical Enginneering, Yazd Branch, Islamic Azad University, Yazd, Iran.

- Email: nayeri@iauyazd.ac.ir
- ORCID: 0000-0003-0479-2431
- Web of Science Researcher ID: AAO-8626-2021
- Scopus Author ID: 35810875000
- Homepage: NA

Mehdi Dolatshahi was born in Isfahan, Iran in 1980. He received the B.Sc and M.Sc degrees in Electrical Engineering in 2003, 2006 respectively. He received Ph.D degree in Electrical Engineering in 2012 from Science and Research Branch, Islamic Azad University, Tehran, Iran. He has been with the Department of Electrical Engineering of Najafabad Branch, Islamic Azad University , since 2006 where he is currently an assistant professor. His research interests include VLSI and CMOS low-voltage, low-power analog and mixed-signal integrated circuit design and optimization as well as CMOS optical communications circuit design.

- Email: Dolatshahi@iaun.ac.ir
- ORCID: 0000-0002-5948-7277
- Web of Science Researcher ID: NA
- Scopus Author ID: 53063559900
- Homepage: http://research.iaun.ac.ir/pd/dolatshahi

Ali Moftakharzadeh was born in Yazd, Iran in 1981. He received the B.Sc. degree in Electrical Engineering from K. N. Toosi University of Technology, Tehran, Iran, in 2002 and the M.Sc. and Ph.D. degrees in Electrical Engineering from Sharif University of Technology, Tehran, Iran, in 2004 and 2009, respectively. In 2010, he joined the Department of Electrical Engineering, Yazd University, Yazd, Iran. His current research interests include VLSI, digital system design, DSP, Image processing, and linear/nonlinear circuits macromodeling.

- Email: moftakharzadeh@yazd.ac.ir
- ORCID: 0000-0002-5634-6815
- Web of Science Researcher ID: NA
- Scopus Author ID: 24071287500
- Homepage: https://yazd.ac.ir/people/moftakharzadeh

How to cite this paper:
S. Sadeghi, M. Nayeri, M. Dolatshahi, A. Moftakharzadeh, "Novel ultra-low-power mirrored folded-cascade transimpedance amplifie," J. Electr. Comput. Eng. Innovations, 11(1): 217-228, 2023.

DOI: 10.22061/jecei.2022.9015.568

URL: https://jecei.sru.ac.ir/article_1785.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Robust Linear Parameter Varying Fault Reconstruction of Wind Turbine Pitch Actuator Using Second-Order Sliding Mode Observer

M. Mousavi[1], M. Ayati[2,], M. Hairi-Yazdi[2], S. Siahpour[3]*

[1]*Department of Mechanical Engineering, Binghamton University, Binghamton 13902, USA.*

[2]*School of Mechanical Engineering, College of Engineering, University of Tehran, Tehran, Iran.*

[3]*Department of Mechanical Engineering, University of Cincinnati, Cincinnati 45221, USA.*

Article Info

Article History:
Received 15 June 2022
Reviewed 29 July 2022
Revised 19 August 2022
Accepted 26 September 2022

Keywords:
Wind turbines
Pitch actuator faults
Linear parameter varying model
Sliding mode observer
Fault reconstruction

[*]Corresponding Author's Email Address: *m.ayati@ut.ac.ir*

Abstract

Background and Objectives: In this paper, a novel linear parameter varying (LPV) model of a wind turbine is developed based on a benchmark model presented by Aalborg University and KK-electronic a/c. The observability and validity of the model are investigated using real aerodynamic data.
Methods: In addition, a robust fault detection and reconstruction method for linear parameter varying systems using second-order sliding mode observer is developed and implemented on the linear parameter varying model. The fault signal is reconstructed using a nonlinear term named equivalent output error injection during sliding motion and a proper transformation. The effect of uncertainties and incorrect measurements are minimized by employing an oriented method that requires solving a nonlinear matrix inequality. During numerical simulations, an actuator fault in the pitch system is considered and the performance of the method in fault reconstruction is investigated.
Results: Wind speed range is considered from 14 m/s to 16 m/s and it is regarded as a stochastic input exerting aerodynamic torque. Fast and accurate fault reconstruction happens in 0.6 seconds with less than one percent error. The observer performance is not affected by the fault and fault is estimated in 2.5 seconds with an error smaller than 2.48 percent.
Conclusion: Results illustrate fast and accurate fault reconstruction and accurate state estimations in the presence of actuator fault.
In this paper, a novel linear parameter varying (LPV) model of a wind turbine is developed based on a benchmark model presented by Aalborg University and KK-electronic a/c. The observability and validity of the model are investigated using real aerodynamic data. In addition, a robust fault detection and reconstruction method for linear parameter varying systems using a second-order sliding mode observer is developed and implemented on the linear parameter-varying model. The fault signal is reconstructed using a nonlinear term named equivalent output error injection during sliding motion and a proper transformation. The effect of uncertainties and incorrect measurements are minimized by employing an H_∞ oriented method which requires solving a nonlinear matrix inequality. During numerical simulations, an actuator fault in the pitch system is considered, and the performance of the method in fault reconstruction is investigated. Wind speed range is considered from 14 m/s to 16 m/s and it is regarded as a stochastic input exerting aerodynamic torque. Fast and accurate fault reconstruction happens in 0.6 seconds with less than one percent error. The observer performance is not affected by the fault and fault is estimated in 2.5 seconds with an error smaller than 2.48 percent results illustrate fast and accurate fault reconstruction and accurate state estimations in the presence of actuator fault.

Introduction

The costs of wind turbines consist of two parts of implementation and maintenance. The maintenance of large wind turbines is a time-consuming process [1], and a costly procedure, especially in offshore wind farms. It requires the generator disconnection from the power distribution network. Therefore, designing and utilizing a fault detection and isolation (FDI) system to diagnose, isolate, and reconstruct wind turbine faults is highly beneficial and critical in supervisory and maintenance cost reduction. In addition, it increases the lifetime of the turbine components and enhances the power generation due to the fault accommodation and active fault-tolerant control in which a reconfigurable controller is employed to accommodate the effect of faults [2]-[6].

The faults occurring in a large wind turbine are classified into three categories. Sensor faults which include rotor speed, generator speed, generator torque, and pitch angle and they, appear as biased output, random output, fixed output, or no output as discussed in [7]. Component faults such as drive train deficiency [8], mass imbalance of the rotor [9], and the generator system [10] are included in the second class. The third category of wind turbine faults aims at the actuator faults, such as the pitch actuator fault, which is addressed in this paper. The pitch system is responsible for the adjustment of the pitch angle of the rotor blades for the variable-pitch wind turbines. Such systems are important in terms of the amount of wind power captured by blades.

Two types of pitch control systems are used in variable-pitch wind turbines. In the first type, three individual electrical motors are implemented. This is beneficial for the fast reaction of the turbine to wind speed changes and power demand. The second type consists of three individual hydraulic pumps, which are slower but bear more stiffness and have smaller backlash. Therefore, considering large wind turbines, a hydraulic pitch system is suggested for higher reliability. Pitch actuator faults occur for three reasons such as high air content of oil, pump wear, and hydraulic leakage. Hydraulic leakage is an incipient fault and occurs faster compared to the other faults. Thus, it should be considered to reduce cost and energy consumption, decrease operational load, increase power harvesting, and avoid stalling [8], [11]-[13], [37], [38].

The rate of occurrence and the values of faulty and healthy properties are shown in the corresponding columns of Table 1. The state of $\theta = 0$ represents proper situation and $\theta = 1$ is fully faulty operation [7]. In the case of hydraulic fault incidence in each of the individual pitch systems, control efforts may lead to two decisions: (1) generator power exceeds the nominal value (2) output power is reduced, which results in power efficiency reduction. As a result of the leakage in the pitch system, the actuation of the pitch angle becomes slower, and smaller wind power is captured. As a result, fault detection, reconstruction, and fault accommodation are useful decisions to reinforce the control system in a way that energy-related cost functions are satisfied [2]-[6]. Many research projects such as the current work have been conducted for this issue to improve the estimation speed and accuracy of the observer-based fault diagnosis methods [2], [3], [6].

Table 1 Rate of incidence and values of faulty and healthy properties in pitch hydraulic system

	Faulty operation	Rate of fault incidence
No-fault	$\omega_n = 11.11\ rad/s, \zeta = 0.6$	
High air content	$\omega_n = 5.73\ rad/s, \zeta = 0.45$	$\lvert\dot{\theta}\rvert \approx 1/month$
Pump wear	$\omega_n = 7.27\ rad/s, \zeta = 0.75$	$\lvert\dot{\theta}\rvert \approx 1/(20\ years)$
Hydraulic leakage	$\omega_n = 3.42\ rad/s, \zeta = 0.9$	$\lvert\dot{\theta}\rvert \approx 1/(100\ seconds)$

The wind turbine benchmark considered in this paper is developed by Aalborg University and KK-electronic a/c, enabling the simulation of various sensor and actuator faults [14]. This model is nonlinear due to the relation of wind and aerodynamic torque exerting on wind turbine blades. This kind of nonlinearity has been handled in different methods. Linearizing around one or several operating points and switching among them (gain-scheduling control) [15] is one of these methods. In this method, several observers are designed in which for reducing switching effects, bumpless switching between models should be considered. Linear parameter varying (LPV) modelling is another method where nonlinear terms are turned into linear but time-varying parameters (quasi-linear) [16]-[18]. In such methods, nonlinear terms are expressed in LPV form. Generally, LPV models yield higher accuracy for the all operating points. Using LPV models leads to LPV observer design; thus, the advantages of linear system characteristics could be utilized.

