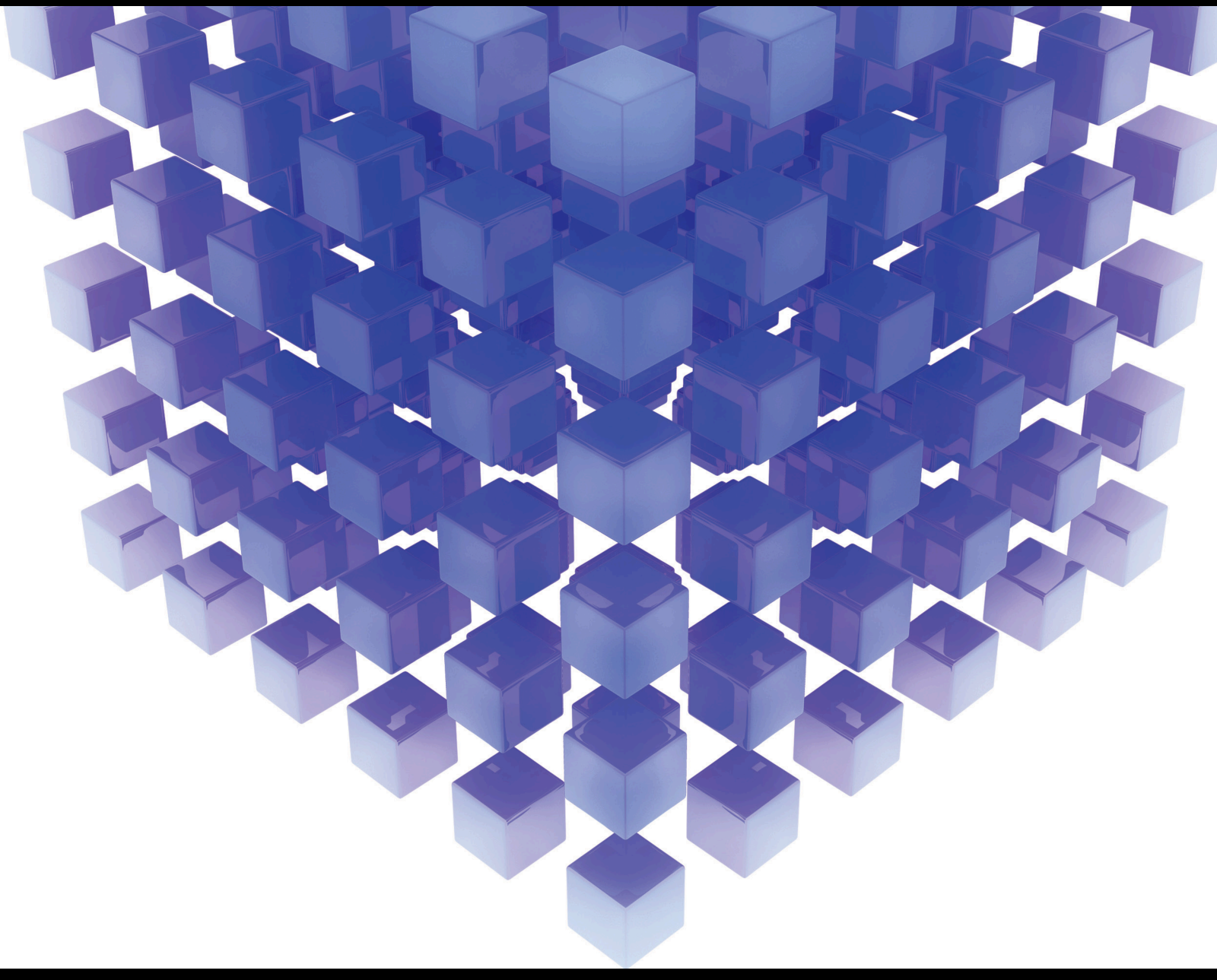
Mathematical Problems in Engineering

Algorithms and Devices for Smart Processing Technology for Energy Saving

Lead Guest Editor: Sanghyuk Lee
Guest Editors: Mohamed Nayel, Pham Van Huy, and Sang-Bong Rhee



Algorithms and Devices for Smart Processing Technology for Energy Saving

Mathematical Problems in Engineering

Algorithms and Devices for Smart Processing Technology for Energy Saving

Lead Guest Editor: Sanghyuk Lee
Guest Editors: Mohamed Nayel, Pham Van Huy, and Sang-Bong Rhee

Contents

Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 9853615, 2 pages
https://doi.org/10.1155/2021/9853615

Editorial

Algorithms and Devices for Smart Processing Technology for Energy Saving

Sanghyuk Lee,[1] **Mohamed Nayel**,[2] **Van Huy Pham**,[3] **and Sang Bong Rhee**[4]

[1]*Xi'an Jiaotong–Liverpool University, High Educational Town, SIP, Suzhou, China*
[2]*Assiut University, Assiut, Egypt*
[3]*Ton Duc Thang University, Ho Chi Minh City, Vietnam*
[4]*Yeungnam University, Daegu, Republic of Korea*

Correspondence should be addressed to Sanghyuk Lee; sanghyuk.lee@xjtlu.edu.cn

Received 19 July 2021; Accepted 19 July 2021; Published 17 August 2021

Smart system and energy saving has become a topic of focus in current research in engineering, business, computer science, and social science with mathematical foundations. The wording "smart" itself has spread widely since 2007, from the revolutionary iPhone. It has changed the mobile business, user platform, data usage, and software/hardware as well. Social media, calling, text, and various applications are all in mobile phones with efficient energy expense afterward. There are also challenges in ubiquitous systems, city planning, and wireless friendly systems for control, management, and industrial application, as well as in unmanned factories [1]. Together with the development of WiFi infrastructure, recently Internet of Things (IoT) has had a lot of emphasis [2]. The recent development of smart devices, such as smart sensors and actuators, has greatly improved the performance of industrial processes such as smart factories, smart farms, and other related systems [3]. With the needs of developing technology, we aim to provide an open forum on this research idea, specifically in smart devices and their application, and related algorithms/applications. The proposed research outputs are becoming an important foundation for the implementation of smart systems with energy-efficient solutions.

By the consideration of the aforementioned necessities, this Special Issue is a timely opportunity to discuss and summarize the latest developments in this area. We especially focus on the research related to algorithms, smart device development, and applications. After summarizing the papers, we feel that the Issue sets the goal of addressing the following major challenges with the advent of many challenging issues concerning smart systems:

(i) Smart processing technologies include applications to complex or networked systems: effective data processing, avoiding signal conflict, and fast and accurate processing.

(ii) Development of high-performance devices through hardware design is done, relevant to Internet of Things (IoT) to extend the connection with networks. As you increment the dimension, there is the so-called curse of dimensionality. Therefore, the solution should be cooperated with big data analysis at the same time.

(iii) Together with smart device and algorithm research, complex system coordination for efficient management is also needed. In this case, research on control and consensus for network systems is needed.

(iv) The link between hardware-level device research and the upper-level management is the final goal to accomplish in the Issue. Despite the fact that there is similar research on "game theoretic approach together with IoT," "consensus for the networked system," and others, they have not yet reached acceptable results in the industrial area. Herewith, we need to make a more open topic and share ideas from the relevant researchers.

From the mentioned challenges, new methods and alternative approaches are needed to solve such tough problems. Recently, AI-oriented research has been provided with its fundamental research on application to networked systems [4] and the heuristic and metaheuristic algorithms also obtained much attention based on nature-inspired optimization algorithms [5].

This Special Issue provides a timely platform to discuss and summarize the latest developments in the relevant area. With this opportunity, we have shared the theoretical methodology and mathematical analysis, although applications concerning smart processing and device development are also addressed. A high number of high-quality submissions were received, totalling 27 papers.

From the Special Issue, we can see that the topics range from smart device development and applications to processing algorithm development and applications as well. Clearly, there are many other interesting developments concerning smart processing research not only in engineering and science but also in social science and administrative decision making for smooth processing. However, most of the papers are focussed on hardware development and application, smart algorithm proposal, and verification. We are sure that whilst our Special Issue can only cover a small fraction of the latest developments, on this foundation, we also hope that further research is inspired in this area.

Conflicts of Interest

The Guest Editors declare no conflicts of interest.

Sanghyuk Lee
Mohamed Nayel
Van Huy Pham
Sang Bong Rhee

References

[1] S. H. Lee, J. H. Han, Y. T. Leem, and Y. Tan, "Towards ubiquitous city: concept, planning, and experiences in the Republic of Korea," in *Knowledge-Based Urban Development: Planning and Applications in the Information Era*IGI Global, Hershey, PA, USA, 2008.

[2] J. Holler, V. Tsiatsis, C. Mulligan, S. Karnouskos, S. Avesand, and D. Boyle, *From Machine-to-Machine to the Internet of Things: Introduction to a New Age of Intelligence*, Academic Press, Cambridge, MA, USA, 2014.

[3] T. Kozai, Ed., *Smart Plant Factory the Next Generation Indoor Vertical Farms*, Springer, Berlin, Germany, 2018.

[4] M. Gilbert, *Artificial Intelligence for Autonomous Networks*, Chapman & Hall/CRC, Boca Raton, FL, USA, 2019.

[5] X. S. Yang, *Nature-Inspired Optimization Algorithms*, Elsevier, Amsterdam, Netherlands, 2014.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 4795187, 10 pages
https://doi.org/10.1155/2020/4795187

Research Article

Source-Word Decomposition for Neural Machine Translation

Thien Nguyen, Hoai Le, and Van-Huy Pham

Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam

Correspondence should be addressed to Thien Nguyen; nguyenchithien@tdtu.edu.vn

Received 20 August 2020; Revised 1 November 2020; Accepted 3 December 2020; Published 16 December 2020

Academic Editor: Yakov Strelniker

End-to-end neural machine translation does not require us to have specialized knowledge of investigated language pairs in building an effective system. On the other hand, feature engineering proves to be vital in other artificial intelligence fields, such as speech recognition and computer vision. Inspired by works in those fields, in this paper, we propose a novel feature-based translation model by modifying the state-of-the-art transformer model. Specifically, the encoder of the modified transformer model takes input combinations of linguistic features comprising of lemma, dependency label, part-of-speech tag, and morphological label instead of source words. The experiment results for the Russian-Vietnamese language pair show that the proposed feature-based transformer model improves over the strongest baseline transformer translation model by impressive 4.83 BLEU. In addition, experiment analysis reveals that human judgment on the translation results strongly confirms machine judgment. Our model could be useful in building translation systems translating from a highly inflectional language into a noninflectional language.

1. Introduction

Neural machine translation (NMT) is an active research field with a lot of newly published works [1–4]. They study different aspects of NMT in order to improve it. In [1], the authors proposed a single model to translate from multiple source languages to multiple target languages. In [2], the author proposed using adversarial input to train the model. Adversarial input is generated from the original input with a small perturbation. In [3], the authors proposed a mechanism to adapt NMT models to new languages and domains. In [4], the authors proposed enriching the training dataset with the predicted sentences. Although these works develop in different directions, all of them are based on end-to-end NMT. End-to-end NMT is a universally applicable translation paradigm. It is complicated from the technical point of view, but very simple from the point of view of linguistics. In contrast to building an effective statistical machine translation system, building a NMT system does not require specialized knowledge of the applied language pair. For all language pairs, regardless of their characteristics, end-to-end NMT takes a sequence of source words from a fixed dictionary, processes, and then returns a sequence of target words from another fixed dictionary. The effectiveness and simplicity in application lead to widespread use of NMT [5–7]. NMT has become a dominant translation paradigm. In ideal circumstances where all words of languages frequently appear in a large training dataset and computational capacity to train translation models is unlimited, end-to-end NMT will work perfectly. In practice, such ideal circumstances do not take place, so the performance of end-to-end NMT is worsened by rare words and out-of-vocabulary problem, which takes place in all translation tasks, especially for low-resource language pairs. To make NMT capable of translating rare words and out-of-vocabulary words, in [8–10], the authors proposed novel NMT models representing words as sequences of subwords, which occur more frequently than the words themselves. For example, according to the byte pair decoding method [8], the Russian word "Цризывают" (meaning: call) is segmented into two subwords "Цризыва@@" and "ют." In the subword-based NMT source, sentences as sequences of source subwords are processed, and then, sequences of target subwords are generated. Generated sequences of target subwords are concatenated to form target sentences, based on characters "@@" informing that containing subwords should be

attached to the following subwords. Their experiment results showed that subword-based NMT delivered a substantial improvement, compared with word-based NMT for high-resource English-German and English-Russian language pairs. The success of subword representation in NMT was further confirmed in the studies [11, 12] for several high-resource language pairs, such as English-Spanish, English-French, and English-Chinese. Recently, a revolutionary NMT model called transformer [13, 14] with the self-attention mechanism significantly outperformed the best previously reported translation models. In cooperation with subword representation, transformer has established itself as the state-of-the-art translation paradigm.

Although subword-based NMT models are able to work well without considering linguistic characteristics of languages, we wonder whether linguistic knowledge helps NMT systems to work more efficiently. Motivated by works [15, 16] in speech recognition with a similar sequence-to-sequence pattern where an original raw input data in the form of discrete speech signal overtime can be represented as a sequence of features, such as log-Mel frequencies, Mel frequency cepstrum, and the knowledge of morphological rich Russian source language and analytic Vietnamese target language, in this work of building a Russian-Vietnamese machine translation system, we experiment the idea of representing each source word in sentence as a combination of features: lemma, grammatical role in sentence, part-of-speech, and morphological features. The decomposition is only deployed on the Russian source side, but not in the Vietnamese target side. The idea comes to mind, since Russian is a morphological rich language, while Vietnamese is an analytic language which lacks morphological marking of case, gender, number, and tense. A Russian sentence consists of tokens infected from lemmas based on their grammatical roles and part-of-speech tags. Inflected tokens are usually called words. By replacing words in the source sentence by a combination of features, we actually increase their appearance frequency in the training dataset; therefore, reduce the severity of rare word and out-of-vocabulary problem in inference. For example, in the training dataset, we have two Russian words, "Пришёл" (meaning: arrived) and "Полюблю" (meaning: fall in love), and in the testing dataset, we have two other words "Приду" (meaning: will arrive) and "Полюбил" (meaning: fell in love). In this case, end-to-end NMT systems are going to recognize two words in the testing dataset as unknown. However, there are close linguistic relationships between words in the training and testing datasets. Both Russian words "Пришёл" and "Приду" are inflected forms of the same lemma "Прийти." The training Russian word "Пришёл" is inflected from the lemma "Прийти," since it is a verb of muscular gender, in singular number and in the past tense, while the testing Russian word "Приду" is inflected from the lemma "Прийти," as it is a verb in singular number and in the future tense. A relationship is also found for two words "Полюблю" and "Полюбил." The training Russian word "Полюблю" is inflected from the lemma "Полюбить," since it is a verb in singular number and in the future tense, while the testing Russian word "Полюбил" is inflected from the lemma "Полюбить," as it is a verb of muscular gender in singular number and in the past tense. If we decompose all these words into features, then in the testing phase, we will have lemmas and grammatical features which are well-known in regard to the training dataset.

In total, this work is dedicated to building a novel transformer-based NMT model taking a sequence of vectors of linguistic features from source words and predicting a sequence of target words.

The rest of this paper is organized as follows. A brief overview of related works is given in the following section. The third section outlines a novel methodology of source-word decomposition for neural machine translation. The fourth section describes materials and methods used in the work. The fifth section shows and analyses experiment results. The final section lists our conclusions from this work.

2. Related Works

This section briefly examines a variety of translation unit representation methodologies used in machine translation systems for several language pairs containing at least one inflectional language, such as Russian, Czech, German, and English.

Before the emergence of NMT, phrase-based SMT used to be very popular. There is a wide range of literature studying phrase-based SMT. Among these studies, factored phrase-based SMT models [17–19] are the most worthy to mention in our work. In factored phrase-based SMT linguistic features, such as lemma, part-of-speech and morphological features are integrating into the surface form of word. The factored phrase-based SMT systems improve translation quality over standard SMT systems for multiple language pairs, such as English-German, English-Spanish, and English-Czech. The approach gained further popularity, after it had been implemented in the famous SMT tool called Moses [20]. In [21, 22], the authors continued to apply and develop the approach and achieved good results. Although the approach belongs to a group of obsolete statistical translation paradigms, its success in integrating linguistic features inspires us to take advantage of linguistic information in redefining the translation unit in modern NMT paradigm.

In recent years, there has been growing interest in integrating linguistic features into NMT architectures. In [23], the authors proposed a novel factored subword-based neural model based on recurrent neural networks that learns source translation unit embeddings, leveraging subword embedding, subword-tag embedding, lemma embedding, part-of-speech embedding, dependency label embedding, and morphological label embedding. They used many different linguistic features in addition to subword itself to take advantage of high-resource characteristic of the English-German language pair. They found that the factored subword-based neural model notably improved translation for the high-resource English-German language pair. Our preliminary experiments with the state-of-the-art transformer NMT model confirm their finding. Our subword-based transformer model combining linguistic-feature

embeddings with subword embedding outperforms a standard subword-based transformer model for the Russian-Vietnamese language pair. However, we believe that we can further improve the system, considering our context of the low-resource and linguistically distant language pair. Due to totally different morphology of Russian and Vietnamese, in the training dataset, the number of unique words of highly inflectional Russian in the source side is multiple times larger than the number of unique words of noninflectional Vietnamese in the target side, which leads to a great probability that a Russian word in reference phase is unseen in the training dataset. To solve the problem, instead of integration, we suggest to use replacement for training from Russian into Vietnamese. Specifically, we calculate source translation unit embedding using only linguistic-feature embeddings without word or subword embedding.

Word representation in [24] bears a close resemblance to our translation unit representation. The authors used a combination of lemma and part-of-speech tag to represent a word in translation for multiple language pairs: English-German, English-Turkish, English-Czech, and English-Latvian. The main difference from our technique lies in the side of translation. They applied their technique in the target side of a NMT system, while we redefine the source-side translation unit. According to their method, a source word is translated into a vector of lemma and part-of-speech tag. Based on that vector, the system predicts a surface form of target word. Obviously, their approach is geared towards translation into an inflectional language. In our case of translation from Russian into noninflectional Vietnamese, we could not take advantage of their technique.

Recently, in [25], the authors introduced an approach of modeling word formation in transformer-based NMT for the English-German language pair. They segmented both English source words and German target words as sequences of vectors of subwords and linguistic subword tags. They reported an improvement over a standard system. Unfortunately, their approach is language-specific, since they deployed a morphological analyzer for English-German language pairs only. Although their approach is interesting, it is not applicable for our task, as we have not found any similar subword taggers for our Russian-Vietnamese language pairs.

3. Source-Word Decomposition for Neural Machine Translation

3.1. Base Transformer Model. Our feature-based transformer model is based on the original transformer model [13], which is the state-of-the-art NMT model. The transformer model has the encoder-decoder architecture. In this work, we make a novelty by modifying the embedding representation in the encoder; therefore, in the following, we describe that part of the encoder in more detail. If you are interested in the general architecture of the transformer model, you can read the original paper [13].

The encoder of the model maps a sequence $\mathbf{x} = \{x_i, \text{for } i = 1, \ldots, n\}$ of n source words x_i from a fixed dictionary $x_i \in \Psi_x$ into a sequence $\mathbf{C} = \{\mathbf{c}_i, \text{for } i = 1, \ldots, n\}$ of individual embeddings of a fixed size $\mathbf{c}_i \in \mathbb{R}^d$. The process of mapping is as follows. First, the encoder looks up each source word x_i in a dictionary of embeddings and retrieves its embedding vector $\mathbf{e}_{x_i}$ of a fixed size $\mathbf{e}_{x_i} \in \mathbb{R}^d$. Next, the encoder looks up the position i of x_i in another dictionary of positional embeddings and retrieves a positional embedding $\mathbf{e}_i$, where $\mathbf{e}_i \in \mathbb{R}^d$. Finally, the encoder adds embedding $\mathbf{e}_i$ with $\mathbf{e}_{x_i}$ weighted by a factor $\sqrt{d}$. Applying the mapping process for all source words, the encoder generates a sequence $\mathbf{C}$ of combined embeddings $\mathbf{c}_i$.

Considering the relationship between words in the sentence with the self-attention mechanism, the encoder transforms the sequence of individual embeddings into a sequence $\mathbf{Z} = \{\mathbf{z}_i, \text{for } i = 1, \ldots, n\}$ of context-aware continuous representations $\mathbf{z}_i \in \mathbb{R}^d$.

Based on the sequence $\mathbf{Z}$, the decoder of the model generates a sequence $\mathbf{y} = \{y_i, \text{for } i = 1, \ldots, k\}$ of target words y_i from another fixed dictionary $y_i \in \Psi_y$. The decoder generates one target word at a time, using previously generated target words as additional input. Mathematically, the transformer model can be represented as a composition of functions as follows.

$$\begin{aligned}
\mathbf{e}_{x_i} &= \phi_x[x_i], \quad \text{for } i = 1, \ldots, n,\\
\mathbf{e}_i &= \phi_i[i], \quad \text{for } i = 1, \ldots, n,\\
\mathbf{c}_i &= \sqrt{d} \times \mathbf{e}_{x_i} + \mathbf{e}_i,\\
\mathbf{C} &= \{\mathbf{c}_i, \quad \text{for } i = 0, \ldots, n\},\\
\mathbf{Z} &= g(\mathbf{C}),\\
y_i &= f(\mathbf{Z}, y_1, y_2, \ldots, y_{i-1}), \quad \text{for } i = 1, \ldots, k.
\end{aligned} \tag{1}$$

In the above equations, notations g and f stand for trainable functions, while notations ϕ_x and ϕ_i are the dictionaries of trainable embeddings with dimension d.

3.2. Source-Word Decomposition. Unlike the basic transformer model, which take source words x_i from a fixed dictionary $x_i \in \Psi_x$ as input, our proposed transformer model takes tuples of linguistic features:

(1) Lemma a_i from a fixed dictionary $a_i \in \Psi_a$

(2) Dependency label b_i from a fixed dictionary $b_i \in \Psi_b$

(3) Part-of-speech tag u_i from a fixed dictionary $u_i \in \Psi_u$ and

(4) Morphological features v_i from a fixed dictionary $v_i \in \Psi_v$

In the place of corresponding source words x_i, consider the fact that the Russian source word is, in fact, a surface form of a lemma, which is inflected on the basis of its grammatical role and part-of-speech. In the other words, a source word can be viewed as a combination of lemma, dependency label, part-of-speech tag, and morphological features.

The grammatical role of a word in sentence is expressed through dependency label assigned to the word. Dependency labels are presented in the study [26]. Part-of-speech types

and corresponding tags of words are listed alphabetically in Table 1, customized for use with Russian from Version 2 of Universal Dependency (https://universaldependencies.org/u/pos/index.html).

Each part-of-speech follows its own morphology rule. For instance, from a lemma "любовь," which is an inanimate noun of feminine gender, we can generate many surface forms according to grammar rules for a noun which has 6 cases (nominative, accusative, genitive, dative, instrumental, and prepositional), two numbers (singular and plural). A noun in Russian also has gender and animate features. Each gender (masculine, feminine, and neuter) has its own inflection rule. Similarly, each animate feature (animate and inanimate) has its own inflection rule. All surface forms inflected from lemma "любовь" are presented in Table 2.

Replacing source words with a combination of the features, we replace the input dictionary Θ_x of the cardinality $|\Theta_x|$ with the tuple of input dictionaries $\Theta_a, \Theta_b, \Theta_u, \Theta_v$ of the summarized cardinality $|\Theta_a| + |\Theta_b| + |\Theta_u| + |\Theta_v|$. From the above analysis, we can see that the summarized cardinality is many times smaller than the cardinality of dictionary of source words—combinations of the features. In the extreme counting, the cardinality of dictionary of source words as combinations of the features is the product $|\Theta_a| \times |\Theta_b| \times |\Theta_u| \times |\Theta_v|$. The reduced input dictionary actually helps to reduce the severity of rare word and out-of-vocabulary problem in inference.

Example of applying source-word decomposition to the Russian word "Последние" (meaning: last) in a sentence is given in Table 3. The application results in a vector of linguistic features: "Последний" (lemma), "amod"(dependency label: an adjectival modifier of a noun), "ADJ" (part-of-speech tag: adjective) and "Animacy = Inan, Case = Acc, Degree = Pos, Number = Plur" (morphological features: inanimate, accusative case, positive degree of comparison, and plural number).

3.3. Feature-Based Transformer Model. In order to decompose source words into tuples of features, we make changes in the encoder of the transformer model, so that it takes vectors of linguistic features as inputs in place of source words. The modified encoder requires four sequences of linguistic features including lemma a_i, dependency label b_i, part-of-speech u_i, and morphological features v_i, for $i = 1, \ldots, n$. Each linguistic-feature tag j is considered as a string in a corresponding dictionary Θ_j. Trainable embeddings $\mathbf{e}_j$ of all linguistic-feature tags are looked up in corresponding dictionaries Θ_j by the modified encoder. As proposed in [23], we apply concatenation operation () on embeddings of all linguistic-feature labels of each source word. The concatenation results in a concatenated embedding corresponding to each source word in sentence. Positional embedding of each source word is then added to the concatenated embedding to form the final embedding representing source word. Given a sequence of final embeddings representing source words in a sentence, following steps of the modified encoder are essentially the same as in the standard encoder.

Table 1: Part-of-speech types and corresponding tags of Russian source words.

Tag	Meaning
ADJ	Adjective
ADP	Adposition
ADV	Adverb
AUX	Auxiliary
CCONJ	Coordinating conjunction
DET	Determiner
INTJ	Interjection
NOUN	Noun
NUM	Numeral
PART	Particle
PRON	Pronoun
PROPN	Proper noun
PUNCT	Punctuation
SCONJ	Subordinating conjunction
SYM	Symbol
VERB	Verb

Table 2: Russian lemma любовь (love) in different cases and numbers.

	Singular number	Plural number
Nominative case	"любовь"	"любови"
Genitive case	"любви"	"любовей"
Dative case	"любви"	"любовям"
Accusative case	"любовь"	"любови"
Instrumental case	"любовью"	"любовями"
Prepositional case	"любви"	"любовяч"

Mathematically, the feature-based transformer model can be represented as a composition of functions as follows.

$$\begin{aligned}
\mathbf{e}_{a_i} &= \phi_a[a_i], \quad \text{for } i = 1, \ldots, n,\\
\mathbf{e}_{b_i} &= \phi_b[b_i], \quad \text{for } i = 1, \ldots, n,\\
\mathbf{e}_{u_i} &= \phi_u[u_i], \quad \text{for } i = 1, \ldots, n,\\
\mathbf{e}_{v_i} &= \phi_v[v_i], \quad \text{for } i = 1, \ldots, n,\\
\mathbf{e}_i &= \phi_2[i], \quad \text{for } i = 1, \ldots, n,\\
\mathbf{c}_i &= \sqrt{d} \times \left(\mathbf{e}_{a_i} \| \mathbf{e}_{b_i} \| \mathbf{e}_{u_i} \| \mathbf{e}_{v_i}\right) + \mathbf{e}_i,\\
\mathbf{C} &= \{\mathbf{c}_i, \quad \text{for } i = 0, \ldots, n\},\\
\mathbf{Z} &= g(\mathbf{C}),\\
y_i &= f(\mathbf{Z}, y_1, y_2, \ldots, y_{i-1}), \quad \text{for } i = 1, \ldots, k.
\end{aligned} \tag{2}$$

In the above equations, notations ϕ_j for $j \in \{a, b, u, v\}$ are the dictionaries of trainable embeddings. It is worth to mention that the sum of dimensions of the embeddings is equal to d to make the concatenated embeddings compatible with the dimension of positional embeddings $\mathbf{e}_i$.

4. Materials and Methods

4.1. Materials. We created our corpus very much in the same way as indicated in the works [27, 28], which are dedicated to study another low-resource language pair in the form of

Table 3: Applying source-word decomposition to Russian word "Цоследние" in a sentence.

	Tag	Meaning
Word	"Цоследние"	Last
In sentence	"в Цоследние Ґоды у Цолитическич Цартий была Цлочая реЦутация в Црессе, и для этоҐо есть веские" Цричины"	"In the last years, political parties had a bad reputation in the press, and there are good reasons for this"
Lemma	"Цоследний"	Last
Dependency label	"Amod"	An adjectival modifier of a noun
Part-of-speech tag	"ADJ"	Adjective
Morphological features	Animacy = Inan	Inanimate
	Case = Acc	Accusative case
	Degree = Pos	Positive degree of comparison
	Number = Plur"	Plural number

Chinese-Vietnamese. First, we picked 33,027 Russian sentences from News Commentary data (http://www.statmt.org/wmt13/training-parallel-nc-v8.tgz) of shared task: Machine Translation of ACL 2013 Eighth Workshop on Statistical Machine Translation. Next, we translated the Russian sentences into Vietnamese. The translation was carried out as follows. First, we used Google translate service to translate all the Russian sentences into Vietnamese. Then, we corrected the translation results, so they not only reflect the meaning of the Russian source sentences but also sound naturally, taking advantage of the fact that we are native speakers of Vietnamese and understand Russian. As a result, we had 33027 Russian-Vietnamese sentence pairs. We then randomly arranged the sentence pairs. From the shuffled corpus, we first took out 1500 sentence pairs to form the testing dataset. After that, we took another 1500 sentence pairs to form the development dataset. The remaining 30027 sentence pairs were used as the training dataset. Summary of the datasets is demonstrated in Table 4 [29]. The summary reveals the huge difference in the dictionary size between Russian and Vietnamese. The number of unique tokens in the Russian training dataset is over 8.5 times larger than the one in the Vietnamese side. The difference can be explained by the fact that Russian is a morphological rich language, while Vietnamese is a noninflectional language. On the other hand, the summary also highlights the difference in average sentence length between Russian and Vietnamese. The number of tokens per Vietnamese sentence is over 1.5 times larger than the one in the Russian side. In other words, to express the same idea, on average, we need to use more Vietnamese words than Russian words.

4.2. Methods. To evaluate feature-based NMT, we performed four experiments. In each experiment, we built and assessed a NMT model. Input and output of each model are presented in Table 5. We used deep learning library PyTorch [30] to build the NMT models with the required input and output by altering an implementation (https://github.com/bentrevett/pytorch-seq2seq) of the state-of-the-art transformer. The source codes of the proposed feature-based and baseline NMT models are provided at GitHub page (https://github.com/ThienCNguyen/Russian-Vietnamse) of the first author.

In the first experiment, we built the baseline W2W model which takes a sequence of Russian words as input and predicts a sequence of Vietnamese words as output. The W2W model is comprised of an encoder and a decoder. The encoder has a 256-dimensional embedding layer, 256-dimensional hidden states, three sublayers consisting of 8-head self-attention layer and 512-dimensional feedforward layer, and dropout layers with level = 10%. The decoder has a similar configuration as the encoder. We used tokenized training and development datasets to train the W2W model. We tokenized Russian sentences in the training and development datasets to produce corresponding sequences of Russian words by using space delimiters between Russian words. We tokenized Vietnamese sentences in the training and development datasets to produce corresponding sequences of Vietnamese words by using a tool provided in [31]. Using Adam optimizer with a learning rate = 0.0005 as reported in [32], we trained the W2W model in 20 epochs of the training dataset. Then, we chose the parameters of the model providing the least cross-entropy loss in the development dataset.

In the second experiment, we built the S2S model which takes a sequence of Russian subwords as input and predicts a sequence of Vietnamese subwords as output. The S2S model has the same configuration and optimization procedure as the baseline W2W model. To produce sequences of subwords for building the model, we tokenized sentences in the training and development datasets by using a tool provided in [33].

In the third experiment, we built the SnF2S model which takes a sequence of Russian subwords and their features (subword tag, lemma, dependency label, part-of-speech tag, and morphological label) as input and predicts a sequence of Vietnamese subwords as output. Subword tag is one of four types: *B, I, E*, and *O*, corresponding to four types: beginning part, inside part, ending part, and full word. Linguistic features of a subword are the same as the ones of containing words which are generated by a deep learning tool Stanza [34]. The SnF2S model is an improvement on the model proposed in [23]. We substituted recurrent neural networks with the state-of-the-art transformer. The SnF2S model also has a similar configuration and optimization procedure as the S2S model except for the encoder embedding layer and dimension of hidden states. The encoder embedding layer is

Table 4: Summary of the parallel datasets used in the study.

Number of	Russian			Vietnamese		
	Training	Development	Testing	Training	Development	Testing
Sentences	30,027	1,500	1,500	30,027	1,500	1,500
Tokens	438,875	21,820	21,941	693,681	34,436	34,651
Tokens per sentence	14.6	14.5	14.6	23.1	23.0	23.1
Unique tokens	46,789	7,520	7,450	5,402	1,985	2,058

Table 5: Input and output of NMT models.

Experiment	Model name	Input unit	Output unit
1	W2W	Word	Word
2	S2S	Subword	Subword
3	SnF2S	Subword and features	Subword
4	F2W	Features only	Word

composed of six embedding sublayers: 352-dimensional subword embedding, 7-dimensional subword-tag embedding, 117-dimensional lemma embedding, 12-dimensional dependency label embedding, 12-dimensional part-of-speech-tag embedding, and 12-dimensional morphological label embedding. We chose the dimension of embeddings, following the ratio recommended in [23]. We applied 512-dimensional hidden states to make them compatible with embedding dimension.

In the second and third experiments with sequences of Vietnamese subwords as output, we applied a postprocessing of concatenating subwords to form a sequence of words in the same way as proposed in [23].

In the fourth experiment, we built the proposed feature-based NMT model called S2F, which takes sequences of features of Russian source words (lemma, dependency label, part-of-speech tag, and morphological label) as input and predicts a sequence of Vietnamese words as output. In turn, the S2F model has a similar configuration and optimization procedure as the SnF2S model except for the encoder embedding layer and dimension of hidden states. The encoder embedding layer is composed of four embedding sublayers: 190-dimensional lemma embedding, 22-dimensional dependency label embedding, 22-dimensional part-of-speech-tag embedding, and 22-dimensional morphological label embedding. We chose the 256-dimension for hidden states to be compatible with embedding dimension.

In all experiments, we used the same assessment procedure for all NMT models. First, we fed Russian sentences of the testing dataset to the models. Then, we compared the predictions by the NMT models with reference Vietnamese sentences in the testing dataset in terms of the lowercase BLEU score which is calculated by the natural language toolkit NLTK [35].

5. Results and Analysis

Primary translation results are provided at GitHub page (https://github.com/ThienCNguyen/Russian-Vietnamse) of the first author. In this section, we analyze translation results for Russian-Vietnamese. We compare the performance of the proposed feature-based NMT with baseline NMT models. We also present human judgment of translation results.

5.1. Machine Judgment. Figure 1 shows the corpus-level BLEU scores of translation results from the testing dataset by the NMT models. We can observe that, among the baseline models, the SnF2S model yields the best result. In comparison with the word-based W2W model, the subword-based S2S model improves by 2.54 BLEU. Compared with the subword-based S2S model, the subword-based feature-added SnF2S model provides an improvement of 3.83 BLEU. This result suggests that we should compare it with the SnF2S model which is the strongest baseline model in order to prove the effectiveness of our proposed model. In comparison with the strongest baseline SnF2S model, our feature-based F2W model outperforms by an impressive 4.83 BLEU. Nevertheless, on the sentence level, the proposed F2W model does not always prove itself better than the SnF2S model. Among 1500 sentences in the testing dataset, the F2W model worsens the translation quality in 41.13% cases, while it improves the BLEU score in 57.6% cases. Detail of the comparison is presented in Figure 2.

5.2. Human Judgment. In addition to machine judgment, we also applied human judgment on translation results by NMT models. We made human analysis to have a more complete assessment on translation results. Specifically, we randomly picked 5 cases from the testing dataset. Here, we present the selected cases and human analysis on translation results. Description of each case consists of a Russian source sentence, its meaning in English, Vietnamese reference, translation results by NMT models, and their corresponding sentence-level BLEU scores.

Table 6 shows translation results by NMT models from a simple source sentence. Although the source sentence is simple, two models, W2W and SnF2S, give wrong translations. Their translations with the meanings "Europe is still in place of Barack Obama, Barack Obama has gone" and "Europe is still impressed with the tragedy of Barak Obama" are far from the initial meaning of the source sentence. On the other hand, two NMT models, S2S and the proposed F2W, perform pretty well for this source sentence. The meaning of the translation by S2S is "Europe is still impressed with the impression of the visit to Obama," which is close to the meaning of the source sentence. The result still has a flaw. Repeated phrase "ấn tượng"g(meaning: impression) in the translation result may make it more difficult

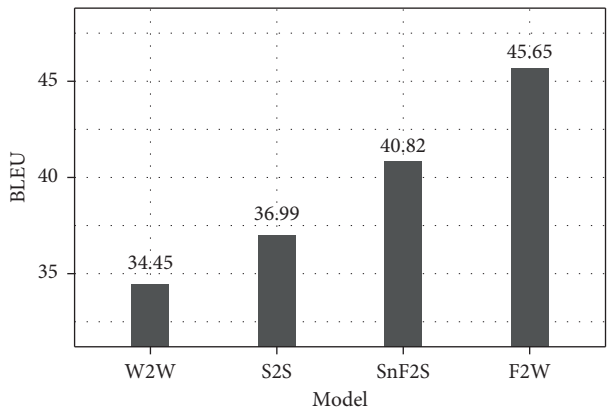


Figure 1: BLEU scores of predicted Vietnamese sentences by NMT models.

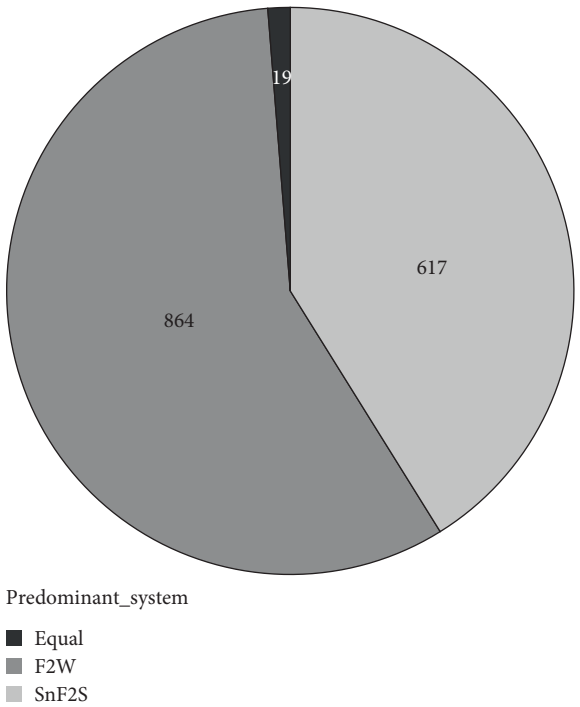


Figure 2: Comparison between F2W and SnF2S models on the sentence level BLEU score.

for us to catch the meaning. Compared to other models, the proposed F2W model gives the best translation result. The meaning of the translation is "Europe is still impressed with the visit of Barack Obama," which bears the closest resemblance with the meaning of the source sentence.

Table 7 shows translation results by NMT models from a more complicated source sentence where the subject has a singular form but plays a plural role. The quality of translation by NMT models is very different in this case. Three baseline models, W2W, S2S, and SnF2S, provide translation results with the wrong meanings "but most books and Stalin's red is a positive light in light," "But most of Stalin's book and press pretend," and "But most of Stalin's books and authors were light cakes under the light." At the same time, our proposed F2W model gives a translation identical to the gold reference.

Table 8 shows translation results by NMT models from a complex source sentence where an infinitive clause is used as subject complement. This long complex sentence is a challenge for NMT models. There is no translation model that gives a good enough result for this case. Furthermore, quality of the translation results by NMT models for this example is the perfect reflection of overall machine judgment on NMT models. Among the baseline models, the SnF2S model gives the best result. Specifically, it partly translates the key phrase "красныч чмеров" (meaning: Khmer Rouge) into "đỏ," while other baseline models mistranslate the phrase. Compared to the best baseline SnF2S model, the proposed F2W model also partly translates that key phrase and improves translation by successfully translating the other key phrase "диШломатическич усилий" (meaning: diplomatic efforts).

Table 9 shows translation results by NMT models from a sentence where the proposed F2W model slightly worsens the translation quality in terms of the BLEU score. In comparison with the best baseline SnF2S model (62.69 BLEU), the proposed F2W model provides a slower BLEU score (60.29 BLEU). From the human perspective, the meaning of the translation result by the F2W model (meaning: we operate with the private sector, not compete with it) is very close to the meaning of the translation result by the SnF2S model (meaning: we work with the private sector, not competing with it) and the reference itself.

Table 10 shows translation results by NMT models from a sentence where the proposed F2W model significantly worsens the translation quality in terms of the BLEU score. In comparison with the best baseline SnF2S model (55.39 BLEU), the proposed F2W model provides a far slower BLEU score (20.69 BLEU). Nevertheless, from the human perspective, the translation result by the F2W model (meaning: from the Persian Gulf, the oil and gas import region in the United States) partially reflects the meaning of the reference, while the SnF2S model mistranslates the Russian source sentence. The meaning of the translation result by the SnF2S model is "a third of oil exported only to the United States."

In overall, both machine and human judgments prove the superiority of the proposed feature-based transformer

Table 6: Translation results by NMT models from a simple source sentence.

Tag	Content
Russian	“евроПа все еще начодится Под вПечатлением визита барака обамы”
Meaning	“Europe is still under the impression of Barack Obama’s visit”
Reference	“châu âu vẫn còn ấn tượng bởi chuyến thăm của Barack Obama”
W2W (19.56)	“châu âu vẫn đang ở vị trí của Barack Obama, Barack Obama đã đi”
S2S (53.73)	“châu âu vẫn còn ấn tượng với sự ấn tượng của chuyến thăm Obama”
SnF2S (45.06)	“châu âu vẫn đang ấn tượng với bi kịch của Barak Obama”
F2W (53.04)	“châu âu vẫn đang ở ấn tượng với chuyến thăm của Barack Obama”

Table 7: Translation results by NMT models from a more complicated source sentence where the subject has a singular form but plays a plural role.

Tag	Content
Russian	“но большинство книГ и авторов изображают сталина в Положительном свете”
Meaning	“But most books and authors portray stalin in a positive light”
Reference	“nhưng hầu hết các cuốn sách và tác giả miêu tả stalin trong một ánh sáng tích cực”
W2W (27.86)	“nhưng hầu hết các sách và màu đỏ của stalin là một ánh sáng tích cực trong ánh sáng”
S2S (36.85)	“nhưng hầu hết cuốn sách và giả vờ báo chí của stalin”
SnF2S (29.19)	“nhưng hầu hết cuốn sách và tác giả của stalin được bánh sáng dưới ánh sáng”
F2W (100)	“nhưng hầu hết các cuốn sách và tác giả miêu tả stalin trong một ánh sáng tích cực”

Table 8: Translation results by NMT models from a complex source sentence where an infinitive clause is used as subject complement.

Tag	Content
Russian	“официальное решение создать трибунал для красныч чмеров является большим достижением После десятилетия диПломатическич усилий”
Meaning	“The official decision to establish a tribunal for the Khmer Rouge is a great achievement after a decade of diplomatic efforts”
Reference	“quyết định chính thức thành lập toà án cho người Khmer đỏ là một thành tựu lớn sau một thập kỷ nỗ lực ngoại giao”
W2W (30.34)	“toà án chính thức để tạo ra toà án <unk> <unk> là một thành tựu lớn sau những nỗ lực ngoại giao ngoại giao”
S2S (16.12)	“quyết định chính thức tạo ra toà án chính thức cho những” gìn mederon là thành tựu lớn sau những nỗ lực ngoại giao
SnF2S (16.28)	“một quyết định chính thức để tạo ra toà án đỏ là thành tựu to lớn sau thập kỷ to”
F2W (38.92)	“quyết định chính thức để tạo ra một toà án đỏ cho các <unk> là thành tựu lớn sau nhiều nỗ lực ngoại giao”

Table 9: Translation results by NMT models from a sentence where the proposed F2W model slightly worsens the translation quality in terms of the BLEU score.

Tag	Content
Russian	“мы работаем с частным сектором, а не конкурируем с ним”
Meaning	“We work with the private sector, not compete with it”
Reference	“chúng tôi hợp tác với khu vực tư nhân, không cạnh tranh với nó”
W2W (27.53)	“chúng tôi đã đối phó với khu vực tư nhân, không phải là một <unk> với nó”
S2S (49.89)	“chúng tôi làm việc với khu vực tư nhân chứ không phải là một người cạnh tranh với họ”
SnF2S (62.69)	“chúng tôi làm việc với khu vực tư nhân, không phải cạnh tranh với nó”
F2W (60.29)	“chúng ta hoạt động với khu vực tư nhân, không phải cạnh tranh với nó”

Table 10: Translation results by NMT models from a sentence where the proposed F2W model significantly worsens the translation quality in terms of the BLEU score.

Tag	Content
Russian	“из ПерсидскоГо залива ПостуПает лишь Пятая часть имПортируемой в соединенные штаты нефти”
Meaning	“Only one fifth of the oil imported into the United States comes from the Persian Gulf”
Reference	“chỉ một phần năm lượng dầu nhập khẩu vào hoa kỳ đến từ vịnh ba tư”
W2W (26.15)	“từ vịnh ba tư chỉ có một phần của hoa kỳ đã được hưởng dầu được hưởng lợi từ dầu”
S2S (29.85)	“trong phần trăm năm, hoa kỳ chỉ làm nhập khẩu dầu”
SnF2S (55.39)	“một phần ba của dầu chỉ xuất khẩu chỉ nhập khẩu hoa kỳ”
F2W (20.69)	“từ vịnh ba tư, khu vực nhập khẩu dầu khí ở hoa kỳ”

model in comparison to other available transformer translation models for translating from Russian into Vietnamese.

6. Conclusions and Perspectives

In this paper, we have successfully integrated linguistic knowledge into the state-of-the-art transformer translation model. We have introduced the feature-based transformer model, which replaces source words by combinations of their features comprising of lemma, dependency label, part-of-speech tag, and morphological label. We have empirically compared the proposed model with other baseline models. Experiment result for the Russian-Vietnamese language pair shows that our model outperforms other models by great distances.

Based on the translation results and our knowledge of the investigated Russian and Vietnamese languages and their relations to other languages, we strongly recommend the feature-based NMT model for building systems translating from highly inflectional synthetic Slavic languages including Russian, Belarusian, Ukrainian, Polish, Bulgarian, Czech, and Serbian into noninflectional analytic languages, such as Vietnamese and Chinese.

Data Availability

The dataset used to support the findings of this study is available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

References

[1] R. Aharoni, M. Johnson, and O. Firat, "Massively multilingual neural machine translation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 3874–3884, Minneapolis, MN, USA, June 2019.

[2] Y. Cheng, L. Jiang, and W. Macherey, "Robust neural machine translation with doubly adversarial inputs," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4324–4333, Florence, Italy, August 2019.

[3] A. Bapna and O. Firat, "Simple, scalable adaptation for neural machine translation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 1538–1548, Hong Kong, China, November 2019.

[4] W. Zhang, Y. Feng, F. Meng, D. You, and Q. Liu, "Bridging the gap between training and inference for neural machine translation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4334–4343, Florence, Italy, August 2019.

[5] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA, May 2015.

[6] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proceedings of the Eighth Syntax, Semantics and Structure in Statistical Translation*, vol. 103, Doha, Qatar, October 2014.

[7] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, Lisbon, Portugal, September 2015.

[8] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 1715–1725, Berlin, Germany, August 2016.

[9] K. Vylomova, T. Cohn, X. He, and G. Haffari, "Word representation models for morphologically rich languages in neural machine translation," in *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, vol. 103, Copenhagen, Denmark, September 2017.

[10] T. Kudo, "Subword regularization: improving neural network translation models with multiple subword candidates," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 66–75, Melbourne, Australia, July 2018.

[11] Y. Wu, M. Schuster, Z. Chen et al., "Google's neural machine translation system: bridging the gap between human and machine translation," *CoRR*, 2016.

[12] M. Johnson, M. Schuster, Q. V. Le et al., "Google's multilingual neural machine translation system: enabling zero-shot translation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017.

[13] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, pp. 5998–6008, Long Beach, CA, USA, December 2017.

[14] A. Vaswani, S. Bengio, E. Brevdo et al., "Tensor2tensor for neural machine translation," in *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas*, vol. 1, pp. 193–199, Boston, MA, USA, March 2018.

[15] H. Sarma, N. Saharia, and U. Sharma, "Development and analysis of speech recognition systems for assamese language using htk," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 17, no. 1, pp. 1–14, 2017.

[16] S. Bhattacharya, D. Manousakas, A. G. C. P. Ramos, S. I. Venieris, N. D. Lane, and C. Mascolo, "Countering acoustic adversarial attacks in microphone-equipped smart home devices," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 2, pp. 1–24, 2020.

[17] P. Koehn and H. Hoang, "Factored translation models," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 868–876, Prague, Czech Republic, June 2007.

[18] A. Birch, M. Osborne, and P. Koehn, "Ccg supertags in factored statistical machine translation," in *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 9–16, Prague, Czech Republic, June 2007.

[19] P. Koehn and J. Schroeder, "Experiments in domain adaptation for statistical machine translation," in *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 224–227, Prague, Czech Republic, June 2007.

[20] P. Koehn, H. Hoang, A. Birch et al., "Moses: open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and*

Poster Sessions, pp. 177–180, Prague, Czech Republic, June 2007.

[21] D. Kolovratnık, N. Klyueva, and O. Bojar, "Statistical machine translation between related and unrelated languages," in *Proceedings of the Conference on Theory and Practice on Information Technologies*, pp. 31–36, Citeseer, Dolný Kubín, Slovakia, September 2009.

[22] S. Huet, E. Manishina, and F. Lefèvre, "Factored machine translation systems for Russian-English," in *Proceedings of the Eighth Workshop on Statistical Machine Translation*, Sofia, Bulgaria, August 2013.

[23] R. Sennrich and B. Haddow, "Linguistic input features improve neural machine translation," in *Proceedings of the First Conference on Machine Translation*, vol. 1, pp. 83–91, Berlin, Germany, August 2016.

[24] M. García-Martínez, O. Caglayan, W. Aransa, A. Bardet, F. Bougares, and L. Barrault, "Lium machine translation systems for wmt17 news translation task," 2017, http://arxiv.org/abs/1707.04499.

[25] M. Weller-Di Marco and A. Fraser, "Modeling word formation in English–German neural machine translation," in *Proceedings of the 58th Annual Meeting of the Association For Computational Linguistics*, pp. 4227–4232, January 2020, https://www.aclweb.org/anthology/2020.acl-main.389/.

[26] M.-C. De Marneffe, T. Dozat, N. Silveira et al., "Universal stanford dependencies: a cross-linguistic typology," *Language Resources and Evaluation*, vol. 14, pp. 4585–4592, 2014.

[27] P. Tran, D. Dinh, and H. T. Nguyen, "A character level based and word level based approach for Chinese-Vietnamese machine translation," *Computational Intelligence and Neuroscience*, vol. 2016, Article ID 9821608, 7 pages, 2016.

[28] P. Tran, D. Dinh, and L. H. B. Nguyen, "Word re-segmentation in Chinese-Vietnamese machine translation," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 16, no. 2, pp. 1–22, 2016.

[29] T. Nguyen, H. Nguyen, and P. Tran, "Mixed-level neural machine translation," *Computational Intelligence and Neuroscience*, vol. 2020, Article ID 8859452, 7 pages, 2020.

[30] A. Paszke, S. Gross, F. Massa et al., "Pytorch: an imperative style, high-performance deep learning library," in *Proceedings of the Advances In Neural Information Processing Systems*, pp. 8024–8035, Vancouver, Canada, December 2019.

[31] T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras, and M. Johnson, "Vncorenlp: a Vietnamese natural language processing toolkit," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 56–60, New Orleans, LA, USA, January 2018.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171–4186, Minneapolis, MN, USA, June 2019.

[33] R. Sennrich, B. Haddow, and A. Birch, "Edinburgh neural machine translation systems for wmt 16," in *Proceedings of the First Conference on Machine Translation*, vol. 2, pp. 371–376, Berlin, Germany, August 2016.

[34] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages," in *Proceedings of the 58th Annual Meeting of the Association For Computational Linguistics*, July 2020, https://www.aclweb.org/anthology/2020.acl-demos.14/.

[35] E. Loper and S. Bird, "NLTK: the natural language toolkit," in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies For Teaching Natural Language Processing and Computational Linguistics*, pp. 63–70, Philadelphia, PA, USA, July 2002.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 8942096, 8 pages
https://doi.org/10.1155/2020/8942096

Research Article

Open-Ended Coaxial Cable Selection for Measurement of Liquid Dielectric Properties via the Reflection Method

Jingchen Wang, Eng Gee Lim, Mark Paul Leach, Zhao Wang, and Ka Lok Man

School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China

Correspondence should be addressed to Eng Gee Lim; enggee.lim@xjtlu.edu.cn

Received 30 July 2020; Accepted 24 August 2020; Published 6 October 2020

Guest Editor: Mohamed Nayel

An open-ended coaxial cable is used to measure the dielectric properties of lossy liquid. The method which is based on the measurement of the reflection coefficient of the open-ended cable makes it easy to operate and postprocess. To meet the accuracy requirements, the dimensions of the coaxial cable need to be taken into consideration; therefore, it is necessary to select an appropriate coaxial cable for the measurement. This paper investigates the influence of cable dimensions on dielectric measurement accuracy. With careful choice of the coaxial cable, the relative error of calculated results can be less than 0.1%.

1. Introduction

Obtaining information regarding material dielectric properties is an important part of material analysis. Multiple approaches are available to acquire this information as discussed in [1]. In medical dosimetry, biological tissue properties such as dielectric constant, conductivity, and loss factor are of significance [2, 3]. Considering the increasing interest in dielectric properties of biomedical tissues due to advances in medical applications and for research, it is necessary to find an effective solution for the measurement of these dielectric properties. A commonly used method to determine the complex permittivity of materials at radio and microwave frequencies [4–7] is based on the open-ended coaxial cable due to its relative simplicity and accuracy [8]. The calibration of an open-ended coaxial probe for dielectric measurements is proposed in [9, 10].

Open-ended coaxial lines are particularly attractive for in vivo or in vitro measurements of biological material [11] as they offer merits including accuracy, the foundation of a simplified equivalent circuit with the form of capacitive impedance [12], and negligible radiation losses [5].

The capacitive nature of the cable changes with cable size; thus, there is an optimum cable size at a given frequency and permittivity, which offers the minimum measurement uncertainty [13]. Using a reference liquid, with well-known dielectric properties, the cable can be optimized to provide high measurement accuracy. Measurements are made of an unknown liquid using a comparison technique to a known reference liquid. In this work, liquid properties relating to the human body are of primary interest, though the technique could be extended to measure the dielectric properties of any liquid over a wide frequency range. In the literature, the complex dielectric structure of the body is represented by one set of properties: relative permittivity $\varepsilon_r = 56$ and conductivity $\sigma = 0.8$ S/m at 403 MHz. Being able to reproduce liquids with the same properties is a common problem, for which the process described here could provide a reliable, rapid, and cost-effective solution.

In Section 2, the equivalent circuit for the measurement is illustrated. From the circuit, the complex permittivity of the object under test is analysed. In Section 3, investigations into the properties of the coaxial cable are performed. With the knowledge that the cable radius and length are related to the reflection coefficient and conductivity, analyses of how the accuracy of the measurements is influenced by the coaxial cable's dimensions are provided. Finally, conclusions are drawn in Section 4.

2. Equivalent Circuit

Several conventional methods are applied in permittivity measurements, such as the lumped capacitance method [13], short monopole antenna [7], and quasi-static analysis [14]. The method used in this paper is the open-ended coaxial cable reflection method.

The open-ended coaxial cable reflection method is based on the measurement of the complex reflection coefficient at a single network port. From this measurement, the complex dielectric constant of the object under test can be calculated. The conductivity of the object under test can also be found based on the relationship between the measured reflection parameters and the dielectric properties [15]. The coaxial cable has an inner conductor surrounded by a dielectric material (Teflon) and an outer conductor [16]. The configuration of the coaxial cable and the measurement system is shown in Figure 1.

An equivalent circuit for the measurement system is shown in Figure 2. The capacitance C_f represents the electric field concentration, which is inside the filled dielectric material part of the coaxial cable. The capacitance $C(\varepsilon)$ is the fringing field concentration in the dielectric, and the conductance $G(\varepsilon)$ is radiation into the dielectric surrounding the cable. The value of the conductance (G) is frequency (f) dependent. The equivalent admittance of the open-ended coaxial cable Y can be written as [18]

$$Y = j\omega C_f + j\omega C(\varepsilon) + G = j\omega C_f + j\omega C_0\varepsilon + G_0\varepsilon^{5/2}, \quad (1)$$

where C_0 represents the capacitance of the air-filled parallel plate capacitance, ε represents the permittivity of the object under test, and G_0 represents the external radiation conductance of a coaxial cable. From transmission line theory [7], terminal admittance Y based on the reflection coefficient Γ can be defined as follows:

$$Y = \frac{1-\Gamma}{1+\Gamma}\frac{1}{Z_0}, \quad (2)$$

where Z_0 represents the characteristic impedance of the coaxial line (50 Ω). The external radiation conductance of a coaxial cable is very small. As a result, it can be ignored ($G_0 \approx 0$). Hence, the relationship between the measured complex permittivity of the object under test and the reflection coefficient Γ is as follows:

$$Y = \frac{1-\Gamma}{1+\Gamma}\frac{1}{Z_0} = j\omega C_f + j\omega C_0\varepsilon. \quad (3)$$

The relative dielectric constant ε' and the loss factor ε'' are calculated using (4) and (5):

$$\varepsilon' = \frac{1}{2\pi f Z_0 C_0} \times \frac{-2|\Gamma|\sin(\varphi)}{1+2|\Gamma|\cos(\varphi)+|\Gamma|^2} - \frac{C_f}{C_0}, \quad (4)$$

$$\varepsilon'' = \frac{1}{2\pi f Z_0 C_0} \times \frac{1-|\Gamma|^2}{1+2|\Gamma|\cos(\varphi)+|\Gamma|^2}, \quad (5)$$

where Γ and φ are the modulus and phase of the input reflection coefficient, respectively. The complex permittivity ε of the object under test and the relationship between loss factor ε'' and conductivity σ can be expressed as follows:

$$\varepsilon = \varepsilon' - j\varepsilon'', \quad (6)$$

$$\sigma = \omega\varepsilon_0\varepsilon'', \quad \varepsilon_0 = 8.548 \times 10^{-12}\ \text{F/m}. \quad (7)$$

The relationship between the sample capacitance measured by this method and the dielectric constant is linear, which can be found in [19]. In equations (4) and (5), the values of C_f and C_0 are calculated using equations (8) and (9) and the dielectric properties of the reference materials used, such as the deionized water or methanol.

$$C_f = \frac{1}{2\pi f Z_0} \frac{-2|\Gamma|\sin(\varphi)}{1+2|\Gamma|\cos(\varphi)+|\Gamma|^2} - \varepsilon' C_0, \quad (8)$$

$$C_0 = \frac{1}{2\pi f Z_0 \varepsilon''} \frac{1-|\Gamma|^2}{1+2|\Gamma|\cos(\varphi)+|\Gamma|^2}. \quad (9)$$

The complex permittivity of the object under test can be obtained using (4) and (5) and the measured reflection coefficient from the object under test. Finally, the dielectric constant and conductivity of the object under test can be calculated from the complex permittivity.

3. Permittivity Measurements

3.1. The Influence of Coaxial Cable Radius. A previous work has shown that the use of a reference liquid with similar dielectric properties to that of the test subject significantly improves measurement accuracy [18]. The dielectric properties of a body-equivalent phantom liquid (used to represent the average dielectric properties of the entire human body) are $\varepsilon_r = 56$ and $\sigma = 0.8$ S/m at 403 MHz. To measure the body-equivalent phantom liquid, for the purposes of simulation and investigation, a reference liquid with dielectric properties of $\varepsilon_r = 54$ and $\sigma = 0.6$ S/m is selected (close to the desired values). To observe the effect of the various cable radii, three radius (R) sizes of the cable are used. Each of the cables has an impedance of 50 Ω and a dielectric constant of 2.17, and they all have a fixed length of 100 mm.

The cable is located within the reference liquid, as shown in Figure 3. Commercial software (CST Studio Suite) has been used to simulate the measurement procedure and calculate the complex permittivity. For each radius of the coaxial cable, the modulus and phase of the reflection coefficient against frequency are shown in Figure 4. The modulus of the reflection coefficient shows a gentle increase with frequency for each cable. The phase of the reflection coefficient covers more than one cycle over the frequency range.

From these results, values of C_f and C_0 can be calculated using equations (6) and (7), and they are shown in Figure 5. It can be seen that the values of capacitance are different for each cable. Furthermore, the capacitances have a peak value within the frequency range. The peak value occurs at the frequency where the phase of the reflection coefficient is equal to −180° and can be seen from Figure 5.

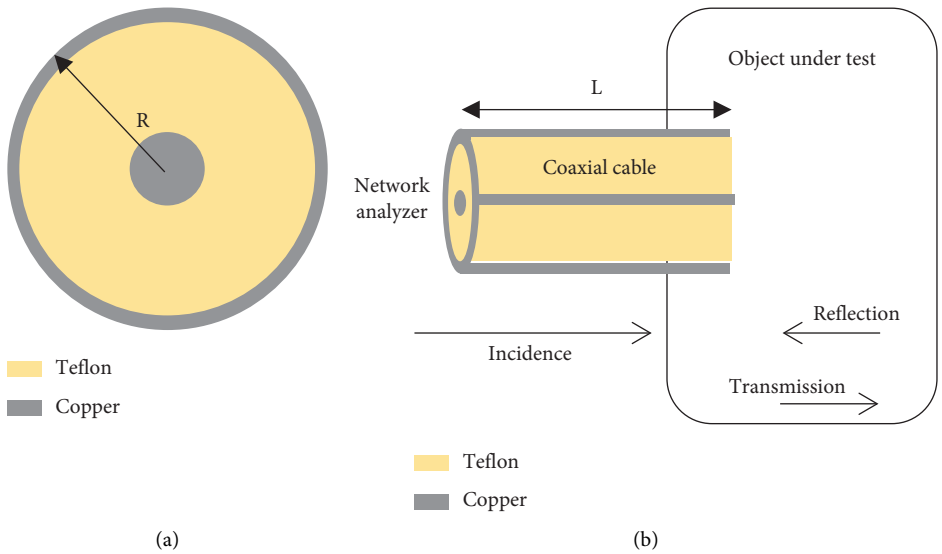


Figure 1: (a) The configuration of a coaxial cable. (b) Configuration of a permittivity measurement system using reflection methods [17].

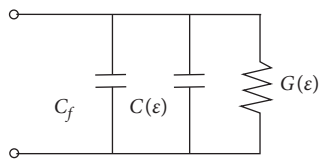


Figure 2: Cable equivalent circuit.

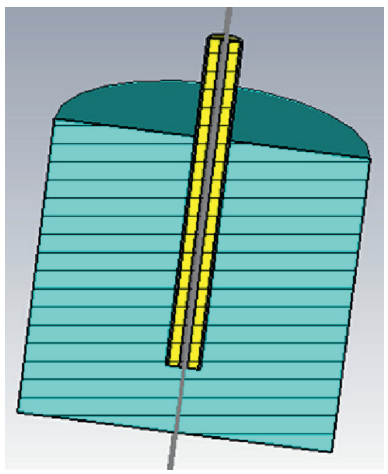


Figure 3: Cross section of the measurement model.

The same process is then followed using the liquid under test. The modulus and phase of the reflection coefficient against frequency obtained from the new simulation for each of the coaxial cables are shown in Figure 6.

As would be expected, the modulus of the reflection coefficient decreases with respect to the results obtained for the reference liquid due to the increase in liquid permittivity and conductivity. The dielectric constant and conductivity of the liquid under test can be predicted from the reflection coefficient using the calculated complex permittivity, as shown in Figure 7.

From Figure 7, it can be seen that cable B with a radius of 2.5 mm offers the best accuracy in comparison with the actual values ($\varepsilon_r = 56$, $\sigma = 0.8$ S/m). The cable with a radius of 1 mm does not function suitably; therefore, it could not be used as a dielectric probe. It is clear from the results that the radius of the cable has a significant influence on the accuracy of the measurement. There is significant variation in the simulation for all 3 cables over the frequency range 200–600 MHz; this is caused by reaching the resonant frequency of the probe in the liquid. This cannot be predicted without knowing the dielectric properties of the liquid in advance; hence, measurements are made over a wide band of frequencies, and the dielectric characteristics are calculated at frequencies between resonances in the region where the calculated dielectric properties appear approximately linear.

3.2. Influence of Coaxial Cable Length. The influence of the cable length (L) on measurement accuracy is now investigated using the cable with 2.5 mm radius from the first investigation with lengths of 100 mm, 50 mm, and 15 mm. The modulus and phase of the reflection coefficient obtained from the simulation for each cable length in the reference liquid are shown in Figure 8.

It can be seen from Figure 8 that the length of the cable has little influence on the modulus of the reflection coefficient; however, it has a significant influence on its phase. The shorter cable has a smaller phase range, whereas the longer cable has a larger phase range.

The values of C_f and C_0 are again calculated using equations (6) and (7), as shown in Figure 9. It is seen that the cables with lengths of 100 mm and 50 mm have a peak value of capacitance at the frequency where the phase is equal to -180°. The 15 mm long cable has a stable value of capacitance over the whole simulation frequency range. This results from the phase range being smaller than 180°.

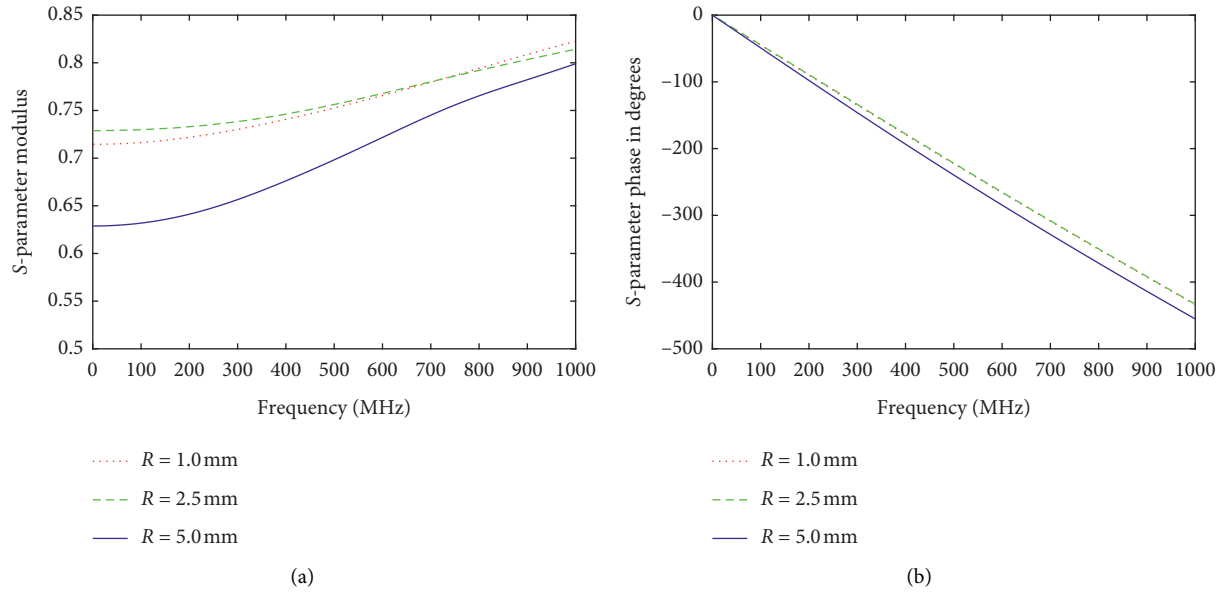


Figure 4: (a) The modulus and (b) phase of the reflection coefficient in the reference liquid for coaxial cables with different radii.

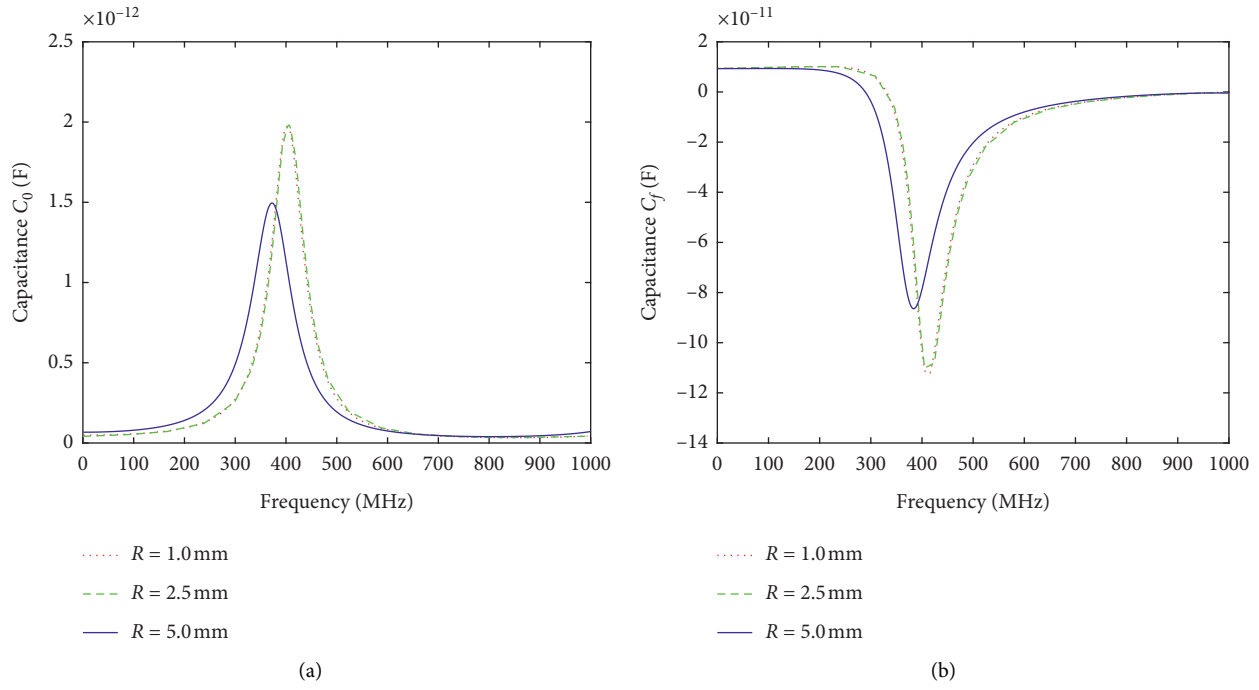


Figure 5: (a) Calculated value of C_0 and (b) calculated value of C_f for coaxial cables with different radii when cables are located in the reference liquid.

Similarly, the simulated modulus and phase of the reflection coefficient for the liquid under test for each cable length are shown in Figure 10. Similar trends in the modulus and phase of the reflection coefficient are observed as that in the simulation using the reference liquid. The reflection coefficient modulus increases from 0.64 to 0.77 for each of the cables over the simulation frequency range. However, the change in phase is affected dramatically by cable length. The 15 mm cable has less than 180° of reflection coefficient phase variation, whereas the 50 mm and 100 mm extend far beyond that.

The complex permittivity is again used to find the dielectric constant and conductivity of the liquid under test, as shown in Figure 11. It can be seen that when the length of the coaxial cable is 15 mm, the average of the calculated results ($\varepsilon_r = 56.55$, $\sigma = 0.83$ S/m) are the closest to the true dielectric

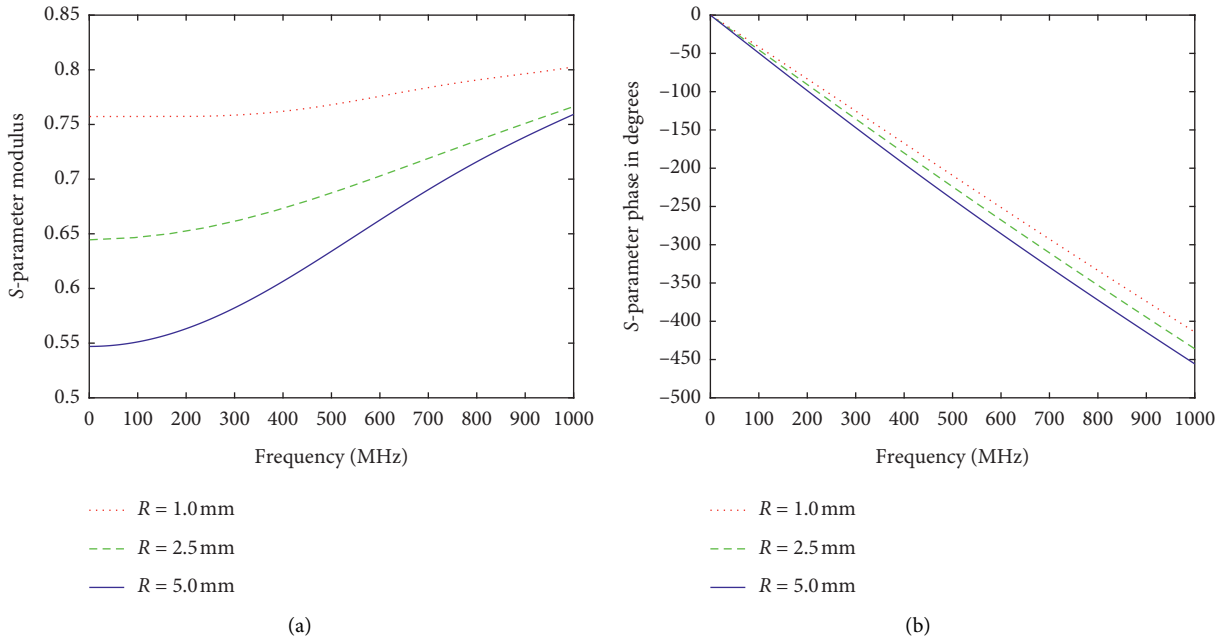


Figure 6: (a) The modulus and (b) phase of the reflection coefficient for the liquid under test with different cable radii.

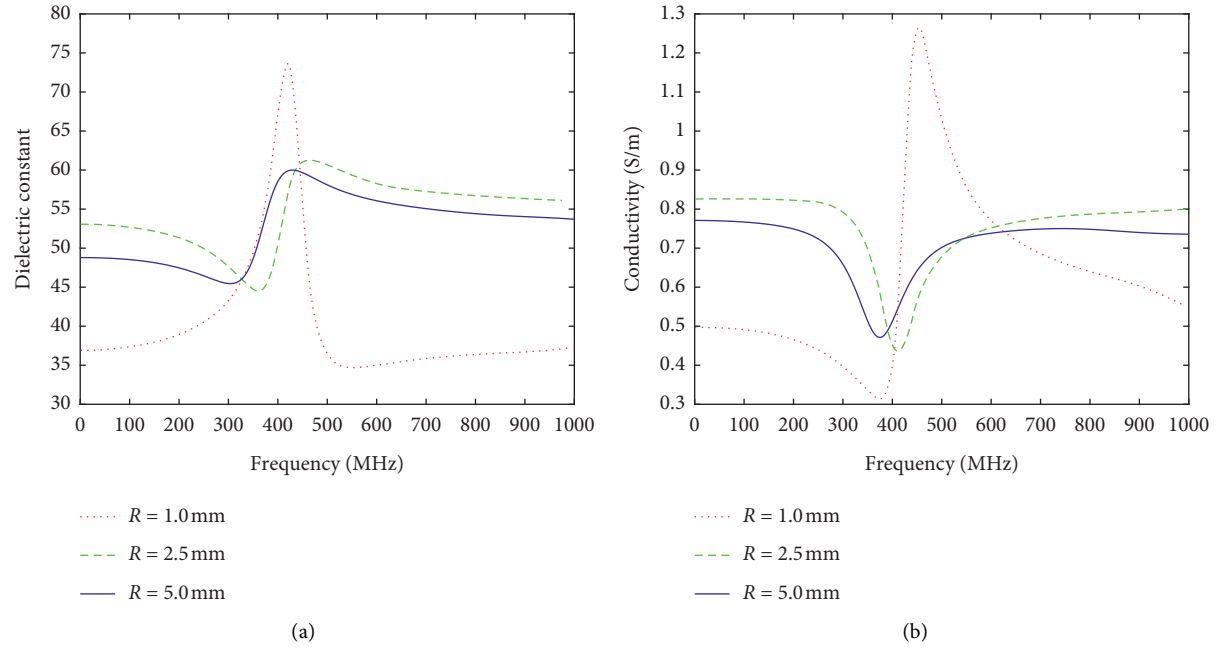


Figure 7: (a) Calculated dielectric constant ε_r and (b) calculated conductivity σ for the liquid under test for each coaxial cable.

constant and conductivity of the liquid under test ($\varepsilon_r = 56$, $\sigma = 0.8$ S/m). Over the simulation frequency range, the predicted values of dielectric constant and conductivity for the liquid under test are generally constant when using the 15 mm cable. Conversely, there is a big change in both the dielectric constant and the conductivity when using the cables with lengths of 50 mm and 100 mm. Therefore, cable length has a significant impact on measurement accuracy.

3.3. Analysis and Discussion. Through simulation, the relationships between the coaxial cable radius and length with conductivity and reflection coefficient prediction accuracy have been investigated and the importance of radius/length is emphasized in this regard.

The radius of the probe is related to the equivalent capacitance C_f and C_0, so the optimum cable radius can guarantee the accuracy of the calculated results. The most

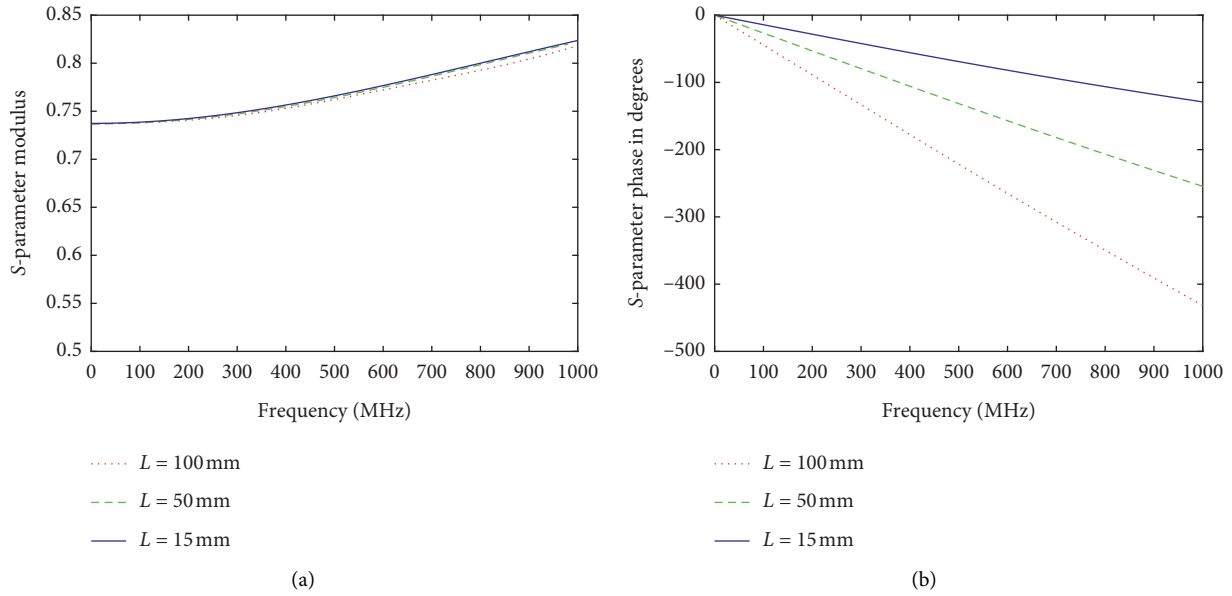


Figure 8: (a) The modulus and (b) phase of the reflection coefficient for the reference liquid with varying cable length.

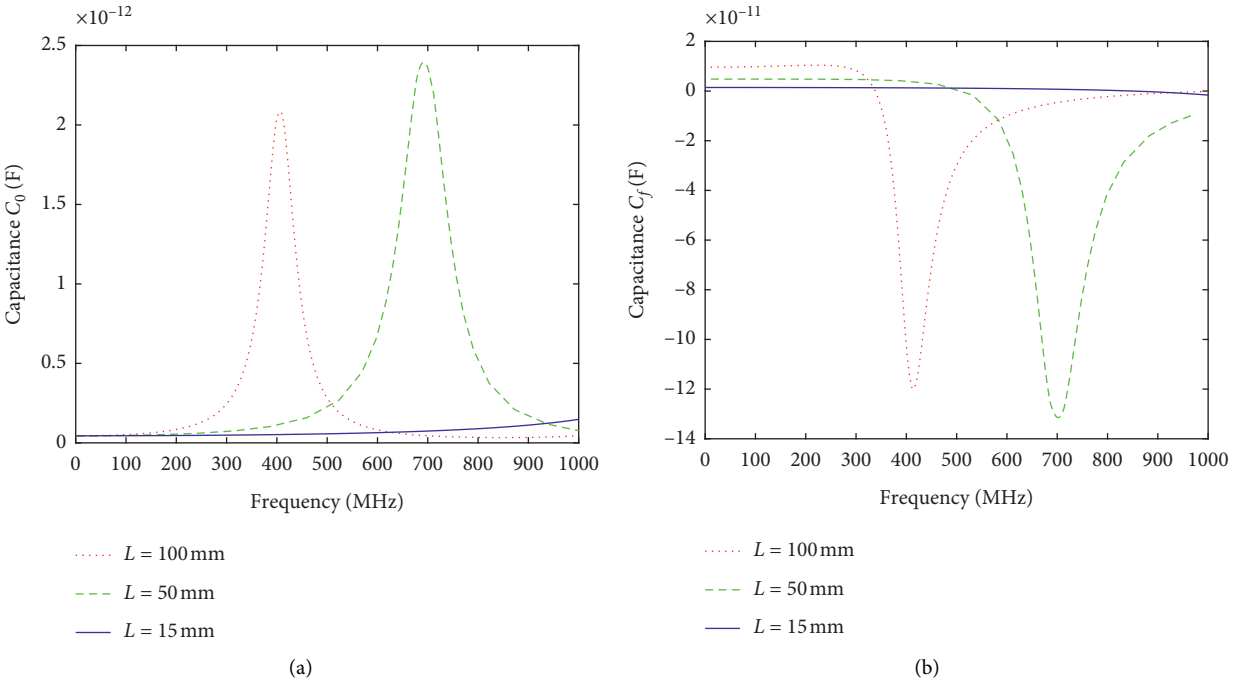


Figure 9: Calculated values of (a) C_0 and (b) C_f with varying cable length.

effective choice of cable radius was found to be 2.5 mm in this case. The length of the probe is related to the equivalent length of the coaxial cavity in the equivalent circuit model and is also related to frequency. A shorter probe has a longer period, which can lead to more stable results over a restricted frequency range. The optimal value of the cable length in this work was found to be 15 mm.

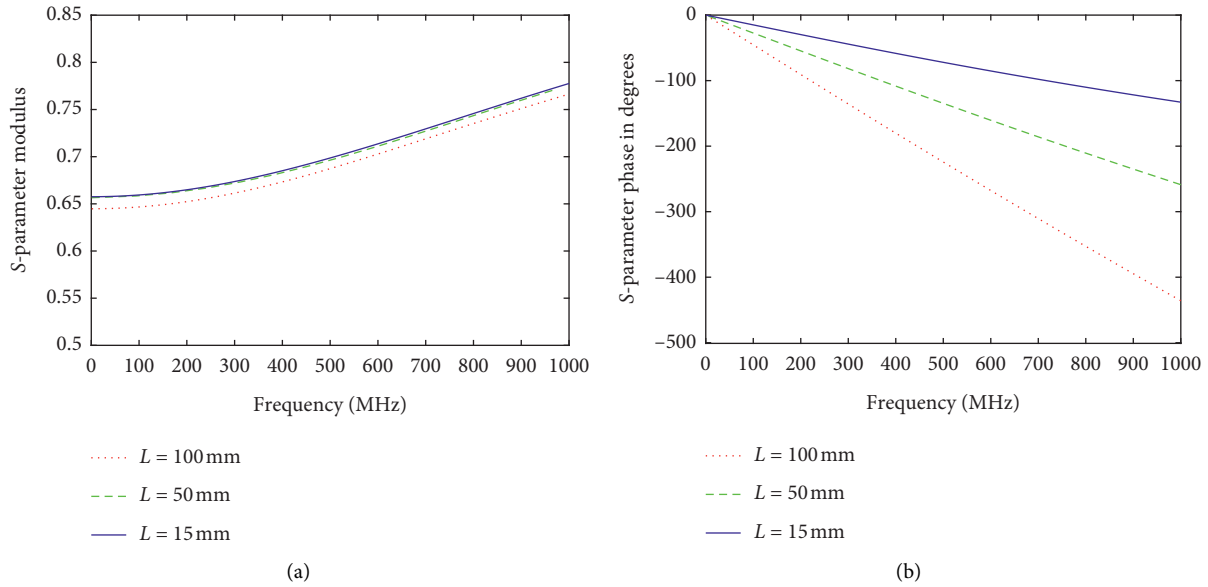


Figure 10: (a) Modulus and (b) phase of the reflection coefficient for the liquid under test with varying cable length.