In this paper, fault detection and reconstruction are covered. An LPV model of the wind turbine is developed and a model-based robust second-order sliding mode observer is applied to the LPV wind turbine model. LPV model is valid in the entire operating trajectory and does not require linearization around one or several operating point(s) [19]-[24]. Once the observer gains are obtained, the observer and fault reconstruction formula are

attainable for all the wind turbine operating regions. Actually, LPV methods attempt to parametrize the model closer to real world at the cost of larger computational effort and complication. Model-based methods are preferable in fault detection and reconstruction studies where physical components' parameters of the plant are accessible. Some surveys in model-based wind turbine FDI have been carried out in [22], [24]-[32].

The proposed observer of this paper includes an LTI gain for linear output error signal and an LPV gain for nonlinear residual signals. The reconstructed actuator fault is generated once the sliding motion takes place using a nonlinear residual signal called "equivalent output injection". Observer design matrices are obtained using H_∞ concepts and solving a nonlinear matrix inequality in which the effect of uncertain and imperfect measurements is minimized.

This paper is structured as follows. Section 2 describes the wind turbine benchmark model. Section 3 presents the development of the methodology and observer design procedure. Then, the pitch actuator fault description is presented in Section 4. Section 5 explains the LPV system description and Section 6 is dedicated to numerical results and energy analysis. Finally, Section 7 is the conclusion.

Wind turbine benchmark model

An overview of the wind turbine model in the benchmark developed by Aalborg University and KK-electronic a/c [14] is illustrated in Fig. 1. The variables are introduced in the following subsections.

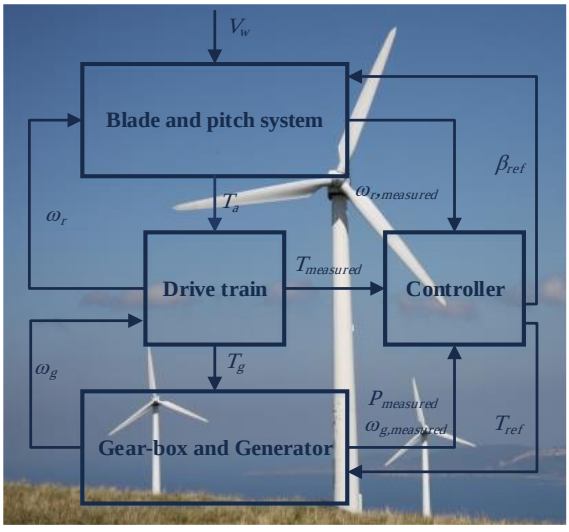


Fig. 1: Wind turbine benchmark developed by Aalborg University and KK-electronic a/c (background photo [31]).

Using aerodynamic principles, the correlation between the wind speed, rotor speed, blades' pitch angle, and the aerodynamic torque exerting the blades is shown in (1). The mentioned correlation is derived considering two assumptions.

1. The wind speed is constant all over the surface of the blades.

2. The wind speed is perpendicular to the rotor plane.

ρ_{air}, R, $\beta(t)$, $\omega_r(t)$, and $V_w(t)$ are air density, blade radius, pitch angle, rotor speed, and average wind speed, respectively. $\lambda(t)$ is the tip speed ratio which is defined in (2). The aerodynamic torque is approximated in (1) using a factor named aerodynamic torque coefficient $C_q(\lambda(t),\beta(t))$.

$$T_a = \frac{\rho_{air}\,\pi R^3 C_q(\lambda,\beta) V_w^2}{2} \tag{1}$$

$$\lambda(t) = \frac{R\omega_r(t)}{V_w(t)} \tag{2}$$

The pitch system consists of three identical hydraulic pumps as the actuator for adjusting the blades' angle by rotation. Three internal controllers are adopted for each actuator giving proper input signals to the actuators. In addition, a second-order transfer function is considered for each of the pitch actuators correlating control input (β_{ref}) to the pitch angle (β). The damping ratio and natural frequency of this model are $\zeta(t)$ and $\omega_n(t)$, respectively. These properties might be time-varying in the event of faults for each system. β_{ref} is pitch control input signal entering each pitch actuator.

$$\begin{aligned}\ddot{\beta}(t) = &-2\zeta(t)\omega_n(t)\dot{\beta}(t) + \omega_n^2(t)\beta_{ref} \\ &- \omega_n^2(t)\beta(t)\end{aligned} \tag{3}$$

The drive-train of the wind turbine consists of two shafts as the low-speed (driver) and the high-speed shaft (driven). The shafts are connected using a gear-box and the aerodynamic power is transferred to the generator through a high-speed shaft. The coupled dynamic equations of the shafts which are considered as a mass-spring model are expressed in (4), (5).

$$\begin{aligned}J_r\,\dot{\omega}_r(t) = &\,T_a(\omega_r,\beta,V_w,t) - K_{dt}\,\theta(t) \\ &- (B_{dt} + B_r)\omega_r(t) \\ &+ \frac{B_{dt}}{N_g}\omega_g(t)\end{aligned} \tag{4}$$

$$\begin{aligned}J_g\dot{\omega}_g\,(t) = &\frac{\eta_{dt}K_{dt}}{N_g}\theta(t) + \frac{\eta_{dt}B_{dt}}{N_g}\omega_r(t) \\ &- \left(\frac{\eta_{dt}B_{dt}}{N_g^2} + B_g\right)\omega_g(t) \\ &- T_g\end{aligned} \tag{5}$$

$$\dot{\theta}(t) = \omega_r(t) - \frac{1}{N_g}\omega_g(t) \tag{6}$$

where, ω_r, ω_g are the rotor and generator speeds, respectively, $\theta(t)$ is the torsion angle of the drive train, T_g is generator torque, J_r and J_g are rotor and generator moments of inertia, K_{dt} and B_{dt} are torsional stiffness

and damping, B_r and B_g are the rotor and generator viscous friction, N_g is the gear ratio and η_{dt} is the efficiency of drive-train. The generator and converter subsystem are modeled by first-order transfer functions:

$$\frac{T_g(s)}{T_{g,ref}(s)} = \frac{\alpha_{gc}}{s + \alpha_{gc}} \tag{7}$$

where, α_{gc} is the generator and converter model parameter and $T_{g,ref}$ is the control output signal of the converter. Parameter values of the benchmark are: $J_r = 55e6$ kg. m^2, $K_{dt} = 2.7e9$ m/rad, $B_g = 3.034$ N. m. s/rad, $N_g = 95$, $\eta_{dt} = 0.92$, $R = 57.5$ m, $J_g = 390$ kg. m^2, $\rho_{air} = 1.225$ kg/m^3, $B_r = 27.8$ kNm/(rad/s), $B_{dt} = 945$ kN. m/(rad/s).

Observer Design

Nowadays, condition monitoring is attracting more attention in technology advancements. It is implemented to prevent serious failures by detecting faults. Condition monitoring in advanced engineering instruments analyses the deviations from normal conditions and detects the existence of faults and failures. On the other hand, Fault reconstruction is an online fault detection method that offers additional information about the size, location, and severity of faults. Such data are useful in the choice of proper action during faulty conditions. Moreover, control efforts are configured considering reconstructed fault data (fault accommodation) to provide better performances during faulty conditions.

A. LPV System Description

An uncertain LPV plant that is subjected to actuator faults is described by

$$\begin{aligned} \dot{x}(t) &= A(\boldsymbol{\rho})x(t) + B(\boldsymbol{\rho})u(t) + M(\boldsymbol{\rho})f_i(t) \\ &\quad + Q\xi(\boldsymbol{\rho}, x, t) \\ y(t) &= Cx(t) + \vartheta(t) \end{aligned} \tag{8}$$

where, $A(\boldsymbol{\rho}) \in \boldsymbol{R}^{n\times n}$, $B(\boldsymbol{\rho}) \in \boldsymbol{R}^{n\times m}$, $M(\boldsymbol{\rho}) \in \boldsymbol{R}^{n\times s}$ are the linear parameter varying matrices of the model. $x(t) \in \boldsymbol{R}^n$, $(t) \in \boldsymbol{R}^m$, $f_i(t) \in \boldsymbol{R}^s$, $\xi(\boldsymbol{\rho}, x, t) \in \boldsymbol{R}^k$, and $y(t), \vartheta(t) \in \boldsymbol{R}^p$, are model states, control input signal, faults of actuators, model uncertainty, and measurement noises, respectively. s and p are the lengths of fault and output signal. It is supposed that s is smaller than p ($s < p$). $\boldsymbol{\rho}$ is the varying parameter vector and is measured or estimated. C is an LTI and full-rank matrix. Also, for Lyapunov stability [32], $\xi(\boldsymbol{\rho}, x, t)$ and $\dot{\xi}(\boldsymbol{\rho}, x, t)$ are assumed to be bounded.

Assumption 1. The actuator fault matrix might be parameter varying ($M(\boldsymbol{\rho})$). It is assumed $M(\boldsymbol{\rho})$ is made up of a parameter invariant matrix (M_{inv}) multiplied by a nonsingular parameter varying matrix ($M_{var}(\boldsymbol{\rho})$) in a way that

$$M(\boldsymbol{\rho}) = M_{inv}M_{var}(\boldsymbol{\rho}) \tag{9}$$

where, $M_{inv} \in \boldsymbol{R}^{n\times s}$ and $M_{var}(\boldsymbol{\rho}) \in \boldsymbol{R}^{s\times s}$. Defining a new variable $\sigma(\boldsymbol{\rho}, t) = M_{var}(\boldsymbol{\rho})f_i(t)$, (8) is rewritten in the form of

$$\begin{aligned} \dot{x}(t) &= A(\boldsymbol{\rho})x(t) + B(\boldsymbol{\rho})u(t) \\ &\quad + M_{inv}\,\sigma(\boldsymbol{\rho}, t) \\ &\quad + Q\xi(\boldsymbol{\rho}, x, t) \\ y(t) &= Cx(t) + \vartheta(t) \end{aligned} \tag{10}$$

$\sigma(\boldsymbol{\rho}, t)$ is the new fault vector which will be converted to the actual fault vector ($f_i(t)$) after being estimated. The new fault signal is bounded due to the Lyapunov stability proof [32].

Assumption 2. $\vartheta(t)$ presents the corruption of sensor measurements and is assumed

$$\vartheta(s) = D(s)\varphi(s) \tag{11}$$

$$D(s) = \frac{a_f}{s + a_f} \tag{12}$$

$D(s)$ is a stable transfer function and $\varphi(t)$ is an unknown but bounded signal [30]. Using this assumption, the effect of output noises on estimations is optimized (this will be discussed later). Then, substituting (11) into (12) yields

$$\dot{\vartheta}(t) = -a_f\vartheta(t) + a_f\varphi(t) \tag{13}$$

Assumption 3. $rank(CM) = s$. This condition determines whether the effect of the fault signal is observable in outputs or not. This is a necessary condition for the fault estimation method presented by Tan and Edwards [33].

B. Second-order LPV Sliding Mode Observer

The LPV sliding mode observer is in the form of

$$\begin{aligned} \dot{\hat{x}}(t) &= A(\boldsymbol{\rho})\hat{x}(t) + B(\boldsymbol{\rho})u(t) \\ &\quad + H_{eq}(\boldsymbol{\rho})e_y(t) \\ &\quad + H_{sw}w(t) \\ \hat{y}(t) &= C\hat{x}(t) \end{aligned} \tag{14}$$

where, $H_{eq}(\boldsymbol{\rho})$ and H_{sw} are observer design matrices and $w(t)$ represents discontinuous output error injection to induce a sliding motion [33]. $e(t)$ and $e_y(t)$ as state estimation error and output estimation error are expressed as

$$e(t) = \hat{x}(t) - x(t) \tag{15}$$

$$e_y(t) = \hat{y}(t) - y(t) = Ce(t) - \vartheta(t) \tag{16}$$

The design steps are expressed in the following. First, the system is transformed in a way that the states are classified into measured states (outputs) and unmeasured states [34]. It is proved that the measured states are estimated in a finite time defining the sliding surface as $S = \{e(t) \in R^n : e_y(t) = 0\}$ [34]. Then, with the appropriate choice of parameters, unmeasured states are estimated asymptotically. Finally, the faults are reconstructed using $w(t)$.

As stated before, sliding mode observer gains are divided into an equivalent gain for linear terms and switching motion gain for nonlinear terms. The equivalent gain ($H_{eq}(\boldsymbol{\rho})$) and its corresponding linear signal are existed to force the incidence of the sliding motion. Such an action is called the reaching phase [27].

Furthermore, the switching gain (H_{sw}) and its corresponding nonlinear signal ($w(t)$) are responsible for the maintenance of sliding motion which is called the sliding phase.

There exists a coordinate-transformation $x_f(t) \rightarrow T_f x(t)$ which changes the output matrix to the form of $C_f = [0_{p\times(n-p)} \; T_{p\times p}]$ in which T is an orthogonal nonsingular matrix.

Here, the index 'f' refers to the system of $x_f(t)$. Also, for an invertible square matrix M_0, the fault matrix becomes in the form of (17).

$$M_{inv,f} = \begin{bmatrix} 0_{(n-p)\times s} \\ 0_{(p-s)\times s} \\ M_{0_{s\times s}} \end{bmatrix} \tag{17}$$

Then, the following structure is obtained after the first transformation.

$$y_f(t) = [0 \;\; T]\begin{bmatrix} x_{1,f} \\ x_{2,f} \end{bmatrix} + \vartheta(t) \tag{18}$$

$$A_f(\boldsymbol{\rho}) = \begin{bmatrix} A_{11,f} & A_{12,f} \\ A_{21,f} & A_{22,f} \end{bmatrix} \tag{19}$$

$$Q_f = \begin{bmatrix} Q_{1,f} \\ Q_{2,f} \end{bmatrix} \tag{20}$$

Rewriting (8) and using the structures of (17)-(20) yields

$$\begin{aligned} \begin{bmatrix} \dot{e}_{1,f} \\ \dot{e}_{2,f} \end{bmatrix} = A_f \begin{bmatrix} e_{1,f} \\ e_{2,f} \end{bmatrix} &- Q_f \xi(\boldsymbol{\rho}, x, t) \\ &- H_{eq,f}(\boldsymbol{\rho}) e_y(t) \\ &+ H_{sw,f} w(t) \\ &- \begin{bmatrix} 0_{(n-s)\times s} \\ M_{0_{s\times s}} \end{bmatrix} \sigma(\boldsymbol{\rho}, t) \end{aligned} \tag{21}$$

Using (8) and (21)

$$e_y(t) = T e_{2,f} - \vartheta(t)$$

Thus, if the sliding motion takes place, from the definition $e_y(t) = 0$ and then $T e_{2,f} = \vartheta(t)$. Using (21) and regarding T as an orthogonal matrix gives:

$$\begin{aligned} \dot{e}_{1,f}(t) &= A_{11,f} + A_{12,f} T^T \vartheta(t) - L T^T w_{eq}(t) \\ &\quad - Q_{1,f} \xi(\boldsymbol{\rho}, x, t) \end{aligned} \tag{22}$$

$$\begin{aligned} 0 &= T A_{21} e_{1,f}(t) - T Q_{2,f} \xi(\boldsymbol{\rho}, x, t) \\ &\quad + T A_{22} T^T \vartheta(t) - \dot{\vartheta}(t) \\ &\quad + w_{eq}(t) - T \begin{bmatrix} 0_{(p-s)\times s} \\ M_{0_{s\times s}} \end{bmatrix} \sigma(\boldsymbol{\rho}, t) \end{aligned} \tag{23}$$

$w_{eq}(t)$ is the equivalent output error injection i.e. the same as $w(t)$ after the sliding motion. As a definition, for a design matrix $Y \in \boldsymbol{R}^{s\times(p-s)}$ and the structure of $W = [\mathrm{Y} \quad M_0^{-1}]$ the new fault signal is reconstructed as

$$\hat{\sigma}(\boldsymbol{\rho}, t) = W T^T w_{eq}(t) \tag{24}$$

In the system of (21), the LTI observer gain is considered in the form of

$$H_{sw,f} = \begin{bmatrix} -L T^T \\ T^T \end{bmatrix} \tag{25}$$

L is the design matrix and is of the form

$$L = [Z \;\; 0_{(n-p)\times s}] \text{ and } Z \in \boldsymbol{R}^{(n-p)\times(p-s)} \tag{26}$$

Z improves the sliding motion incidence and is synthesized by solving some matrix inequalities (it will be discussed later). Joining measurement noises signal and $e_{1,f}(t)$ together as a new state vector, an assembled state-space is obtained.

$$\begin{aligned} e_a(t) &= A_a(\boldsymbol{\rho}) e_a(t) + B_a(\boldsymbol{\rho}) \xi_a(t) \\ \hat{\sigma}(\boldsymbol{\rho}, t) - \sigma(\boldsymbol{\rho}, t) &= C_a e_a(t) + F_a \xi_a(t) \end{aligned} \tag{27}$$

where,

$$\begin{aligned} e_a^T(t) &= \left[T^T \vartheta(t) \quad e_{1,f}(t)\right]^T, \\ \xi_a^T(t) &= [\xi(t) \quad T^T \varphi(t)]^T \\ A_a(\boldsymbol{\rho}) &= \begin{bmatrix} -a_f I_p & 0_{p\times(n-p)} \\ A_{12}(\boldsymbol{\rho}) + L A_{22}(\boldsymbol{\rho}) + a_f L & A_{11} + L A_{21}(\boldsymbol{\rho}) \end{bmatrix} \end{aligned} \tag{28}$$

$$B_a(\boldsymbol{\rho}) = \begin{bmatrix} 0_{p\times(n-p)} & -a_f I_p \\ -Q_{1,f} - L Q_{2,f} & -a_f L \end{bmatrix} \tag{29}$$

$$\begin{aligned} C_a(\boldsymbol{\rho}) \\ = [-(W A_{22}(\boldsymbol{\rho}) + a_f W) \quad -W A_{21}(\boldsymbol{\rho})] \end{aligned} \tag{30}$$

$$F_a = [W Q_{2,f} \quad a_f W] \tag{31}$$

The effect of uncertainty and measurement noises are minimized if there exist positive definite and symmetric matrices ${P_{af}}_{p\times p}$ and ${P_{11}}_{(n-p)\times(n-p)}$ such that the following matrix inequalities hold

$$\boldsymbol{X}(\boldsymbol{\rho}) = \begin{bmatrix} P_1 A_a + {A_a}^T P_1 & P_1 B_a \Delta & C_a^T \\ (B_a)^T P_1 & -\gamma I & (F_a)^T \\ C_a & F_a \Delta & -\gamma I \end{bmatrix} < 0 \tag{32}$$

$$P_1 = \begin{bmatrix} P_{11} & 0 \\ 0 & P_{af} \end{bmatrix} > 0 \tag{33}$$

where, $\Gamma > 0$ means Γ is a positive definite matrix; then, using bounded real lemma $\|\hat{\sigma}(\boldsymbol{\rho},t) - \sigma(\boldsymbol{\rho},t)\| < \gamma\|\xi_a(\boldsymbol{\rho},t)\|$. Also,

Remark 1. It should be mentioned that there exist parameter varying terms in the matrix inequality (32) which make it confusing to solve the matrix inequality to determine a unique minimized γ. Therefore, we are unable to obtain the second transformation and the design matrices and the range of variation of $\boldsymbol{\rho}$ has to be considered. As a result, the effects of uncertainty and measurement noises are almost minimized.