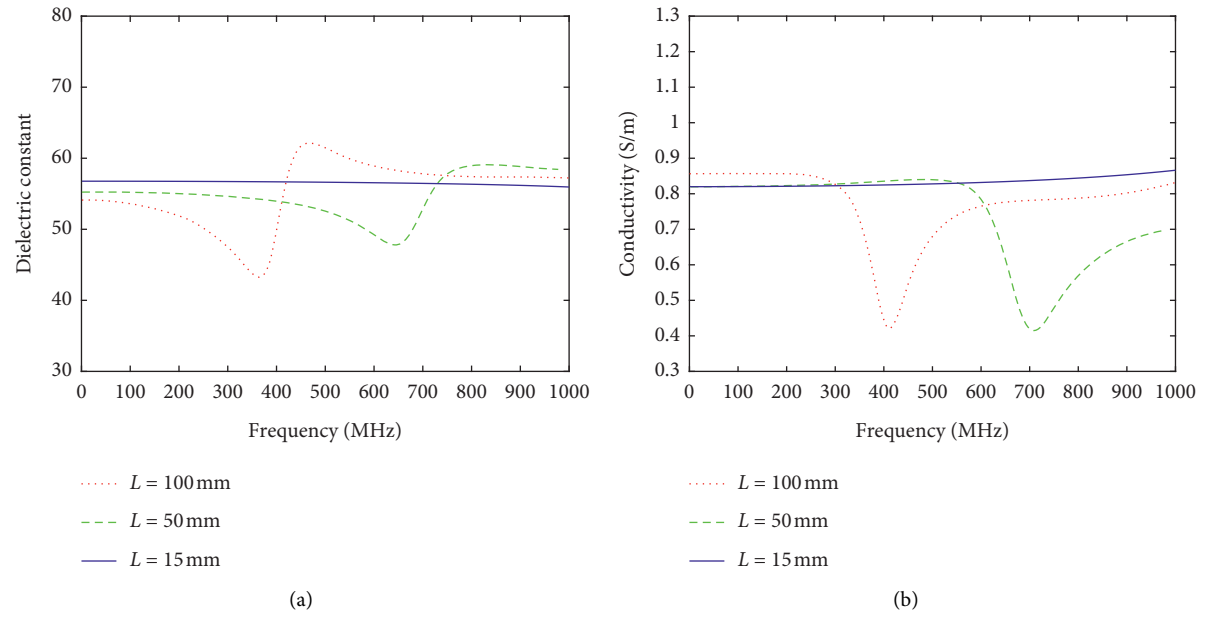


Figure 11: (a) Dielectric constant ε_r and (b) conductivity σ of object under test with varying cable length.

4. Conclusion

This paper investigates the influence of the dimensions of a coaxial cable on the accuracy of measured liquid dielectric properties from two aspects: radius and length. It was found that the accuracy of the cable has a strong dependence on the radius, with very small radii offering no accuracy. It was also found that decreasing cable length improved accuracy. The measured results can be close to the actual value over a restricted frequency range after the dimensions of cables are appropriately selected. It can be seen that, for the coaxial cable, selection of cable size and measurement frequency range are important factors when measuring the dielectric properties of a liquid using the return loss method.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was partially supported by the Natural Science Foundation of Jiangsu Province (grant no. BK20131183) and Suzhou Science and Technology Bureau (grant no. SYG201211) and partially supported by the AI University Research Centre (AI-URC) through XJTLU Key Programme Special Fund (KSF-P-02) and XJTLU Research Development Fund (PGRS-13-03-06, RDF-14-03-24, RDF-14-02-48, and PGRS2006007). The authors would like to express their sincere gratitude to CST AG for providing the CST STUDIO SUITE® electromagnetic simulation software package under the China Key University Promotion Program and Suzhou Municipal Key Lab for Wireless Broadband Access Technologies in the Department of Electrical and Electronics Engineering, Xi'an Jiaotong-Liverpool University, for research facilities.

References

[1] B. Neves, "Permittivity/dielectric constant: do the math," *Circuitree*, vol. 116, 1996.

[2] G. Brix, M. Reinl, and G. Brinker, "Sampling and evaluation of specific absorption rates during patient examinations performed on 1.5 Tesla MR systems," *Magnetic Resonance Imaging*, vol. 19, no. 6, pp. 769–779, 2001.

[3] P. A. Bottomley, R. W. Redington, W. A. Edelstein, and J. F. Schenck, "Estimating radiofrequency power deposition in body NMR imaging," *Magnetic Resonance in Medicine*, vol. 2, no. 4, pp. 336–349, 1985.

[4] T. P. Marsland and S. Evans, "Dielectric measurements with an opened coaxial probe," *IEE Proceedings*, vol. H134, no. 4, pp. 341–349, 1987.

[5] G. S. Smith and J. D. Nordgard, "Measurement of the electrical constitutive parameters of materials using antennas," *IEEE Transactions on Antennas & Propagation*, vol. AP-33, pp. 783–792, 1985.

[6] M. M. Brady, S. A. Symons, and S. S. Stuchly, "Dielectric behaviour of selected animal tissues in vitro at frequencies from 2 to 4 GHz," *IEEE Transactions on Biomedical Engineering*, vol. BME-28, pp. 305–307, 1981.

[7] E. C. Burdette, F. L. Cain, and J. Seals, "In vivo probe measurement technique for determining dielectric properties at VHF through microwave frequencies," *IEEE Transactions on Microwave Theory and Techniques*, vol. 28, pp. 414–427, 1980.

[8] H. L. Gong, S. D. Chen, and D. J. Qian, "Rebuilt and modified model for complex permittivity measurement using the coaxial line," *Information and Electronic Engineering*, vol. 6, no. 6, pp. 448–451, 2008.

[9] G. P. Otto and W. C. Chew, "Improved calibration of a large open-ended coaxial probe for dielectric measurements," *IEEE Transactions on Instrumentation and Measurement*, vol. 40, no. 4, pp. 742–746, 1991.

[10] L. Zhang, P. G. Liu, X. T. Shi et al., "A comparative study of a calibration method for measuring the dielectric properties of biological tissues on electrically small open-ended coaxial probe," in *Proceedings of the 2012 International Conference on Biomedical Engineering and Biotechnology*, Macao, China, May 2012.

[11] T. W. Athey and M. A. Stuchly, "Measurement of radio frequency permittivity of biological tissues with an open-ended coaxial line (parts 1 and 2)," *IEEE Transactions on Microwave Theory and Techniques*, vol. MTT-30, pp. 82–92, 1982.

[12] M. A. Stuchly, M. M. Brady, S. S. Stuchly, and G. Gajda, "Equivalent circuit of an open-ended coaxial line in a lossy dielectric," *IEEE Transactions on Instruction and Measurement*, vol. IM-31, no. 2, pp. 116–119, 1982.

[13] M. A. Stuchly and S. S. Stuchly, "Coaxial line reflection methods for measuring dielectric properties of biological substances at radio and microwave frequencies—a review," *IEEE Transactions on Instrumentation and Measurement*, vol. IM-29, no. 3, pp. 176–183, 1980.

[14] D. K. Misra, "A quasi-static analysis of open-ended coaxial lines," *IEEE Transactions on Microwave Theory and Techniques*, vol. MTT-35, no. 10, pp. 925–928, 1987.

[15] L. Zhang, "Study on measurement of biological tissue dielectric properties in high frequency using reflection method on the open-ended coaxial," Master thesis, National University of Defense Science and Technology, Changsha, China, 2011.

[16] T. T. Khan, "A method for measuring dielectric properties of non-magnetic liquids and predicting their contamination level," in *Proceedings of the 2012 IEEE/OSA/IAPR International Conference on Informatics, Electronics & Vision*, pp. 445–450, Dhaka, Bangladesh, 2012.

[17] J. Wang, M. Leach, E. G. Lim, and Z. Wang, "An open-ended coaxial cable reflection method for measuring dielectric properties of biological liquid," in *Proceedings of the 7th International Conference on Information, System and Convergence Applications*, Ho Chi Minh, Vietnam, February 2020.

[18] D. Berube, F. M. Ghannouchic, and P. Savard, "A comparative study of four open-ended coaxial probe models for permittivity measurements of lossy dielectric biological materials at microwave frequencies," *IEEE Transactions on Microwave Theory and Techniques*, vol. 44, no. 10, pp. 192-193, 1996.

[19] V. Hartwig, G. Giovannetti, N. Vanello et al., "An electrodeless system for measurement of liquid sample dielectric properties in radio frequency band," in *Proceedings of the 2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, New York, NY, USA, August 2006.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 4724310, 13 pages
https://doi.org/10.1155/2020/4724310

Research Article

Improved ORB-SLAM2 Algorithm Based on Information Entropy and Image Sharpening Adjustment

Kaiqing Luo,[1,2,3] **Manling Lin**,[1] **Pengcheng Wang**,[1] **Siwei Zhou**,[1] **Dan Yin**,[1] **and Haolan Zhang**[4]

[1]*School of Physics and Telecommunication Engineering, South China Normal University, Guangzhou 510006, China*
[2]*Guangdong Provincial Engineering Research Center for Optoelectronic Instrument, South China Normal University, Guangzhou 510006, China*
[3]*SCNU Qingyuan Institute of Science and Technology Innovation Co., Ltd., Qingyuan 511517, China*
[4]*The SCDM Center, NingboTech University, Ningbo 315100, China*

Correspondence should be addressed to Kaiqing Luo; kqluo@scnu.edu.cn and Haolan Zhang; haolan.zhang@nit.zju.edu.cn

Received 31 July 2020; Revised 27 August 2020; Accepted 7 September 2020; Published 23 September 2020

Academic Editor: Sanghyuk Lee

Simultaneous Localization and Mapping (SLAM) has become a research hotspot in the field of robots in recent years. However, most visual SLAM systems are based on static assumptions which ignored motion effects. If image sequences are not rich in texture information or the camera rotates at a large angle, SLAM system will fail to locate and map. To solve these problems, this paper proposes an improved ORB-SLAM2 algorithm based on information entropy and sharpening processing. The information entropy corresponding to the segmented image block is calculated, and the entropy threshold is determined by the adaptive algorithm of image entropy threshold, and then the image block which is smaller than the information entropy threshold is sharpened. The experimental results show that compared with the ORB-SLAM2 system, the relative trajectory error decreases by 36.1% and the absolute trajectory error decreases by 45.1% compared with ORB-SLAM2. Although these indicators are greatly improved, the processing time is not greatly increased. To some extent, the algorithm solves the problem of system localization and mapping failure caused by camera large angle rotation and insufficient image texture information.

1. Introduction

Simultaneous Localization and Mapping [1] (SLAM) is described as follows: the robot enters an unknown environment, uses laser or visual sensors to determine its own posture information, and reconstructs a three-dimensional map of the surrounding environment in real time. The system is a hot spot in robot research at present, and it has important theoretical significance and application value for the realization of autonomous control and mission planning of robots.

The current SLAM system is mainly divided into two categories according to different sensor types: laser SLAM and visual SLAM. A SLAM that uses lidar as an external sensor is called a laser SLAM, and a SLAM that uses a camera as an external sensor is called a visual SLAM. The advantage of a lidar is its wide viewing range, but the disadvantage is the high price. The angular resolution is not high enough, which affects the accuracy of modelling. The sensor camera of the visual SLAM has the characteristics of high cost performance ratio, wide application range, and rich information collection. Therefore, the visual SLAM has developed rapidly since the 21st century.

According to the classification of visual sensors, visual SLAM systems divided into monocular SLAM, binocular SLAM, and RGB-D SLAM. According to the image processing methods, they are divided into direct method and indirect method, such as feature point method and outline feature method. According to the construction of the map, the degree of sparseness can be divided into sparse, dense, semidense, etc. The iconic research results of visual SLAM are Mono-SLAM, PTAM, ORB-SLAM, ORB-SLAM2, etc.

In 2007, Andrew Davison proposed Mono-SLAM [2], which is the real-time implementation of SfM (Structure from Motion), so it is also known as Real-Time Structure from Motion. Mono-SLAM is a SLAM based on probabilistic mapping and has closed-loop correction function. However, Mono-SLAM can only handle small scenes in real time, and only has good effects in small scenes. In the same year, Georg Klein and David Murray proposed PTAM [3] (Parallel Tracking and Mapping), and the innovation is to divide the system into two threads: tracking and mapping, and the concept of key frames is proposed. PTAM is no longer a processing sequence, but a key frame that contains a large amount of information. In 2015, MUR-ARTAL R. et al. proposed ORB-SLAM. In 2016, ORB-SLAM2 was also proposed. ORB-SLAM [4] is an improved monocular SLAM system based on feature points of PTAM. The algorithm can run quickly in real time. As the indoor environment and the wide outdoor environment are robust to vigorous sports, ORB-SLAM2 [5] increases the scope of application on the basis of ORB-SLAM. It is a set that is based on monocular, binocular, and RGB-D. ORB-SLAM2 is more accurate than the previous solution and can work in real time on a standard CPU.

ORB-SLAM2 has improved greatly in running time and accuracy, but there are still some problems to be solved [6, 7]. The feature point extraction is poorly robust in an environment with sudden changing in lighting, too strong or too weak light intensity and too little texture information, which cause the problem of system tracking failure under dynamic environment; the feature points lose when the camera rotates at a large angle. Common cameras often produce image blur when moving quickly, and solving motion blur becomes an important direction in visual SLAM. Image restoration technology can reduce the ambiguity of the image and can restore it from the original image to a relatively clear image. The deblurring method is based on the maximum posterior probability and the norm of sparse representation is used for deblurring [8]. The computational efficiency is improved, but the computational cost is still very large. The image is segmented into many image blocks with the same fuzzy mode, and the point diffusion function of each image block is predicted by convolution neural network. Finally, the clear image is obtained by deconvolution, but the processing time is too long [9]. The literature [10] uses multiscale convolutional neural networks to restore clear images in end-to-end form and optimizes the results by multiscale loss function. Because the same network parameters are used for different fuzzy kernels, the network model is too large. The literature [11] reduces the network model and designs a spatial change cyclic neural network, which can achieve good deblurring effect in dynamic scene, but the generalization ability is general. It also has semantic segmentation architectures and deep learning methods based on deep Convolutional Neural Network (CNN) for semantic segmentation to identify relevant objects in the environment [12].

Because the SLAM system needs real time, the computation of neural network is large, the processing time is long, and the generality is lacking, and another method is adopted in this paper. The algorithm in this paper starts from the perspective of increasing the image richness. This paper adds a sharpening adjustment algorithm based on the information entropy threshold in the feature point extraction part of the tracking thread. By preprocessing the image block, the information entropy of the image block is compared with the threshold. If the information entropy is below the threshold, it will be sharpened to enhance the richness of the information and facilitate subsequent processing. If the information entropy is above the threshold, the richness of the information is considered sufficient and no processing is performed. In order to increase the generality, this paper proposes the algorithm will automatically calculate the information entropy threshold of the scene according to different scenes. Experimental results show that the improved algorithm improves the accuracy and robustness of ORB-SLAM2 in the case of the rotation at a large angle when the texture information is not rich.

2. The Framework of the Algorithm

The ORB-SLAM2 system is divided into three threads, namely, Tracking, Local Mapping, and Local Closing. In the entire process, the ORB algorithm is used for feature point detection and matching, and the BA (Bundle Adjustment) algorithm is used to perform nonlinear iterative optimization of the three procedures to obtain accurate camera pose and 3D map data. Since the ORB-SLAM2 system relies heavily on the extraction and matching of feature points, running in an environment with rich textures or dealing with video will fail to obtain enough stable matching point pairs. Thus, the beam adjustment method lacks sufficient input information. The posture deviation cannot be effectively corrected [13]. The algorithm of this paper preprocesses the image before extracting the feature points of the tracking thread and adds a sharpening adjustment algorithm based on the information entropy threshold.

2.1. Overall Framework. The overall system framework of the improved algorithm is shown in Figure 1. There are two steps in our program. The first step automatically determines the information entropy threshold of the scene based on the adaptive information entropy algorithm. The second step compares the threshold and the information entropy of each image block and performs the following operations.

The system first accepts a series of image sequences from the camera, passes the image frames to the tracking thread after initialization, extracts ORB features from the images in this thread, performs pose estimation based on the previous frame, tracks the local map, optimizes the pose, determines the key frame according to the rules, and passes it to the Local Mapping thread to complete the construction of the local map. Finally, Local Closing performs closed-loop detection and closed-loop correction. In the algorithm of this paper, the sharpening adjustment based on the information entropy threshold is added to the image preprocessing part before the ORB feature detection in the Tracking thread.

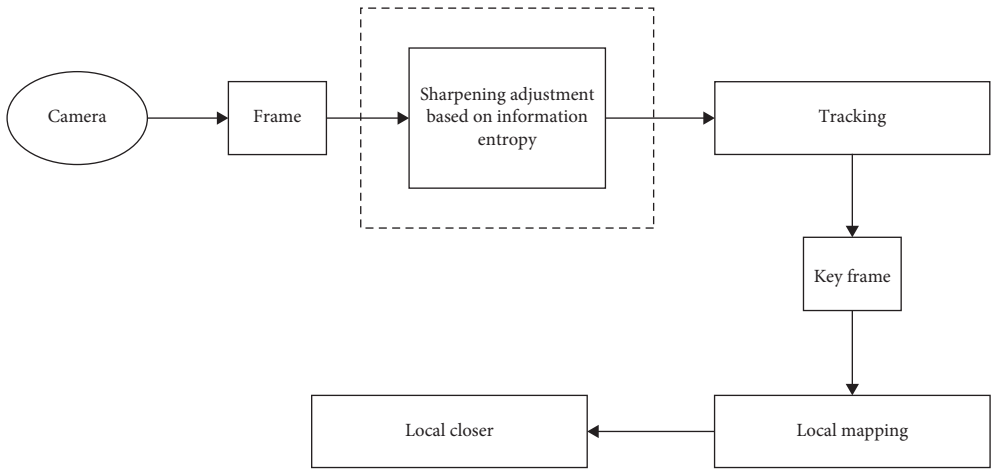


Figure 1: Algorithm system framework.

2.2. Tracking Thread. The input to the tracking thread is each frame of the video. When the system is not initialized, the thread tries to initialize using the first two frames, mainly to initialize and match the feature points of the first two frames. There are two initialization methods in the ORB-SLAM2 system, namely, the initialization of monocular camera, binocular camera, and RGB-D camera. The experiment in this article mainly uses monocular camera.

After the initialization is complete, the system starts to track the camera position and posture. The system first uses the uniform velocity model to track the position and posture. When the uniform velocity model fails, the key frame model is used for tracking. In the case when both types of tracking fail, the system will trigger the relocation function to relocate the frame. After the system obtains each frame, it will use feature points to detect and describe and use feature descriptors for tracking and pose calculation. In the process of feature point extraction, image quality plays an important role. Therefore, this article creatively adds a preprocessing algorithm to improve image quality before the image feature point extraction step.

At the same time, the system will track the local map, match the current frame with related key frames to form a set, and find the corresponding key points. The beam adjustment method is used to track the local map points to minimize the reprojection error, thereby optimizing the camera posture of the current frame. Finally, according to the conditions of whether to generate a key frame, a judgment is made to generate a key frame. Specify the current frame that meets certain conditions as a key frame. The key frames selected in the "tracking" thread need to be inserted into the map for construction. The key frame contains map points as feature points. When more than a certain number of key frames are collected, their key points will be added to the map and become map points. Delete the map points that do not meet the conditions.

2.3. Local Mapping Thread. The input of the thread is the key frame inserted by the tracking thread. On the basis of newly added key frames, maintain and add new local map points. The key frame input by the thread is generated according to the rules set during tracking, so the result of Local Mapping thread is also closely related to the quality of the image used for feature point extraction in the tracking process. Image adaptive preprocessing algorithm also improves Local Mapping thread when optimizing the extraction of feature points.

During the construction of the map, local BA optimization will be performed. Use BA to minimize reprojection errors and optimize Map points and poses. Because BA requires a lot of mathematical operations, which are related to key frames, Local Mapping will delete redundant key frames in order to reduce time consumption.

2.4. Local Closing Thread. In the continuous movement of the camera, it is completely consistent with the actual that the camera pose calculated by the computer and the map points is obtained by the triangulation. There is a certain error between them. While the number of frames increases, the error gradually accumulates. In order to reduce these accumulated errors, the most effective method is closed-loop correction. ORB-SLAM2 uses the closed-loop detection method. When the camera re-enters the previous scene, the system detects Closed-loop, and global BA optimization is performed to reduce the cumulative error. Therefore, when the ORB-SLAM2 system is applied to a large-scale scene, it shows higher robustness and usability. The input of the thread is the key frame screened by the Local Mapping thread. The word bag vector of the current key frame is stored in the database of the global word bag vector to speed up the matching of subsequent frames. At the same time, it detects whether there is a loopback, and if it occurs, it passes the pose Graph optimization to optimize the pose of all key frames and reduce the accumulated drift error. After the pose optimization is completed, a thread is started to execute the global BA to obtain the most map points and key pose results of the entire system.

2.5. The Process of Sharpening Adjustment Algorithm Based on Information Entropy. Motion-blurred images are more disturbed by noise, which impact the extraction of

feature points in the ORB-SLAM2 system. The number of stable feature points found on these images is insufficient. The feature points cannot even be found in the images with too much blurs. The insufficient number will then affect the accuracy of the front-end pose estimation. When the feature points are less than the number specified by the system, the pose estimation algorithm cannot be performed so that the back-end cannot obtain the front-end information and the tracking fails. Based on this background, this paper proposes a sharpening adjustment algorithm based on information entropy, preprocessing the image before the tracking thread.

In images with poor texture and blurred images, the sharpening of the image can make the corner information of the image more prominent. Because when the ORB feature is extracted in the tracking thread, it is easier to detect the feature point and enhance the system's stability. And the screening based on the information entropy threshold is added, and the image sharpening adjustment is performed only when the information entropy of the image block is less than the threshold. It can reduce the time of sharpening adjustment, ensure the real-time performance of the system, and maintain the integrity of the image information. If the image block information entropy is bigger than the threshold, it will not be processed.

The process of sharpening adjustment algorithm based on information entropy threshold is as follows, and the algorithm flow chart is shown in Figure 2.

(1) At first, an image input by the frame is converted into a grayscale image, and the image is expanded into an 8-layer image pyramid under the effect of the scaling factor, and then each layer of the pyramid image is divided into image blocks.

(2) Calculate the information entropy E of the image block, and compare the obtained information entropy E with the information entropy threshold $E0$. The image block with the information entropy E less than the threshold $E0$ indicates that the image block contains less effective information and the effect of ORB feature extraction is poor, so you need to sharpen first to enhance the detail in the image.

(3) After the sharpening process is performed on the image block with the information entropy less than the threshold, ORB feature points are extracted with the image block with the information entropy greater than the threshold. The feature points are extracted using the FAST feature point extraction algorithm in the pyramid. The quadtree homogenization algorithm is used for homogenization processing, which makes the distribution of extracted feature points more uniform. It avoids the phenomenon of clustering of feature points; thus, the algorithm becomes more robust.

(4) Then, perform a BRIEF description of the feature points to generate a binary descriptor of the feature points after the homogenization process. The feature points generated with the BRIEF descriptor at this time are called ORB features, which have invariance of viewpoints and invariance of lighting ORB features which are used in the later graph matching and recognition in the ORB-SLAM2 system.

3. The Novel Image Preprocessing Algorithm

We propose and add a sharpening adjustment algorithm based on information entropy. Before the tracking thread, the image is preprocessed. The above algorithm framework roughly describes the specific operation of the algorithm. The following will explain in detail the relevant principles and details of the improved algorithm. Algorithms used in this article are proposed including the ORB algorithm in ORB-SLAM2, the sharpening adjustment algorithm, the principle and algorithm of information entropy, and the adaptive information entropy threshold algorithm.

3.1. ORB Algorithm. As explained in the previous article, the ORB-SLAM2 system relies heavily on the extraction and matching of feature points. Therefore, running in an environment with rich texture or video processing will not be able to obtain enough stable matching point pairs, resulting in loss of system position and attitude tracking. Case: in order to better explain the improved system in this article, we first introduce the feature point extraction algorithm used by the system.

The focus of image information processing is the extraction of feature points. The ORB algorithm combines the two methods: FAST feature point detection and BRIEF feature descriptor [14]. To make further optimization and improvement on their basis, in 2007, the FAST corner detection algorithm was proposed. Compared with other feature point detection algorithms, the FAST algorithm has better real-time detection and robustness. The core of the FAST algorithm is to take a pixel and compare it with the gray value of the points around it. For comparison, if the gray value of this pixel differs from most of the surrounding pixels, it is considered to be a feature point.

The specific steps of FAST corner detection feature points: each pixel is detected, with the detected pixel as the center p; 3 pixels with a radius of 16 pixels on a circle, set a gray value threshold t. Comparing the 16 pixels on the circle, when there are consecutive n (12) pixels, the gray value is greater than $I_p + t$ or less than $I_p - t$ (I_p is the gray value of point p); then, p is determined as a feature point. In order to improve the efficiency of detection, a simplified judgment is first performed. It is only necessary to detect whether the gray values of 1, 5, 9, and 13 satisfied the above conditions; when at least 3 points are met, continue to detect the remaining 12 points (Figure 3).

The concept of mass points is introduced to obtain the scale and rotation invariance of FAST corner points. The centroid refers to the gray value of the image block as the center of weight. The specific steps are as follows.

(1) In a small image block B, the moment defining the image block B is

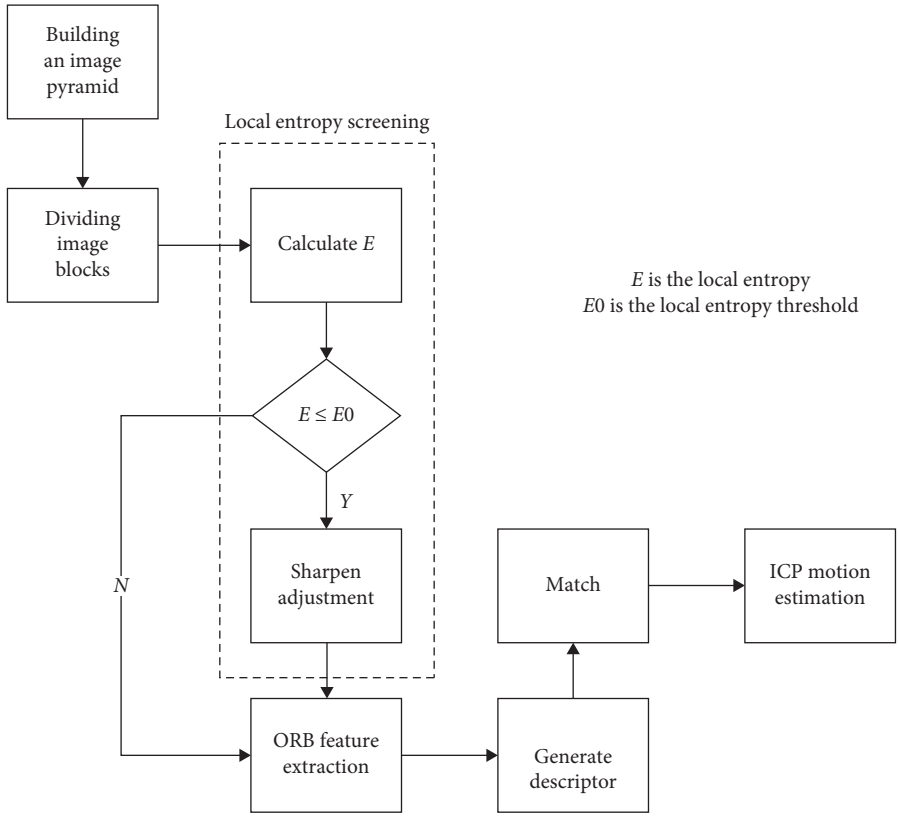


Figure 2: Process of sharpening adjustment algorithm based on information entropy.

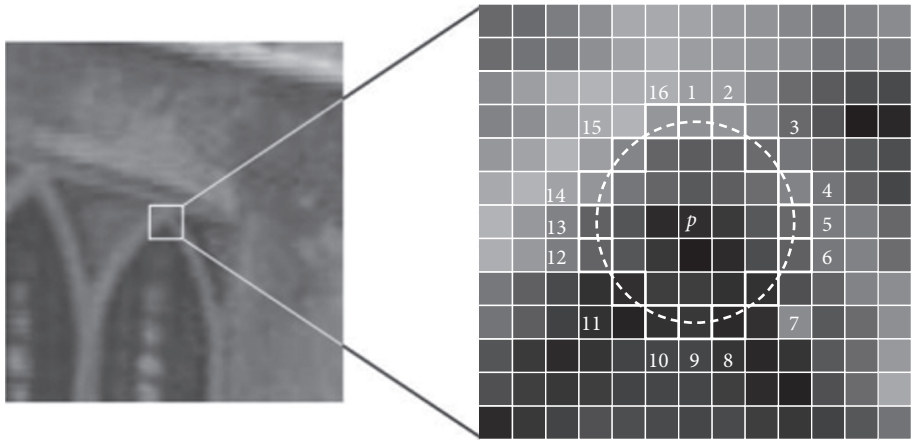


Figure 3: FAST feature point detection.

$$m_{\mathrm{pq}} = \sum_{x,y \epsilon B} x^p y^q I(x, y), \quad p, q = \{0, 1\}. \tag{1}$$

(2) You can find the centroid of the image block by moment:

$$C = \left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}} \right). \tag{2}$$

(3) Connect the geometric center O and the centroid C of the image block to get a direction vector $\overrightarrow{OC}$. Then, the feature point direction can be defined as follows:

$$\theta = \arctan\left(\frac{m_{01}}{m_{10}}\right). \tag{3}$$

Through the above methods, FAST corners have a description of scale and rotation, which greatly improves the robustness of their representation between different images. So, this improved FAST is called Oriented FAST in the ORB.

After the feature point detection is completed, the feature description needs to be used to represent and store the feature point information to improve the efficiency of the subsequent matching. The ORB algorithm uses the BRIEF descriptor whose main idea is to distribute a certain probability around the feature point. Randomly select several point pairs and then combine the gray values of these point pairs into a binary string. Finally use this binary as the descriptor of this feature point.

The BRIEF feature descriptor determines the binary descriptor by comparing the pair of gray values. As shown in Figure 4, if the gray value at the pixel x is less than the gray value at y, the binary value of the corresponding bit is assigned to 1, otherwise it is 0:

$$\tau(p; x, y) = \begin{cases} 1, & I(x) < I(y), \\ 0, & I(x) \geq I(y). \end{cases} \tag{4}$$

Among them, $I(x)$ and $I(y)$ are the gray values at the corresponding corners after image smoothing.

For n (256) (x, y) test point pairs, the corresponding n-dimensional BRIEF descriptor can be formed according to the following formula:

$$f_n(p) = \sum_{1 \leq i \leq n} 2^{i-t} \tau(p; x, y). \tag{5}$$

However, the BRIEF descriptor does not have rotation invariance, the ORB algorithm improves it. When calculating the BRIEF descriptor, the ORB will solve these corresponding feature points in the main direction, thereby ensuring that the rotation angle is different. The point pairs selected for the same feature point are the same.

3.2. Sharpening Adjustment. In the process of image transmission and conversion, the sharpness of the image is reduced. In essence, the target contour and detail edges in the image are blurred. In the extraction of feature points, what we need to do is to traverse the image, and the pixels with large difference in surrounding pixels are selected as the feature points, which requires the image's target contour and detail edge information to be prominent. Therefore, we introduced image sharpening.

The purpose of image sharpening is to make the edges and contours of the image clear and enhance the details of the image [15]. The methods of image sharpening include statistical difference method, discrete spatial difference method, and spatial high-pass filtering method. Generally, the energy of the image is mainly concentrated in the low frequency. The frequency where the noise is located is mainly in the high-frequency band. The edge information of the image is mainly concentrated in the high-frequency band. Usually smoothing is used to remove the noise at high frequency, but this also makes the edge and contour of the image blurred, which affects the extraction of the feature points. The root cause of the blur is that the image is subjected to an average or integral operation. So, performing an inverse operation (a differential operation) on the image can restore the picture to clear.

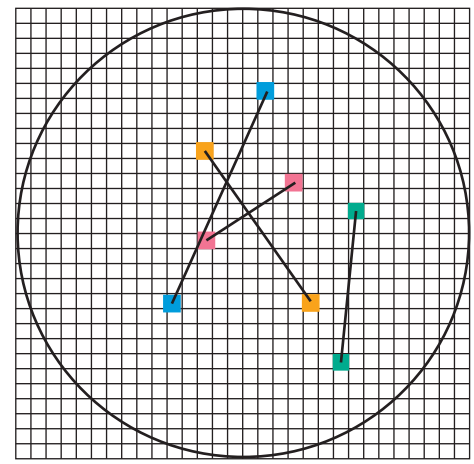


Figure 4: BRIEF description.

In order to reduce the adverse effects, we adopt the method of high-pass filtering (second-order differential) in the spatial domain and use the Laplacian operator to perform a convolution operation with each pixel of the image to increase the variance between each element in the matrix and the surrounding elements to achieve sharp image. If the variables of the convolution are the sequence $x(n)$ and $h(n)$, the result of the convolution is

$$y(n) = \sum_{i=-\infty}^{\infty} x(i)h(n-i) = x(n) * h(n). \tag{6}$$

The convolution operation of the divided image blocks is actually to use the convolution kernel to slide on the image. The pixel gray value is multiplied with the value on the corresponding convolution kernel. Then, all the multiplied values are added as the gray value of the pixel on the image corresponding to the middle pixel of the convolution kernel. The expression of the convolution function is as follows:

$$\begin{aligned} \mathrm{dst}(x, y) = & \sum_{\substack{0 \leq x' < \mathrm{kernel.cols} \\ 0 \leq y' < \mathrm{kernel.rows}}} \mathrm{kernel}(x', y') \\ & * \mathrm{src}(x + x' - \mathrm{anchor}.x, y + y' - \mathrm{anchor}, y), \end{aligned} \tag{7}$$

where anchor is the reference point of the kernel; kernel is the convolution kernel, and the convolution template here is the matrix form of the Laplacian variant operator, as shown below:

$$H = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}. \tag{8}$$

The Laplacian deformation operator uses the second derivative information of the image and is isotropic.

(a) (b)

Figure 5: Sharpening adjustment.

The results of image discrete convolution on the classic Lenna image using the sharpening algorithm adopted in this paper are shown in Figure 5.

From the processing results, it can be seen that the high-pass filtering effect of the image sharpening algorithm is obvious, which enhances the details and edge information of the image, and the noise becomes small, so this sharpening algorithm is selected.

3.3. Screening Based on Information Entropy. In 1948, Shannon proposed the concept of "information entropy" to solve the problem of quantitative measurement of information. In theory, entropy is a measure of the degree of disorder of information, which is used to measure the uncertainty of information in an image. The larger the entropy value is, the higher the degree of disorder. In image processing, entropy can reflect the information richness of the image and display the amount of information contained in the image. The information entropy calculation formula is

$$H(x) = -\sum_{0}^{255} p(x_i)\log_2 p(x_i), \tag{9}$$

where $p(x_i)$ is the probability of a pixel with grayscale $i(i = 0 \ldots 255)$ in the image. When the probability is closer to 1, the uncertainty of the information is smaller, and the receiver can predict the transmitted content. The amount of information in the belt is less, and vice versa.

If the amount of information contained in the image is expressed by information entropy, then the entropy value of an image of size $m \times n$ is defined as follows:

$$p_{ij} = \frac{f(i,j)}{\sum_{i=1}^{M}\sum_{j=1}^{N} f(i,j)}, \tag{10}$$

$$H = -\sum_{i=1}^{M}\sum_{j=1}^{N} p_{ij}\log_2 p_{ij}, \tag{11}$$

where $f(i, j)$ is the gray level at the point (i, j) in the image, p_{ij} is the gray distribution probability at the point, and H is the entropy of the image. If $m \times n$ is taken as (i, j) a local neighborhood in the center, then H is called the information entropy value of the image [16].

From the perspective of information theory, when the probability of image appearance is small, the information entropy value of the image is large, indicating that the amount of information transmitted is large. In the extraction of feature points, the information entropy reflects the texture contained in the partial image information richness or image pixel gradient change degree. The larger the information entropy value is, the richer the image texture information is and more obvious the changing of the image pixel gradient is [17].

Therefore, the effect of ORB feature point extraction is good, and the image block does not require detail enhancement. The local information entropy value is low, so the effect of ORB feature point extraction is poor which is shown in Figure 6. After sharpening adjustments to enhance details, we optimize the effect of feature point extraction [18], which is shown in Figure 7. By comparing the feature point extraction of ORB-SLAM2, it can be intuitively seen that the optimized extraction algorithm accuracy is better.

3.4. Adaptive Information Entropy Threshold. Because the number of the information entropy is closely related to the scene, different video sequences in different scenes have different information richness, so the information entropy threshold of different scenes must also be different. In each different scene, it requires repeated experiments to set the information entropy threshold multiple times for matching calculations to get the corresponding threshold value. However, the experience value in different scenes differs greatly. The threshold value is not universal, which will affect the image preprocessing and feature extraction. It will result in the failure of quickly obtaining better matching results. So, the adaptive algorithm of information entropy is particularly important [19].

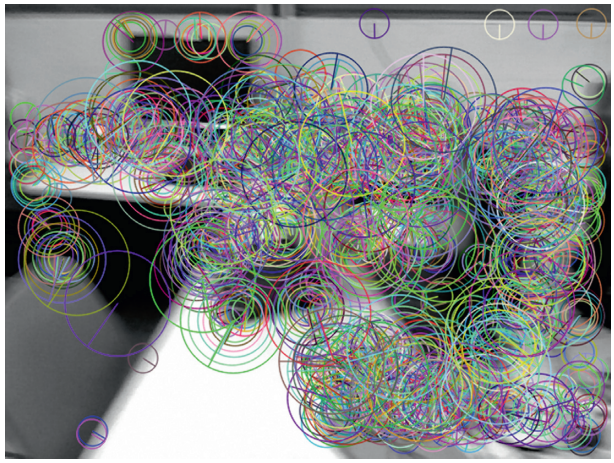


Figure 6: Feature point extraction based on ORB-SLAM2.

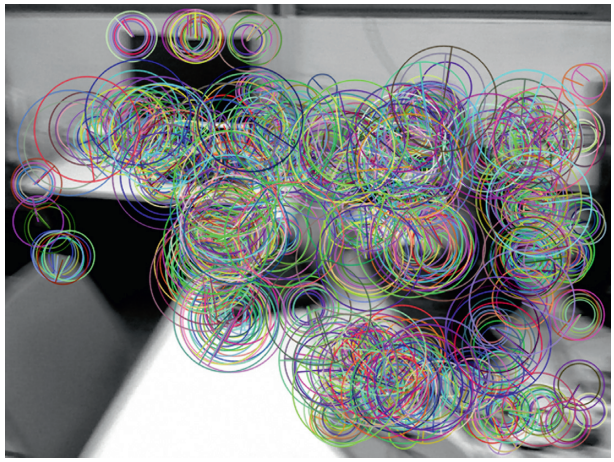


Figure 7: Extraction of sharpening adjustment feature points based on information entropy.

In view of the above problems, we propose an adaptive method of information entropy threshold, which adjusts the threshold according to different scenarios. The self-adjusting formula is

$$E_0 = \frac{H(i)_{\text{ave}}}{2} + \delta, \tag{12}$$

where $H(i)_{\text{ave}}$ is the average value of the information entropy in this scene, which can be obtained by obtaining the information entropy of each frame in a video in the scene under the first run and then divided by the number of frames, i is the number of frames in the video sequence, and δ is the correction factor; after experiment, it is 0.3 which works best. Through the above formula, the calculated E_0 is the information entropy threshold of the scene.

4. Experimental Results and Analysis

All the experiments are conducted on a laptop computer. The operating system is Linux 16.04 64 bit, the processor Intel (R) Core (TM) i5-7300 U CPU @ 2.50 GHz, the operating environment is CLion 2019 and opencv3.3.0, and the program uses C++ Compilation. The back-end uses G2O for posture optimization based on posture maps to generate motion trajectories. The image sequence reading speed of this algorithm is 30 frames per second, compared with the open source ORB-SLAM2 system.

4.1. Comparison of Two-Frame Image Matching. In this paper, the rgbd_dataset_freiburg1_desk dataset is selected, from which two images with blurry images and large rotation angles are extracted, and the image detection and matching processing are performed using the ORB algorithm, ORB + information entropy, and ORB + information entropy + sharpening extraction algorithm. And compare the data such as extraction effect, key points, matching number, matching number, and matching rate after RANSAC filtering. The matching result statistics of the image are shown in Table 1, and the matching effect is shown in Figures 8, 9, and 10.

Table 1 shows the results of the feature point detection and matching through the ORB, ORB + information entropy, and ORB + information entropy + sharpening. It proves that the correct matching points and the matching rate of ORB and ORB + information entropy are basically the same. The matching effect diagram of the latter shows a mismatch with a large error. For ORB + information entropy + sharpening, the number of correct matching points of the extraction algorithm has increased significantly, and its matching rate, compared to ORB + information entropy and ORB + information entropy + sharpening extraction algorithms, have been, respectively, improved by 36.7% and 38.8%.

The results show that adding the information entropy screening and sharpening process proposed by this algorithm can effectively increase the number of matches and improve the matching rate. The accuracy of image feature detection and matching are improved, which provide more accurate and effective for subsequent process as tracking, mapping, and loop detection.

4.2. Comparison of Image Matching in Visual Odometer. For the timestamp sequence of different blur degrees and rotation angles of the image (that is, a sequence of frames 170–200), ORB-SLAM2 algorithm and the feature detection and matching algorithm in the improved algorithm of this paper are introduced to perform two adjacent frame's feature detection and matching. To compare the matching points of the two algorithms in different images, the image is clear between 170 and 180 frames, and the motion blur starts between 180 and 190 frames. Maximum rotation angle is between 190 and 200 frames. The most dramatic motion blur is shown in Figure 11.

It can be seen from Figure 12 that the improved algorithm can match more feature points than the original algorithm. The number of matching points of the improved algorithm is more than 50, and tracking loss does not occur, even if the image is blurred due to fast motion. When the resolution of the image is reduced, the

Table 1: Comparison of matching rates of three feature detection and matching algorithms.

Model	Key points	Matches	RANSCK	Match rate (%)
ORB	1000/1004	254	87	34.25
+information entropy	1000/1002	255	86	33.73
+sharpening	1000/983	284	133	46.83

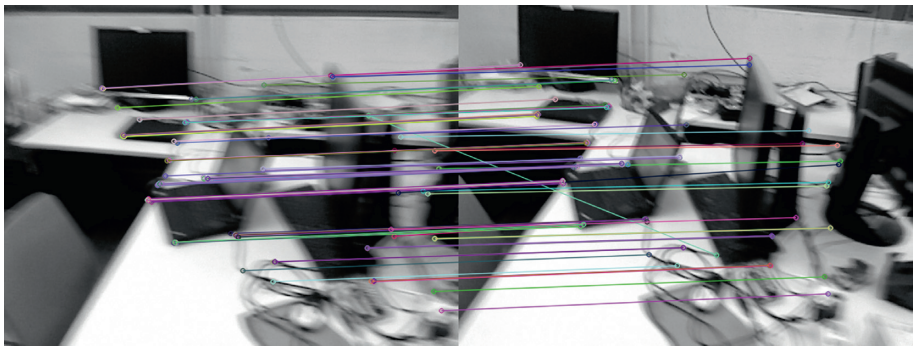


Figure 8: Matching effect of ORB.

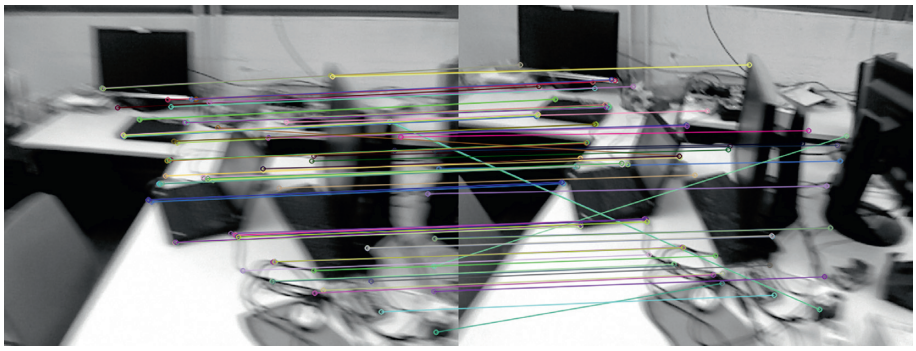


Figure 9: Matching effect of ORB + information entropy.

Figure 10: Matching effect of ORB + information entropy + sharpening.

improved algorithm still has more matching feature points and has strong anti-interference ability and robustness.

4.3. Analysis of Trajectory Tracking Accuracy. The gbd_dataset_freiburg1_desk dataset in the TUM data is selected in the experiment. We use the real trajectory of the robot provided by TUM to compare it with the motion trajectory calculated by the algorithm to verify the improvement effect of the proposed algorithm. Here, we take the average absolute trajectory error. The average relative trajectory error and the average tracking time are standard. The comprehensive trajectory tracking accuracy and real-time performance are compared between the ORB-SLAM2 system and the improved SLAM system.

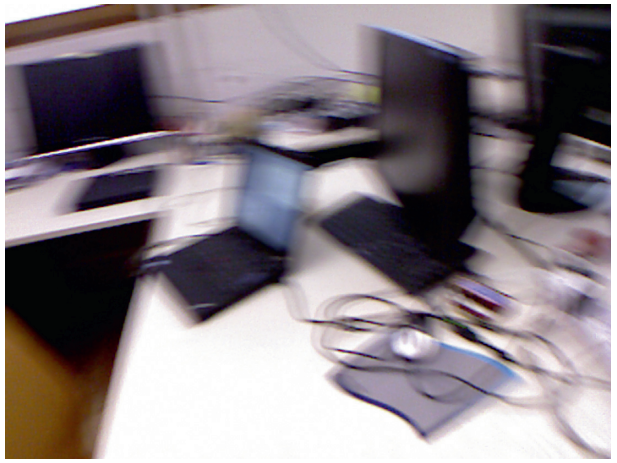


Figure 11: 200th frame image.

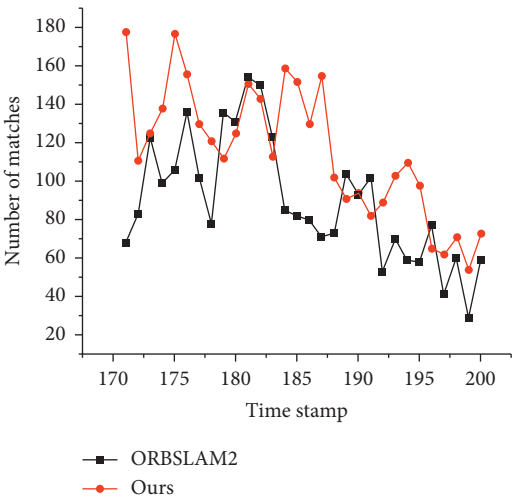


Figure 12: Comparison between the algorithm of this paper and ORB-SLAM2 motion blur.

We use the information entropy of image blocks to determine the amount of information. The algorithm sharpens the image blocks with small information entropy, enhances local image details, and extracts local feature points that can characterize image information for the back-end as adjacent frames. The correlation basis of key frame matching enhances robustness and reduces the loss of motion tracking caused by the failure of interframe matching at the same time. Based on the matching result, the R and t transformation relationship between frames is calculated, and the back-end uses g2o to optimize the pose based on the position and attitude map. Finally the motion trajectory is generated, which is shown in Figure 13.

In terms of trajectory tracking accuracy, the absolute trajectory error reflects the difference between the true value of the camera's pose and the estimated value of the SLAM system. The absolute trajectory error root mean square RMSE (x) is defined as follows:

$$\mathrm{RMSE}(x) = \sqrt{\frac{\sum_{i=1}^{n} \left\| x_{e,i} - x_{s,i} \right\|^2}{n}}. \tag{13}$$

Among them, $x_{e,i}$ represents the estimated position of the ith frame in the image sequence and $x_{s,i}$ represents the standard value of the position of the ith frame in the image sequence. Taking the rgbd_dataset_freiburg1_desk dataset as an example, the error between the optimized motion trajectory and the standard value is shown in Figure 14.

The average absolute trajectory error refers to directly calculating the difference between the real value of the camera's pose and the estimated value of the SLAM system. The program first aligns the real value and the estimated value according to the time stamp of the pose and then calculates each pair of poses. The difference is very suitable for evaluating the performance of visual SLAM systems. The average relative trajectory error is used to calculate the difference between the pose changes on the same two timestamps. After alignment with the timestamps, both the real pose and the estimated pose are calculated at every same period of time; then, the difference is calculated. It is suitable

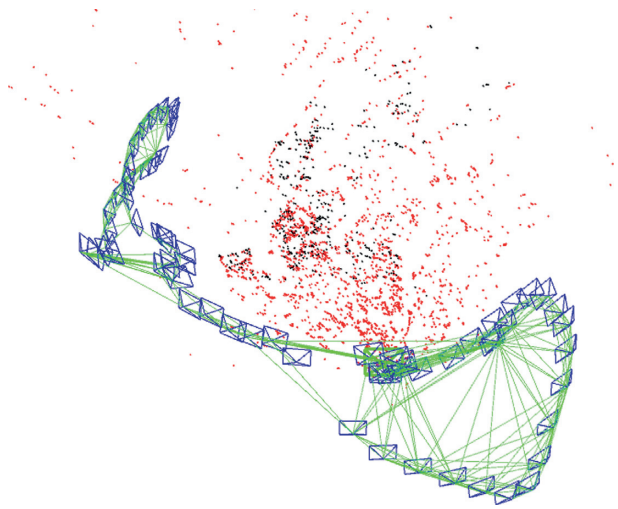


Figure 13: Movement track.

for estimating the drift of the system. The definition of the absolute trajectory error in frame i is as follows:

$$F_i = Q_i^{-1} S p_i. \tag{14}$$

Among them, Q_i represents the real pose of the frame, and pi represents the algorithm estimated pose of the frame, $S \in \mathrm{Sim}\ (3)$ is the similarity transformation matrix from estimated pose to true pose.

This experiment comprehensively compares the three datasets rgbd_dataset_freiburg1_desk/room/360, of which the desk dataset is an ordinary scene; the 360 dataset is the dataset collected when the camera performs a 360-degree rotation motion, and the room dataset is a scene collected when the image sequence texture is not rich. The ORB-SLAM2 generates different trajectory results each time. In order to eliminate the randomness of the experimental results, this article uses 6 experiments on each dataset under the same experimental environment to calculate the average absolute trajectory error, the average relative trajectory

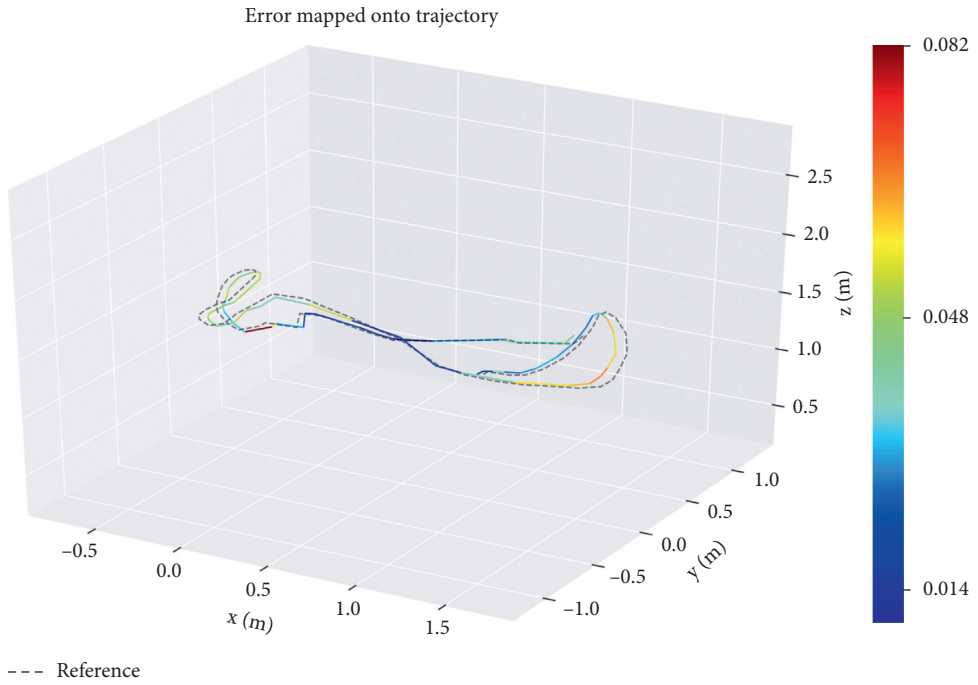


Figure 14: Absolute trajectory error.

Table 2: Average absolute trajectory error.

Average absolute trajectory error difference/m	ORB-SLAM2	The proposed algorithm
Desk	0.057630667	0.029899333
360	0.113049667	0.072683167
Room	0.229222167	0.111185333

Table 3: Average relative trajectory error.

Average relative trajectory error difference/m	ORB-SLAM2	The proposed algorithm
Desk	0.0362615	0.029899
360	0.1534532	0.078611
Room	0.1592775	0.092167

Table 4: Average tracking time.

Average relative trajectory error difference/m	ORB-SLAM2	The proposed algorithm
Desk	0.0244937	0.0289299
360	0.0186606	0.0225014
Room	0.0220836	0.0282757

error, and the average tracking time. Tables 2–4 are plotted in Figures 15–17 to compare the algorithm performance by analyzing the trajectory.

Under the evaluation standard of absolute trajectory error, the algorithm in this paper has certain advantages. In the scene where the camera rotates 360°, the absolute trajectory error is increased by 35%, and in the ordinary scene, the absolute trajectory error is increased by 48%. In the case of not rich texture, the absolute trajectory error improvement is the largest, indicating that the algorithm in this paper has a considerable improvement on scenes with rich image sequence textures. There is also a certain improvement in the situation of large-angle camera rotation.

There is an advantage in trajectory error: in ordinary scenes, it is 17.5% smaller than the average relative trajectory error of the ORB-SLAM2 system. In the scene where the camera rotates 360° and the image sequence texture is not rich, the algorithm in this paper is better than the ORB-SLAM2 system, which has been improved by more than 40%. It proves that the algorithm in this article is indeed improved compared to the ORB-SLAM2 system.

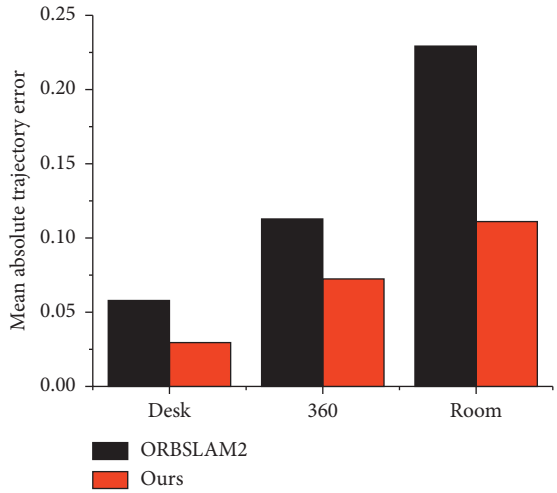


Figure 15: Average absolute trajectory error.

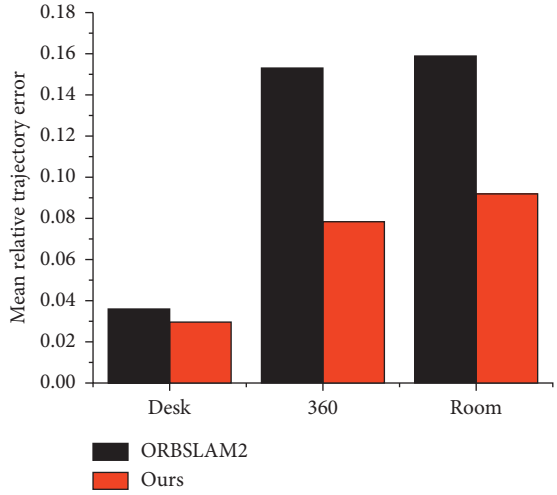


Figure 16: Average relative trajectory error.

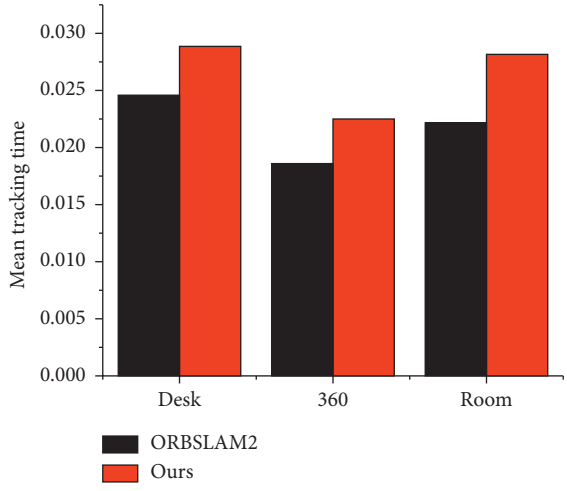


Figure 17: Average tracking time.

Due to the sharpening of the local image, the average tracking time is slightly increased but the increase is not large.

5. Conclusion

We propose an improved ORB-SLAM2 algorithm, which is based on adaptive information entropy screening algorithm and sharpening adjustment algorithm. Our advantage is that the information entropy threshold can be automatically calculated according to different scenes with generality. Sharpening the picture blocks below the information entropy threshold increases the clarity in order to better extract the feature points. The combination of information entropy screening image block and sharpening algorithm solves the camera's large-angle rotation and the system positioning and mapping failure caused by the lack of texture information. Although the processing time increases, it improves the processing accuracy. We compare the average absolute trajectory error, the average relative trajectory error, and the average tracing time in the trajectory error with ORB-SLAM2. The experimental results show that the improved model based on information entropy and sharpening improved algorithm achieve better results compared with ORB-SLAM2. This article innovatively combines the ORB-SLAM2 system with adaptive information entropy. The combination of sharpening algorithms improves the accuracy and robustness of the system without a noticeable gap in processing time.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC)-Guangdong big data Science Center Project under Grant U1911401 and in part by the South China Normal University National Undergraduate Innovation and Entrepreneurship Training Program under Grant 201910574057.

References

[1] D. Schleicher, L. M. Bergasa, R. Barea, E. Lopez, and M. Ocana, "Real-time simultaneous localization and mapping using a wide-angle stereo camera," in *Proceedings of the IEEE Workshop on Distributed Intelligent Systems: Collective Intelligence and its Applications (DIS'06)*, pp. 2090–2095, Beijing, China, October 2006.

[2] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "MonoSLAM: real-time single camera SLAM," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1052–1067, 2007.

[3] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proceedings of the 2007 6th IEEE and*

ACM International Symposium on Mixed and Augmented Reality, pp. 225–234, Nara, Japan, November 2007.

[4] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics*, vol. 29, no. 6, pp. 1052–1067, 2007.

[5] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[6] K. Di, W. Wan, and J. Chen, "SLAM visual odometer optimization algorithm based on information entropy," *Journal of Automation*, vol. 47, no. 6, pp. 770–779, 2020.

[7] M. Quan, H. Wei, and J. Chen, "SLAM visual odometer optimization algorithm based on information entropy," *Journal of Automation*, vol. 1, no. 8, pp. 180–198, 2020.

[8] L. Xu, S. Zheng, and J. Jia, "Unnatural L0 sparse representation for natural image deblurri," in *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1107–1114, Portland, OR, USA, June 2013.

[9] J. Sun, W. Cao, Z. Xu, and J. Ponce, "Learning a convolutional neural network for non-uniform motion blur removal," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 769–777, Boston, MA, USA, June 2015.

[10] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 257–265, Honolulu, HI, USA, July 2017.

[11] J. Zhang, J. Pan, J. Ren et al., "Dynamic scene deblurring using spatially variant recurrent neural networks," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2521–2529, Salt Lake City, UT, USA, June 2018.

[12] G. Chen, W. Chen, H. Yu et al., "Research on autonomous navigation method of mobile robot based on semantic ORB-SLAM2 algorithm," *Machine Tool & Hydraulics*, vol. 48, no. 9, pp. 16–20, 2020.

[13] X.. Qiu and C. Zhao, "Overview of ORB-SLAM system optimization framework analysis," *Navigation Positioning and Timing*, vol. 6, no. 3, pp. 46–57, 2019.

[14] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proceedings of the European Conference on Computer Vision*, pp. 430–443, Graz, Austria, 2006.

[15] A. Foi, V. Katkovnik, and K. Egiazarian, "Pointwise shape-adaptive DCT for high-quality denoising and deblocking of grayscale and color images," *IEEE Transactions on Image Processing*, vol. 16, no. 5, pp. 1395–1411, 2007.

[16] Y. Yu and W. H. Tardós, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[17] R. Xu, S. Liu, and J. Chen, "The rationality and application of information entropy description in images," *Information Technology*, vol. 11, no. 5, pp. 59–61, 2005.

[18] Y. Zhu, "Overview of image enhancement algorithms," *Information and Computer (Theoretical Edition)*, vol. 16, no. 6, pp. 104–106, 2017.

[19] Z. Yuan, "Image threshold automatic selection based on one-dimensional entropy," *Journal of Gansu University of Technology*, vol. 1, no. 1, pp. 84–88, 1993.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 3943041, 11 pages
https://doi.org/10.1155/2020/3943041

Research Article

STAP Optimization of Airborne Phased Array Radar in Nonuniform Environment Based on EFA Algorithm

Bin Tang,[1,2] **Xiaoxia Zheng**,[1,2] **Mingxin Liu**,[1,2] **and Mengxu Fang**[1,2]

[1]*Chengdu Aeronautic Polytechnic, Chengdu 610100, China*
[2]*College-enterprise Joint Avionics Innovation Base in Sichuan, Chengdu 610100, China*

Correspondence should be addressed to Bin Tang; liumaocheng@mail.sdufe.edu.cn

Received 15 March 2020; Revised 5 August 2020; Accepted 12 August 2020; Published 28 August 2020

Guest Editor: Sanghyuk Lee

EFA (extended factored approach) algorithm is the main method of space-time adaptive processing technology (STAP) for airborne phased array radar, but it is faced with many problems, such as large number of samples and large amount of calculation. Therefore, this paper uses a method of spatial data dimensionality reduction processing based on cyclic iterative calculation to optimize its STAP. The final experimental results show that, after spatial data dimensionality reduction processing optimization, the STAP performance of EFA algorithm is further expanded in the range of sample number adaptation; especially in the case of small sample number, the optimized STAP performance has been basically close to the ideal compared with other optimization schemes; tap performance also proves that the optimization scheme in this paper has better convergence speed and STAP performance.

1. Introduction

In modern air war, airborne radar is the key military technology to win the war, especially the modern early warning aircraft with airborne phased array radar as the main detection means, which is also called the air combat force multiplier of modern war, so the research on airborne phased array radar technology has been endless. The biggest advantage of airborne phased array radar is that its platform can bring it into the high altitude and then scan the airspace through the radio pulse beam in the high altitude, so its detection range and distance are greatly increased. However, in the process of airborne phased array radar scanning the airspace, the transmitted radio pulse beam does not only reflect the objects of interest but also reflect all objects in the airspace. The reflected radio pulse beam either scatters to the surrounding environment or feeds back to the phased array radar antenna array. The radio pulse beam fed back to phased array radar antenna array is processed by signal machine to form scanning information, which includes not only interested target information but also clutter information and interference information. How to eliminate these clutter and interference information has become the key direction to improve the performance of airborne phased array radar. In this study, the main lobe of Doppler filter in space-time adaptive processing technology is easy to be affected by clutter penetration. Two methods, EFA algorithm and spatial decomposition dimension reduction, are used to improve the space-time adaptive processing technology. It is hoped that through this improvement, on the one hand, the computational complexity of space-time adaptive processing technology can be reduced, and on the other hand, the influence of clutter penetration on space-time adaptive processing technology can be reduced.

In the process of processing clutter information and jamming information, the main technology of airborne phased array radar is space-time adaptive processing technology (STAP), but STAP technology itself has some defects. Especially in the sample size, clutter and noise processing has great limitations. Therefore, based on the analysis of EFA algorithm and STAP technology, this paper proposes an innovative spatial dimension reduction method to optimize the STAP technology performance, so that it can break through the limitation of sample size, further improve the

ability of STAP to process clutter, and improve the detection ability of airborne phased array radar.

This research includes three parts. The first part is the summary and analysis of STAP technology by domestic and foreign research institutions and scholars. The second part mainly introduces the airborne phased array radar and its clutter model, the related concepts of EFA technology, as well as the detailed airspace dimensionality reduction measures. The third part is to simulate and verify the improved STAP technology of this study and confirm whether the improvement measures of this study can effectively improve the performance of STAP.

2. Related Work

Since STAP technology was proposed by Brenna and reed, the related research work has not stopped, and with the emergence of phased array radar technology and frequent military conflicts around the world, more and more attention has been paid and research has been done. In the process of studying the waveform design of fully adaptive radar, scholars Setlur and Rangaswamy used independent iteration to optimize the STAP filter, thus effectively improving the performance of the filter [1]; Jingwei Xu et al. designed a new STAP radar with FDA as the launch array, which can effectively identify the clutter with fuzzy range, and designed a secondary range correlation compensation method to reduce the dependence of the modified radar clutter on the secondary range Aiming at the defects of STAP algorithm, fan team optimized STAP technology by combining principal component analysis and subcpi technology, which greatly improved its convergence speed and proved that this improved method is more practical than traditional technology [2]; Xu et al. designed a new STAP radar with FDA as the launch array, which can effectively identify the clutter with fuzzy range, and designed a secondary range correlation compensation method to reduce the dependence of the modified radar clutter on the secondary range [3]; when McDonalds and Cerutti-Maori study the radar detection technology in the sea clutter environment, they use the actual sea clutter data to simulate the sea targets, and through the two-component clutter model fitting method, they overcome the false alarm problem in the radar detection process and improve the radar detection performance Wang et al. used antenna pulse as the research basis, designed a method that can effectively improve the detection speed of STAP radar system for low target performance, and used min-max iterative algorithm to overcome the problem of antenna pulse selection [4]; Xu et al. proposed a robust adaptive beamforming method to improve the performance of fda-stap in detecting high-speed targets [5]; Kang et al. in the research of radar system considered that the directivity and resolution of the sensor can be effectively improved when the multiantenna elements are coherent processing the multipulse signal, while the STAP technology can effectively suppress the interference signal [6].

Aiming at the diversity of the transmitting waveform characteristics of airborne MIMO radar, Yidou et al. proposed three kinds of reduced dimension STAP algorithms for MIMO Radar based on the framework structure of reduced dimension STAP, which improved the antijamming performance of MIMO radar [7]. Wang et al. used antenna pulse as the research basis, designed a method that can effectively improve the detection speed of STAP radar system for low target performance, and used min-max iterative algorithm to overcome the problem of antenna pulse selection. Zhou et al. [8] put forward a subspace method based on radar parameters and ellipsoidal wave function to solve the problem that STAP's suppression performance decreases due to the influence of heterogeneous clutter. This method effectively improves STAP's suppression performance in heterogeneous clutter environment [9]; Lin et al. put forward a fast dimension reduction algorithm when studying the clutter suppression of airborne MIMO radar, which can effectively improve the performance of STAP in clutter suppression [10]; when Shi wa et al. studied the problem of target azimuth information hiding caused by the noise interference of STAP radar in the power condensing mode, they proposed a scattering wave interference method, which can effectively improve the performance of STAP radar [11]; Gao et al. put forward a STAP algorithm based on knowledge-assisted sparse recovery to solve the problem that the STAP covariance matrix of airborne radar based on the estimation of clutter covariance matrix of training snapshot which easy to be damaged by the target class signal. This algorithm can effectively identify the clutter [12]; Bruce l team proposed a linear transformation process to solve the clutter spectrum dispersion problem of bistatic STAP radar based on the focus matrix, which can effectively improve the clutter suppression ability of STAP radar in nonuniform clutter environment [13].

Wen et al. proposed an enhanced 3D joint domain local STAP method to solve the problem of performance degradation of traditional phased array space-time adaptive processing pseudotarget jamming. This method effectively reduced the radar performance degradation caused by deception jamming by prewhitening filter [14]. When the s team studied the amplitude of time output signal in the detection of weak target by Doppler pre-STAP radar, it was found that linear prediction can improve the performance of Doppler pre-STAP radar more effectively than binomial distribution time weight when determining the weight [15]; Li aiming at the problem of radar performance degradation caused by space-time adaptive processing space-time steering vector distortion, an anticross coupling space-time adaptive processing method of subarray clutter covariance matrix is proposed. This method can effectively suppress mutual coupling, target component, and target space-time steering vector mismatch in SCM by using target covariance matrix [16]; Chen proposed a two-dimensional amplitude and phase estimation algorithm based on the space-time sliding window of the test unit, which can effectively suppress clutter and interference signals [17]; aiming at the problem that the amount of training data and calculation in 3D space-time adaptive processing exceeds the actual demand, a method is proposed to convert the plane array data into azimuth and elevation linear array, respectively, so as to form the equivalent cross array before STAP. This method

can effectively reduce the amount of training data and calculation [18]; the team of Zhi Qi thinks that the STAP algorithm determines the covariance matrix through training sample data, but this process will be destroyed by outliers. Therefore, a robust STAP algorithm is designed based on the joint sparse recovery of clutter spectrum, which can effectively overcome the problem brought by outliers [19]; Jia et al. proposed an online antenna pulse selection method suitable for space-time adaptive processing. The method uses the structural clutter covariance matrix and considers antenna pulse selection and covariance matrix estimation to provide a good objective function for antenna pulse. The experimental results show that this method is feasible [20]. From the comprehensive analysis, there are many researches on STAP performance but few researches on phased array radar, which is one of the important purposes of this paper.

3. Optimization Measures of Space-Time Adaptive Processing Technology

3.1. Clutter Model of Airborne Phased Array Radar. The difference between airborne phased array radar and ground-based radar is that compared with the ground, the airborne phased array radar is in motion because of the platform, so the Doppler frequency of ground clutter in the signal received by the airborne phased array radar will have a serious problem of expansion. In some cases, the repetition rate of Doppler broadband pulse of ground clutter will be doubled, and the appearance of STAP technology effectively improves this problem. However, STAP is limited by computation and clutter plus noise covariance matrix, so it has great limitations.

Suppose that the antenna of airborne phased array radar is as shown in Figure 1 and the antenna of radar is arranged in the way of uniform linear array. Then, suppose the number of array elements is N, the speed of radar platform is v_a, the angle between the flight direction and the array direction of radar antenna is θ_p, the wavelength and pulse repetition frequency of radar are λ and f_r respectively, and the number of radar transmitted pulses within the unit coherent processing interval (CPI) is K. Then, in a CPI, the clutter data reflected from any clutter scattering point on the ground to the airborne phased array radar system can be described as follows: the airborne phased array radar will meet two conditions during operation, one is the distance between the carrier moving distance in unit CPI and the unit to be detected (cut), and the distance between the two is unchanged; the other is that the clutter echo of each scattering point is statistically independent At the same time, it satisfies the Gauss distribution. Assuming that the distance loop of the point is l, the corresponding pitch angle is φ_l and the azimuth angle of the radar antenna is θ_l, then the space-time sampling data of the radar antenna n received in the k transmission pulse can be expressed by the following formula:

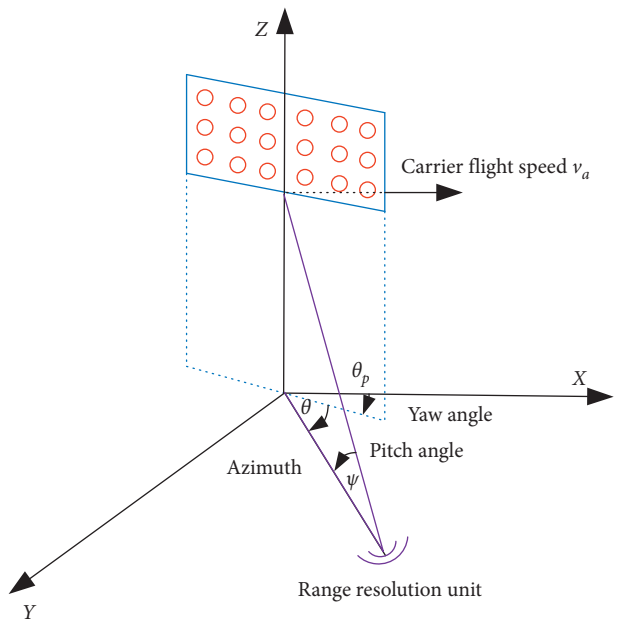


Figure 1: Basic structure of airborne phased array radar.

$$c_{n,k}(\theta,\varphi) = \frac{\rho_i}{R_l^2} F(\theta_i,\varphi_l) G(\theta_i) \exp\{j2\pi[(n-1)\omega_s(\theta_i,\varphi_l) + (k-1)\omega_t(\theta_i,\varphi_l)]\}, \quad n = 1,2,\ldots,N; \quad k = 1,2,\ldots,K. \tag{1}$$

In the formula, R_i represents the distance between radar and scattering point in l, ρ_i is the emission coefficient of clutter, $F(\theta_i,\varphi_l)$ and $G(\theta_i)$ represent the direction pattern of signal transmission and reception, $\omega_s(\theta_i,\varphi_l)$ represents the normalized spatial frequency and satisfies $\omega_s(\theta_i,\varphi_l) = (d/\lambda)\cos\theta_i\varphi_l$, and $\omega_t(\theta_i,\varphi_l)$ represents the Doppler frequency and satisfies $\omega_t(\theta_i,\varphi_l) = (2v/\lambda f_r)\cos(\theta_i+\theta_p)\cos\varphi_l$. From these two calculation methods, it can be seen that the normalized airspace frequency is affected by carrier speed and carrier distance, but the carrier is in flight state, so the Doppler frequency is different in different time periods. When $\theta_p = 0$, the relationship between Doppler frequency and azimuth cosine of clutter on cut is linear; otherwise, it is nonlinear, as shown in Figures 2–4.

Then, for a cut, the clutter echo received by the radar is the sum of all the clutter echoes of scattering points in the cell, and the formula is as follows:

$$c_{n,k}(\varphi) = \int_0^{\pi} \frac{\rho_i}{R_l^2} F(\theta_i,\varphi_l) G(\theta_i) \exp\{j2\pi[(n-1)\omega_s(\theta_i,\varphi_l) + (k-1)\omega_t(\theta_i,\varphi_l)]\}\mathrm{d}\theta, \quad n = 1,2,\ldots,N; \quad k = 1,2,\ldots,K. \tag{2}$$

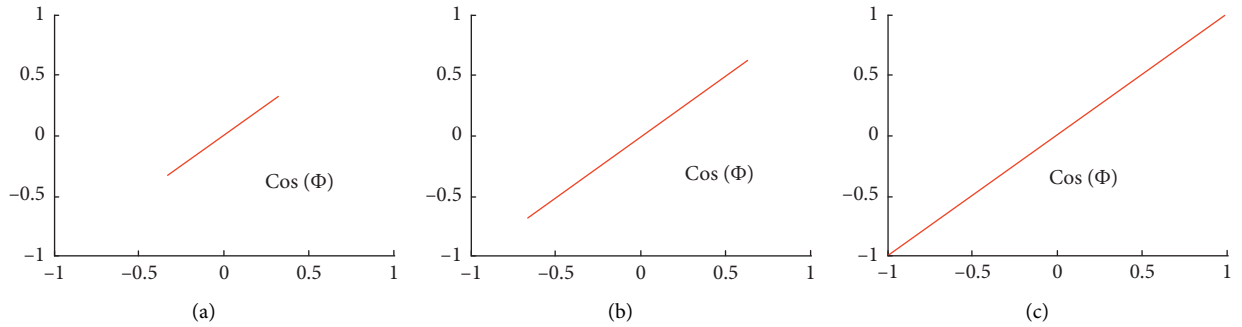


Figure 2: Clutter trace distribution (front side view array $\theta p = 0$). (a) Short range. (b) Medium range. (c) Long range.

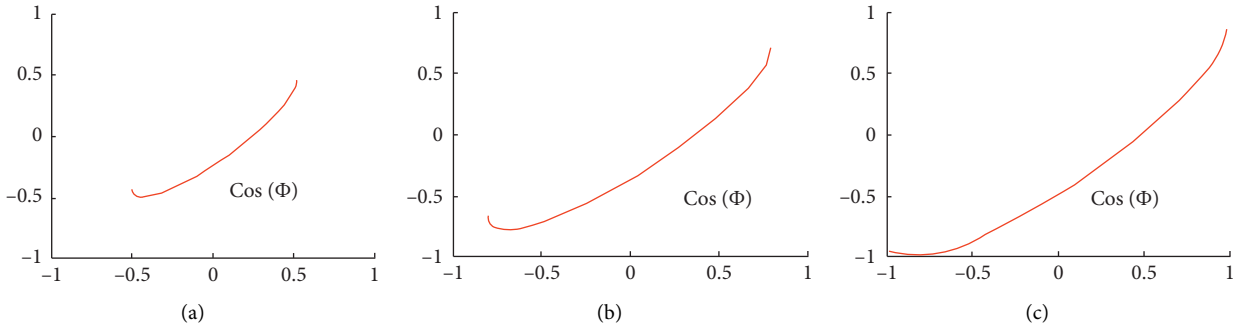


Figure 3: Clutter trace distribution (squint array $\theta p = 0$). (a) Short range. (b) Medium range. (c) Long range.

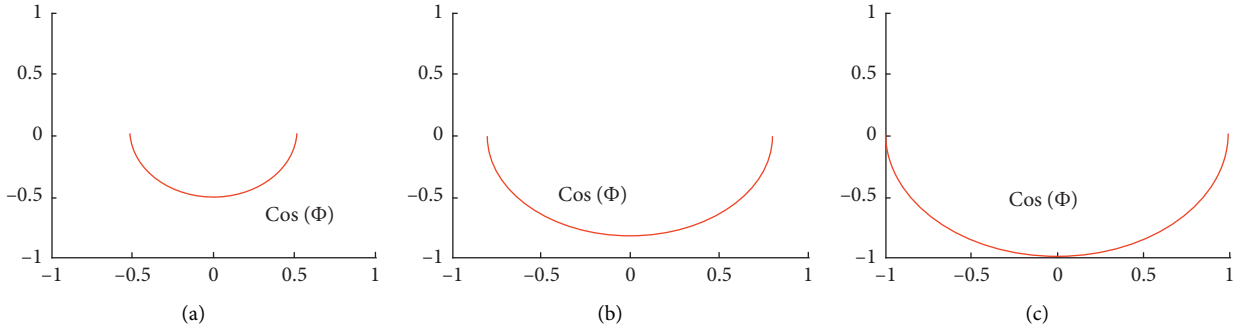


Figure 4: Clutter trace distribution (forward looking array $\theta p = 0$). (a) Short range. (b) Medium range. (c) Long range.

The result is the sum of the echoes in the range of 0 to π. The rest of the range cannot be received because the radar backplane is blocked. When the radar receives the sum of the echoes, it will divide the clutter scattering in a cut into N_c units. Then, formula (2) can be divided into

$$c_{n,k}(\varphi) = \sum_{i=1}^{Nc} \frac{\rho_i}{R_l^2} F(\theta_i, \varphi_l) G(\theta_i) \exp\{j2\pi[(n-1)\omega_s(\theta_i, \varphi_l) + (k-1)\omega_t(\theta_i, \varphi_l)]\}, \quad n = 1, 2, \ldots, N; \; k = 1, 2, \ldots, K. \tag{3}$$

Then, the clutter received by the radar whose array element number is N and the transmitted pulse number is K can be expressed by the array vector:

$$c(l) = \left[c_{1,1}(l), c_{1,2}(l), \ldots, c_{1,K}(l), c_{2,1}(l), \ldots, c_{n,k}(l), \ldots, c_{N,K}(l),\right]^T, \quad n = 1, 2, \ldots, N; k = 1, 2, \ldots, K, \tag{4}$$

where $(\cdot)^T$ is the transpose. In view of l and φ_l, $c_{n,k}(l)$ can be expressed as $c_{n,k}(\varphi)$. In general, there will be noise when the clutter echo is reflected back to the radar, so the echo received in a cut is represented as $x(l) = c(l) + n$, n as Gaussian white noise signal, and then the covariance matrix between the clutter and noise can be represented as

$$R = E\left[x(l)x(l)^H\right], \tag{5}$$

where $(\cdot)^H$ represents conjugate transposition and $E[\cdot]$ represents expectation. In general, the covariance matrix is determined by sample estimation. If the clutter distribution characteristics of a cut and its adjacent cut are the same, then its covariance matrix can be estimated by the covariance matrix of adjacent cells, as follows:

$$\widetilde{R} = \frac{1}{L}\sum_{\rho=1}^{L} x(p)x(p)^H, \tag{6}$$

where L represents the number of independent units. However, using this method will lead to the loss of spurious ratio (SCNR) of the output signal. The RMB criterion points out that only when the number of samples exceeds 2 times of matrix dimension can SCNR be less than 3 dB, so the number of matrix samples must be more than 2 times. Finally, the matrix can be transformed into radar clutter spectrum by using space-time Capon spectrum, and the clutter characteristics can be observed by clutter spectrum. The definition of space-time Capon spectrum is

$$p(\varphi,\theta) = \frac{1}{s^H(\varphi,\theta)R^{-1}s(\varphi,\theta)}, \tag{7}$$

where $s(\varphi,\theta)$ represents two-dimensional search oriented vector and $(\cdot)^{-1}$ represents matrix inversion.

The above is the process of building the clutter model of airborne phased array radar.

3.2. Space-Time Adaptive Processing Technology and Extended Factored Approach. In general, the signals reflected back to the airborne phased array radar can be represented by the following two assumptions:

$$\begin{aligned} H_0 &: s(l) = c(l) + n, \\ H_1 &: x(l) = c(l) + n + \beta s, \end{aligned} \tag{8}$$

where β is the normalized Doppler frequency of the target and other symbols are the same as the above, where $s = s_t \otimes s_s$, s_t, s_s represents the time domain and space domain guidance vectors of the target, respectively, which are calculated by the following formula:

$$\begin{aligned} s_t &= \left[1, \exp j2\pi\omega_t, \ldots, \exp j2\pi(K-1)\omega_t\right], \\ s_s &= \left[1, \exp j2\pi\omega_s, \ldots, \exp j2\pi(N-1)\omega_s\right]. \end{aligned} \tag{9}$$

In the reflected signal, clutter, noise, and target signal are all independent of each other. The radar samples the spatial and temporal data by the combination of antenna array and transmitting pulse, and STAP filters the data. Suppose w is the weight vector of STAP processor, then the output result is

$$z = w^H x(l). \tag{10}$$

The final purpose of STAP is to identify the target signal, so the selection of the optimal weight vector of STAP processor can be solved by the following formula:

$$\begin{cases} \min & E\left|w^H x(l)\right|^2 = w^H R w \\ \text{s.t.} & w^H s = 1. \end{cases} \tag{11}$$

The main purpose of this formula is to suppress clutter when the signal energy is constant, so as to minimize the output capacity of clutter and noise. The solution of w is obtained by Lagrange multiplier method, as follows:

$$w = \gamma R^{-1} s, \tag{12}$$

where γ represents the normalized constant. According to the above formula, STAP processor is a generalized Wiener filter. The improvement factor (IF) measures the key indicators of STAP processor, which can be defined as

$$\mathrm{IF} = \frac{\mathrm{SCNR}_{\mathrm{out}}}{\mathrm{SCNR}_{\mathrm{in}}}, \tag{13}$$

where $\mathrm{SCNR}_{\mathrm{out}}$ and $\mathrm{SCNR}_{\mathrm{in}}$ represent the signal to noise ratio of output and input, respectively, where

$$\mathrm{SCNR}_{\mathrm{out}} = s^H R^{-1} s. \tag{14}$$

If the input power of clutter and noise is set to $\sigma_c$and σ_w, respectively, then

$$\mathrm{SCNR}_{\mathrm{out}} = \frac{1}{\sigma_c + \sigma_w} = \frac{1}{(\mathrm{CNR}+1)\sigma_w}, \tag{15}$$

where CNR in represents the noise ratio, and then IF of the optimal STAP is expressed as

$$\mathrm{IF} = (\mathrm{CNR}+1)\sigma_w s^H R^{-1} s. \tag{16}$$

The above is the whole process of STAP, but the airborne phased array radar directly uses STAP technology to process the clutter signal, which requires a lot of hardware consumption, computation, and clutter covariance matrix. Therefore, in the practical application of STAP dimension reduction processing, other methods should be combined to reduce the use cost of STAP. This paper first introduces the EFA algorithm used in this paper.

In the process of STAP application, the Doppler filter with high out band attenuation can effectively suppress clutter, but the main lobe of Doppler filter will have the problem of clutter infiltration, so we need to take adaptive processing method to solve this problem. The main purpose of EFA algorithm is to use the data of Doppler channel and its adjacent channels to complete the adaptive processing. The basic principle of its structure is shown in Figure 5. The adjacent channel is the auxiliary channel, and the rest is the main channel, so as to improve the robustness of the time domain.