Once L is obtained, the observer gain $H_{sw,f}$ is calculated and reverted to a system of (10) using

$$H_{sw} = T_f^{-1} \times H_{sw,f} \tag{34}$$

After obtaining the design matrix L, a second transformation is applied to the system of (21) where C_f is reformed into $C_s = [0_{p\times(n-p)} \;\; I_p]$. 's' represents second coordinate transformation.

$$\begin{bmatrix} x_{1,s} \\ x_{2,s} \end{bmatrix} = T_s \begin{bmatrix} x_{1,f} \\ x_{2,f} \end{bmatrix}$$
$$\text{where } T_s = \begin{bmatrix} I_{(n-p)\times(n-p)} & L \\ 0_{p\times(n-p)} & T \end{bmatrix} \tag{35}$$

$$A_s(\boldsymbol{\rho}) = \begin{bmatrix} A_{11,s} & A_{12,s} \\ A_{21,s} & A_{22,s} \end{bmatrix} \tag{36}$$

$H_{eq}(\boldsymbol{\rho})$ in the second coordination is obtained online after the calculation of L in a parameter varying structure of

$$H_{eq,s}(\boldsymbol{\rho}) = \begin{bmatrix} A_{12,s}(\boldsymbol{\rho}) \\ A_{22,s}(\boldsymbol{\rho}) + k_2 I_p \end{bmatrix} \tag{37}$$

$H_{eq,s}(\boldsymbol{\rho})$ is reverted to the main coordinates (14) by

$$H_{eq}(\boldsymbol{\rho}) = T_f^{-1} \times T_s^{-1} \times H_{eq,s}(\boldsymbol{\rho}) \tag{38}$$

The notation of j'th component of a vector $\vec{V}$ is defined as an index, e.g. V_j. Then, the equivalent output injection signal is calculated cell by cell separately in

$$w_j(t) = -k_1 sign\left(e_{y,j}(t)\right)\left|e_{y,j}(t)\right|^{\frac{1}{2}} + z_j(t) \tag{39}$$

$$\dot{z}_j(t) = -k_3 sign\left(e_{y,j}(t)\right) - k_4 e_{y,j}(t) \tag{40}$$

for *j=1,2,...,p*

k_1, k_2, k_3, k_4 are design parameters in satisfying inequalities below. By choosing proper values of scalars k_1, k_2, k_3, k_4, second-order sliding motion takes place in a finite time and the fault reconstruction process begins where proof by the Lyapunov method is explained in [35].

$$k_1 > 2\sqrt{\epsilon}$$
$$k_2 > 0 \tag{41}$$
$$k_3 > \epsilon$$
$$k_4 > \frac{k_2^2[(k_1)^3 + 1.25(k_1)^2 + 2.5(k_3 - \epsilon)]}{k_1(k_3 - \epsilon)}$$

ϵ is the bound of fault incidence rate or $\left|\dot{f}_i(t)\right| < \epsilon$. Substituting (40) into (39), the output injection signal is obtained by

$$w_j(t) = -k_1 sign\left(e_{y,j}(t)\right)\left|e_{y,j}(t)\right|^{\frac{1}{2}} + \int \left[-k_3 sign\left(e_{y,j}(s)\right) - k_4 e_{y,j}(s)\right] ds \tag{42}$$

By substituting (9) into (24), the fault estimation signal becomes

$$\hat{f}_i(t) = M_{var}^{-1}(\boldsymbol{\rho})\hat{\sigma}(\boldsymbol{\rho},t) = M_{var}^{-1}(\boldsymbol{\rho})WT^T w(t) \tag{43}$$

It should be mentioned that fault reconstruction in (43) is obtainable if only the sliding motion takes place.

Remark 2. The estimation of the new fault signal is enhanced by exploiting a low-pass filter in the form of

$$F(s) = \frac{b}{s+b} \tag{44}$$

Such a filter lowers the high frequency of measurement noises to enhance the new fault estimation signal. Filter reduces the amplitude of noises and results in a smoother estimation signal.

Remark 3. The method explained in Section 3 does not require any redundant instruments. Usually, wind turbine sensors consist of three pitch angles, generator and rotor speed, generator torque, and effective wind speed [16]. Using the filtered sensors' data and a microcomputer, the proposed FDI algorithm could be implemented and states and pitch faults are calculated.

Observability of the wind turbine model is inspected which is baffling due to the LPV description of wind turbine plant (Section 5) [36], where using Simulink, we watched the rank of LPV observability matrix online. For all the wind turbine operating regions $rank(\boldsymbol{O}_{LPV}) = 6$; therefore, the system is observable and the states can be estimated utilizing a proper observer.

Pitch Actuator Fault Model

It is assumed that an identical performance and fault occur in all the pitch systems, and we only look through one system.

A second-order transfer function is assumed for each of the pitch actuators. Thus, the pitch angle and pitch rate are regarded as the system states.

The matching condition mentioned in assumption 3 does not hold unless a change of variables in dynamic equations of the pitch system is performed.

$$\begin{bmatrix}\beta(t)\\ \dot{\beta}'(t)\end{bmatrix} \to \bar{T}\begin{bmatrix}\beta(t)\\ \dot{\beta}(t)\end{bmatrix} \quad (45)$$

where, $\bar{T} = \begin{bmatrix}1 & 0\\ 0 & \frac{1}{\omega_n^2(t)}\end{bmatrix}$ and the pitch system equations are transformed to

$$\begin{bmatrix}\dot{\beta}(t)\\ \ddot{\beta}'(t)\end{bmatrix} = \begin{bmatrix}0 & \omega_n^2(t)\\ -1 & -2\zeta(t)\omega_n(t)\end{bmatrix}\begin{bmatrix}\beta(t)\\ \dot{\beta}'(t)\end{bmatrix} \quad (46)$$

Therefore, the assumption of $rank(CM_{inv}) = 1 = s \le p = 1$ still holds.

Hydraulic leakage affects the pitch actuator properties such as the natural frequency and damping ratio of each actuator. The system properties are $\omega_{n,h}$ and ζ_h as healthy and $\omega_{n,f}$ and ζ_f as a faulty situation. Then, considering the fault as changing properties in a linear fraction of both healthy and faulty mode, the incidence of the fault is modeled using two varying parameters $\theta_1$and θ_2. $\theta_1 = \theta_2 = 0$ means no fault in the system and $\theta_1 = \theta_2 = 1$ indicates a totally faulty situation.

$$\begin{aligned}\omega_n^2(t) &= \theta_1(t)\omega_{n,f}^2 + (1-\theta_1(t))\omega_{n,h}^2\\ &= \omega_{n,h}^2 + \theta_1(t)\left[\omega_{n,f}^2 - \omega_{n,h}^2\right]\end{aligned} \quad (47)$$

$$\begin{aligned}-2\zeta(t)\omega_n(t) &= -2\zeta_f\omega_{n,f}\theta_2(t)\\ &\quad + \left(1-\theta_2(t)\right)(-2\zeta_h\omega_{n,h})\\ &= -2\zeta_h\omega_{n,h}\\ &\quad + \theta_2(t)\left[-2\zeta_f\omega_{n,f}\right.\\ &\quad \left. + 2\zeta_h\omega_{n,h}\right]\end{aligned} \quad (48)$$

It is assumed that the set of $\{\omega_{n,h}, \zeta_h\}$ or the set of $\{\omega_{n,f}, \zeta_f\}$ occur, simultaneously. When the pitch system performance is normal, its properties are in a healthy situation. Malfunctioning of the pitch system means a faulty situation for both properties. Thus, $\theta_1(t) \approx \theta_2(t)$. Substituting (47) and (48) into (46), yields

$$\begin{aligned}&\begin{bmatrix}0 & \omega_n^2(t)\\ -1 & -2\zeta(t)\omega_n(t)\end{bmatrix}\\ &= \begin{bmatrix}0 & \omega_{n,h}^2 + \theta_1(t)\left(\omega_{n,f}^2 - \omega_{n,h}^2\right)\\ -1 & -2\zeta_h\omega_{n,h} + \theta_1(t)\left(-2\zeta_f\omega_{n,f} + 2\zeta_h\omega_{n}\right.\end{bmatrix}\\ &= \begin{bmatrix}0 & \omega_{n,h}^2\\ -1 & -2\zeta_h\omega_{n,h}\end{bmatrix}\\ &+ \begin{bmatrix}0 & \theta_1(t)\left(\omega_{n,f}^2 - \omega_{n,h}^2\right)\\ 0 & \theta_1(t)\left(-2\zeta_f\omega_{n,f} + 2\zeta_h\omega_{n,h}\right)\end{bmatrix}\end{aligned} \quad (49)$$

Multiplying (50) with the states gives,

$$\begin{aligned}&\begin{bmatrix}0 & \theta_1(t)\left(\omega_{n,f}^2 - \omega_{n,h}^2\right)\\ 0 & \theta_1(t)\left(-2\zeta_f\omega_{n,f} + 2\zeta_h\omega_{n,h}\right)\end{bmatrix}\begin{bmatrix}\beta(t)\\ \dot{\beta}'(t)\end{bmatrix}\\ &= \begin{bmatrix}\omega_{n,f}^2 - \omega_{n,h}^2\\ -2\zeta_f\omega_{n,f} + 2\zeta_h\omega_{n,h}\end{bmatrix}\dot{\beta}'(t)\theta_1(t)\\ &= M_{inv}M_{var}(\boldsymbol{\rho})f_i(t)\end{aligned} \quad (50)$$

where, $M_{inv} = \begin{bmatrix}\omega_{n,f}^2 - \omega_{n,h}^2\\ -2\zeta_f\omega_{n,f} + 2\zeta_h\omega_{n,h}\end{bmatrix}$ and $M_{var}(\boldsymbol{\rho}) = \rho_4$, $f_i(t) = \theta_1(t)$. Therefore, the process fault is modeled as an actuator additive fault in (50).

Wind Turbine LPV Description

Tables may place within the texts or just before the figures. All quantities in tables should be accompanied by their units. Table footnotes should be indicated by letters a, b, c, etc. The states of the wind turbine model are rotor rotational speed, generator speed, torsion angle of the drive-train, generator torque, pitch angle, and pitch angle rate. One of the purposes of the first transformation is reshaping the output matrix. Therefore, we rearrange the order of the states in a way that the specific structure takes place. The state vector is defined as

$$\begin{aligned}&x(t)\\ &= \left[\dot{\beta}'(t) \quad \theta(t) \quad \omega_r(t) \quad \omega_g(t) \quad T_g(t) \quad \beta(t)\right.\end{aligned} \quad (51)$$

Among the considered states in the model, the drive-train equation is severely nonlinear. Such a nonlinear differential equation can be linearized and expressed in a linear parameter varying manner where the model matrices change with a varying parameter. The drive-train equation is shown in (52).

$$\begin{aligned}\dot{\omega}_r(t) &= \frac{1}{J_r}T_a(\omega_r, \beta, V_w, t) - \frac{K_{dt}}{J_r}\theta(t)\\ &\quad - \frac{B_{dt} + B_r}{J_r}\omega_r(t)\\ &\quad + \frac{B_{dt}}{J_r}\omega_g(t)\end{aligned} \quad (52)$$

The nonlinear term is $T_a(\omega_r, \beta, V_w, t)$ which is the aerodynamic torque exerted by the wind turbine and estimated in the form $T_a(\omega_r, \beta, V_w, t) = \frac{1}{2}\rho_{air}\pi R^3 C_q(\omega_r, \beta, V_w)V_w^2$. C_q is a torque coefficient table and consists of aerodynamic experimental data. Using surface fitting by the well-known software MATLAB, a fifth-degree (quintic) polynomial is derived in which ω_r, β, V_w are the variables. The interpolated form of $\frac{1}{J_r}T_a(\omega_r, \beta, V_w, t)$ becomes

$$\begin{aligned}T_a(\omega_r, \beta, V_w, t)/J_r &= p_1V_w\omega_r + p_2\omega_r^2\\ &\quad + p_3\frac{\omega_r^3}{V_w} + p_4\frac{\omega_r^4}{V_w^2} + p_5\frac{\omega_r^5}{V_w^3}\\ &\quad + p_6V_w\omega_r\beta + p_7V_w\omega_r\beta^2\\ &\quad + p_8V_w\omega_r\beta^3 + p_9V_w\omega_r\beta^4\\ &\quad + p_{10}V_w^2\beta + p_{11}\omega_r^2\beta\\ &\quad + p_{12}V_w^2\beta^2 + p_{13}V_w^2\beta^3\\ &\quad + p_{14}V_w^2\beta^4 + p_{15}V_w^2\beta^5\\ &\quad + p_{16}\omega_r^2\beta^2 + p_{17}\omega_r^2\beta^3\\ &\quad + p_{18}\frac{\omega_r^3\beta}{V_w} + p_{19}\frac{\omega_r^3\beta^2}{V_w}\\ &\quad + p_{20}\frac{\omega_r^3\beta}{V_w} + p_{21}V_w^2\end{aligned} \quad (53)$$

To validate the accuracy of the model, the parameters of the benchmark and presented model are compared. Fig. 2 demonstrates real data points (colored surface) and some sampled data from the LPV model (black points).

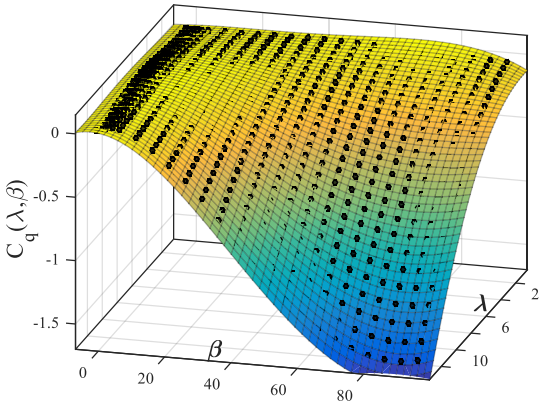


Fig. 2: Torque aerodynamic coefficient (10) (colored surface) and sample points from the proposed model (black points).

The black points approximately coincide with the real data surface showing the validity of the model. The varying parameters are capsulated in $\boldsymbol{\rho} = [V_w, \omega_r, \beta, \dot{\beta}']^T$. Then, a new structure is obtained by rearranging (53). Table 2 consists of the LPV coefficients.

Table 2: Coefficients of the wind turbine LPV model

$p_1 = 2.13e-5$	$p_2 = -8.18e-3$	$p_3 = -3.92e-5$
$p_4 = 2.89e-1$	$p_5 = -3.70e-7$	$p_6 = 2.17e-10$
$p_7 = -1.10e-11$	$p_8 = 2.47e-13$	$p_9 = -2.63$
$p_{10} = 9.26$	$p_{11} = -11.42$	$p_{12} = -1.27e-4$
$p_{13} = 8.89e-7$	$p_{14} = 1.44e-2$	$p_{15} = -3.34e-2$
$p_{16} = -3.15e-4$	$p_{17} = 1.31e-5$	$p_{18} = -2.36e-7$
$p_{19} = 1.53e-7$	$p_{20} = -1.2e-9$	

$$
\begin{aligned}
&T_a(\omega_r, \beta, V_w, t)/J_r \\
&\quad = \Big[p_1\rho_1 + p_2\rho_2 + p_3\frac{\rho_2^2}{\rho_1} + p_4\frac{\rho_2^3}{\rho_1^2} \\
&\quad + p_5\frac{\rho_2^4}{\rho_1^3} + p_6\rho_1\rho_3 + p_7\rho_1{\rho_3}^2 \\
&\quad + p_8\rho_1{\rho_3}^3 + p_9\rho_1{\rho_3}^4\Big]\omega_r \\
&\quad + \Big[p_{10}V_w^2\beta + p_{11}\rho_2^2 + p_{12}\rho_1^2\rho_3 \\
&\quad + p_{13}\rho_1^2{\rho_3}^2 + p_{14}\rho_1^2{\rho_3}^3 + p_{15}\rho_1^2{\rho_3}^4 \\
&\quad + p_{16}\rho_2^2\rho_3 + p_{17}\rho_2^2{\rho_3}^2 + p_{18}\frac{\rho_2^3}{\rho_1} \\
&\quad + p_{19}\frac{\rho_2^3\rho_3}{\rho_1} + p_{20}\frac{\rho_2^3}{\rho_1}\Big]\beta + p_{21}V_w^2 \\
&\quad = \Delta A_{3,3}\omega_r + \Delta A_{3,6}\beta + p_{21}V_w^2
\end{aligned}
\tag{54}
$$

Substituting into (52) results in

$$
\begin{aligned}
\dot{\omega}_r(t) = &-\frac{K_{dt}}{J_r}\theta(t) - \frac{B_{dt}+B_r}{J_r}\omega_r(t) \\
&+ \frac{B_{dt}}{J_r}\omega_g(t) \\
&+ \Delta A_{3,3}(\boldsymbol{\rho})\omega_r(t) \\
&+ \Delta A_{3,6}(\boldsymbol{\rho})\beta(t) + p_{21}V_w^2
\end{aligned}
\tag{55}
$$