Assuming $\widetilde{y}_k(l) = \left[\widetilde{x}_{K-1}^T(l),\ \widetilde{x}_k^T(l),\ \widetilde{x}_{k+1}^T(l)\right]^T$ and $\widetilde{S}_k = \left[\widetilde{s}_{k-1}^T,\ \widetilde{s}_k^T,\ \widetilde{s}_{k+1}^T(l)\right]^T$, the constraint equations of three Doppler channels in signal processing can be expressed as follows:

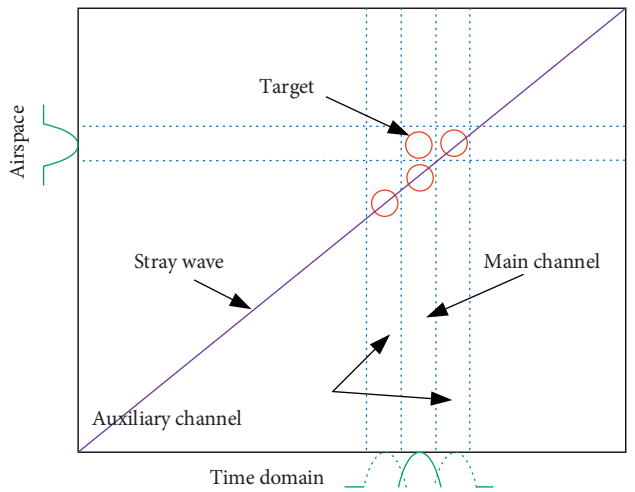


Figure 5: EFA fundamentals.

$$\begin{cases} \min & E\left[\left|\tilde{w}_{\mathrm{EFA}}^{H}\tilde{y}_k(l)\right|\right]^2 \\ \text{s.t.} & \tilde{w}_{\mathrm{EFA}}^{H}\tilde{S}_K = 1. \end{cases} \tag{17}$$

The optimal solution of the formula is expressed as

$$\tilde{w}_{\mathrm{EFA}}^{H} = \frac{R_{\tilde{y}_k}^{-1}\tilde{S}_k}{\tilde{S}_k^{H}R_{\tilde{y}_k}^{H}\tilde{S}_k}. \tag{18}$$

Corresponding $R_{\tilde{y}_k} = E[\tilde{y}_k(l)\tilde{y}_k^{H}(l)]$, according to the RMB criterion, the number of independent training samples required by EFA is well controlled, and the amount of calculation is further reduced.

3.3. Optimization of Space-Time Adaptive Processing Technology. The main process of post-Doppler adaptive processing (FA) is to suppress the main clutter by using time-domain Doppler and then carry out spatial adaptive processing. Therefore, the adaptive processing can be carried out in multiple Doppler channels at the same time. Here, it is assumed that the filter coefficient vector of Doppler channel k, $k = 1, 2, 3, \ldots, K$, is $f_k = [w_k^0, w_k^1, w_k^2, \ldots, w_k^{K-1}]^T$, where $w_k^l = \exp(j2\pi\overline{f}_k l)$, $\overline{fk} = (k - 1/K)f_r$ in the formula and $l = 0, 1, 2, \ldots, K-1$. In the combination of multiple Doppler filters, the center frequency of each filter is different, but the amplitude characteristics are the same. Then, the response function of amplitude is a periodic function with a period of f_r, as shown in Figure 6. Figure 6 is the amplitude response curve of a Doppler filter with a center frequency of $0.5f_r$. Between the first pair of zeros, the main response width of the filter is $2f_r/N$, while the width of the corresponding half power is about $0.9f_r/N$, while the height ratio of the main lobe to the side lobe of the filter is 13.2 db. In this process, the high sidelobe is easy to cause clutter leakage and ultimately affect the performance of Doppler filter adaptive processing, so it must be suppressed.

After Doppler channel filtering, the data of clutter and noise in the signal can be expressed by the following formula:

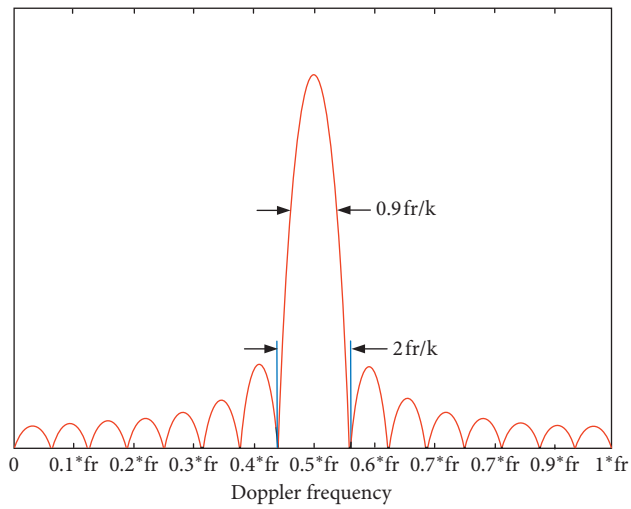


Figure 6: Amplitude response of a single number of Doppler filters.

$$\tilde{x}_k l = f_k^K \otimes I_N x(l). \tag{19}$$

The formula $N \times N$ represents unit matrix. Suppose that the target's guiding vector $s = s_t \otimes s_s$ and s_t, s_s represent the target's guiding vector in time domain and space domain, respectively. See formula (9) for the calculation method of s_t, s_s, and the data of the target's guiding vector after k filtering will become

$$\tilde{x}_k = \left(f_k^K \otimes I_N\right)s. \tag{20}$$

According to the RMB criterion, FA processing is used to keep the energy of k target signal unchanged, and then the output ability of clutter is reduced to achieve the purpose of clutter suppression:

$$\begin{cases} \min & E\left[\tilde{w}_{\mathrm{FA}}^{H}\tilde{x}_k(l)\right]^2 \\ \text{s.t.} & \tilde{w}_{\mathrm{EFA}}^{H}\tilde{S}_K = 1. \end{cases} \tag{21}$$

The optimal solution can also be obtained by using Lagrange formula:

$$\tilde{w}_{\mathrm{FA}}^{H} = \frac{R_{\tilde{x}_k}^{-1}\tilde{S}_k}{\tilde{S}_k^{H}R_{x_k}^{-1}\tilde{S}_k}, \tag{22}$$

where $R_{\tilde{x}_k} = E[\tilde{x}_k(l)\tilde{x}_k^{H}(l)]$ is the covariance matrix of the Doppler channel k. For EFA, according to formulas (17) and (18), the optimal solution can also be obtained, and there is $R_{\tilde{y}_k} = E[\tilde{y}_k(l)\tilde{y}_k^{H}(l)]$. However, in the actual processing, the clutter and noise data of the unit distance unit adjacent to the target unit of the covariance matrix can be obtained. Then, according to formula (6), $R_{\tilde{x}_k}$ and $R_{\tilde{y}_k}$ can be obtained:

$$\begin{aligned} \hat{R}_{\tilde{x}_k} &= \frac{1}{L}\sum_{l=1}^{L}\tilde{x}(l)\tilde{x}_k^{H}(l), \\ \hat{R}_{\tilde{y}_k} &= \frac{1}{L}\sum_{l=1}^{L}\tilde{y}(l)\tilde{y}_k^{H}(l), \end{aligned} \tag{23}$$

where L has the same meaning as in formula (6). Theoretically, when the number of training samples of Doppler filter bank $L \longrightarrow \infty$, there should be $R_{\overline{y}k} = \widehat{R}_{\overline{y}_k}$ and $R_{\overline{x}k} = \widehat{R}_{\overline{x}_k}$, However, from a practical point of view, the number of samples cannot be infinite and some studies have proved that, when the number of samples is more than twice the number of covariance matrix, the output loss of signal to noise ratio will not be greater than 3 dB. Although the number of samples and the amount of calculation have decreased, but when the number of radar antenna elements increases, the number of samples will increase, eventually leading to the reduction of clutter suppression performance. Therefore, further dimension reduction is needed.

The clutter data processed by Doppler channel k can be expressed by the following formula when dimension reduction is carried out from the perspective of spatial data:

$$\widetilde{z}_k(l) = \sum_{i=1}^{Nc} \alpha_i c_{s,i}, \tag{24}$$

where $\alpha_i = (\sigma_i(\theta_l, \varphi_i)/R_l^2) F\theta_l, \varphi_i G\theta_l f_k^H c_{t,i}$ is a constant. It can be seen from the above formula that the filtered data become the product of the spatial guiding vector of the clutter and a constant, which belongs to the column vector, and the size of the column vector is $N \times 1$, while $c_{s,i}$ can continue to be represented by the Kronecker product of the short vector, as follows:

$$\begin{aligned} c_{s,i} &= \left[1, \ldots, \exp\left(2j\pi(N_1 - 1)N_2\omega_{s,i}\right)\right]T \\ &\quad \otimes \left[1, \ldots, \exp j2\pi(N_2 - 1)\omega_{s,i}\right]^T, \\ a_i &= \left[1, \exp\left(2j\pi N_2\omega_{s,i}\right), \ldots, \exp\left(2j\pi(N_1 - 1)N_2\omega_{s,i}\right)\right]^T, \\ b_i &= \left[1, \exp\left(j2\pi\omega_{s,i}\right), \ldots, \exp\left(j2\pi(N_2 - 1)\omega_{s,i}\right)\right]T. \end{aligned} \tag{25}$$

The above two formulas are the column vectors of $N_1 \times 1$, $N_2 \times 1$, and $N = N_1 N_2$; then, the formula (24) can be expressed as

$$\widetilde{z}_k(l) = \sum_{i=1}^{Nc} \alpha_i c_{s,i} = \sum_{i=1}^{Nc} \alpha_i a_i \otimes b_i. \tag{26}$$

Then, formula (17) is in the form of

$$\widetilde{w}_{\mathrm{FA}} = u \otimes v, \tag{27}$$

where $u = [u_1, u_2, \ldots, u_{N1}]^T v = [v_1, v_2, \ldots, v_{N2}]^T$; substituting formula (27) into formula (17), we can get

$$\begin{cases} \min & E\left[\left|\widetilde{w}_{\mathrm{FA}}^H \widetilde{x}_k(l)\right|\right]^2 = E\left[\left|(u \otimes v)^H \widetilde{x}_k(l)\right|^2\right] \\ \text{s.t.} & \widetilde{w}_{\mathrm{FA}}^H \widetilde{s}_k = (u \otimes v)^H \widetilde{s}_k = 1. \end{cases} \tag{28}$$

Formula (28) becomes a biquadratic cost function of vector u and vector v. Then, according to the theory of cycle minimization, u, or v, get the initial value randomly and then substitute it into formula (28) to solve until the final error value is less than the threshold value. That is, fix the vector u first, give it a specified initial value, write it down as $u(0)$, and then substitute it into formula (28), you can get

$$\begin{cases} \min & E\left[\left|(u(0) \otimes v)^H \widetilde{x}_k(l)\right|^2\right] = E\left[\left|v^H (u(0) \otimes I_{N2}) \widetilde{x}_k(l)\right|^2\right] \\ \text{s.t.} & (u(0) \otimes v)^H \widetilde{s} = v^H (u(0) \otimes I_{N2})^H \widetilde{s} = 1, \end{cases} \tag{29}$$

where IN_2 is the unit matrix of $N_2 \times N_2$, and if it is named $R_v = (u(0) \otimes I_{N2})^H R_{\widehat{x}_k} (u(0) \otimes I_{N2})$, $s_v = (u(0) \otimes I_{N2})^H \widetilde{s}$, then we can get

$$v(1) = \frac{R_v^{-1} s_v}{s_v^H R_v^{-1} s_v}. \tag{30}$$

Then, the obtained $v(1)$ value is substituted into formula (28) to continue the iterative calculation, and

$$\begin{cases} \min & E\left[\left|(u \otimes v(1))^H \widetilde{x}_k(l)\right|^2\right] = E\left[\left|u^H (I_{N1} \otimes v(1))^H \widetilde{x}_k(l)\right|^2\right] \\ \text{s.t.} & (u \otimes v(1))^H \widetilde{s} = u^H (I_{N1} \otimes v(1))^H \widetilde{s} = 1, \end{cases} \tag{31}$$

where IN_1 is $N_1 \times N_1$'s unit matrix. Similarly, $R_u = (I_{N1} \otimes v(1))^H R_{\widehat{x}_k} (I_{N1} \otimes v(1))$, $s_u = (v(1) \otimes I_{N1})^H \widetilde{s}$, we can get

$$u(1) = \frac{R_u^{-1} s_u}{s_u^H R_u^{-1} s_u}. \tag{32}$$

Cycle the steps of the above iteration until the k step is reached. $\|vk - vk - 1\|/\|vk\| < \delta$, δ is the error threshold and $\|\cdot\|$ is the 2 norm of the vector. In the process of iteration, the Doppler covariance matrix of phased array meets $R_v \in C^{N_2 \times N_2}$ and $R_u \in C^{N_1 \times N_1}$, respectively. Then, by decomposing the spatial data, the training samples of covariance matrix only need to be larger than $2\max(N_1, N_2)$, and the closer N_1, N_2 value is, the smaller the sample size is.

4. Result Analysis

4.1. Convergence and Computational Analysis. In this paper, the convergence and computation of the optimized STAP are verified and analyzed. First, the life cost function is

$$J(u, v) = E\left[\left|(u \otimes v)^H \widetilde{x}_k(l)\right|^2\right] + \mu\left[(u \otimes v)^H \widetilde{s}_k - 1\right]. \tag{33}$$

According to the formula, it can be seen that the cost function is continuous and differentiable. Substituting $u(0), \ldots, u(k)$, $v(1), \ldots, v(k)$ is calculated in the previous section with the cost function $J(u, v)$; the following results can be obtained:

$$\begin{aligned} J(u(0), v(1)) \geq J(u(1), v(1)) \geq \cdots \geq J(u(k-1), v(k-1)) \\ \geq J(u(k-1), v(k)) \geq J(u(k), v(k)). \end{aligned} \tag{34}$$

In the above results, it is obvious that the results of each cycle iteration satisfy the monotonic decrease. A $\|v\| = 1$, you can get the following results:

$$J(u(t), v(t)) = \overline{J}(w) = w^H R_{\tilde{x}k} w \geq \lambda_{\min} \|w\|, \tag{35}$$

where $\lambda_{\min}$ represents the minimum eigenvalue of the covariance matrix $R_{\tilde{x}_k}$ of STAP. This result shows that there is a constant of $0 < c < +\infty$, let $\lambda_{\min}\|w\| \leq \overline{J}(w) = w^H R_{\tilde{x}_k} w < c$ hold, this result shows that the cost function is bounded. According to the above analysis, we can know that $J(u(t), v(t))$ is an Lyapunov function, and it further shows that the iterative algorithm in the previous section has the characteristics of progressive convergence.

After it is proved that the iterative cycle adopted in this paper for STAP optimization has convergence, the sample size calculated by several algorithms is shown in Table 1, and the last sdd-fa and sdd-efa are the optimized algorithms in this paper. It can be seen from the table that the number of full-dimensional space-time adaptive samples is 2048 and the calculation amount reaches 109, while the number of samples of FA algorithm and EFA algorithm is 128 and 384, respectively. In this paper, the number of samples of FA algorithm and EFA algorithm after spatial decomposition and dimension reduction optimization is 16 and 32, respectively, and the calculation amount is also within 103. The results show that the number of samples and the amount of computation of FA algorithm and EFA algorithm are greatly optimized after the optimization processing of spatial decomposition and dimensionality reduction, which shows that the STAP performance of FA algorithm and EFA algorithm is greatly improved.

4.2. Performance Simulation of Improvement Factor. In this paper, the optimized STAP performance is simulated. In general, the performance of STAP is measured by improvement factor. Therefore, it is assumed that the data of airborne phased array radar and other relevant data involved in the experiment are shown in Table 2, in which the array of experimental airborne phased array radar is 64×64, the number of pulses in the unit coherent processing compartment $K = 16$, the pulse repetition frequency $f_r = 2000$ Hz, the wave length of radar emission $\lambda = 0.2$ m, the flight height and speed of radar carrier are 9000 m and 100 M/s, respectively, the spurious ratio is 60 dB, and the clutter relative broadband is set as 0.02.

Figure 7 shows the results of improvement factor performance at 500 samples. The purple dotted line opt in the figure represents the optimal STAP, and this curve is mainly used as a reference. The rest are several algorithms involved in the experiment. It can be seen from the figure that, as the normalized Doppler frequency increases from small to large, the improvement factors of all algorithms appear as a decline rise process. When the normalized Doppler frequency value is 0, the improvement factors of all algorithms are minimized. But compared with the opt curve, we can see that the EFA algorithm and the EFA algorithm (sdd-efa), which are optimized by spatial data dimensionality reduction, have the best improvement factor. The improvement factor data of FA algorithm and EFA algorithm (sdd-fa) are basically the same, but they are far from the ideal data.

Table 1: Performance comparison between different algorithms.

Algorithm	Index	
	Amount of computation	Sample size
Full-dimension STAP	$O(10^9)$	2048
FA	$O(10^5)$	128
EFA	$O(10^6)$	384
SDD-FA	$O(10^2)$	16
SDD-EFA	$O(10^3)$	32

Table 2: Experimental phased array radar and other relevant data.

Name	Symbol	Numerical value
Airborne phased array radar array	—	64×64
Number of pulses	K	16
Pulse repetition rate	fr	2000 Hz
Radar wavelength	λ	0.2 m
Flight altitude of carrier	h_a	9000m
Carrier flight speed	v_a	100 m/s
Clutter relative broadband	B_r	0.02
Heterozygosity ratio	CNR	60 dB

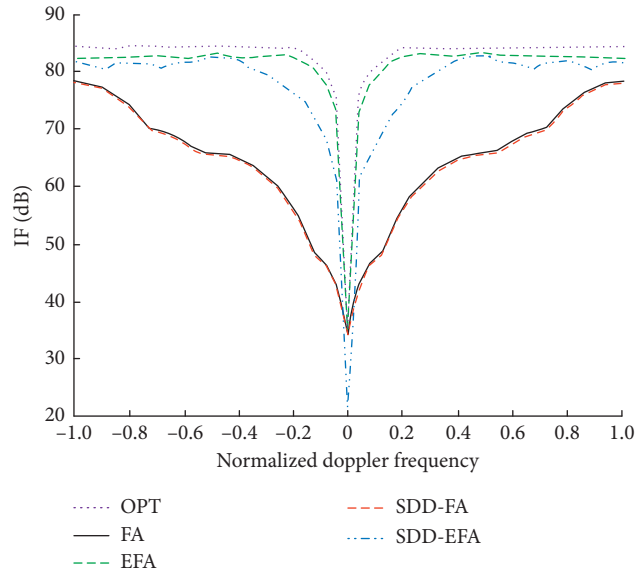


Figure 7: Performance comparison of improvement factors under 500 samples.

Figure 8 shows the statistical results of the improvement factor curve of the four algorithms and the ideal improvement factor curve when the number of samples is 50. It can be seen from the figure that, in the case of a small number of samples, the improvement factor curves of the four algorithms are greatly different. The FA algorithm is the farthest from the ideal curve, and the improvement factor curve of the sdd-efa algorithm is the closest to the ideal curve, which shows that the performance of the algorithm is the closest to the ideal STAP performance. Sdd-fa algorithm and EFA algorithm are relatively similar. The performance of STAP in different normalized Doppler frequencies is staggered. The STAP performance of EFA

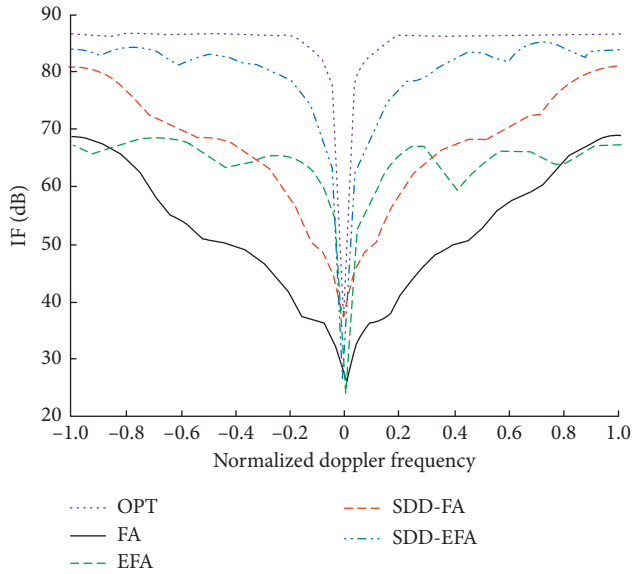


Figure 8: Comparison of improvement factor performance under a number of samples.

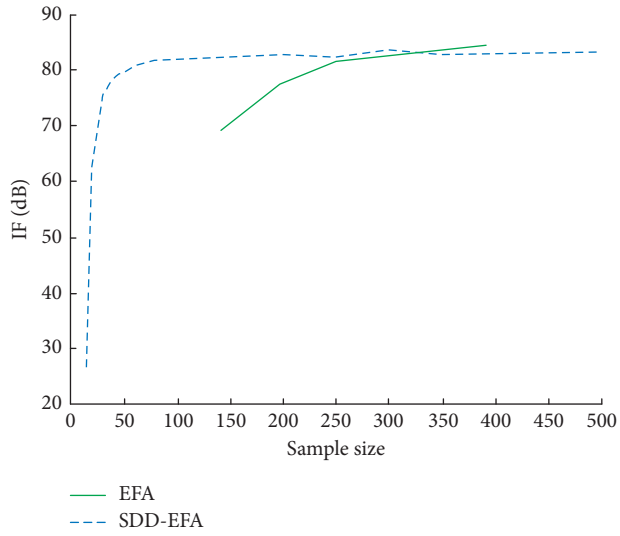


Figure 9: Performance comparison of EFA algorithm and sdd-efa algorithm under different training samples.

algorithm in Doppler frequency −0.2~0.2 is better. The rest of the frequency curves are more prominent in sdd-fa algorithm, but the STAP performance of EFA in different frequencies is more stable.

Figure 9 shows the performance comparison of EFA algorithm and sdd-efa algorithm under different training samples. As can be seen from the figure, with the increase of the number of training samples, the improvement factors of the sdd-efa algorithm show an upward stable trend. When the number of samples is 50, the curve area is stable, which shows that the STAP performance of the sdd-efa algorithm tends to be stable when the number of samples exceeds 50, while the EFA algorithm starts to appear when the number of samples is 200 and tends to be stable when the number of samples is 350. The comparison shows that the STAP sample of sdd-efa algorithm has a wider range of applicability.

Figure 10 shows the relationship between the sdd-efa algorithm and the number of iterations. It can be seen from the figure that with the increase of the number of iterations, the improvement factor of the sdd-efa algorithm increases gradually and reaches a stable stage after nine iterations. The result shows that the performance of the SATP tends to be stable when the number of iterations exceeds nine.

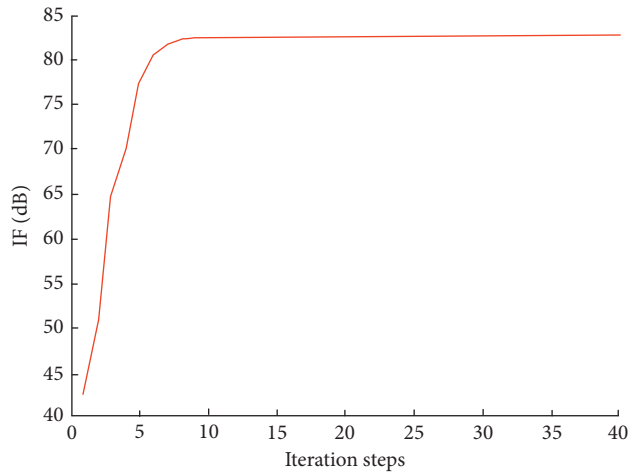


Figure 10: Relationship between the sdd-efa algorithm and the number of iterations.

4.3. Performance Analysis of Clutter Removal. The experimental results of STAP performance in the sidelobe region of phased array radar are shown in Figures 11–14. Figure 11 shows the statistical results of STAP performance of four algorithms in the Sidelobe region. It can be seen from the figure that, with the gradual increase of the normalized Doppler frequency, the STAP performance of the four algorithms also experienced a downward to upward trend, and the improvement factors of all algorithms reached the lowest near 0.6. But in the whole stage, the biggest improvement factor is still the sdd-efa algorithm, followed by the sdd-fa algorithm. The two are close in most of the normalized Doppler frequency range, but between 0.3 and 1.0, the performance of sdd-efa algorithm is better. FA fluctuates with the frequency of normalized Doppler, and the performance of EFA is very unstable.

Figure 12 shows the change trend of improvement factors of EFA algorithm and sdd-efa algorithm in sidelobe clutter area under different training samples. It can be seen from the figure that the trend of STAP performance of the two algorithms in the side lobe clutter area with the number of samples is basically similar to the result shown in Figure 9, that is, the STAP performance of the sdd-efa algorithm in the side lobe clutter area is more extensive in the range of the number of training samples, and the advantages are more obvious.

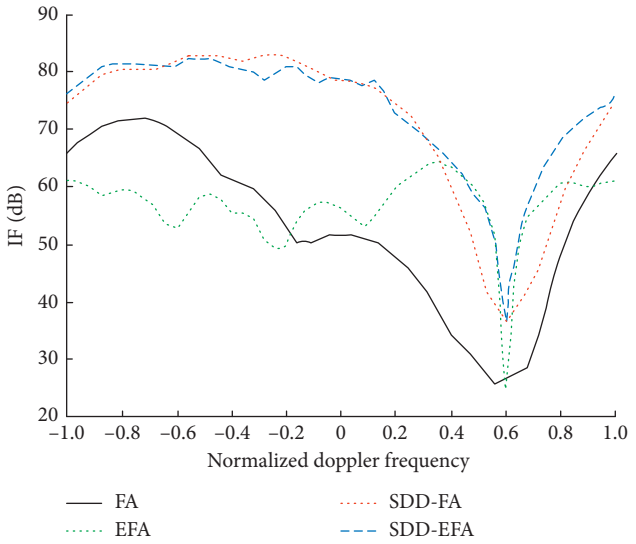


Figure 11: Performance comparison of STAP improvement factors in side lobe area.

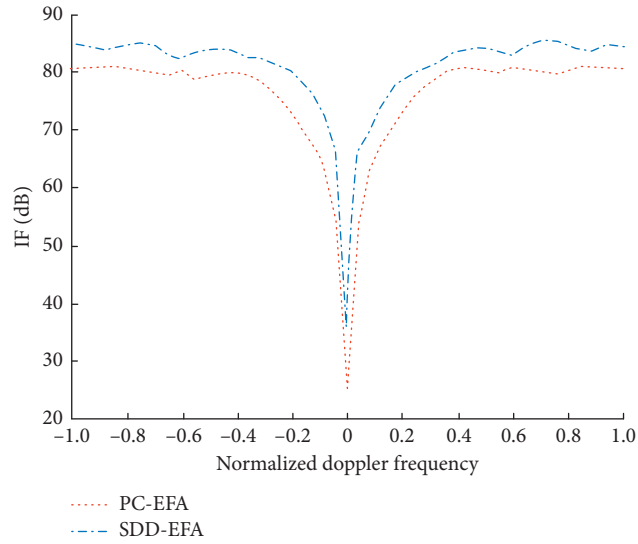


Figure 13: Comparison of morning STAP performance of two different optimization schemes with 50 samples.

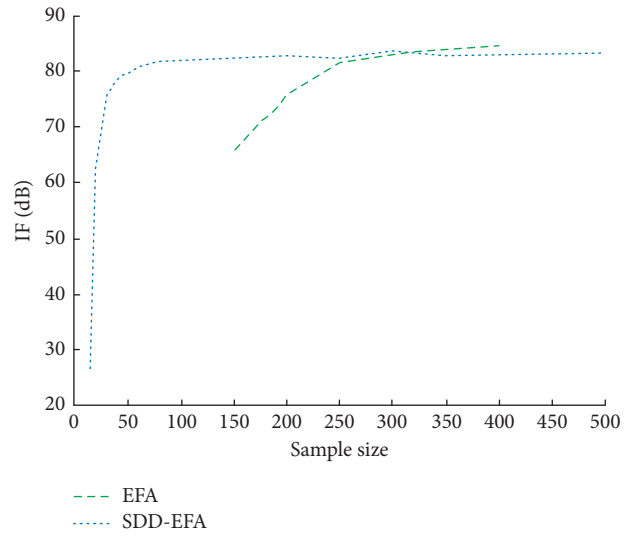


Figure 12: Comparison of sidelobe clutter STAP convergence performance.

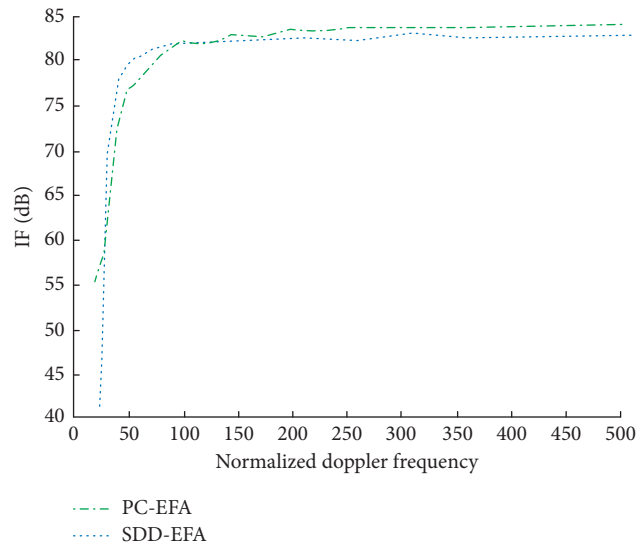


Figure 14: STAP performance comparison of two different optimization schemes under different sample numbers.

In order to verify the advantages and disadvantages of this optimization scheme compared with other optimization schemes, we choose to compare it with pc-efa algorithm. Figure 13 shows the improvement factor comparison between pc-efa algorithm and sdd-efa algorithm in 50 samples. It can be seen from the figure that the improvement factor values of the two optimization schemes are relatively close, but the improvement factor values of the optimization scheme in this paper are larger than those of the pc-efa algorithm, which is basically kept at about 80 dB, while the pc-efa algorithm is kept at about 75 db.

Figure 14 shows the improvement factor performance comparison of the two optimization schemes at different sample numbers. It can be seen from the figure that, in the initial stage, the improvement factor performance of pc-efa algorithm is better than that of this algorithm, but with the increase of the number of samples, the improvement factor performance of this optimization scheme gradually exceeds that of pc-efa algorithm, and the two are basically close when the number of samples is 100. In terms of the convergence speed, the convergence speed of this optimization scheme is faster and the convergence performance is better than that of pc-efa algorithm.

5. Conclusion

Because the airborne phased array radar is flying at high altitude, it will receive a lot of clutter when it scans at high altitude. Therefore, clutter suppression becomes the main direction to improve the performance of phased array radar.

At present, space-time adaptive processing technology (STAP) is the main technology to suppress clutter of phased array radar. However, STAP itself has many problems, so it can not be directly applied and so this paper proposes to optimize it. In this paper, EFA algorithm in STAP needs a large number of samples and computation, so a method of spatial data dimensionality reduction based on cyclic iterative calculation is adopted to optimize EFA algorithm. Finally, the experimental results show that the STAP performance of the optimized STAP algorithm is greatly improved, especially in the case of a small number of samples; the STAP performance of the EFA algorithm optimized in this paper is basically close to the ideal STAP performance; for clutter suppression in sidelobe clutter area, the optimization scheme in this paper also shows more excellent STAP performance; and compared with other optimization schemes, the optimization scheme in this paper is better in STAP performance and convergence speed, which shows that the optimization scheme in this paper is successful. However, the optimization scheme in this paper has only carried out simulation experiments, there is lack of practical application proof, and it is need to be further verified.

Data Availability

The raw/processed data required to reproduce these findings cannot be shared at this time as the data also form part of an ongoing study.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This project was granted financial support from the Key Project of the Sichuan Provincial Department of Education under Grant no. 18ZA0034.

References

[1] P. Setlur and M. Rangaswamy, "Waveform design for radar STAP in signal dependent interference," *IEEE Transactions on Signal Processing*, vol. 64, no. 1, pp. 19–34, 2016.

[2] Y. Fan, Y. Liu, Y. Liu, J. An, and X. Bu, "Robust two-stage reduced-dimension STAP algorithm and its performance analysis," *Journal of Systems Engineering and Electronics*, vol. 27, no. 5, pp. 954–960, 2016.

[3] J. Xu, S. Zhu, and G. Liao, "Range ambiguous clutters suppression for airborne FDA-STAP radar," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 8, p. 1, 2015.

[4] M. K. Mcdonald and D. Cerutti-Maori, "Coherent radar processing in sea clutter environments, part 1: modelling and partially adaptive STAP performance," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 52, no. 4, pp. 1797–1817, 2016.

[5] J. Xu, G. Liao, L. Huang et al., "Robust adaptive beamforming for fast-moving target detection with FDA-STAP radar," *IEEE Transactions on Signal Processing*, vol. 65, no. 4, pp. 973–984, 2016.

[6] B. Kang, "Student research highlight estimation of structured covariance matrices for radar STAP," *IEEE Aerospace and Electronic Systems Magazine*, vol. 30, no. 2, pp. 24-25, 2015.

[7] G. Yiduo, G. Jian, H. Darong et al., "Unified theoretical frame of a joint transmitter-receiver reduced dimensional STAP method for an airborne MIMO radar," *Journal of Radars*, vol. 5, no. 5, pp. 517–525, 2016.

[8] X. Wang, E. Aboutanios, and M. G. Amin, "Reduced-rank STAP for slow-moving target detection by antenna-pulse selection," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1150–1160, 2015.

[9] Y. Zhou, Z. Chen, L. R. Zhang et al., "Radar parameters aided STAP for heterogeneous clutter suppression," *Journal of the University of Electronic Science and Technology of China*, vol. 45, no. 6, pp. 905–910, 2016.

[10] H. Lin, Z. Yong-Shun, L. I. Zhe et al., "A dimension-reduced rapid STAP algorithm for airborne MIMO radar," *Electronics Optics & Control*, vol. 7, no. 22, pp. 14–16, 2015.

[11] C. Shi-Wa, Z. Qing-Song, Z. Jian-Yun et al., "Signal characteristic analysis of scattered-wave fake clutter jamming method for STAP radar," *Acta Electronica Sinica*, vol. 46, pp. 401–409, 2018.

[12] Z. Gao and H. Tao, "Robust STAP algorithm based on knowledge-aided sparse recovery for airborne radar," *Iet Radar Sonar & Navigation*, vol. 11, no. 2, pp. 321–329, 2016.

[13] B. L. McKinley, L. Kristine, and Bell, "Range-dependence compensation for bistatic STAP using focusing matrices," in *Proceedings of the IEEE National Radar Conference-Proceedings*, pp. 1750–1755, Arlington, VA, USA, May 2015.

[14] C. Wen, Y. J. Peng, and J. Wu, "Enhanced three-dimensional joint domain localized STAP for airborne FDA-MIMO radar under dense false-target jamming scenario," *IEEE Sensors Journal*, vol. 18, no. 10, pp. 4154–4166, 2018.

[15] S. Audrey, W. L. Melvin, and D. B. Williams, "Linear prediction based temporal weighting for pre-Doppler STAP," in *Proceedings of the IEEE National Radar Conference*, pp. 478–483, Arlington, VA, USA, May 2015.

[16] Z. Li, Y. Zhang, W. Liu et al., "Robust STAP against unknown mutual coupling based on middle subarray clutter covariance matrix reconstruction," *IEEE Access*, vol. 7, p. 1, 2019.

[17] Z. Chen, Y. Zhou, L. Zhang et al., "Single data set-STAP algorithm in heterogeneous environment," *Hangkong Xuebao/Acta Aeronautica et Astronautica Sinica*, vol. 36, no. 12, pp. 3938–3946, 2015.

[18] K. Duan, H. Xu, H. Yuan et al., "Reduced-DOFs three-dimension STAP via subarray synthesis for non-sidelooking planar array airborne radar," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 56, no. 4, p. 1, 2019.

[19] G. Zhi-Qi, T. Hai-Hong, and Z. Ji-Chao, "Robust STAP algorithm based on joint sparse recovery of clutter spectrum for airborne radar," *Acta Electronica Sinica*, vol. 44, no. 11, pp. 2796–2801, 2016.

[20] F. Jia, Y. He, and R. Li, "Online antenna-pulse selection for STAP by exploiting structured covariance matrix," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E102.A, no. 1, pp. 296–299, 2019.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 5914062, 9 pages
https://doi.org/10.1155/2020/5914062

Research Article

Improved Error Correction Methods for Filterless Digital Class D Audio Power Amplifier Based on FCLNF

Li Li,[1,2] **Hong-jie Li**,[2] **and Yan-jing Sun**[1]

[1]*School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China*
[2]*Department of Electronic Information and Electrical Engineering, Anyang Institute of Technology, Anyang 455000, China*

Correspondence should be addressed to Li Li; lilifkb@163.com

Received 4 May 2020; Revised 19 July 2020; Accepted 22 July 2020; Published 12 August 2020

Guest Editor: Sanghyuk Lee

Aiming at correcting the error caused by the nonlinear and power supply noise of the bridge-tied-load (BTL) power stage of the filterless digital class D power amplifier, an error correction method was proposed based on feedforward power supply noise suppression (FFPSNS) and first-order closed loop negative feedback (FCLNF) techniques. This method constructed the first-order LCLNF loop for the power stage and further reduced the impact of the power supply noise on the power amplifier output by using FFPSNS technology to introduce the power supply noise into the feedback loop at the same time. The 0.35 μm CMOS process is used for analysis and comparison in Cadence. Cadence simulation results indicate that PSRR at the power supply noise frequency of 200 Hz is improved with 36.02 dB. The power supply induced intermodulation distortion (PS-IMD) components are decreased by approximately 15.57 dB and the signal-to-noise ratio (SNR) of the power amplifier is increased by 17 dB. The total harmonic distortion + noise (THD + N) of the power amplifier is reduced to 0.02% by FCLNF + FFPSNS.

1. Introduction

In recent years, the country has promoted high-efficiency and energy-saving technologies and encouraged the strengthening of high-efficiency and energy-saving technological transformations and key technological breakthroughs. With the rise of audio and video equipment for low-power applications, filter-free digital class D audio power amplifiers have high power efficiency and easy interface with digital audio sources, which are favored by researchers in the industry [1]. And in the 11 key technologies evaluated by IEEE Spectrum in the past ten years, it is predicted that high-efficiency class D audio amplifiers will eventually unify the audio amplifier market [2]. However, its own nonideal state and power supply noise can cause serious distortion of the output signal of the power amplifier.

Digital class D power amplifiers have received increasing attention due to their advantages of high power efficiency, easy system transplantation, and resistance to external interference [3]. The traditional digital class D power amplifier is usually composed of a digital switching signal modulator, a power stage, and an inductive capacitor (LC) analog low-pass filter. The digital switching signal modulator mostly uses uniform-sampling pulse width modulation (UPWM) technology realization. The LC low-pass filter of the traditional digital class D power amplifier occupies about 75% of the volume of the entire power amplifier system and consumes about 30% of the cost, which severely hinders the portable application of digital class D power amplifiers [4]. The filter-free digital class D power amplifier, as a new type of digital class D power amplifier, can drive the loudspeaker to work without the need of an LC low-pass filter through a special modulation technique and maintain high power efficiency to meet the needs of digital audio with the development of equipment miniaturization; the filter-free digital class D power amplifier has become a research hotspot in the current power amplifier field [5].

In order to enable the digital class D audio power amplifier to achieve high power supply rejection ratio (PSRR), power supply-induced intermodulation Distortion (PS-IMD), signal-to-noise rate (SNR), and total harmonic distortion + noise (THD + N), two large error sources need

to be corrected for the nonidealities and the power level source noise introduced by the power level [6].

For filter-free digital class D power amplifiers, the errors introduced by the nonlinearity and nonideal working conditions of the power stage, especially the errors introduced by the power supply noise of the power stage, seriously affect the output performance of the power amplifier [7, 8]. Although the power amplifier's power supply rejection ratio (PSRR) can theoretically reach infinity when the upper and lower half bridges of the bridge-tied-load (BTL) power stage of the power amplifier are completely matched, the power supply interference is caused by the power supply noise. The modulation distortion (PS-IMD) is still large. In response to this problem, Donida et al. [7] and Mostert et al. [8] introduced the output signal of the power amplifier to the input of the power amplifier by establishing a global closed-loop negative feedback loop including a digital switching signal modulator and a power stage. Precorrection is performed in the terminal or switching signal modulator to achieve the purpose of correcting the power level error, but an additional analog-to-digital converter (ADC) is required, which leads to a substantial increase in system cost. Chen et al. [9] and Cellier et al. [10] built a local closed loop negative feedback (LCLNF) loop containing only the power level to correct the power level error. This method is ensuring the system. In a stable condition, the ability to suppress the noise of the power level power supply is weak. In summary, there are currently fewer error correction methods for the power stage of the filter-free digital class D power amplifier, and the existing digital class D power amplifier power stage error correction method may have a higher cost of implementation, or the power stage power error. The correction ability is weak.

Aiming at the above two major error sources, Dong Jun Lee and Jinho Noh et al. used the local negative feedback technology to establish a power-level noise suppression module between the switch signal modulator and the power level, but it would cause an increase in the power level switching frequency [11, 12]. There are also the use of analog negative feedback control loop schemes and the use of power-level power supply noise feedforward precorrection technology [13–15]. The scheme is complex and requires the use of ADC to convert power source noise into digital signals. In [16], a feedforward noise technology for class D audio amplifier with single end (SE) output structure is proposed, but this method is applied to class D audio amplifier with SE output structure, and the system contains an off-chip LC low-pass filter. This paper designs a first-order closed-loop negative feedback (FCLNF) error correction method for the filter-free digital class D audio power amplifier bridge-tied-load (BTL) power stage and the FCLNF suitable for the filter-free digital class D audio power amplifier BTL power stage plus feedforward power supply noise suppression (FCLNF + FFPSNS) error correction method.

2. FCLNF + FFPSNS Error Correction

Class D power stage is one of the key modules of the filter-free digital class D amplifier. Its function is to amplify the weak PWM signal to drive the low impedance speaker. Its input signal and output signal are both digital PWM signals. Since the ripple noise contained in the power supply and the nonideal state of the power stage itself affect the quality of the output signal, it is necessary to perform error correction on the open loop power stage.

Filter-free digital class D audio power amplifier is mainly composed of digital UPWM modulator and BTL power stage. The former is a digital part and the latter is an analog part. Its structure is shown in Figure 1. It can be seen from Figure 1 that the filter-free digital class D audio power amplifier can be divided into UPWM modulator, open-loop BTL power stage, and loudspeaker as a whole. The UPWM modulator is a digital circuit, and the open-loop BTL power stage is an analog circuit. UPWM modulator consists of the digital interpolation filter, in-phase and inverted unity gain buffer, Sigma-Delta modulator, and UPWM generator.

2.1. Principle Analysis of FCLNF Error Correction. FCLNF circuit is composed of compensation module, remodulation module, BTL power level, and some passive components [3, 17]. Its circuit and equivalent model are shown in Figure 2. There V_{IN_P} and V_{IN_N} are, respectively, the input signals and V_{CM} is the common mode voltage or reference voltage.

F_1 and F_2 are transfer functions from the input signal of V_{IN_P} and V_{IN_N} to the inverting input of the integrator. G_{int1} and G_{int2} are the integrator gains of the upper and lower bridges, respectively. H_1 and H_2 are the feedback factors of the upper and lower bridges, respectively. G_{M1} and G_{M2} are the combined linear gain of the power level and the remodulation module.

Under ideal conditions,

$$G_{M1} = G_{M2}, V_{IN\_P} = V_{IN\_N} = 0, \tag{1}$$

if the deviation between the two feedback paths is ε.

The mismatch factor between resistances is $r\,(0 < r < 1)$, mismatch factor between capacitors is $c\,(0 < c < 1)$, and ε is equivalent to $r + c$,

The PSRR of power amplifier using the FCLNF method is

$$\mathrm{PSRR}_{FCLNF} = 20\lg\left[LG_2(s)\right] - 20\lg(r + c). \tag{2}$$

2.2. Principle Analysis of FCLNF + FFPSNS Error Correction. FFPSNS feeds the power stage power supply noise forward to the power stage front-end module for precorrection through certain means. The feedforward control network subtracts the feedforward power supply noise from the PWM input signal and then quantifies the power supply noise through the comparator. The power supply noise is converted into the effective duty cycle of PWM signal, so as to achieve the purpose of correcting the power supply noise. This proposal is composed of FCLNF module, FFPSNS module, remodulation module, and BTL power level [18, 19]. Its schematic diagram and equivalent model are shown in Figure 3.

There V_{IN_P} and V_{IN_N} are, respectively, the input signals of the upper and lower half bridges of the power port in the power amplifier. V_{CM} is the common mode voltage or

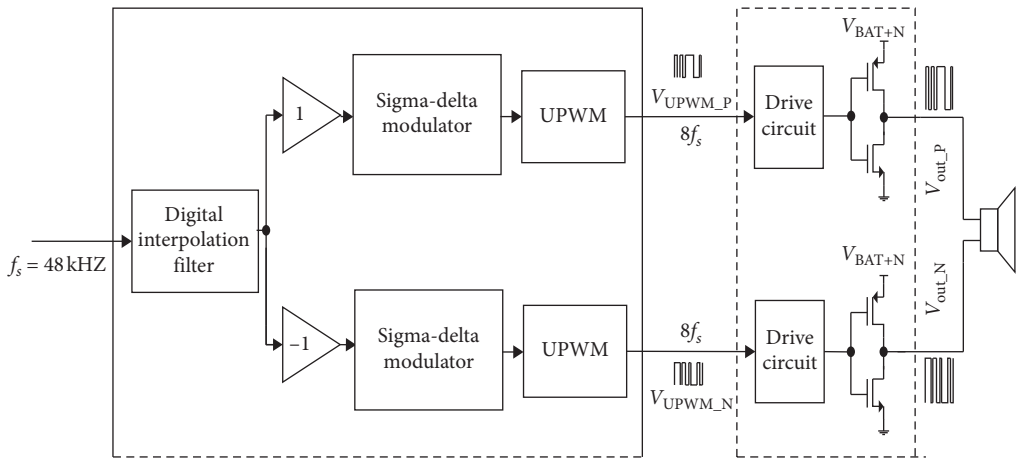


Figure 1: Block diagram of filterless digital class D power amplifier.

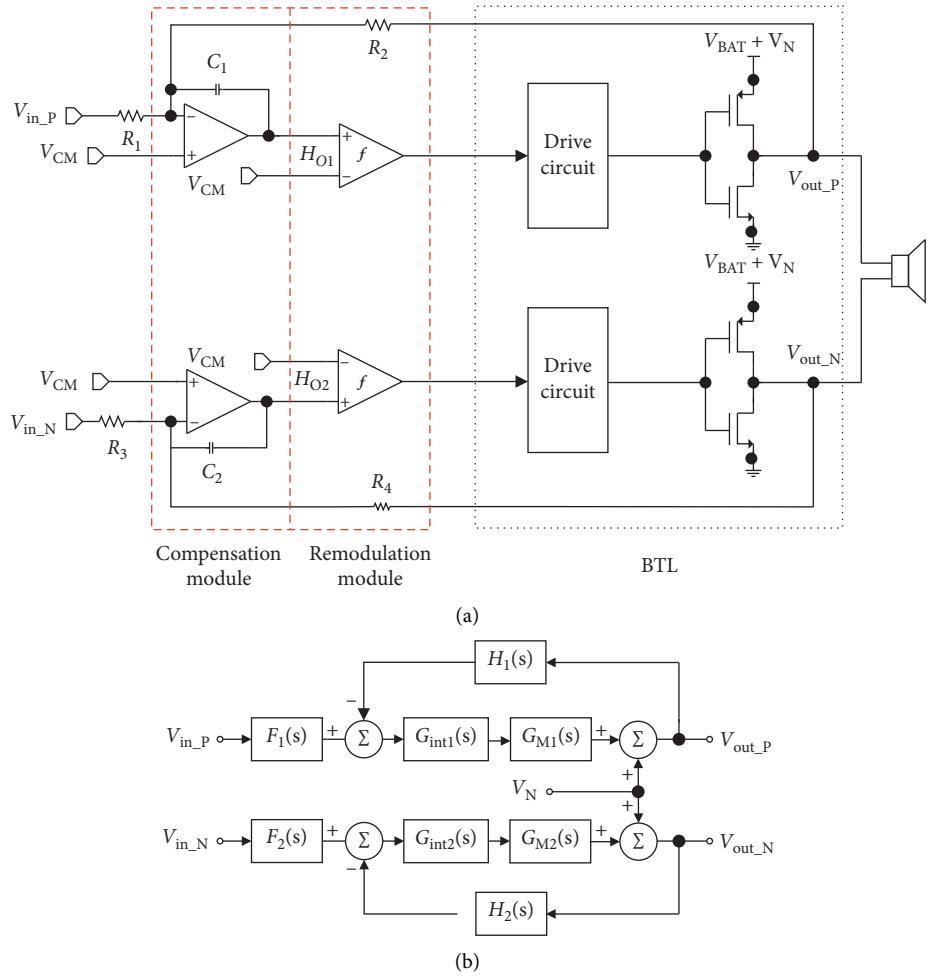


Figure 2: Diagram of FCLNF correction circuit. (a) Block diagram of FCLNF correction. (b) Mathematical model diagram of FCLNF correction.

reference voltage. R_f is feedback resistance. V_{A1} and V_{A2} are the output signals of the integrator. V_{F1} and V_{F2} are the output signals of FFPSNS module. $R_1 - R_5$, R_f, R_i, C_1 and $R_1' - R_5'$, R_f', R_i', C_1' have the same physical meaning and value.

F_1 and F_2 are transfer functions from the input signal of V_{IN_P} and V_{IN_N} to the inverting input of the integrator. $G_{int\,1}$ and $G_{int\,2}$ are the integrator gains of the upper and lower bridges, respectively. H_1 and H_2 are the feedback factors of the upper and lower bridges, respectively. G_{M1} and

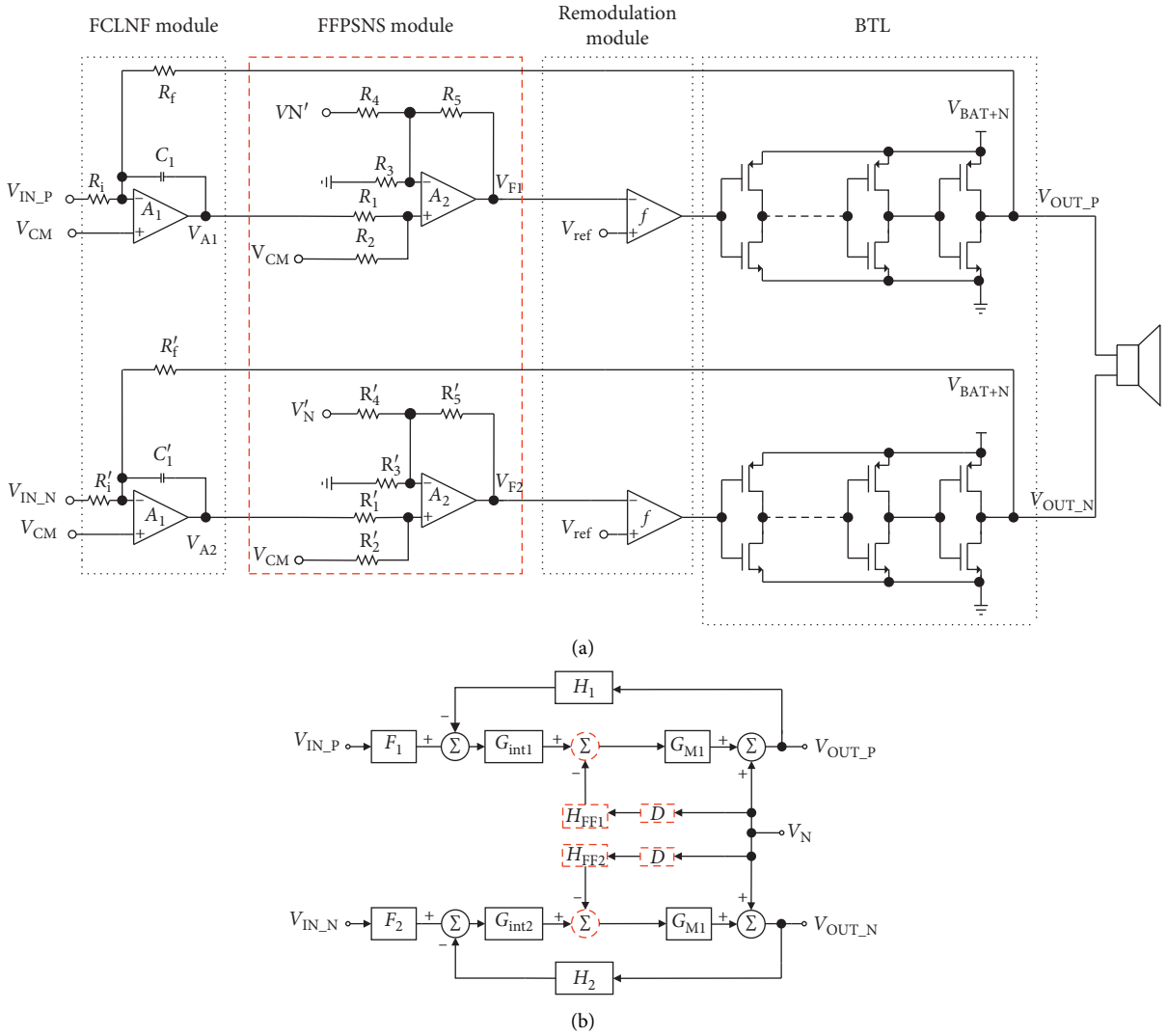


FIGURE 3: Diagram of FCLNF + FFPSNS correction circuit. (a) Block diagram of FCLNF + FFPSNS correction. (b) Mathematical model diagram of FCLNF + FFPSNS correction.

G_{M2} are the combined linear gain of the power level and the remodulation module. D is the power supply noise scaling factor. H_{FF1} and H_{FF2} are the transfer function of feed-forward path.

Under ideal conditions, $H_{FF1} = H_{FF2}$, $G_{M1} = G_{M2}$, and $V_{IN\_P} = V_{IN\_N} = 0$ if the deviation between the two feedback paths is ε.

The deviation between H_{FF} theory and actual situation is β. The mismatch factor between resistances is $r\,(0 < r < 1)$, mismatch factor between capacitors is $r\,(0 < c < 1)$, and ε is equivalent to $r + c$.

The PSRR of power amplifier using the FCLNF + FFPSNS method is

$$\begin{aligned}&\mathrm{PSRR}_{\mathrm{FCLNF+FFPSNS}}\\&\quad = 20\lg\left[LG_2(s)\right] - 20\lg(r+c) - 20\lg(|\beta|).\end{aligned} \tag{3}$$

2.3. Circuit Design of Improved Error Correction Method. The double-input single-output folding cascode two-stage Miller compensation operational amplifier used in this paper is shown in Figure 4.

Based on the proposed power stage error correction method, the circuit implementation of the filter-free digital class D power amplifier BTL power stage with error correction is implemented. It can be seen from Figure 1 that the operational amplifier and voltage comparator are the key modules for this method. The operational amplifier A_1 uses a double-input single-output folding cascode two-stage Miller compensation architecture, as shown in Figure 4. In the figure, V_{DD} is the power supply voltage; V_{b1}, V_{b2}, V_{b3}, and V_{b4} are the bias voltages of the operational amplifier; V_O is the output signal of the operational amplifier; V_{in1} and V_{in2} are the input signals of the operational amplifier; C_m is the Miller compensation capacitor; and resistance is R_m. By

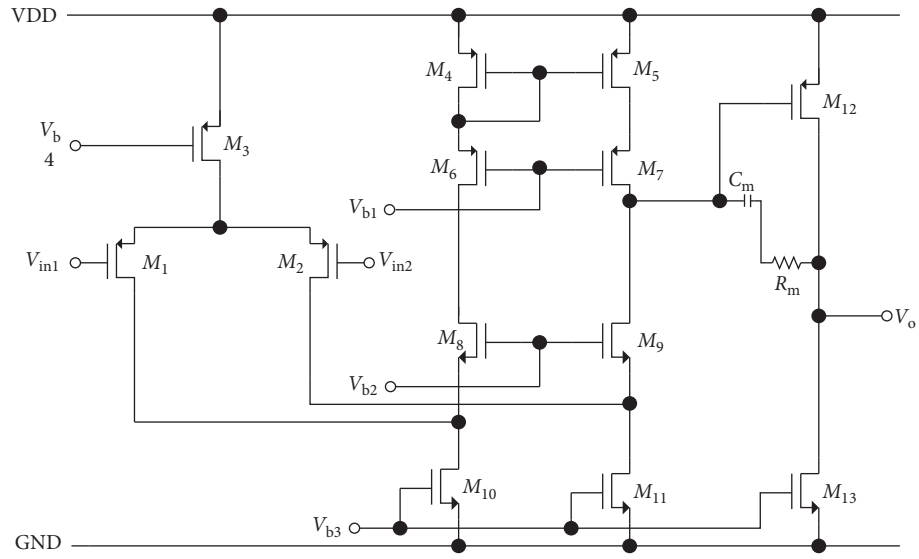


Figure 4: Structure of the operational amplifier.

connecting with C_m in parallel to improve the stability of the operational amplifier, the unity gain bandwidth and phase margin of the operational amplifier A_1 are 13 MHz and 70.3°, respectively, and the open loop gain is 109.9 dB, and the PSRR at 100 Hz is 95 dB. The feedforward operational amplifier A_2 also uses the architecture of Figure 4, with unity gain bandwidth and phase margin of 10 MHz and 68°, respectively.

The voltage comparator circuit structure is relatively simple, and its circuit is shown in Figure 5. In the figure, V_{in1} and V_{in2} are the input signals of the voltage comparator and V_O is the output signal of the voltage comparator. The comparator uses a two-stage amplification architecture. The first stage of amplification uses a differential amplifier circuit structure and a differential circuit to suppress common mode interference. The second-stage amplification adopts a cascode circuit to control the offset voltage well. The output stage adopts the push-pull output to increase the driving capability of the output, while adding an inverter to the output stage can enhance the response time of the comparator. After simulation, the output waveform of the comparator has a rise time of 560 ps and a fall time of 720 ps.

2.4. Simulation and Result Analysis. In order to verify the error correction effect of the above scheme, the simulation experiment route adopted in this paper is shown in Figure 6. The experimental simulation in this paper is based on Matlab and Cadence. Specifically, Matlab is used to generate the two PWM signal sources required by Cadence simulation, and the signal data exported by Matlab is saved as.dat file. In Cadence, V_{source} module is used to store the waveform data in pwl type in Cadence and use it as the signal for Cadence simulation Source. FCLNF control circuit containing BTL power level was built on Cadence platform, and then Cadence Spectre simulator was used for transient simulation of the system. Finally after sampling in Cadence, import the.csv file into Matlab for spectrum analysis and performance calculation.

In this scheme, ASMC 0.35 μm CMOS process is adopted to design and build the circuit on Cadence platform. The simulation conditions are mos_tt process angle and the temperature is 27°C. FCLNF control circuit works under 5 V DC power supply voltage, that is, V_{DD} is 5 V, while the BTL power stage works under 10 V DC power supply voltage. VCM is half of the working voltage of the negative feedback control circuit, and an 8Ω resistor is used as the load RL of the BTL power stage. The input signal of power amplifier is 24-bit sinusoidal single frequency digital signal (frequency is 1 kHz and amplitude is −5 dB). The UPWM waveform data generated by Matlab is input into the simulation circuit through the V_{source} component in Cadence. The transient simulation time is 42.7 ms. The simulation output waveform is sampled and exported at an interval of 40.69 ns (the sampling frequency is 24.6 MHz). The simulation result is shown in Figure 7.

In Figure 7, after FCLNF correction, the power amplifier's PSRR at 200 Hz power source noise frequency was equal to 82.8 dB, and the PSRR increased by 36.02 dB. The PS-IMD of the power amplifier is approximately equal to −80.63 dB, reducing by 34.73 dB. The SNR was increased to 85.7 dB and the SNR performance was much higher than that of the output signal of the uncorrected scheme (58.59 dB), which improved by 27.11 dB. The output THD + N performance of the power amplifier was reduced to 0.0356%, far less than the output THD + N (0.464%) of the power amplifier without correction scheme. Figure shows that the PSRR of FCLNF + FFPSNS corrected power amplifier at 200 Hz power noise frequency is 84.3 dB, which is close to the performance of PSRR in FCLNF.

The influence of power level source noise amplitude on the performance of power amplifier PSRR and PS-IMD is shown in Figures 8 and 9.

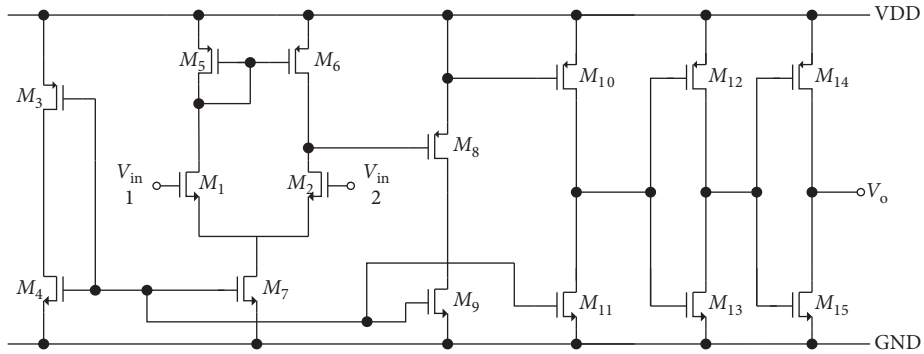


Figure 5: Structure of the voltage comparator.

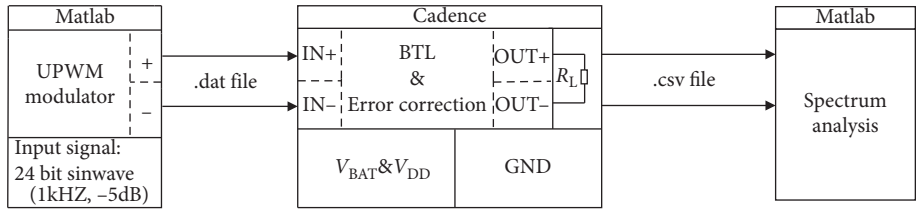


Figure 6: Block diagram of experimental.

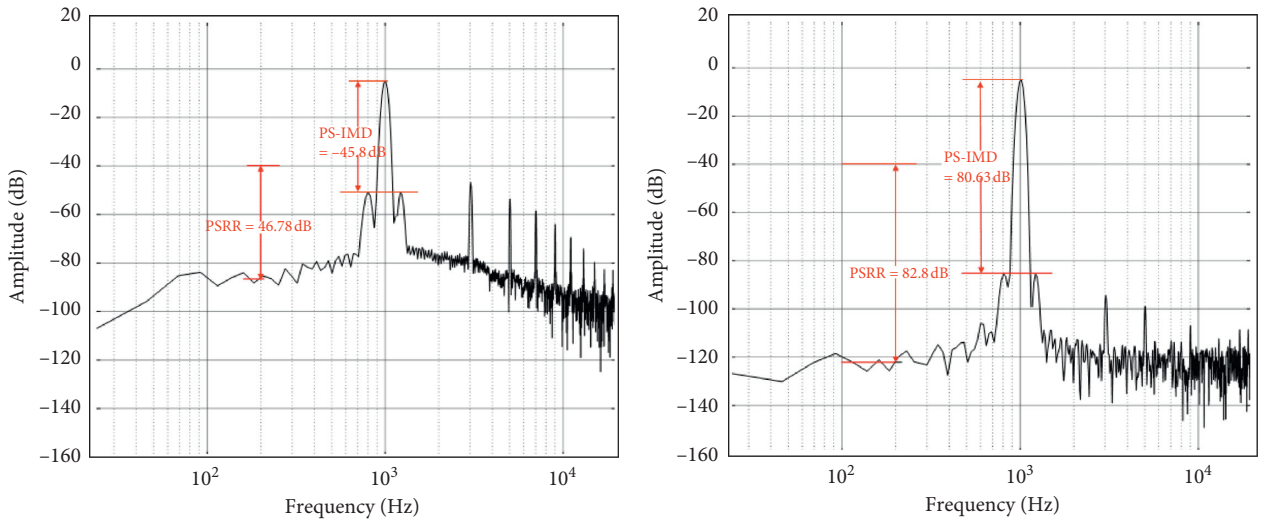
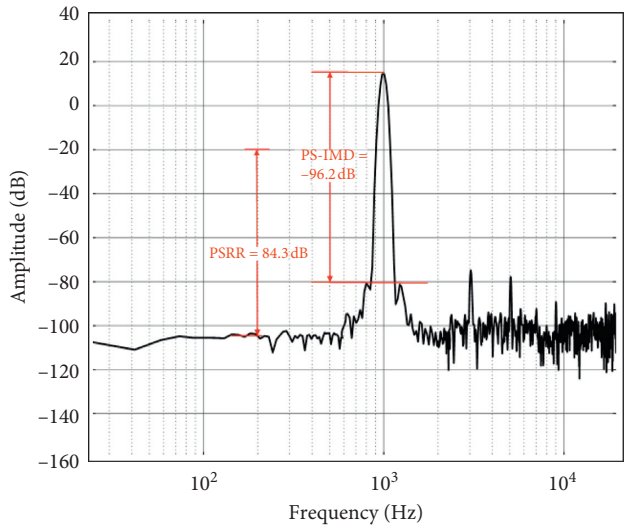


Figure 7: Continued.



(c)

Figure 7: Spectrum diagram of power output signal. (a) Uncorrected. (b) FCLNF corrected. (c) FCLNF + FFPSNS corrected.

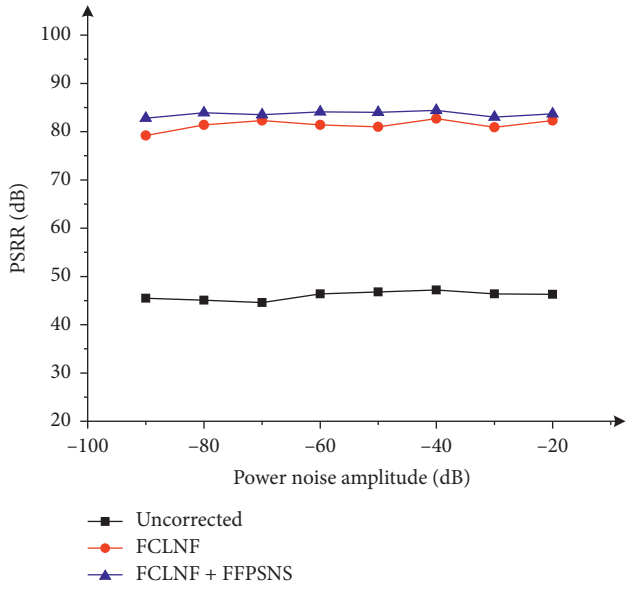


Figure 8: Contrast diagram of PSRR vs. power noise amplitude.

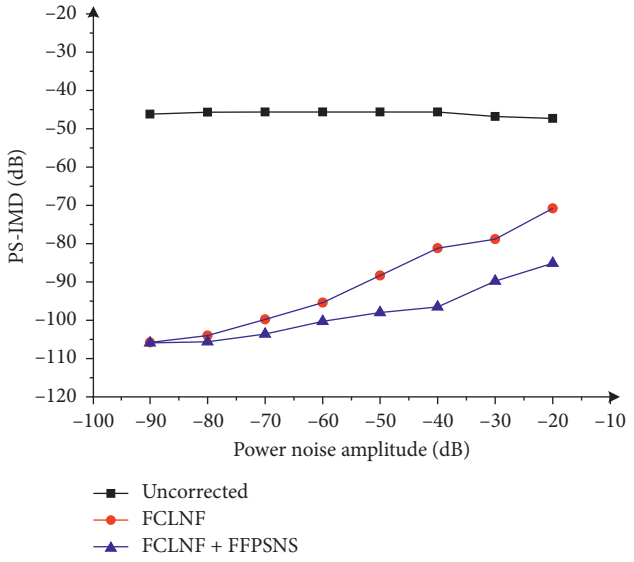


Figure 9: Contrast diagram of PS-IMD vs. power noise amplitude.

The power supply noise amplitude increased from −90 dB to −20 dB. Regardless of whether it is calibrated or not, the power amplifier's PSRR remained basically unchanged with the increase of power supply noise amplitude, and the PSRR corrected by FCLNF and FCLNF + FFPSNS is significantly higher than that by uncorrected. The PSRR characteristics of the two correction schemes are basically the same.

The PS-IMD of the power amplifier does not change with the noise amplitude of the power supply. After FCLNF correction and FCLNF + FFPSNS correction, when the power supply noise amplitude is low (<−70 dB), PS-IMD is basically unchanged. This is because of the background noise in the output signal spectrum. The amplitude of the intermodulation components generated by the power supply noise and the input signal is basically the same as the background noise amplitude. However, as the noise amplitude of power supply continues to increase, PS-IMD of power amplifier starts to increase, and PS-IMD performance declines. This is due to the intermodulation component generated by the power supply noise and the input signal is higher than the background noise when the power supply noise amplitude increases to a certain extent and the PS-IMD characteristics of the two correction schemes are basically equivalent.

Figure 10 compares the changes in the PSRR of the power amplifier with the power supply noise frequency when the three solutions are used. It can be seen from Figure 6 that, whether corrected or not, the PSRR of the

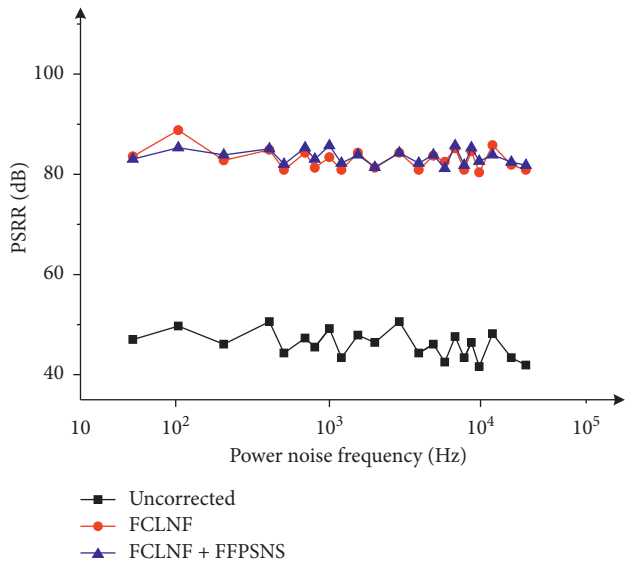


Figure 10: Contrast diagram of PSRR vs. power noise frequency.

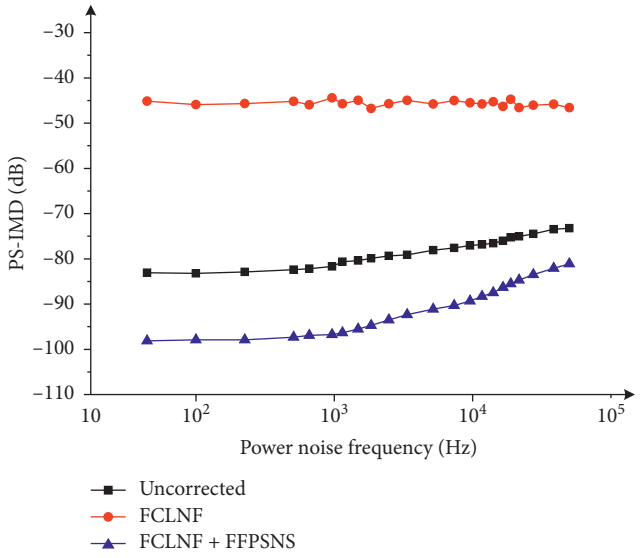


Figure 11: Contrast diagram of PS-IMD vs. power noise frequency.

power amplifier will not change drastically as the noise frequency of the power stage power supply increases, and the PSRR of the power amplifier after correction by LCLNF + FFPSNS and LCLNF is significantly higher than the PSRR of the uncorrected power amplifier.

Figure 11 compares the changes of the PS-IMD of the power amplifier with the power supply noise frequency when the three schemes are used. As the power supply noise frequency increases, the PS-IMD of the uncorrected power amplifier is relatively large and basically remains unchanged, while the PS-IMD of the power amplifier after correction using LCLNF and LCLNF + FFPSNS changes slowly at low frequencies (<1000 Hz). At high frequencies (>1000 Hz), the PS-IMD increases significantly, which is caused by the limited gain bandwidth product of the operational amplifier in the correction circuit, but the PS-IMD is smaller than the uncorrected one. At the same time, the PS-IMD of the power amplifier after LCLNF + FFPSNS correction is significantly smaller than the LCLNF correction scheme.

In summary, the PSRR of the power amplifier after correction by LCLNF and LCLNF + FFPSNS is basically not affected by the amplitude and frequency of the power stage power supply noise. With the change of the noise amplitude and frequency of the power stage power supply, the PS-IMD of the power amplifier after correction by LCLNF + FFPSNS is significantly lower than that after correction by LCLNF. At the same time, the power-stage power supply noise amplitude has a more significant impact on the PS-IMD of the power amplifier than the power supply noise frequency.

3. Concluding Remarks

Under the same experimental conditions, the two correction schemes were compared and analyzed. The results show that compared with the uncorrected power level, the performance of PSRR at 200 Hz source noise frequency is improved by 36.02 dB after FCLNF correction. The PS-IMD component was approximately reduced by 34.73 dB. SNR performance improved by 27.11 dB, and THD + N decreased significantly. Compared with only using FCLNF error correction scheme, the performance of the power amplifier using FCLNF + FFPSNS error correction scheme is comparable at 200 Hz power supply noise frequency. However, the PS-IMD of the power amplifier decreased by 15.57 dB. SNR was further improved by 17 dB, and THD + N dropped further down to 0.02%.

Data Availability

The whole result data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by Key Problems in Science and Technology of Henan Provincial (172102310671), Anyang Science and Technology Project and Anyang Institute of Technology Cultivation Project (YPY2019004), and Key Discipline Project of Henan Education Department (2018 [NO:119]).

References

[1] C. K. Lam, M. T. Tan, S. M. Cox, and K. S. Yeo, "Class-D amplifier power stage with PWM feedback loop," *IEEE Transactions on Power Electronics*, vol. 28, no. 8, pp. 3870–3881, 2013.

[2] Z. Yu, F. Wang, and Y. Fan, "A power supply error correction method for single-ended digital audio class D amplifiers," *International Journal of Electronics*, vol. 103, no. 12, pp. 2110–2124, 2016.

[3] Z. Zhou, Y. Shi, X. Ming, B. Zhang, Z. Li, and Z. Chen, "Modelling and analysis of a high-performance class D audio amplifier using unipolar pulse-width-modulation," *International Journal of Electronics*, vol. 99, no. 2, pp. 163–177, 2012.

[4] H. Zheng, Z. Zhu, and R. Ma, "A. 0.02% THD and 80 dB PSRR filterless class D amplifier with direct lithium battery hookup in mobile application," *Journal of Semiconductors*, vol. 38, no. 7, pp. 56–63, 2017.

[5] H. He, Y. Kang, and J. Yu, "A novel low-power high-efficiency 3-state filterless bang-bang class D amplifier," in *Proceedings of IEEE International Conference On Electronics*, Cairo, Egypt, December 2015.

[6] H. Zheng, Q. Wang, and Y. Liu, "A low THD + N and high PSRR class-D power amplifier in mobile application," *Guti Dianzixue Yanjiu Yu Jinzhan/Research and Progress of Solid State Electronics*, vol. 37, no. 2, pp. 134–139, 2017.

[7] A. Donida, P. Malcovatip, A. Nagari, R. Cellier, and A. Baschirotto, "A 40-nm CMOS, 1.1-V, 101-dB DR, 1.7-mW continuous-time ΣΔ ADC for a digital closed-loop class-D amplifier," in *Proceedings of 2013 IEEE 20th International Conference On Electronics, Circuits, and Systems (ICECS)*, Abu Dhabi, UAE, December 2013.

[8] F. Mostert and D. Schinkel, W. Groothedde, A 5×80 W 0. 004% THD+N automotive multiphase class-D audio amplifier with integrated low-latency ΔΣ ADCs for digitized feedback after the output filter," in *Proceedings of IEEE International Solid-State Circuits C.Solid-state Circuits Conference*, San Francisco, CA, USA, February 2017.

[9] X. Chen, H. Qu, Z. Yu, C. Zhang, and E. Zhang, "A filterless digital audio class-D amplifier based on grow-left double-edge pulse width modulation," in *Proceedings of 2nd International Conference on Integrated Circuits and Microsystems*, Nanjing, China, November 2017.

[10] R. Cellier, G. Pillonnet, and N. Abouchi, "Analysis and design of an analog control loop for digital input class D amplifiers," in *Proceedings of IEEE 18th International Conference on Electronic,Circuits and Systems*, Beirut, Lebanon, December 2011.

[11] D. Lee, J. Noh, and J. Lee, "A class-D amplifier for a digital hearing aid with 0.015% total harmonic distortion plus noise," *ETRI Journal*, vol. 35, no. 5, pp. 819–826, 2013.

[12] J. Noh, D. Lee, J.-G. Jo, and C. Yoo, "A class-D amplifier with pulse code modulated (PCM) digital input for digital hearing aid," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 2, pp. 465–472, 2013.

[13] K. Philips and J. Van Den, "Power DAC: a single-chip audio DAC with a 70%-efficient power stage in 0.5 μm CMOS," in *Proceedings of Digest Of Technical Papers Of the 1999 IEEE International Solid-State Circuits Conference*, San Francisco, CA, USA, February 1999.

[14] S. Ramaswamy, J. Krishnan, B. Forejt, J. Joy, M. Burns, and G. Burra, "A high-performance digital-input class-D amplifier with direct battery connection in a 90 nm digital CMOS process," in *Proceedings of Digest Of Technical Papers Of the 2008 IEEE International Solid-State Circuits Conference*, San Francisco, CA, USA, February 2008.

[15] S.-H. Chien, Y.-W. Chen, and T.-H. Kuo, "A low quiescent current, low THD + N class-D audio amplifier with area-efficient PWM-residual-aliasing reduction," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 12, pp. 3377–3385, 2018.

[16] C. W. Lin and B. S. Hsieh, "Dynamic power efficiency improvement for PWM class-D amplifier," *IEICE Electronics Express*, vol. 10, no. 6, 2013.

[17] A. I. Colli-Menchi, J. Torres, and E. Sanchez-Sinencio, "A feed-forward power-supply noise cancellation technique for single-ended class-D audio amplifiers," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 3, pp. 718–728, 2014.

[18] V. Adrian, K. Cui, B. H. Geww, and J. S. Chang, "A Randomized Modulation scheme for filterless digital class D audio amplifiers," in *Proceedings of IEEE International Symposium on Circuits and Systems*, Melbourne VIC, Australia, June 2014.

[19] F. Guanzurolif, R. Bassoli, C. Crippa, D. Devecchi, and G. Nicollini, "A 1 W 104 dB SNR filter-less fully-digital open-loop class D audio amplifier with EMI reduction," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 3, pp. 686–698, 2012.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 8951353, 11 pages
https://doi.org/10.1155/2020/8951353

Research Article

Optimisation Analysis of Structural Parameters of an Annular Slot Ejector Based on the Coanda Effect

Fengliang Wu[1] **and Zhisheng Li**[1,2]

[1]*College of Safety Science and Engineering, Xi'an University of Science and Technology, Xi'an 710054, China*
[2]*Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu 610031, China*

Correspondence should be addressed to Fengliang Wu; wufl@xust.edu.cn

Received 13 March 2020; Revised 2 July 2020; Accepted 16 July 2020; Published 11 August 2020

Guest Editor: Sanghyuk Lee

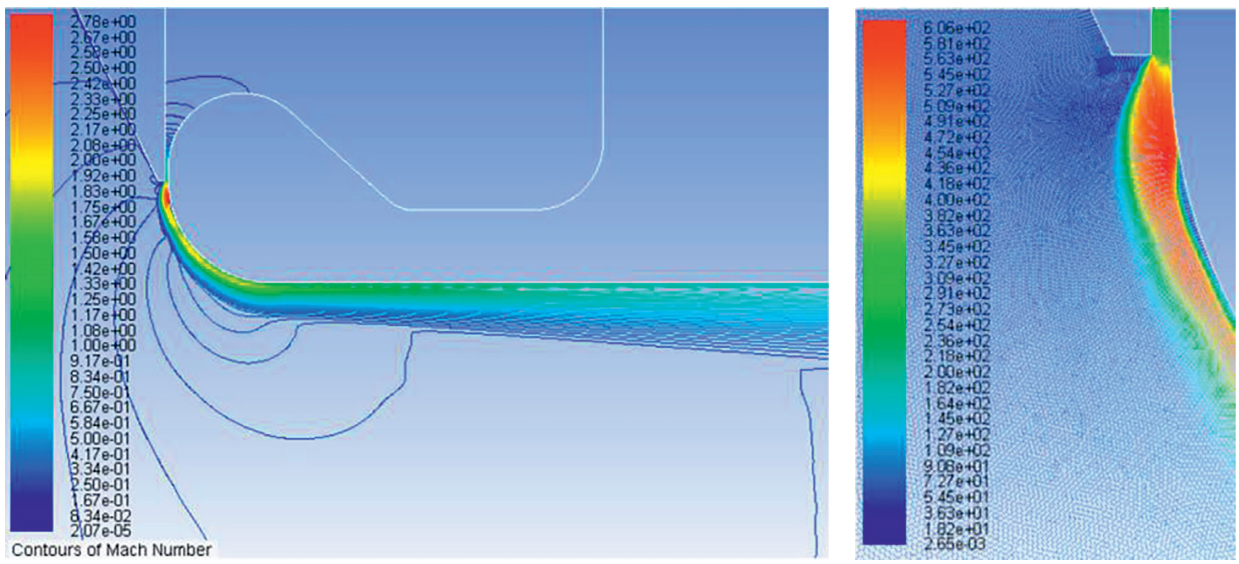
To investigate the effect of structural parameters on the performance of an annular slot ejector, a series of numerical simulations were conducted with single-factor analysis. Moreover, a multifactor grey relational analysis was applied to examine the correlations between the structural parameters and entrainment ratio. Subsequently, the optimised model was verified by comparing the simulated results with experimental data. Results show that the performance of the optimised ejector model was improved. The RNG k-ε turbulent transport mode can simulate the internal field characteristics of an annular slot ejector, and the corresponding simulated results, as verified by experiment, satisfy engineering requirements. In addition, a quantitative correlation between structural parameters and entrainment ratio was obtained as follows: e (nozzle clearance) > θ (diffusing chamber angle) > H_L (mixing chamber length) > d (throat diameter) > K_L (diffusing chamber length). This work may provide a certain guiding significance for the design and application of annular slot ejectors.

1. Introduction

An annular slot ejector is a mechanical device which is different from a central jet [1–4]. A large volume of low-pressure fluid is dragged by high-pressure fluid due to the special structure of the annular slot ejector [5]. The working principle of an annular slot ejector is known as the Coanda effect [6, 7]. The phenomenon is described as a fluid flowing not along the original direction but along a curved surface. Subsequently, the mainstream flow passes through the nozzle clearance, it deflects and flows along the wall surface, and then induces a large amount of air from its surroundings [8]. Gregory-Smith and Gilchrist described three main features of the Coanda effect including the nonviscous effect, viscous effect, and a stronger entrainment capability [9]. Besides, another major feature of the Coanda effect is the Coanda flare, also known as a shock wave when the high-pressure gas reaches a certain speed (Figure 1).

The Coanda effect has received extensive attention in aviation [11], medicine [12], acoustics [8], robotics [13], and leafless fans [14]. Dong-Won et al. utilised a Coanda nozzle to cause jet deflection and wall shearing [15]. Their work showed that when the pressure increases, the potential splash zone tends to move downstream. In recent years, the introduction of the Coanda effect has promoted research into, and development of, annular ejectors [10, 16–25].

Ameri proposed a semiempirical formula for the section velocity based on a new ejector model by conducting a set of experiments with an LDV (laser Doppler velocimeter) [26]. In his study, it was assumed that the flow between the pressure inlet and the nozzle must satisfy the isentropic condition; however, the isentropic condition cannot be applied at high pressures. Guerriero investigated the influence of structural parameters on the ejector performance by conducting a set of experiments. They suggested that nozzle clearance has a significant influence on the ejector performance [27] and the corresponding results under the



(a) (b)

Figure 1: A Coanda flare (a) contours of Mach number at the ejector throat and (b) contours of velocity [10].

same model and working conditions were verified by Kim et al. employing CFD methods. Moreover, they pointed out that the stagnation pressure ratio is another important parameter affecting ejector performance [28, 29]. Alexandru et al. developed a semiempirical formula for two-dimensional Coanda flow while the curvature of the tangential momentum equation was neglected [16]. Sierra del Rio et al. designed a two-ejector model with varying nozzle clearance (0.3 mm and 0.8 mm) to investigate the effect of nozzle clearance on the flow velocity by means of CFD methods [30]; their results showed that the velocity increases with the increase of nozzle clearance. Similar trends were obtained by Lowry et al. based on a new Coanda ejector [31]. Jain et al. developed a new ejector model with two nozzles to investigate the relationship between the structural parameters (nozzle clearance and throat diameter) and the flow characteristic. They also suggested that the nozzle clearance has a significant influence on the flow velocity, and the larger the throat diameter is, the faster the mixing layer is developed [10].

Although many studies have been conducted, previous studies focused on single-factor analysis and few studies, where all structural parameters were varied simultaneously, have been undertaken. Moreover, different optimal sizes of the model were obtained due to the differences in the structure being modelled. Besides, in the early literature only the nozzle clearance is deemed to have been an important parameter that influences ejector performance, but correlations between other geometric parameters and the entrainment ratio were ignored. In the present work, a set of numerical simulations were conducted using single-factor analysis and multifactor analysis to investigate annular slot ejector performance including five structural parameters (namely, mixing chamber length, diffusion chamber length, diffusion chamber angle, throat diameter, and nozzle clearance). Moreover, the optimised model was verified and analysed by conducting a series of experiments to compare with the results of numerical simulation.