We are unable to merge the last term of the (55) in the system matrix and it is considered as model uncertainty. The LPV description of the model is expressed in

$$
\begin{aligned}
&A(\boldsymbol{\rho}) \\
&= \begin{bmatrix}
-2\zeta_h\omega_{n,h} & 0 & 0 & 0 & 0 & -1 \\
0 & 0 & 1 & -\frac{1}{N_g} & 0 & 0 \\
0 & -\frac{K_{dt}}{J_r} & a_{33}(\boldsymbol{\rho}) & \frac{B_{dt}}{N_g J_r} & 0 & a_{36}(\boldsymbol{\rho}) \\
0 & \frac{\eta_{dt}K_{dt}}{N_g J_g} & \frac{\eta_{dt}B_{dt}}{N_g J_g} & a_{44} & -\frac{1}{J_g} & 0 \\
0 & 0 & 0 & 0 & -\alpha_{gc} & 0 \\
\omega_{n,h}^2 & 0 & 0 & 0 & 0 & 0
\end{bmatrix} \\
&B = \begin{bmatrix} 0 & 0 & 0 & 0 & \alpha_{gc} & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T, \\
&C = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \\
&Q = [0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0]^T, \\
&M_{inv} = [m_1 \quad 0 \quad 0 \quad 0 \quad 0 \quad m_6]^T, \\
&M_{var}(\boldsymbol{\rho}) = \rho_4, \; u(t) = [\tau_{g,r}(t) \quad \beta_r(t)]^T, f_i(t) = \theta_1(t)
\end{aligned}
\tag{56}
$$

A summary of the wind turbine fault reconstruction and observation is illustrated in Fig. 3.

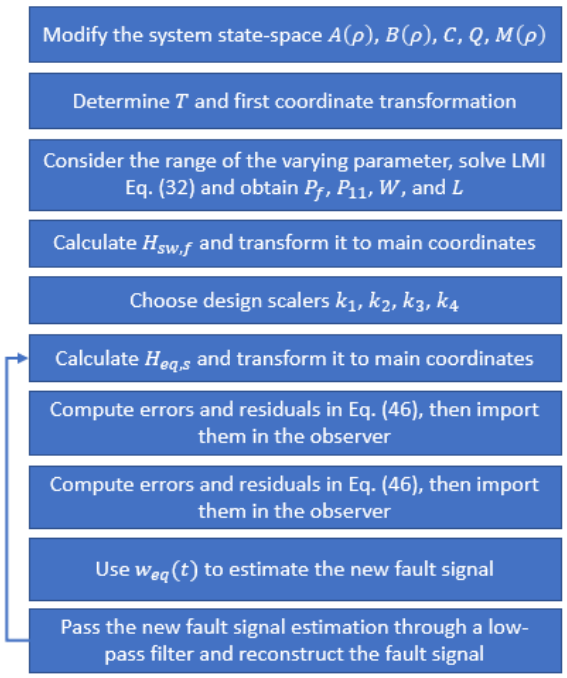


Fig. 3: Design algorithm flowchart for wind turbine fault reconstruction.

Results and Discussion

In this section, reconstruction of pitch actuator fault is investigated. The observability of the proposed model has been checked in Section 3.3. The observability matrix (51) is full-rank during the operation which means the model is observable even by the time-varying nature of $\boldsymbol{\rho}$. Considering the structure in (17), $T = I_4$ and

$$T_f = \begin{bmatrix} 46.7 & 0 & 0 & 0 & 0 & 3 \\ 15.6 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Then, solving matrix inequality (32) gives P_f, P_{11}, L and W. Once L is calculated, the first observer gain $H_{eq}(\boldsymbol{\rho})$ is built in the structure of (41). The second transformation is then possible substituting L in (39). The fault is reconstructed substituting W and T in (47).

$$P_{af} = \begin{bmatrix} 3.41 & -0.0019 & 3.35 & 1.61e-8 \\ -7.854 & -4.31e-10 & 3.35 & -10.82 \\ 3.35 & 3.35 & 5.12e-4 & -2.88e-2 \\ 1.61e-8 & -2.88e-2 & -7.44e-9 & 7.11 \end{bmatrix}$$

$$L = \begin{bmatrix} 0.0011 & -1.86e-5 & 5.22e-24 & 0 \\ -0.0075 & -2.48e-5 & 3.26e-24 & 0 \end{bmatrix}$$

$$W = [8.81e-5 \quad -1.06e-7 \quad 0 \quad -8.95e-3]$$

$$P_{11} = \begin{bmatrix} 17.2 & -35.9 \\ -35.9 & 112.8 \end{bmatrix}$$

$$k_1 = 10, k_2 = 1, k_3 = 50, k_4 = 100, a_f = 50$$

Wind speed, as stochastic input data, has been manipulated using a stochastic input profile gathered from the actual wind speed measurements of a wind park [37] An average of 15 m/s has been chosen for wind speed which is included in region 3 of operation regions.

A high-slope ramp (nearly unit step) between 8th to 14th seconds is considered as an actuator fault in the pitch system caused by hydraulic leakage. Hydraulic leakage is an incipient and fast decaying fault [7]. The range and the rate of such a fault are illustrated in Table 1. To emphasize the paper methodology i.e. robust LPV fault reconstruction using second-order sliding mode observers, a different fault scenario is inspected. To exhibit the rapid and precise fault recognition and reconstruction, it is supposed that an abrupt fault takes place faster than 1/(100 seconds) (according to Table 1) in the simulation.

Fig. 4 shows the reconstruction of the LPV fault indicator in the presence of model uncertainty and measurement noise during 6 seconds. A step is considered as the fault indicator θ_1. Also, the estimation of pitch system properties (natural frequency and damping ratio) are illustrated in Fig. 4.

As shown in Fig. 4, once an actuator fault occurs, it is reconstructed accurately and immediately. The effects of noises and uncertainty have been minimized in (32). However, the fault estimation signal conveys noise between the 8th to 14th seconds. The estimation of the new fault signal in (24) is enhanced by exploiting a low-pass filter in the form of

$$F(s) = \frac{10}{s+10} \tag{60}$$

As mentioned, the new fault signal in (10) is passed through a low-pass filter $F(s)$ in order to become smoother in (44). The measurement noise in all of the sensors appears as fluctuations. The measurement noises' powers are chosen as $\|\varphi_{\omega_r}\| = 10^{-5}$, $\|\varphi_{\omega_g}\| = 10^{-1}$, $\|\varphi_{\tau_g}\| = 10^{+3}$, and $\|\varphi_{\beta}\| = 10^{-4}$. It should be noted that the robustness of the proposed observer is not disturbed by the amplitude of the noises with known bounds. In addition, all the measurements could be properly filtered to avoid the harmful effects of the noises in state estimation.

Figs. 5 & 6 illustrate the performance of second-order sliding mode observers in the presence of pitch actuator fault. The estimation signals converge to the system states in at most 2.5 seconds with less than 2.48 percent estimation error in generator speed.

Rahnavard et al. [22] used an LTI first-order sliding mode observer to reconstruct wind turbine faults. Compared with this paper, both methods are fast and accurate and also overcome the output noises' effects. But, the proposed LPV method covers all the wind turbine operating regions and no linearization approximations are required at the cost of heavier computations. While in [22], the nonlinear equations are linearized around an operating point. It requires gain-scheduling and switching among the models which reduces the robustness and model-accuracy of the method.

Sloth et al. [17], used robust theory for LPV active-fault-tolerant control of a benchmark model, similar to this paper. The LPV model in [17], contains linearizing the aerodynamic torque around a floating trajectory. The LPV model of the current paper uses a quintic multi-variable polynomial instead of linearization which improves the accuracy.

A time-invariant sliding mode observer (with the same procedure for designing observer gains) is carried out for a nonlinear model in which the aerodynamic torque is regarded as the uncertainty ($\xi_{NL} = T_a/J_r$). It should be mentioned that $\xi_{LPV} = p_{21}V_w^2$ from (55) is considered as the uncertainty signal, exerting the proposed LPV model.

Fig. 7 illustrates the comparison between the fault reconstruction performance of the nonlinear model (in which the aerodynamic torque is regarded as an uncertainty) and the LPV model of a wind turbine from Section 5. As expected, the LPV model provides a more

accurate estimation of the actuator faults using the same filters (60).

As an average value, the amplitude of fluctuations of fault reconstruction in the nonlinear model is approximately 11.5 times larger than fluctuations of fault reconstruction in the LPV model.

In addition, the mean value of fault reconstruction of the nonlinear model is biased about 3 percent before fault occurrence which demonstrates the weaker performance of the observer, affected by large uncertainty ($\xi_{NL} = T_a/J_r$).

It arises from the severe nonlinearity of aerodynamic torque which is handled using LPV methods. In this case, the magnitude of uncertainty in (58) is much less than that in the nonlinear model i.e. $\|\xi_{LPV}\| = 0.0288\, rad/s^2$ in comparison with $\|\xi_{NL}\| = 0.8979\, rad/s^2$. In addition, the observer LPV gain is calculated in an adaptive manner (38).

Conclusion

The application of SMO on a wind turbine system that contains a parameter varying model is investigated. The results of this paper show that the sliding mode fault reconstruction method is applicable for LPV systems. It is interesting to note that the severity of the pitch hydraulic pump is estimated while the model parameters are varying, sensors are faulty, and some parameters are uncertain.

The LPV model of a wind turbine is derived by surface interpolation and fitting a quintic polynomial for the aerodynamic torque coefficient. The LPV model predicts the real behavior of the system with the highest error of 3.2 percent.

The robustness of the estimation and reconstruction scheme is another merit of the proposed SMO as shown in the results. Wind speed range is considered from 14 m/s to 16 m/s and it is regarded as a stochastic input exerting aerodynamic torque. Fast and accurate fault reconstruction happens in 0.6 seconds with less than one percent error.

Compared to the previous works, the proposed observer performance is less affected by the pitch actuator fault, and the fault severity is estimated in 2.5 seconds with an error smaller than 2.48 percent. The LPV observer presented in this work covers both of the wind turbine operating regions which are more appropriate for large-scale applications.

As a prospective, the proposed method will be applied to a real large wind turbine to assess its performance. In addition, the tuning mechanism of the design parameters of the proposed observer can be determined by using optimization methods such as genetic algorithm which results in better performance and smaller errors.

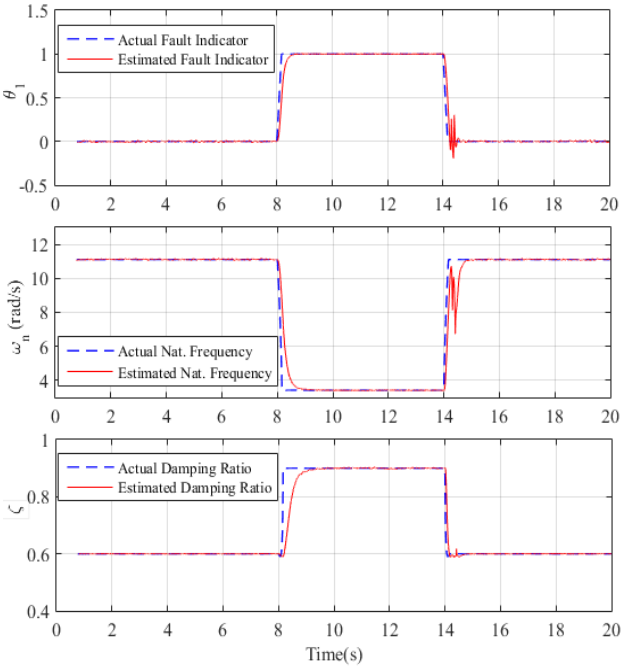


Fig. 4: Fault indicator, natural frequency, and damping ratio of pitch actuator. The Blue (dashed) line is the actual signal and the red (solid) line is the estimation.

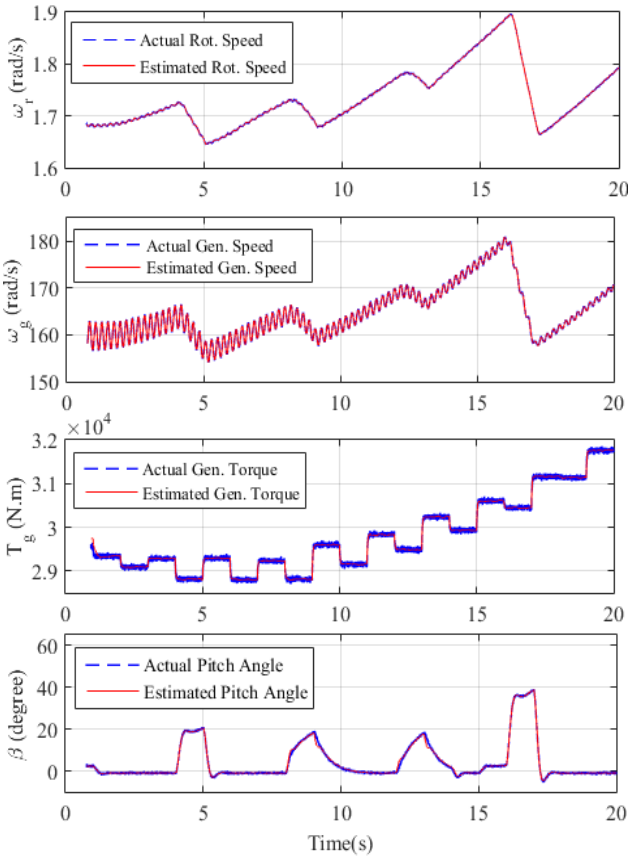


Fig. 5: Measured state estimation (output estimation) using second-order SMO in the presence of pitch actuator.

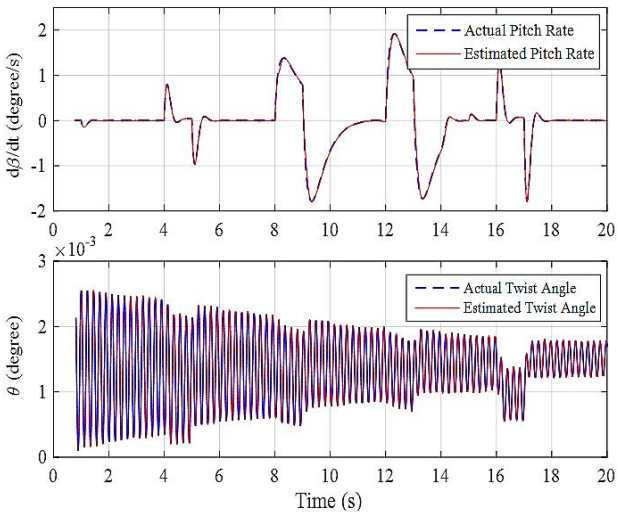


Fig. 6: Unmeasured state estimation using second-order SMO in the presence of pitch actuator fault.

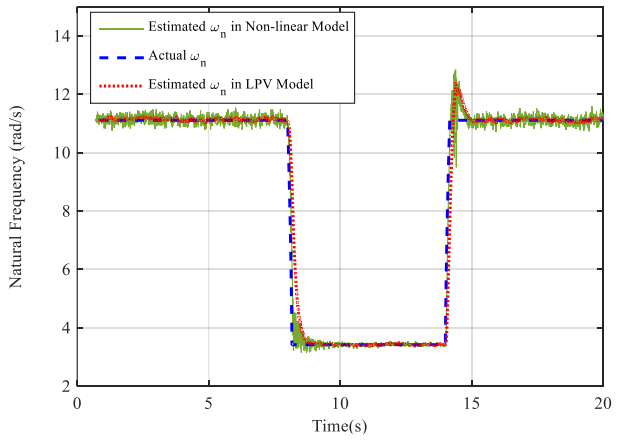


Fig. 7: Comparison of fault reconstruction results in nonlinear model and LPV model.

Author Contributions

M. Mousavi carried out data analysis and simulations. M. Mousavi, and M. Ayati, interpreted the results and wrote the manuscript. M. Mousavi, M. Ayati, M. Hairi-Yazdi, and S. Siahpour revised the manuscript.

Acknowledgment

There is no acknowledgement.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

References

[1] Y. Wentao, G. Hua, X. Shuai, Y. Geng, "Nonlinear individual pitch control of large wind turbines for blade load reduction," in Proc. International Conference on Renewable Power Generation, 2015.

[2] M. Rahnavard, M. Ayati, M. R. Hairi Yazdi, "Robust actuator and sensor fault reconstruction of wind turbine using modified sliding mode observer," Trans. Inst. Meas. Control, 41(6): 1504-1518, 2019.

[3] J. Lan, R. J. Patton, X. Zhu, "Fault-tolerant wind turbine pitch control using adaptive sliding mode estimation," Renewable Energy, 116: 219-231, 2018.

[4] Z. Rafiee, A. Mosahebfard, M. H. Sheikhi, "High-performance ZnO nanowires-based glucose biosensor modified by graphene nanoplates," Mater. Sci. Semicond. Process., 26(10) 115, 2020.

[5] D. Kim, D. Lee, "Hierarchical fault-tolerant control using model predictive control for wind turbine pitch actuator faults," Energies, 12(16): 3097, 2019.

[6] H. Habibi, H. Rahimi Nohooji, I. Howard, "Optimum efficiency control of a wind turbine with unknown desired trajectory and actuator faults," J. Renewable Sustainable Energy, 9(6): 063305, 2017.

[7] T. Esbensen, C. Sloth, S. L. Pedersen, J. M. Holm, T. N. Jensen, M. Philipsen, "Fault diagnosis and fault-tolerant control of wind turbines", Master Thesis, Aalborg University, 2009.

[8] B. Lu, "A review of recent advances in wind turbine condition monitoring and fault diagnosis," in Proc. 2009 IEEE Power Electronics and Machines in Wind Applications: 1-7, 2009.

[9] M. R. Wilkinson, F. Spinato, P. J. Tavner, "Condition monitoring of generators and other subassemblies in wind turbine drive trains," in proc. 2007 IEEE International Symposium on Diagnostics for Electric Machines, Power Electronics and Drives: 388-392, 2007.

[10] S. Khodakaramzadeh, M. Ayati, M. R. Haeri Yazdi, "Fault diagnosis of a permanent magnet synchronous generator wind turbine," J. Electr. Comput. Eng. Innovations (JECEI), 9(2): 143-152, 2021.