2. Numerical Model

2.1. Turbulent Model. The flow of gas inside the annular ejector contains turbulence, and the velocity gradient of the mainstream gas at the throat of the ejector changes significantly which may generate more vortices. Amel et al. analysed the variation of the flow characteristic for both single-phase flow and two-phase flow mode inside the ejector based on a supersonic ejector using CFD methods. They suggested that the RNG (renormalisation group) k-ε model be applied to simulate a supersonic ejector [32]. Victor and Steven also verified the RNG k-ε turbulence model as being able to simulate the flow characteristic based on a Coanda ejector. Therefore, in the present work, the RNG k-ε double equation model [33] was applied to the annular slot ejector as follows.

k equation:

$$\frac{\partial(\rho k)}{\partial t} + \frac{\partial(\rho k u_i)}{\partial x_i} = \frac{\partial\left(\alpha_k \mu_{\text{eff}}\left(\partial k/\partial x_j\right)\right)}{\partial x_j} + G_k - \rho\varepsilon, \quad (1)$$

where ρ is the fluid density, k is the turbulent kinetic energy, t is the time, α_k is the turbulent Prandtl number of k, μ_i and μ_{eff} are the viscosity coefficients, x_i and x_j are coordinate vectors, G_k represents the turbulent kinetic energy generated by the laminar velocity gradient, and ε is the dissipation rate.

ε equation:

$$\frac{\partial(\rho\varepsilon)}{\partial t}+\frac{\partial(\rho\varepsilon u_i)}{\partial x_i}=\frac{\partial\left(\alpha_\varepsilon\mu_{eff}\partial\varepsilon/\partial x_j\right)}{\partial x_j}+C_{1\varepsilon}*\frac{\varepsilon}{k}G_k-C_{2\varepsilon}\rho\frac{\varepsilon^2}{k},$$
$$C_{1\varepsilon}*=C_{1\varepsilon}-\frac{\eta\left(1-(\eta/\eta_0)\right)}{1+\beta\eta^3},\quad(2)$$
$$\eta=\sqrt{2E_{ij}}\,\frac{k}{\varepsilon},$$
$$E_{ij}=\frac{1}{2}\left(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right),$$

where $C_{1\varepsilon}$ and $C_{2\varepsilon}$ are constants, indicating the effect of buoyancy on the dissipation rate, and α_ε is the turbulent Prandtl number of ε. The values of the coefficients in the formula are as follows:

$$\begin{aligned}\alpha_k&=\alpha_\varepsilon=1.39,\\ \beta&=0.012,\\ \eta_0&=4.377,\\ C_\mu&=0.0845,\\ C_{1\varepsilon}&=1.42,\\ C_{2\varepsilon}&=1.68.\end{aligned}\quad(3)$$

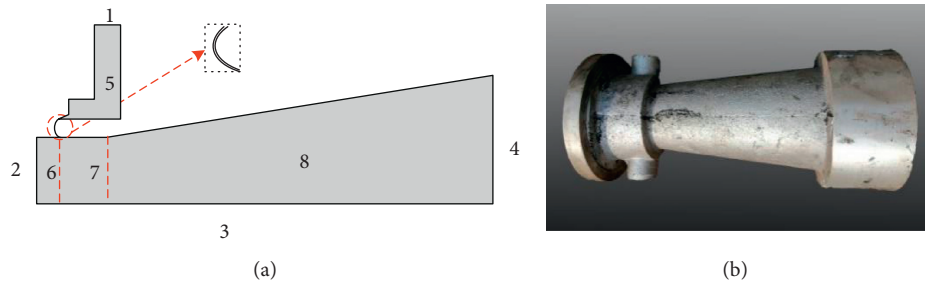
2.2. Modelling and Meshing for an Annular Slot Ejector. An annular slot ejector is usually an axisymmetric structure, which includes eight parts (Figure 2). Moreover, the 3D model can be simplified to a 2D model according to the flow characteristic. The mainstream gas flows at high speed along the wall after passing through the nozzle clearance. Meanwhile, a secondary flow will be induced into the mixing chamber. Thereafter, the mixing gas flows outward through the diffusion chamber; however, the velocity near the wall is different from that near the centreline due to the Coanda effect, which causes a physical gradient inside the ejector; therefore, the corresponding mesh mapped using grid-generating software (ICEM) is encrypted, to guarantee the accuracy of the numerical simulation results. All elements are quadrilaterals with about 200,000 in each mesh. The mesh size near the wall boundary is 0.1 mm and it gradually increases to 1 mm (Figure 3(a)). The independence of the grid has been analysed by obtaining a resulting mesh size of 0.75 mm. Three meshes (a coarse grid, medium grid, and fine grid) were used to evaluate grid size-independence; the three grid sizes tested were 0.75 mm, 1 mm, and 1.25 mm, respectively. Figure 3(b) shows that secondary mass flow increases first and then decreases as the primary pressure increases from 0.3 MPa to 0.7 MPa; the difference in secondary mass flow for the fine and medium grid is observed to be slight, but there was some improvement compared with the case modelled using a grid size of 1 mm. Therefore, the simulation model with a 1-mm grid was selected to reduce the computational time and ensure accuracy in subsequent simulations.

2.3. Boundary Conditions. The operating fluid is a compressed gas, and the ejector fluid is from the surrounding air. In the present work, both fluids are treated as ideal gases, as carried out by other scholars [29, 34]. An implicit solution method and a hybrid initialisation method were applied to the flow calculation of compressible gases. Both the high-pressure inlet and the low-pressure inlet of the annular slot ejector were set as a pressure inlet boundary condition, and the mixed fluid outlet is configured as a pressure outlet. Other boundaries are wall surfaces, and the insulation between the wall and the environment is assumed to be adiabatic, nonpermeable, and nonslip boundaries [35] (Table 1).

3. Analysis of Factors Affecting Ejector Performance

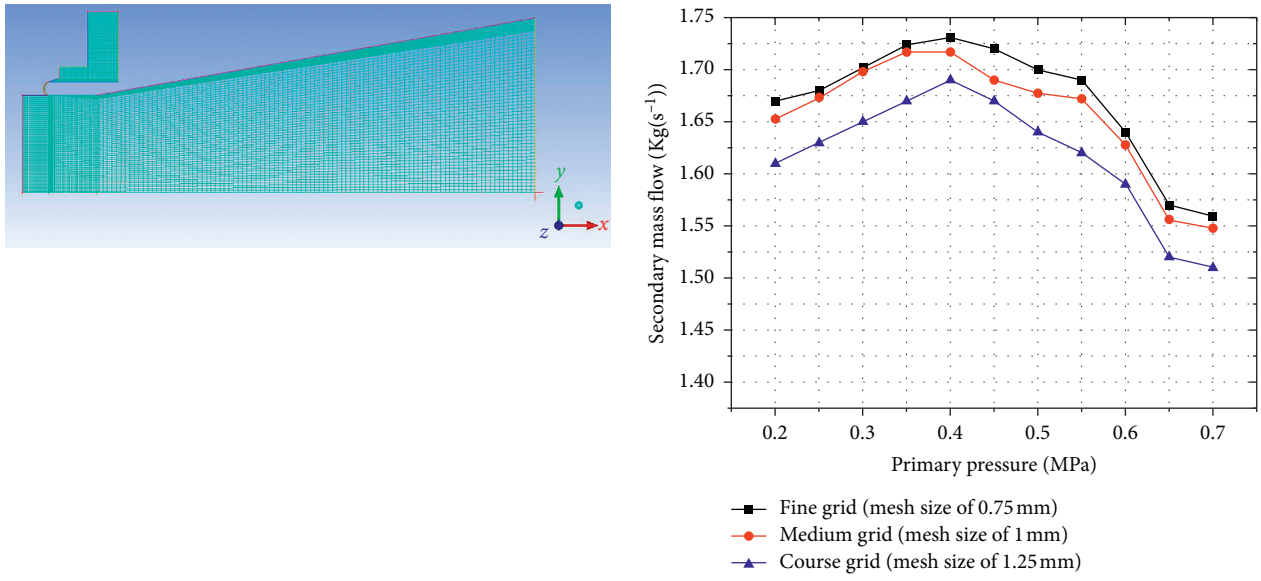
3.1. Influence of Mixing Chamber Length. Figure 4 shows that the entrainment ratio raises first and then decreases as the length of the mixing chamber increases from 10 mm to 100 mm, and the corresponding maximum entrainment ratio is obtained when the mixing chamber length is 40 mm. There is a certain buffer developed during the mixing process of primary and secondary flows. More importantly, when the mixing chamber is short, mainstream gas does not transmit momentum to the ejector fluid well, resulting in the insufficient mixing of the two fluids; moreover, when the length of the mixing chamber is too large, the impact loss of the fluid increases. Figure 5 indicates that the distribution of the flow velocity under different primary pressures include 0.3 MPa, 0.4 MPa, 0.5 MPa, 0.6 MPa, and 0.7 MPa, respectively. It is seen that the flow velocity away from the wall shows a certain gradient, and there is good turbulent mixing at the interface between the primary gas and the secondary gas [36]. In addition, for different primary pressures, the flow velocity near the wall is always much larger than that near the central axis, which corresponds to the main features of the Coanda effect [37].

3.2. Influence of Diffusion Chamber Length. Figure 6 shows the relationship between diffusion chamber length and entrainment ratio when the mainstream pressure is 0.3 MPa; the entrainment ratio continuously increases as the diffusion chamber length increases from 120 mm to 400 mm. This can be explained by their different physical flow processes; the flow of the mixed fluid in the diffusion chamber is a process in which static pressure recovery quickly, and the rapid decline of velocity and the diffusion area increases with a certain gradient during the process (Figure 7). The high-pressure jet layer is wider when the diffusion chamber length is 400 mm than at other sizes, suggesting good turbulent mixing. Moreover, the diffusion area contributes to the amount of induced air during the process, and it may be necessary to use more compressed gas to complete turbulent mixing with the induced fluid.



(a) (b)

Figure 2: The annular slot ejector model. (a) Schematic diagram of annular slot ejector and (b) photograph of the physical model. 1. High-pressure inlet. 2. Secondary inlet. 3. Symmetry axis. 4. Outlet. 5. Storage room. 6. Suction. 7. Mixing chamber. 8. Diffusion chamber.



(a) (b)

Figure 3: (a) Meshing for annular slot ejector and (b) secondary mass flow with different grid densities.

Table 1: Boundary conditions.

Location	Pressure	Turbulence intensity	Hydraulic diameter (m)	Total temperature (K)
Primary inlet	$3–7 \times 10^5$ Pa	1	0.025	
Secondary inlet	1×10^5 Pa	1	0.08	300
Outlet	1×10^5 Pa	5	0.16	

3.3. Influence of the Diffusion Chamber Angle. Figure 8 indicates that the entrainment ratio decreases as the angle of the diffusing chamber increases from 6° to 16°; the high-pressure jet is not separated from the wall surface as the angle increases due to the Coanda effect. Nevertheless, the high-pressure jet layer becomes thinner and the corresponding mixed boundary layer moves towards the wall because the adsorption capacity decreased. Moreover, due to the increase of the diffusion area, the amount of induced air rises (in relative terms), and the traction force on the induced fluid is reduced, thereby causing the entrainment ratio to decrease at the same primary pressure. Figure 9 indicates the relationship between the velocity and the angle of the diffusing chamber; the flow velocity decreases upon widening of the diffusion chamber angle. Besides, when the diffusion angle is large enough, a local counterflow zone appears near the inlet section of the diffusion chamber and the surrounding countercurrent zone still contains higher-energy fluid, which may cause energy loss. Therefore, the attenuation of the jet flow may be greater, further resulting in a lower entrainment ratio.

3.4. Influence of Throat Diameter. Figure 10 shows that the mass flow both primary and secondary gas increases with the increase of the throat diameter, and the growth rate of the

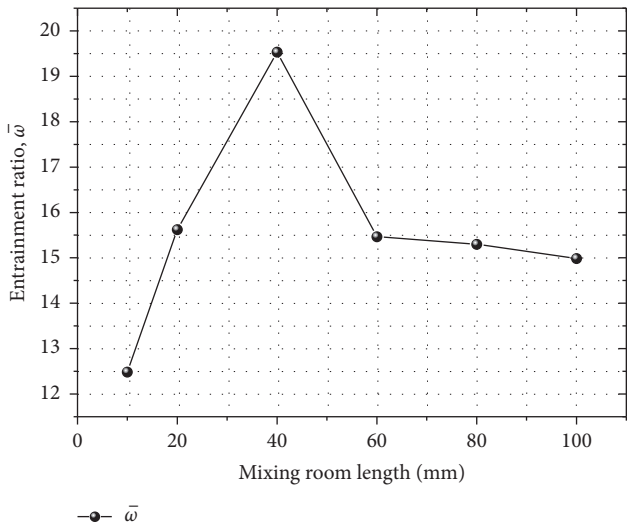


Figure 4: Mixing chamber length v entrainment ratio.

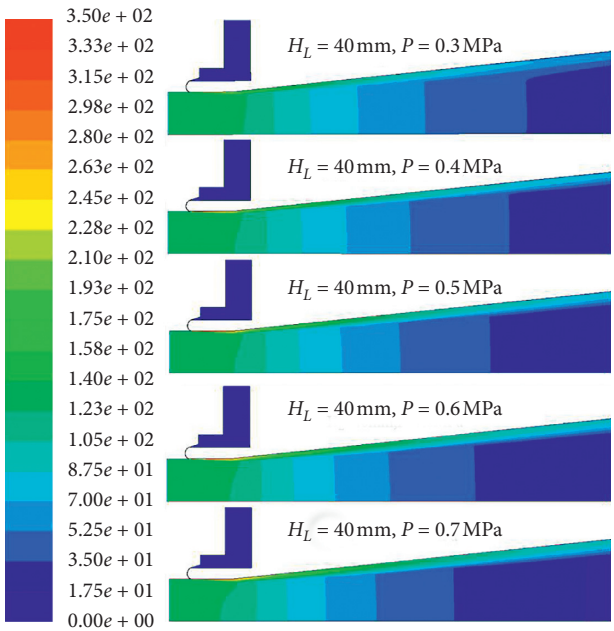


Figure 5: Velocity contours inside the ejector at different primary pressures.

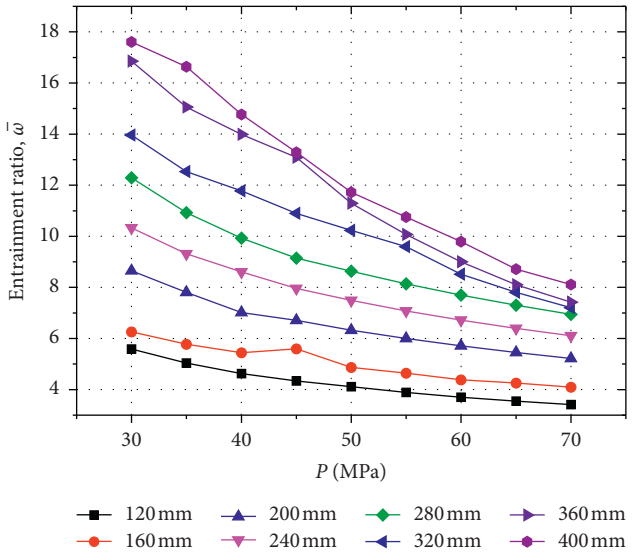


Figure 6: Diffusion chamber length v entrainment ratio.

secondary mass flow rate is significantly higher than that of the primary flow. This can be explained by considering that the aspect ratio both primary and secondary flows increase as throat diameter rises, inducing more air, and resulting in a greater mass flow rate. Figure 11 shows that the entrainment ratio decreases significantly as the primary pressure increases from 0.3 MPa to 0.7 MPa, which can be explained by the fact that the larger primary pressure is, the more mass flow is available for primary flow, resulting in a lower entrainment ratio; however, the entrainment ratio shows a complicated trend in behaviour upon variation of the throat diameter, which may lie in the complex physical processes including turbulent flow mixing in both primary and secondary flows. In addition, as the aspect ratio increases, the more shearing force will be needed to drag the secondary flow in. Thereafter, the primary and secondary flows enter the mixing chamber to achieve static pressure matching, completing the mixing process of the two flows inside the mixing chamber, reducing the impact loss between the two fluids. At the same time, as the throat diameter increases from 60 mm to 160 mm, more air (by volume) is induced into the ejector. In previous studies, Kim et al. analysed the influence of throat diameter on the ejector coefficient based on a Coanda ejector [28]. Their results showed that the velocity decreases when the throat diameter increases from 35 mm to 50 mm; however, the higher velocity does not mean a larger mass flow due to the decline in diameter resulting in a lower flow. In the present work, the mixing entrainment ratio is obtained when the throat diameter is 160 mm under different primary pressures.

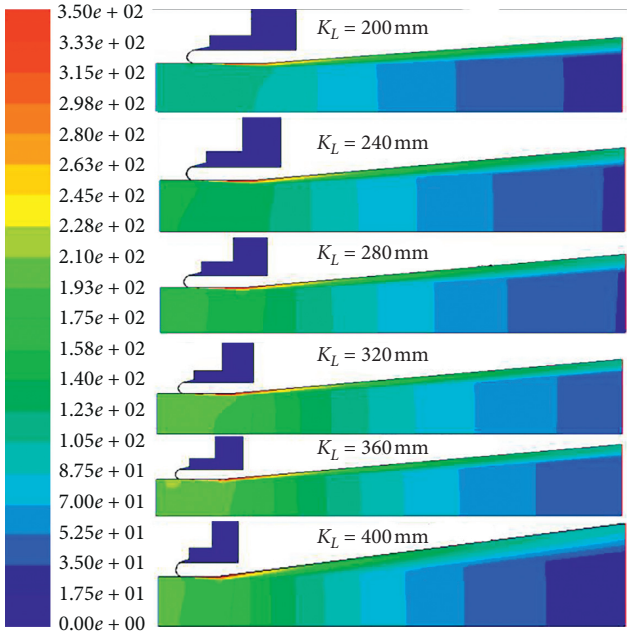


Figure 7: Velocity contours inside the ejector for different diffusion chamber lengths.

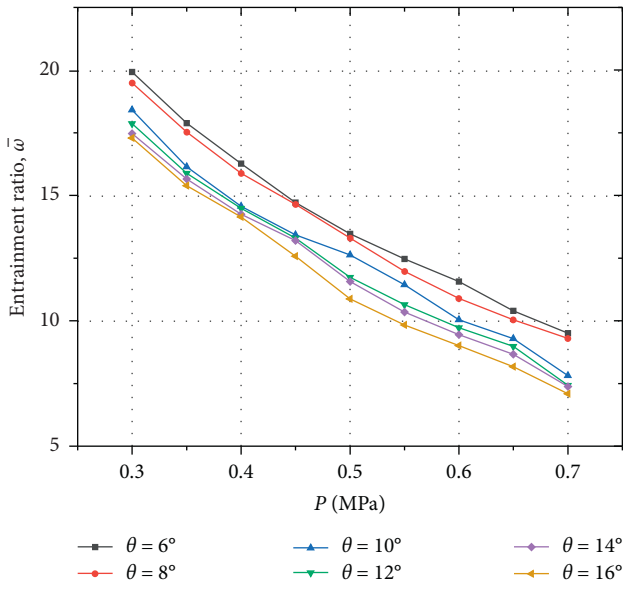


Figure 8: Diffusion chamber angle v entrainment ratio.

3.5. Influence of Nozzle Clearance. Figure 12 shows that the entrainment ratio gradually decreases as the primary pressure increases from 0.3 MPa to 0.7 MPa. When the mainstream pressure is fixed, the entrainment ratio decreases continuously as the nozzle clearance increases from 0.1 mm to 0.5 mm; a similar trend was obtained elsewhere [30]. It also can be seen that when the primary pressure is 0.3 MPa, the entrainment ratio increases by 42.3% as the nozzle clearance increases from 0.1 mm to 0.15 mm, and the entrainment ratio rises by 85.9% when the nozzle clearance increases from 0.1 mm to 0.5 mm. Therefore, the nozzle

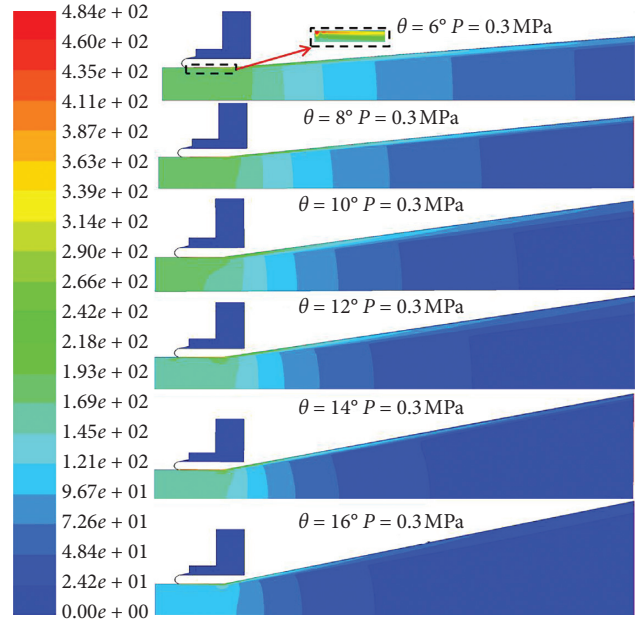


Figure 9: Velocity contours inside the ejector for different angles.

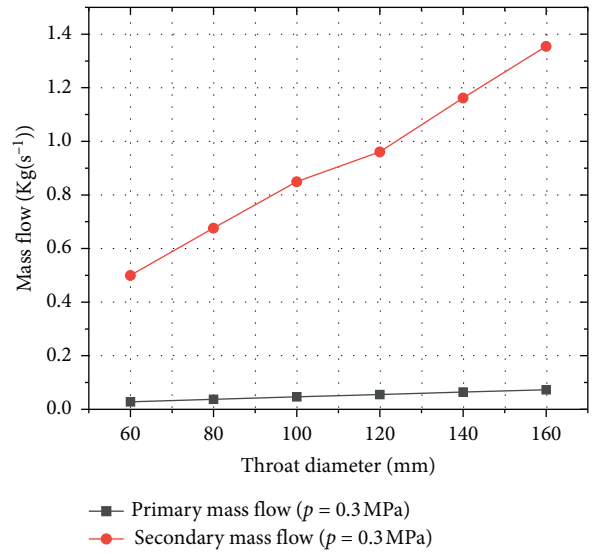


Figure 10: Throat diameter of diffusion chamber v mass flow.

clearance affects the performance of the annular slot ejector to a significant extent.

3.6. Multifactor Analysis of Ejector Performance Employing the Grey Relational Analysis Method. The above analysis shows the influence of the geometry on the entrainment ratio when one parameter changes but other parameters are fixed. Nozzle clearance is the more important parameter; nevertheless, the importance of the other parameters in terms of their influence on ejector performance is unclear; therefore, it is necessary to analyse ejector performance when the five geometric parameters are varied simultaneously. The grey relational analysis method involves the analysis of an

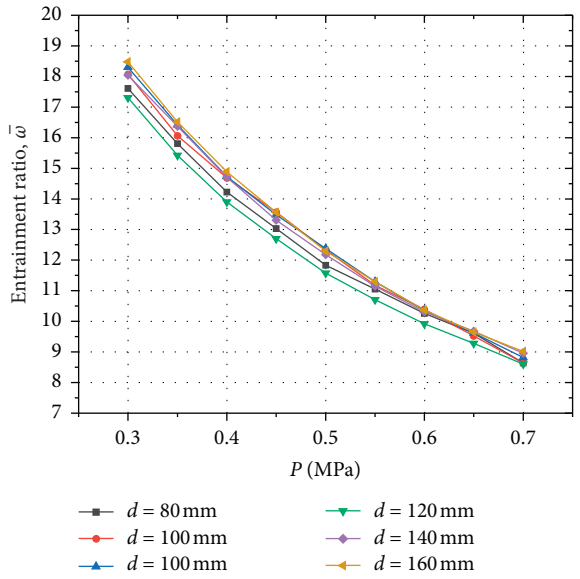


Figure 11: Throat diameter v entrainment ratio.

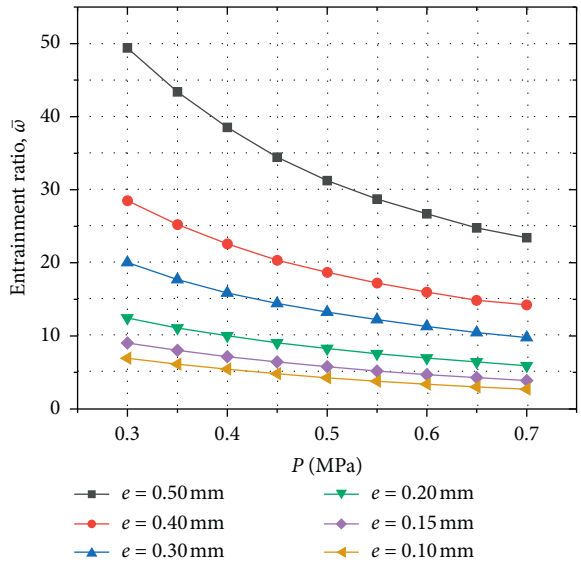


Figure 12: Nozzle clearance v entrainment ratio.

abstract system or phenomenon, which makes up for the deficiencies in systematic analysis using mathematical statistical methods. It is also applicable to any number of samples and works irrespective of a parametric distribution being known a priori. Ju-Long [38] proposed a theoretical model for analysing the correlation between samples:

$$\gamma_{0i}(k) = \frac{\min_i \min_k \left|(X_0(k) - X_i(k))\right| + \xi \max_i \max_k \left|(X_0(k) - X_i(k))\right|}{\left|(X_0(k) - X_i(k))\right| + \xi \max_i \max_k \left|(X_0(k) - X_i(k))\right|}, \tag{4}$$

where $X_i(k)$ is observed data on sequence k, $X_0(k)$ reflects the behaviour of the system characteristics, k can also represent the time serial number and the index number, $\gamma_{0i}(k)$ is the grey correlation of X_i and X_0, and ξ is the resolution factor.

In the present work, the grey relational analysis method was applied to study ejector performance without fixed geometric parameters. A set of numerical simulations were conducted to investigate the ejector performance under constant pressure (Table 2). Taking the first set of parameters from Table 2 as a reference sequence, the six sets of parameters were normalised and initialised (Table 3). Finally, the sensitivity of each structure parameter to the entrainment ratio was obtained (Table 4) which can be sorted as follows: nozzle clearance (e) > diffusion chamber angle (θ) > mixing chamber length (H_L) > throat diameter (d) > diffusion chamber length (K_L).

4. Experimental Optimisation Model of an Annular Slot Ejector

4.1. Experiments. Although a series of numerical calculations were conducted to optimise the ejector structure, the optimised model still needs further experimental verification. The optimised model and dimensions were obtained based on the aforementioned simulated results, and the comparison between the original model and the optimised model is shown in Figure 13 and Table 5 (where D_{main} is the diameter of the primary inlet, D_{sec} is the diameter of the secondary inlet, and D_{out} is the diameter of the outlet).

In this experiment, the logarithmic linear measurement method [39] was used to measure the average dynamic pressure in the ejector inlet section, and then the average velocity and mass flow of the ejector inlet section were obtained. The experiments were conducted in a relatively closed indoor environment (Figure 14). An air compressor was adopted to provide mainstream gas flow under different pressures and to act as a buffer airflow. The gas collection tank was used to supply a continuous flow of compressed gas. Moreover, the compressed air was dried before entering the ejector. The pressure transmitter measured the mainstream pressure, and the gas turbine flowmeter was used to measure the primary mass flow and a Pitot tube was utilised to measure the dynamic pressure at each measurement point on the cross-section of the ejector exit and to obtain the average flow velocity and the total mass flow. These three parameters were transmitted through a sensor to a computer program for simultaneous counting with a pulse counter.

4.2. Analysis of Experimental Results. Nine groups of experimental tests present a set of parameters for ejector performance at different primary pressures (Table 6); the primary mass flow (G_1) rises continuously while the secondary mass flow (G_2) tends to first increase, then decrease as the primary pressure increases from 0.3 MPa to 0.7 MPa. Moreover, the dynamic pressure and velocity in the ejector exit show the same trend as the secondary flow; this can be explained by considering that the excessive primary pressure developed poor turbulent mixing, forming a local vortex, thereby leading to a decrease in the secondary mass flow rate. The entrainment ratio ($\bar{\omega}$) decreases continuously with increasing primary pressure (Figure 15). Moreover, there is good consistency found by comparing simulated results and

Table 2: Preliminary calculated entrainment ratios.

H_L	K_L	θ	d	e	Entrainment ratio
10	200	6	80	0.1	37.431
20	240	8	100	0.15	20.756
40	280	10	120	0.2	21.461
60	320	12	140	0.3	14.286
80	360	14	160	0.4	10.987
100	400	16	180	0.5	9.301

Table 3: Pretreatment results.

X_0	X_1	X_2	X_3	X_4	X_5
1	1	1	1	1	1
0.554514	0.5	0.833333	0.75	0.8	0.666667
0.573346	0.25	0.714286	0.6	0.666667	0.5
0.381654	0.166667	0.625	0.5	0.571429	0.333333
0.293524	0.125	0.555556	0.428571	0.5	0.25
0.248495	0.1	0.5	0.375	0.444444	0.2

Table 4: Relevance ranking results.

Structural parameter	Correlation	Rank
Mixing chamber length	0.586878	3
Diffusion chamber length	0.51222	5
Diffusion chamber angle	0.66573	2
Throat diameter	0.563725	4
Nozzle clearance	0.767563	1

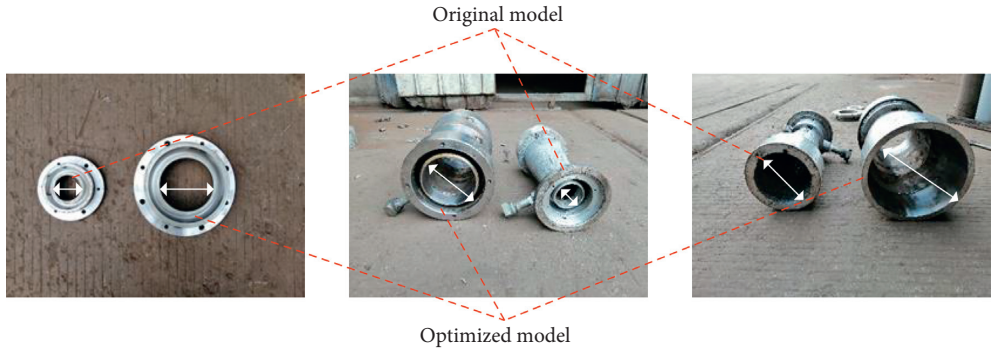


Figure 13: Comparison of annular ejector models. (a) Gland, (b) secondary inlet, and (c) outlet.

Table 5: Improved model dimensions for an annular slot ejector.

Structural parameter	Origin size (mm)	Optimised size (mm)
D_{main}	25	25
D_{sec}	80	160
D_{out}	145	290
d	80	160
e	0.3	0.1
θ	6°	6°
H_L	20	40
K_L	300	400

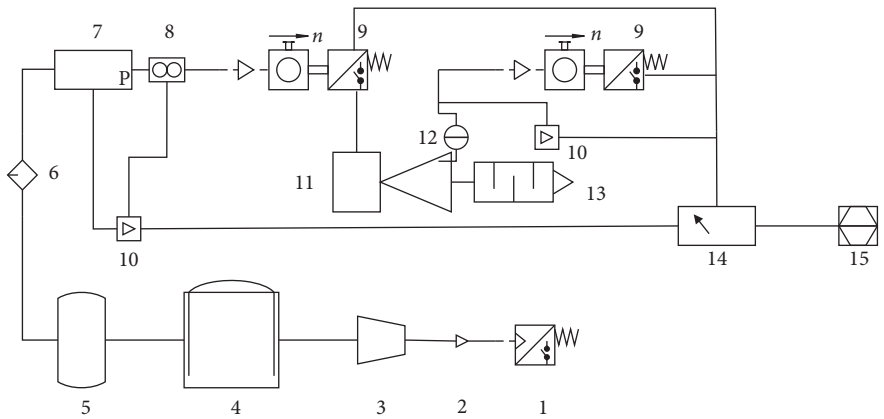


Figure 14: Schematic diagram of the experimental flow regime in the annular ejector. 1. Power switch. 2. Starting device. 3. Air compressor. 4. Gas tank. 5. Buffer gas tank. 6. Desiccator. 7. Pressure transmitters. 8. Gas vortex flowmeter. 9. Pulse counter. 10. Sensor. 11. Annular slot ejector. 12. L-type pitot tube and differential manometer. 13. Silencer. 14. Monitor. 15. Computer.

Table 6: Experimental results.

P_1 (MPa)	G_1 (kg/s)	G_2 (kg/s)	ϖ	P (Pa)	V (m^3/s)
0.30	0.0234	0.991	41.367	107.28	13.154
0.35	0.0276	1.061	37.428	122.97	14.083
0.40	0.0298	1.097	35.805	131.45	14.561
0.45	0.0341	1.220	34.780	162.59	16.194
0.50	0.0375	1.291	33.440	182.06	17.136
0.55	0.0409	1.257	31.678	172.60	16.685
0.60	0.0435	1.248	27.678	170.15	16.566
0.65	0.0463	1.209	25.118	159.67	16.048
0.70	0.0576	1.164	19.201	148.01	15.451

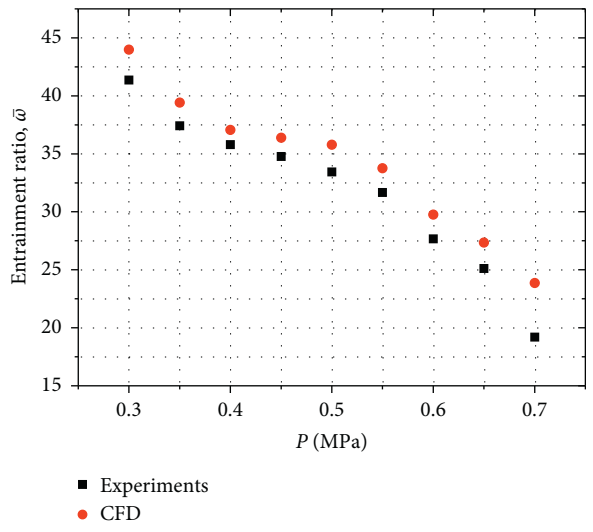


Figure 15: Comparison of experimental and numerical simulation results.

experimental data, which also indicates that the RNG-k-ε turbulent transport mode can simulate the flow characteristics of the gas in the ejector. Besides, it could be seen that numerical simulation results are always slightly higher than experimental test values, which can be explained using the Boussinesq hypothesis, ensure solution closure, and the working flow was set to that of an ideal gas.

5. Conclusion

To investigate the annular slot ejector performance, a two-dimensional ejector structure model was constructed employing Fluent 15.0. Five factors (e, θ, H_L, d, and K_L) were selected to analyse ejector performance; while one parameter was changed, the others were fixed. Then, the grey correlation analysis was used to study ejector performance when the five structural parameters were changed simultaneously, and the correlation of the performance parameters affecting the annular ejector was obtained. Finally, the optimised model was verified by comparing the numerical results with experimental data. The main conclusions were as follows.

ϖ first rises, then falls when increasing H_L from 10 mm to 100 mm, and when H_L was 40 mm, the maximum value of ϖ was obtained; there is a nonmonotonic trend seen when D rises from 80 mm to 160 mm. When θ increases from 6° to 16°, there is a continuously decreasing trend; however, when varying K_L (from 120 mm to 400 mm) and e (from 0.1 mm to 0.5 mm), ϖ increased at all times. Therefore, the dimensions of the optimised model were as follows: $H_L = 40$ mm, $K_L = 400$ mm, $\theta = 6°$, $d = 160$ mm, and $e = 0.5$ mm. The grey correlation between the five factors and the ejector performance was obtained as follows: $e > \theta > H_L > d > K_L$; therefore, the nozzle clearance is the most important parameter among the five factors affecting ejector performance. Nine groups of primary pressure conditions (from 0.3 MPa to 0.7 MPa) were

analysed using a set of experiments based on the optimised model to verify the accuracy of the simulated entrainment ratio; the corresponding results indicated that the two methods were consistent. Besides, the feasibility of the use of the RNG-k-ε turbulent transport mode was verified in simulating the flow characteristics of the gas in the ejector.

Data Availability

The data used to support the findings of this study are included within the manuscript.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant nos. 51974232 and 51574193) and Fundamental Research Funds of Shaanxi Province, China (Grant no. 2017JM5039).

References

[1] Y. Han, L. Guo, X. D. Wang, and A. C. Y. Yuen, "A steam ejector refrigeration system powered by engine combustion waste heat: part 1. characterization of the internal flow structure," *Applied Sciences*, vol. 9, no. 20, p. 4275, 2019.

[2] Y. Han, X. D. Wang, L. Guo, and A. Chun, "A steam ejector refrigeration system powered by engine combustion waste heat: part 2. understanding the nature of the shock wave structure," *Applied Sciences*, vol. 9, no. 20, 2019.

[3] Y. Han, X. Wang, A. C. Y. Yuen et al., "Characterization of choking flow behaviors inside steam ejectors based on the ejector refrigeration system," *International Journal of Refrigeration*, vol. 113, pp. 296–307, 2020.

[4] Y. Wu, H. Zhao, C. Zhang, L. Wang, and J. Han, "Optimization analysis of structure parameters of steam ejector based on CFD and orthogonal test," *Energy*, vol. 151, pp. 79–93, 2018.

[5] S. Da-Wen and W. E. Ian, "Recent developments in the design theories and applications of ejectors-a review," *Fuel & Energy Abstracts*, vol. 36, no. 5, pp. 361–370, 1995.

[6] D. J. Pieiro, "Henry Marie Coanda and the "Coanda effect"" *Revista Portuguesa De Pneumologia*, vol. 80, no. 1, p. 4, 2010.

[7] C. Ionică, D. Sorin, and G. Virgil, "Theoretical approaches regarding the gasodynamic phenomena in asymmetric flows," *Advanced Material Research*, vol. 1128, pp. 364–371, 2015.

[8] C. Smith, "On some recent applications of the Coanda effect to acoustics," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, p. 16, 2010.

[9] D. G. Gregory-Smith and A. R. Gilchrist, "The compressible Coanda wall jet--an experimental study of jet structure and breakaway," *International Journal of Heat and Fluid Flow*, vol. 8, no. 2, p. 9, 1987.

[10] S. Jain, Shashi, and S. Kumar, "Numerical studies on evaluation of smoke control system of underground metro rail transport system in India having jet injection system: a case study," *Building Simulation*, vol. 4, no. 3, pp. 205–216, 2011.

[11] I. Cîrciu and S. Dinea, *Review of the Air Force Academy*, "Henri Coandă" Air Force Academy, Braşov Romania, 2010.

[12] A. Perrig, F. Avellan, J.-L. Kueny, M. Farhat, and E. Parkinson, "Flow in a pelton turbine bucket: numerical and experimental investigations," *Journal of Fluids Engineering*, vol. 128, no. 2, pp. 350–358, 2006.

[13] E. Natarajan and N. O. Onubogu, "Application of Coanda effect in robots–a review," *Mechanical Engineering and Technology*, vol. 125, pp. 411–418, 2012.

[14] L. Guoqi, H. Yongjun, and Y. Yingzi, "Influence of Coanda surface curvature on performance of bladeless fan," *Journal of Thermal Science*, vol. 23, no. 5, p. 10, 2014.

[15] L. Dong-Won, H. Jae-Gun, K. Young-Doo, and K. Soon-Bum, "A study on the air knife flow with Coanda effect," *Journal of Mechanical Science and Technology*, vol. 21, no. 12, p. 7, 2007.

[16] D. Alexandru, F. Frunzulica, and C. I. Tudor, *Mathematical Modelling and Numerical Investigations on the Coanda Effect, Nonlinearity, Bifurcation and Chaos-Theory and Applications*, Intech, London, UK, 2012.

[17] D. Alexandru, F. Frunzulica, F. Frunzulica, and T. Ionescu, "Coanda effect on the flows through ejectors and channels," *Scientific Research and Education in the Air Force*, vol. 20, pp. 161–174, 2018.

[18] A. Dumitrache, F. Frunzulica, and O. Preotu, "Flow analysis in various ejectors configurations," in *Proceedings of the 2017 Fourth International Conference on Mathematics and Computers in Sciences and in Industry*, p. 7, Corfu, Greece, August 2017.

[19] A. Dumitrache, "Numerical investigation of the flow in a Coanda ejector," in *Proceedings of the 4th European Conference For Aerospace Sciences (EUCASS)*, p. 12, Saint Petersburg, Russia, July 2011.

[20] V. Rajalakshmi, K. Kavitha, and D. Lavanya, "Design and optimization of single head planar Coanda gripper," *Advances in Natural and Applied Sciences*, vol. 11, no. 4, p. 8, 2017.

[21] V. Benche and V. Benche, "Transient proceses for vent-ejectors assisted by Coanda effect," in *Proceedimgs of the 6th International Conference on Hydraulic Machinery and Hydrodynamics Timisoara*, p. 6, Timişoara, Romania, October 2004.

[22] T.-H. Kim, *A Study on the Characteristics of Coanda Nozzle Flow*, Saga University, Saga, Japan, 2007.

[23] P. M. Weston, V. Sharifi, and J. Swithenbank, "Destruction of tar in a novel Coandă tar cracking system," *Energy & Fuels*, vol. 28, no. 2, pp. 1059–1065, 2014.

[24] H. C. Yang, "Horizontal two-phase jet behavior with an annular nozzle ejector in the water tank," *Journal of Visualization*, vol. 18, no. 2, pp. 359–367, 2014.

[25] C. P. Lubert, "Some recent experimental results concerning turbulent Coanda wall jets," *The Journal of the Acoustical Society of America*, vol. 136, no. 4, p. 2137, 2015.

[26] A. Mohammad, *An Experimental and Theretical Study of Coanda ejectors*, Case Western Reverse University, Cleveland, OH, USA, 1993.

[27] V. Guerriero, "Numerical solutions of compressible flow mixing in Coanda ejectors," in *Proceedings of the Eighth Symposium on Fluid Control, Measurement and Visualization*, China Society of Theoretical and Applied Mechanics, Chengdu, ChinaChina Society of Theoretical and Applied Mechanics, Chengdu, China, August 2005.

[28] H. D. Kim, G. Rajesh, T. Setoguchi, and S. Matsuo, "Optimization study of a Coanda ejector," *Journal of Thermal Science*, vol. 15, no. 4, pp. 331–336, 2006.

[29] G. Rajesh, "A computational study of the gas flow in a Coanda ejector," in *Proceedings of the Korean Society of Mechanical Engineers Conference*, Busan, Korea, June 2005.

[30] J. A. Sierra del Rio, J. G. Ardila Marin, S. Vélez Garcia, M. Londoño, and D. A. Hincapie Zuluaga, "Simulation analysis of a coanda-effect ejector using CFD," *Tecciencia*, vol. 12, no. 22, pp. 17–25, 2016.

[31] K. P. Lowry, R. Y Coley, D. L Miglioretti et al., "Effect of Coanda nozzle clearance on the flow characteristics of air amplifier," in *Proceedings of the 2014 6th International Symposium on Fluid Machinery and Fluid Engineering*, Wuhan, China, March 2014.

[32] H. Amel, F. Henry, and S. Leclaire, "CFD analysis of a supersonic air ejector. Part I: experimental validation of single-phase and two-phase operation," *Applied Thermal Engineering*, vol. 29, no. 8, p. 9, 2009.

[33] Y. Victor and A. O. Steven, "Renormalization group analysis of turbulence I. basic theory," *Journal of Scientific Computing*, vol. 1, no. 1, pp. 3–51, 1986.

[34] D. Valentín, A. Guardo-Zabaleta, and E. Egusquiza, "Use of Coandă nozzles for double glazed fac¸ ades forced ventilation," *Energy & Buildings*, vol. 62, p. 10, 2013.

[35] A. Li, A. C. Y. Yuen, T. B. Y. Chen, and C. Wang, "Computational study of wet steam flow to optimize steam ejector efficiency for potential fire suppression application," *Applied Sciences*, vol. 9, no. 7, 2019.

[36] E. F. Schum, P. M. Bevilaqua, and S. V. Patankar, *Computation of the Turbulent Mixing in Curved Ejectors*, Rockwell International Corporation, Milwaukee, WI, USA, 1980.

[37] A. R. Gilchrist, *The Development and Breakaway of a Compressible Air Jet with Streamline Curvature and its Application to the Coanda*, Durham University, Durham, UK, 1985.

[38] D. Ju-Long, "Control problems of grey systems," *Systems & Control Letters*, vol. 1, no. 5, pp. 288–294, 1982.

[39] J. Zhou, J. Ren, and C. Yao, "Multi-objective optimization of multi-axis ball-end milling inconel 718 via grey relational analysis coupled with RBF neural network and PSO algorithm," *Measurement*, vol. 102, p. 15, 2017.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 8718597, 13 pages
https://doi.org/10.1155/2020/8718597

Research Article

Machining Parameters and Toolpath Productivity Optimization Using a Factorial Design and Fit Regression Model in Face Milling and Drilling Operations

Gustavo M. Minquiz,[1,2] **Vicente Borja**,[3] **Marcelo López-Parra**,[3] **Alejandro C. Ramírez-Reivich**,[3] **Leopoldo Ruiz-Huerta**,[4,5] **R. C. Ambrosio Lázaro**,[1] **Alejandro Shigeru Yamamoto Sánchez**,[6] **H. Vazquez-Leal**,[7] **María-Esther Pavon-Solana**,[1] **and J. Flores Méndez**[1,2]

[1]*Benemérita Universidad Autónoma de Puebla-Ciudad Universitaria, Blvd. Valsequillo y Esquina, Av. San Claudio s/n, Col. San Manuel, C.P. 72570, Puebla, Pue, Mexico*
[2]*Tecnológico Nacional de México/I.T, Puebla Av. Tecnológico No. 420, Maravillas, C.P 72220, Puebla, Pue, Mexico*
[3]*Universidad Nacional Autónoma de México, Facultad de Ingeniería, Av. Universidad No. 3000, C.P. 04510, Ciudad de México, Mexico*
[4]*Universidad Nacional Autónoma de México, Instituto de Ciencias Aplicadas y Tecnología (ICAT), Circuito Exterior s/n, Ciudad Universitaria AP 70-186, C.P. 04510, Ciudad de México, Mexico*
[5]*National Laboratory for Additive and Digital Manufacturing, (MADiT), Circuito Exterior s/n, Ciudad Universitaria, A.P. 70-186, C.P 04510, Ciudad de México, Mexico*
[6]*Sandvik Coromant México-Parque Industrial Querétaro, Av. Cerrada de la Estacada #550 C, Santa Rosa Jaúregui, C.P. 76220, Mexico*
[7]*Facultad de Instrumentación Electrónica, Universidad Veracruzana, Cto. Gonzalo Aguirre Beltrán S/N 91000, Xalapa-Veracruz, Mexico*

Correspondence should be addressed to J. Flores Méndez; xavier_snk@hotmail.com

Received 10 February 2020; Revised 29 May 2020; Accepted 30 May 2020; Published 27 July 2020

Guest Editor: Mohamed Nayel

Very commonly, a mechanical workpiece manufactured industrially includes more than one machining operation. Even more, it is a common activity of programmers, who make a decision in this regard every time a milling and drilling operation is performed. This research is focused on better understanding the power behavior for face milling and drilling manufacturing operations, and the methodology followed was the design of experiments (DOEs) with the cutting parameters set in combination with toolpath evaluation available in commercial software, having as main goal to get a predictive power equation validated in two ways, linear or nonlinear, and understanding the energy consumption and the quality surface in face milling and final diameter in drilling. The results show that it is possible to find difference in a power demand of 1.52 kW to 3.9 kW in the same workpiece, depending on the operations (face milling or drilling), cutting parameters, and toolpath chosen. Additionally, the equations modelled showed acceptable values to predict the power, with p values higher than 0.05 which is the significance level for the nonlinear and linear equations with an R square predictive of 98.36. Some conclusions established that optimization of the cutting parameters combined with toolpath strategies can represent an energy consumption optimization higher than 0.21% and the importance to try to find an energy consumption balance when a workpiece has different milling operations.

1. Introduction

According to Energy Information Administration (EIA) [1], 27203 Trillion BTU was the total energy consumed by the industrial sector in 2018. The aerospace, automotive, and plastics industries use machining processes to produce metallic parts for different manufacturing purposes [2]. To improve energy consumption efficiency in manufacturing systems, several researchers have focused their work on identifying the primary variables involved in machining [3]. Several analyses to assess the environmental impacts of machining processes have been carried out to better understand the behavior of the energy consumed. Some authors have found that the energy required for material removal is very low when compared to the total energy of the machine tool operation [4]. Therefore, several researchers have focused their work on finding ways to reduce this total energy rate. Guo et al. [5–7] analyzed the surface quality and energy consumption of a turning process; the aforementioned research tested a workpiece made of steel 11SMnPb30 and found that the best cutting parameters ranged as follows: a table feed of 0.05–0.3 mm/*r*, an axial depth of cut 0.5–1 mm, and a cutting speed of 80–800 m/min. To understand and minimize energy behavior in the milling processes, different operations have been studied. These previous works have proposed face milling improvements based on efficient toolpaths and energy consumption [8–10]. These improvements were such as the optimization of surface milling regarding cycle time, as well as the study of toolpath selection as a critical factor [11, 12]. Investigation in the optimization of machining parameters trying A6061 aluminum and using the response surface methodology based on particle swarm optimization not only led to finding an optimal combination of machining parameters but also that the spindle speed variable is the most influencing in surface roughness, power consumption, and cutting force. This alternative study showed ±7% of error [13]. Mathematical power prediction has been proposed, based on experimental models of infinitesimal cutting force during the machining process in order to obtain the power consumption, where the data were divided into spindle rotational, feed motion, and idle power, and the error of mean power was 0.208% [14]. Other studies aimed at the development of a mathematical model to know the power demand in face milling has considered both the cutting parameters and the flank wear [15]. Additionally, an artificial intelligence model was proposed in order to monitor the main drive power in the CNC machine in real time. The surface roughness and measuring the flank wear of the cutting tool were the main output variables considered, and random forest techniques had an accuracy between 33 and 44% [16]. Other techniques have considered the multi-optimization analysis, using the same roughness surface methodology, using titanium Ti-6A1-4V, as well as the surface roughness values. This model made it possible to know the power consumption with a correlation coefficient higher than 90% [17]. Other relevant studies consider the laser surface texturing (LST) technique in combination with the addition of nanoparticles (NPs) to the cutting fluid for getting better results in terms of reducing the spindle load (power demand) and reducing the tool wear in relation to surface roughness and thus finding that it is possible to improve the milling process up to 11% [18]. Some studies have successfully used milling processes with more than one machining type, determining different cutting parameters for the primary geometry features of their case studies [19, 20]. Behrendt et al. [19] propose a procedure for machine tool selection based on energy consumption, while Helu et al. [20] evaluate green machining technologies based on sustainability and cost. Li et al. [21] have investigated workpieces with different machined features to identify relevant factors. They found that critical factors are cutting parameters optimization, process planning, and job shop scheduling. The analysis that they have carried out has proved that the toolpath selection impacts both the material removal power and the air cutting times. For a process, which includes more than one milling feature, it is necessary to select different cutting parameters for each feature, based on material, cutting tool, and toolpath [22, 23]. The power required by a milling operation depends on parameters like spindle speed, feed rate, depth of cut, and transversal cut [24]. Different investigations have proposed energy consumption predictive models based on cutting parameters to estimate surface roughness [25, 26] and tool wear [27]. In order to optimize machining processes, different statistical techniques were used. With the use of either nonlinear or linear polynomial models, it is possible to model power behavior. Amongst these techniques are design of experiments (DOEs), Taguchi, nondominated sorting genetic algorithm (NSGA), and response surface.

This paper reports on a research aimed at modelling the power consumption of the face milling and drilling operations in terms of cutting parameters. This is relevant because such power demand models assist in defining cutting parameters in order to achieve the required quality of surface, which is key in manufacturing processes and adequate productivity rate. The energy consumption is also considered as a factor. Besides, the work reported successfully identifies facing and drilling toolpaths with the lowest power demands and employs DOE to define tests that were carried out to calculate the constants of the desired models. The methodology used by the authors identified that nonlinear equations provided a better fit than linear equations for the machining operations studied. Additionally, the combination of the cutting parameters and toolpath is more suitable to improve a machining process.

2. Power Demand Background

Several researchers have worked on understanding power consumption on machine tools. Draganescu et al. attempted to develop a model for the spindle motor's efficiency η using the following equation, which relates the p_c, minimal cutting power, with p_{mc}, the power consumed by the spindle drive motor [28]:

$$\eta = \frac{p_c}{p_{\mathrm{mc}}}. \tag{1}$$

Gutowski et al. [4] determined the power consumed in a machining operation using the following equation, which estimates p, total power, with p_o, idle power, $\dot{m}$, the rate of material processing, and k, a constant:

$$p = p_o + k\dot{m}. \tag{2}$$

Different authors [4, 7] have analyzed machining power demand composed of three components:

(i) *Standby Power*. It is the constant power consumed by a machine tool whether material is being cut or not. The electrical components of a machine that always consume power are its controller, its computer panel, and its lights.

(ii) *Component Power*. It is the power required to achieve the desire spindle speed and feed rate and is needed to carry out each tool change and power the coolant pump.

(iii) *Machining Power*. It is the power used to remove material.

There exist different alternatives to name each part of the power used by a CNC machine. For example, equation (3) shows the total power consumed by a machine tool. The constant power is the equivalent to the standby state which includes unloaded motor, coolant pump, computer, and fans. The variable power includes the required power to move the spindle and carrying out jog and tool change. Finally, the force necessary to remove the material is the cutting power, Cutting P in equation (3), which is equivalent to the machining power. On the other hand, the constant power, Constant P, plus the variable power, Variable P, equal the tare power [4]:

$$T.\ \text{Power} = \text{Constant}\ P + \text{Variable}\ P + \text{Cutting}\ P. \tag{3}$$

2.1. Models to Predict Surface Roughness and Power Consumption. Different statistical techniques [29, 30] have been used to model the relationship between output variables and cutting parameters in machining processes. These techniques have showed their advantages on accuracy and simplicity. To establish a regression model for a particular phenomenon or behavior, it is necessary to set up selective experiments, considering output variables as dependent and input variables as independent [31]. Next, data are collected from running experiments, and an equation, i.e., a regression model, is defined based on a statistical analysis of measurements. There exist several regression models, i.e., first order, second order, and nonlinear models. In the particular case of nonlinear models, their constants can be calculated using the method of least square on the results of experiments [31]. Fang and Safi-Jahanshahi [26] proposed a nonlinear model (fit regression model), to predict the surface roughness as a variable dependent on the cutting speed, feed per tooth, and depth of cut, as follows:

$$y = x_1 \cdot a^{x_2} \cdot b^{x_3} \cdot c^{x_4}. \tag{4}$$

In equation (4), the dependent variable is y; x_1, x_2, x_3, and x_4 are constants; and a, b, and c are independent variables. The experiments documented in [26] reported an average error for equation (4) of 3.86%, and constants x_1, x_2, x_3, and x_4 started with a value of one. The Levenberg–Marquardt algorithm was used to find the optimum fit values. The following equation was used to solve the Levenberg–Marquardt algorithm:

$$\delta(k) = \left(V^T V + kD\right)^{-1} V^T (y - \eta), \tag{5}$$

where k is a conditioning factor and D is a diagonal matrix with elements of $V^T V$. The direction of $\delta(k)$ is intermediate between the direction of the Gauss–Newton increment ($k \longrightarrow 0$) and the direction of steepest descent, as follows:

$$\frac{V^T (y - \eta)}{\left\| V^T (y - \eta) \right\|}, \quad (k \longrightarrow \infty). \tag{6}$$

This paper contributes to the understanding of the power and energy behaviors as functions of machining process strategies and cutting parameters. Two milling operations are studied, facing and drilling. In order to identify which one consumes less power, three strategies are tested for facing and two for drilling. Next, DOE is used to define experiments, and the Levenberg–Marquardt algorithm with a least square estimation is employed to help in calculating the constants of linear and nonlinear models that lead to the final total power estimation. Additionally, measurements of surface finish achieved during the experiments and a nonlinear model for tare power are reported.

3. Experiment Setup

Figure 1 shows the use of the DOE methodology followed in this work to model the energy consumed by face milling and drilling.

3.1. Case Study Specification and Cutting Tool Selection. The process comprised two operations: first, face milling, consistent on getting a flat area, perpendicular to the axis of the spindle, with a depth of cut of 3 mm; second, a drilling operation applied on nine Ø 14 mm holes with a depth of 25 mm, in a rectangular array of 3×3. The machining process to manufacture the part was defined using the commercial software Mastercam® × 8. The face milling operation was performed using different toolpath strategies for each of them: dynamic, zigzag, and one-way. This operation was carried out in one step, i.e., only one facing stage, while the drilling operation was defined first using the counterbore toolpath cutting option and then employing the chip break. Table 1 shows the experiment about material and cutting tool.

The experiments here reported (Tables 2 and 3) implemented the use of air to remove chips during the process because the color of the chips was blue, which meant that due to the settings selected for the cutting parameters, the friction between the tool and the material was not too high.

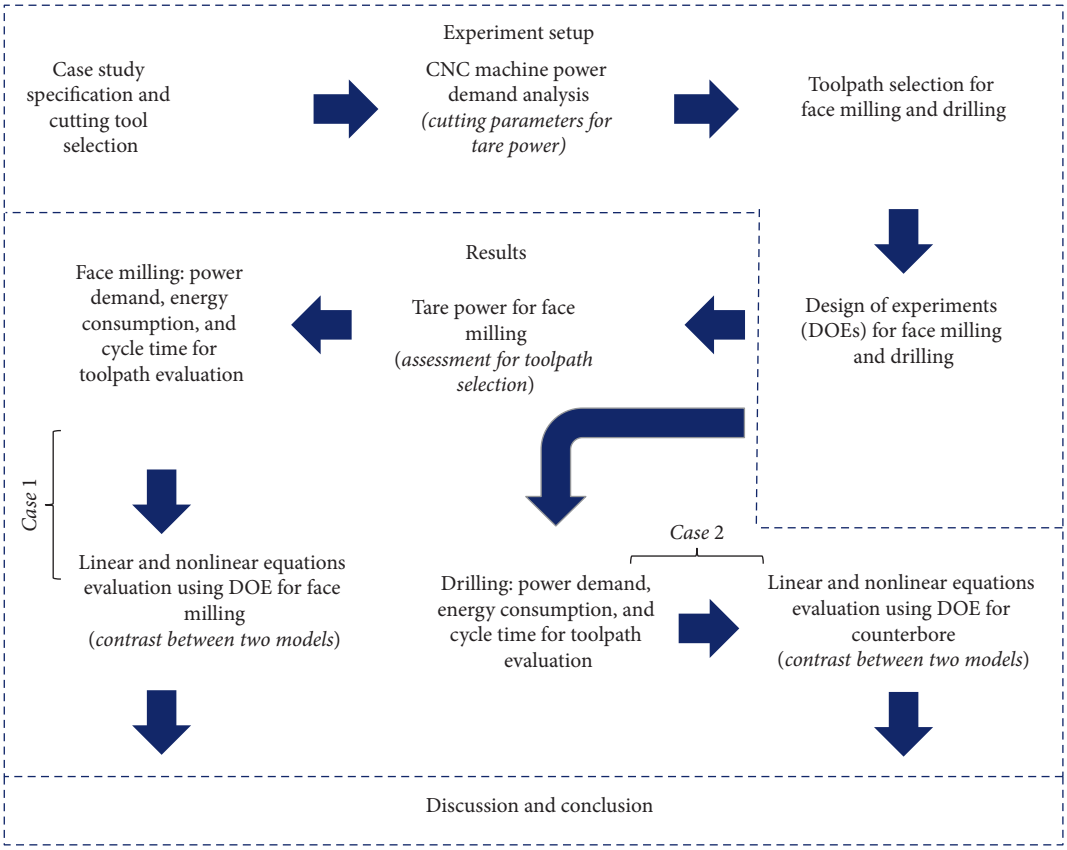


Figure 1: Methodology followed for this research.

Table 1: Details of the experiment.

Workpiece dimensions	180 mm × 150 mm, depth 30 mm, AISI 1045 medium carbon steel, hardness, Brinell 163
Workpiece material	Carbon 0.43–0.5% max, manganese 0.6–0.9% max, silicon 0.05% max, phosphorus 0.04% max, and sulphur 0.050% max
Chemical composition	ISO code IP240-1200-XA 1630 a solid carbide square end mill with effective cutting edge: 4, cutting diameter: 12 mm, flute helix angle: 35 deg, corner chamfer width: 0.2 mm
Tool material	ISO code 860.1-1400-040A1-PM 4234 a solid carbide drill with cutting diameter: 14 mm, hole tolerance: H8, point angle: 147°, chip flute length: 60 mm

Table 2: Cutting parameters used in the DOE for tare power.

Parameters	1	2	3
Spindle speed (rpm)	2005	5777	9549
Feed rate (mm/min)	802	2607	4412

3.2. CNC Machine Power Demand Analysis, Surface Roughness, and Drilling Holes. The experiments were set up on a milling machine Haas VM3. The power was measured using a three-phase energy analyzer Fluke® 430 series II, which is a portable device, complemented by alligator clips and current clamps to connect it (Figure 2). The measurement equipment took data every second, sending information to the computer, and analyzed in the Fluke® PowerLog software.

A portable surface roughness tester Mitutoyo® SJ-210 (Figure 2) was used; the instrument was set to measure under ISO 4287 : 1997 the arithmetical mean roughness value (Ra) with a resolution of 0.006 μm, cutoff wavelength λc of 0.8, roughness sampling length of 0.8 mm, the mean groove spacing RSm (mean peak width) of ×5 and 0.05 mm/s speed, stylus tip radius of 5 μm, and measuring force of 4 mN to measure the surface quality. The procedure can be described as follows: the first step consists on taking three samples in

Table 3: Cutting parameters for the DOE for the face milling and drilling operations.

Parameters		Level		
		Low	Medium	High
Face milling	v_c, cutting speed (m/min)	315	350	385
	a_p, depth of cut (mm)	0.5	1.0	2.0
	f_z, feed per tooth (mm)	0.15	0.175	0.2
Drilling	Cutting speed (m/min)	315	350	385
	Feed (mm/r)	0.15	0.175	0.2

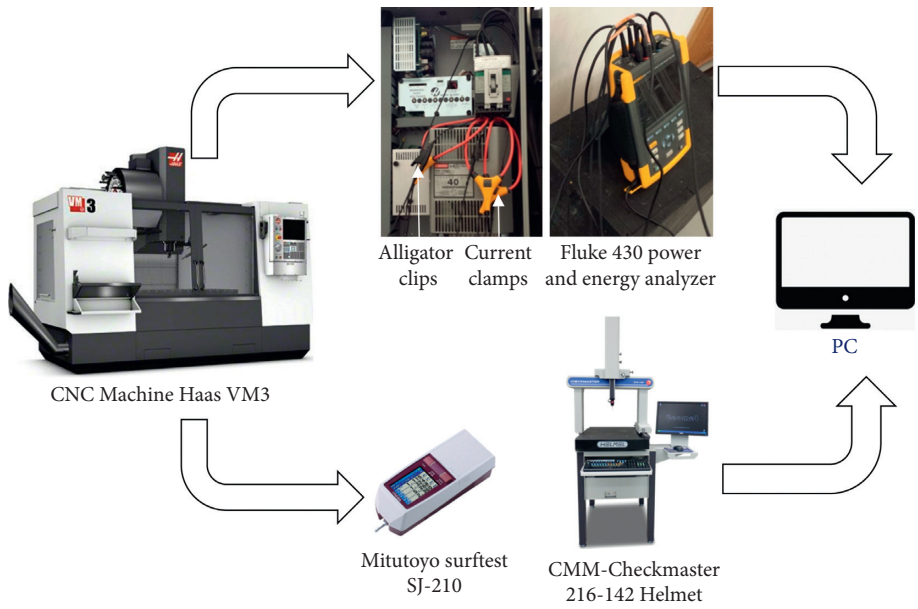


Figure 2: Schematic representation for collecting data.

different points along the flat facing machined area; the second step is to repeat three times the measure in the same points and finally is getting the average between the measurements that were obtained. The drilling holes were measured with a Checkmaster® 216 142-Helmet with resolution (Figure 2) 0.5 μm, coordinate measuring machine using a Renishaw sensor adapted to M2 Ø 4 mm ruby ball, stainless steel stem, and L 20 mm; the data were collected and analyzed using Geomet® 101 software. The procedure followed begins with collecting the first 3 data points at or near the bottom along the axis of a cylinder, secondly collecting the last 3 data points at or near the top, then repeating twice the measure in the same points, and finally getting the average between the measurements found.

The breakdown of the power of the CNC machine used was analyzed using the measurements that were obtained with the power analyzer.

3.2.1. Total Power (kW). The total power consumed in each one of the experiments (Tables 2 and 3) was measured. Different tests in each experiment were carried out to estimate the coefficients x_1, x_2, and x_3, while a_p, v_c, and f_z are values of the cutting parameters chosen by the user to model the power consumption with a first-order model (see equation (7)) and a nonlinear model (fit regression model) (see equation (8)) and identifying the convenience of both:

$$\mathrm{kW}_{\mathrm{Total}} = a_p \cdot x_1 + v_c \cdot x_2 + f_z \cdot x_3, \tag{7}$$

$$\mathrm{kW}_{\mathrm{Total}} = x_1 \cdot a_p^{x_2} \cdot v_c^{x_3} \cdot f_z^{x_4}. \tag{8}$$

3.2.2. Constant Power. His factor was measured setting the machine to the standby state. Its value was 0.55 kW. During the experiments, the coolant water pump was turned off because the metal chips were removed by using air instead.

3.2.3. Tare Power. Face milling was used for modelling the tare power. It was measured setting the machine tool in a running condition, without removing material configured with the dynamic toolpath strategy. This toolpath was used in the study (Table 2) considering the low, middle, and high ranges that could reach the spindle speed and feed rate. So, nine tests were defined in this experiment. The data collected were used to calculate the constants of the nonlinear model proposed to estimate this energy. The model was established with spindle speed, n, and feed rate, v_f, as independent variables, shown as follows:

$$\mathrm{kW}_{\mathrm{Tare}} = x_1 + n^{x_2} + v_f^{x_3}. \tag{9}$$

3.2.4. Cutting Power. This is the necessary power to remove material. It could be determined by subtracting the tare power from the total power.

The parameters chosen in Table 2 were set under parameters mentioned in the data sheet of the CNC machine, the maximum spindle speed is 1200 rpm, while the maximum feed rate is 18 000 mm/min, and also the parameters are according to cutting tool operations suggested by the supplier. The coolant was turned off because it was not removed material.

3.3. Toolpath Selection for Face Milling and Drilling. This research considered the face milling for modelling the tare power. Cutting parameters were set as recommended by the tool supplier [32] for the facing operation. Three toolpath strategies were used: dynamic, zigzag, and one-way (Figure 3); the toolpath with the lowest power demand was identified.

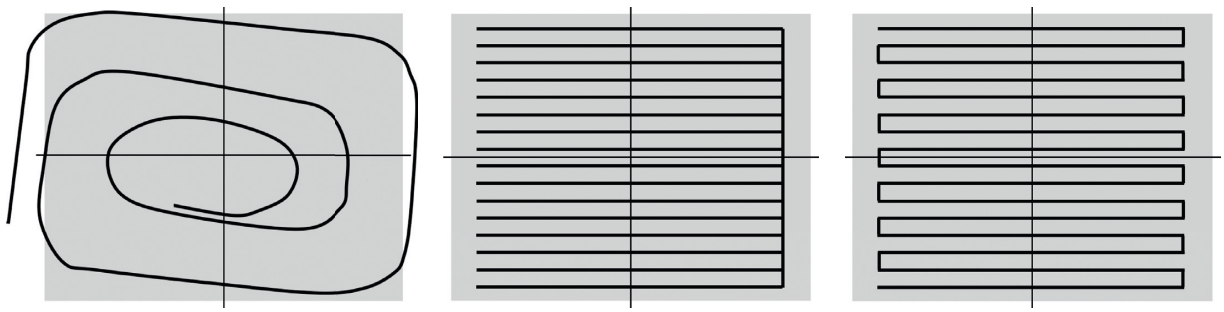


Figure 3: Toolpath strategies: (a) dynamic; (b) one-way; (c) zigzag. The drill counterbore and chip break toolpath cutting options were tested to identify the tool path strategy that required the lowest power demand.

The cutting parameters were the same for the three strategies (i.e., feed per tooth 0.2 mm, cutting speed 385 m/min, and depth of cut 0.5 mm), reaching a final depth of 3.0 mm.

3.4. Design of Experiments for Face Milling and Drilling. The cutting parameters for face milling were defined using a DOE 3^3. Table 3 shows these values. In order to obtain a satisfactory roughness, the levels of the depth of cut were set to low values: 0.5, 1, and, finally, 2 mm. The quality of surface finish achieved in the workpiece was measured.

For drilling, cutting parameters were defined with an experiment 2^3, as shown in Table 3. This operation was set with the coolant applied through the tool.

As mentioned before, a DOE 3^3 was used for face milling, and this generates 27 possible combinations of cutting parameters for the machining operation. All the values of the low, medium, and high levels for face milling of Table 3 are within the ranges suggested by the tool supplier. A DOE 2^3 was selected for drilling, which implies 9 possible combinations of cutting parameters. Again, all the values of the low, medium, and high levels for face milling of Table 3 are within the ranges suggested by the tool supplier. The results were used for the power demand and energy consumption statistical models proposed by the authors as described below.

4. Results

4.1. Tare Power for Face Milling. The statistical software Minitab® was used to carry out the nonlinear and linear regressions of the data obtained with the experiments shown in Table 3. Each one of the 27 defined tests were carried out once, to calculate the constants for equations (10) and (11). Table 4 shows the data of the nonlinear analysis solved with the Levenberg–Marquardt method and based on the results of the facing experiments using the CNC machine operating with a dynamic toolpath strategy. Table 4 also includes the estimated parameter values and their standard error.

The nonlinear predictive model equation (10), which is based on equation (9), presents kW_{Tare} as a function of n, spindle speed, and v_f, feed rate. In the following equation, kW_{Tare} is the tare power (i.e., constant power plus variable power):

Table 4: Nonlinear regression analysis carried out with the Levenberg–Marquardt algorithm.

Parameter	Estimate	SE estimate
x_1	0.003014	0.003773
x_2	0.639114	0.123164
x_3	0.070207	0.078683

p value: 0.623; maximum iterations: 200; tolerance: 0.00001; SE: standard error.

$$kW_{Tare} = 0.00301386 \cdot n^{0.63914} \cdot v_f^{0.0702071}. \quad (10)$$

Figure 4 shows over the X-axis the three spindle speeds chosen in Table 2. However, each speed is repeated three times because it was necessary to make nine tests in this experiment. For example, the spindle speed of 2005 rpm and the feed rate of 802 mm/min were tested with the three toolpaths chosen: dynamic, zigzag, and one-way; this is what the first three bars show in Figure 4, and the next three bars show the same spindle speed but with a feed rate of 2647 mm/min and the same toolpath mentioned. The tare power over the y axis, of the CNC machine, increases as the spindle speed goes up. The three spindle speeds and feed rate values specified by Table 2 were tested in the experiment. The values calculated with equation (10) to predict the tare power demand for the strategy dynamic are represented by the dotted line.

4.2. Case 1 Face Milling: Power Demand, Energy Consumption, and Cycle Time for Toolpath Evaluation. Three different toolpath strategies, i.e., cutting options, were tested with the face milling operation using the medium values defined in Table 3, the cutting parameters were tested three times, and Table 5 shows the average values obtained.

As shown in Table 5, the toolpath dynamic consumed 0.121 kWh, which was the lowest value, and the quality surface obtained was 0.22 μm. Therefore, considering the power and energy values, the dynamic cutting option resulted in being the most convenient for energy saving.

4.3. Nonlinear and Linear Equations Evaluation Using DOE. Table 6 presents the results of the statistical analysis for the experiment defined in Table 3. Twenty-seven tests were

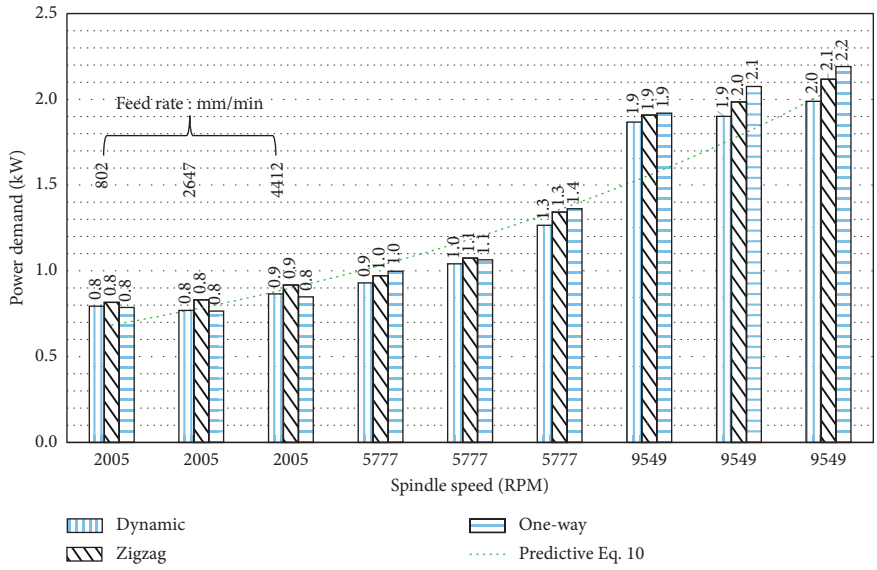


Figure 4: Experimental and theoretical tare power for the face milling operation.

Table 5: Face milling operation comparative with different cutting options.

Cutting option	Cycle time (hrs)	Power demand (kW)	Energy consumption (kWh)	Quality surface R_a (μm)	
				X-axis	Y-axis
Dynamic	0.08	1.52	0.121	0.22	0.24
Zigzag	0.10	1.55	0.154	0.20	0.28
One-way	0.07	1.99	0.136	0.21	0.22

Table 6: Nonlinear regression analysis carried out with the Levenberg–Marquardt algorithm.

Parameter	Estimate	SE estimate
x_1	0.119379	0.0296690
x_2	0.586379	0.0127844
x_3	0.660086	0.0414910
x_4	0.501722	0.0290065

Maximum iterations: 200; tolerance: 0.00001; SE: standard error.

carried out using the dynamic cutting options because it had the lowest power demand. The maximum power demand measured in the experiments was 4.10 kW, the medium value was 2.98 kW, and the lowest was 1.98 kW.

As mentioned in the previous section, the results of Table 6 were used for the model presented by equation (11), where $\mathrm{kW}_{\mathrm{Total}}$, the total power demand, can be estimated from the cutting parameters a_p, v_c, and f_z:

$$\mathrm{kW}_{\mathrm{Total-facing}} = 0.119379 \cdot a_p^{0.586379} \cdot v_c^{0.660086} \cdot f_z^{0.501722}. \quad (11)$$

Figure 5 shows that there is a normal distribution because the residual model falls on the straight line, which can be seen in the normal probability plot. Additionally, the histogram shows that the data are skewed, and the residuals are randomly distributed, which means constant variance is possible as seen in the versus fits chart. The residuals are independent from one another and are randomly distributed around the center line which means a normal behavior in the versus order chart. This tendency demonstrates the suitability of the experimental results.

Table 7 summarizes the results of the 27 experiments carried out with the DOE specified in Table 3 and shows the ANOVA data. The significance of each factor a_p, v_c, and f_z was tested. The analysis was carried out with a confidence level of 95% for all intervals. In Table 7, the p values of the three factors are significant because they are lower than 0.05, which means that the three factors are determinant for the power demand.

The surface quality produced by the 27 tests carried out, which combined different cutting parameters, was measured. The minimum average surface roughness obtained, R_a, was 0.01 μm and the maximum 0.183 μm. Therefore, these values according to indication of surface texture in technical product documentation fell between $N1$ and $N3$, which represent superfinishing and high level of finish [33].

Equation (12) shows the linear regression equation to predict the power in kW. This includes the coefficients determined using the least squares estimation, which represents the estimated change in mean response for each unit change in the predictor value:

$$\mathrm{kW}_{\mathrm{Total-facing}} = -2.296 + 1.1881 a_p + 0.005681 v_c + 8.621 f_z. \quad (12)$$

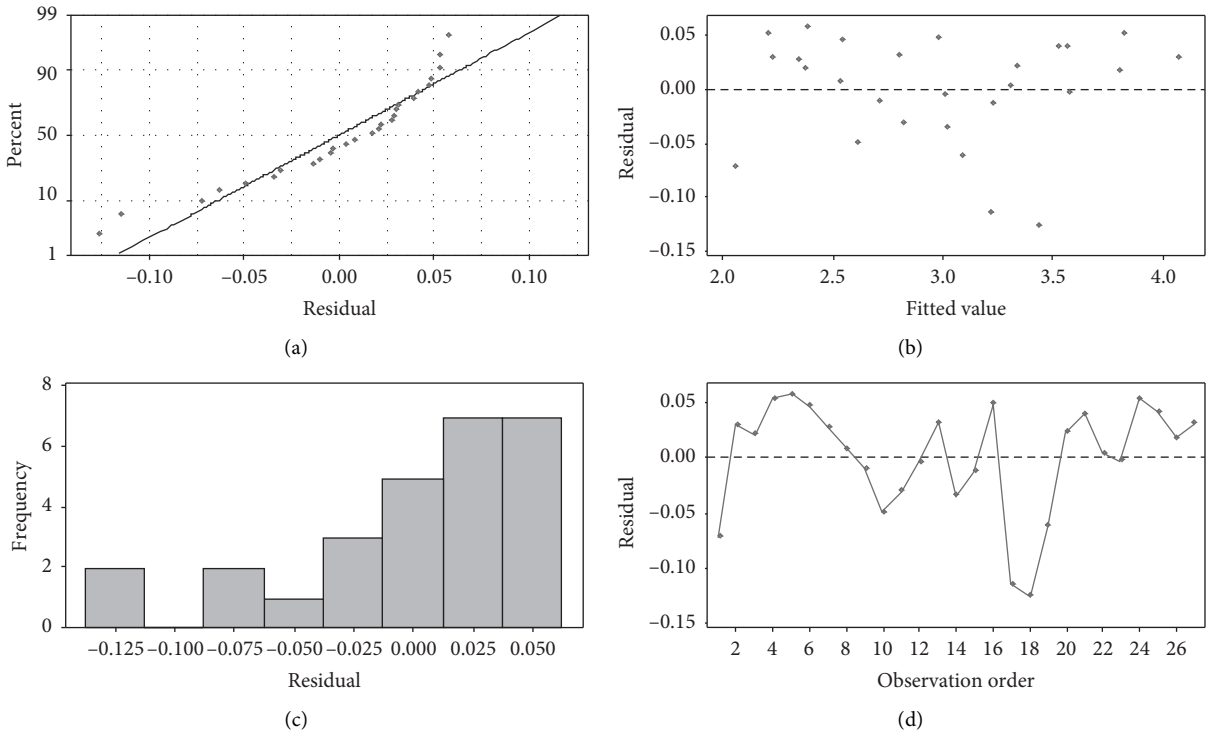


Figure 5: Residual plots for the face milling operation: (a) normal probability plot; (b) versus fits; (c) histogram; (d) versus order.

Table 7: Variance analysis of the experiment (Table 3).

Source	DF	Adj. SS	Adj. MS	F value	p value
Regression	3	7.89958	2.63319	669.94	0.00001
a_p	1	6.35191	6.35191	1616.07	0.00001
v_c	1	0.71156	0.71156	181.04	0.00001
f_z	1	0.83611	0.83611	212.72	0.00001
Error	23	0.09040	0.00393		
Total	26	7.98998			

R^2 (pred): 98.36%.

Figure 6 shows a comparison between the real power demands (denoted with green color in figure), with the linear equation prediction (equation (12), denoted with blue line in figure) and with the nonlinear equation prediction (equation (11), denoted with yellow color in figure). The figure compares the values obtained in each one of the 27 tests carried out during the experiment. Clearly, the green line is completely aligned with the yellow line, whereas the blue line shows different slight values in comparison with the real power behavior. For example, in test 2, the real value is 2.25 kW, which contrasts with the 1.97 kW estimated by the linear predictive equation.

4.4. Case 2 Drilling: Power Demand, Energy Consumption, and Cycle Time for Toolpath Evaluation. The two different drilling cutting options tested, counterbore and chip break, were compared. The tool machined the hole with a single movement when the counterbore cutting option was used. The hole was machined progressively, i.e., with a series of penetration and withdrawal movements, using the chip break option.

The toolpath counterbore consumed 0.011 kWh to produce a hole with a radius of 7.001 mm (Table 8). Therefore, considering the lower energy value, the counterbore toolpath option resulted to be just more convenient for energy saving.

4.5. Nonlinear and Linear Equations Evaluation Using DOE for Counterbore. Table 9 presents the results of the statistical analysis for the experiment defined in Table 3. Nine tests were carried out using the counterbore drilling option because it had the lowest power demand. The maximum power demand measured in the experiments was 4.12 kW, the medium value was 3.46 kW, and the lowest was 2.94 kW.

The modelling of the drilling operation was carried out for the drill counterbore toolpath based on the results of the regression analysis presented in Table 9. In the model of the operation, kW_{Total} represents the total power demand and it is a function of the cutting drilling parameters v_c and f_n, as follows:

$$kW_{Total-drilling} = 0.27747 \cdot v_c^{0.486936} \cdot f_n^{0.0197716}. \quad (13)$$

Figure 7 shows that there is a normal distribution because the residuals for the model fall on the straight line, making it possible to see this in a normal probability plot. Also, the histogram shows that the data are skewed, and the residuals are randomly distributed which means constant variance is possible seeing in versus fits charts; the residuals are independent from one another, those randomly around the center line mean a normal behavior in versus order chart.

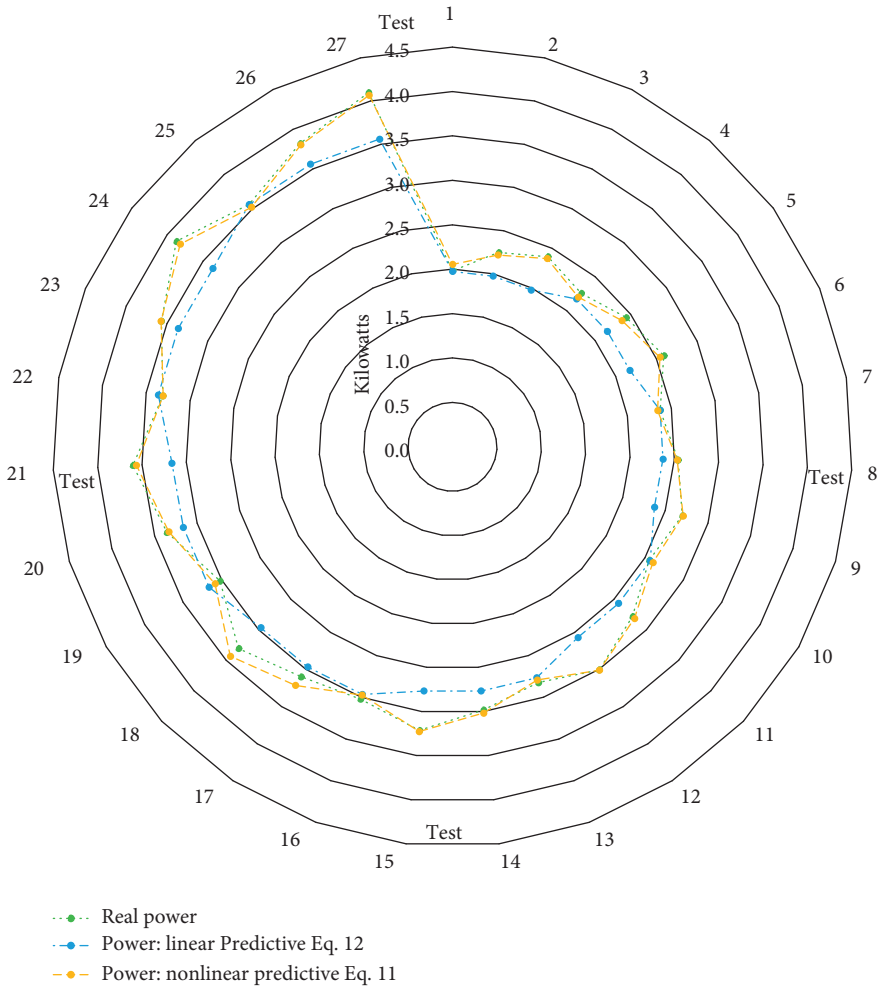


Figure 6: Face milling power real value compared against the results obtained with the nonlinear and linear predictive models.

Table 8: Drilling operation comparative with different cutting options.

Cutting option	Cycle time (hrs)	Power demand (kW)	Energy consumption (kWh)	Radii (mm)
Counterbore	0.003	3.930	0.011	7.001
Chip break	0.003	3.256	0.014	7.002

Table 9: Nonlinear regression analysis carried out with the Levenberg–Marquardt algorithm.

Parameter	Estimate	SE estimate
x_1	0.277470	0.0820493
x_2	0.486936	0.0552161
x_3	0.019772	0.0419983

Maximum iterations: 200; tolerance: 0.00001.

This tendency demonstrates the suitability of the experimental results. The data collected are the support to construct the model of equation (13) with its respective variables.

Table 10 shows the results of the 9 tests carried out with the parameters of the DOE displayed in Table 3 and also shows the ANOVA data. The significance of each factor v_c and f_n was tested. The analysis was carried out with a confidence level of 95% for all intervals. The p value of the v_c factor is significant, which means that this parameter has an important impact on power. However, the f_n factor is higher than 0.05, implying a low impact on power.

Equation (14) shows the model to predict the power generated with the regression analysis. It can be seen that drilling operations include two cutting parameters, cutting speed (v_c), and feed per revolution (f_n):

$$\mathrm{kW}_{\mathrm{Total-drilling}} = 1.661 + 0.008873 v_c + 0.367 f_n. \quad (14)$$

Figure 8 shows a comparative between the measured values of power, real powers (green color), with the values obtained with the nonlinear equation (equation (13), denoted with blue color in figure) and with the linear equation (equation (14), denoted with yellow color in figure). The blue line is aligned with the yellow line which are the predictive equations, whereas green line shows slightly

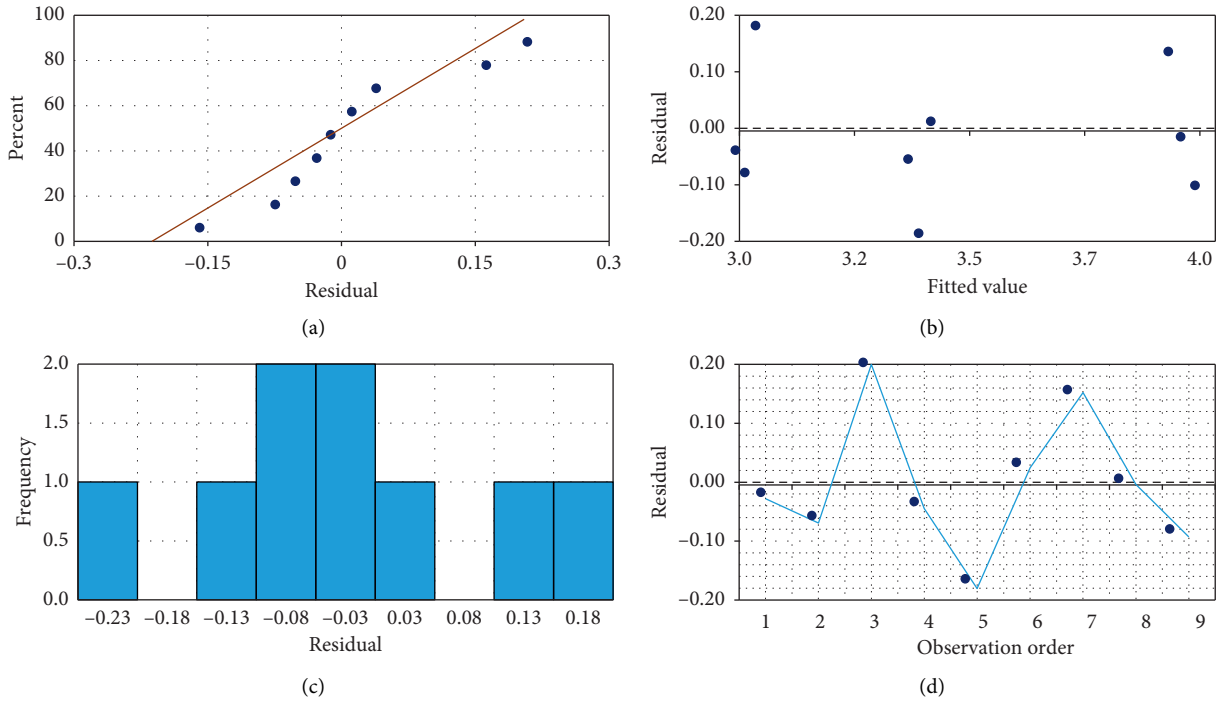


Figure 7: Residual plots of the drilling operation: (a) normal probability plot; (b) versus fits; (c) histogram; (d) versus order.

Table 10: Variance analysis.

Source	DF	Adj. SS	Adj. MS	F value	p value
Regression	2	1.47609	0.73805	47.09	0.00001
v_c	1	1.46422	1.46422	93.42	0.00001
f_n	1	0.01187	0.01187	0.76	0.418
Error	6	0.09404	0.01567		
Total	8	1.57013			

R^2 (pred): 93.45%.

different values in tests 3 and 7. For example, the real value in test 3 is 3.23 kW, in contrast to 3.03 kW obtained with the linear predictive equation and 3.07 kW with the nonlinear equation.

5. Discussion

Tool path strategies and cutting parameters are important factors in the power and energy behavior. It is also known that the maximum power demanded by a CNC machine is tare power [7], but it is not constant. However, if the tool path does not change, the spindle speed and the feed rate will vary depending on cutting parameters. This could be modelled with linear and nonlinear equations, in the idle CNC machine tool condition. The work reported in this paper considered the dynamic tool path strategy for facing as a basis for developing a model because its power demand was inferior to the other two milling strategies (Table 5). There was no representative difference of the surface finish obtained with each strategy, as displayed in Table 4; along the X-axis, R_a varied up to 0.02 μm and up to 0.04 μm, in the Y-direction.

As the tool path was constant in facing, the energy performance depended on the spindle speed and feed rate. Tare power was modelled with the dynamic tool path. However, the same procedure reported in this paper can be used with other machining processes and different tool paths to define the appropriate model for tare power. A nonlinear equation was used to represent the tare power as a function of cutting parameters because of the comparison of the results obtained with linear and nonlinear models reported in Figures 5 and 7.

The data collected through different experiments allowed the authors to choose cutting parameters to achieve workpiece design specifications and, at the same time, take into account low energy consumption operations. According to the data collected using different tool paths for facing, the dynamic tool path reported 1.52 kW and 0.121 kWh, the zigzag 1.55 kW and 0.154 kWh, and the one-way tool path reported 1.99 kW and 0.136 kWh. This behavior was determinant to select the results of the dynamic tool path strategy to define the model for the operation because of its low power demand.

From the face milling 27 experiments, the cutting parameters of the three tests (A, B, and C) carried out with the dynamic toolpath with the lowest power demand and energy consumption were as follows: A—$a_p = 1$ mm, $v_f = 350$ m/min, and $f_z = 0.20$ mm per tooth; B—$a_p = 1.5$ mm, $v_f = 350$ m/min, and $f_z = 0.20$ mm per tooth; and $\mathbf{C}$—$a_p = 2$ mm, $v_f = 350$ m/min, and $f_z = 0.20$ mm per tooth. The power, energy, and roughness surface were $A = 2.59$ kW, 0.030 kWh, and 0.085 μm; $B = 3.21$ kW, 0.038 kWh, and 0.56 μm; and $C = 3.87$ kW, 0.046 kWh, and 0.66 μm. Based on these results, the values for v_f and f_z could be selected,

Figure 8: Drilling power real value compared against the results obtained with the nonlinear and linear predictive models.

and there are two options for a_p: the lowest power and energy consumption; therefore, the lowest cycle time was obtained with option *A*, but the best quality surface was produced with option *C*. So, if the surface finish achieved with *A* satisfies the workpiece requirements, this is the best option. Still, *C* is also adequate.

Drilling was tested with coolant applied to the tool, especially for removing the chip. In addition, the highest variation in ratio was 0.08 mm. However, drill counterbore required the highest value in power with 2.9 kW, reaching 3.9 kW, and on the contrary, this operation showed the lowest energy value with 0.009 kWh until reaching 0.011 kWh. Besides, the same behavior observed in the drill counterbore operations was found using the chip break strategy. Power values ranged from 2.8 kW to 3.2 kW and energy from 0.01 kWh to 0.014 kWh. The results above-mentioned showed that, in the case of drill counterbore, the variation between the lowest and the highest power was estimated in 0.97 kW and the figure for energy was estimated in 0.002 kWh. When drilling with chip break, the power was 0.376 kW and the energy 0.004 kWh. The power variation was an important measurement in this operation, especially because this behavior is a consequence of all power peaks that take place during machining.

6. Conclusions

Based on the comprehensive outcome of our investigation, the following conclusions can be drawn:

(1) The approach used in this research in terms of cutting parameters as demonstrated by the nonlinear power model is validated by the fact that the p value for the lack-of-fit test is 0.623 (Table 4), which means that this value is larger than the significance level of 0.05 or the linear model with R^2 (pred) 98.36 (Table 7).

(2) The power demand and energy consumption differences of the facing and drilling operations for the workpiece analyzed varied from 2 to 4 kW and 0.125 kWh to 0.011 kWh. In order to achieve a good balance of consumption, i.e., without big differences or peaks, it is necessary to carefully select the tool-path trajectory, cutting parameters, and cutting tools

for each machining operation required by a workpiece.

(3) Different behaviors in energy consumption derived from choosing one or another toolpath available in commercial software. The values found for dynamic toolpaths (tare power) are 1.21 kWh, as a minimum, to 1.54 kWh, as a maximum. This could represent an optimization value of 0.21% (Table 5). Moreover, regarding the behaviors in energy consumption shown when using one or another cutting parameters, the values found for dynamic toolpaths are 0.030 kWh, as a minimum, to 0.046 kWh, as a maximum (Discussion). This could mean an optimization value of 0.35%.

Further research will focus to find a correct energy balance, using equation prediction by varying the cutting parameters. More studies are necessary in order to better fit the quality of surface with energy consumption.

Abbreviations

a_p:	Depth of cut
DOEs:	Design of experiments
f_n:	Feed per revolution
f_z:	Feed per tooth
kW:	Kilowatts
kWh:	Kilowatts hour
n:	Spindle speed
R_a:	Roughness average
rpm:	Revolutions per minute
v_c:	Cutting speed
v_f:	Feed rate
ANOVA R^2 (pred):	Analysis of variance R^2, predicted.

Data Availability

The nature of the data obtained in the present investigation are of the experimental type. The equations presented in the article are of the own authorship of all those who participated in the writing of the same and come from the result of studying the theories related to the design of experiments. Addition information can be made available from the correspondence author (xavier_snk@hotmail.com) or the principal author (gminquiz@yahoo.com) upon request. There is no restriction to access such data.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors thank the support provided by Sandvik Coromant S.A. de C.V. (México) to carry out the experiments; in particular, the collaboration of Pablo Flores is appreciated. The research reported in this paper was sponsored by the Universidad Nacional Autónoma de México (Grant nos. UNAM-DGAPA-PAPIIT IT101816 and 101718), Consejo Nacional de Ciencia y Tecnología (Grant 288411 and support: 355272), and Project SEP-CONACyT-Basic Science (no. 83239).

References

[1] U. S. Administration, "Energy information," *Monthly Energy Review Journal*, vol. 35, no. 5, 2017.

[2] T. N. Nghiep, A. Sarhan, and A. A. D. Hideki, "Analysis of tool deflection errors in precision CNC end milling of aerospace aluminum 6061-T6 alloy," *Measurement*, vol. 125, pp. 476–495, 2018.

[3] L. Aoyama, *Strengthened Manufacturing Research at NAMRC 45*, SME, Dearborn, MI, USA, 2017, https://www.sme.org/technologies/articles/2017/may/strengthened-manufacturing-research-at-namrc-45/.

[4] T. Gutowski, J. Dahmus, and A. Thiriez, "Electrical energy requirements for manufacturing processes," in *Proceedings of the 13th CIRP International Conference of Life Cycle Engineering*, Leuven, Belguim, May 2006.

[5] Y. Guo, J. Loenders, J. Duflou, and B. Lauwers, "Optimization of energy consumption and surface quality in finish turning," *Procedia CIRP*, vol. 1, pp. 512–517, 2012.

[6] P. T. Mativenga and M. F. Rajemi, "Calculation of optimum cutting parameters based on minimum energy footprint," *CIRP Annals*, vol. 60, no. 1, pp. 149–152, 2011.

[7] N. Diaz, S. Choi, M. Helu et al., *Machine Tool Design and Operation Strategies for Green Manufacturing*, Laboratory for Manufacturing and Sustainability UC Berkeley, Berkeley, California, 2010.

[8] A. Rangarajan and D. Dornfeld, "Efficient tool paths and Part Orientation for face milling," *CIRP Annals*, vol. 53, no. 1, pp. 73–76, 2004.

[9] N. Diaz, M. Helu, and D. Dornfeld, "Design and operation strategies for green machine tool development," *Laboratory for Manufacturing and Sustainability*, Berkeley, CA, USA, 2010.

[10] K. Erkorkmaz, S. E. Layegh, I. Lazoglu, and H. Erdim, "Feedrate optimization for freeform milling considering constraints from the feed drive system and process mechanics," *CIRP Annals*, vol. 62, no. 1, pp. 395–398, 2013.

[11] A. Aramcharoen and P. T. Mativenga, "Critical factors in energy demand modelling for CNC milling and impact of toolpath strategy," *Journal of Cleaner Production*, vol. 78, pp. 63–74, 2014.

[12] G. M. Minquiz, V. Borja, M. Lopez-Parra, A. C. Ramírez-Reivich, M. A. Domínguez, and A. Alcaide, "A comparative study of CNC part programming addressing energy consumption and productivity," in *Proceedings of the 6th CIRP International Conference on High Performance Cutting*, pp. 581–586, Springer, Berlin, June 2014.

[13] R. Lmalghan, M. C. K. Rao, S. ArunKumar, S. S. Rao, and M. A. Herbert, "Machining parameters optimization of AA6061 using response surface methodology and Particle Swarm optimization," *International Journal of Precision Engineering and Manufacturing*, vol. 19, no. 5, pp. 695–704, 2018.

[14] X. Luan, S. Zhang, and G. Li, "Modified power prediction model based on infinitesimal cutting force during face milling process," *International Journal of Precision Engineering and Manufacturing-Green Technology*, vol. 5, no. 1, pp. 71–80, 2018.

[15] D. Y. Pimenov, "Mathematical modeling of power spent in face milling taking into consideration tool wear," *Journal of Friction and Wear*, vol. 36, no. 1, pp. 45–48, 2015.

[16] D. Y. Pimenov, A. Bustillo, and T. Mikolajczyk, "Artificial intelligence for automatic prediction of required surface roughness by monitoring wear on face mill teeth," *Journal of Intelligent Manufacturing*, vol. 29, no. 5, pp. 1045–1061, 2018.

[17] N. Kumar Sahu and A. B. Andhare, "Modelling and multi-objective optimization for productivity improvement in high speed milling of Ti–6Al–4V using RSM and GA," *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, vol. 39, no. 39, pp. 5069–5085, 2017.

[18] L. Peña-Parás, D. Maldonado Cortés, M. Rodriguez-Villalobos, A. G. Romero Cantu, and O. E. Montemayor, "Enhancing tool life, and reducing power consumption and surface roughness in milling processes by nanolubricants and laser surface texturing," *Journal of Cleaner Production*, vol. 253, p. 2020.

[19] T. Behrendt, A. Zein, and S. Min, "Development of an energy consumption monitoring procedure for machine tools," *CIRP Annals*, vol. 61, no. 1, pp. 43–46, 2012.

[20] M. Helu, J. Rühl, D. Dornfeld, P. Werner, and G. Lanza, "Evaluating trade-offs between sustainability, performance, and cost of green machining technologies," in *Proceedings of the 18th CIRP International Conference on Life Cycle Engineering*, Braunschweig, Germany, May 2011.

[21] L. Li, C. Li, Y. Tang, and Q. Yi, "Influence factors and operational strategies for energy efficiency improvement of CNC machining," *Journal of Cleaner Production*, vol. 161, pp. 220–238, 2017.

[22] L. Wen, *Efficiency of Manufacturing Processes*, Springer, Berlin, Germany, 2015.

[23] S. Kalpakjian and S. Schmid, *Manufactura, Ingeniería Y Tecnología*, Pearson Education, London, UK, 2006.

[24] G. Quintana, J. Ciurana, and J. Ribatallada, "Modelling power consumption in ball-end milling operations," *Materials and Manufacturing Processes*, vol. 26, pp. 746–756, 2013.

[25] Z. H. Wang, J. T. Yuan, T. T. Liu, J. Huang, and L. Qiao, "Study on surface roughness in high-speed milling of AlMn1Cu using factorial design and partial least square regression," *The International Journal of Advanced Manufacturing Technology*, vol. 76, no. 9–12, pp. 1783–1792, 2014.

[26] X. D. Fang and H. Safi-Jahanshahi, "A new algorithm for developing a reference-based model for predicting surface roughness in finish machining of steels," *International Journal of Production Research*, vol. 35, no. 1, pp. 179–199, 1997.

[27] H. Ganesan and G. Mohankumar, "Optimization of machining techniques in CNC turning centre using genetic algorithm," *Arabian Journal for Science and Engineering*, vol. 38, no. 6, pp. 1529–1538, 2013.

[28] F. Draganescu, M. Gheorghe, and C. V. Doicin, "Models of machine tool efficiency and specific consumed energy," *Journal of Materials Processing Technology*, vol. 141, no. 1, pp. 9–15, 2003.

[29] J. Taneja, M. Bector, and R. R. Kumar, "Application of Taguchi method for optimizing turning process by the effects of machining parameters," *International Journal of Engineering and Advanced Technology*, vol. 2, 2012.

[30] W. M. Ponnala and K. L. N. Murthy, "Modeling and optimization of end milling machining process," *International Journal of Research in Engineering and Technology*, vol. 1, 2012.

[31] D. C. Montgomery and G. C. Runge, *Applied Statistics and Probability for Engineers, United States of America*, John Wiley & Sons, Hoboken, NJ, USA, 2003.

[32] Sandvik, *Rotating Tools Milling*, Sandvik, Stockholm, Swede, 2012.

[33] ISO 1302:2002, *Geometrical Product Specifications (GPS)—Indication of Surface Texture in Technical Product Documentation*, ISO 1302:2002, Los Angeles, CA, USA, 2012.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 6483194, 11 pages
https://doi.org/10.1155/2020/6483194

Research Article

Generation-Load Coordinative Scheduling considering the Demand-Response Uncertainty of Inverter Air Conditioners

Wei Hu,[1] Jin Yang,[2] Yi Wu,[1] Weiguo Zhang,[3] Xueming Li,[3] Xiaorong Li,[1] and Ciwei Gao [4]

[1]*Power Dispatch Center, State Grid Jiangsu Electric Power Company, Nanjing 211106, China*
[2]*Jiangsu Xingli Construction Group Company, Nanjing 210000, China*
[3]*Power Consumption Department, Nari Technology Company, Nanjing 211106, China*
[4]*School of Electrical Engineering, Southeast University, Nanjing 210096, China*

Correspondence should be addressed to Ciwei Gao; ciwei.gao@seu.edu.cn

Received 31 May 2020; Accepted 22 June 2020; Published 26 July 2020

Guest Editor: Sanghyuk Lee

Inverter air conditioners (IACs) have gradually become the mainstream of resident air-conditioning equipment. Similar to traditional fixed-frequency air conditioners, IACs have the potential for demand response and load scheduling. However, the uncertainty of IACs is nonnegligible in generation-load scheduling. In this paper, the uncertain demand-response cost of IACs is studied for the first time. Meanwhile, based on the cost, a generation-load coordinative day-ahead scheduling model is proposed. In the scheduling, an IACs aggregator and traditional generators are coordinately dispatched to minimize the expected scheduling cost of the power system. The case study shows that the coordinative scheduling model can reduce the scheduling cost of the power system and encourage the IACs aggregator to improve their responsiveness or reduce their uncertainty, so as to improve the economy and reliability of power scheduling.

1. Introduction

Increasing renewable energy generation [1, 2] and load [3] have led to difficulty in power dispatch, maintenance, and safe operation. To address the problem, more reserve capacity and generating capacity are invested, however, which lack reliability in the near term and economy in the long term [4]. In recent years, air-conditioning load is gradually increasing, and it becomes a considerable component of the maximum load. Particularly, in the summer of East China, the proportion of air-conditioning load in the maximum load can reach 50% [5]. From the perspective of generation and load scheduling, air-conditioning load can adjust demand and provide demand response to schedulers with appropriate control strategies [6, 7]. For example, air-conditioning load can cut peak and fill valley [8], provide system backup [9], and stabilize the fluctuation of renewable energy output [10]. In summary, it is a kind of cost-effective resources suitable for system operators scheduling.

In recent years, load scheduling is gradually incorporated into the day-ahead economic dispatch of power system. A residential load-scheduling approach is proposed to offset the uncertainty caused by renewable generation in [11]. Sohet et al. [12] design an incentive mechanism to foster engagement of electric vehicles in demand response and load scheduling. An optimal day-ahead scheduling of power-to-gas (PtG) storage and gas load management is proposed in [13], where PtG is regarded as a schedulable resource. In literature, load-scheduling optimization is used for the economic evaluation of demand-response load. Since air conditioners have the great potential to offer demand-response service, some studies have proposed the methodology of scheduling optimization for them [14–16]. However, inverter air conditioners are rarely incorporated into the day-ahead scheduling of power system as a kind of schedulable load or response-demand resource, where the control strategy, controllable duration, and uncertainty should be all considered in it.

Dynamic performance, stability, sensitivity, and robustness are intensively studied in [17]; IACs are proved to be suitable for providing regulation capacity to power systems. A price-responsive model-based optimal demand-response control strategy is proposed in [18] for IACs to reduce electricity costs

and the peak power demands. Although the modeling and control strategies of IACs are intensively studied in the literature, the uncertainty in demand response of IACs is ignored, which is harmful to the stability of power system or causes more expensive cost in purchasing reserve.

Meanwhile, the modeling, control strategies, and reserve optimization of inverter air conditioners are studies for demand response in [19, 20] and for frequency regulation in [21–23]. Although, in the literature, the feasibility and technologies of inverter air conditioners to provide demand-response service have been fully studied, the economic evaluation and load-scheduling methodology are still lack of consideration, especially when the uncertainty is regarded as a kind of risk cost for the scheduler. Hence, the uncertain-risk cost should be modeled and incorporated into the scheduling optimization of inverter air conditioners for power system.

To address the above problems, this paper proposes a risk evaluation-based generation-load collaborative scheduling optimization model, where the objective function is to minimize the scheduling cost of power system and the solution is subject to the operational reliability of power system. The contributions are as follows:

(1) From the perspective of the scheduler, this paper proposes a payment cost model of IACs considering the uncertainty in demand response for the first time, which includes the payment cost of demand response and uncertain-risk cost. It can be applied to evaluate the scheduling cost of IACs with the scheduling potential of different degrees. This provides methodology of economic evaluation for generation-load coordinative scheduling.

(2) Based on the payment cost model of IACs, a generation-load coordinative scheduling optimization is also formulated, where the operational reliability, controllable duration, and actual response are all taken into account, so as to efficiently evaluate the scheduling cost of power system and keep as close as possible to optimal solution.

2. Load-Scheduling Modeling of IACs Aggregator considering Demand-Response Uncertainty

2.1. Stochastic Modeling of IAC. When an IAC is turned on, the relationship between the temperature of the room and the cooling capacity is given in

$$\begin{aligned} T_{\text{in}}[(n+1)\Delta t] &= e^{-\Delta t/\text{RC}} T_{\text{in}}(n\Delta t) \\ &\quad + \left(1 - e^{-\Delta t/\text{RC}}\right)\left[T_{\text{out}}(n\Delta t) - Q_{\text{AC}}(n\Delta t)R\right] \\ &\quad + V(n\Delta t), \end{aligned} \tag{1}$$

where $T_{\text{in}}(n\Delta t)$ and $T_{\text{out}}(n\Delta t)$ denote the indoor and outdoor temperature (°C) at $n\Delta t$; $Q(n\Delta t)$ denotes the cooling capacity (kW) of the IAC at $n\Delta t$; Δt denotes the temperature-monitoring interval (s) of the IAC; R is the equivalent thermal resistance (Ω) of the room; C is the equivalent heat capacity (J/°C) of the room; $V(n\Delta t)$ denotes the random temperature deviation caused by the influence of outdoor airflow movement/sunlight exposure, window opening and closing, indoor personnel flow, and other factors.

During the operation of IACs, the cold/hot air intake is controlled by adjusting the operating frequency of compressor, so as to continuously control the power. The relationship between the operating frequency, cooling capacity, and electric power of an IAC is shown in

$$Q_{\text{AC}}(n\Delta t) = \alpha f(n\Delta t) + \beta, \tag{2}$$

$$P_{\text{AC}}(n\Delta t) = \gamma f(n\Delta t) + \delta, \tag{3}$$

where $P_{\text{AC}}(n\Delta t)$ denotes the cooling power (kW) of the IAC at $n\Delta t$; $f(n\Delta t)$ denotes the operating frequency of the compressor (Hz) at $n\Delta t$; α and β denote the constant coefficient of the cooling capacity; γ and δ denote the constant coefficient of the active power.

2.2. Modeling of Demand-Response Uncertainty for IACs Aggregator. In this paper, we adopt a statistical methodology to evaluate expected uncertain-risk cost of IACs-load scheduling. Also, the confidence level is considered in risk-related constraints. To apply the methodology to IACs-load scheduling, we consider stochastic processes in two aspects: one is the responsive deviation of signal IAC; the other is the responsive deviation of IACs aggregator. The detailed methodology and application are as follows: we derive demand-response potential (the expected responsive power) and risk (the distribution and variance of responsive power) of single IAC from the modeling as presented in formulas (1)–(9); then the stochastic uncertain-risk cost of IACs aggregators is deduced as presented in formulas (10)–(14).

Given that the outdoor temperature T_{out} is unchanged during the minimum scheduling period and the IAC operates under the setting temperature T_{set} before load scheduling, the load power P_{AC} can be formulated, as expressed in formula (4); when the scheduling instruction is issued, taking load reduction as an example, the IAC will operate under the lowest frequency $f_{\min}$; the load power P_{AC} can be formulated, as expressed in formula (5); when the indoor temperature gradually rises to the upper bound $T_{\max}$ beyond the range of comfortable temperature for users, the scheduler will release the control and readjust the frequency of the IAC to the f_{set} corresponding to T_{set}. The adjustment process is presented in Figure 1:

$$\begin{aligned} &P_{\text{AC}}\left(T_{\text{set}}, T_{\text{out}}, V(n\Delta t)\right) \\ &\quad = \gamma\left(\frac{T_{\text{out}} - T_{\text{set}} + \left(V(n\Delta t)/\left(1 - e^{-\Delta t/\text{RC}}\right)\right)}{\alpha R} - \frac{\beta}{\alpha}\right) + \delta, \end{aligned} \tag{4}$$

$$P_{\text{AC}}(f_{\min}) = \gamma \cdot f_{\min} + \delta. \tag{5}$$

Under the control mode, the controllable duration t_c of an IAC can be expressed in

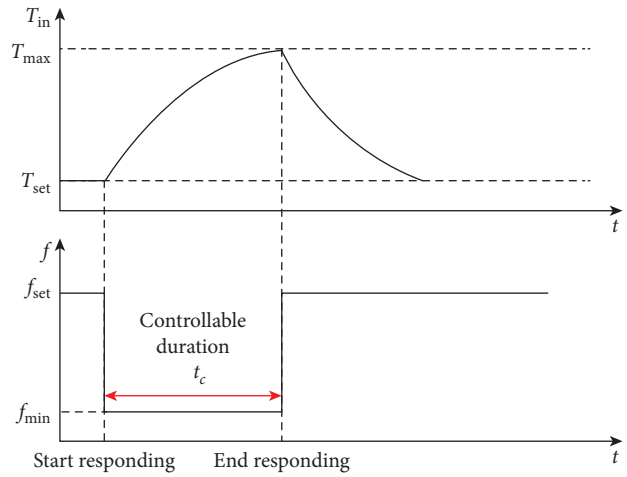


Figure 1: The load-scheduling process of a IAC.

$$t_c = \mathrm{RC} \cdot \ln \frac{T_{out} - T_{set} - RQ_{min} + V(n\Delta t)}{T_{out} - T_{max} - RQ_{min}}. \quad (6)$$

The load reduction or reserve capacity of an IAC can be expressed in

$$\Delta P_{AC} = P_{AC}(T_{set}, T_{out}, V(n\Delta t)) - P_{AC}(f_{min}) = \gamma \cdot \left(\frac{T_{out} - T_{set}}{\alpha R} - \frac{\beta}{\alpha}\right) - \gamma \cdot f_{min} + \frac{\gamma \cdot V(n\Delta t)}{\alpha R(1 - e^{-\Delta t/RC})}. \quad (7)$$

$\gamma \cdot V(n\Delta t)/\alpha R(1 - e^{-\Delta t/RC})$ is a stochastic function, denoting responsive deviation of the IAC, denoted as $\xi(n\Delta t)$:

$$\xi(n\Delta t) = \frac{\gamma}{\alpha R(1 - e^{-\Delta t/RC})} V(n\Delta t). \quad (8)$$

In this paper, the normal distribution is used to evaluate the stochastic responsive deviation of a single IAC.

For an aggregator with N IACs, the mean value ξ_i of each IAC is 0, and the variance of each IAC is denoted by $\sigma_{\xi,i}(i = 1, 2, \cdots, N)$. Then, the mean responsive deviation of the aggregator is also 0; the variance is expressed by

$$\sigma = \sqrt{\sum_{i=1}^{N} \sigma_{\xi,i}^2}. \quad (9)$$

Here, the demand-response-implementation mechanism is introduced as follows. Before participating in load scheduling, an IACs aggregator will contract with IAC users to control over them during load-peak periods. Only during the periods, when the aggregators receive demand-response (DR) signals, they can control the power of the IACs. After DR, the aggregators should pay to the users or give a discount on electricity fee according to the contracts. Moreover, to evaluate the DR potential of IACs, IACs should update and report their operational status to the aggregator every scheduling period.

3. Stochastic Cost of IACs Aggregator for Load Scheduling

3.1. Payment Cost. In the traditional load-scheduling contract between the system operator and users, after the air-conditioning load responds to the scheduling instruction, the system operator needs to pay a certain fee to compensate the load according to the actual response of the load. The responsive deviation of IACs aggregator will affect the final clearing expenses; hence, the uncertainty of the payment cost should be also considered by the scheduler. The payment cost C_{pay} of lth IACs aggregator at tth scheduling period consists of two parts: one is compensation fee paid to users according to the actual response; the other is the penalty fee caused by the responsive uncertainty of the aggregator, which is revenue for schedulers, as expressed in

$$C_{pay}(l,t) = P_{real}(l,t) \cdot v_c(l,t) + \xi_l \cdot v_p(l,t), \quad (10)$$

where v_c and v_p denote the capacity price and penalty price for IACs aggregator, respectively; P_{real} denotes the actual responsive power of the IACs aggregator, which is the difference between expected load reduction P_{de} and responsive deviation, as expressed in formula (11). ξ_l denotes the mean responsive deviation of lth IACs aggregator:

$$P_{real}(l,t) = P_{de}(l,t) + \xi_l. \quad (11)$$

Take (11) into (10) and derive the following formula:

$$C_{pay}(l,t) = P_{de}(l,t) \cdot v_c(l,t) + \xi_l \cdot \left[v_c(l,t) + v_p(l,t)\right], \quad (12)$$

where $P_{de}(l,t) \cdot v_c(l,t)$ is denoted by C_{fix} and $\xi_l \cdot [v_c(l,t) + v_p(l,t)]$ is denoted by C_ξ. Then, the expected payment cost C_{pay} of lth IACs aggregator at tth scheduling period is expressed by

$$E\left[C_{pay}(l,t)\right] = C_{fix}(l,t) + E\left[C_\xi(l,t)\right], \quad (13)$$

where $E(C_\xi)$ can be expressed by

$$E\left[C_\xi(l,t)\right] = \int_{-\infty}^{-P_{de}(l,t)} \left[\xi_l \cdot v_p(l,t) - P_{de}(l,t) \cdot v_c(l,t)\right] \cdot f(\xi_l) d\xi_l + \int_{-P_{de}(l,t)}^{0} \xi_l \cdot \left[v_c(l,t) + v_p(l,t)\right] \cdot f(\xi_l) d\xi_l, \quad (14)$$

where $f(\xi_l)$ denotes the probability density function of responsive deviation of lth IACs aggregator.

3.2. Uncertain-Risk Cost. There is always a deviation between the actual response and the expected response of IACs aggregators, which brings a certain risk and corresponding cost to the scheduler. When the responsive power is less than the expected responsive power, the system may not be able to supply enough power to users, resulting in power failure. To supply enough power and maintain supply-demand balance, schedulers need to bear the payment risk of direct load control (DLC). Hence, power-failure loss can be evaluated by the compensation cost of DLC; when the aggregator

overresponds, the power supply is greater than the demand, and the power grid needs to bear the loss of reduced electricity sales. In summary, for a scheduler, the uncertain-risk cost C_{risk} caused by the responsive uncertainty of lth IACs aggregator at tth scheduling period can be expressed by

$$C_{\text{risk}}(l,t) = \int_{-\infty}^{0} -\xi_l \cdot v_{\text{re}}(l,t) \cdot f(\xi_l) d\xi_l + \int_{0}^{+\infty} \xi_l \cdot v_{\text{sell}}(l,t) \cdot f(\xi_l) d\xi_l, \tag{15}$$

where v_{re} denotes the power-failure loss per capacity (RMB/kWh); v_{sell} denotes the loss of reduced electricity sales per capacity (RMB/kWh).

4. Generation-Load Coordinated Scheduling Optimization

Figure 2 illustrates the generation-load coordinated scheduling steps in this paper. First of all, before day-ahead scheduling, IACs aggregators should submit demand-response parameters to scheduler, such as bidding price $v_c(l,t)$ (capacity price), mean responsive deviation ξ_l, and the probability density function of responsive deviation $f(\xi_l)$, which are derived according to formulas (1)–(9). At the same time, traditional generators also submit generating parameters to the scheduler. Secondly, the optimal scheduling plan will be calculated according to formulas (16)–(27) and then be issued to the generators and IACs aggregators. Finally, the generators and IACs aggregators will execute the scheduling plan. Finally, scheduler will conduct settlement according to actual generating capacity of generators and responsive deviation of IACs aggregators.

4.1. Objective Function. The objective function of the generation-load coordinated scheduling optimization is to minimize the expected generation cost and load-scheduling cost:

$$F = \min \sum_{t=1}^{T} \sum_{j=1}^{N_G} C_{G_j,t}(P_G(j,t)) + E \sum_{t=1}^{T} \sum_{l=1}^{N_L} [C_{\text{total}}(l,t)], \tag{16}$$

where T is the number of scheduling periods; N_G and N_L denote the number of generators and IACs aggregators, respectively; $C_{Gj,t}(P_G(j,t))$ is the generation-cost function of jth generator at tth period, which is presented in

$$C_{Gj,t}(P_G(j,t)) = a_j P_G(j,t)^2 + b_j P_G(j,t) + c_j s_G(j,t), \tag{17}$$

where $P_G(j,t)$ denotes the power of jth generator at tth period; $S_G(j,t)$ denotes the operational status of jth generator at tth period, which is equal to 0 when it is off and 1 when it is on; a_j, b_j, and c_j denote the parameters of the generation-cost function of jth generator.

The scheduling cost C_{total} of IACs aggregators consists of the payment cost and the uncertain-risk cost, as presented in

$$C_{\text{total}}(l,t) = C_{\text{fix}}(l,t) + C_{\xi}(l,t) + C_{\text{risk}}(l,t). \tag{18}$$

Above all, the objective function can be simplified as expressed by

$$F = \min \sum_{t=1}^{T} \sum_{j=1}^{N_G} C_{Gj,t}(P_G(j,t)) + \sum_{t=1}^{T} \sum_{l=1}^{N_L} C_{\text{fix}}(l,t) + E\left[C_{\xi}(l,t)\right] + C_{\text{risk}}(l,t). \tag{19}$$

The risk factor λ is introduced into formula (19), denoting the sensitivity of the scheduler to the risk. The objective function can be expressed as follows:

$$F = \min \sum_{t=1}^{T} \sum_{j=1}^{N_G} C_{Gj,t}(P_G(j,t)) + \sum_{t=1}^{T} \sum_{l=1}^{N_L} (1-\lambda) C_{\text{fix}}(l,t) + \lambda\left\{E\left[C_{\xi}(l,t)\right] + C_{\text{risk}}(l,t)\right\}. \tag{20}$$

4.2. Constraints

4.2.1. Power-Balance Constraint. Generation should be equal to consumption at each scheduling period:

$$\sum_{j=1}^{N_G} P_G(j,t) + \sum_{l=1}^{N_L} P_{\text{de}}(l,t) = P_D(t), \tag{21}$$

where $P_D(t)$ denotes forecasted load at t^{th} period.

4.2.2. Constraints of Generators

(1) Generation capacity constraint is as follows:

$$P_{Gj,\min} \le P_G(j,t) \le P_{Gj,\max}, \tag{22}$$

where $P_{GJ,\min}$ and $P_{GJ,\max}$ denote the upper and lower bound of jth generating power, respectively.

(2) Ramping constraint is as follows:

$$-D_{Rj} \le P_G(j,t) - P_G(j,t-1) \le U_{Rj}, \tag{23}$$

where U_{Rj} and D_{Rj} denote the maximum ramp-up and ramp-down rate of jth generator, respectively.

4.2.3. Constraints of IACs Aggregator

(1) Responsive power constraint is as follows:

$$0 \le P_{\text{de}}(l,t) \le P_{l,\max}, \tag{24}$$

where $P_{l,\max}$ denotes the maximum responsive power of lth aggregator.

(2) Controllable duration constraint is as follows:

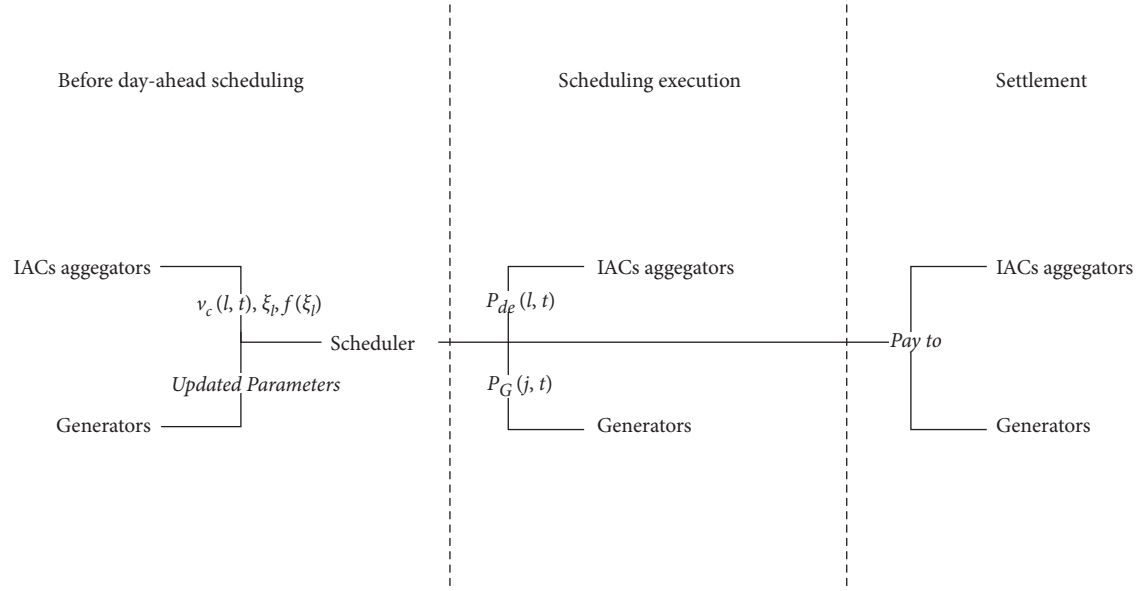


Figure 2: Coordinated scheduling steps.

$$t_{\text{control}}(l,t) \geq \Delta T, \tag{25}$$

where $t_{\text{control}}(l,t)$ denotes the controllable duration of lth aggregator during tth scheduling period; ΔT denotes the scheduling interval, which is 30 min in this paper.

(3) Responsive stability constraint is as follows:

To ensure the stability of system operation, chance constraint method is adopted to constrain the responsive stability of IACs aggregators. Namely, the probability that the responsive-deviation rate of IACs aggregator is less than or equal to a set value ε^* is not less than β:

$$P_r\left[\varepsilon_l \leq \varepsilon^*\right] \geq \beta, \tag{26}$$

where ε_l denotes the responsive-deviation rate of lth aggregator, which is the ratio of the variance of the responsive-deviation power to expected load reduction:

$$\varepsilon_l = \frac{\sigma(\xi_l)}{P_{\text{de}}(l,t)}. \tag{27}$$

Formulas (19)–(27) constitute a generation-load coordinated scheduling optimization model considering the uncertainty of air-conditioning response. The decision variables include the output of the traditional generating units $P_G(j,t)$ and the expected load reduction of IACs aggregators $P_{\text{de}}(j,t)$. The model is a nonlinear stochastic optimization based on chance constraints.

5. Case Study

5.1. Numerical Setting. Taking IACs aggregators participating in system scheduling in a typical high-temperature day of summer as a scenario, the scheduler forecasts the load power of the next day in the region and obtains the daily load prediction curve as shown in Figure 3. It is assumed that there are 3 traditional generating units in this area; see Table 1 for details. The sum of the maximum output of the three units is 2000 kW, which cannot meet the power supply during the peak-load period. There will be a power supply gap between 12:00 and 15:00. In the period of power supply shortage, the traditional generating units and IACs aggregators are jointly scheduled to meet the load requirements.

According to the set temperature value and actual responsive characteristics, the IACs are aggregated into three aggregators, and the specific information of each aggregator is shown in Table 2.

5.2. Potential of Load Reduction. According to the weather forecast, the outdoor temperature of the next day when there is a power supply gap in the area is obtained, as shown in Figure 4. Based on the responsive uncertainty model of IACs aggregator, the load-reduction potential, controllable duration, and the scheduling responsive-deviation distribution of IACs aggregators are calculated, as shown in Table 3.

It can be observed from Table 3 that, in each period of power supply gap, the controllable durations of aggregators A, B, and C are more than 30 minutes, which can cover the duration of each scheduling interval and can meet the comfort experience of IAC users in the process of scheduling.

5.3. Results and Analysis. Based on the generation-load coordinated scheduling optimization model considering the uncertainty of IACs, the output of the traditional generators at each time is optimized, and the results are shown in Figure 5. The overall coordinated scheduling results of IACs aggregators and traditional units are shown in Figure 6.

It can be seen from Figure 6 that the output of traditional generating units is different from each other at all periods,

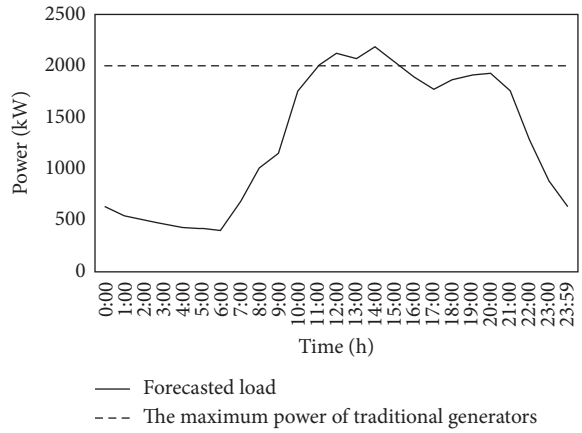


Figure 3: Daily load prediction curve.

Table 1: Temperature and wildlife count in the three areas covered by the study.

Index	Parameters of generation-cost function			Maximum power (kW)
	a_j	b_j	c_j	
1	0.0015	0.03	30	1000
2	0.0021	0.02	60	550
3	0.0038	0.04	45	450

but the relationship between the output of three units is the same at each period (The Output of Generator 1 > The Output of Generator 2 > The Output of Generator 3), which is related to the cost coefficients of each unit. The smaller the cost coefficient, the more the output.

Most of the time, the traditional generating units can meet the load demand of the region. In the peak-load period, the sum of the maximum output of the traditional generating units is less than the forecasted load. In Figure 6, the red shaded part is the power supply gap in the peak-load period. In the power supply gap period, traditional units and aggregators need to be jointly scheduled to ensure the power supply and demand balance. See Table 4 for the total expected load reduction of aggregators in each power supply gap period.

From Table 4, the total load-reduction potential of three aggregators in each power supply gap period is greater than the power supply gap, which provides the possibility for IACs aggregators to participate in grid coordinated scheduling.

5.4. Comparison of Optimal Scheduling under Different Objectives. Two cases with different objective functions are given as follows.

Case 1. To minimize the payment cost Ctotal of IACs aggregators.

Case 2. To minimize the payment cost Cfix of IACs aggregators.

The simulation results, namely, the optimal load-reduction distributions of aggregators under different objectives, are shown in Figure 7. See Table 5 for the cost of aggregators under different objectives.

It can be seen from Figure 7 that, in Case 2, the scheduling sequence of aggregators is arranged according to the user compensation cost at each period, and aggregator with the lowest user compensation cost is called first, while, in Case 1, the user compensation cost and responsive deviation of each aggregator are taken into account for optimizing scheduling. Namely, the smaller the responsive deviation is, the more likely the aggregator will be called.

In Case 2, since the scheduler only considers the user compensation cost of aggregators, the compensation cost (1018 RMB) under the scheduling optimization is lower than the cost in Case 1. It can also be seen from Figure 7 that, in Case 2, the scheduler preferentially calls the aggregator with lower compensation cost. However, the aggregators have higher uncertain-risk cost, which causes the scheduler to bear higher payment cost (1275.7 RMB). Although the compensation cost (1031 RMB) of the aggregators in Case 1 is higher, the scheduler calls the aggregators with less responsive deviation, and the lower uncertain-risk cost makes the payment cost (1132.4 RMB) borne by the scheduler lower than that in Case 2, which proves that generation-load coordinated scheduling optimization model proposed considering the uncertainty of demand response can reflect the scheduling cost of IACs aggregators more reasonably.

5.5. Sensitivity Analysis

5.5.1. Payment Prices. The payment prices of schedulable load in different regions are different from each other. We assume that the payment prices of aggregators B and C are fixed, which are 3 RMB/kWh and 4RMB/kWh, respectively. The payment price of aggregator A ranges from 1.2 to 3.2 RMB/kWh. The simulation results show the total scheduling cost and payment cost of aggregator A, as presented in Figure 8.

It can be observed from Figure 8 that the payment cost and payment cost of aggregator A increase with the increase of user payment price. This is because the payment price of aggregator A is lower than that of aggregator C, while its responsive deviation is between B and C; hence, its scheduling priority is still lower than that of aggregator B. With the increase of user payment price, the load reduction of aggregator A remains unchanged, the expected uncertain-risk cost remains unchanged, and the fixed payment cost increases, so its payment cost and payment cost increase accordingly.

5.5.2. Penalty Prices. Similarly, suppose that the penalty price of aggregators B and C is fixed, which are 15 RMB/kWh and 5 RMB/kWh, respectively. The penalty price of aggregator A ranges from 5 to 30 RMB/kWh. The simulation results show the total scheduling cost and payment cost of aggregator A, as presented in Figure 9.

From Figure 9, it can be observed that the payment cost and payment cost of aggregator A decrease with the increase

Table 2: The parameters of IACs aggregators.

	Aggregator A	Aggregator B	Aggregator C
Minimum operating frequency (Hz)		20	
Upper limit of user comfortable temperature (°C)		28	
Setting temperature (°C)	23	24	25
Numbers of IACs	100	300	200
Payment prices (RMB/kW)	2	3	4
Punishment prices (RMB/kW)	10	15	5
Confidence values of responsive stability constraint		0.9	

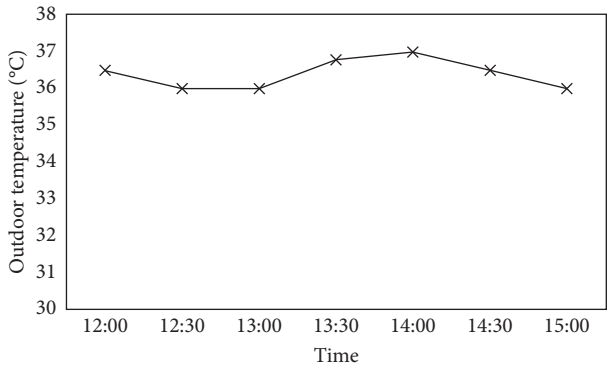


Figure 4: Outdoor temperature during power supply gap.

Table 3: The potential of IACs aggregators.

Time		Aggregator A	Aggregator B	Aggregator C
12:00	The potential of load reduction (kW)	41	90	40
14:30	Controllable duration (min)	42.8	40.3	37.4
	Total load reduction (kW)	171		
12:30	The potential of load reduction (kW)	35	75	28
13:00	Controllable duration (min)	48.6	43.5	38.7
15:00	Total load reduction (kW)	138		
	The potential of load reduction (kW)	44	99	44
13:30	Controllable duration (min)	41.5	39	33
	Total load reduction (kW)	187		
	Potential of load reduction (kW)	46	105	50
14:00	Controllable duration (min)	38.6	34.2	30.2
	Total load reduction (kW)	201		
Variance of responsive deviation of single IAC		0.15	0.06	0.14
Variance of responsive deviation of IACs aggregators		1.5	1.0	2.0

of user penalty price. With the increase of penalty price, the scheduler gradually reduces the load reduction of aggregator A for optimizing cost, so the payment cost and payment cost of aggregator A are reduced.

5.5.3. Confidence of Responsive Stability. The confidence of responsive stability reflects the constraint on responsive deviation of aggregators. The larger the confidence value is, the more strict the constraint on responsive deviation of aggregators is. Suppose that the confidence values of aggregators B and C are all set to 0.9 and the confidence values of aggregator A vary from 0.9 to 0.95. The simulation results show the payment cost and payment cost of aggregator A under different response-stability confidence, as presented in Figure 10.

It can be seen from Figure 10 that, with the increase of the confidence value of the responsive stability of aggregator A, the response of aggregators becomes more stable, the risk of aggregator A is smaller, the uncertain-risk cost is reduced, and the payment cost is thus reduced. At the same time, the penalty cost of aggregator A decreases, and the payment cost of aggregator A increases.

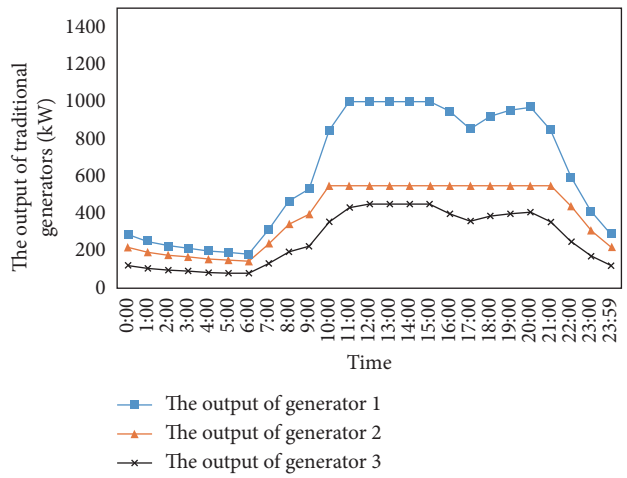


Figure 5: The output of the traditional generators.

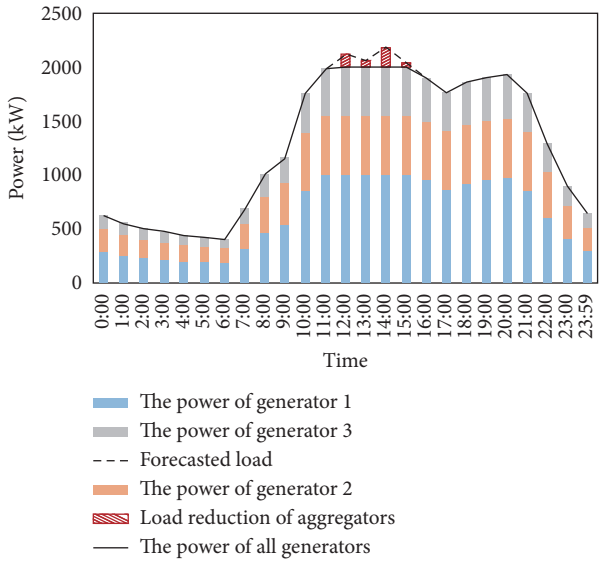


Figure 6: The output of the traditional generators and IACs aggregators.

Due to the influence of payment price, penalty price, and response-stability constraint, the payment cost and payment cost of aggregator A decrease with the increase of penalty price and the decrease of penalty price, while with the increase of response-stability confidence, the payment cost decreases and the payment cost increases.

Based on the generation-load coordinated scheduling optimization model, the grid scheduling department is more inclined to call the air-conditioning users with low payment price, high penalty price, and more reliable demand response for their own interests.

For IACs aggregators, when the payment price and penalty price are bound by the contract between schedulers and schedulable load, improving the reliability and stability of its demand response can increase the revenue of the aggregators. Therefore, according to the scheduling model in this paper, schedulers can reduce the scheduling cost as well as motivating air-conditioning users to improve their responsive stability, so as to improve the economy and reliability of power grid scheduling operation.

5.5.4. Risk Aversion Factors. The risk aversion factor λ in the objective function can represent the degree of the concern of scheduler to the risk caused by the demand-response uncertainty of air-conditioning aggregators. With (20) as the objective function, the payment price, and response-deviation variance of each IACs aggregator in Table 6, as well as risk aversion factor ranging from 0 to 1, we can derive the payment cost of air-conditioning aggregators through simulation, as shown in Figure 10. Obviously, from Figure 11, when the risk aversion factor is set to 0.6, the payment cost is the smallest.

The column in Figure 11 represents the load reduction of each IACs aggregators during the power supply gap period from 13:30 to 14:00. We can observe the following:

(1) When $\lambda = 0$, in order to minimize the total scheduling cost, the scheduler will call up aggregator A and then aggregator B according to the order of small to large payment price. However, the responsive reliability of aggregator A and aggregator B is poor, and the uncertain-risk cost is high, so the payment cost is high in this case.

(2) When λ ranges from 0 to 0.5, with the increase of λ, the scheduler will gradually pay more attention to the uncertain-risk cost of IACs aggregators. Therefore, it will gradually increase the call of IACs aggregators with smaller expected responsive deviation, then the uncertain cost will reduce, and hence the payment cost will gradually decrease.

(3) When $\lambda = 0.6$, the scheduler will preferentially call aggregator B with lower payment price and more reliable demand response, and the payment cost will also reach the minimum at this time.

(4) When λ ranges from 0.6 to 1, the scheduler pays more attention to the risk than the payment price. Meanwhile, with the increase of λ, more and more prefer to call the air-conditioning aggregators with higher responsive reliability. However, due to the fact that the payment price of the aggregators with higher responsive reliability is higher, the payment cost is on the rise with the increase of λ.

Table 4: Total load reduction of aggregators at each time.

Time	Load reduction (kW)
12:00	115
12:30	88
13:00	59
13:30	120
14:00	178
14:30	108
15:00	41

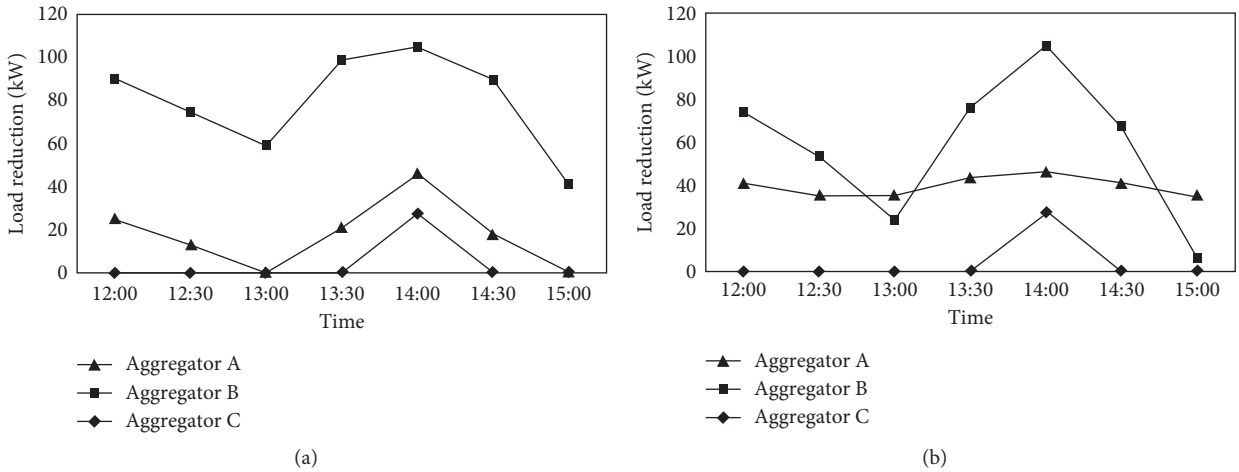


Figure 7: Load-reduction distribution of aggregators in (a) Case 1 and (b) Case 2.

Table 5: The cost of aggregators in both cases.

Payment cost (RMB)	Compensation cost (RMB)	Payment cost (RMB)
Case 1		
1132.4	1031	1062.94
Case 2		
1275.7	1018	1162.22

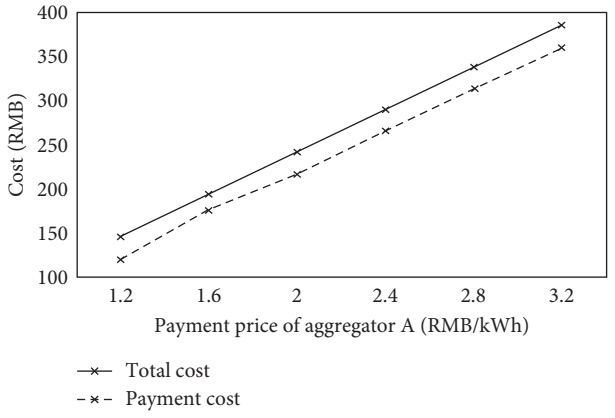


Figure 8: The payment cost and payment cost of aggregator A under different payment price.

(5) When $\lambda = 1$, in order to avoid the risk, the scheduler will call air-conditioning aggregators according to the order of responsive reliability. However, the aggregators with better reliability have higher payment cost; hence the payment cost is not the minimum.

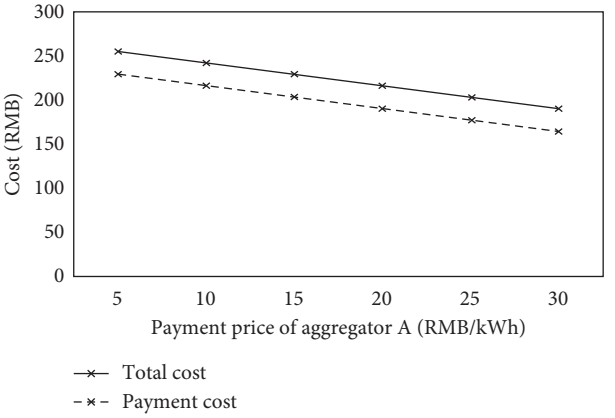


Figure 9: The payment cost and payment cost of aggregator A under different penalty price.

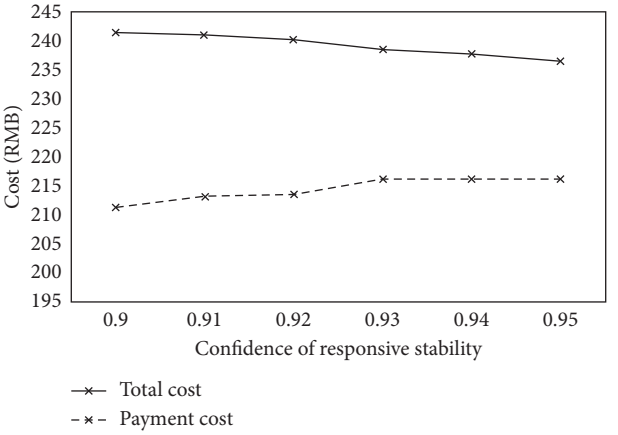


Figure 10: The payment cost and payment cost of aggregator A under different confidence of responsive stability.

Table 6: The numerical setting of payment price and variance of responsive deviation.

	Aggregator A	Aggregator B	Aggregator C
Payment price (RMB/kwh)	2	3	4
Variance of responsive deviation	2.0	1.5	1.0

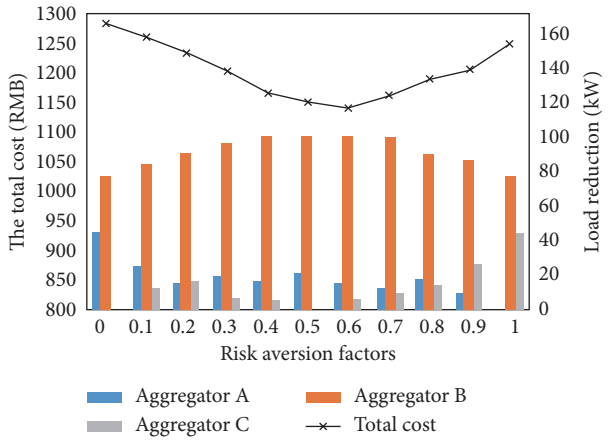


Figure 11: The payment cost and load reduction under different risk aversion factors.

6. Conclusions

Based on the evaluation of uncertain-risk cost and opportunity constraints, this paper proposes a day-ahead generation-load coordinated scheduling optimization model. According to the load forecasting, traditional generating units and IACs aggregators are coordinated to meet the power supply and demand balance in the period of insufficient power supply. Through this study, the following conclusions are derived:

(1) Based on the generation-load coordinated scheduling optimization, the scheduler is more inclined to call those IACs aggregators with low payment price, high penalty price, and more reliable demand response for their own interests. For IACs aggregators, under the condition that the payment price and penalty price are bound by the contract, improving the reliability and stability of demand response can increase the operational economy.

(2) The proposed coordinated scheduling optimization can reduce the scheduling cost of the power grid as well as motivating air-conditioning users to improve their responsive stability, so as to improve the economy and reliability of system operation.

(3) The risk aversion factor is very important to the economic operation of the system. Too large or too small risk factors are not conducive to optimizing the economic operation of the system.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This project was partly funded by the National Science Fundation of China (51577029) and Research Project of Jiangsu Power Company.

References

[1] N. K. Dhaliwal, F. Bouffard, and M. O'Malley, "A fast flexibility-driven generation portfolio planning method for sustainable power systems," *IEEE Transactions on Sustainable Energy, (Early Access)*, 2020.

[2] M. C. Carreton and M. Carrion, "Generation capacity expansion considering reserve provision by wind power units," *IEEE Transactions on Power Systems, (Early Access)*, 2020.

[3] C. Gu, W. Yang, Y. Song, and F. Li, "Distribution network pricing for uncertain load growth using fuzzy set theory," *IEEE Transactions on Smart Grid*, vol. 7, no. 4, pp. 1932–1940, 2016.

[4] S. Pineda and J. M. Morales, "Chronological time-period clustering for optimal capacity expansion planning with storage," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 7162–7170, 2018.

[5] C. Li, J. Shang, S. Zhu, L. Du, C. Wu, and Q. Wang, "An analysis of energy consumption caused by air temperature-affected accumulative effect of the air conditioning load," *Automation of Electric Power System*, vol. 34, no. 20, pp. 30–33, 2010.

[6] N. Mahdavi and J. H. Braslavsky, "Modelling and control of ensembles of variable-speed air conditioning loads for demand response," *IEEE Transactions on Smart Grid (Early Access)*, 2020.

[7] K. Meng, Z. Y. Dong, Z. Xu, Y. Zheng, and D. J. Hill, "Coordinated dispatch of virtual energy storage systems in smart distribution networks for loading management," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 4, pp. 776–786, 2019.

[8] X. Hu and J. Nutaro, "A priority-based control strategy and performance bound for aggregated HVAC-based load shaping," *IEEE Transactions on Smart Grid (Early Access)*, 2020.

[9] W. Cui, Y. Ding, H. Hui et al., "Evaluation and sequential dispatch of operating reserve provided by air conditioners considering lead-lag rebound effect," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 6935–6950, Nov. 2018.

[10] X. Gao, K. Meng, D. Wang, G. Liang, F. Luo, and Z. Y. Dong, "Optimal wind turbine and air conditioner loads control in distribution networks through MILP approach," in *Proceedings of 2016 IEEE Power and Energy Society General Meeting*, IEEE, Boston, MA, USA, pp. 1–5, July 2016.

[11] T. Remani, E. A. Jasmin, and T. P. I. Ahamed, "Residential load scheduling with renewable generation in the smart grid: a reinforcement learning approach," *IEEE Systems Journal*, vol. 13, no. 3, pp. 3283–3294, 2019.

[12] B. Sohet, Y. Hayel, O. Beaude, and A. Jeandin, "Coupled charging-and-driving incentives design for electric vehicles in

urban networks," *IEEE Transactions on Intelligent Transportation Systems (Early Access)*, 2020.

[13] H. Khani and H. E. Z. Farag, "Optimal day-ahead scheduling of power-to-gas energy storage and gas load management in wholesale electricity and gas markets," *IEEE Transactions on Sustainable Energy*, vol. 9, no. 2, pp. 940–951, 2018.

[14] T. Lee, M. Cho, Y. Hsiao, P. Chao, and F. Fang, "Optimization and implementation of a load control scheduler using relaxed dynamic programming for large air conditioner loads," *IEEE Transactions on Power Systems*, vol. 23, no. 2, pp. 691–702, 2008.

[15] F. Luo, J. Zhao, Z. Y. Dong et al., "Optimal dispatch of air conditioner loads in southern China region by direct load control," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 439–450, 2016.

[16] N. Chakraborty, A. Mondal, and S. Mondal, "Intelligent scheduling of thermostatic devices for efficient energy management in smart grid," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 6, pp. 2899–2910, Dec. 2017.

[17] H. Hui, Y. Ding, T. Chen, S. Rahman, and Y. Song, "Dynamic and stability analysis of the power system with the control loop of inverter air conditioners," *IEEE Transactions on Industrial Electronics (Early Access)*, 2020.

[18] M. Hu and F. Xiao, "Price-responsive model-based optimal demand response control of inverter air conditioners using genetic algorithm," *Applied Energy*, vol. 219, pp. 151–164, 2018.

[19] Y. Che, J. Yang, Y. Zhou, Y. Zhao, W. He, and J. Wu, "Demand response from the control of aggregated inverter air conditioners," *IEEE Access*, vol. 7, pp. 88163–88173, 2019.

[20] B. Wang, T. Zhang, X. Hu, Y. Bao, and H. Su, "Consensus control strategy of an inverter air conditioning group for renewable energy integration based on the demand response," *IET Renewable Power Generation*, vol. 12, no. 14, pp. 1633–1639, 2018.

[21] H. Hui, Y. Ding, Z. Lin, P. Siano, and Y. Song, "Capacity allocation and optimal control of inverter air conditioners considering area control error in multi-area power systems," *IEEE Transactions on Power Systems*, vol. 35, no. 1, pp. 332–345, 2020.

[22] H. Hui, Y. Ding, and M. Zheng, "Equivalent modeling of inverter air conditioners for providing frequency regulation service," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 2, pp. 1413–1423, 2019.

[23] X. Zhuang, C. Ye, Y. Ding, H. Hui, and B. Zou, "Data-driven reserve allocation with frequency security constraint considering inverter air conditioners," *IEEE Access*, vol. 7, pp. 120014–120022, 2019.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 5206082, 15 pages
https://doi.org/10.1155/2020/5206082

Research Article

Strategies for Datacenters Participating in Demand Response by Two-Stage Decisions

Yuling Li, Xiaoying Wang, and Peicong Luo

State Key Laboratory of Plateau Ecology and Agriculture, Department of Computer Technology and Applications, Qinghai University, Xining 810016, China

Correspondence should be addressed to Xiaoying Wang; wxy_cta@qhu.edu.cn

Received 1 May 2020; Accepted 19 June 2020; Published 22 July 2020

Guest Editor: Van Huy Pham

Modern smart grids have proposed a series of demand response (DR) programs and encourage users to participate in them with the purpose of maintaining reliability and efficiency so as to respond to the sustainable development of demand-side management. As a large load of the smart grid, a datacenter could be regarded as a potential demand response participant. Encouraging datacenters to participate in demand response programs can help the grid to achieve better load balancing effect, while the datacenter can also reduce its own power consumption so as to save electricity costs. In this paper, we designed a demand response participation strategy based on two-stage decisions to reduce the total cost of the datacenter while considering the DR requirements of the grid. The first stage determines whether to participate in demand response by predicting real-time electricity prices of the power grid and incentive information will be sent to encourage users to participate in the program to help shave the peak load. In the second stage, the datacenter interacts with its users by allowing users to submit bid information by reverse auction. Then, the datacenter selects the tasks of the winning users to postpone processing them with awards. Experimental results show that the proposed strategy could help the datacenter to reduce its cost and effectively meet the demand response requirements of the smart grid at the same time.

1. Introduction

With the development of the world economy and the depletion of natural resources, energy crisis and environmental pollution have become serious issues of sustainable development. The Energy Development Corporation envisions the world powered by clean energy and calls on the energy sector to make the necessary shift to clean and low-carbon energy to alleviate the energy issues [1]. As a large energy-consuming country, China has proposed a series of corresponding energy utilization strategies. For example, *Jiangsu* province first experimented with a dynamic seasonal electricity pricing policy in 2015, using a potential way to encourage users to participate in demand response [2]. On the other hand, demand-side management as an important part of the smart grid is a critical way and an inevitable choice to achieve energy-saving and sustainable development of power enterprises. In order to meet various requests of the power users, the smart grid also needs to continuously deepen the power demand-side response and guidance mechanism to enhance the energy-saving and environmental protection effects [3]. The smart grid needs to propose a series of incentive mechanisms and effective demand response strategies to increase the flexibility of demand-side management to improve its sustainability. Smart grid programs should be carefully designed and can integrate the use of renewable energy, reduce electricity use, and lower CO_2 emissions. Hence, some research studies have turned to focus on the design of a new hybrid smart grid performance framework for evaluation from the perspective of sustainability [4].

Currently, an effective way for the smart grid to maintain sustainable development is to implement a more effective demand response strategy to realize the balance of power supply and demand. In order to improve the reliability, stability, and sustainability of the smart grid, power

companies in some countries have initiated a series of demand response (DR) programs to encourage their users to participate in with the aim of reducing the peak load of the grid and thus improving the reliability of the power grid [5]. In general, there are two main types of demand response programs of the smart grids, including incentive-based and price-based programs. In particular, incentive-based demand response programs include direct load control and interruptible load control. Price-based demand response programs include time-of-use electricity prices, real-time electricity prices, peak electricity prices, and other price-related programs [6]. A great number of power companies are also trying to implement special demand response programs to achieve the goal of peak load shaving. Demand response programs also provide users with incentives to encourage user participation to adapt power consumption to demand response goals.

On the other hand, with the increasing complexity of computing services and storage services provided by datacenters, and with the development of technology and the advancement of communication technologies, datacenters are more and more widely used as the basic computing and storage infrastructure. In 2017, there were approximately 8 million datacenters around the world, which consumed 416.2 terawatt-hours of electricity [7]. Hence, the power requirements of the datacenters increase dramatically, and high energy consumption becomes increasingly serious, which can lead to high carbon emissions of datacenters, which are not conducive to environmental protection. The datacenter is a large power consumer in the smart grid environment, and its load is usually highly flexible and adjustable. Datacenter participating demand response is envisioned as a promising approach for alleviating operational instability faced by smart grids and on the one hand, the datacenter can also save energy in response to the increase of electricity costs. In this case, the datacenter can be regarded as a potential participant of demand response programs in the smart grid [8, 9]. On the other hand, most datacenters support various types of workloads. Moreover, some workloads have the characteristic of high flexibility. For example, delay-tolerant tasks can be scheduled later as long as they could be completed before their deadlines [10]. This proves that datacenters can give a response by adjusting their own workloads when participating in demand response.

As an important part of the smart grid, demand-side management is a significant way for the sustainability of power companies. Therefore, the power grid should conduct more plans to attract more users to participate in demand response projects. Demand response conducts economic incentives aimed at balancing energy demand during critical demand periods [11]. As a commodity, electricity has its own particularity. One is that electricity is not easily stored. Generally, electricity is consumed as much as it generates. The other is that the cost of electricity is relatively high during peak periods of electricity consumption. Therefore, it is critical for power companies to propose kinds of demand response programs to alleviate the load pressure. Generally, consumers will be encouraged to shift their loads to off-peak hours in demand response process [12]. Real-time electricity prices track the price of the entire electricity market dynamically and guide users to avoid electricity consumption behaviors during peak periods, thereby reducing their electricity demand during peak periods [13]. However, datacenters are high demanders of electricity, making them potential participants in demand response. Datacenter participation in the demand response programs would bring many benefits, which could not only reduce the load pressure on the power grid but also reduce the power cost of the datacenter.

In this paper, we aim to reduce the total cost of the datacenter by involving the datacenter to participate in demand response program provided by the smart grid and in turn to improve the sustainability of the demand side of the grid. We proposed a two-stage decision process to figure out particular actions. In the first stage, the power grid sent the DR signals to the datacenter according to the variation of real-time electricity prices, while sending incentive information to the datacenter. The second stage was the bidding selection between the datacenter and its users based on a reverse auction mechanism. The user submitted bidding information firstly, including the number of tasks that could be postponed and the cost that the datacenter needed to pay due to the delayed execution of some tasks. Then, the datacenter tried to find the optimal solution by stochastic searching, delayed the execution of the task of the winning users, and rewarded them. Using the two-stage strategy, the purpose of reducing the peak load pressure of the smart grid while reducing the power cost of the datacenter could be achieved. This paper is an extended version of our prior work [14]. Based on the original version, we used the neural network for forecasting the real-time electricity price to judge the grid operation state in advance. In addition, we conducted a more detailed and comprehensive analysis of experiment results, including analyzing the impact of different forecasting accuracy on datacenter participants in DR, discussing the efficiency of the proposed method, and comparing the reward paid from the grid to the datacenter under different strategies.

The remainder of this paper is organized as follows. Section 2 introduces some related works about datacenter participation in demand response programs. Section 3 shows the system architecture and the optimization problems defined in this paper. Section 4 describes the two-stage decision process and strategies we designed to solve the defined problem. Section 5 analyzes the experiment results by comparing three different strategies. At last, Section 6 concludes the whole paper and discusses the possible future work.

2. Related Work

In recent years, the topic of how datacenter participates in demand response programs for the purpose of balancing the supply and demand for electricity has been discussed a lot. Different researchers keep different opinions and design diverse strategies to participate in demand response programs. Some of the researchers have focused on the benefits

for both supply and demand sides by proposing a game-theoretic framework between the datacenter and the smart grid. There were also some researchers who realized the purpose of participating in demand response by transferring the load of the datacenter to other distributed datacenters for execution. In addition, some works proposed letting the datacenter actively participate in DR programs by reducing the load by shifting or delaying its own tasks.

2.1. Taking a Game-Theoretic Approach to Participate in Demand Response. Zhou et al. [15] used the real-time electricity price as a demand response signal in the deregulated electricity market and modeled the datacenter decisions on power company selection and workload scheduling as a many-to-one game model. This algorithm not only improves the revenue of power companies but also reduces the average contract cost of the datacenter. They turned the game problem into an easy-to-solve mixed-integer linear programming problem for solving. In [16], the authors adopted time-varying rewards to motivate the willingness of users to participate in demand response projects. They used a game theory framework to model the game between a single datacenter and its users. In addition, they extended the proposed strategy by combining two demand response strategies: server shutdown and local renewable energy generation. Zhou et al. [17] aimed at the two-way electricity flow between smart grids and datacenters with hybrid green energy generation capability and geo-distributed nature of large cloud system; they proposed a pricing scheme designed for geo-distributed green datacenters from a multileader single-follower game point of view. At the cloud side, in quest for scalability, robustness, and performance, the energy cost minimization problem is solved in a distributed manner, based on the technique of alternating direction method of multipliers. Bahrami et al. [18] considered the real-time pricing (RTP) and modeled the decisions of datacenters select utility company and workload scheduling as a many-to-one matching game with externalities. They developed a distributed algorithm that guarantees converging to a stable outcome.

Through the analysis of the above pieces of literature, these studies are basically to establish a game model between the datacenter and the power company. Finally, they proposed an effective algorithm and obtained an equilibrium solution.

2.2. Participating in Demand Response by Geographic Load Balancing. Wang et al. [19] proposed a price-based incentive method. In order to motivate users to participate in their demand response projects, power companies set different power prices in different datacenters, encouraging users to shift the load to datacenters with lower electricity prices. Wang and Ye [20] regarded the datacenter operations as a problem of minimizing energy costs. They designed a distributed demand response algorithm. Administrators send optimized messages by broadcasting messages to each datacenter, and then each datacenter itself optimizes towards its own objective. Tran et al. [21] studied the distributed datacenter demand response method and proposed a two-stage pricing scheme based on the Stackelberg game. Each power company maximizes its profit by setting real-time prices in the first stage. In the second phase, the datacenter service provider uses workload shifting and dynamic server allocation to minimize its costs according to the electricity price. Liu et al. [22] thought predicted-based pricing is an appealing market design and showed that it outperforms more traditional supply function bidding mechanisms in situations where market power is an issue. These researches started with electricity prices from power companies and proposed incentive methods or other demand response methods for achieving the objective of reducing electricity costs in datacenters. Chau and Yang [23] took the advantage of the electricity price varying by region; they proposed a competitive online algorithm to achieve the load balance that can be applied to a limited time range and small-to-medium-scale energy storage. Liu et al. [24] explored the feasibility of using fully renewable energy to power Internet-scale systems. They conducted a trace-based study to assess three issues related to achieving this goal: the impact of geographic load balancing, the role of storage, and the optimal combination of renewable energy. Rao et al. [25] aimed at the location and time diversity of the electricity prices; they studied the issue of minimizing the total electricity cost in multiple electricity market environments while ensuring service quality. They modeled the problem as a constrained mixed-integer programming to solve.

2.3. Combining Multiple Methods to Participate in Demand Response. Guo et al. [26] aimed at the problem of a serious carbon footprint that resulted from using a backup diesel generator to supply power when datacenter participates in demand response. They focused on the effective and environment-friendly demand response of the datacenter. By designing economic incentives, they encourage tenants to reduce the load of the emergency demand response (EDR) in emergency situations and use the Nash equilibrium theory to analyze the relationship between datacenter operators and tenants. Chen et al. [27] argued that datacenters had the ability to provide large capacity reserves to emerging smart grid programs and, thus, provided opportunities for sustainable datacenter growth and cost savings as well as more flexibility for the grid. Therefore, they proposed policies to deliver datacenter demand response for peak shaving, regulation services, and frequency control programs. Cupelli et al. [28] combined the datacenter heating, ventilation, and air-conditioning equipment and delayed IT load and battery storage systems to coadjust the demand response projects. They proposed an optimization framework based on MPC (Model Predictive Control) forecasting to ensure the reliable operation of the datacenter and also analyzed the cost reduction in the case of incentives and prices. Paul et al. [29] considered that different types of jobs are distributed in different layers of the datacenter server, and, in particular, some of them can be delayed. They proposed rolling scheduling based on Receding Horizon Control (RHC). Based on the discrete-time model, they used two different

methods to formulate and solve the optimal server configuration problem: minimizing operating costs and jointly minimizing capital and operating costs between service configurations at different layers. In [30], they took into account the risk of energy demand and core business processes of the datacenter in participating in demand response process, and they used a genetic heuristic algorithm for energy consumption prediction. Guo et al. [31] proposed a coordinated management method and model for datacenter computing and cooling load. In the datacenter energy supply system with distributed photovoltaic power generation, they established a datacenter energy supply system under the time-of-use electricity price environment to respond to demand with the goal of minimizing the cost of electricity consumption. In response to the economic demand response of the custodian, they [32] proposed a reward-reduction incentive mechanism, which reflects the interaction between the DRP (demand response provider) of custodian compensation to the custodian and the colocation operator reward to tenants. Chen et al. [33] proposed a pricing mechanism for mandatory and voluntary EDR (emergency demand response) projects, which is a parameterized supply function based on bidding, both when tenants are price-taking and when they are price-anticipating.

Overall, some of the previous works only considered the power adjustment of the datacenter as a holistic entity to participate in demand response, and a lot of them were realized by the mutual transfer of load between geographically distributed datacenters. In addition, some researchers designed game-theoretic models between the datacenter and the power grid and sought to find an equilibrium solution between them. Considering the emergency of sending demand response signals, we adopt forecasting methods to obtain demand response signals in advance in order to schedule the tasks of the datacenter in time. In this paper, we design a two-stage decision strategy to make the datacenter participate in demand response. Compared with prior research work, the main contributions of this paper include the following:

- Establishing interactions between the smart grid, the datacenter, and the users and designing a two-stage decision process to participate in DR
- Predicting DR requirements by following real-time electricity price variation in the first interaction stage and designing a corresponding incentive mechanism
- Designing the second stage of the datacenter interaction mechanism with the users by reverse auction, in which users submit bid information to the datacenter voluntarily and the datacenter selects some users by analyzing the optimal choice

3. System Architecture and Problem Modeling

In this section, we mainly introduce the interaction between the grid, the datacenter, and the users. We aim at the problem of datacenter participating in demand response to establish the optimization problem. Figure 1 shows the interaction between the smart grid, the datacenter, and the users. The grid sends demand signals and incentive information at some unspecific moments. Then, the datacenter selects users according to their bidding information and postpones tasks of selected users after receiving the demand response signals from the smart grid. Assume that the datacenter consists of N hosts, denoted as host 1 to host N. We also assume the electricity price at time t is $\sigma(t)$ (\$/Wh) and the revenue brought by executing a task is γ (\$).

3.1. Grid Side. As shown in Figure 1, the operations of the grid side include sending DR signals, incentive information, and charging electricity bills to datacenters. In this paper, we use the variation of real-time electricity prices to determine the moment of demand response.

Generally, smart grids propose incentive methods in order to attract more users to participate in demand response programs for balancing the supply and demand for electricity. Considering that the datacenter (DC) participates in demand response, the grid side should give the datacenter a certain reward based on the behavior of DC. In general, the rewards given to users participating in demand response by the power grid are calculated according to the execution rate of power reduction. We define a parameter of execution rate to reflect the degree of datacenter participation in DR. Here, we use the average value of the power consumption of the datacenter at each time slot in the past 5 days as the reference value p_{base}. If the actual power consumption reduction after the DC participates in the demand response is p_{red}, then the execution rate is defined as $r_e = (p_{\text{red}}/p_{\text{base}})$. We refer to the reward method proposed in [34] and define the reward at time t as equation (1), denoted as $\theta(t)$, wherein the baseline is a constant which is a base value preset for rewards:

$$\theta(t) = \begin{cases} r_e \cdot \text{baseline}, & 0 \le r_e < 1, \\ \text{baseline}, & r_e \ge 1. \end{cases} \tag{1}$$

3.2. User Side. From the perspective of datacenter users, they can voluntarily decide whether to participate in the demand response or not. Here we consider two types of tasks, including interactive workloads and batch workloads, respectively. At each time slot, users submit tasks to the datacenter, including the tasks that need to be responded immediately (interactive tasks) and the tasks that can be postponed to process (batch-type tasks). Denote $I_u(t)$ and $B_u(t)$ as the number of interactive-type tasks and batch-type tasks, respectively, which are submitted at time slot t by user u. Furthermore, we divided batch-type tasks into two categories, including delay-sensitive and delay-insensitive tasks according to the time that can be delayed. Specifically, delay-sensitive tasks can be postponed for 4 time intervals, and delay-insensitive tasks can be postponed for 8 time intervals.

If the user decides to participate in the demand response, he/she can submit the bid information defined as a tuple consisting of the number of task candidates and the corresponding charge, denoted as $<\phi_u(t,t'),\ \alpha_u(t)>$. Specifically, the first element $\phi_u(t,t')$ denotes the number of tasks

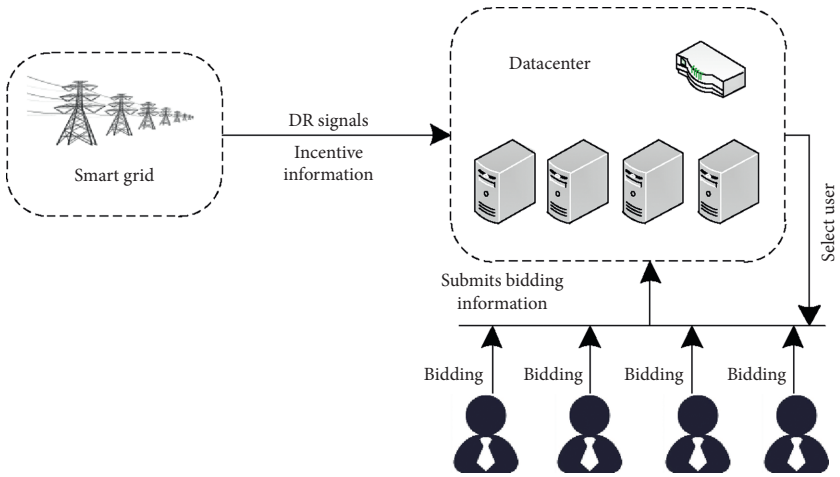


Figure 1: Interaction between the smart grid, the datacenter, and users.

that can be delayed to process from time t to t' and the second element $\alpha_u(t)$ denotes the expected charge resulting from postponing these tasks of user u. We use a binary variable $x_{u,t}$ to record whether the user u wins the bid at the current time slot, which indicates that user u is selected or not.

3.3. Datacenter Side. As stated previously, the datacenter can be regarded as a large load of the smart grid and could participate in demand response programs provided by the grid. From the perspective of the datacenter, effectively scheduling tasks submitted by its users could be a useful means to help it adjust its own power consumption. This can also help the datacenter owner to reduce the electricity cost.