[11] S. M. Hosseini, M. Manthouri, "Type 2 adaptive fuzzy control approach applied to variable speed DFIG based wind turbines with MPPT algorithm," Iran. J. Fuzzy Syst., 19(1): 31-45, 2021.

[12] V. Fazlollahi, F. A. Shirazi, M. Taghizadeh, S. Siahpour, "Robust wake steering control design in a wind farm for power optimization using adaptive learning game theory (ALGT) method," International Journal of Control, (Just-accepted): 1 , 2021.

[13] Z. Dehghani Arani, S. A. Taher, M. H. Karimi, M. Rahimi, "Coordinated model predictive DC-Link voltage, current, and electromagnetic torque control of wind turbine with DFIG under grid faults," J. Electr. Comput. Eng. Innovations (JECEI), 8(2): 201-218, 2020.

[14] P. F. Odgaard, J. Stoustrup, M. Kinnaert, "Fault tolerant control of wind turbines – a benchmark model," in Proc. 7th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes, 42(8): 155-160, 2009.

[15] H. Badihi, Y. Zhang, H. Hong, "Fuzzy gain-scheduled active fault-tolerant control of a wind turbine," J. Franklin Inst., 351(7): 3677–706, 2014.

[16] Y. Yin, P. Shi, F. Liu, "Gain-scheduled robust fault detection on time-delay stochastic nonlinear systems," IEEE Trans. Ind. Electron., 58(10): 4908–16, 2011.

[17] C. Sloth, T. Esbensen, J. Stoustrup, "Robust and fault-tolerant linear parameter-varying control of wind turbines," Mechatronics, 21(4): 645–59, 2011.

[18] F. D. Bianchi, R. J. Mantz, C. F. Christiansen, "Gain scheduling control of variable-speed wind energy conversion systems using quasi-LPV models," Control Eng. Pract., 13(2): 247–55, 2005.

[19] M. Sami, R. J. Patton, "An FTC approach to wind turbine power maximisation via T-S fuzzy modelling and control," in Proc. 8th IFAC

symposium on fault detection, supervision and safety of technical processes, Mexico City, Mexico, 45(20): 349-354, 2012.

[20] A. A. Ozdemir, P. Seiler, G.J. Balas, "Wind turbine fault detection using counter-based residual thresholding," in Proc. IFAC world congress, 44(1): 8289-8294, 1998.

[21] M. R. Alizadeh Pahlavani, H. Damroodi, "LPV Control for speed of permanent magnet synchronous motor (PMSM) with PWM Inverter," J. Electr. Comput. Eng. Innovations (JECEI), 4(2): 185-193, 2016.

[22] M. Rahnavard, M. R. Hairi-Yazdi, M. Ayati, "On the development of a sliding mode observer-based fault diagnosis scheme for a wind turbine benchmark model," Energy Equip. Syst., 5(1): 13–27, 2017.

[23] J. Liu, D. Xu, X. Yang, "Sensor fault fetection in variable speed wind turbine system using $H-/H\infty$ method," in Proc. 7th World Congress on Intelligent Control and Automation: 4265–4269, 2008.

[24] J. Blesa, P. Jiménez, D. Rotondo, "An interval NLPV parity equations approach for fault detection and isolation of a wind farm," IEEE Trans. Ind. Electron., 62(6): 3794–3805, 2015.

[25] H. Badihi, Y. Zhang, H. Hong, "Fault-tolerant cooperative control in an offshore wind farm using model-free and model-based fault detection and diagnosis approaches," Appl. Energy, 201: 284-307, 2017.

[26] H. Badihi, Y. Zhang, H. Hong, "Active fault tolerant control in a wind farm with decreased power generation due to blade erosion / debris build-up," IFAC, 48(21): 1369–74.

[27] M. Mousavi, M. Rahnavard, M. R. Hairi-Yazdi, M. Ayati, "On the development of terminal sliding mode observers," in Proc. 26th Iranian Conference on Electrical Engineering (ICEE 2018), 2018.

[28] M. Rahnavard, M. Ayati, M. R. Hairi-Yazdi, M. Mousavi, "Finite time estimation of actuator faults, states, and aerodynamic load of a realistic wind turbine," Renew. Energy, 130: 256–267, 2019.

[29] M. Mousavi, M. Rahnavard, M. Ayati, M. R. Hairi Yazdi, "Terminal sliding mode observers for uncertain linear systems with matched disturbance," Asian J. Control, 21(1): 377–386, 2019.

[30] M. Mousavi, M. Rahnavard, S. Haddad, "Observer based fault reconstruction schemes using terminal sliding modes," Int. J. Control, 93(4): 881-888, 2018.

[31] Wonderful Pictures, "Wind turbine pictures from around the world", 2017. [Available online].

[32] A. Pisano, E. Usai, "Globally convergent real-time differentiation via second order sliding modes Globally convergent real-time differentiation via second order sliding modes," Int. J. Syst. Sci., 38(10): 37–41, 2007.

[33] C. P. Tan, C. Edwards, "Sliding mode observers for reconstruction of simultaneous actuator and sensor faults," in Proc. IEEE Conference on Decision and Control, 2 :1455–1460, 2003.

[34] H. Alwi, C. Edwards, "Robust fault reconstruction for linear parameter varying systems using sliding mode observers," Int. J. Robust Nonlinear Control, 24(14): 1947-1968, 2013.

[35] J. A. Moreno, M. Osorio, "Strict lyapunov functions for the super-twisting algorithm," IEEE Trans. Autom. Control., 57(4): 1035–1040, 2012.

[36] S. Sundaram, C. N. Hadjicostis, "Structural controllability and observability of linear systems over finite fields with applications to multi-agent systems," IEEE Trans. Autom. Control, 85(1): 60-73, 2012.

[37] A. Kumar, K. Stol, "Simulating feedback linearization control of wind turbines using high-order models," Wind Energy, 13(5): 419-432, 2010.

[38] G. Pujol-Vazquez, L. Acho, J. Gibergans-Báguena, "Fault detection algorithm for wind turbines' pitch actuator systems," Energies, 13(11): 2861, 2020.

[39] D. DSL, L. PW, "Simulation model of wind turbine 3p torque oscillations due to wind shear and tower shadow," IEEE Trans. Energy Convers., 21(3): 717–724, 2006.

Biographies

Mohammad Mousavi received his B.Sc. degree in mechanical engineering from Shiraz University, Shiraz, Iran, in 2015. In 2017, he graduated from the School of Mechanical Engineering, University of Tehran, Tehran, Iran. He researches in the fields of wind turbines and model-based fault detection in the Advanced Instrumentation Laboratory, at the School of Mechanical Engineering. His current research interests are industrial application of observers, theorical enhancement of sliding mode observers, linear parameter varying observer, and controller design.

- Email: smousav1@binghamton.edu
- ORCID: 0000-0003-2410-4595
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Moosa Ayati received his B.Sc. degree from Isfahan University of Technology, Isfahan, Iran, in 2004, and his M.Sc. and Ph.D. degrees in 2006 and 2011 from K. N. Toosi University of Technology, Tehran, Iran, all in Electrical Engineering with the first rank honors. He spent two years as a Post-Doctoral Fellow at the University of Tehran School of Electrical Engineering, College of Engineering, Tehran, Iran, working on fault detection systems. He is currently head of the Advanced Instrumentation Laboratory (AIL) and an Associate Professor with the control division, School of Mechanical Engineering, College of Engineering, at the University of Tehran, Tehran, Iran. His area of interest includes adaptive control and system identification, fault detection systems, instrumentation and industrial automation, Mechatronics, and hybrid systems. Professor Ayati is a member of the Iranian Society of Instrumentation and Control, Iranian Society of Mechanical Engineers, and the National Society of Mechatronics.

- Email: m.ayati@ut.ac.ir
- ORCID: 0000-0001-9943-739X
- Web of Science Researcher ID: AFQ-6437-2022
- Scopus Author ID: 25027078800
- Homepage: https://rtis2.ut.ac.ir/cv/m.ayati/?lang=en-gb

Mohammad Reza Hairi-Yazdi received his B.Sc. and M.Sc. degrees in Mechanical Engineering from Amir Kabir University of Technology, Tehran, Iran in 1985 and 1987 respectively. He received his Ph.D. degree from Imperial College London in 1992 and since then he has been at the University of Tehran, Tehran, Iran where he is now a Professor at the School of Mechanical Engineering. His main research interests include design, simulation, manufacturing and control of dynamic systems.

- Email: myazdi@ut.ac.ir
- ORCID: 0000-0002-2507-313X
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Shahin Siahpour received his B.S. degree in mechanical engineering from Shiraz University, Shiraz, Iran, in 2015, and the M.S. degree in mechanical engineering from University of Tehran, Tehran, Iran, in 2017. He is currently pursuing the Ph.D. degree in mechanical engineering with the University of Cincinnati, Cincinnati, OH, USA. His research interests include deep learning, prognostics, health management, and industrial AI.

- Email: siahposn@mail.uc.edu
- ORCID: 0000-0002-5359-7731
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:
M. Mousavi, M. Ayati, M. hairi-yazdi, S. Siahpour, "Robust linear parameter varying fault reconstruction of wind turbine pitch actuator using second order sliding mode observer," J. Electr. Comput. Eng. Innovations, 11(1): 229-241, 2023.

DOI: 10.22061/jecei.2022.8179.500

URL: https://jecei.sru.ac.ir/article_1786.html