In general, datacenter should guarantee the QoS (quality of service) of users when datacenter processes the requests of users. In this section, we define a penalty model for the datacenter if it proactively postponed the execution of the user tasks. In other words, if tasks submitted by users are postponed due to the datacenter participation in the DR programs, the datacenter should pay the users for the delay. We adopted the penalty model adopted in [35], using t_{sub} and t_{exec} to represent the submission time and the actual execution time of the task, respectively. Then, the penalty for delaying task i can be defined as follows:

$$\mu(i) = \left(t_{\text{exec}} - t_{\text{sub}}\right) \cdot \beta, \tag{2}$$

where β is a constant factor used to reflect the penalty.

Hence, denote $\pi(t)$ and $r_{\text{user}}(t)$ as the number of tasks that violate the QoS restriction and the total fee paid to users by the datacenter, respectively. Therefore, $r_{\text{user}}(t)$ consists of the penalty due to the postponed tasks in order to respond to the DR signal and it can be calculated as

$$r_{\text{user}}(t) = \sum_{i=1}^{\pi(t)} \mu(i). \tag{3}$$

In addition, in the process of the datacenter interacting with its users, the datacenter needs to pay the users who actively bid the cost and are willing to participate in demand response. Then, the total charge $\omega(t)$ of datacenter that needs to pay for the users can be calculated as

$$\omega(t) = \sum_{u=1}^{m} \alpha_u(t) \cdot x_{u,t}. \tag{4}$$

In general, computing power consumption is the dominant part of the total datacenter power consumption [36], which is mainly related to the utilization ratio of the computing devices. Denote P_n^t as the computing power consumption of host n in the t^{th} slot. Here, we use a linear calculation method to calculate power consumption [37], as shown in the following equation:

$$P_n^t = P_n^{\max}\left(c + (1-c) \cdot u_{n,t}\right), \tag{5}$$

where $P_n^{\max}$ is the maximum power consumption of the server, c is the percentage of static power consumption, and $u_{n,t}$ is the CPU utilization of the server n at time t. Hence, the total power consumption of datacenter at time t can be calculated as

$$P_t = \sum_{n=1}^{N} P_n^t. \tag{6}$$

3.4. Establishment of the Optimization Problem. Hence, the total cost of datacenter consists of the penalty due to the postponed tasks in order to respond to the DR signal, the cost paid to some users who actively bid, the revenue from performing tasks, and the energy charge. Therefore, the total cost of the datacenter at time slot t can be calculated as follows:

$$C(t) = E_t \cdot \sigma(t) + r_{\text{user}}(t) + \omega(t) - \lambda(t) \cdot \gamma - \theta(t), \tag{7}$$

where $\lambda(t)$ and E_t represent the total amount of tasks that need to be executed at time t and energy consumption of datacenter at time slot t, respectively.

Overall, the target optimization problem is to minimize the cost of datacenter under certain constraints, as shown in equations (9) and (10), wherein equation (9) means that the

total reward asked by the users for postponing their tasks should be less than the revenue brought by finishing these tasks and equation (10) limits the valid range of the variables.

$$\text{minimize } C(t) \tag{8}$$

$$\text{subject to} \sum_{u=1}^{m} \alpha_u(t) \cdot x_{u,t} <= \sum_{u=1}^{m} \phi_u(t, t') \cdot \gamma, \tag{9}$$

$$x_{u,t} \in [0,1], \quad 0 \le t \le \tau,\ u = 1, 2, \ldots, m. \tag{10}$$

The notations used throughout this paper are listed in Table 1.

4. Strategies and Methods

To address the issue mentioned in Section 3, we proposed a two-stage decision process to participate in DR for cost reduction of datacenter and peak load shaving of the power grid and also implemented two other methods for comparison.

4.1. Two-Stage Decision Process (OP). In this subsection, we will show in detail the issues involved in the proposed two-stage approach.

4.1.1. Stage I: The Interaction between the Grid and the Datacenter. The main goal of this stage is to improve grid reliability and reduce peak load pressure. At this stage, the grid will send DR signals based on the price of electricity. In order to allow grid users to have sufficient preparation time to adjust their own loads to participate in demand response, we adopt the method of predicting real-time electricity prices to send DR signals. The grid will send incentive information to electricity users according to equation (1) so as to attract more entities to participate in demand response.

Since the datacenter might not know the accurate time when the DR request will arrive, forecasting techniques were used here to predict the coming of DR signals according to the real-time electricity price of the smart grid. Here, we use the neural network of long short-term memory (LSTM) network model to learn the characteristics of the real-time electricity price [38]. The model is a recurrent neural network trained using back-propagation time, and it can overcome the problem of gradient disappearance of traditional neural networks [39]. Moreover, LSTM uses four neural network layers and interacts with each other by a special method. Specifically, we use the historical real-time electricity price data of the previous k days to predict the data on the $k+1$ day.

4.1.2. Stage II: The Interaction between the Datacenter and Its Users. At this stage, a reverse auction mechanism is employed for the interaction between the users and the datacenter. Users perform competitive bidding to the datacenter when the datacenter receives demand response signals. After the bidding process, the datacenter selects the winning users, pays the required fee to them, and reallocates the tasks of these users according to the auction results.

(1) Reverse auction model. In this process, we regard the datacenter as the buyer and the users as the sellers. Each user u voluntarily submits bid information represented as a tuple $< \phi_u(t, t'), \alpha_u(t) >$. Then, the datacenter selects the winning users according to the objective function as shown in equation (8) while meeting the constraints shown in equation (9) and equation (10).

After the auction between users and the datacenter, the tasks performed at DR periods will be rescheduled. The tasks can be shifted in the time dimension to reduce the electricity cost of datacenter. Denote m as the number of users who submit tasks to the datacenter at time t. Denote $\delta(t)$ as the number of batch-type tasks executed at time t, including both the batch-type tasks submitted at time t and the batch-type tasks postponed from previous time slots, which can be calculated as

$$\delta(t) = \sum_{u=1}^{m} B_u(t) - \sum_{u=1}^{m} \phi_u(t, t') + \sum_{u=1}^{m} \sum_{t_b=0}^{t-1} \phi_u(t_b, t) \cdot x_{u,t}, \tag{11}$$

where $t' > t, t_b < t$.

Denote $\lambda(t)$ as the total amount of tasks that the datacenter needs to process at the time slot t, which can be calculated as

$$\lambda(t) = \sum_{u=1}^{m} I_u(t) + \delta(t). \tag{12}$$

(2) Solution. We adopt a stochastic searching algorithm based on the concept of the *Genetic Algorithm* (GA) to solve the optimization problem as defined in Section 3. GA is a kind of random parallel search algorithm based on the principles of natural selection and genetic genetics. It starts with randomly initializing the population, and after multiple iterations, it constantly generates individuals with higher fitness and finally obtains the optimal individual. We use binary encoding, where an example of chromosome representation is shown in Table 2. If a user wins the bid, the corresponding coding value is 1; otherwise it is 0. In the example shown in the table, $u1$ did not win the bidding but $u2$ won it.

In this method, the datacenter not only participates in the demand response at the first stage but also uses the reverse auction mechanism to reschedule tasks of users in the second stage. Users submit bidding information to the datacenter in a bidding manner, and then the datacenter selects the winning user by stochastic searching to find the close-to-optimal solution. At last, the datacenter can effectively participate in DR programs by actively interacting with the users and reschedule the workloads. The pseudocode of the datacenter decision in the user bidding process is shown as follows (Algorithm 1).

In Algorithm 1, M is the population size, p_c represents the crossover probability, p_m is the mutation probability, I represents the iteration times and i is the evolutionary generation counter, P_0 is the original population, *ind_best* is used to record the best individual, and *newPop* is the current population updated with new generated individuals during

Table 1: Parameter notations.

Symbol	Description
N	The number of hosts in datacenters
τ	The number of time slots
$\sigma(t)$	The electricity price at time slot t
γ	The revenue of a task executed by the datacenter
p_{base}	Basic power consumption
p_{red}	The actual power consumption reduction
r_e	The execution rate when participating in DR
$\theta(t)$	The reward is given to datacenters by the grid at time slot t
$\mu(i)$	The penalty of violating the QoS of task i
t_{exec}	The execution time of a task
t_{sub}	The submission time of a task
β	The penalty factor
P_n^t	The power consumption of host n at time slot t
$P_n^{\max}$	The maximum power consumption of host n
c	The percentage of static power consumption
$U_{n,t}$	The CPU utilization of host n at time slot t
P_t	The total power consumption of datacenter at time slot t
E_t	The total energy consumption of datacenter at time slot t
$I_u(t)$	The number of interactive tasks submitted at time slot t by user u
$B_u(t)$	The number of batch-type tasks submitted at time slot t by user u
$\phi_u(t, t')$	The number of tasks from user u that can be delayed to process from time t to t'
$\alpha_u(t)$	The charge resulted from postponing the execution of some tasks from user u at time slot t
$x_{u,t}$	A binary variable indicating user u is selected or not at time slot t
$\delta(t)$	The number of batch-type tasks executed at time t
$\lambda(t)$	The total number of tasks needed to be processed at time slot t
$\pi(t)$	The number of tasks that violate QoS at time slot t
$r_{\text{user}}(t)$	The reward given to the users by the datacenter at time slot t

each round. $f(x)$ is a function for calculating the fitness value, which is the opposite of the optimization objective as defined in equation (7). Line 5 evaluates the fitness value of each individual ind_j in population P_i and line 10 selects two individuals from the current population P_i according to roulette selection operation to generate two new individuals.

Here we give on example to demonstrate the working process of the two-stage strategy. Assume that currently a DR moment is predicted through the decision in the first stage and 5 users (denoted as $u1$~$u5$) have submitted their bidding information at this moment. Then, the datacenter will select winning users according to the bidding information submitted by the users based on Algorithm 1. In the selection process, assume that $u1$, $u3$, and $u4$ won, while $u2$ and $u5$ failed. In this case, the datacenter will postpone some deferrable tasks based on the task information submitted by $u1$, $u3$, and $u4$ with awards, while the tasks of $u2$ and $u5$ are still processed at the current time slot.

4.2. Best Effort (BS). In this strategy, the datacenter makes its best effort to participate in DR programs of the grid by selectively delaying some tasks according to the needs of the datacenter. No auction or bidding mechanisms are employed here. Therefore, the datacenter might violate the QoS of the users and thus a certain penalty would be charged for delaying the execution of some user tasks arbitrarily regardless of the task characteristics. In other words, this strategy does not consider the active participation potential of the users and their other requirements.

Table 2: Chromosome coding example.

User	$u1$	$u2$	$u3$	$u4$	$u5$
Bidding situation	0	1	1	0	1

4.3. Static Method (ST). In this strategy, the datacenter does not participate in the demand response programs and just manages the workload as usual. For the datacenter, no particular actions will be performed after receiving DR signals.

5. Experimental Results and Analysis

In order to verify the effectiveness of the proposed method, we used the *CloudSim Plus* tool [40] to simulate the interaction between the datacenter and its users. The datacenter used in our experiment consists of 2000 hosts, and we assume that the number of users is 10. The detailed parameters settings are shown in Table 3. In the experiment, we simulate a whole day (24 hours) and divide it into 24 slots. The data of real-time electricity price is derived from the website [41].

In the experiment, we adopt the *Google-trace* workload for simulation [42], and we assigned tasks to 10 users. During data processing, we use the *Pandas* [43] library of the *Python* [44] language to process and statistics of the workload data of *Google-trace* in the *task-usage* file. Tasks contain two categories, interactive-type and batch-type tasks, with a ratio of 1 : 1.5 [45].

Input: $I_u(t)$, $B_u(t)$, p_{base}, $\alpha_u(t)$, $\phi_u(t, t')$
Output: the best individual *ind_best*
Begin
(1) **Initialization**
(2) **Initialize** M; p_c; p_m; I; *newPop*; P_0; *ind_best*; $i \leftarrow 1$; $j \leftarrow 1$;
(3) **while** $i \leq I$ **do**
(4) **for** $j = 1$ to size(P_i) **do**
(5) evaluate $f(ind_j)$ for P_i
(6) **end for**
(7) *ind_best*←the best individual
(8) **while** size(*newPop*) < M **do**
(9) **if** random(0, 1) < p_c **do**
(10) choose individual *par*1, *par*2 from P_i
(11) determine the crossover point randomly
(12) crossover *par*1 and *par*2
(13) generate two new individuals $x1$, $x2$
(14) **end if**
(15) **if** random(0, 1) < p_m **do**
(16) choose individual *par*3 from P_i
(17) mutate *par*3 at a random point
(18) generate a new individual $x3$
(19) **end if**
(20) *newPop*←[*newPop*, $x1$, $x2$, $x3$]
(21) **end while**
(22) $i \leftarrow i + 1$
(23) P_i←*newPop*
(24) **end while**
(25) **return** *ind_best*
End

Algorithm 1: The process of the second stage.

Table 3: Parameter settings.

Baseline	γ	β	K	N
5000	0.9	1.2	31	2000

Moreover, batch-type tasks are further divided into two types, delay-sensitive and delayed-insensitive tasks, respectively, which are delay-tolerant if being postponed for 4 time slots and 8 time slots. Figure 2 shows the detail of tasks used in the experiment, wherein the blue part represents interactive tasks, and the green part and yellow part represent delay-sensitive tasks and delay-insensitive tasks, respectively.

In our work, as stated in Section 3.1, the method of calculating the execution rate defined in this paper is based on the historical power consumption of the datacenter. Hence, we use the average power value of the 5-day historical data in *Google-trace* as the base value used for calculating the execution rate. Figure 3 shows the detailed average power consumption.

5.1. Forecasting Electricity Prices. Since the DR signal is usually sent from the smart grid according to the variation of the electricity loads, in order to proactively participate in demand response, we attempted to predict the electricity price ahead to prepare for better scheduling. Here, the electricity price data used in the following experiments derives from power companies [41]. We use the historical data of January 2018 as a training set to predict the real-time electricity price situation on 1 February and use the LSTM time series prediction model for forecasting. We use the *AdaDelta* optimizer; the number of iterations is 100. We also combined machine learning *scikit-learn* library for data normalization and training. The forecasting result is shown in Figure 4, and the error rate is basically kept within 10%. It can be clearly found that there are six peak electricity price moments, which are 6: 00, 7: 00, 17: 00, 18: 00, 19: 00, and 20: 00, respectively. In the experiment, DR signals are sent during these six moments.

5.2. Task Scheduling Details. Figure 5 shows the task scheduling conditions under the three strategies. As mentioned, under ST strategy, the datacenter just ignored the DR signals and did not perform any particular task-scheduling actions. Compared with the ST method, under BS and OP, there were fewer tasks processed during DR moments, including 6 : 00–7:00 and 17 : 00–20 : 00. This is because the datacenter actively reacted to the DR signals in the first stage and conducted task scheduling. Although the BS strategy can

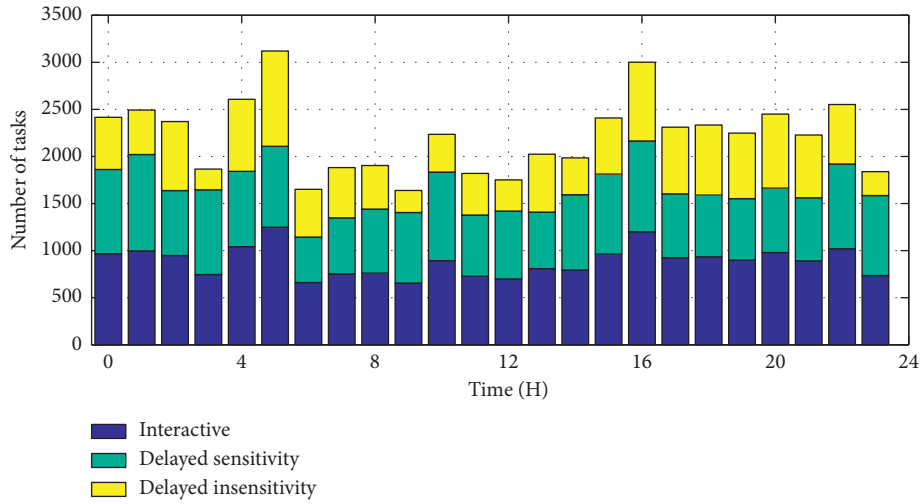


Figure 2: The number of tasks used in the experiment.

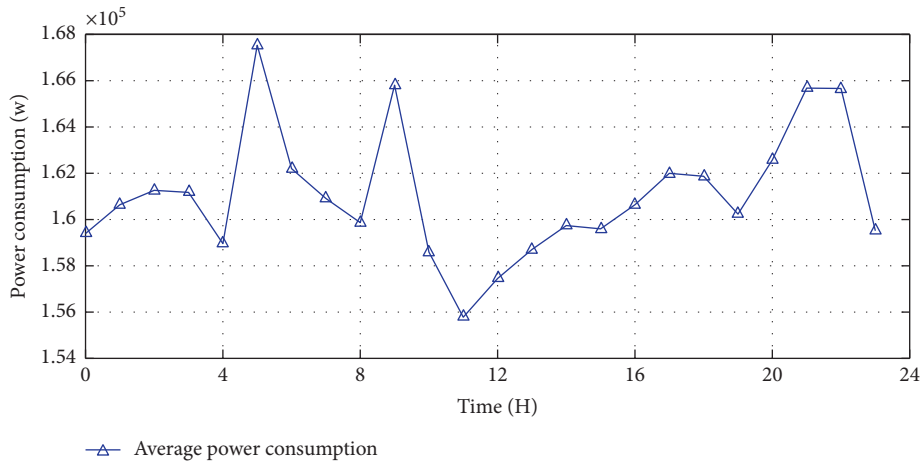


Figure 3: Average power consumption of the datacenter in past 5 days.

reduce the execution of tasks at the DR time, the datacenter may be punished by reducing the execution of some tasks. This is because the BS strategy does not consider whether the task can be postponed. For example, delaying interactive task execution will violate the QoS of tasks and result in a penalty.

Here, in order to prove that the OP strategy can indeed give a better response at the peak electricity price, we define a parameter PTA (peak average tasks to average tasks of all time slots) to measure the response situation of the three methods. Here, in our experiments, PTA is calculated as the average load at the moment of demand response periods divided by the average load of the entire day. Figure 6 shows the PTA value of the three strategies. Compared to the BS strategy, the PTA under the OP strategy is the lowest, which can show that the proposed OP strategy can achieve a better goal of participating in demand response and reduce load pressure during peak hours of the grid.

5.3. Power Consumption Condition. Figure 7 shows the detailed power consumption of the datacenter at each moment. The red-dashed line represents the variation of the real-time electricity price from the grid. There are two peak electricity price periods. It can be seen from the figure that the power consumption has been reduced under the BS and OP strategies during the time period of peak electricity prices, including 6 : 00–7:00 and 17 : 00–20 : 00. Results also showed that OP could lead to the largest reduction of power consumption compared with the other two strategies.

5.4. Revenues and Costs of the Datacenter. We also examined the detailed situation of the revenues and costs during the two-stage demand response actions conducted by the datacenter. As shown in Figure 8(a), the red pillar represents the fee paid by the datacenter to the users who win the bidding in reverse auctions under the OP strategy. In comparison, the blue pillars are the penalty in the BS strategy due to the postponed tasks submitted by the users, which resulted from the datacenter DR decisions. In other words, the datacenter delayed some of the interactive tasks so that their QoS was violated. Therefore, the datacenter could pay the users less money using the OP strategy than the BS strategy, since the penalties were reduced by means of the

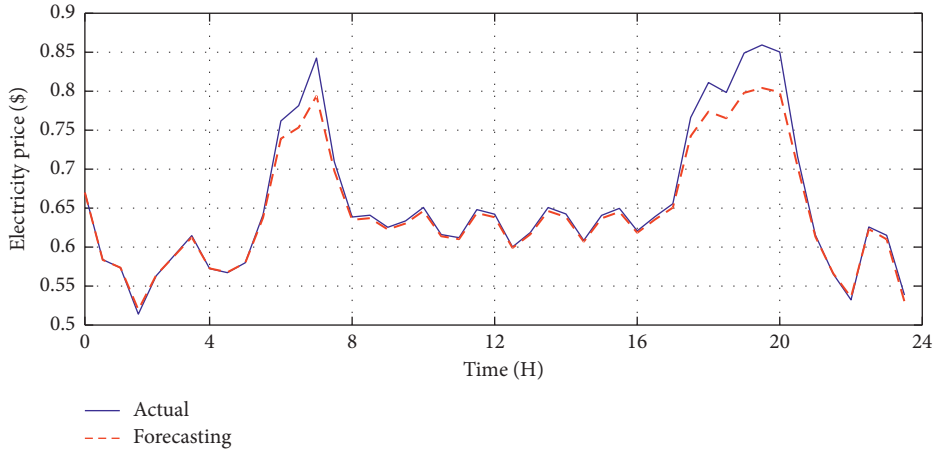


Figure 4: Forecasting the electricity price.

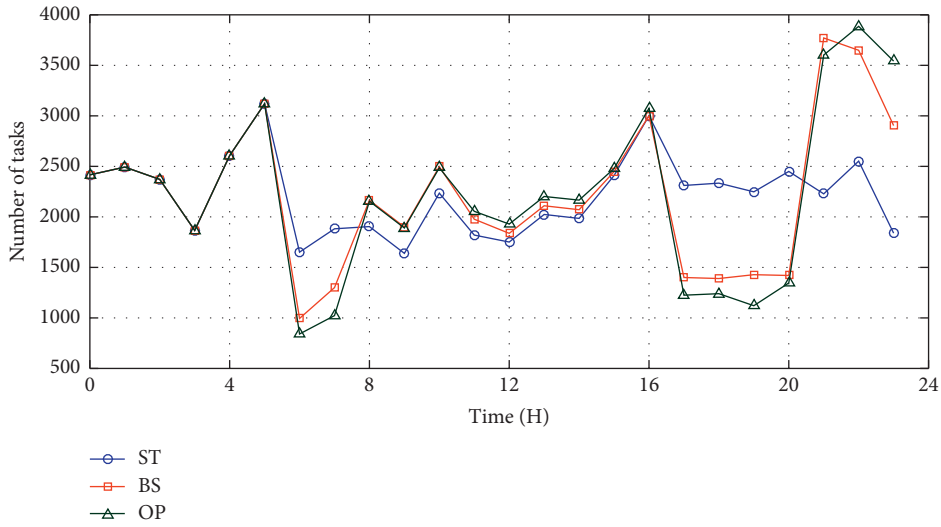


Figure 5: Task execution under three strategies.

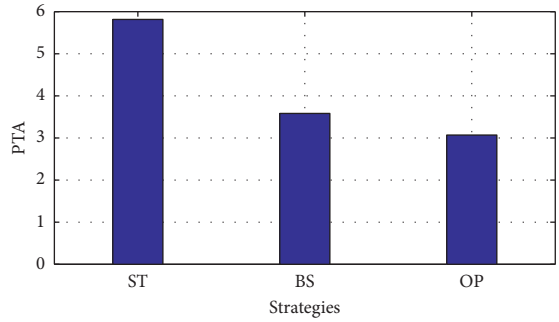


Figure 6: The PTA values under three different strategies.

second-stage auction mechanism. Figure 8(b) shows the rewards given to the datacenter from the smart grid during the periods of demand response. It can be obviously seen that the rewards under the OP strategy are higher than the BS strategy, which means that under the OP strategy the DR execution rate is higher.

Figure 9 shows the costs and rewards paid between the smart grid and the datacenter. As shown in Figure 9(a), from the results of the electricity cost under the three strategies, it can be seen that ST leads to the highest electricity cost, while the OP strategy leads to the lowest cost. This is because the OP strategy actively participates in the demand response at

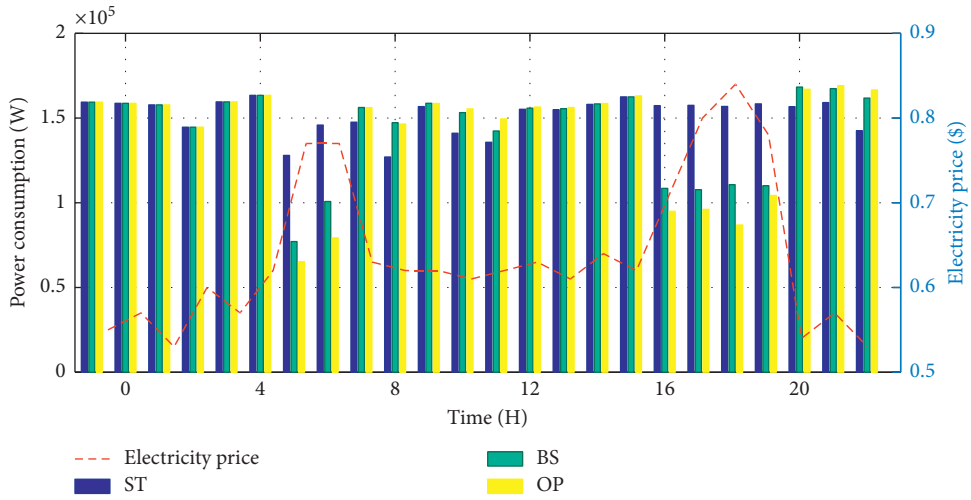


Figure 7: Hourly power consumption of the datacenter.

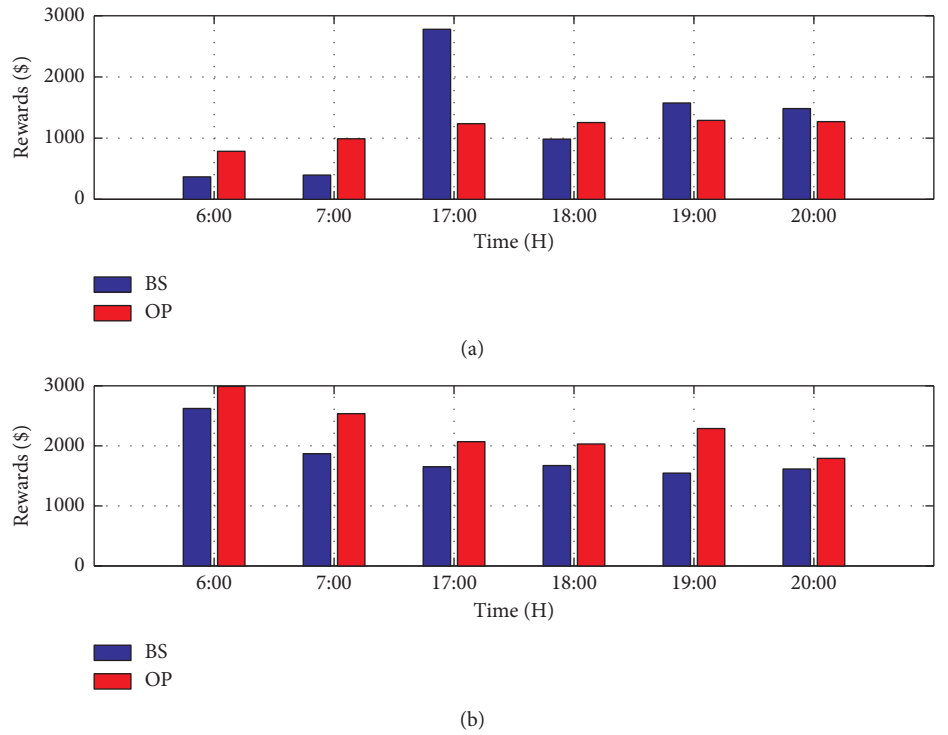


Figure 8: Detailed penalties and rewards in DR periods. (a) The rewards/penalties given to users by the datacenter. (b) The rewards given to the datacenter by the smart grid.

the moments when the electricity price is relatively high. In particular some tasks are postponed until the time when the electricity price is low, so the OP strategy leads to the lowest electricity costs. Figure 9(b) shows the rewards paid to the datacenter from the smart grid due to its participation in the demand response. It can be observed that the OP strategy has the highest reward. This is because the execution rate of the OP strategy was relatively high, which indicates that the demand response participation was effective and could bring revenue for the datacenter.

Furthermore, Figure 10 shows the total cost reduction of BS and OP compared to the ST strategy. It can be observed that they both reduced the total cost than ST, while OP brings much more reduction than BS. This is because the datacenter participated in the demand response, and some tasks were postponed for later execution during DR periods, thereby reducing the cost of electricity included in the total cost. Compared with the BS strategy, since the OP strategy used the reverse auction mechanism to negotiate with users, the penalty became less and the rewards obtained from the

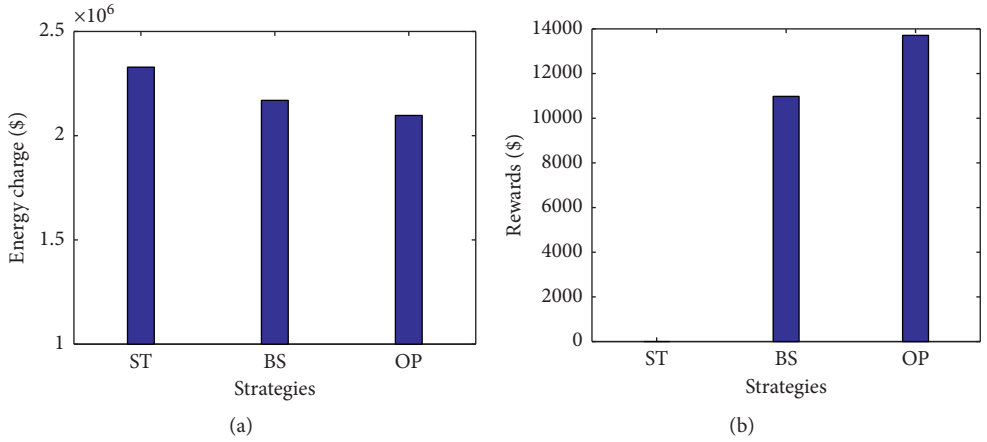


Figure 9: Electricity costs and total rewards of the three strategies. (a) Electricity costs of the datacenter under the three strategies. (b) The total rewards received from smart grid.

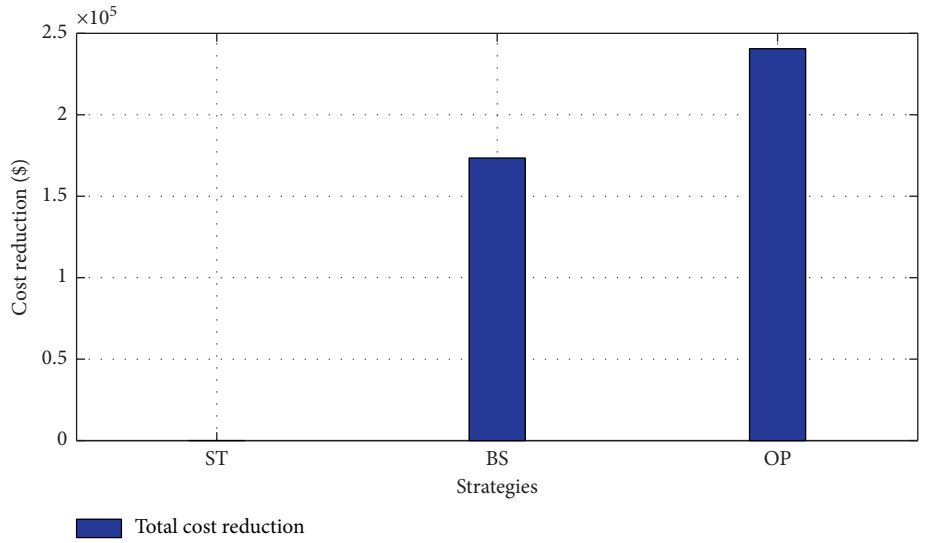


Figure 10: Cost reduction under three different strategies.

grid side could be higher. In contrast, the BS strategy resulted in higher penalties because the datacenter delayed the tasks without considering the requirements of users and thus violated their QoS. Overall, employing the OP strategy could minimize the total cost of the datacenter according to the above results and analysis.

In addition, we also analyzed the impact of the real-time electricity price forecasting results on datacenter participation in demand response when the training dataset size is in different settings. Figures 11(a) and 11(b) show the forecasting results when the size k of the training set is 31 and 15, respectively, in which the average error in Figure 11(a) is within 10% and that in Figure 11(b) is within 25%. The DR moments can be accurately predicted when the size k of the training set is 31, so the datacenter can respond at the correct DR moments. However, when the training set size k is 15 and the iteration time is 20, the forecasting results are not very accurate. Under this setting, not all of the DR periods were predicted in advance. For example, the datacenter should respond to the demand response of the power grid at the two moments of 6 : 00 and 17 : 00, but it did not actually perform any actions because of the wrong judgment. This resulted in the datacenter not fully participating in DR program and being unable to effectively reduce its own electricity costs. From the results shown in Figure 11(c), we can obviously obtain that the total cost of the datacenter is the highest when there is no forecasting, while it is the least when the forecasting is accurate. This can prove that improving the accuracy of the DR signal prediction can reduce the total cost of the datacenter more.

5.5. Algorithm Performance Evaluation. Figure 12 shows the relationship between the iteration times and the convergence of the algorithm. In general, the faster the convergence speed of the algorithm is, the better the algorithm is considered, which can save the calculation execution time and in turn achieve high efficiency. From Figure 12, we can obtain

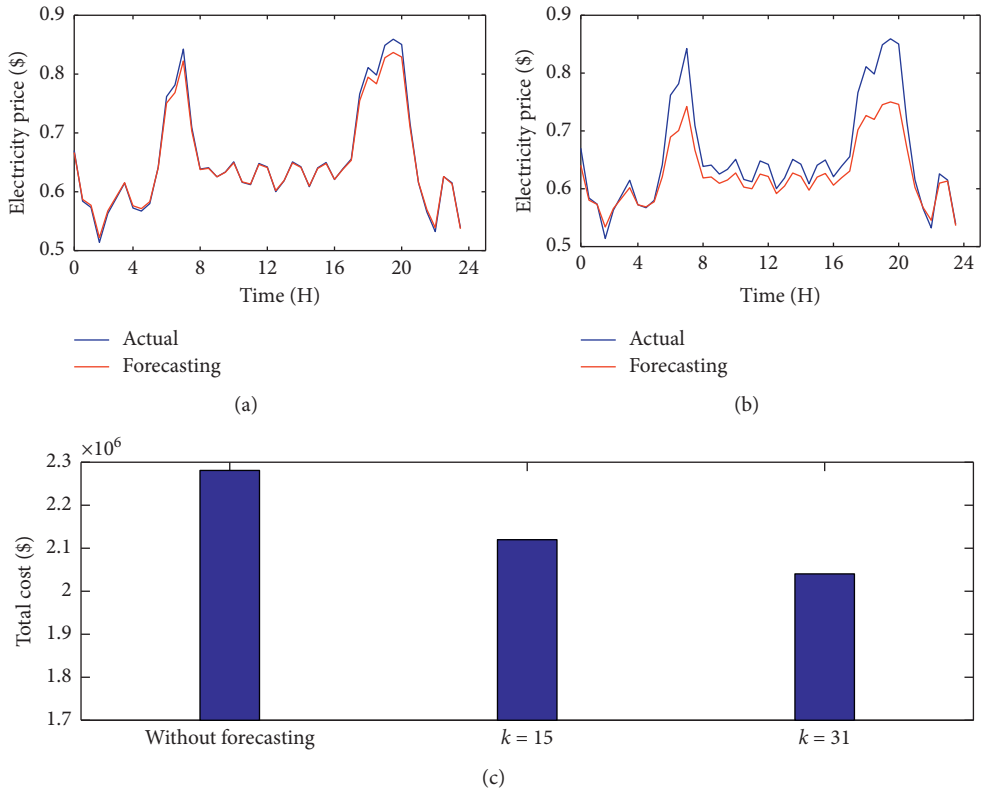


Figure 11: The total cost of the datacenter under different forecasting accuracies. (a, b) The forecasting results when k is set to 31 and 15, respectively. (c) The total cost under different settings.

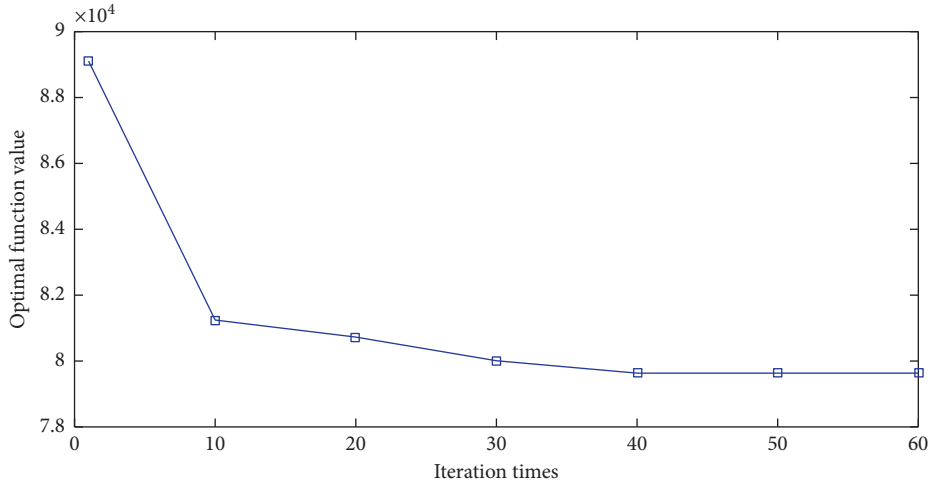


Figure 12: The relationship between the convergence and the iteration times.

that when the iteration times reach about 50 times, the optimal function value tends to be stable and get to the minimum. This illustrates that the method proposed in this paper is efficient for finding close-to-optimal solutions. As we know, datacenters usually have a very large number of processing requests at each moment. Therefore, using the proposed method to solve the problem is more efficient when the amount of requests becomes large.

6. Conclusions and Future Work

In this paper, we proposed a two-stage datacenter demand response strategy towards the smart grid environment. The first stage is the interaction between the smart grid and the datacenter, which mainly includes the grid sending DR signals to the datacenter according to real-time electricity load. In the second stage, the datacenter interacts with its users by employing the reverse auction mechanism. Users

submit bidding information to the datacenter, including the number of tasks that can be postponed and the corresponding charge for postponing these tasks. Then, the datacenter delays the execution of tasks of the winning users and pays the relevant fee after the bid selection. Simulation experiments illustrated that the strategy proposed in this paper could help the datacenter to reduce its total operational cost by effectively participating in demand response while facilitating the smart grid to shave the peak load and keep itself stable. Overall, the strategy proposed in this paper can adjust the power demand according to the DR signals to a certain extent. Therefore, the method can provide a way to help the sustainable development of power enterprises and also facilitate the decarbonization of the datacenters.

As a next step in the future, we are planning to study the formulation of the smart grid incentive mechanism and demand response items for the datacenter load characteristics, so that datacenters can better participate in demand response while ensuring that datacenter users have a higher service quality. We also plan to develop a datacenter charging agreement for users based on demand response signals on the grid side. In this way, the datacenter and the grid benefit at the same time.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (nos. 61762074, 91847302, and 61862053) and National Natural Science Foundation of Qinghai Province (nos. 2019-ZJ-7034 and 2020-ZJ-943Q).

References

[1] Energy, 2020, https://www.energy.com.ph/sustainability/.

[2] Demand Response in Jiangsu Province, 2019, https://max.book118.com/html/2018/0627/5201014120001300.shtm.

[3] H. Guo, *Development of Power Demand Side Management in the Smart Grid Based on the Analysis of Power Load*, in Chinese, http://cdmd.cnki.com.cn/Article/CDMD-10079-1014371242.htm, 2014.

[4] Z. Huiru and L. Nana, "Performance evaluation for sustainability of strong smart grid by using stochastic AHP and Fuzzy topsis methods," *Sustainability*, vol. 8, no. 2, p. 129, 2019.

[5] P. Palensky and D. Dietrich, "Demand side management: demand response, intelligent energy systems, and smart loads," *IEEE Transactions on Industrial Informatics*, vol. 7, no. 3, pp. 381–388, 2011.

[6] M. H. Albadi and E. F. El-Saadany, "Demand response in electricity markets: an overview," in *Proceedings of the 2007 IEEE Power Engineering Society General Meeting*, Tampa, FL, USA, June 2007.

[7] Datacenter Energy Consumption and Efficiency Issues, 2020, http://www.jifang360.com/news/2018731/n3530105256.html.

[8] A. Wierman, Z. Liu, I. Liu, and H. "Mohsenian-Rad, "Opportunities and challenges for data center demand response," proceedings of the international green computing conference," in *Proceedings of the International Green Computing Conference*, Dallas, TX, USA, November 2014.

[9] L. Zhang, S. Ren, C. Wu, and C. Li, "A truthful incentive mechanism for emergency demand response in colocation data centers," in *Proceedings of the 2015 IEEE Conference on Computer Communications (INFOCOM)*, Kowloon, Hong Kong, April 2015.

[10] S. S. Gill and R. Buyya, "A taxonomy and future directions for sustainable cloud computing: 360 degree view," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–33, 2019.

[11] Z. Khalid, G. Abbas, M. Awais, T. Alquthami, and M. B. Rasheed, "A novel load scheduling mechanism using artificial neural network based customer profiles in smart grid," *Energies*, vol. 13, no. 5, p. 1062, 2020.

[12] S. Caron and G. Kesidis, "Incentive-based energy consumption scheduling algorithms for the smart grid," in *Proceedings of the 2010 First IEEE International Conference on Smart Grid Communications*, Gaithersburg, MD, USA, October 2010.

[13] H. Huang, L. Deng, F. Wen, and F. Wang, "Customer response behavior based on real-time pricing," *Electric Power Construction*, vol. 37, pp. 63–68, 2016, in Chinese.

[14] Y. Li, X. Wang, P. Luo, D. Huang, and M. Zhao, "A two-stage demand response strategy for datacenters in the smart grid environment," in *Proceedings of the 2020 the 3rd International Conference on Energy, Electricity and Power Engineering in Chongqing*, pp. 24–26, Chongqing, China, April 2020.

[15] Z. Zhou, F. Liu, and Z. Li, "Pricing bilateral electricity trade between smart grids and hybrid green datacenters," *ACM SIGMETRICS Performance Evaluation Review*, vol. 43, no. 1, pp. 443-444, 2015.

[16] Y. Zhan, D. Xu, H. Yu, and Y. Sh, *Incentivizing Users of Data Centers Participate in the Demand Response Programs via Time-Varying Monetary Rewards*, https://arxiv.org/abs/1604.01950, 2016.

[17] Z. Zhou, F. Liu, and Z. Li, "Bilateral electricity trade between smart grids and green datacenters: pricing models and performance evaluation," *EEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3993–4007, 2016.

[18] S. Bahrami, V. W. S. Wong, and J. Huang, "Data center demand response in deregulated electricity markets," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2820–2832, 2019.

[19] H. Wang, J. Huang, X. Lin, and H. "Mohsenian-Rad, "Proactive demand response for data centers: a win-win solution," *IEEE Trans. Smart Grid*, vol. 7, no. 3, pp. 1–13, 2015.

[20] H. Wang and Z. Ye, "Renewable energy-aware demand response for distributed data centers in smart grid," in *Proceedings of the 2016 IEEE Green Energy and Systems Conference*, pp. 1–8, Long Beach, CA, USA, November 2016.

[21] N. H. Tran, D. H. Tran, S. Ren, Z. Han, E. N. Huh, and C. S. Hong, "How geo-distributed data centers do demand response: a game-theoretic approach," *IEEE Trans. Smart Grid*, vol. 7, no. 2, pp. 937–947, 2015.

[22] Z. Liu, I. Liu, S. Low, and A. Wierman, "Pricing data center demand response," *ACM SIGMETRICS Performance Evaluation Review*, vol. 42, no. 1, pp. 111–123, 2014.

[23] C. K. Chau and L. Yang, "Competitive online algorithms for geographical load balancing in data centers with energy

storage," in *Proceedings of the 5th International Workshop on Energy Efficient Data Centres—E2DC '16*, Waterloo, Canada, June 2016.

[24] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. H. Andrew, "Geographical load balancing with renewables," *ACM SIGMETRICS Performance Evaluation Review*, vol. 39, no. 3, pp. 62–66, 2011.

[25] L. Rao, X. Liu, L. Xie, and W. Liu, "Minimizing electricity cost: optimization of distributed Internet data centers in a multi-electricity-market environment," in *Proceedings of the INFOCOM 2010. 29th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies*, pp. 15–19, San Diego, CA, USA, March 2010.

[26] Y. Guo, H. Li, and M. Pan, "Colocation data center demand response using Nash bargaining theory," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 4017–4026, 2018.

[27] H. Chen, M. C. Caramanis, and A. K. Coskun, "Reducing the data center electricity costs through participation in smart grid programs," in *Proceedings of the International Green Computing Conference*, Dallas, TX, USA, November 2014.

[28] L. Cupelli, T. Schütz, P. Jahangiri, M. Fuchs, A. Monti, and D. Müller, "Data center control strategy for participation in demand response programs," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 11, pp. 5087–5099, 2018.

[29] D. Paul, W. D. Zhong, and S. K. Bose, "Demand response in data centers through energy-efficient scheduling and simple incentivization," *IEEE Systems Journal*, vol. 11, no. 2, pp. 613–624, 2015.

[30] A. V. Vesa, T. Cioara, I. Anghel et al., "Energy flexibility prediction for data center engagement in demand response programs," *Sustainability*, vol. 12, no. 4, p. 1417, 2010.

[31] J. Guo, H. Yang, P. Zhang, X. Ye, T. Xiong, and C. Gong, "Optimal dispatching of data center energy-suppling system with demand response," *Journal of Electric Power Science and Technology*, vol. 33, pp. 45–51, 2018, in Chinese.

[32] N. H. Tran, T. Z. Oo, S. Ren, Z. Han, E.-N. Huh, and C. S. Hong, "Reward-to-reduce: an incentive mechanism for economic demand response of colocation datacenters," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3941–3953, 2016.

[33] N. Chen, X. Ren, S. Ren, and A. Wierman, "Greening multi-tenant data center demand response," *Performance Evaluation*, vol. 91, pp. 229–254, 2015.

[34] Tai Power Company, https://www.taipower.com.tw/tc/index.aspx, 2019.

[35] S. K. Garg, S. K. Gopalaiyengar, and R. Buyya, "SLA-based resource provisioning for heterogeneous workloads in a virtualized cloud datacenter," in *Proceedings of the International Conference on Algorithms and Architectures for Parallel Processing*, pp. 371–384, Berlin, Heidelberg, October 2011.

[36] Y. Liu, E. Yang, and J. Xu, "Energy consumption management in cloud-oriented data centers," *Journal of Telecommunication Science*, vol. 12, pp. 102–108, 2012, in Chinese.

[37] Q. Meng, Z. Du, Y. Chen, and X. Wang, "A green-aware virtual machine migration strategy for sustainable datacenter powered by renewable energy," *Simulation Modelling Practice and Theory*, vol. 58, pp. 3–14, 2015.

[38] S. Baike, https://baike.sogou.com/v140767477.htm?fromTitle=LSTM, 2020.

[39] LSTM model interpretation, 2020, https://blog.csdn.net/qian99/article/details/88628383.

[40] Cloudsim-plus, 2019, https://github.com/manoelcampos/cloudsim-plus.

[41] Australian Energy Market Operator, https://www.aemo.com.au/, 2019.

[42] Google-Trace, https://commondatastorage.googleapis.com/clusterdata-2011-2/SHA256SU, 2019.

[43] Pandas Library, https://pandas.pydata.org/, 2020.

[44] Python language, 2020, https://www.python.org/.

[45] Z. Liu, Y. Chen, C. Bash et al., "Renewable and cooling aware workload management for sustainable data centers," in *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems—SIGMETRICS '12*, pp. 11–15, London, UK, June 2012.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 8149383, 13 pages
https://doi.org/10.1155/2020/8149383

Research Article

An Innovative Design of Decoupled Regenerative Braking System for Electric City Bus Based on Chinese Typical Urban Driving Cycle

Yanfeng Xiong, Qiang Yu, Shengyu Yan, and Xiaodong Liu

School of Automobile, Chang'an University, Xi'an 710064, Shanxi, China

Correspondence should be addressed to Yanfeng Xiong; xiongyanfeng118@126.com

Received 24 March 2020; Accepted 19 June 2020; Published 18 July 2020

Guest Editor: Sanghyuk Lee

This paper proposes a novel decoupled approach of a regenerative braking system for an electric city bus, aiming at improving the utilization of the kinetic energy for rear axle during a braking process. Three contributions are added to distinguish from the previous research. Firstly, an energy-flow model of the electric bus is established to identify the characteristic parameters which affect the energy-saving efficiency of the vehicle, while the key parameters (e.g., driving cycles and the recovery rate of braking energy) are also analyzed. Secondly, a decoupled braking energy recovery scheme together with the control strategy is developed based on the characteristics of the power assistance for electric city bus which equips an air braking system, as well as the regulatory requirements of ECE R13. At last, the energy consumption of the electric city bus is analyzed by both the simulation and vehicle tests, when the superimposed and the decoupled regenerative braking system are, respectively, employed for the vehicle. The simulation and actual road test results show that compared with the superposition braking system of the basic vehicle, the decoupled braking energy recovery system after the reform can improve the braking energy recovery rate and vehicle energy-saving degree. The decoupled energy recovery system scheme and control strategy proposed in this paper can be adopted by bus factories to reduce the energy consumption of pure-electric buses.

1. Introduction

Insufficient mileage has become the main obstacle that restricts the development of the pure-electric vehicle due to its layout, battery cost, energy density, and other restrictions, especially for an electric city bus. Under the premise of constant external constraints, improving the vehicle energy-saving efficiency is of great significance for extending the vehicle driving range [1–5]. There are many approaches to improve vehicle energy efficiency. This paper systematically analyzes the factors that affect vehicle energy efficiency and puts forward an engineering solution to improve energy efficiency, which is helpful to reduce the energy consumption of the vehicle.

1.1. Literature Review. Superposition type of braking energy recovery system is mainly used in an electric city bus, which applies motor braking synchronously on the original mechanical braking without changing the braking system of the original vehicle. For passenger vehicles, in addition to the superposition type, a decoupled braking energy recovery system is adopted for some models. Decoupled braking energy recovery products have been launched by Bosch, TRW, and other enterprises [6, 7]. At present, the researches on the braking energy of electric buses are comprehensive and in-depth. The main factors affecting braking energy recovery in pure-electric passenger vehicles were analyzed, which include braking strength, brake hydraulic cylinder pressure, and initial vehicle speed in literature [8]. The attachment coefficients of different road surfaces were analyzed, which would affect the braking energy recovery rate in the literature [9]. In literature [10], in addition to the braking energy recovery rate, two other braking energy recovery evaluation indicators were established, which are energy-saving contribution rate and driving range contribution rate. The former reflects the contribution of braking energy recovery to vehicle energy consumption, and the latter refers to

the increased driving range of vehicles with braking energy recovery function compared with the vehicles without this function. The definition of braking energy recovery contribution rate was proposed, which is the ratio of the recovered portion of the braking energy which can be converted into kinetic energy at the wheel end to the braking energy recovered by the whole driving condition in literature [11]. A specific test method for calculating the braking energy recovery rate was provided by literature [12], and the real-vehicle tests were carried out on three pure-electric vehicles. However, the relationship of the braking energy recovery rate vehicle driving range was not analyzed, which is analyzed quantitatively. Although the methods to improve the braking energy recovery rate were not mentioned in the literature [13, 14], predecessors have done a lot of useful exploration. The fuzzy rules for braking energy recovery based on practical and simulation experience were designed in the literature [15], without verified on a real vehicle. A fuzzy control strategy for braking energy recovery was proposed by the literature [16, 17], which aimed at small and medium braking strength. A decoupled braking energy recovery strategy for pure-electric passenger vehicles was proposed by literature [18]. VCU analyzed the driver's intention of the braking pedal and calculated the distribution of braking force between front and rear axles. The motor braking had the priority, and the insufficient part is provided by the hydraulic control unit. However, braking stability is not considered in this literature. Based on the known cycle condition, the model of braking prediction and distribution was developed in literature [19]. On the premise of braking stability, with the target of the maximum motor braking capacity and recovery efficiency, the model realized offline calculations in advance which met the real-time requirements of the table looking up online. Under the actual working condition, the robustness of the model still needs to be verified. A fuzzy control strategy for braking energy recovery was proposed, which took SOC and vehicle speed as input variables, and took a braking force of motor as output variables in literature [20]. It is helpful to recover more braking energy, but due to the unchanged braking system structure, the improvement is limited. The braking system structure was changed, in which the traditional hydraulic braking was replaced by EMB in literature [21]. Under the condition that ABS was not triggered and the braking strength was little, motor braking met the vehicle braking requirement. The validity of the strategy was verified in the HILLS environment, without verification for a real vehicle. A superposition braking energy recovery strategy for small pure-electric four-wheel-drive passenger vehicles was proposed in the literature [22]. Under small and medium braking strength, four motors recovered braking energy at the same time in order to increase the braking energy recovery rate. This power system configuration is less commonly used in pure-electric city buses. A braking energy recovery system scheme was proposed in literature [23], based on composite power supply combining power battery with supercapacitor. This scheme improves the capacity of braking energy recovery under large braking strength, but it is not applicable for an urban condition which is mainly small braking strength. Literature [24] optimized vehicle speed fluctuations to reduce energy consumption based on adaptive cruise function for pure-electric passenger vehicles. However, it is not applicable to frequent acceleration and deceleration conditions of urban buses. Pure-electric buses mainly operate on urban roads, and the small-strength braking condition is more than passenger vehicles. It is beneficial for improving the braking energy recovery rate of a pure-electric city bus by realizing motor braking in advance based on the control strategy of decoupled braking energy recovery. In addition, a pure-electric city bus is rear-motor rear-wheel-drive configuration with a large load of the rear axle, so the improvement of energy consumption on the braking-decoupled rear axle is better than that of the braking-decoupled front axle.

1.2. Motivation and Innovation. Pure-electric city bus driving form is basically rear-motor rear-wheel-drive configuration. Aiming at the energy dissipation path, the factors affecting vehicle energy-saving rate and braking energy recovery rate are hackled. In addition to constants such as transmission system and powertrain efficiency, the variables are mainly the ratio of motor braking force to mechanical braking force. Considering that the motor only acts on the rear axle which has a large load, it is helpful for increasing vehicle energy-saving rate and braking energy recovery rate by reducing mechanical braking force and action time on the rear axle. In order to realize the reduction just mentioned, a relief valve is increased on the braking gas road of the rear axle. Considering that axle-load distribution of small braking strength is not required in ECE braking regulations when the braking strength is small, the relief valve is closed and the rear axle is only under motor-braking, and when the braking strength is large, the relief valve is opened, and the rear axle is only under mechanical braking. This innovation is applicable for frequent small braking strength conditions of a city bus. Compared with the vehicle with the scheme and control strategy of this paper to superposition braking energy recovery scheme, the energy consumption is reduced by 3–7%, and the larger the rear axle loads, the more energy consumption decreases. It is proved to be feasible and effective by performance simulation and real-car verification. The new parts selected in this scheme are all mature, which are convenient for engineering.

1.3. Organization of the Paper. The remainder of the paper is organized as follows. Section 2 analyzes the influence factors of the energy-saving effect for the electric city bus. The scheme and control strategy of the decoupled braking energy recovery system is proposed in Section 3. The experimental verification of the decoupled braking system solution and the results discussion is implemented in Section 4, followed by the conclusions in Section 5.

2. Factors Affecting Vehicle Energy Conservation Effect

2.1. Vehicle Energy Dissipation Path. Considering working conditions, cost, and other factors, the driving configuration of a single motor and rear-motor rear-wheel-drive configuration is adopted by the electric city bus. Because the motor only acts on the rear axle, front axle braking can only be

completed by mechanical braking. The energy dissipation path is shown in Figure 1.

According to Figure 1, vehicle energy dissipation path is divided into three parts:

Part 1: Rear-wheel, half shaft, drive reducer, and transmission shaft. In this stage, the energy is transmitted in mechanical form.

Part 2: Motor and inverter. In this stage, the energy is transmitted in electrical form.

Part 3: Battery. In this stage, the energy is stored in chemical form.

When the vehicle is driven, battery-electric quantity E_B passes through the electric driving system, transmission shaft, and rear axle, and finally outputs driving energy E_T. The driving energy E_T needs to overcome driving resistance of the whole vehicle and then converts into mechanical energy. The mechanical energy of the front axle cannot be recovered, usually, that of the rear axle cannot be recovered totally. Braking energy recovery rate λ is defined as the ratio of motor recoverable energy E_{gen} to a total consumption of brake without the function of braking energy recovery E_T. Therefore, the recoverable energy at the position of the flange where the rear axle and the transmission shaft are connected is $\lambda E_T \eta_{tra}$, and the energy finally stored in the battery is E_B'. In the driving process, the output energy of transmission shaft is E_D, the output energy of the motor is E_{TM}, and that of battery is E_{dchg}; in the braking process, the output energy of transmission shaft is E_D', the recovered energy of the motor is E_{gen}, the recovered energy of the battery is E_{chg}, the mechanical efficiency of transmission shaft is η_{drv}, the mechanical efficiency of the rear axle is η_{tra}, the driving efficiency of the motor is η_{TM}, the recovery efficiency of the motor is η_{gen}, the discharging efficiency of the battery is η_{dchg}, and the recovery efficiency of the battery is η_{chg}. The vehicle energy flow distribution is shown in Figure 2.

As shown in Figure 2, the driving process energy consumption E_B is shown in

$$E_B = \frac{E_T}{\eta_{dchg}\eta_{TM}\eta_{dr\,v}\eta_{tra}}. \tag{1}$$

Energy recovery during braking is shown in

$$E_B' = \lambda E_T \eta_{tra}\eta_{drv}\eta_{gen}\eta_{chg}. \tag{2}$$

In order to analyze the vehicle energy-saving potential, the energy-saving degree of the whole vehicle η_{reg} refers to the ratio of the effective electric quantity recovered by braking to the electric quantity consumed by the whole vehicle without the function of braking energy recovery, which is shown in.

$$\eta_{reg} = \frac{E_B'}{E_B} = \lambda \cdot \eta_{gen}\eta_{TM}\eta_{chg}\eta_{dchg}\eta_{drv}^2\eta_{tra}^2. \tag{3}$$

The more energy-saving the degree of the whole vehicle is, the more energy can be reused. From (3), it can be seen that the vehicle energy-saving degree η_{reg} is related to driving efficiency η_{TM} and feedback efficiency η_{gen} of the motor system, discharge efficiency η_{dchg} and feedback efficiency η_{chg} of the battery system, efficiency η_{drv} of the transmission shaft, and mechanical efficiency η_{tra} of the rear axle and other constants and is directly proportional to the recovery rate of the braking energy λ.

The force analysis of the vehicle braking process is shown in Figure 3. In this figure, the normal reaction of the front axle is F_{zf}, the normal reaction of the rear axle is F_{zr}, the height of the vehicle mass center is h_g, the distance from the mass center to the front axle is a, the distance from the mass center to the rear axle is b, the load transfer of the front and rear axle during braking is G_z, the vehicle gravity is G, and the distance between the front and rear axle is L.

Combined with Figures 2 and 3, the braking energy recovery rate λ is shown in

$$\lambda = \frac{E_{gen}}{E_T} = \frac{\int P_{reg}\mathrm{d}t}{\int \left(P_{reg} + P_{hyd}\right)\mathrm{d}t} = \frac{F_{reg}}{F_{reg} + F_{xbf} + F_{xbr}}. \tag{4}$$

In (4), P_{reg} and F_{reg} refer to the braking power and force of the motor acting on the rear axle, respectively. P_{hvd} refers to the force of the mechanical braking system acting on the whole vehicle, consisting of F_{xbf} and F_{xbr}. The front axle braking force of the electric city bus is mechanical braking, so increasing the motor power P_{reg} and reducing the mechanical braking force of the rear axle F_{xbr} is beneficial to improving the braking energy recovery rate λ. The braking force of the whole vehicle, including F_{reg} and P_{hvd}, is strongly related to the driver's braking intention. The driver's braking intention refers to the driver's expectation of deceleration $\mathrm{d}v/\mathrm{d}t$ at current speed v, while the speed v and deceleration parameters $\mathrm{d}v/\mathrm{d}t$ are strongly related to the driving conditions of the vehicle, furthermore affecting the braking energy recovery rate λ.

2.2. Condition Parameters Analysis. The driving condition represents the speed-time of a vehicle in a certain area. There are many characteristic parameters in a driving condition, among which the speed parameters have a great impact on energy consumption [25]. Speed is mainly affected by morning and evening peak periods, road conditions, driving habits and other factors [26–29]. The ideal vehicle development model is to generate typical conditions by collecting the actual road conditions and carry out development verification based on typical conditions. However, considering the progress of the development and regulatory requirements, the test condition is always employed for actual vehicle development, while some actual conditions are also supplemented as reference.

There are two kinds of test conditions for city buses in China now: WTVC urban condition and CTUDC condition. The latter comes from China's automobile industry-standard named QC/T 759 "Urban Operation Cycle for Automobile test." The parameters of the actual conditions for Changchun, Jinan, Hangzhou, Dalian, and other typical

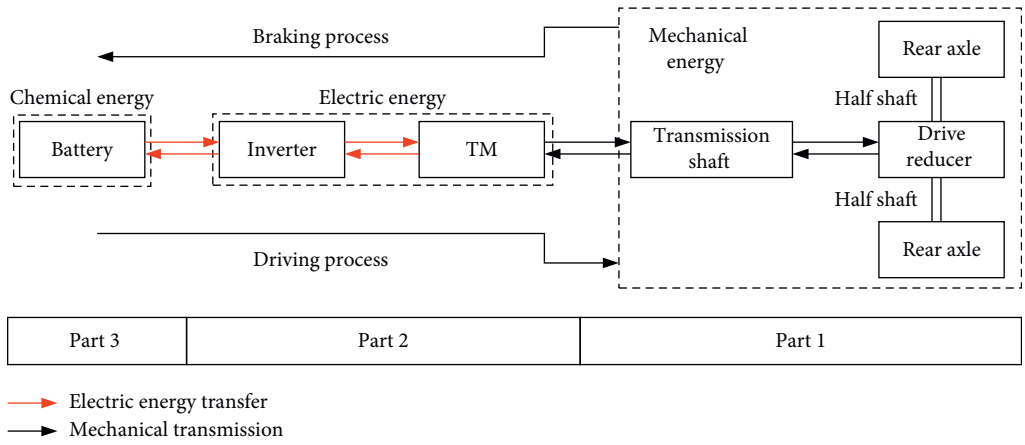


Figure 1: Vehicle energy dissipation path.

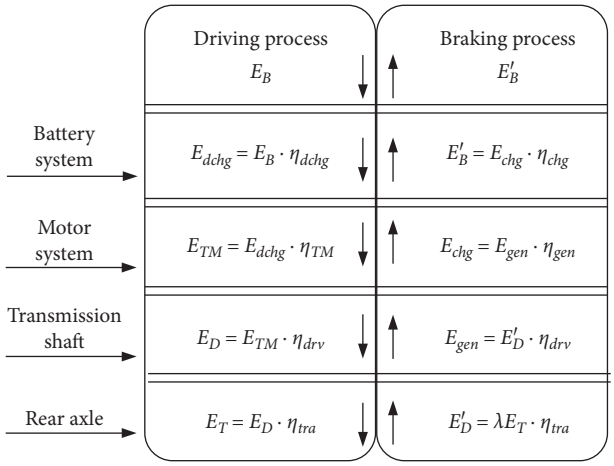


Figure 2: Vehicle energy flow distribution.

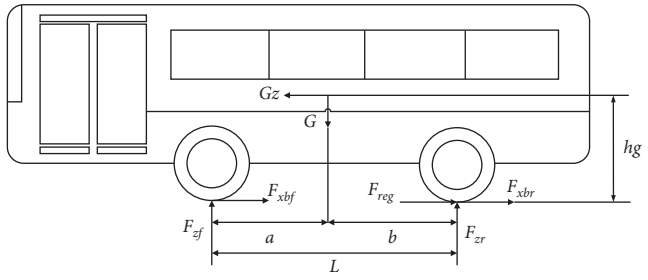


Figure 3: Analysis of the braking process of the vehicle.

cities in China are contrasted with the WTVC and CTUDC test conditions. The comparison results are shown in Table 1.

As is shown in Table 1, compared with the condition of WTVC, the characteristic parameters of CTUDC conditions are closer to the four actual typical cities in China, such as speed, proportion, acceleration, and other parameters. Due to the fact that CTUDC driving condition is more representative, it is adopted in this paper. The electric city bus studied in this paper is driven by a single motor, so the motor needs to meet the torque and power requirements when the whole vehicle is driven and braked. The distribution of speed and acceleration frequency based on CTUDC driving condition is shown in Figures 4 and 5.

As is shown in Figures 4 and 5, the low-speed part of 0~20 km/h accounts for 58.58% and the medium speed part of 20~40 km/h accounts for 35.44%, which means the vehicle speed mainly concentrate on medium and low speed. The acceleration reflects the driver's intention to change the speed. Considering the security of the standing passengers, the deceleration of the city bus is small. The maximum deceleration is 1.05 m/s^2, and the braking condition accounts for 26.57%. The condition features mentioned above are the data basis of the research on braking energy recovery in this paper. Whether the motor torque can cover the braking demand of the whole vehicle is the premise of realizing the braking energy recovery function. The frequency distribution of the vehicle braking torque demand under different vehicle speeds is shown in Figure 6.

As is shown in Figure 6, the average driving torque demand in the driving process is 1390 Nm, and the working time exceeding 2850 Nm accounts for 2.92% of the total driving condition. The peak torque of the motor which is 2850 Nm meets the vehicle driving demand under most of the driving conditions. During the braking process, the average braking torque demand is −2030 Nm, and the working time when the braking torque demand exceeds 2850 Nm accounts for 5.82%. The cost of the motor is directly proportional to the motor torque, and the whole vehicle does not recover part of the energy exceeding the motor peak torque.

2.3. Braking Regulations. In the process of braking energy recovery, the rear axle increases the motor braking force, which changes the braking force distribution between the front and rear axles. To change the threshold value, the ECE R13 braking regulations must be followed to ensure braking safety. For the double-axle bus, braking regulation is stipulated in ECE R13 that when the braking strength is less than 0.15 g, there is no requirement for the distribution of front and rear axle load. As is shown in Table 1, it can be seen that the braking strength of the city bus is less than 0.15 g no matter for the test or the actual conditions. In this case, the

Table 1: Characteristic parameters of typical driving conditions.

Name	Parameter	CTUDC	WTVC urban	Changchun	Jinan	Hangzhou	Dalian
Velocity class	Maximum speed (km/h)	60.00	66.20	50.50	39.00	47.80	50.20
	Average speed (km/h)	15.89	22.92	11.50	13.28	13.31	12.95
Proportion class	Idle time ratio (%)	29.00	16.70	16.79	34.43	32.71	29.3
	Constant speed time ratio (%)	8.70	24.60	1.64	2.36	2.85	1.63
Acceleration class	Maximum deceleration (m/s^2)	−1.05	−1.03	−0.77	−1.4	−1.21	−1.2

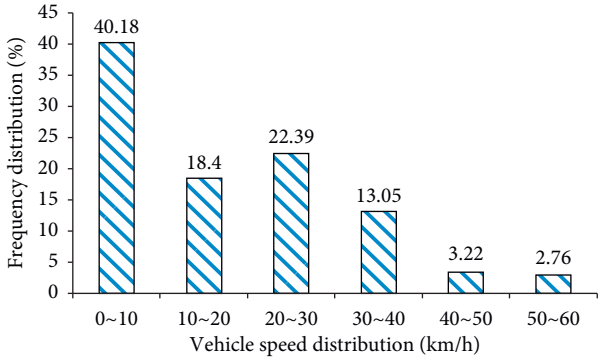


Figure 4: Vehicle speed frequency distribution of CTUDC.

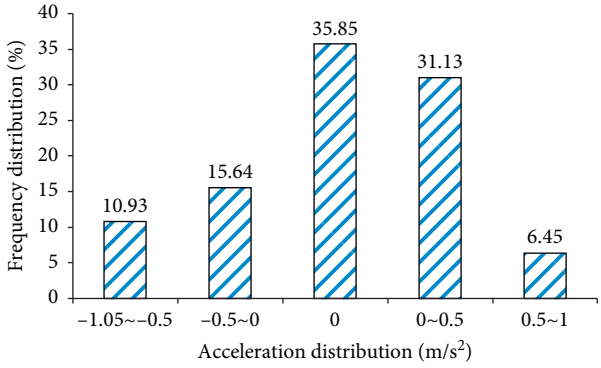


Figure 5: Acceleration frequency distribution of CTUDC.

regulation has no restriction on the distribution of braking force between two axles. Braking force distribution coefficient β is introduced as shown in

$$\beta = \frac{F_{xbf}}{F_{reg} + F_{xbf} + F_{xbr}} = 1 - \lambda - \frac{F_{xbr}}{F_{reg} + F_{xbf} + F_{xbr}}. \quad (5)$$

As is shown in Figure 3, the front axle load increases due to the forward movement of axle load during braking, and β increases accordingly. In order to improve the braking energy recovery rate, the rear axle braking force F_{xbr} should be reduced as much as possible. The superposition braking energy recovery scheme does not change the original vehicle braking system, which means mechanical and electric braking act at the same time, so the braking energy recovery has a common effect. In the decoupled scheme, electric braking takes priority, which can reduce the mechanical braking force of the rear axle to zero, so the recovery effect is better than the former.

There is no requirement for β in the CTUDC condition of the electric city bus, which can meet the requirements of braking regulations. The actual braking force is the smaller of the whole vehicle braking force and the road braking force. Considering that urban road in China is mainly asphalt pavement, the adhesion coefficient of dry asphalt pavement is usually 0.85, and that of wet asphalt is 0.5 [30]. Therefore, the adhesion coefficient of the road surface can meet the demand for vehicle braking energy recovery. The decoupled braking energy recovery scheme of an electric city bus is developed focusing on the reduction of the braking force for the rear axle, which is the key point to improve the braking energy recovery rate.

3. Scheme and Control Strategy of the Decoupled Braking Energy Recovery System

3.1. Decoupled Braking Energy Recovery Scheme. The superposition type of braking energy recovery scheme of the electric city bus does not change the original braking system, and the electric brake is applied synchronously on the basis of the mechanical brake of the rear axle. The advantage is that it does not increase any cost, and the disadvantage is that if the motor brake force applied is little, the recovery will have a common effect. Moreover, when the force is large, the rear axle will be easy to be locked first; then the vehicle may side out, which will affect the braking safety. The decoupled scheme needs to change the original braking system, which means a driver's braking intention will be judged by VCU, and the priority of motor braking will be realized.

A 12-meter electric bus is investigated in this paper. The main parameters of the vehicle and key assembly are shown in Table 2.

As is shown in Table 2, the curb weight of the vehicle is large, and the proportion between the front and rear axle loads is close to 1 : 2, when the vehicle is fully loaded. Making full use of the rear axle braking energy will be beneficial to improving the braking energy recovery rate. Therefore, a decoupled braking energy recovery scheme for electric city buses is proposed as shown in Figure 7.

As is shown in Figure 7, a relief valve, a check valve, and an air pressure sensor are installed at the front relay valve of the rear braking circuit of the bus. The relief valve is used to cut off the rear braking pipeline when braking with little intensity (e.g., less than 0.15 g), so as to realize the independent brake of the rear axle motor. On the contrary, the relief valve is opened by the large air pressure when braking with large intensity, and the electric and mechanical brake of the rear axle will work simultaneously. The check valve is used to realize that the air returns to the rear brake and ensure the air pressure balance in the brake pipeline when the braking strength changes from large to little. The air

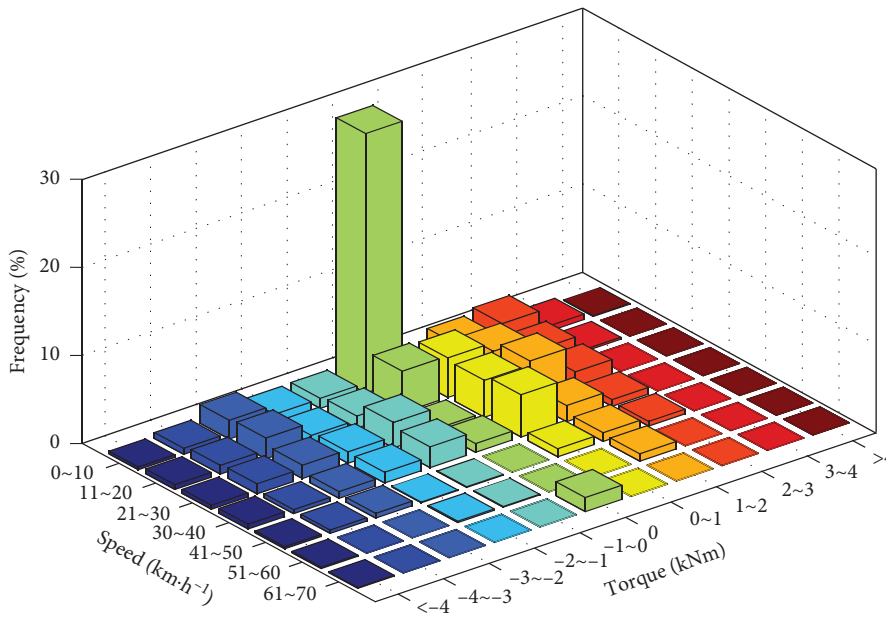


Figure 6: Frequency distribution of vehicle torque demand at different speeds.

Table 2: Main parameters of the vehicle key assembly.

Name	Parameters	Value
Vehicle	Wheel base (mm)	6100
	Distance between the centroid and front axle at curb/full weight (mm)	4095/3895
	Height of centroid at curb/full weight (mm)	758/1000
	Front and rear axle load at curb weight (kg)	4300/8800
	Front and rear axle load at full load (kg)	6500/11500
TM	Maximum torque (Nm)	2850
Battery	Capacity (Ah)	420
	Voltage (V)	384
	Maximum power (kW)	175
	Depth of discharge (%)	20~100
Transmission system	Drive shaft efficiency (%)	99
	Rear axle efficiency (%)	95

pressure sensor is used to identify the switch state of the relief valve and obtain the air pressure value of the rear brake chamber. The advantage of the scheme is that it does not change the distribution coefficient of the braking force and has good security. It is especially suitable for the little intensity braking condition of a city bus, which can ensure the priority of electric braking of the rear axle, thereby, improving the recovery rate of braking energy. Besides, it is easy to be realized in engineering with fewer mature parts be newly added; thus, it also has higher reliability and lower cost.

The whole vehicle is divided into two states during the braking process. (i) Based on the small braking strength: when the brake pedal is pressed, the front axle brake air chamber works normally while whether the rear axle brake air chamber works should be judged, owing to the fact that an overflow valve is equipped on the air pressure pipeline leading to the rear axle brake air chamber. If the air pressure of the pipeline is not enough to push the pressure limiting spring of the overflow valve, the rear axle brake air chamber will not enter the high-pressure gas. At this time, braking intention is identified by VCU through the brake pedal, and VCU judges whether the relief valve is open through a pressure sensor. According to the opening of the brake pedal and the air pressure difference between the front axle and the rear axle, VCU sends a request to the motor for the corresponding motor braking torque. (ii) Based on the large braking strength: in this case, the relief valve is open. VCU obtains the changes from the rear brake circuit, by a pressure sensor, stops motor braking, and recovers mechanical braking. At this time, the check valve is opened to realize the air pressure balance of the front and rear brake circuits.

3.2. Decoupled Braking Energy Recovery Strategy. The driver's braking intention is judged by VCU according to the brake pedal signal. Then, the braking torque of the motor acting on the rear axle is calculated by VCU in accordance with the ABS system, available battery power, vehicle speed, motor speed, and other information, while the front axle

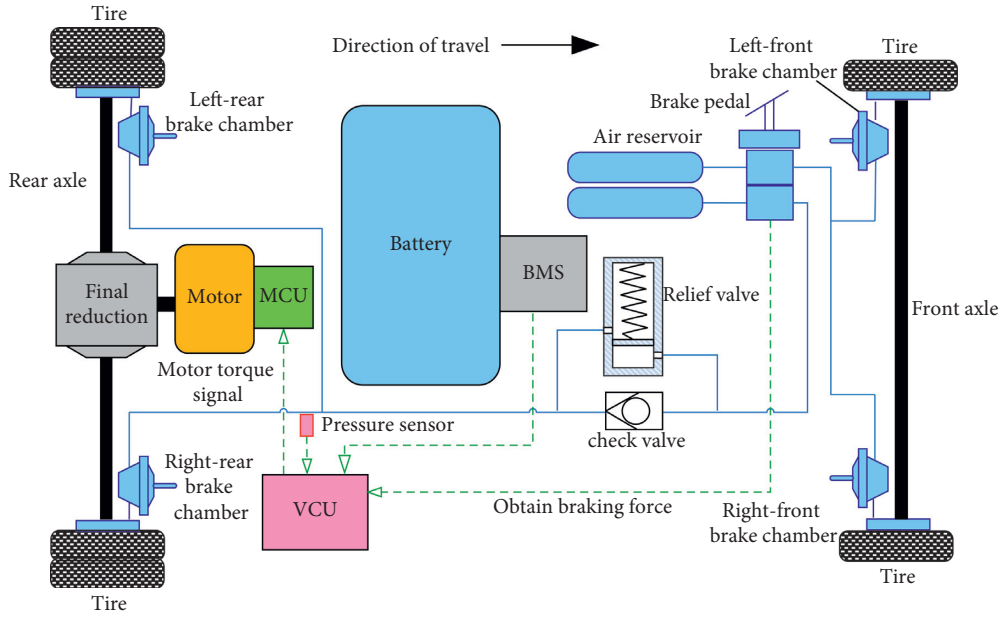


Figure 7: Decoupled braking energy recovery system for electric bus.

remains mechanical braking. The strategy is shown in Figure 8.

As is shown in Figure 8, the input signals of the braking energy recovery module include brake pedal opening signal, vehicle speed signal, ABS working signal, battery SOC, air pressure sensor signal, and the output signals including motor braking torque. The core of the algorithm is to compare the difference between rear axle braking torque demand and motor available torque. The strategy is shown in Figure 9.

As is shown in Figure 9, the driver's demand for braking force is reflected by the opening of the brake pedal. If the opening is confirmed, F_{br_req} will be confirmed accordingly. Air pressure of the brake wheel cylinder of the rear axle is represented by air pressure sensor through voltage value, and the actual air pressure value of the rear axle is uploaded to VCU. According to the conversion relationship between the air pressure of the rear axle and the mechanical braking torque, the actual mechanical braking torque of the rear axle F_{br_fric} can be obtained by the method of linear interpolation. The braking torque demand of the motor F_{reg_req} can be obtained by calculating the difference of the total braking torque of the rear axle F_{br_req} and the mechanical braking torque of the rear axle F_{br_fric}. The available torque of the motor F_{reg_act} is restricted by the output capacity of the motor and the feedback capacity of the battery. The former is mainly affected by the external characteristics and temperature of the motor, and the latter is mainly restricted by SOC and temperature of the battery cell. In order to ensure the battery life, its performance usually declines when SOC is high and the temperature is low. If the required braking torque of the motor F_{br_req} is higher than the available torque of the motor F_{reg_act}, it will be adopted as the required braking torque of the motor, and the insufficient part will be provided by the mechanical braking torque of the rear axle. If the required braking torque of the motor F_{br_req} is lower than the available torque of the motor F_{reg_act}, the required braking torque of the rear axle F_{br_req} will be provided totally by motor, and, at this time, the actual mechanical braking torque of the rear axle F_{br_fric} is zero, which is shown in

$$\begin{cases} \text{if } F_{br\_req} > F_{req\_act}; \quad sgn\left(F_{br\_fric}\right) = F_{br\_req} - F_{req\_act}, \\ \text{if } F_{br\_req} < F_{req\_act}; \quad sgn\left(F_{br\_fric}\right) = 0. \end{cases} \tag{6}$$

4. Results and Discussion

4.1. Simulation Verification. The main parameters of the whole vehicle and key assembly of electric bus studied in this paper are shown in Table 2, and the whole vehicle simulation model is built in AVL-CRUISE, as shown in Figure 10.

The map of motor efficiency used in the model is shown in Figure 11.

As is shown in Figure 11, during the driving process, the motor is in the high-efficiency area when the speed is 1000~1500 r/min (corresponding to the vehicle speed of 25~36 km/h) and the torque is 800–1500 Nm, which is basically consistent with the distribution of vehicle torque demand shown in Figure 6, and the driving efficiency of the motor is equivalent to the feedback efficiency. The battery system can be regarded as an equivalent circuit composed of open-circuit voltage and equivalent internal resistance in series [31]. The open-circuit voltage and equivalent internal resistance are functions of battery SOC, and the characteristics of the battery cells are shown in Figure 12.

As is shown in Figure 12, with the increase of SOC, the open-circuit voltage of the cell also becomes larger, and the internal resistance of the battery system is the same. When SOC is more than 20%, the internal resistance of the battery system is relatively small and stable. When SOC is less than

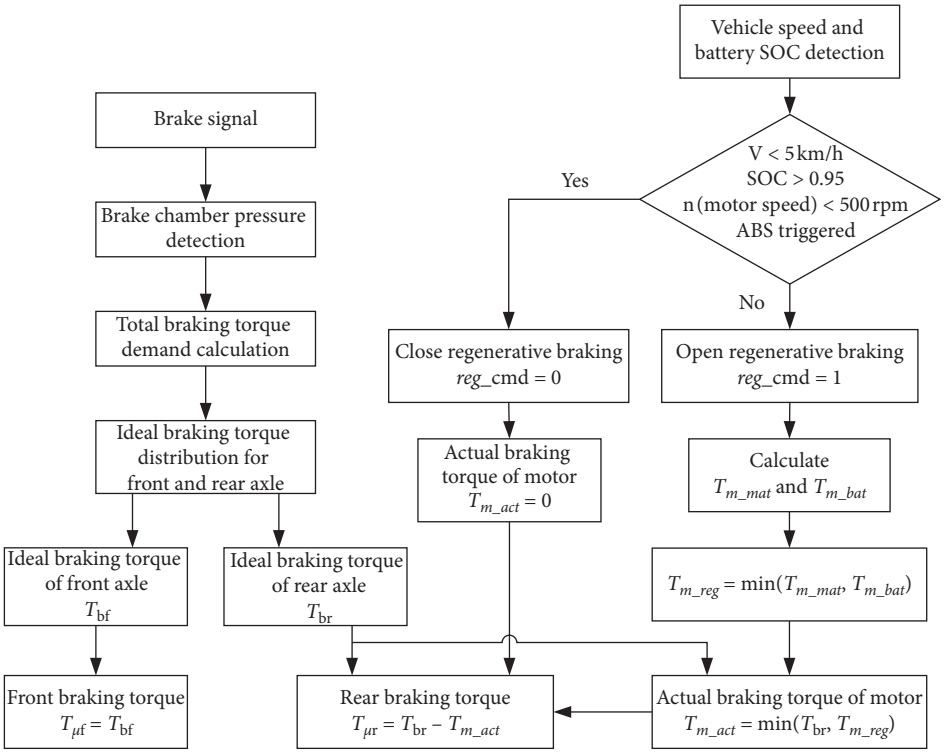


Figure 8: Decoupled braking energy recovery strategy flow chart.

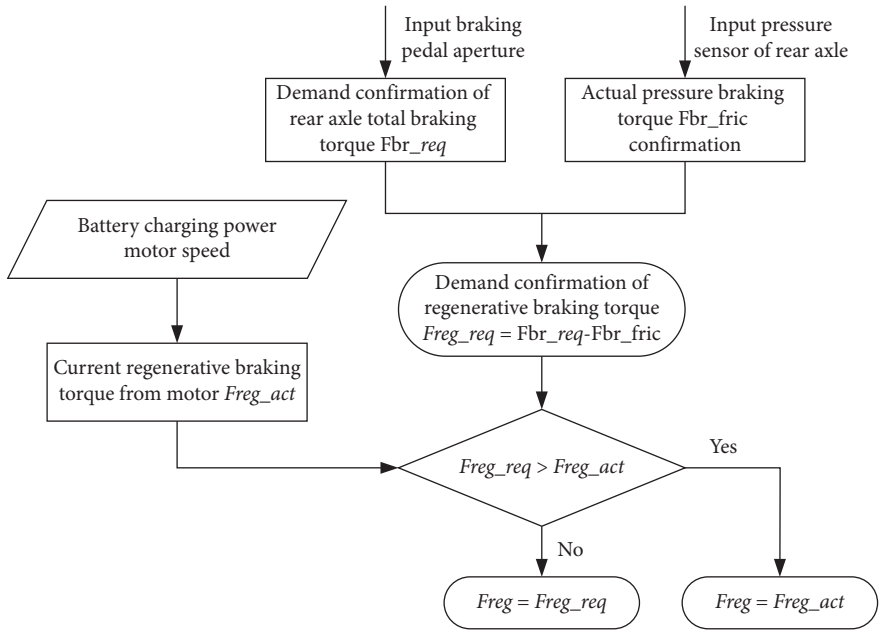


Figure 9: Decoupled control strategy of braking energy recovery.

20%, the internal resistance of the battery system increases sharply, corresponding to 80% of DOD as shown in Table 2. The decoupled braking energy recovery strategy is shown in Figures 9 and 10. The superposition braking energy recovery strategy calculates the braking torque of the motor based on the linear interpolation according to vehicle speed and brake pedal opening. Combined with the actual brake pedal opening which is usually 15~30% [32], the braking torque of the motor acting on the rear axle at different vehicle speeds is shown in Figure 13.

As is shown in Figure 13, when the vehicle speed is less than 10 km/h, the recoverable mechanical energy of the whole vehicle is small and the motor braking is not involved considering the vehicle braking smoothness. Under the same

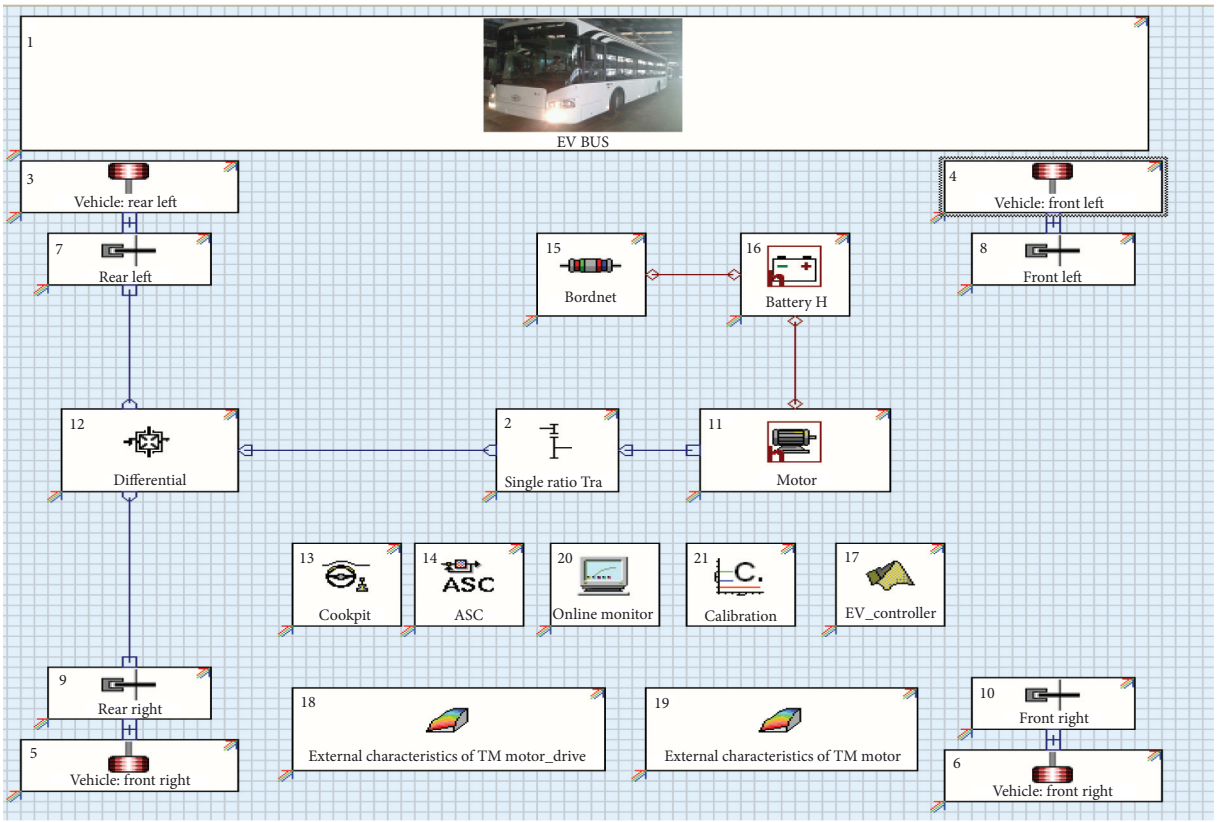


Figure 10: Physical model of performance simulation.

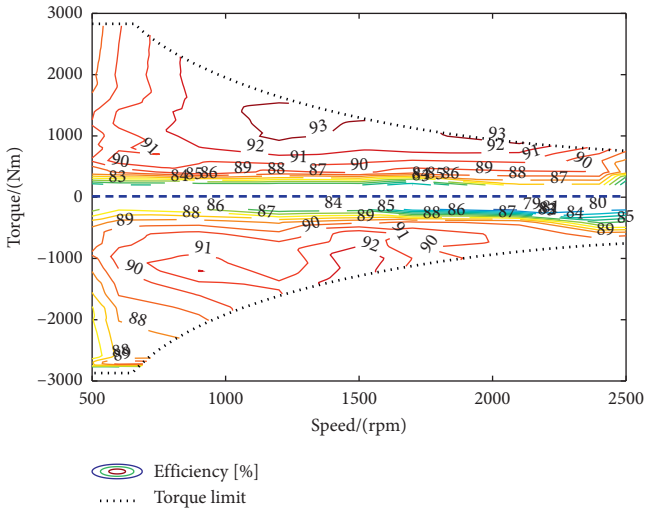


Figure 11: Motor efficiency MAP.

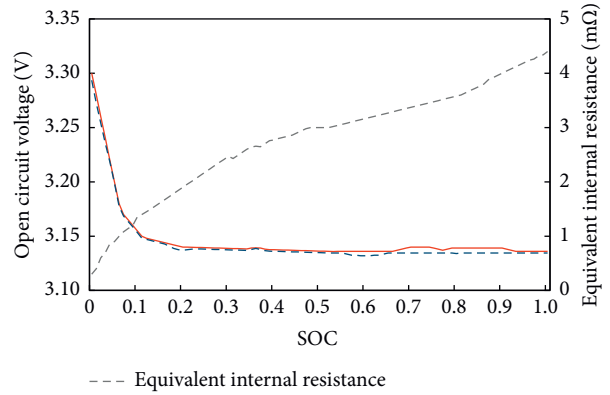


Figure 12: Changes of open-circuit voltage of the cell and internal resistance with SOC.

brake pedal opening, when the vehicle speed is in 10~55 km/h, the motor braking torque becomes larger correspondingly with the increase of the vehicle speed. When the vehicle speed is more than 55 km/h, the motor braking torque decreases correspondingly with the increase of the vehicle speed, mainly because the motor torque starts to decrease after the motor speed is higher than the base speed point. Based on CTUDC condition, vehicle energy consumption results of different loads under decoupled and superposition types of the braking energy recovery scheme are shown in Table 3.

As is shown in Table 3, as for the energy-saving effect compared with superposition type, the larger the vehicle curb weight is, the more significant the decoupling braking energy recovery scheme is. When the vehicle is fully loaded, there is 8.01% as power saving, while the half load and empty load, respectively, correspond to 5.17% and 3.01%. The main reasons are as follows. (i) The braking strength under CTUDC condition is small, and the motor peak torque

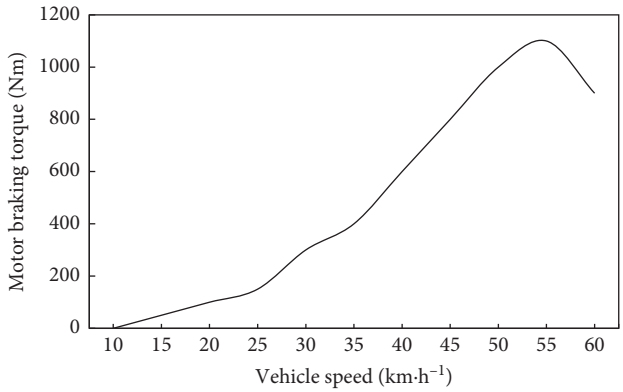
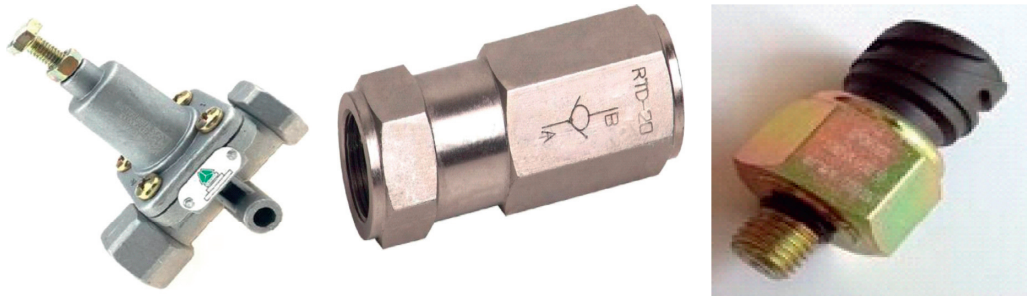


Figure 13: Braking torque of motor at different speeds.

Table 3: Vehicle energy consumption under different loads.

Load type	Simulation power consumption/(kWh/100 km)	
	Superimposing	Decoupling
Full load	93.6	86.1
Half load	79.3	75.2
Empty load	66.5	64.5



(a) (b) (c)

Figure 14: New parts. (a) Relief valve. (b) Check valve. (c) Air pressure valve.

covers the vehicle braking demand more than 95%. Compared with the superposition braking energy recovery scheme, the motor of the decoupled type can be more fully involved in the rear axle braking, while the mechanical braking of the rear axle is less involved. (ii) As is shown in Table 2, with the axle load change from zero to full, the rear axle load increases by 500 kg compared with the front axle load. The main reason is that except for sitting passengers, standing passengers are mainly concentrated in the wheelbase and the rear suspension. The rear axle load increases more than the front axle load, and the base of mechanical energy conversion at the rear axle is more than that with an empty load, so there is more braking energy that can be recovered.

4.2. Test Verification. The original electric bus adopts the superposition braking energy recovery scheme. Under CTUDC condition, it completes the energy consumption tests without braking energy recovery function and with superposition recovery function. After that, it is transformed into a sample vehicle with a decoupling braking energy recovery function. The restructuring scheme is shown in Figure 7, and the newly added parts are shown in Figure 14.

The actual road test was carried out in the test field of Tongliao, Jilin Province, China. The vehicle speed and mileage information were collected by VBox, and the driving data was collected by CANoe. CTUDC is adopted for the whole vehicle test condition, and the actual test speed follows the curve shown in Figure 15.

As is shown in Figure 15, the difference between the test speed and CTUDC condition speed is less than 2 km/h, which means the actual speed follows CTUDC well and the test data is valid. Statistics are made for the power consumption of the whole vehicle in three states: without braking energy recovery function, superposition, and decoupled braking energy recovery functions, respectively. The results are shown in Table 4.

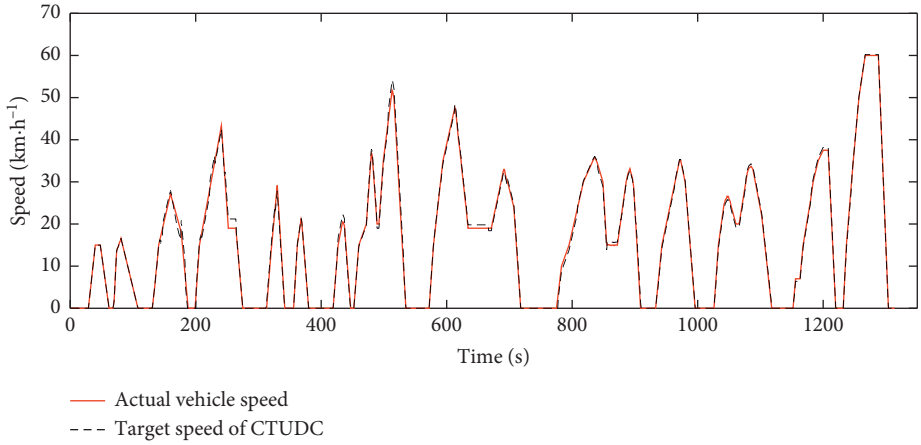


Figure 15: Test condition and actual speed.

Table 4: Test data under different braking energy recovery strategies.

Load type	Measured power consumption/(kWh/100 km)		
	Without braking energy recovery function	Superposing	Decoupling
Full load	122.14	89.28	82.97
Half load	110.04	81.36	76.09
Empty load	91.43	67.89	64.99

As is shown in Table 4, compared with the superposition braking energy recovery scheme, the decoupled one has 7.07% as energy-saving at full load and 6.48% and 4.27% at half load and empty load, respectively. The error between the measured results and the simulation results is within 1.5%, which verifies the feasibility and effectiveness of the scheme. Combined with Table 4 and (3), the energy-saving degree of the whole vehicle η_{reg} can be calculated. The energy-saving degree of the whole vehicle with the superposition braking energy recovery function is 26.91% at full load, 26.06% at half load, and 25.75% at empty load, and that with decoupled function is 32.07% at full load, 30.85% at half load, and 28.92% at empty load. Aiming at energy-saving degree, based on the same vehicle model, only under the premise of changing braking energy recovery function, compared with superposition type, the decoupled one energy-saving efficiency increases by 5.17% at full load, 4.79% at half load, and 3.17% at empty load. Under the load situation mentioned above, the energy-saving efficiency of the whole vehicle accounts for about 74% of the braking energy recovery rate. According to the energy-saving efficiency model of the whole vehicle, there are efficiency losses of the motor, battery, and transmission systems. The energy dissipation path is the same, so the overall efficiency loss is basically the same, which also verifies the effectiveness of the vehicle energy-saving model.

5. Conclusions

(1) Through identifying the energy dissipation path of the electric city bus, analyzing the whole vehicle force in the braking process, and establishing the whole vehicle energy-saving model, which contains the model constant of assembly efficiency, such as motor, battery, and transmission system, the whole vehicle energy-saving rate is proportional to the braking energy recovery rate, and the braking energy recovery rate is strongly related to the condition parameters such as speed and deceleration. Through analyzing the characteristic parameters of the urban condition and the single motor with rear-motor rear-wheel-drive configuration, it is concluded that reducing the mechanical braking force of the rear axle can effectively improve the recovery rate of braking energy.

(2) A decoupled braking energy recovery scheme and algorithm is proposed. In this scheme, relief valve, check valve, and air pressure sensor are added to the rear brake air circuit of the whole vehicle, so that the mechanical braking of the rear axle is not involved in the case of small braking intensity. The core of the algorithm is to get the actual mechanical braking torque of the rear axle according to the air pressure value of the rear brake air circuit and calculate the difference of total braking torque demand of the rear axle and the actual mechanical demand, compared with the available torque of the motor, and the ideal braking torque of the motor acting on the rear axle can be obtained. This scheme is particularly suitable for urban conditions, in which the newly added three parts are mature products with low cost and high reliability, which is conducive to engineering.

(3) Based on the scheme and algorithm mentioned above, the single factor energy consumption test is

carried out on the same vehicle. The road test results show that compared with superposition type, the decoupling type braking energy recovery rate is increased by 4~7% under different loads, and the energy-saving degree is increased by 3~5%, which verifies the feasibility and effectiveness of the scheme. The ratio of the test results of the whole vehicle energy-saving degree and the braking energy recovery rate both is 74%, which verifies the validity of the vehicle energy efficiency model.

(4) The decoupled braking energy recovery scheme proposed in this paper realizes the decoupled brake of the rear axle. In the next step, the front axle brake air circuit is to be modified in order to further improve the energy-saving rate and braking energy recovery rate and reduce the energy consumption of the whole vehicle. Meanwhile, braking smoothness of the decoupled scheme should be further researched.

Abbreviations

WTVC: World transient vehicle cycle
HAS: Hydraulic actuation system
DP: Dynamic programming
VCU: Vehicle control unit
TM: Traction motor
EMB: Electronic mechanical brake
DOD: Depth of discharge
SCB: Slip control boost
CTUDC: Chinese typical urban driving cycle
ECB: Electric city bus
SOC: State of charge
HILS: Hardware in the loop simulation.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Nature Science Foundation of China with Grant no. 51507013.

References

[1] M. Ehsani, Y. Gao, and A. Emadi, *Modern Electric,hybrid lectric and Fuel Cell Vehicles: Fundamentals, Theory, and Design*, CRC Press, New York, NY, USA, 2nd edition, 2009.

[2] S. Hano and M. Hakiai, "New challenges for brake and modulation systems in hybrid electric vehicles and electric vehicles," in *Proceedings of the 39th JSAE Conference*, Tokyo, Japan, May 2011.

[3] G. Sovran, "The impact of regenerative braking on the powertrain delivered energy required for vehicle propulsion," *SAE Paper*, vol. 12, no. 4, pp. 891–906, 2011.

[4] J. Tong and L. Zhang, "Analysis of domestic bus market performance of 2018 and outlook of 2019," *Commercial Vehicle*, vol. 38, no. 1, pp. 25–28, 2019.

[5] K. Nanda and C. Shankar, "Cooperative control of regenerative braking and friction braking for a hybrid electric vehicle," *Journal of Automobile Engineering*, vol. 230, no. 1, pp. 103–116, 2016.

[6] T. Okano, S. Sakai, and T. Uchida, "Braking performance improvement for hybrid electric vehicle based on electric motor's quick torque response," in *Proceedings of the 19th Electric Vehicle Symposium*, EVS, Pusan, Republic of Korea, October 2002.

[7] D. Kim and H. Kim, "Vehicle stability control with regenerative braking and electronic brake force distribution for a four-wheel drive hybrid electric vehicle," *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, vol. 220, no. 6, pp. 683–693, 2006.

[8] Q. Guo, Z. Zhao, P. Shen, X. Zhan, and J. Li, "Adaptive optimal control based on driving style recognition for plug-in hybrid electric vehicle," *Energy*, vol. 186, no. 11, pp. 824–837, 2019.

[9] K. Itani, A. De Bernardinis, Z. Khatir, and A. Jammal, "Comparison between two braking control methods integrating energy recovery for a two-wheel front driven electric vehicle," *Energy Conversion and Management*, vol. 122, no. 8, pp. 330–343, 2016.

[10] L. Chu, D. Liu, H. Liu et al., "A Study on the evaluation method of braking energy recovery in battery electric vehicle," *Automotive Engineering*, vol. 39, no. 4, pp. 471–479, 2017.

[11] B. Qiu and Q. Chen, "Evaluation method of regenerative braking for electric city bus," *Journal of Mechanical Engineering*, vol. 48, no. 16, pp. 80–85, 2012.

[12] L. Chu, J. Cai, and Z. Fu, "Research on brake energy regeneration evaluation and test method of pure electric vehicle," *Journal of Hua Zhong University of Science and Technology*, vol. 42, no. 1, pp. 18–22, 2014.

[13] S. Gino, "The impact of regenerative braking on the powertrain-delivered energy required for vehicle propulsion: 2011-01-0891," in *Proceedings of the SAE 2011 World Congress & Exhibition*, Detroit, MI, USA, April 2011.

[14] S. Pan, Z. Song, and X. Wang, "Control strategy for electromechanical braking of electric vehicle," *Control Engineering of China*, vol. 24, no. 2, pp. 309–314, 2017.

[15] Y. F. Xin, T. Z. Zhang, H. G. Zhang, Q. Zhao, J. Zheng, and C. Wang, "Fuzzy logic optimization of composite brake control strategy for load-isolated electric bus," *Mathematical Problems in Engineering*, vol. 2019, Article ID 9735368, 14 pages, 2019.

[16] S. Q Li, B Yu, and X. Y Feng, "Research on braking energy recovery strategy of electric vehicle based on ECE regulation and I curve," *Science Progress*, vol. 17, no. 1, pp. 762–779, 2019.

[17] D. Zhao, L. Chu, N. Xu, C. Sun, and Y. Xu, "Development of a cooperative braking system for front-wheel drive electric vehicles," *Energies*, vol. 11, no. 2, pp. 377–401, 2018.

[18] H. Guo, H. He, and X. Xiao, "A predictive distribution model for cooperative braking system of an electric vehicle," *Mathematical Problems in Engineering*, vol. 2014, Article ID 828269, 11 pages, 2014.

[19] Y. Lu, L. Tong, Y. Chen et al., "Study of regenerative braking strategy for electric vehicle based on fuzzy logic control," *China Science Paper*, vol. 12, no. 4, pp. 398–402, 2017.

[20] J. W. Ko, S. Y. Ko, I. S. Kim, D. Y. Hyun, and H. S. Kim, "Cooperative control for regenerative braking and friction braking to increase energy recovery without wheel lock," *International Journal of Automotive Technology*, vol. 15, no. 2, pp. 253–262, 2014.

[21] G. Li and Z. Yang, "Energy saving control based on motor efficiency map for electric vehicles with four-wheel independently driven in-wheel motors," *Advances in Mechanical Engineering*, vol. 10, no. 8, pp. 64–82, 2018.

[22] E. Xing, P. Wang, X. Shang et al., "Regenerative braking technology for electric vehicle," *Science Technology and Engineering*, vol. 18, no. 25, pp. 116–127, 2018.

[23] J. Brady and M. O'Mahony, "Development of a driving cycle to evaluate the energy economy of electric vehicles in urban areas," *Applied Energy*, vol. 177, no. 5, pp. 165–178, 2016.

[24] S. Zhang and X. T. Zhuan, "Study on adaptive cruise control strategy for battery electric vehicle," *Mathematical Problems in Engineering*, vol. 2019, Article ID 7971594, 14 pages, 2019.

[25] L. Yao, L. Chu, F. Zhou, M.-H. Liu, Y.-S. Zhang, and W.-R. Wei, "Simulation and analysis of potential of energy-saving from braking energy recovery of electric vehicle," *Journal of Jilin University*, vol. 43, no. 1, pp. 6–11, 2013.

[26] Y. Xiong, Q. Yu, S. Yan et al., "Research on the influence of working conditions characteristic parameters on energy consumption for plug-in hybrid electric vehicle," *Automobile Technology*, vol. 12, no. 12, pp. 23–28, 2019.

[27] J. Wu, C. Zhang, and N. Cui, "Fuzzy energy management strategy for a hybrid electric vehicle based on driving cycle recognition," *International Journal of Automotive Technology*, vol. 13, no. 7, pp. 1159–1167, 2012.

[28] R. Günther, T. Wenzel, M. Wegner, and R. Rettig, "Big data driven dynamic driving cycle development for busses in urban public transportation," *Transportation Research Part D: Transport and Environment*, vol. 51, no. 5, pp. 276–289, 2017.

[29] B. Lorenzo, D. Massimo, and P. Marco, "Development of driving cycles for electric vehicles in the context of the city of Florence," *Transportation Research Part D: Transport and Environment*, vol. 47, no. 3, pp. 299–322, 2016.

[30] L. Yang, Y. Hu, and B. Yan, "Optimal charge depleting control of plug-in hybrid electric vehicles based on driving condition," *Automotive Safety and Energy*, vol. 8, no. 1, pp. 87–96, 2017.

[31] B. Zhou, C. Tian, Y. Song, and X. Wu, "Control strategy of AFS based on estimation of tire-road friction coefficient," *Journal of Hunan University Natural Sciences*, vol. 44, no. 4, pp. 16–23, 2017.

[32] V. H. Johnson, "Battery performance models in ADVISOR," *Journal of Power Sources*, vol. 110, no. 2, pp. 321–329, 2002.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 2903157, 11 pages
https://doi.org/10.1155/2020/2903157

Research Article

PCOI: Packet Classification-Based Optical Interconnect for Data Centre Networks

Rab Nawaz Jadoon,[1,2] **Mohsin Fayyaz**,[3] **WuYang Zhou**,[1] **Muhammad Amir Khan**,[3] **and Ghulam Mujtaba**[3]

[1]*School of Information Science and Technology, University of Science and Technology of China, Hefei 230000, China*
[2]*Department of Computer Science, COMSATS University, Islamabad-Abbottabad Campus, Islamabad 22060, Pakistan*
[3]*Department of Electrical and Computer Engineering, COMSATS University Islamabad, Abbottabad Campus, Islamabad, Pakistan*

Correspondence should be addressed to Rab Nawaz Jadoon; rabnawaz@mail.ustc.edu.cn and WuYang Zhou; wyzhou@ustc.edu.cn

Received 1 January 2020; Revised 21 June 2020; Accepted 30 June 2020; Published 17 July 2020

Guest Editor: Van Huy Pham

To support cloud services, Data Centre Networks (DCNs) are constructed to have many servers and network devices, thus increasing the routing complexity and energy consumption of the DCN. The introduction of optical technology in DCNs gives several benefits related to routing control and energy efficiency. This paper presents a novel Packet Classification based Optical interconnect (PCOI) architecture for DCN which simplifies the routing process by classifying the packet at the sender rack and reduces energy consumption by utilizing the passive optical components. This architecture brings some key benefits to optical interconnects in DCNs which include (i) routing simplicity, (ii) reduced energy consumption, (iii) scalability to large port count, (iv) packet loss avoidance, and (v) all-to-one communication support. The packets are classified based on destination rack and are arranged in the input queues. This paper presents the input and output queuing analysis of the PCOI architecture in terms of mathematical analysis, the TCP simulation in NS2, and the physical layer analysis by conducting simulation in OptiSystem. The packet loss in the PCOI has been avoided by adopting the input and output queuing model. The output queue of PCOI architecture represents an M/D/32 queue. The simulation results show that PCOI achieved a significant improvement in terms of throughput and low end-to-end delay. The eye-diagram results show that a good quality optical signal is received at the output, showing a very low Bit Error Rate (BER).

1. Introduction

DCNs are indispensable entities that enable many of today's services like social networking, search engines, e-mail, and so on. DCNs should be able to satisfy the Quality of Service (QoS) requirements of a huge number of customers in companies like Microsoft, Google, Yahoo, eBay, IBM, and so on. These companies have data centres which have at least 50,000 nodes in a single data centre. Such a massive scale infrastructure needs energy to work. In 2010, the energy consumed by data centres was 1.5 percent of global power consumption. In 2012, the energy consumed by data centres was 120 billion kilowatts [1]. Apart from energy consumption, the DCNs are also growing in the annual network traffic. According to Cisco, the DCNs' traffic is growing at a compound rate of 25 percent up to 2019, reaching to 10 zettabytes (ZB) per year [2]. The data centre traffic is now measured in ZB, and by 2021 more than 95 percent of the total traffic will be terminated and originated in the DCNs. The major concern in designing a DCN is reduction in its power consumption, which impacts the operational expenditure (OPEX). Optical switch technologies are the choice of future [3].

Such challenges of growing traffic demand and energy consumption can be addressed by the introduction of optical interconnects. The optical interconnects in DCNS are able to

provide ultrahigh transmission bandwidth in an energy- and cost-efficient way. There are some obvious benefits of using optical interconnects in DCNs such as large port count, long reach communication, ability to reconfigure, large extinction ratio, and dealing with traffic heterogeneity. Optical DCNs have many communicating nodes, so they have large arrays of similar optical components. Optical components used in optical data centres follow fixed physical laws related to light. Interaction between optical components depends upon the changes in the properties of light brought by the optical components. Optical switching is a key component of today's high-performance communication networks. Optical interconnect is a box with N inputs and N outputs. At any time, the internal interconnection of optical components establishes paths from the inputs to the outputs. Congestion can occur when the number of source nodes trying to access a destination node exceeds the capacity of the destination node, which makes queuing unavoidable. Queuing can be done on the input side, the output side, or both [4].

The signal degradation due to optical components can introduce Bit Error Rate (BER) in optical interconnects. The signal degradation can be measured by analysing the eye diagram of the signal. The eye diagram is created by superimposition of zeros and ones of the optical digital data stream. The eye diagram of the optical signal can give information about jitter, noise, signal amplitude, duty cycle distortion, fall time, and bit period. Different types of signal degradations include reduced amplitude, changed shape, introduction of noise, and change in the fall time or rise time of pulse. Different types of optical components are responsible for such signal degradations. Optical components used in optical data centres include couplers, Directly Modulated Masers (DML), Vertical Cavity Surface Emitting Laser (VCSEL), Fibber Delay Lines (FDL), Fiber Bragg Grating (FBG), and SOA. Couplers cause reduction in amplitude, DMLs and VCSELs cause change in shape of signal, FDLs cause delay in bits, FBGs cause addition of noise, and SOAs introduce spikes in the optical signal.

The signal degradation compensation can be achieved using various methods. Reduction in amplitude is compensated by using amplifiers or increasing the amplitude of transmitted signal. Optical amplifiers can reduce the BER by boosting the amplitude of weak optical signal [5]. Use of DML reduces the cost of communication system as it eliminates the need for separate modulator. However, they introduce nonlinear changes in shape of optical signal which need to be compensated. These changes can be compensated by using predistortion circuits [6]. Changes in shape of optical signal are also introduced by VCSEL. Feedforward equalization is used to compensate these changes [7].

When SOAs are used in optical DCNs, they cause spikes at the start of bit. These spikes gradually reduce with time, depending upon the transition time of SOA. Such distortion can be compensated by increasing the length of bit period, such that it is greater than the transition time of SOA. This results in the reduction of supported bit rate of the system. The equalizer adaptation algorithm can mitigate the timing jitter caused by FDL, which works bitwise to detect the amount of time shift [8]. Uniform noise introduced by FBG can be mitigated by either amplifying the optical signal or increasing the transmitter power.

The objective of this paper is the proposal of a novel architecture which presents various benefits over existing architectures and investigate its performance on the basis of the following techniques:

(1) Mathematical analysis of the input and output queuing is performed

(2) Simulation of TCP protocol

(3) Eye-diagram analysis is also conducted to measure the signal degradation and BER at physical layer

In the rest of paper, Section 2 describes related work of various optical architectures, Section 3 describes the PCOI architecture, Section 4 describes the system model for input and output queue analysis, Section 5 describes the input and output queuing analysis of PCOI architecture, Section 6 describes the TCP simulation of PCOI architecture, Section 7 presents the physical layer analysis in terms of eye diagram and Bit Error Rate (BER), and Section 8 presents the conclusions and future work.

2. Related Work

Various architectures have been proposed in the literature which encounter various problems of optical networks and exploit benefits of optical components. These architectures can be classified into the categories which include (a) reconfigurable architectures, (b) low latency architectures, (c) low blocking probability architectures, (d) low power consumption architectures, (e) scaling link bandwidth architectures, and (f) high radix architectures.

2.1. Reconfigurable Architectures. These architectures have the ability to change the network topology based on change in traffic patterns. Reconfigurable architecture mentioned in the literature includes [9–11] as explained subsequently.

In [9], the architecture presented is based on Wavelength Selective Switch (WSS) and MEMS switch. Every server has a unique wavelength, which is multiplexed together according to destination rack. WSS groups the wavelengths based on destination racks and sends groups of wavelengths to MEMS ports. The topology manager is responsible for configuring WSS and MEMS switch for proper function.

In [10], a reconfigurable architecture based on Arrayed Waveguide Grating Router (AWGR) and Tunable Wavelength Converter (TWC) is presented. Header is extracted from each packet and sent to the control plane. The payload of packet keeps on waiting in FDL until the control unit makes decision about the wavelength to be set for each packet through TWC. Based on change in wavelength, the AWGR routes the packet to proper destination port.

The architecture in [11] is made up of MEMS switches. The control unit is responsible for configuration of MEMS switches. As MEMS switches are slow to configure, so this architecture is more suitable to circuit-switched applications as compared to packet-switched application.

The architecture in [12] uses Software Defined Network (SDN) with optical switching. The building block is a pod, which hosts several racks. The ToRs are connected using a star topology. The switching within the same pod is performed passively using optical filtering. The network can be scaled by using pods in a ring topology. Each ring has WDM traffic and add/drop multiplexing to and from ring is performed on a per-wavelength basis. The data plane exists in TDMA, where time slots are accessed for rack to rack communication.

2.2. Low Latency Architecture. Low latency architectures are the ones which have distributed control which include [10, 13].

In [10], a broadcast-select Spanke-type architecture is presented which minimizes the control decisions in the network to reduce the latency. The packets from source nodes are broadcast to all destination nodes. At the destination node, Wavelength Selector (WS) selects the appropriate destination node. This architecture is scalable without affecting the latency.

In [13], the network is flattened into three stages. The first stage is called input module (IM), the second stage is called central module (CM), and the third stage is called output module (OM). Each module is made up of AWGR and a scheduler. The scheduler at each stage configures the wavelength of flows to be directed to appropriate output port of AWGR. This is also scalable without affecting the latency of the system.

2.3. Low Blocking Probability Architecture. These are the architecture which show a high number of successfully transmitting nodes out of total transmitting nodes at the same time. These architectures include [14–16].

The architecture in [14] is a three-stage network. The first and third stages are based on AWGR, whereas the second stage is a time buffer. Collisions are avoided by using the time buffers. The packets wait in the second stage for contention to be resolved in the third stage.

In [15], the contention is avoided by using space-wavelength multiplexing and broadcast-select scheme. The select unit has two functions: first to select the correct spatial group and second to select correct wavelength. The select units are made up of SOAs. The central scheduler controls the SOA gates. There are two receivers on each destination node, which further reduces the blocking probability.

In [16], the Reflective SOA (RSOA) is behaving as the mutex element. In case of contention, when multiple input ports are trying to access the same destination port, the RSOA grants access to only one input port to transmit. The sender node only starts to transmit the data if a positive acknowledgment is given by RSOA. The acknowledgment of RSOA is made up of reflected optical power. In this case, multiple sender nodes are trying to send to the same output port the reflected power if RSOA drops.

2.4. Low Power Consumption Architectures. The architectures with low energy consumption are [17, 18]. The architecture in [17] has highly distributed control, which makes it scalable to large port counts. Electronic buffer is implemented on each node, which has multiple queues, and each queue has a distinct wavelength. Multiple flows can arrive at a single output port, which can be resolved by Wavelength Selector (WS) partially. When packets from different queues of the same node go to a single destination node, the contention cannot be resolved and only one flow is forwarded. However, when multiple flows from different nodes arrive at the destination, such contention is resolved by WS by changing its wavelength.

In [19], the architecture is made up of clusters, boards, and nodes. Each board has multiple nodes and multiple boards are connected using optical wavelength multiplexing. Vertical Cavity Surface Emitting Lasers (VCSELs) are used to eliminate the need for external modulator, which conserves energy.

In [18], the two main optical communicating devices are AWGR and Microring Resonators (MRR). AWGRs are passive optical devices, which do not need any external energy input. MRRs are also very low power and high bandwidth devices. This architecture is entirely made up of low energy consumption devices, which save energy.

In [20], an architecture is presented which minimizes energy consumption by using a combination of optical cross-connects and WDM rings. Optical circuit switches provide dedicated nonblocking circuits. WDM rings are interrack switching elements.

The architecture in [21] uses space and wavelength multiplexing. The architecture is divided into cards and each card has multiple nodes. Each node has a distinct wavelength. It uses multiple level addressing. The card is selected using couplers in space domain and port is selected in wavelength domain by the wavelength.

2.5. Scaling Link Bandwidth Architectures. These architectures include [9, 19, 22]. In [19], Vertical Cavity Surface Emitting Lasers (VCSELs) are used, which behave as both the source of light and the modulator. VCSELs eliminate the need for external modulators in the transmitters. The unused VCSELs are shared among transmitters to increase their transmitting bandwidth, which makes this architecture a scaling bandwidth architecture.

In [23], the architecture uses the ring topology and wavelength multiplexing to interconnect the nodes. The add/drop multiplexer and Wavelength Selective Switch (WSS) are used in the ring to select the wavelength for destination node and add the wavelength to send data by sender node. This architecture is not much scalable.

In [24], an architecture is presented in which AWGRs are used but the interconnection between racks is based on passive optical components. The mixed linear integer programming model is used for wavelength assignment, which ensures that a single wavelength is assigned between two server groups. The directionality of AWGR in ensured so

that flows are always directed from input ports to output ports.

The work in [25] presents an architecture which uses optics and commodity switches. The backplane is a switchless core made up of bus-based fiber rings. The architecture is divided into sectors and each sector is made up of ToR switches and interconnection pods. Within a sector, electronic switching is performed. The absence of switches implies full bisection bandwidth for single-hop communication.

2.6. High Radix Architectures. High port count in the architectures is achieved if the network performance is not affected by increasing the number of nodes in the network.

In [26], a distributed and scalable optical packet switch architecture is presented. It is based on Arrayed Waveguide Grating Router (AWGR) and SOA. A single AWGR can connect multiple ToRs. It has a modular structure and nonblocking nature of AWGR makes it scalable. It shows lower latency and low blocking probability.

In [9], an architecture is presented which can achieve dynamic configuration of link bandwidth. For overall connectivity, Microelectromechanical System (MEMS) switch is used, which uses micromirrors to deflect the light to the output port. Wavelengths are grouped using Wavelength Selective Switch and directed to respective destination racks.

The architecture in [10] presents a reconfigurable architecture, which is based on AWGR. Wavelength conversion mechanism is used at the input port to rout the light paths to respective destination ports. Tunable Wavelength Converters (TWC) are used for this purpose to change the wavelength of light signal. The labels are processed to extract the destination address, which is used to change the wavelength of input signal.

The architecture in [13] flattens the networks, which reduces the number of hops in the network. It is a multistage architecture, which is achieved using three modules: the input module, central module, and output module. Each module is composed of TWC and AWGR. Wavelength switching is used in each module to direct the light path to appropriate input port of next module.

The PCOI architecture was proposed to minimize the routing complexity by exploiting packet classification and use of queuing. It further reduced energy consumption by using passive optical components. The next section describes the PCOI architecture briefly.

3. The PCOI Architecture

The proposed architecture which avoids contention under various traffic patterns is shown in Figure 1. The nodes are arranged in the form of racks. Each rack is assigned a unique wavelength. This architecture exploits the benefits of input and output queuing, wavelength multiplexing, and space multiplexing. Each rack has a classify module. This module classifies the packets on the basis of destination rack and puts them in respective queues. The packets from queues are converted into optical domain by using Electrical-to-Optical (E/O) conversion. The optical flows from a single rack have a unique wavelength, thus requiring as many wavelengths as the number of nodes. The flows of a single destination rack from multiple sender racks have different wavelength, which allows simultaneous data to be received from multiple sender nodes, making the all-to-one communication possible. Optical power combiner at receiver rack collects various wavelength flows, which are later demultiplexed and given to $1 \times N$ switch at receiver rack. The optical power combiner is used because of its passive nature and low cost; it can combine various frequency optical signals without the need for external source. The packets of each source rack are kept separate using different wavelengths.

The demultiplexed wavelengths contain the packets of respective source racks. Out-of-band signalling is used to reduce the header extraction time. At the destination rack, packets of each wavelength go through a $1 \times N$ switch, which is also shown in Figure 2. The switch is designed after header detection packet goes to its destination node, where they are collected by the output queue in case of contention.

The $1 \times N$ switch in Figure 2 is a high port optical switch. The header information and payload of packet are separated. The header is sent to control unit, whereas the payload is sent to SOA after some delay, which is equal to the processing time of control unit. The delay is provided by the FDL which exits before the SOA. The payload keeps traveling in FDL until the decision is made by control unit to turn on the respective SOA of destination port. The rest of SOAs are turned off by the control unit. As a result, the payload only travels to the destination port of switch as determined by the destination address of header.

4. System Model

The PCOI architecture represents a nonblocking switch as it involves self-routing mechanism, where a packet finds its path to the desired output port. This is shown in Figure 3. The transmitter and receiver nodes represent the input and output queue system. Head-of-Line (HOL) blocking problem arises when multiple source nodes try to send to the same destination node, which results in stoppage of transmission of packets from head of input queue which are sending packets to that destination. This is shown in Figure 4. N input and output ports are assumed, and transmission is synchronous, which means that packets are sent from the input ports to different output ports at the same time. When the buffer of the output port is full, it sends backpressure signal to the respective input queue. After receiving the backpressure signal, the input queue stops transmitting the packets to that output port. The packets of different input queues that want to send packets to the same output buffer are made to wait in the input queue. The backpressure mechanism prevents overflow in the output buffer. The effect of backpressure on the performance of optical interconnect for different buffer sizes is analysed. The number of ports is assumed to be large, that is, $N \longrightarrow \infty$. The PCOI architecture achieves speed-up at the output port by receiving multiple flows using wavelength, code, and space multiplexing. It results in significant improvements in delay reduction and throughput enhancement. The throughput

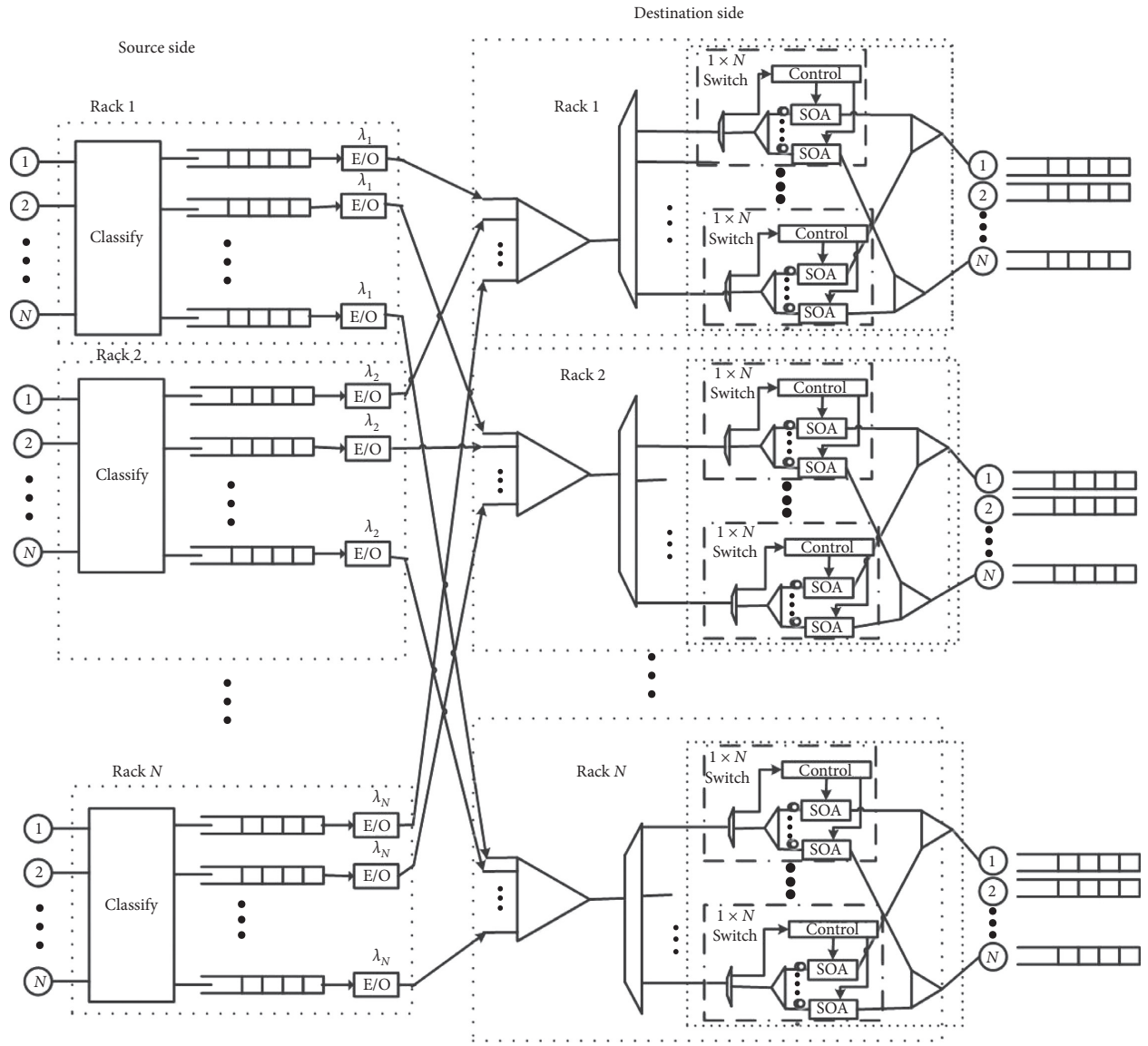


Figure 1: PCOI architecture.

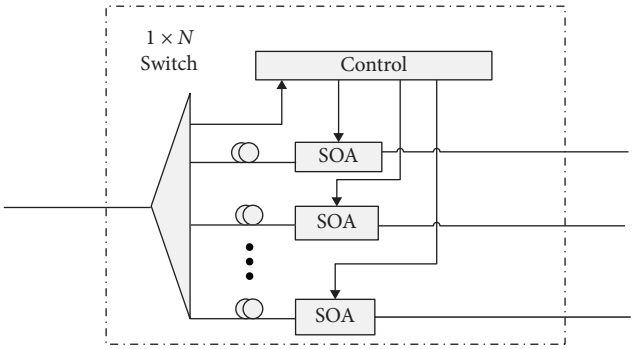


Figure 2: $1 \times N$ switch.

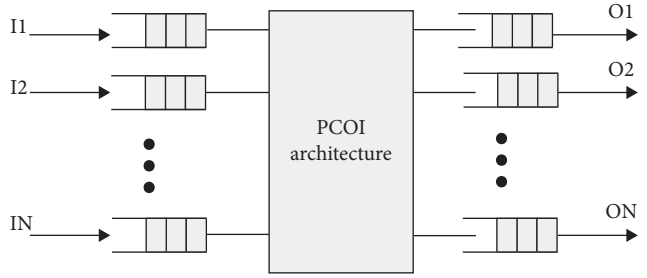


Figure 3: Switch model for the PCOI architecture throughput enhancement.

enhancement is evident from Figure 4, where each destination ports 32 servers in the queuing model. It can process 32 streams at the same time, which increases the throughput and reduces the queuing delay.

5. The Input and Output Queue Analysis of PCOI Architecture

The PCOI architecture behaves as an N\times N switch. The packets are assumed to be of fixed length and fixed transmission time. The traffic is randomly distributed, which

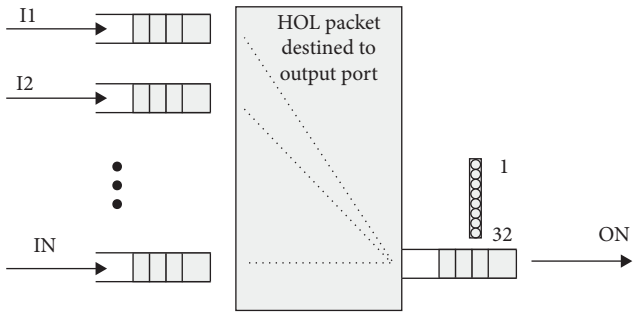


Figure 4: Queue model used to evaluate the PCOI architecture.

means that a packet from an input port can be sent to any of the N output ports with equal probability, which is $1/N$. The output queue operates on First Come First Serve (FCFS) basis. If the output port is idle, the packets go through the PCOI architecture directly. The performance of the PCOI architecture is affected by the output buffer size "b." The arrival process is Bernoulli, as the probability of arrival in a time slot defines the load. The efficiency of the PCOI architecture is measured, based on maximum throughput and average delay. The packet delay is made up of three components: (i) waiting time in input queue until the head; (ii) waiting time at the head of the input queue due to HOL, and (iii) waiting time at the output queue due to contention of the output port.

The traffic of input port on average is "p" packets per unit time. Packets of the input ports are independent of each other. There is a possibility that more than "b" packets are trying to access the same output port. The output port of the PCOI architecture can receive 32 multiplexed streams of packets. The waiting time at the input buffer until it begins the transmission to the output port represents an M/D/32 system. The M/D/32 queue can be treated as M/D/1 queue by scaling basic unit of time; the p of M/D/1 queue becomes $32p$ for M/D/32 queue [27].

The analysis is based on [28], which represents an M/D/1 queue. It is modified for the RPL architecture by replacing p with $32p$. Closed form expressions for the average delay and maximum throughput are derived. Thus, the total average delay, which is shown in Figure 5, is given by

$$D = \frac{\overline{Q_b^2} + (64p-1)\overline{Q_b} + (32p)^2}{64p\left(1-32p-\overline{Q_b}\right)} + \frac{\overline{Q}}{32p}, \tag{1}$$

$$\overline{Q_b} = \overline{Q} - b\left(1-32p_0\right) + \sum_{j=1}^{b}(b+1-j)p_j, \tag{2}$$

$$\overline{Q_b^2} = \overline{Q^2} - 2b\overline{Q} + b^2\left(1-32p_0\right) - \sum_{j=1}^{b}(b+1-j)^2 p_j, \tag{3}$$

where $\overline{Q}$ ad $\overline{Q^2}$are the first and second moments of waiting customers of M/D/c queue and p_j are the steady-state probabilities given by

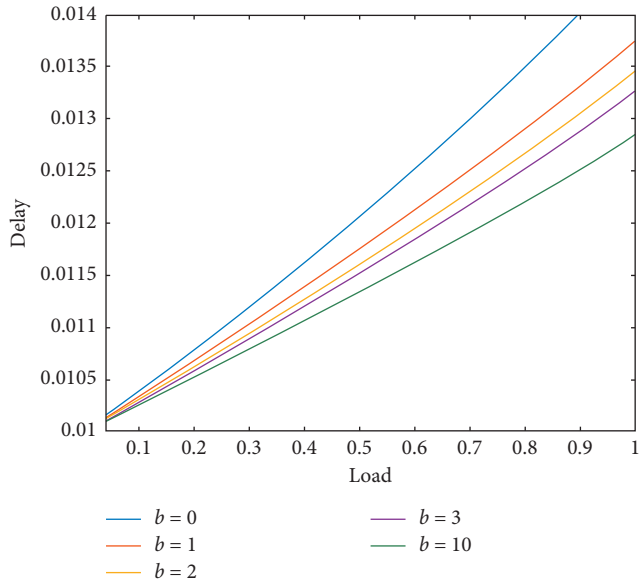


Figure 5: Delay analysis of the PCOI architecture.

$$\overline{Q} = \frac{(32p)^2}{2(1-32p)}, \tag{4}$$

$$\overline{Q^2} = \frac{(32p)^2\left((32p)^2 - 32p + 3\right)}{6(1-32p)^2}, \tag{5}$$

$$p_0 = (1-32p), \tag{6}$$

$$p_1 = (1-32p)\left(e^{32p}-1\right), \tag{7}$$

$$p_j = (1-32p)\sum_{i=1}^{j}(-1)^{j-i}e^{i(32p)}\left[\frac{(i\,32p)^{j-i}}{(j-1)!} + \frac{(i\,32p)^{j-i-1}}{(j-i-1)!}\right]. \tag{8}$$

The corresponding p. g. f for the steady-state probabilities is

$$P(z) = \frac{(1-32p)(1-z)}{1-ze^{32p(1-z)}}. \tag{9}$$

The probability that a packet arrives at the head of the input queue and experiences delay due to backpressure, which is shown in Figure 6, is given by

$$P_{b,s} = \frac{1}{32p}\left(1 - \sum_{i=0}^{b+1} p_i\right). \tag{10}$$

6. Simulation

The performance of the PCOI architecture is measured for different output queue sizes using TCP. Simulation setup consists of a 512×512 network. The network is implemented in NS2, a discrete event simulator developed for research and educational use [29]. It is an open-source software. For many-to-many communication pattern, each node

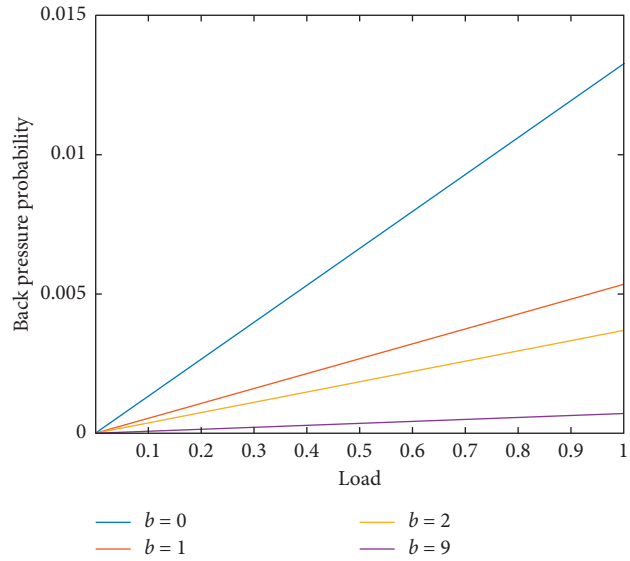


Figure 6: Backpressure analysis of the PCOI architecture.

transmits to a randomly chosen destination node. The random destination node address is generated according to a uniform random variable. Figure 7 shows the number of sender nodes trying to access the destination nodes. This number fluctuates between 1 and 5. The median of these two numbers is 3, so for many-to-many communication pattern, it is assumed that 3 nodes are trying to access a destination node.

TCP is used as the transmission protocol. The traffic has a constant bit rate. The load is varied from 1 Mbps to 10 Mbps, which accounts for the normalized load from 0.1 to 1 in the figures in the following sections. The link delay is assumed to be 10 ms. TCP is responsible for 90 percent of Internet traffic. TCP can adapt the transmission of packets according to bandwidth; it avoids congestion and retransmits lost packets. By keeping the output queue buffer size constant, the network initially shows a higher average delay due to the slow-start phase of the TCP. TCP has a slow-start phase because all the nodes try to transmit at the same time and congestion causes packets of input nodes to wait to avoid packet loss. Performance of TCP is analysed by changing the buffer size of the output queue in NS2. The Drop-Tail queue is implemented in which the last arrived packet is dropped if the queue is full. The overall throughput of TCP increases by increasing the buffer size, whereas the average delay of the packets decreases by increasing the buffer size of the output queue. Constant bit rate (CBR) application is used to generate traffic on the nodes. TCP uses reliable congestion control, in which acknowledgment is created by the destination to know whether packets have been received. Lost packets are interpreted as a congestion signal. Initially, the average delay is high but as the load increases, the average delay decreases. It uses a dynamic congestion window, which grows rapidly initially and then increases slowly as it reaches a threshold. When congestion is detected, it drops rapidly. The output of the NS2 simulation is in the form of a trace file. The required information of delay and throughput is obtained from the trace file by performing text processing using a Perl script. The small Perl programs are used as filters to extract the required information from the text. The trace file has 12 fields, the first field is of event type, the second field is the time at which the event occurs, the third field is the input node of link at which event occurs, the fourth field gives output node of link, the fifth field is of the packet type (TCP, UDP, AGT, etc.), the sixth field gives the packet size in bytes, the seventh field includes flags, the eighth field gives the flow id, the ninth field gives the address in the form "node.port," the tenth field gives the address in the same form, the eleventh field gives the packet sequence number of the network layer, and the twelfth field shows the unique id of the packet.

Figure 8 shows the TCP delay for different output buffer sizes. Initially delay is large due to slow-start phase of the TCP and it reduces later.

Figure 9 shows the throughput performance of the TCP for different buffer sizes. It is seen that for lower buffer size there is sudden drop in throughput for higher load because of sudden change in congestion window size. For lower value of output buffer size, the congestion occurs early by increasing the load.

Figure 10 shows the maximum achievable throughput as a function of output queue buffer size. All-to-one communication pattern is the worst traffic encountered in any network.

Figures 11 and 12 show that PCOI architecture shows performance benefit for this traffic pattern for all-to-one and many-to-many communication paradigm, respectively.

7. Physical Layer Analysis

The physical layer of PCOI architecture is simulated in the OptiSystem [30], to measure the BER and signal degradation. The signal degradation can be analysed from the eye diagram, which is shown in Figure 13. The simulation in

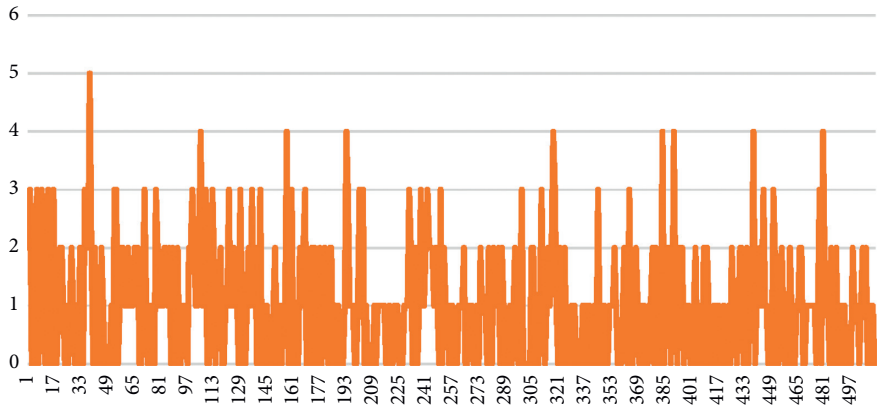


Figure 7: Number of sender nodes trying to access a destination node.

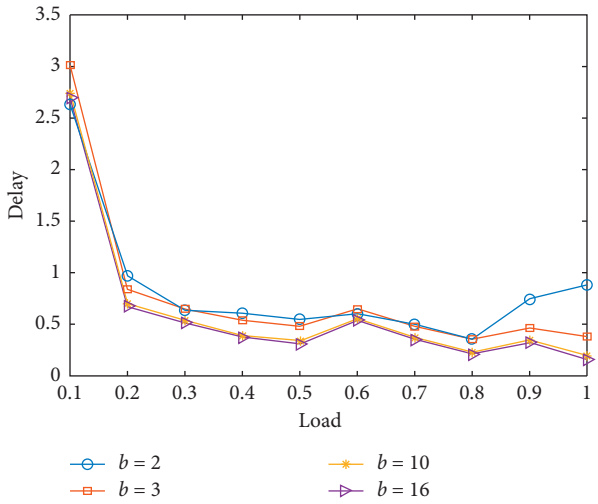


Figure 8: Delay analysis of the PCOI architecture.

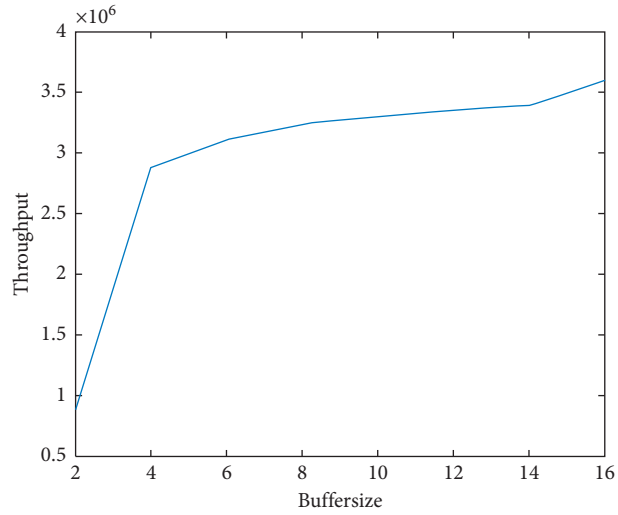


Figure 10: Maximum throughput simulation of the PCOI architecture.

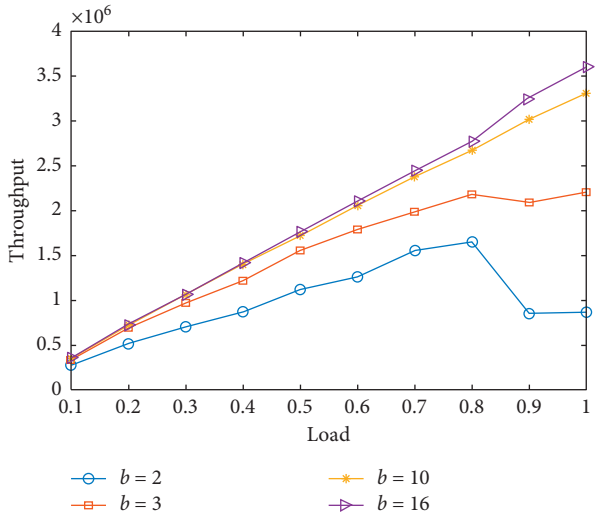


Figure 9: Throughput simulation of the PCOI architecture.

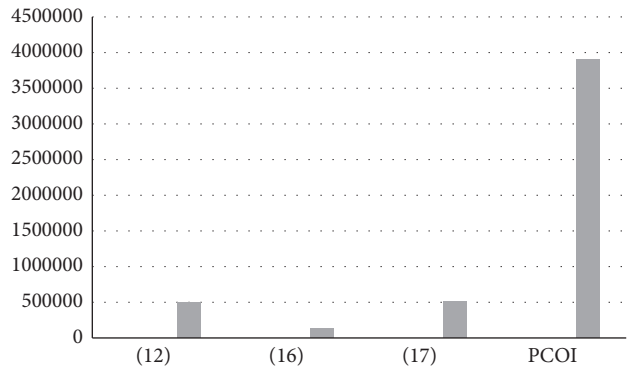


Figure 11: Throughput comparison for all-to-one communication pattern.

OptiSystem is carried out by analysing the optical path of each architecture. The optical signal consists of pseudo-random bit sequence. For bit generation, non-return-to-zero (NRZ) pulse is used. NRZ pulse is used because the pulses have more energy and have additional rest state besides zeros and ones, which gives a bigger margin between two logic levels. Mach Zehnder (MZ) modulator generates the optical

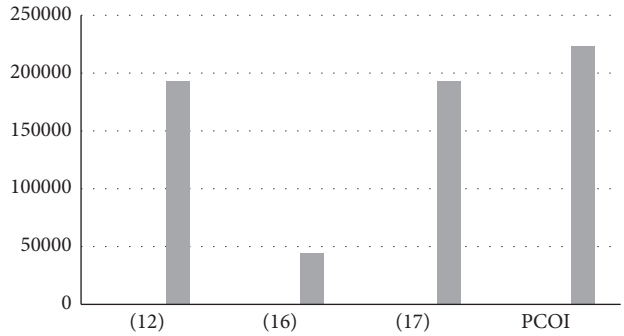


Figure 12: Throughput comparison for many-to-many communication pattern.

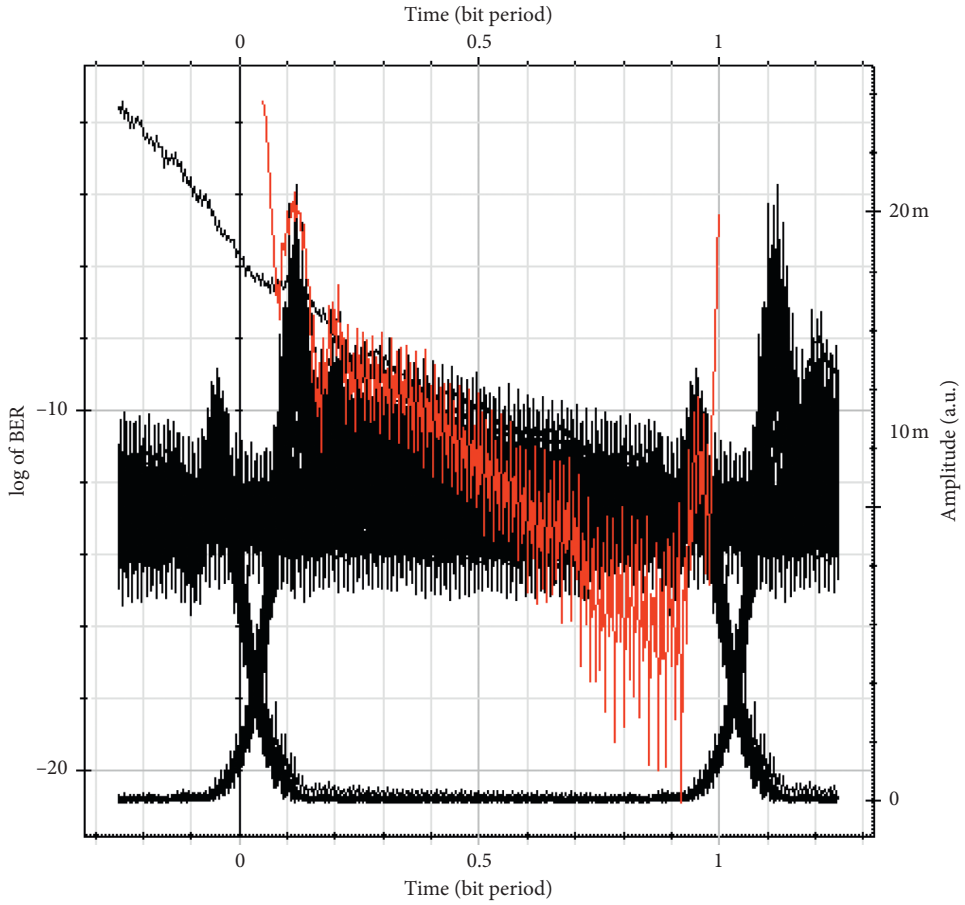


Figure 13: Eye diagram of the PCOI architecture.

signal which corresponds to the bitstream. MZ modulator is popular for low power, compact size, and monolithic integration. The transmit power is varied from −12 dBm to 5 dBm. The eye diagram of the PCOI architecture in Figure 13 is calculated using the 0 dBm transmit power. It has shown a very good eye opening. This shows that received signal in the PCOI architecture has enough signal quality to be detected. The eye opening shows that it has low jitter. The log (BER) of −10 is achieved for this eye diagram.

8. Conclusions and Future Work

The modelling of optical data centres is very important, as it helps in making important decisions about their performance. It is very important to consider at design time of optical data centres which optical components to use. The choice of optical components directly affects the quality of received signal. If the quality of received signal is bad, it can adversely affect the Bit Error Rate of the system. Ambiguity

in logic levels makes it difficult for the receiver to distinguish between bits. However, if an optical component is used that degrades the signal quality, then the mitigation techniques should be used to counter the effect of optical components. There are two ways to improve the performance of optical data centres, either reducing the signal degradation or making the design of optical data centre to reduce contention.

The main advantages of the PCOI architecture are routing simplicity, reduced energy consumption, scalability to large port count, packet loss avoidance, and all-to-one communication support. Routing simplicity is achieved in PCOI by using the packet classifier at the sender side, which classifies the packets based on destination rack and puts them in a queue. The reduced energy consumption is achieved in PCOI by use of passive optical components. Passive optical components are those which do not need any external power source for their working. The PCOI architecture is scalable to large port count due to lack of central controller. The PCOI architecture can avoid packet loss in worst communication patterns by exploiting the redundancy of optical components and queues to temporary store packets with collisions. All-to-one communication pattern is the worst communication pattern in the communication system, in which all the sender nodes try to access a destination node or rack at the same time. PCOI architecture supports all-to-one communication pattern by using queues and passive optical components.

The general problems seen in the PCOI architecture are the use of large number of optical components and the signal degradation. There are two main types of optical signal degradations in PCOI: one is caused by passive optical components which is simply the reduction in optical power as the optical signal passes through a passive optical component and the second is the change in the shape of optical signal as it goes through SOA. Due to signal degradation, it is not possible to improve the BER beyond a certain limit.

The future work is related to the introduction of new modulation formats to the performance of PCOI and the analysis of signal degradation imposed by new modulation formats.

Data Availability

No data were used to support this study. The authors have conducted the simulations to evaluate the performance of proposed protocol. However, any query about the research conducted in this paper is highly appreciated and can be answered by the principal authors Rab Nawaz Jadoon (rabnawaz@cuiatd.edu.pk) and Mohsin Fayyaz (mohsinf@cuiatd.edu.pk) upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

Dr. Rab Nawaz Jadoon personally acknowledges COMSATS University, Islamabad-Abbottabad Campus, and School of Information Science and Technology, USTC, Hefei, China, which extended their full support by providing the authors all key resources during the implementation and all afterward phases of this project. This work was financially supported by National Natural Science Foundation of China (Grant no. 61631018).

References

[1] B.-C. Lin, "Generalization of an optical asa switch," *Applied Sciences*, vol. 9, no. 6, pp. 1096–6, 2019.

[2] M. Moralis-Pegios, N. Terzenidis, G. Mourgias-Alexandris, and K. Vyrsokinos, "Silicon photonics towards disaggregation of resources in data centers," *Applied Sciences*, vol. 8, no. 1, p. 83, 2018.

[3] P.-A. Blanche, L. La Comb, Y. Wang, and M. Wu, "Diffraction-based optical switching with mems," *Applied Sciences*, vol. 7, no. 4, p. 411, 2017.

[4] M. J. Karol, M. G. Hluchyg, and S. P. Morgan, "Input versus output queueing on a space-division packet switch," *IEEE Transactions on Communication*, vol. 35, no. 12, 1987.

[5] S. Ruiz-Moreno, G. Junyent, M. J. Soneira, and J. R. Usandizaga, "Statistical analysis of nonlinear optical amplifier in high saturation," *IEE Proceedings J Optoelectronics*, vol. 135, no. 1, pp. 34–38, 1988.

[6] X. Lu, T. Xu, T. Liu et al., "Reduction of intermodulation distortion in directly modulated lasers: rf predistortion, electronics," in *Proceedings of the International Conference on Communications and Control (ICECC)*, pp. 4437–4439, Ningbo, China, September 2011.

[7] A. V. Rylyakov, C. L. Schow, B. G. Lee, F. E. Doany, C. Baks, and J. A. Kash, "Transmitter predistortion for simultaneous improvements in bit rate, sensitivity, jitter, and power efficiency in 20 Gb/s cmos-driven vcsel links," *Journal of Loghtwave Technology*, vol. 30, no. 4, pp. 399–405, 2012.

[8] A. C. Carusone, "An equalizer adaptation algorithm to reduce jitter in binary receivers," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 53, no. 9, pp. 807–811, 2006.

[9] A. Singh, K. Ramachandran, L. Xu, and Y. Zhang, "Proteus: a topology malleable data center network," in *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*, Monterey, CA, USA, 2010.

[10] X. Ye, Y. Yin, S. J. B. Yoo, P. Mejia, R. Proietti, and V. Akella, "DOS: a scalable optical switch for datacenters," in *Proceedings of the ANCS'10 ACM/IEEE Symposium on Architectures for Networking and Communications Systems*, La Jolla, CA, USA, October 2010.

[11] Q. Kong, S. Huang, Y. Zhou et al., "MMTDnet: a novel optical circuit switch architecture for data center networks," in *Proceedings of the Asia Communications and Photonics Conference*, Beijing China, November 2013.

[12] P. Bakopoulos, K. Christodoulopoulos, and G. Landi, "NEPHELE: an end-to-end scalable and dynamically reconfigurable optical architecture for application-aware SDN cloud data centers," *IEEE Communication Magazine*, vol. 56, no. 2, pp. 178–188, 2018.

[13] K. Xia, Y.-H. Ka, M. Yang, and H. Jonathan Chao, "A petabit bufferless optical Switch for data center networks," *Optical Networks*, Springer, Berlin, Germany, pp. 135–154, 2013.

[14] J. Gripp, J. E. Simsarian, J. D. LeGrange, P. Bernasconi, and D. T. Neilson, "Photonic terabit routers: the IRIS project, optical fiber communication (OFC)," in *Proceedings of the IEEE Conference on Optical Fiber Communication (OFC)*, pp. 1–3, San Diego, CA, USA, March 2010.

[15] R. Luijten, W. E. Denzel, R. R. Grzybowski, and R. Hemenway, "Optical interconnection networks: the osmosis project," in *Proceedings of the 17th Annual Meeting of the IEEE Lasers and Electro-Optics Society*, Rio Grande, Puerto Rico, November 2004.

[16] R. Proietti, C. J. Nitta, Y. Yin, R. Yu, S. J. B. Yoo, and A. Venkatesh, "Scalable and distributed contention resolution in awgr based data center switches using rsoa based optical mutual exclusion," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 19, no. 2, 2013.

[17] S. Di Lucente, R. P. Centelles, H. J. S. Dorren, and N. Calabretta, "Study of the performance of an optical packet switch architecture with highly distributed control in data center environment," in *Proceedings of the IEEE 16th International Conference on Optical Network Design and Modeling (ONDM)*, pp. 1–6, Colchester, UK, April 2012.

[18] B. Neel, R. Morris, D. Ditomaso, and A. Kodi, "Sprint: scalable photonic switching fabric for high performance computing (HPC)," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 4, no. 9, 2012.

[19] A. K. Kodi and A. Louri, "Energy-efficient and bandwidth-reconfigurable photonic networks for high-performance computing (HPC) systems," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 17, no. 2, pp. 384–395, 2011.

[20] S. Basu, C. McArdle, and L. P. Barry, "Scalable ocs-based intra/inter data center network with optical tor switches," in *Proceedings of the 18th International Conference on Transparent Optical Networks (ICTON)*, Trento, Italy, July 2016.

[21] O. Liboiron, I. Cerutti, R. P. Giorgio, A. Nicola, and P. Castoldi, "Energy efficient design of scalable optical multi plane interconnection architecture," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 17, no. 2, pp. 377–383, 2011.

[22] P. N. Ji, D. Qian, K. Kanonakis, C. Kachris, and I. Tomkos, "Design and evaluation of a flexible bandwidth OFDM intra data center interconnect," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 19, Article ID 3700310, 2013.

[23] F. Nathan, F. Nathan, A. Forencich et al., "A 10 μs hybrid optical/circuit electrical packet network for datacenters," in *Proceedings of the Optical Fiber Communication Conference/ National Fiber Optic Engineers Conference*, Anaheim, CA, USA, March 2013.

[24] H. Ali, E. Taisir, H. El-Gorashi, and J. M. H. Elmirghani, "High performance awgr pons in data centre networks," in *Proceedings of the 17th International Conference on Transparent Optical Networks (ICTON)*, Budapest, Hungary, July 2015.

[25] A. Kushwaha, A. Gumaste, T. Das, S. Hote, and Y. Wen, "Flexible interconnection of scalable systems integrated using optical networks (fission) data-center-concepts and demonstration," *Journal of Optical Communications and Networking*, vol. 9, no. 7, pp. 585–600, 2017.

[26] S. Chen, S. Huang, Q. Kong et al., "A distributed and scalable optical packet switch architecture for data centers," in *Proceedings of the 13th International Conference on Optical Communications and Networks (ICOCN)*, Suzhou, China, November 2014.

[27] D. Gross, C. M. Harris, J. F. Shortle, and J. M. Thompson, *Fundamentals of Queueing Theory*, John Wiley & Sons, , John Wiley & Sons, Hoboken, NJ, USA, Third edition.

[28] I. Iliadis and W. E. Denzel, "Analysis of packet switches with input and output queuing," *IEEE Transactions on Communications*, vol. 41, no. 5, pp. 731–740, 1993.

[29] Network Simulator 2 (Wiki), 2018, http://nsnam.sourceforge.net/wiki/index.php/User_Information.

[30] OptiSystem, 2018, http://www.optiwave.com.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 6541782, 22 pages
https://doi.org/10.1155/2020/6541782

Research Article

Multiobjective Optimized Dispatching for Integrated Energy System Based on Hierarchical Progressive Parallel NSGA-II Algorithm

Aidong Zeng,[1,2] **Sipeng Hao**,[1,2] **Jia Ning**,[1,2] **Qingshan Xu**,[3] **and Ling Jiang**[4]

[1]*School of Electric Power Engineering, Nanjing Institute of Technology, Nanjing 211167, China*
[2]*Jiangsu Collaborative Innovation Center for Smart Distribution Network, Nanjing 211100, China*
[3]*School of Electrical Engineering, Southeast University, Nanjing 210096, China*
[4]*State Grid Tianjin Electric Power Company Electric Power Research Institute, Tianjin 300010, China*

Correspondence should be addressed to Aidong Zeng; zengaidong@foxmail.com

Received 24 March 2020; Accepted 7 May 2020; Published 31 May 2020

Guest Editor: Sanghyuk Lee

Considering the importance of reducing system operating costs and controlling pollutant emissions by optimizing the operation of the integrated energy system, the energy supply structure of the integrated energy system and the joint multiobjective optimization dispatching structure is analyzed in this paper based on a day-ahead economic optimization dispatching model of the integrated energy system. Afterwards, the multiobjective optimization model of the integrated energy system is studied and multiobjective hierarchical progressive parallel algorithm based on improved NSGA-II is proposed according to the characteristics of the model. The algorithm improves the nondominated layer sorting algorithm, changes the convergence judgment condition while introducing the target reaching method to accelerate convergence, and introduces parallel computing technology according to the characteristics of the algorithm. The case shows that the proposed algorithm not only has advantages on the diversity in searching solutions but also can achieve better results in many aspects such as the iteration time and algorithm convergence which are required in practical engineering projects.

1. Introduction

Facing the pressure of energy shortage, energy saving, and emission reduction, building a clean, reliable, interactive, and efficient distributed cooling, heating, and power integrated energy system has become an important means to promote the innovation of efficient energy-use models and develop a low-carbon economy.

Research and development of optimized dispatching and operation technologies for integrated heating and cooling power systems is a necessary condition for vigorously developing the efficient use of distributed energy systems and also the technical basis for carrying out a multienergy system demonstration project that uses electric power as the core and optimizes the operation of a variety of energy sources, which can provide support for realizing the transformation of energy utilization ways and a prerequisite for promoting changes in production and lifestyle, and becomes an important part of the country's carrying and promoting energy reform and the third industrial revolution.

The optimal dispatching goal of the combined heat and power system is not only economical, commonly, the minimum exchange power with the bus and the minimum adjustment of equipment operating power are also available. In [1], an island-type integrated energy system in eastern China was taken as the research object, and the optimization goals were set as economical optimization and minimum power exchange with the tie bus. The operation scheme of the integrated energy system was obtained by solving the optimal scheduling model. In [2], an active distribution network (ADN) with integrated energy system was taken as the research object, and a method for managing the

congestion of ADN through reasonable dispatch of integrated energy system was proposed. In the method, a new probability-based risk multiple objective model to achieve the optimal balance between feeder congestion and economic goals was defined. In [3], a comprehensive energy system optimization plan with the goal of optimizing the performance of the combined heat and power (CHP) plant was proposed. The optimization goals in the plan were to maximize power production, minimize fuel consumption, maximize the heat used in the slurry process, and improve the CHP plant's power production efficiency.

In recent years, a number of scholars have taken environmental protection factors into the system's day-ahead scheduling model. For example, ultralow-energy consumption buildings were used as research objects in [4] to model their energy management and control their operating costs while optimizing their emission of greenhouse gases. In [5], a penalty function of polluting gas emissions was added to the system's optimization objective function, and a multiobjective opportunity planning model was established by considering factors comprehensively such as the operating costs of the combined heat and power system and overall pollutant emissions. Finally, an improved particle swarm algorithm with local and global memory was adopted to solve the optimization model.

In [6], HOMER Pro software was used to simulate the optimization of microgrid based on CHP plants. The microgrid architecture included PV arrays, diesel generators, and batteries (operating in grid-connected and island mode). Multiple operational goals included total net present cost (TNPC), cost of energy (COE), and annual greenhouse gas (GHG) emissions generated, as well as the maximization of annual waste heat recovery (WHR) and annual grid sales (GS). In [7], greenhouse gas emissions and pollutant discharge costs were incorporated into the system's operating cost function, and a multiobjective optimization model was established with the lowest comprehensive operating cost of the integrated cooling and heating power system as the objective function, and the abovementioned model was solved by particle swarm optimization algorithm. In [8], a multiobjective thermodynamic model of a paper mill's combined heat and power (CHP) system was established. In the multiobjective optimization model, three optimization goals were considered, including CHP power generation efficiency, total cost of system products, and CO_2 of the entire plant emissions, and genetic algorithms are utilized to optimize the objective function.

As mentioned above, optimal operation scheduling of an integrated energy system is no longer limited to the optimum economic dispatch; more factors such as sustainable development of environment, prevention of feeder congestion, and energy efficiency improvement need to be taken into account.

In terms of optimization model solving methods, there are currently two main types of solving methods: one is to convert multiple optimization targets into a single optimization target and then perform single-objective optimization, and the other is to directly apply intelligent multiobjective algorithms for optimization.

In the first method, for example, a multiobjective optimal scheduling model that considers both the economic and environmental objectives of a combined heat and power CHP system was proposed [9]. The optimization problem was transformed into a single-objective optimization problem by maximum fuzzy satisfying method, and then the model was solved by using an improved single-objective genetic algorithm. Finally, the validity of the proposed model and algorithm was verified by a multiobjective optimization example including a wind-solar-storage system and a CHP unit. Energy equivalent factor which converts three kinds of energy (cold, heat, and electricity) into each other equivalently was introduced [10], and a multiobjective energy-saving scheduling model for a combined heat and cold power system that includes system fuel costs, environmental emission costs, and energy coordination costs was established. Target membership function was used to convert multiple targets into single targets, and finally, quadratic programming algorithm was utilized to solve the model.

To solve the shortcoming of lacking objective basis for the selection of weight factors among multiple objective functions, the negative ideal point method was proposed in [11], which introduced subjective and objective weighting methods to determine the weight factors among various objective functions. The algorithm was applied to a combined heat and power system with the core of a gas internal combustion engine. An equal performance coefficient from the perspective of pollutant emissions was proposed to reflect the energy consumption of the system in [12], and a multiobjective optimal scheduling model of the integrated cooling, heating, and power system was established. The optimization objectives of the model include power generation costs and emission costs. The two optimization goals were transformed into a single goal for processing by the membership function. Finally, the model was solved using a nonlinear programming algorithm. Distributed energy resources (DER) and microgrids were taken as research objects in [13], and an optimization model for optimal management of MG was proposed based on the combined heat and power (CHP) system considering economic, environmental, and reliability aspects. The exchange market algorithm (EMA) and weighted factors method were used to combine three conflicting targets and treat the multiobjective problem as a single target problem, and the fuzzy satisfaction method was applied to select the best compromise solution.

The biggest problem in converting a multiobjective problem into a single-objective optimization is that it cannot fully find a set of solutions that meets the multiobjective optimization problem. The single target solution obtained is often a weight of multiple targets with less effective in balancing trade-offs or preferences for each goal. Therefore, in recent years, more and more researches have shifted the solution method to the direct use of intelligent algorithms such as genetic algorithms and particle swarm optimization to directly solve multiobjective problems and find the appropriate problem solution set.

Among the multiobjective genetic algorithms, the widely used algorithms are nondominated sorting genetic

algorithm (NSGA) and nondominated sorting genetic algorithm-II (NSGA-II) [14–16]. The sorting method of NSGA is simple and intuitive, that is, to traverse the entire population, compare the individuals one by one and then select individuals that meet the requirements until the layering work is completed in the cycle. In an m-dimensional multiobjective optimization problem, the time complexity of the sorting part is $O(mN^3)$ [17]. The NSGA is simple and easy to understand, but there are three major problems [18]: (1) it has high time complexity; (2) it does not adopt elite retention strategy; (3) it needs to specify the shared radius parameter in advance when adjusting the virtual fitness value. As an improved algorithm of NSGA, NSGA-II solved the abovementioned three problems by adopting better accounting and elite strategies which reduced the time complexity of the sorting algorithm from $O(mN^3)$ to $O(mN^2)$ [19]. NSGA-II reduces the calculation time compared with NSGA, but in some large optimization projects, there is still a long calculation time problem [20].

A cold, heat, and power integrated energy system with the core of a small biogas internal combustion engine was taken as the research object in [21]. A multiobjective optimal scheduling model with comprehensive economic benefits, environmental benefits, and energy utilization benefits was established based on energy flow characteristics and operating constraints of the system. The system's multiobjective optimization scheduling model is finally solved by using the PSO multiobjective algorithm with constraint factors to obtain the system's Pareto solution set. Time-of-use electricity price was considered and a multiobjective optimal scheduling model of a combined heat and power system with the lowest fuel cost, the lowest pollutant emissions, and the best electricity purchase and sales from public networks was established in [22]. The optimal problem was solved by using a spatially coupled PSO optimization algorithm. An example showed that the algorithm had stronger global search capabilities than traditional PSO algorithms and can obtain more reliable Pareto solutions. Carbon emissions trading was taken as the research point and pollutant emission penalty function was stripped from the system operating costs in [23], then a multiobjective optimal scheduling model for a combined heat and power system that considers carbon emissions trading costs and system overall operating costs was established. A fuzzy self-correcting PSO algorithm was used to solve the model. The example showed that the system can effectively control the carbon emissions under the premise of ensuring normal operation and obtain some transaction benefits, thus improving the system's economic operation level.

A multiobjective optimization model of the CHP integrated energy system was established considering randomness of wind power and load fluctuations in the case of off-grid mode in [24]. The optimization objective considered operating costs, comprehensive benefits, and system power supply reliability. Finally, the model was solved by NSGA algorithm to obtain the optimal solution set. A multiobjective optimization model with the goal of minimizing system operating costs and carbide emissions for a CHP integrated energy system was established in [25]. Considering the nonlinearity and high complexity of the model, modified MOPSO algorithm based on the differential evolution algorithm was used to solve the model. Finally, a laboratory integrated energy system is used as a case to verify the effectiveness of the proposed model and algorithm. Economic and environmental objectives were considered and a dual-objective optimal scheduling model for a CHP system was proposed in [26]. The model was solved using a fuzzy MOPSO algorithm based on chance constraint programming (CCP). The Pareto optimal solution set and the optimal solution under two subobjectives are given in the simulation example.

In summary, the research on the multiobjective optimization operation of the integrated energy system is mainly focused on the selection of multiple targets and the solution method of the multiobjective optimization problem. The single-objective transformation method had the disadvantage of a relatively limited solution set, the diversity of the searched solutions was insufficient, and it was easy to fall into a local optimum. However, using intelligent multiobjective algorithm to solve the problem has the problem that the solution process is lengthy and difficult to integrate into actual engineering applications.

Therefore, based on the establishment of the optimal operation model of the integrated energy system with multiple optimization objectives, this paper proposes a multiobjective hierarchical progressive parallel NSGA-II algorithm according to the characteristics of the model. The algorithm improves the nondominated layer sorting algorithm, changes the convergence judgment condition to introduce the target reaching method to accelerate convergence, and introduces parallel computing technology according to the characteristics of the algorithm, which accelerates the convergence speed of the algorithm while taking into account the diversity of the solutions searched.

The case study shows that the algorithm proposed in this paper not only has advantages on the diversity of solutions searched but also can achieve better results in many aspects such as the iteration time and algorithm convergence which are required in practical engineering projects. In practical engineering applications, the optimal operation scheme can be selected according to the needs of customers to achieve multiobjective optimized operation of the integrated energy system in the smart grid zone.

2. Energy Supply Structure of Integrated Energy System

The integrated energy system is a new type of overall energy supply system on the demand side with many advantages such as various forms of energy supply, numerous energy supply equipment, flexible energy supply schemes, and satisfactory overall economy. Generally, the energy sources of integrated energy systems include power grids, independent distributed power generation equipment, and gas companies. The loads supplied mainly include various

industrial customers, residential customers, and commercial customers around the system. Various types of customers have different cold, heat, and electricity needs, and the integrated energy system can formulate corresponding energy supply plans based on these energy needs to meet customers' load requirements.

The main energy-supplying equipment of the integrated energy system studied in this paper includes a CHP unit, which contains microgas turbines, waste heat boilers, and absorption refrigerators. Refrigeration equipment includes electric refrigerator units, refrigeration air conditioners, and cold storage devices. Heating equipment includes gas boilers, heating air conditioners, and heat storage devices. Pure electrical supply equipment includes storage batteries and photovoltaic cells.

The integrated energy system can freely trade electricity with the public grid, such as purchase electricity, from the grid when the system power supply is insufficient or the grid power price is lower than the system's power generation cost and sell electricity to the grid when the system power supply is surplus. At the same time, it is supposed that there is no gas production in the integrated energy system and all the required gas is purchased from external gas companies. The energy supply structure that reflects the energy flow in the system is shown in Figure 1.

The integrated energy system performs bilateral energy interaction with the external power grid and unidirectional energy interaction with external natural gas companies. For the CHP unit, the microgas turbine converts the chemical energy of natural gas into electricity, and the waste smoke is utilized by a waste heat boiler to generate high-temperature steam, for heating or cooling through an absorption lithium bromide unit. In the cooling system, the electric refrigerator unit can generate cold water required for cooling, the cold storage device can flexibly store and release the cold water medium according to the needs, and the space cooling load can be met by cooling water through the fan coil cooling or directly through the refrigeration air conditioner. In the thermal system, the gas boiler can generate hot water required for heating and the heat storage device can flexibly store and release the hot water medium according to the needs. The space heat load can be met by hot water through the fan tray or directly by air conditioner. The solar and storage power generation unit in the integrated energy system can exchange power with the external power grid through the public bus.

Based on this energy supply structure, this paper establishes a joint multiobjective optimization dispatching model for an integrated energy system. The input conditions for the dispatching model mainly include the next day forecast values of various loads in the system, TOU (time of use) price (electricity and gas) of the energy market, and various types of energy supply equipment models, capacities, and operating conditions in the system.

A Pareto solution set is obtained by solving the joint multiobjective optimization model of the integrated energy system, and then the dispatching instructions are issued by one solution of the Pareto solution set according to the goal preference. The input and output of the model are shown in Figure 2. The mathematical models of the energy supply equipment in the integrated energy system are established in [27].

3. Joint Multiobjective Optimization Dispatching Model of Integrated Energy System

The integrated energy system contains a variety of energy equipment; in order to satisfy various loads in the system efficiently, the energy equipment operation status in the integrated energy system must be optimized for dispatching according to the optimization goal.

Based on the formulation of the energy supply structure of the integrated energy system and the establishment of mathematical models for various energy supply equipment in the system, a multidimensional optimized operation objective function is established considering the time-of-use energy price and various equipment operation constraints. By solving the multiobjective optimization dispatching model, the purpose of reasonably formulating an operation strategy that meets the requirements of the dispatching objective is achieved.

The assumptions of the model are as follows:

(1) It is assumed that the output of the equipment is continuously adjustable within the operating constraints

(2) The equipment has no failures during the optimization operation cycle

3.1. Multidimensional Optimized Operation Objective Function. In the process of joint multiobjective operation optimization of the integrated energy system, not only the economic status of the system is considered but also the environmental indicators of the entire integrated energy system. The environmental indicators are also called emission target..

(1) The economic goal of the integrated energy system is to obtain the lowest daily operating cost which mainly includes three aspects: system fuel costs, power interaction costs with the grid, and system operation maintenance costs:

$$f_1 = \min K_{\text{Total}} = \min\left(K_{\text{Fuel}} + K_{\text{Grid}} + K_{\text{Maintain}}\right), \tag{1}$$

where f_1 is the economic goal of the system; K_{Total} is the daily comprehensive operation cost of the system; K_{Fuel} is the fuel costs of the system; K_{Grid} is the power interaction costs with the grid; and K_{Maintain} is the operation maintenance costs of the system,

wherein the fuel costs function for fuel boilers and microturbines in the integrated energy system is calculated as

$$K_{\text{Fuel}} = \sum_{t=1}^{24}\sum_{i=1}^{n_{\text{CHP}}} c_{\text{Gas}}^{t} f_{\text{CHP}i}\left(P_i^t\right) + \sum_{t=1}^{24}\sum_{i=1}^{n_{\text{GB}i}} c_{\text{Gas}}^{t} F_{\text{GB}i}^{t}, \tag{2}$$

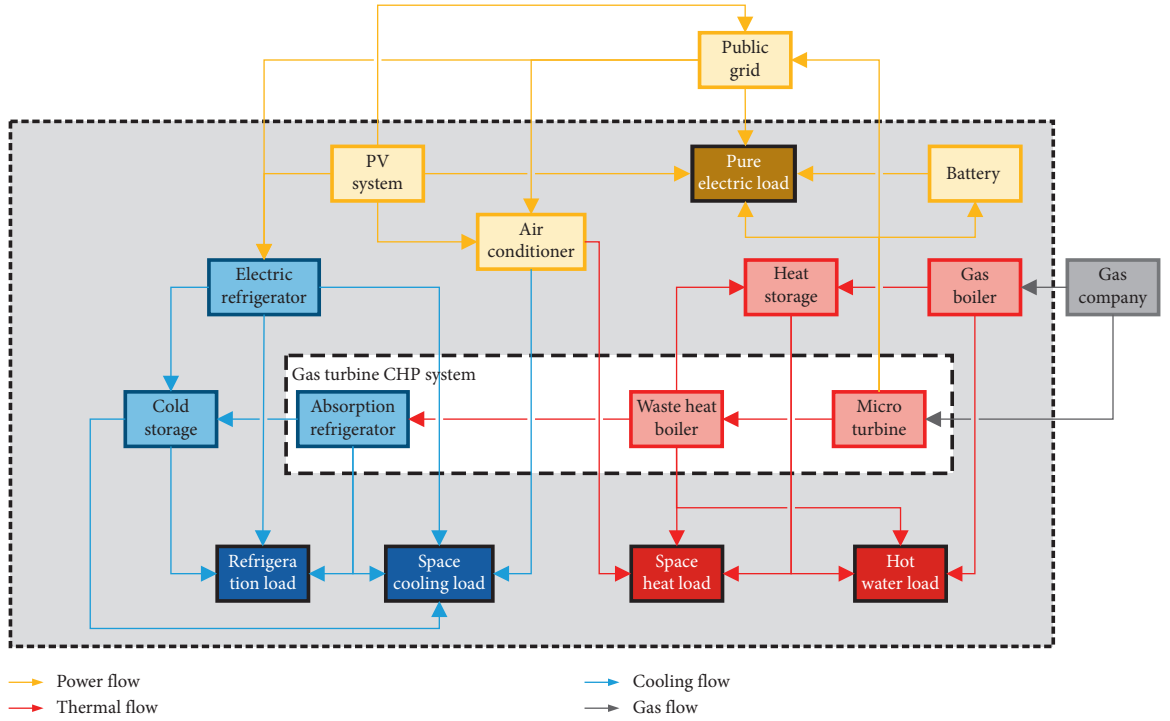


Figure 1: Energy supply structure of the integrated energy system.

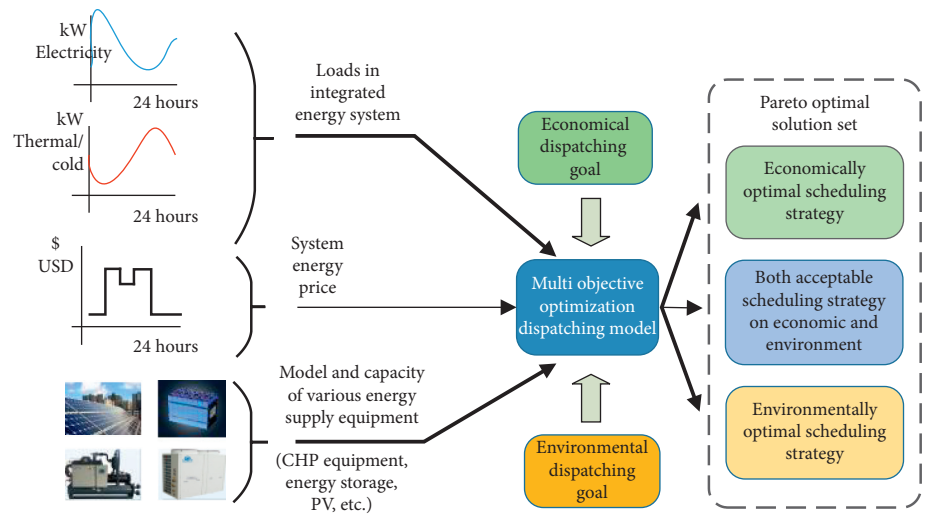


Figure 2: Joint multiobjective optimization dispatching structure of the integrated energy system.

where f_{CHPi} is the energy consumption curve function of i^{th} microturbines; P_i^t is the electric power output of i^{th} microturbines with the unit of kW; c_{Gas}^t is the hourly price of gas with the unit of \$/(kW·h) after conversion according to the calorific value of natural gas; F_{GBi}^t is the energy consumption of the i^{th} gas boiler at time t with the unit of kW; and t is the time period number with the unit of hour.

The power interaction costs function of the system with the grid is calculated as follows:

$$K_{Grid} = \sum_{t=1}^{24} c_{Grid}^t P_{Grid}^t, \tag{3}$$

where c_{Grid}^t is the hourly price of electricity and P_{Grid}^t is the hourly power exchange value of the integrated energy system with external grid.

The operation maintenance costs function of the system is calculated as

$$K_{\text{Maintain}} = \sum_{t=1}^{24} p_{\text{mWB}} Q_{\text{WB}}^{t} + \sum_{t=1}^{24} p_{\text{mAC}} P_{\text{AC}}^{t} + \sum_{t=1}^{24} p_{mEC} P_{\text{chil}}^{t} + \sum_{t=1}^{24} p_{\text{mstor}} H_{\text{in}}^{t} + \sum_{t=1}^{24} p_{\text{mstor}} H_{\text{out}}^{t} + \sum_{t=1}^{24} p_{\text{mAir}} P_{\text{cond}}^{t} + \sum_{t=1}^{24} \sum_{i=1}^{n_{\text{CHP}}} p_{\text{mCHP}i} P_{i}^{t} + \sum_{t=1}^{24} \sum_{i=1}^{n_{\text{distri}}} p_{\text{mdistri}} P_{\text{distri}}^{t} + \sum_{t=1}^{24} p_{\text{mGB}} Q_{\text{GB}}^{t}, \tag{4}$$

where $p_{\text{mCHP}i}$ is the unit power operation maintenance cost of the micro turbine; p_{mdistri} is the unit power operation maintenance cost of the distributed generation unit; p_{mstor} is the unit power operation maintenance cost of the storage devices; p_{mGB} is the unit power operation maintenance cost of the gas boiler; p_{mWB} is the unit power operation maintenance cost of the waste heat boiler; p_{mAC} is the unit power operation maintenance cost of the absorption chiller; p_{mEC} is the unit power operation maintenance cost of the compression electric chiller; p_{mAir} is the unit power operation maintenance cost of the air conditioner; p_i is the electric power output of the i^{th} micro turbine; P_{distri}^{t} is the output of the i^{th} distributed generation unit; H_{in}^{t} and H_{out}^{t} refer to the charging and discharging power of three types of energy storage equipment, respectively; Q_{GB}^{t} is the heating power of the gas boiler; Q_{WB}^{t} is the heating power of the waste heat boiler; P_{AC}^{t} is the refrigeration power of the absorption chiller; P_{chil}^{t} is the operating power of the compression electric chiller; P_{cond}^{t} is the operating power of the air-conditioning system; and the abovementioned variable units have been converted into kW.

(2) The environmental protection goal of the integrated energy system is to obtain the lowest daily operating emissions which mainly include three aspects: nitride, carbide, and sulfide emissions.

$$f_2 = \min\left(PG_{\text{Total}}\right) = \min\left(\delta_1 \times PG_C + \delta_2 \times PG_S + \delta_3 \times PG_N\right), \tag{5}$$

where f_2 is the environmental protection goal of the system; PG_{Total} is the total daily operating emissions of pollutants from the integrated energy system; PG_{C} is the daily carbon emissions of the entire region; PG_{S} is the daily sulfide emissions of the entire region; and PG_{N} is the daily nitride emissions of the entire region and the specific formulas are as follows:

$$\begin{aligned} PG_C &= \sum_{t=1}^{24} \sum_{i=1}^{n} \left(G_i^{\text{CO}} + T_i^{\text{CO}} + Z_i^{\text{CO}}\right) \times \Delta t, \\ PG_S &= \sum_{t=1}^{24} \sum_{i=1}^{n} \left(G_i^{\text{SO}} + T_i^{\text{SO}} + Z_i^{\text{SO}}\right) \times \Delta t, \\ PG_N &= \sum_{t=1}^{24} \sum_{i=1}^{n} \left(G_i^{\text{NO}} + T_i^{\text{NO}} + Z_i^{\text{NO}}\right) \times \Delta t, \end{aligned} \tag{6}$$

where G_i^{CO}, T_i^{CO}, and Z_i^{CO} are the CO_X emissions produced from the region's purchase of electricity, microturbines, and gas boilers with the unit of kg/h; G_i^{NO}, T_i^{NO}, and Z_i^{NO} are the NO_X emissions produced from the region's purchase of electricity, microturbines, and gas boilers with the unit of kg/h; G_i^{SO}, T_i^{SO}, and Z_i^{SO} are the SO_X emissions produced from the region's purchase of electricity, microturbines, and gas boilers with the unit of kg/h; $\delta 1$, $\delta 2$, and $\delta 3$ are the weight coefficients of carbon emissions, sulfur emissions, and nitrogen emissions.

3.2. System Constraints. The joint multiobjective operation of the integrated energy system needs to meet three types of constraints: system operation constraints, environmental emission constraints, and economic constraints.

(1) Operational constraints of the system include electric power balance constraints, cooling and heating power supply constraints, equipment capacity constraints, and equipment operation constraints.

The constraints of electric power balance are as follows:

$$\sum_{i=1}^{n_{\text{CHP}}} P_i^t + \sum_{i=1}^{n_{\text{distri}}} P_{\text{distri}}^t + P_{\text{Grid}}^t + P_{\text{stor}}^t = P_{\text{Load}}^t + P_{\text{cond}}^t, \tag{7}$$

where P_{Grid}^t is the hourly power exchange value of the integrated energy system with public grid; P_{Load}^t is the hourly electrical load; P_{Stor}^t is the charging or discharging power of battery; P_{cond}^t is the hourly operation power of air conditioner; P_{distri}^t is the hourly operation power of distributed generation unit; and P_i^t is the power generated by the i^{th} microturbine.

The constraint of heating power supply is as follows:

$$\sum_{i=1}^{n_{\text{CHP}}} H_i^t + \sum_{i=1}^{n_{\text{boiler}}} H_{\text{boiler}i}^t + \text{COP}_{\text{cond}} \times P_{\text{cond}}^t + H_{\text{out}}^t - H_{\text{in}}^t \geq H_{\text{space}}^t + H_{\text{Water}}^t. \tag{8}$$

The constraint of hot water supply is as follows:

$$\sum_{i=1}^{n_{\mathrm{CHP}}} H_i^t + \sum_{i=1}^{n_{\mathrm{boiler}}} H_{\mathrm{boiler}i}^t + H_{\mathrm{out}}^t - H_{\mathrm{in}}^t \geq H_{\mathrm{Water}}^t, \tag{9}$$

where H_i^t is the retrieved heat value by the i^{th} microturbine through waste heat boiler; H_{boiler}^t is the heat output of the i^{th} gas boiler; H_{in}^t and H_{out}^t are the input and output of the thermal energy storage equipment; $\mathrm{COP}_{\mathrm{cond}}$ is the coefficient of performance of air-conditioning equipment; and H_{space}^t and H_{water}^t are the hourly space heating load and hot water load of the integrated energy system, respectively.

The constraint of cooling power supply is as follows:

$$\sum_{i=1}^{n_{\mathrm{CHP}}} C_i^t + \sum_{i=1}^{n_{\mathrm{boiler}}} C_{\mathrm{chil}}^t + \mathrm{EER}_{\mathrm{cond}} \times P_{\mathrm{cond}}^t + C_{\mathrm{out}}^t - C_{\mathrm{in}}^t \geq C_{\mathrm{space}}^t + C_{\mathrm{Refri}}^t. \tag{10}$$

The constraint of refrigeration power supply is as follows:

$$\sum_{i=1}^{n_{\mathrm{CHP}}} C_i^t + \sum_{i=1}^{n_{\mathrm{boiler}}} C_{\mathrm{chil}}^t + C_{\mathrm{out}}^t - C_{\mathrm{in}}^t \geq C_{\mathrm{Refri}}^t, \tag{11}$$

where C_i^t is the refrigeration power of the i^{th} microturbine manufactured by the absorption refrigeration unit; C_{chil}^t is the refrigeration power generated by the electric chiller; C_{in}^t and C_{out}^t are the power input and output of the cold storage equipment, respectively; $\mathrm{EER}_{\mathrm{cond}}$ is the cooling energy efficiency ratio of air-conditioning equipment; and C_{Space}^t and C_{Refri}^t are the hourly space cooling load and refrigeration load of the integrated energy system, respectively.

The equipment capacity and operating constraints of the integrated energy system are as follows:

for microturbines,

$$P_i^{\min} \leq P_i^t \leq P_i^{\max}, \quad i \in n_{\mathrm{CHP}}. \tag{12}$$

For gas boilers,

$$0 \leq H_{\mathrm{boiler}i}^t \leq H_{\mathrm{boiler}i}^{\max}, i \in n_{\mathrm{boiler}}. \tag{13}$$

For air-conditioning equipment,

$$0 \leq P_{\mathrm{cond}}^t \leq P_{\mathrm{cond}}^{\max}. \tag{14}$$

For heat (cold) energy storage equipment,

$$\begin{aligned} 0 \leq H_{\mathrm{in}}^t &\leq H_{\mathrm{in}}^{\max}, \\ 0 \leq H_{\mathrm{out}}^t &\leq H_{\mathrm{out}}^{\max}, \\ S_{\mathrm{stor}}^{\min} \leq S_{\mathrm{stor}}^t &\leq S_{\mathrm{stor}}^{\max}, \end{aligned} \tag{15}$$

where H_{in}^t and H_{out}^t are the power input and output of the heat (cold) storage equipment at time t; $H_{\mathrm{in}}^{\max}$ and $H_{\mathrm{out}}^{\max}$ are the power input and output limits of the heat (cold) storage equipment; and S_{stor}^t is the state of charge of the heat (cold) storage equipment at time t.

(2) During the operation of the integrated energy system, it is necessary to control the amount of pollutants emitted by the entire system within a certain range, including the total pollutant discharge constraints in the region and the hourly discharge constraints of various pollutants.

Constraints on total carbon emissions in the region are as follows:

$$0 \leq PG_C \leq PG_C^{\max}. \tag{16}$$

Constraints on total sulfide emissions in the region are as follows:

$$0 \leq PG_S \leq PG_S^{\max}. \tag{17}$$

Constraints on total nitride emissions in the region are as follows:

$$0 \leq PG_N \leq PG_N^{\max}, \tag{18}$$

where $PG_C^{\max}$ is the upper limit of the daily carbon emission value of the entire region; $PG_S^{\max}$ is the upper limit of the daily sulfide emission value of the entire region; and $PG_N^{\max}$ is the upper limit of the daily nitride emission value of the entire region.

The hourly carbon emission constraints in the region are as follows:

$$0 \leq \left\{ \sum_{i=1}^{n} \left(G_i^{\mathrm{CO}} + T_i^{\mathrm{CO}} + Z_i^{\mathrm{CO}} \right) \right\}^t \leq PG_C^t. \tag{19}$$

The hourly sulfide emission constraints in the region are as follows:

$$0 \leq \left\{ \sum_{i=1}^{n} \left(G_i^{SO} + T_i^{SO} + Z_i^{SO} \right) \right\}^t \leq PG_S^t. \tag{20}$$

The hourly nitride emission constraints in the region are as follows:

$$0 \leq \left\{ \sum_{i=1}^{n} \left(G_i^{NO} + T_i^{NO} + Z_i^{NO} \right) \right\}^t \leq PG_N^t, \tag{21}$$

where PG_C^t, PG_S^t, and PG_N^t are the upper limits of the carbon, sulfide, and nitride emission value of the entire region at time t.

(3) During the operation of the integrated energy system, the operating costs of the entire system needs to be controlled within a certain range. The economic constraint is as follows:

$$0 \leq K_{\mathrm{Total}} \leq K_{\max}, \tag{22}$$

where K_{Total} is the entire daily operating cost of the integrated energy system and $K_{\max}$ is the upper limit of the daily operating cost.

3.3. Solving Algorithm. Due to the existence of a variety of energy supply equipment in the optimization model and the complex coupling relationship between cold, heat, and electrical energy, the model has a large number of optimization variables and a large volume of constraint conditions, when using traditional global solution space search algorithms such as NSGA-II or multiobjective particle swarm algorithm will make the solution process unacceptably lengthy which cannot meet the time limit requirements of the rolling optimization process within the dispatching system. When the traditional multiobjective optimization algorithm which transitions from single-objective optimization algorithm such as constraint method or weighting method is used to solve the model, the solution obtained is often limited to the local Pareto optimal solution set and it is difficult to obtain a large range of Pareto optimal solution set;, therefore, the traditional multiobjective optimization algorithm cannot provide the dispatching system with a variety of dispatching schemes for the system to make decisions.

Based on the abovementioned analysis, this paper proposes a hierarchical progressive parallel NSGA-II algorithm suitable for solving multiobjective dispatching strategies of integrated energy systems. When solving the optimization model, it can not only make full use of the advantages of the global solution space search algorithm in the balance and diversity of the solution set searched but also achieve fast solution speed for large-scale nonlinear optimization problems which can meet the calculation time limit requirements of the rolling optimization process within the dispatching system. The algorithm has made three improvements based on NSGA-II as follows:

(1) Improve the nondominated layer sorting algorithm to achieve fast layering based on model characteristics.

In the process of solving the joint multiobjective optimization operation model of the integrated energy system, considering that the model only includes dual optimization goals, this paper adopts a more efficient and fast nondominated layer sorting method according to the characteristics of the model to meet the requirements of rapid model solving.

Firstly, the merge-sorting method is used to sort the individuals in the population by ascending order according to first objective function value. The merge sorting method is a sorting method by using the divide and conquer strategy. In divided stages, the method divides the problem into small questions and solves it recursively, while in the conquer stages, the method combines the answers obtained in divided stages together to get the results [28–30]. Suppose there is a population as shown in Figure 3 and each individual has two objective function values. The overall process of sorting the individuals in the

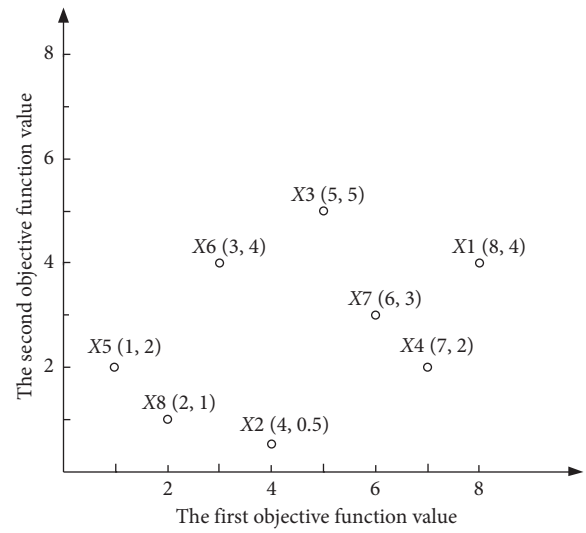


Figure 3: The individuals with two objective function values in a population.

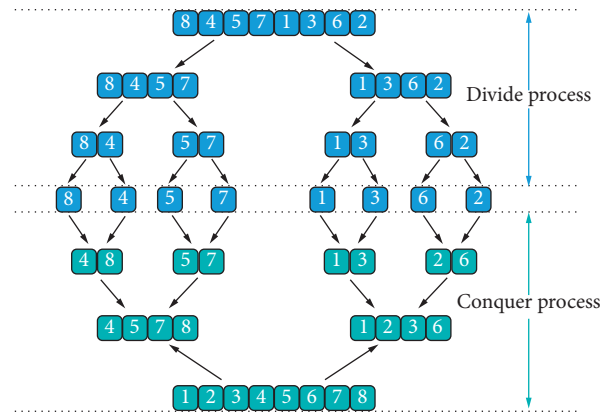


Figure 4: The overall process of the merge-sorting method.

population by ascending order according to first objective function value is shown in Figure 4. In the divided process, this structure is much like a complete binary tree and it can be understood as the process of recursively splitting subsequences. In the conquer process, we need to merge two already ordered subsequences into an ordered sequence, such as the last merge in Figure 4, to merge the already ordered subsequences of [4, 5, 7, 8] and [1–3, 6] into the final sequence [1–8]. The last merge within the conquer process demonstration is shown in Figure 5.

If the first objective function values of the two individuals are the same, the individuals are sorted in ascending order according to the second objective function value and the result of the merge sorting is stored in the array F_i. In this case, the individual with the smallest first objective function value must be an individual in the first layer of the nondominated layer. Then, the adjacent individual of the first individual is searched; if the second objective function

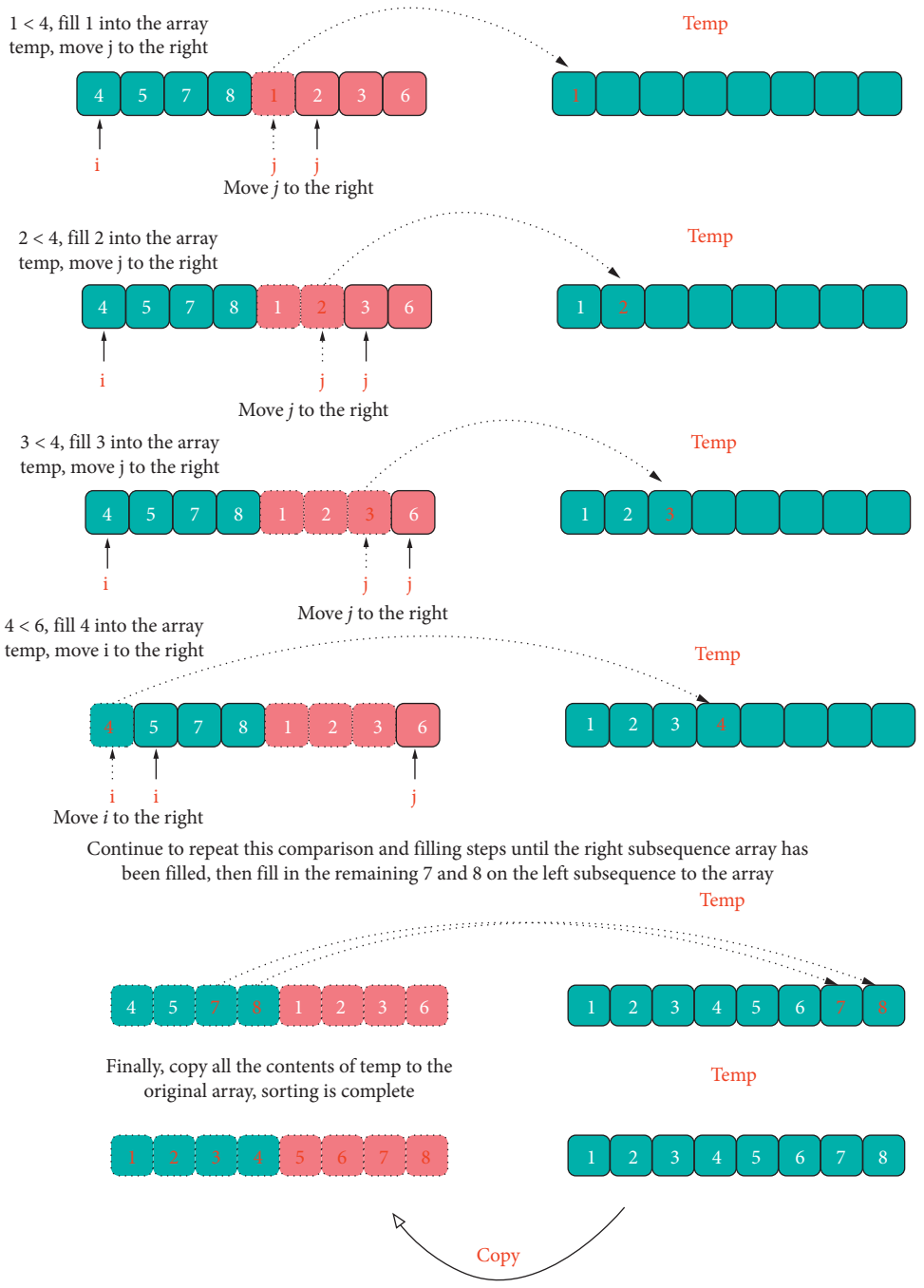


Figure 5: The last merge within the conquer process demonstration.

value of the adjacent individual is smaller than the previous individual, the individual and the previous individual are belonging to the same layer of the nondominated layer; otherwise, the individual is not belonging to the same layer and will be inserted at the end of the array F_i and so on to find all first-level nondominated individuals in the population. Next, start with the first element at the end of the array F_i to find the second layer of nondominated individuals and so on until the population is completely stratified. In the example described above, the process of stratification occurs which is shown in Figure 6.

Since the number of individuals in the next generation is half of that in F_i, an incomplete layering strategy is adopted during the layering process to interrupt layering as soon as the number of individuals in the next generation exceeds the next generation population size, which can avoid many unnecessary layering comparison steps. The schematic diagram of the nondominated layer layering algorithm based on improved NSGA-II is shown in Figure 7:

In the nondominated layer sorting algorithm, the average time complexity of the merge sorting process

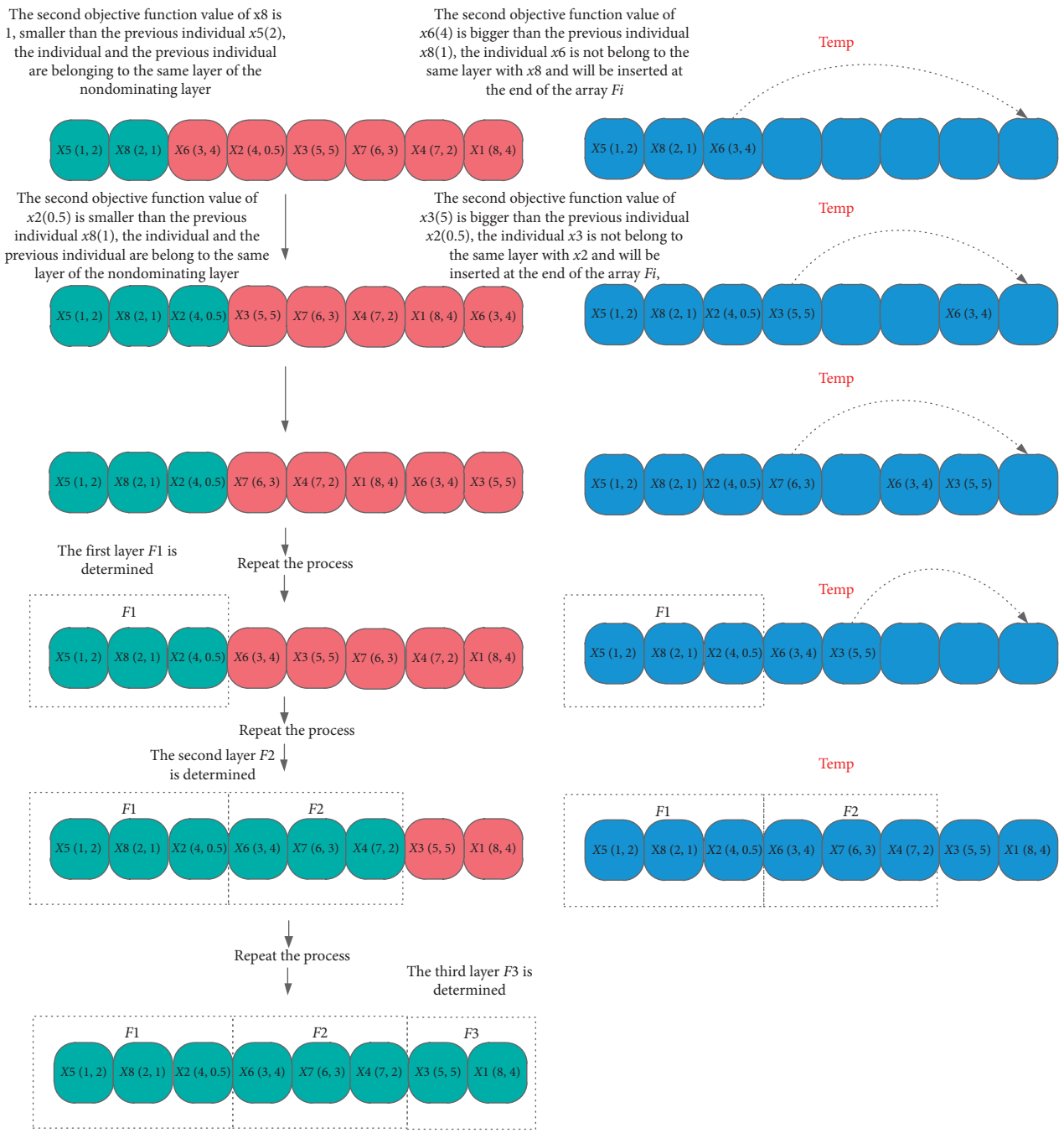


FIGURE 6: The process of stratification in the example.

is $O(N * \log_2 N)$ and the total times of calculations spent in the layering operation are

$$T_{\text{Total}} = \sum_{i=1}^{k} T_i = \sum_{i=1}^{k} \{N - i\} = N - 1 + N - 2 + \cdots + N - k < kN, \tag{23}$$

where N is twice the size of the population size; k is the total number of layers after the nondominated sorting is completed; T_{Total} is the total times of calculations spent in the layering operation after the nondominated sorting is completed; and T_i is the maximum times of comparisons required for generating the i^{th} nondominated layer.

The time complexity of the hierarchical (layering) operation is O (kN) after the nondominated sorting is completed and the total time complexity of the algorithm is O ($N * \log 2\ N + kN$). In general, $k \ll N$, so the total time complexity of the algorithm is O ($N * \log 2\ N$). The sorting principle of NSGA-II is similar to the bubble sorting which needs to compare the individuals within the population in pairs and the average time complexity of the sorting algorithm is

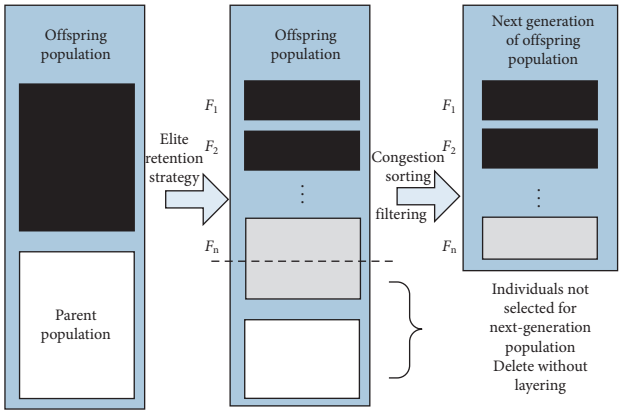


Figure 7: Schematic diagram of nondominated layer layering algorithm based on improved NSGA-II

O (mN^2), where m is the number of objectives. Compared with NSGA-II, the proposed algorithm reduces the time complexity from square order to linear logarithmic order.

(2) Modify the convergence judgment conditions and combine the target reaching method to accelerate the convergence process.

The improved multiobjective genetic algorithm based on NSGA-II can quickly approach the vicinity of the Pareto front set in the process of solving the model. However, in order to achieve the final convergence, it requires a large number of iterations and function comparisons which means longer calculation time. Therefore, a comprehensive hierarchical progressive algorithm is proposed in this paper to speedup the process of finding the Pareto optimal solution set.

Firstly, the aforementioned improved genetic algorithm based on NSGA-II is used for searching solutions close to the Pareto front by setting its termination condition to limited generations. Secondly, the target reaching method is utilized for a secondary search which uses the calculation result of the previous step as the initial value and a local search is performed by the target reaching method to obtain the Pareto solution set.

The standard form of utilizing the target reaching method to solve the model is

$$\begin{cases} \min \gamma \\ s.t. f_i(x) - w_i \times \gamma \le F_i^*, \\ c(x) \le 0, \\ c_{eq}(x) = 0, \\ Ax \le b, \\ A_{eq}x \le b_{eq}, \\ lb \le x \le ub, \end{cases} \tag{24}$$

where optimization variable x is an n-dimensional variable which includes hourly cold, heat, and power output of microturbine-based CHP system, hourly cold, heat, and power output of auxiliary energy supply equipment, hourly cold, heat, and power output of various energy storage equipment, SOC status of energy storage equipment, and hourly interacting power with the grid. $f_i(x)$ is the i^{th} optimization objective function; ω is the weight coefficient of the i^{th} objective function; F_i^* is the optimization target value expected to be achieved by the i^{th} target; γ is the target achievement coefficient which is used for representing the achievement of the predetermined goal; $c(x)$ and $c_{eq}(x)$ are the various nonlinear inequality constraints and nonlinear equality constraints met by the optimization variables; A and b are the coefficient matrices of various linear inequality constraints met by the optimization variables; A_{eq} and b_{eq} are the coefficient matrices of various linear equality constraints met by the optimization variables; lb and ub are the lower and upper limits of the optimization variable x, respectively.

In the target reaching method, the multiobjective optimization problem is transformed into a single-objective optimization problem by using virtual variable γ and the single-objective optimization model is solved by using the interior point method. In the solving process, the interior point method constructs a Lagrangian auxiliary function which meets the Karush–Kuhn–Tucker conditional:

$$L(x,\lambda) = f(x) + \sum \lambda_{c,i} c_i(x) + \sum \lambda_{A,i}(Ax - b) + \sum \lambda_{ceq,i} ceq_i(x) + \sum \lambda_{Aeq,i}\left(A_{eq}x - b_{eq}\right). \tag{25}$$

The second-order derivative Hessian matrix in its correction equation is

$$H = \nabla^2_{xx} L(x,\lambda) = \nabla^2 f(x) + \sum \lambda_i \nabla^2 c_i(x) + \sum \lambda_i \nabla^2 ceq_i(x), \tag{26}$$

where $\lambda_{c,i}\,\lambda_{A,i}\,\lambda_{ceq,i}\,\lambda_{Aeq,i}$ are Lagrange multipliers which possess the same dimensions with respective constraints. x, lb, and ub are $n \times 1$ column vectors; b is a $p \times 1$ column vector; b_{eq} is a $q \times 1$ column vector; A is a $p \times n$ matrix; A_{eq} is a $q \times n$ matrix; Hessian matrix is an $n \times n$ matrix.

Solutions can be achieved by transformation, LDL^T decomposition, iteration, and correction of the matrices in the correction equations. The iteration is automatically stopped when the convergence conditions are met and the solution closest to the expected optimization target in the feasible region is found as a Pareto optimal solution.

(3) Introduce parallel computing technology to accelerate the solution process according to the characteristics of the algorithm.

Because there are a large number of repetitive loop iterative processes in the multiobjective hierarchical progressive algorithm based on improved NSGA-II, in addition, each process has no parameter passing to each other and does not depend on the results of other process, parallel computing technology is introduced to accelerate the convergence process of the algorithm. Among them, the part applicable for parallel computing mainly includes the nondominated layer merge sorting calculation and the call of the target reaching method for achieving fast convergence. In the nondominated layer merge sorting section, the fitness of individuals is first calculated on multiple parallel processing units, and then the individuals of the population to be sorted are evenly distributed to m parallel processing units. The complexity of the merge sorting process is reduced to O $(N/m * \log_2 N)$ under parallel processing and the algorithm speeds up with the increase of the parallel processing units. When the target reaching method is utilized for a secondary search, the method starts from different initial values and there is no parameter passing to each other, so it is also suitable for using parallel processing technology to evenly distribute the computing task to multiple parallel processing units of the computer.

In addition, because the GPU which is designed for parallel operations has more computing units than the CPU, in order to make full use of the GPU in large-scale array matrix processing, multiple operations without complex logical judgments and repeated calculations with little data correlation, the population crossover and mutation operations in the algorithm are assigned to the GPU. As the GPU is limited by the hardware architecture, single-precision calculations are much faster than double-precision calculations; therefore, this paper uses single-precision (single type variable) calculations for various types of population operations.

Through the improvement of the abovementioned three aspects, the hierarchical progressive parallel NSGA-II algorithm can greatly reduce the calculation time in the process of obtaining the Pareto solution set, which can meet the requirements of the rolling optimization process within the dispatching system while ensuring the uniform distribution of the solution and the diversity of the population. The solution set obtained by the algorithm can provide flexible and diverse dispatching schemes for decision makers to choose.

The hierarchical progressive parallel NSGA-II algorithm is written in MATLAB to solve the joint multiobjective optimization dispatching model of the integrated energy system in this paper. The algorithm flowchart is shown in Figure 8.

4. Case Scenario

4.1. Case Situation. Figure 9 shows the daily forecasting cooling, heating, and power load curve of a typical integrated energy system in the Tianjin Sino-Singapore Eco-city smart grid zone. The region adopts time-of-use electricity pricing with peak time from 8 to 11 and 18 to 23, flat time from 7 to 8 and 11 to 18, and valley time from 23 to 7. The time-of-use electricity price is shown in Table 1.

Assuming the natural gas price is 0.4642 \$/cubic meter and the high calorific value is 8571 kcal/cubic meter, thus the gas price is converted to 0.0464 \$/kW·h. The weight coefficients $\delta 1$, $\delta 2$, and $\delta 3$ for carbon, sulfur, and nitrogen emissions are taken as 1,100,100, respectively.

The unit of emissions is converted to g/kWh according to [31] and the pollutant emission data of each equipment in the integrated energy system is shown in Table 2.

4.2. Equipment Parameters in the Case. The main equipment in the integrated energy system contains microturbines, waste heat boilers, absorption refrigerators, electric refrigerators, gas boilers, batteries, heat (cold) storage devices, household air conditioners, and photovoltaic cells. The integrated energy system exchanges power through a centralized power bus with the public grid, purchases power from the grid when the power supply within the system is insufficient, and sells power to the grid when the power supply within the system is surplus. At the same time, it is assumed that there is no gas production in the integrated energy system and all the gas required is supplied by the external gas company. The parameters of the energy supply equipment and energy storage equipment in the case are shown in Tables 3 and 4.

5. Case Study

5.1. Comparisons of Pareto Boundaries for System Optimal Operation under Different Algorithms. Three different algorithms were used to solve the joint multiobjective optimal operation and dispatching model of the integrated energy system, and the Pareto optimal solution sets under different algorithms were obtained.

When using the weighting method, an initial solution is randomly generated within the feasible solution range as the initial iterative value. The weights of the two targets are normalized and 49 sets of weights are taken for iteration. The Pareto optimal solution set boundary obtained is shown in Figure 10. As shown in the figure, the Pareto optimal solution set boundary obtained by the weighting method is relatively concentrated in some areas and it is not very continuous on the entire solution set front; in other words, it has disadvantages in the diversity of the solutions searched.

When the NSGA-II algorithm is used, the initial population size is set to 2000 and single-precision real-number coding is used. Each bit of the individual is the electric, heating, and cooling power output values by various types of energy supply equipment. The crossover operator uses two-point crossover and adjacent individuals cross. The

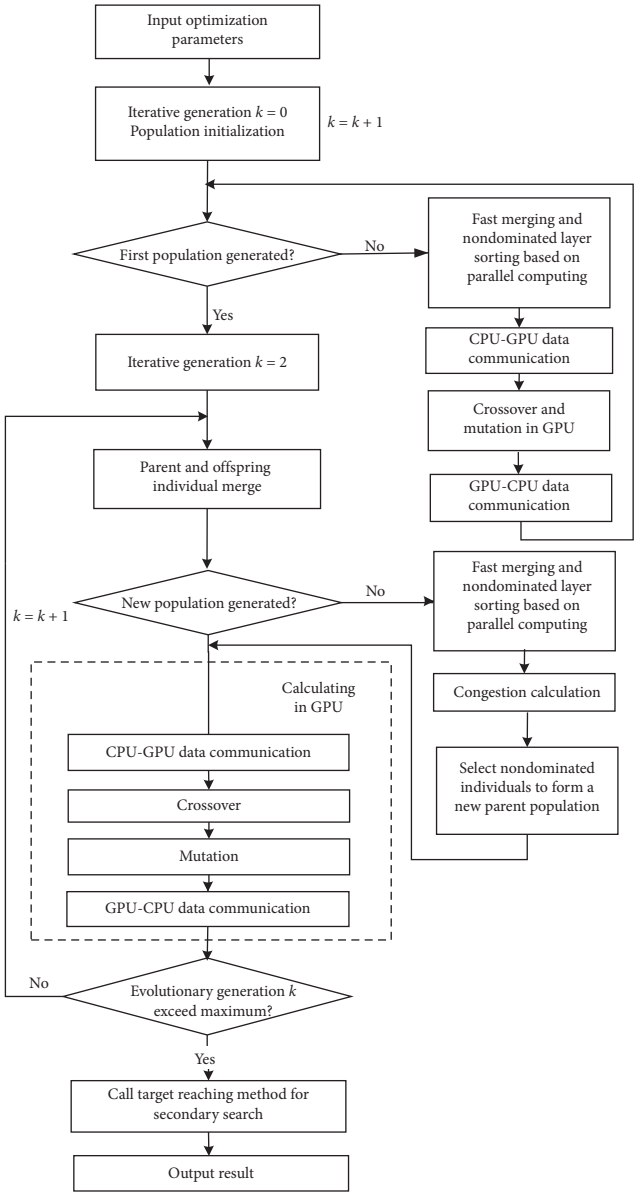


Figure 8: Flowchart of hierarchical progressive parallel NSGA-II algorithm.

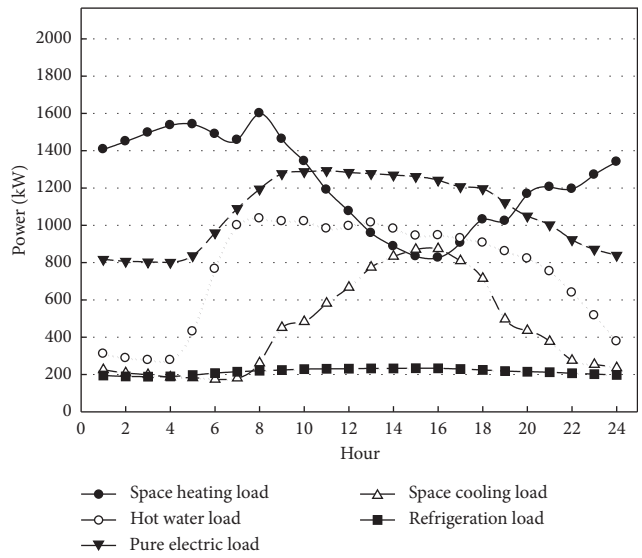


Figure 9: Curve of the forecasting cooling, heating, and power load of a typical integrated energy system in one day.

Table 1: Time-of-use price.

Period	Electricity price ($/kWh)
Peak	0.1159
Valley	0.0491
Flat	0.0816

crossover rate is 0.8 and the mutation probability is 0.04. The maximum iteration generation is set to 200 and the relative change criterion between the two generations of individuals in the Pareto optimal solution set is set to 1e-1. If the difference between the two generations is less than the criterion, the iteration is terminated. The Pareto optimal solution set boundary obtained is shown in Figure 11.

As shown in the figure, the Pareto optimal solution set boundary obtained by the NSGA-II algorithm is relatively scattered and continuous on the entire front of the solution set, which has a prominent advantage in the diversity of the solutions searched, that is, the NSGA-II algorithm can find as many Pareto optimal solutions as possible. However, the disadvantage of NSGA-II algorithm is that the calculation time is very long and the optimization process can take several hours at a time, which is unacceptable in practical engineering applications.

When the multiobjective hierarchical progressive parallel algorithm based on improved NSGA-II is used, the initial population size is set to 2000 and single-precision real-number coding is used. Each bit of the individual is the electric, heating, and cooling power output values by various types of energy supply equipment. The crossover operator uses two-point crossover and adjacent individuals cross. The crossover rate is 0.8, the mutation probability is 0.04, and the maximum iteration generation is set to 200. If the maximum iteration generation is exceeded, the target reaching method is utilized for a secondary search. The relative change criterion between the two generations of individuals in the Pareto optimal solution set is set to 1e-1. If the difference between the two generations is less than the criterion, the iteration is terminated. The test computer is equipped with Intel's third-generation Core processor i5-3230M, which can perform dual-core four-thread parallel computing. The main frequency of each core of the CPU is 2.60 GHz; the configured GPU is NVIDIA GeForce 610M. The programming environment is MATLAB R2012a and the result is shown in Figure 12.

As shown in the figure, the Pareto optimal solution set boundary obtained by the multiobjective hierarchical progressive parallel algorithm based on improved NSGA-II is also relatively scattered, and it is passable continuous on the entire solution set front. Because the target reaching method does not pay attention to the diversity of solutions, the diversity of solutions will be sacrificed in the secondary

Table 2: Pollutant indicators of major gas and fuel equipment in the integrated energy system.

Pollutant	Microturbine	Internal combustion engine	Fuel cell	Gas boiler	Public grid	Absorption refrigeration unit
CO_X	724	543	486	254	922	171
NO_X	0.2	0.15	0.0136	0.54	2.295	0.06
SO_X	0.0036	0.0032	0.3327	0.764	3.583	0.0008

Note. The pollutant indicators of microturbine, internal combustion engine, fuel cell, and public grid are based on the production of 1 kWh electricity; the pollutant indicators of gas boilers are based on the production of 1 kWh heat; the pollutant indicators of the absorption refrigeration unit are based on the production of 1 kWh cooling energy.

Table 3: The parameters of energy supply equipment.

Equipment	Parameters	Values
Capstone C1000 microturbine system	Maximum power generated $P_{c1000,\max}$	1000 kW
	Rated efficiency η_{c1000}	0.33
Waste heat boiler	Maximum power input $P_{\mathrm{EB},\max}$	2000 kW
	Rated efficiency η_{EB}	0.8
Lithium bromide absorption refrigeration unit	Maximum power input $P_{\mathrm{AC},\max}$	2000 kW
	Coefficient of performance $\mathrm{COP}_{\mathrm{AC}}$	1.2
Gas boiler	Maximum power input $P_{\mathrm{GB},\max}$	1000 kW
	Rated efficiency η_{GB}	0.9
Electric refrigerator	Maximum power input $P_{\mathrm{chil},\max}$	500 kW
	Coefficient of performance $\mathrm{COP}_{\mathrm{chil}}$	4.0
Household air conditioner	Maximum power input $P_{\mathrm{cond},\max}$	1000 kW
	Energy efficiency ratio $\mathrm{EER}_{\mathrm{cond}}$	2.6
	Coefficient of performance $\mathrm{COP}_{\mathrm{cond}}$	3.1
Public grid	Maximum power exchange $P_{\mathrm{Bus},\max}$	1500 kW
Photovoltaic cells	Maximum power generated $P_{\mathrm{DG},\max}$	187.8 kW

Table 4: The parameters of energy storage equipment.

Parameters	Storage		
	Lead-acid batteries	Heat storage tank	Cold water storage tank
Charging/(heat, cold) efficiency	0.97	0.95	0.95
Discharge/(heat, cold) efficiency	0.97	0.95	0.95
Maximum charge/(heat, cold) rate	0.2	0.2	0.2
Maximum discharge/(heat, cold) rate	0.3	0.2	0.2
Self-discharge/(heat, cold) rate	0.02	0.03	0.03
Maximum state of charge/energy	0.9	0.9	0.9
Minimum state of charge/energy	0.2	0.1	0.1
Capacity	200 kW·h	1000 kW·h	500 kW·h

The calculation results of the case are analyzed from three aspects.

search process in exchange for the convergence time of the calculation. The advantages of the overall algorithm in the diversity of solutions are not as prominent as the NSGA-II algorithm. However, its advantage lies in the combination of the rapid convergence of the target reaching method and the advantages of the NSGA-II algorithm in searching for the diversity of solutions, which makes the algorithm to find as many and more dispersed Pareto optimal solutions as possible in a short time. The optimization process can be reduced to several minutes on the micro-PC, which greatly improves the engineering practicability of the algorithm.

5.2. Performance Comparison of Different Algorithms. NSGA-II and the multiobjective hierarchical progressive parallel algorithm based on improved NSGA-II proposed in this paper are used to solve the optimization operation model of the integrated energy system. The average individual comparison times in a single layering process within one solution process, nondominated frontier average number of layers, and the average CPU time consumed by a single layering process of the two algorithms are recorded. The parameter settings of the test are the same as the parameter conditions for obtaining the Pareto boundary using NSGA-II. During the test, the experiment was repeated 20 times for each algorithm and the average value of each test data was taken. The results of the time-consuming comparison tests between NSGA-II and the proposed algorithm on a single nondominated layer construction are shown in Table 5.

From the table, the algorithm proposed in this paper has very obvious advantages in terms of the average individual comparison times in a single evolutionary layering process,

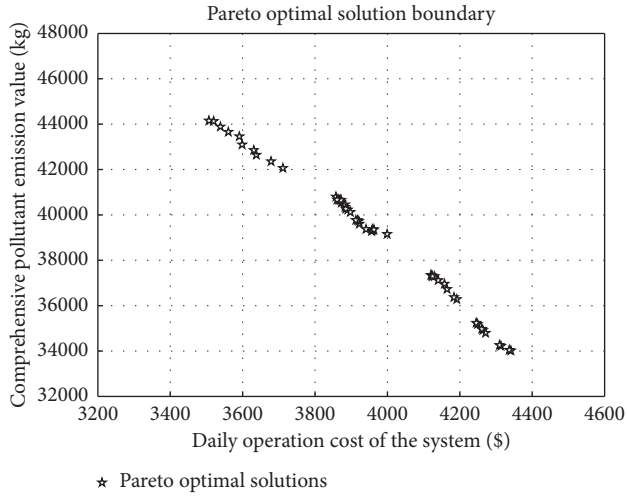


Figure 10: Pareto optimal solution boundary obtained by the weighting method.

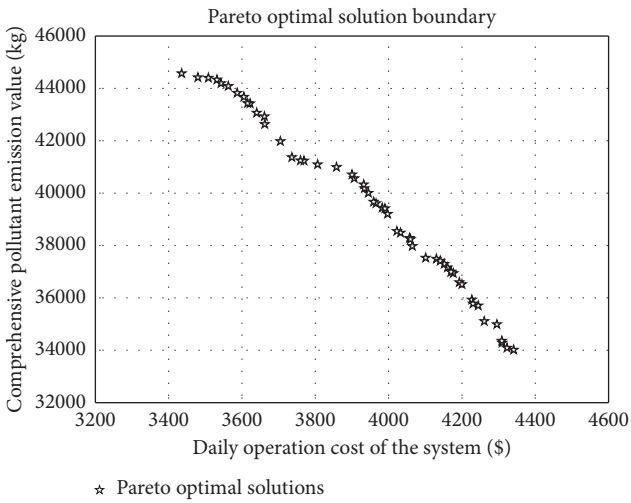


Figure 11: Pareto optimal solution boundary obtained by NSGA-II.

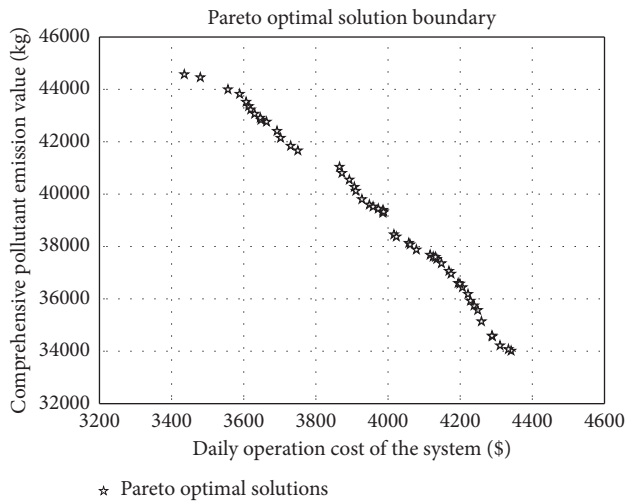


Figure 12: Pareto optimal solution boundary obtained by hierarchical progressive parallel NSGA-II.

the average number of nondominated frontier layering layers, and the average CPU time consumed by a single layering process. The decrease in the average individual comparison times is mainly due to the use of a nondominated frontier construction method based on dual-objective merge sorting. The decrease in average number of layers of the nondominated frontier is mainly due to the practice of stopping the stratification when the individual is full. The significant reduction in the average CPU time consumed by the layering process is mainly due to the reduction in the number of comparisons and the introduction of CPU dual-core parallel computing.

The changes in the number of stratifications in the nondominated layer of the subparent comprehensive population during a typical evolutionary process are as shown in Figure 13.

As shown in Figure 13, the multiobjective hierarchical progressive parallel algorithm based on improved NSGA-II greatly reduces the number of layers in a single evolution process and significantly improved computing efficiency in the process of performing hierarchical selection operations compared with NSGA-II.

Although several improvements have been made for the NSGA-II algorithm, due to the large number of evolutionary generations, the large number of comparisons, and the large amount of calculations, the total calculation time is still quite long and the algorithm execution efficiency is lower than the weighting method. Therefore, this paper applies the comprehensive hierarchical progressive algorithm which modifies the convergence condition judgment to solve the model. Taking advantage of the strong scalability of NSGA-II, by setting a certain truncated evolutionary generation and combining with the target reaching method to perform the final step of convergence calculation, the convergence progress of the entire algorithm is accelerated. The specific test calculation time is shown in Table 6.

As shown in the table, when the truncated evolutionary generation is set to 100, the total calculation time of the entire algorithm can be controlled within 10 minutes, which is much faster than NSGA-II (up to several hours, not listed because the magnitudes are not of average significance). Due to the need for specifying the weight and initial position, the fastest traditional algorithm has the characteristics of fast single optimization process but poor overall convergence. It may take many trials to find a satisfactory Pareto front; overall, it is not as good as the algorithm proposed in this paper. The hierarchical progressive algorithm based on the improved convergence judgment accelerates the process of convergence by combining NSGA-II with the target reaching method, but at the expense of a small portion of the Pareto solution set diversity, which is not only reflected in the number of searched solutions but also in the reduction of the dispersion degree of searched solutions. The diversity is slightly weaker than NSGA-II, but since the initial value used by the target reaching method in the convergence calculation process was obtained from the improved NSGA-II algorithm after iteration for several generations and therefore these initial values still have relatively good dispersion, hence the solution diversity of the proposed algorithm is

Table 5: Comparison of test results between NSGA-II and the proposed algorithm in every construction of a nondominated frontier set.

Number of tests	NSGA-II			The multiobjective hierarchical progressive parallel algorithm		
	The average individual comparison times	Nondominated frontier average number of layers	Average CPU time consumed by a single layering process(s)	The average individual comparison times	Nondominated frontier average number of layers	Average CPU time consumed by a single layering process(s)
1	4861898.4	58.5	93.3	62521.9	29.2	1.126
2	4665847.7	50.6	89.6	73679.1	25.3	1.3
3	4585557.9	47.4	88	61776.7	23.7	1.348
4	4448335.5	41.9	85.4	77458.3	41	1.943
5	5255582.6	74.2	100.9	75764.1	27.1	1.379
6	4999614.9	64	96	70900.7	28	1.694
7	4983007.9	63.3	95.7	69879.4	41.7	1.655
8	5375101.4	79	103.2	74943.3	29.5	1.848
9	5143954.2	69.8	98.8	76288.2	24.9	1.899
10	5286829.5	75.5	101.5	65054.2	27.7	1.472
11	5387153.8	79.5	103.4	45385.8	29.7	0.725
12	4806556.6	56.3	92.3	65837.3	28.1	1.502
13	4718012.1	52.7	90.6	59488.3	26.4	1.261
14	4966130.4	62.6	95.4	48214.9	31.3	0.832
15	4987274.1	63.5	95.8	48977.8	31.7	0.711
16	4883584.6	59.3	93.8	44668.7	29.7	0.697
17	5054736	66.2	97.1	40401.2	23.1	0.535
18	4854897.6	58.2	93.2	65453	29.1	1.487
19	4643834.1	49.8	89.2	70914.4	24.9	1.695
20	4998534.4	63.9	96	50881.1	22	0.933
Average value	4945322.2	61.8	95	62424.4	28.7	1.302

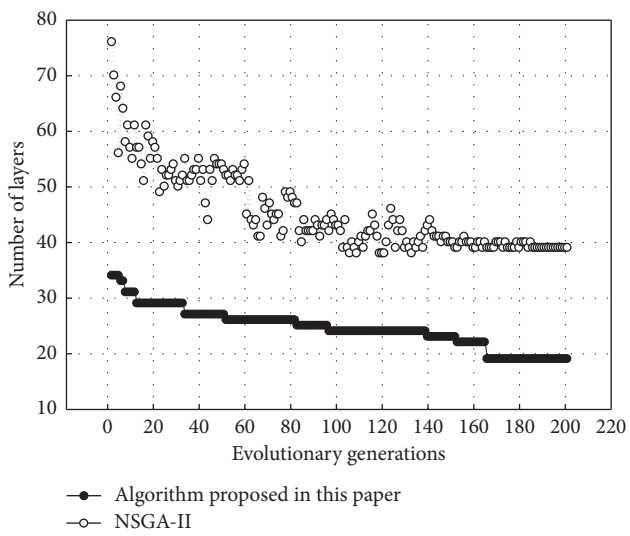


Figure 13: Changes in the number of stratifications in the nondominated layer of the subparent comprehensive population in the evolutionary process.

much better than that of traditional weighting methods. In general, the algorithm proposed in this paper has better comprehensive computing capabilities and advantages in terms of computing time and the diversity of solutions.

In the multiobjective hierarchical progressive algorithm based on improved NSGA-II, parallel computing acceleration mainly includes two parts: CPU dual-core parallel acceleration and GPU + CPU cooperative processing acceleration. The CPU dual-core parallel acceleration part mainly includes the nondominated layer merge sorting calculation and the final step of calling the target reaching method to achieve rapid convergence. According to the description of introducing the parallel calculation to accelerate the solution process, it can be seen that the ideal speedup ratio is 2 considering the calculation amount is evenly divided into the two cores of the CPU. However, because the calculation amount that needs to be decomposed is not necessarily an even number and data splitting and integration, the speedup ratio is lower than 2 in actual tests. The specific average test results are shown in Table 7.

Since the GPU has more arithmetic units (multiplication and addition units of integer and floating point number, special arithmetic units, etc.) than the CPU and its design is more oriented to parallel operations, the GPU has a large advantage over the CPU in large-scale array and matrix processing, multiple simple calculations, and repeated operations with little data correlation in theory. At the specific execution level, GPU + CPU cooperative processing acceleration has advantages in terms of population crossover and mutation operation time in a single iteration. However, the elapsed time of GPU-CPU data transmission and communication needs to be added to the total calculation time; thus, in the testing system, the acceleration ratio can only reach an average of about 1.08 in the cooperative processing acceleration process.

5.3. Dispatching Strategy of Typical Optimal Solution. In the Pareto optimal solution set, the sum of the pollutant emission values of Pareto solution with the best economic

Table 6: Time-consuming comparison in whole solution period of algorithms under different convergence conditions.

Number of tests	Traditional multiobjective weighting algorithm		Multiobjective hierarchical progressive parallel algorithm based on improved NSGA-II				
	Average searching time for a single solution(s)	Total calculation time(s)	Average time taken per iteration(s)	Truncated evolutionary generation, 100		Truncated evolutionary generation, 200	
				Number of Pareto solutions found	Total calculation time (s)	Number of Pareto solutions found	Total calculation time(s)
1	6.2	306.1	2.8	43	434.6	45	716.6
2	6.2	304	2.9	43	445.9	46	740.1
3	6	293.2	3.2	42	467.4	49	788.9
4	5.3	261.3	2.9	41	416.1	46	702.5
5	5.5	268.9	2.9	41	422.3	46	710.6
6	6.5	317.7	3.1	44	470.7	48	782.8
7	6.8	334.5	2.9	45	460.6	46	754.6
8	6.8	331.3	3	44	469.2	47	772.9
9	5.7	279.6	2.9	42	425.5	46	711.2
10	6.4	311.5	3.2	43	474.7	49	794.5
11	7	342.6	3.2	45	486.9	49	802.6
12	5.5	270.7	3.2	41	454.2	49	773.3
13	6	294.7	3.3	43	475.7	53	804.1
14	6.1	299.7	3.2	43	468.9	49	788.4
15	5.9	291.2	2.9	42	433.4	46	721.4
16	5.7	277.6	3.1	42	449.6	48	761.2
17	6.8	333.1	3.1	44	475.4	48	784.3
18	6.8	331.5	3.2	44	484.4	49	803.3
19	6.7	328.7	3	44	461.6	47	759
20	6.9	336.8	3.2	45	483.6	49	799.3
Average	6.2	305.7	3.1	43.1	458	47.8	763.6

Table 7: Algorithmic acceleration test results using parallel computing technology.

Parallel acceleration position in the algorithm		Ideal speedup ratio	Average test results
CPU dual-core parallel acceleration	Nondominated layer merge sorting section	2	1.834
	Calling target reaching method to achieve rapid convergence section	2	1.668
GPU + CPU cooperative processing acceleration	Population crossover and mutation operations	—	1.082

index is 44571.567 kg/day and the operating cost is 3433.16 USD/day. The sum of the pollutant emission values of Pareto solution with the best environmental protection index is 34015.972 kg/day and the operating cost is 4344.79 USD/day. Select a typical Pareto solution that takes into account both economic and environmental protection from the Pareto solutions set; the sum of the pollutant emission values is 42771.233 kg/day and the operating cost is 3667.74 USD/day. The dispatching curve of various load balance and equipment operating status are listed below; the day-ahead optimized dispatching curve of electric load balance is shown in Figure 14.

In Figure 14, the system purchases electricity from the public grid to meet the demand for electricity loads during the time when the electricity price is low from 23 : 00 to 8 : 00 and the purchase amount is less than the case of optimal economy index and greater than the case of optimal environmental protection. When the peak and flat electricity price periods come, the microturbine enters the rated operating state to meet the demand of the electric load and the total power generation amount is between the economically optimal situation and the environmentally friendly optimal situation.

The day-ahead optimized dispatching curve of space heat load balance with both economic and environmental advantages is shown in Figure 15.

In Figure 15, the space heat load is jointly supplied by the air-conditioning system, gas boiler, and microgas turbine system. At the time of valley electricity prices, the space heat load is mainly provided by the air conditioner and during the peak electricity price periods, the waste heat boiler in the microgas turbine system and gas boiler are applied for heating supply.

The day-ahead optimized dispatching curve of hot water load balance with both economic and environmental advantages is shown in Figure 16.

In Figure 16, the hot water load of the integrated energy system in the smart grid zone is met by waste heat boiler and

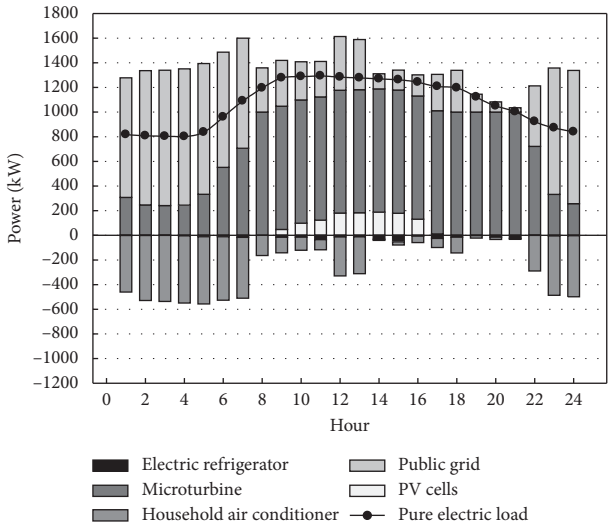


Figure 14: Optimized dispatching curve of electric load balance with both economic and environmental advantages.

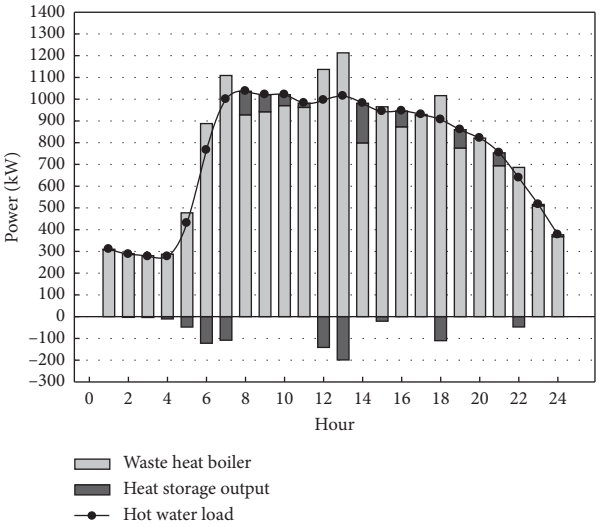


Figure 16: Optimized dispatching curve of hot water load balance with both economic and environmental advantages.

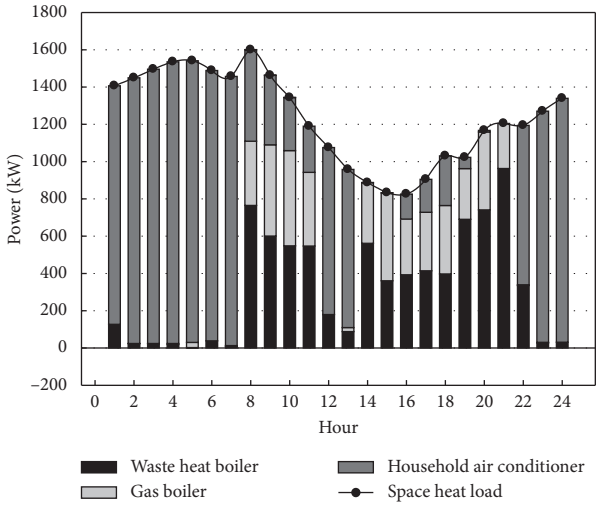


Figure 15: Optimized dispatching curve of space heat load balance with both economic and environmental advantages.

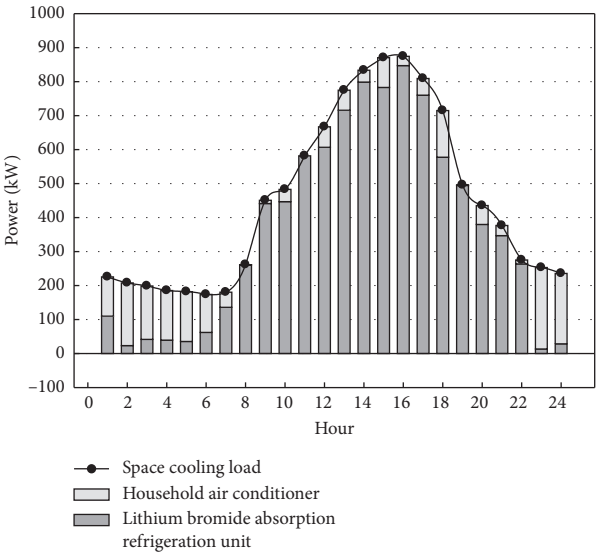


Figure 17: Optimized dispatching curve of space cooling load balance with both economic and environmental advantages.

heat storage equipment. The supply is similar to the dispatching strategy with optimal environmental protection indicators and the difference is reflected in the scheduling of heat energy storage.

The day-ahead optimized dispatching curve of space cooling load balance with both economic and environmental advantages is shown in Figure 17.

In Figure 17, the space cooling load is jointly supplied by the air-conditioning system and lithium bromide absorption refrigeration unit within the microturbine system. At the time of valley electricity prices, the space cooling load is mainly provided by the air conditioner and the remaining time is mainly provided by the lithium bromide absorption refrigeration unit. When the cooling capacity of the lithium bromide absorption refrigerating unit is insufficient, the household air-conditioning system is used as the peak shaving device of the space cooling load.

The day-ahead optimized dispatching curve of refrigeration load balance with both economic and environmental advantages is shown in Figure 18.

As shown in Figure 18, the refrigeration load of the integrated energy system is jointly met by the lithium bromide absorption refrigeration unit, electric refrigerator, and cold storage equipment, and the output of the three types of energy supply equipment is relatively uniform.

The day-ahead optimized dispatching curve of energy storage equipment operating status with both economic and environmental advantages is shown in Figure 19.

The operating status of the energy storage equipment with both economic and environmental advantages is less than the average state of energy under the dispatching

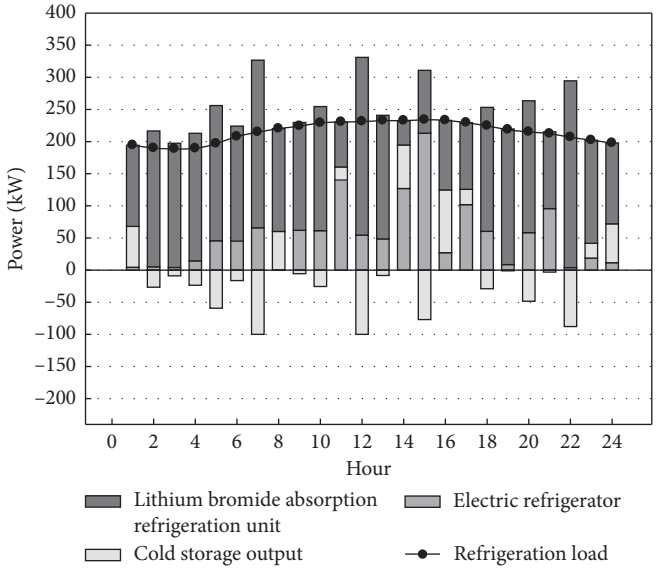


Figure 18: Optimized dispatching curve of refrigeration load balance with both economic and environmental advantages.

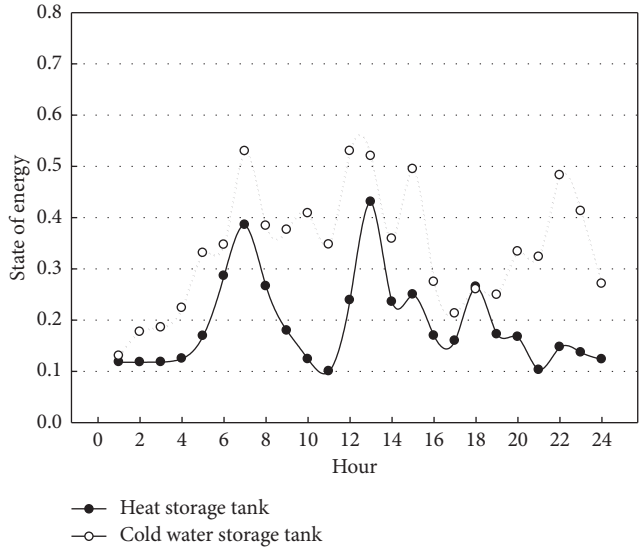


Figure 19: Optimized dispatching curve of energy storage equipment operating status with both economic and environmental advantages.

strategies with optimal economic indicators and environmental protection indicators; this is because it is not necessary to reach the limits of economic or environmental protection goals and energy storage equipment can operate with some margin when both goals are considered.

The pollutant emissions from various types of energy supply equipment and public grid in the smart grid zone with both economic and environmental advantages are shown in Figure 20.

As shown in Figure 20, pollutant emissions mainly come from energy production emissions of public grids, followed by emissions from microturbine, CHP auxiliary systems, and gas boiler in the optimized dispatching strategy with both economic and environmental advantages. The strategy utilizes cold and heat energy storage equipment to store the excess energy generated by the microturbine and release it at the required time. When the pollutants emissions are not unacceptably high, the public grid electricity is used for energy supply during nonpeak pricing hours. The emission of pollutants in the scheme is between the best economic situation and the best environmental protection situation. It does not cause a large degree of pollutant emissions and controls the operating cost of the integrated energy system within a certain range, which is rather a balanced dispatching strategy. In practical engineering applications, the optimal operation scheme can be selected according to the needs of customers such as economic-oriented or environmental protection-oriented, thus achieving the multiobjective optimized operation of the integrated energy system in the smart grid zone.

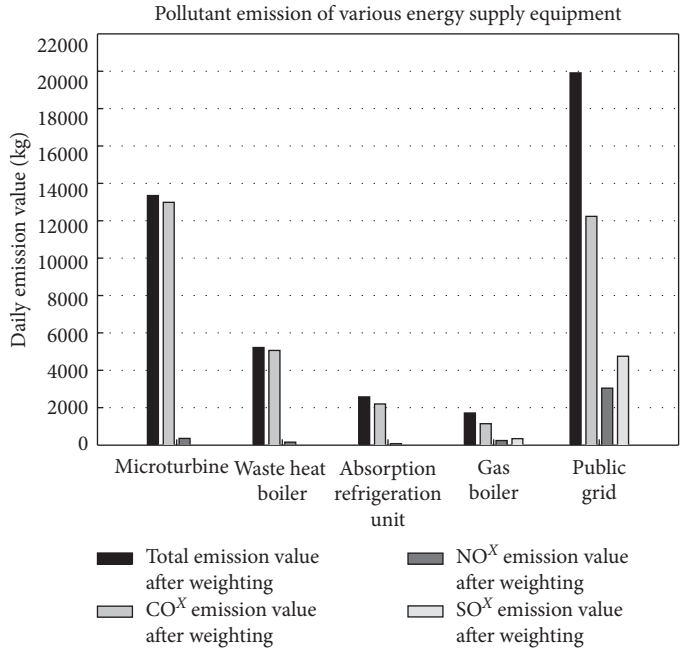


Figure 20: Pollutant emissions from various types of energy supply equipment in the optimized dispatching strategy with both economic and environmental advantages.

6. Conclusion

This paper analyzes the typical energy supply structure of the integrated energy system and the joint multiobjective optimization dispatching architecture, researches on the optimization dispatching model of the integrated energy system under multiple objectives, and proposes multiobjective hierarchical progressive parallel algorithm based on improved NSGA-II according to the characteristics of the model.

The algorithm improves the nondominated layer sorting algorithm, changes the convergence judgment condition while introducing the target reaching method to accelerate convergence, and introduces parallel computing technology according to the characteristics of the algorithm. The algorithm accelerates the convergence speed of the solution process while taking into account the diversity of the solution searched.

By analyzing the application scenarios of the optimized operation of the integrated energy system under multiobjectives, the typical daily operation optimization model of integrated energy systems in the smart grid zone is solved, and system's joint multiobjective optimized operation Pareto boundary is obtained. Finally, the algorithm proposed in this paper is compared with traditional weighting algorithm and NSGA-II algorithm in terms of solution diversity and solution time. The case study shows that the algorithm proposed in this paper not only has advantages on the diversity of solutions searched but also can achieve better results in many aspects such as the iteration time and algorithm convergence which are required in practical engineering projects. In practical engineering applications, the optimal operation scheme can be selected according to the needs of customers to achieve multiobjective optimized operation of the integrated energy system in the smart grid zone.

Data Availability

The data used to support the findings of this work are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Aidong Zeng and Sipeng Hao conceived the research scheme and wrote the manuscript. Jia Ning verified the feasibility of the method and implemented the algorithm. Qingshan Xu checked the data processing results and manuscript. Ling Jiang offered the data in the case study.

Acknowledgments

This work was supported by the Science and Technology Project of State Grid Corporation of China (no. SGTJDK00DWJS1600014), University Natural Science Research General Project of Jiangsu Province (no. 18KJB470013), University-level Research Foundation of Nanjing Institute of Technology (no. YKJ201715), Open Research Fund of Jiangsu Collaborative Innovation Center for Smart Distribution Network, Nanjing Institute of Technology (no. XTCX201802), and University-level Research Foundation of Nanjing Institute of Technology (no. JCYJ201815).

References

[1] Y. Wang, Y. Lu, L. Ju et al., "A multi-objective scheduling optimization model for hybrid energy system connected with wind-photovoltaic-conventional gas turbines, CHP considering heating storage mechanism," *Energies*, vol. 12, no. 3, 2019.

[2] A. Abdolahi, J. Salehi, F. Samadi Gazijahani et al., "Probabilistic multi-objective arbitrage of dispersed energy storage systems for optimal congestion management of active distribution networks including solar/wind/CHP hybrid energy system," *Journal of Renewable and Sustainable Energy*, vol. 10, no. 4, 2018.

[3] S. Seijo, I. Del Campo, J. Echanobe, and J. García-Sedano, "Modeling and multi-objective optimization of a complex CHP process," *Applied Energy*, vol. 161, pp. 309–319, 2016.

[4] M. Stadler, A. Siddiqui, C. Marnay, H. Aki, and J. Lai, "Control of greenhouse gas emissions by optimal DER technology investment and energy management in zero-net-energy buildings," *European Transactions on Electrical Power*, vol. 21, no. 2, pp. 1291–1309, 2011.

[5] W. Gu, Z. Wu, R. Bo et al., "Modeling, planning and optimal energy management of combined cooling, heating and power microgrid: a review," *International Journal of Electrical Power & Energy Systems*, vol. 54, pp. 26–37, 2014.

[6] A. Waqar, M. S. Tanveer, J. Ahmad et al., "Multi-objective analysis of a CHP plant integrated microgrid in Pakistan," *Energies*, vol. 10, no. 10, 2017.

[7] M. Sharafi and T. Y. ELMekkawy, "Multi-objective optimal design of hybrid renewable energy systems using PSO-simulation based approach," *Renewable Energy*, vol. 68, pp. 67–79, 2014.

[8] P. Ahmadi, A. Almasi, M. Shahriyari, and I. Dincer, "Multi-objective optimization of a combined heat and power (CHP) system for heating purpose in a paper mill using evolutionary algorithm," *International Journal of Energy Research*, vol. 36, no. 1, pp. 46–63, 2012.

[9] R. Hemmati, "Optimal cogeneration and scheduling of hybrid hydro-thermal-wind-solar system incorporating energy storage systems," *Journal of Renewable and Sustainable Energy*, vol. 10, no. 1, p. 014102, 2018.

[10] X. Ran, R. Zhou, Y. Yang et al., "The multi-objective optimization dispatch of combined cold heat and power based on the principle of equal emission," in *Proceedings of the 2012 IEEE Power and Energy Society General Meeting*, IEEE, San Diego, CA, USA, pp. 1–5, July 2012.

[11] K. Yang, Y. Ding, N. Zhu, F. Yang, and Q. Wang, "Multi-criteria integrated evaluation of distributed energy system for community energy planning based on improved grey incidence approach: a case study in Tianjin," *Applied Energy*, vol. 229, pp. 352–363, 2018.

[12] F. Fang, Q. H. Wang, and Y. Shi, "A novel optimal operational strategy for the CCHP system based on two operating modes," *IEEE Transactions on Power Systems*, vol. 27, no. 2, pp. 1032–1041, 2011.

[13] P. Pourghasem, F. Sohrabi, M. Abapour, and B. Mohammadi-Ivatloo, "Stochastic multi-objective dynamic dispatch of renewable and CHP-based islanded microgrids," *Electric Power Systems Research*, vol. 173, pp. 193–201, 2019.

[14] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[15] S. Jeyadevi, S. Baskar, C. K. Babulal, and M. Willjuice Iruthayarajan, "Solving multiobjective optimal reactive power dispatch using modified NSGA-II," *International Journal of Electrical Power & Energy Systems*, vol. 33, no. 2, pp. 219–228, 2011.

[16] X. Fang, W. Wang, L. He et al., "Research on improved NSGA-II algorithm and its application in emergency management," *Mathematical Problems in Engineering*, vol. 2018, Article ID 1306341, 13 pages, 2018.

[17] F. Mendoza, J. L. Bernal-Agustin, and J. A. Dominguez-Navarro, "NSGA and SPEA applied to multiobjective design of power distribution systems," *IEEE Transactions on Power Systems*, vol. 21, no. 4, pp. 1938–1945, 2006.

[18] Y. Yang, Z. Wang, B. Yang et al., "Multiobjective optimization for fixture locating layout of sheet metal part using SVR and NSGA-II," *Mathematical Problems in Engineering*, vol. 2017, Article ID 7076143, 10 pages, 2017.

[19] A. Kamjoo, A. Maheri, A. M. Dizqah, and G. A. Putrus, "Multi-objective design under uncertainties of hybrid renewable energy system using NSGA-II and chance constrained programming," *International Journal of Electrical Power & Energy Systems*, vol. 74, pp. 187–194, 2016.

[20] M. Elarbi, S. Bechikh, A. Gupta et al., "A new decomposition-based NSGA-II for many-objective optimization," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 7, pp. 1191–1210, 2017.

[21] F. Wang, L. Zhou, H. Ren et al., "Multi-objective optimization model of source–load–storage synergetic dispatch for a building energy management system based on TOU price demand response," *IEEE Transactions on Industry Applications*, vol. 54, no. 2, pp. 1017–1028, 2017.

[22] F. Wang, L. Zhou, H. Ren, and X. Liu, "Search improvement process-chaotic optimization-particle swarm optimization-elite retention strategy and improved combined cooling-heating-power strategy based two-time scale multi-objective optimization model for stand-alone microgrid operation," *Energies*, vol. 10, no. 12, p. 1936, 2017.

[23] J. Cao, C. Crozier, M. McCulloch et al., "Optimal design and operation of a low carbon community based multi-energy systems considering EV integration," *IEEE Transactions on Sustainable Energy*, vol. 10, no. 3, pp. 1217–1226, 2018.

[24] R. Salem, A. Bahadori-Jahromi, A. Mylona, P. Godfrey, and D. Cook, "Comparison and evaluation of the potential energy, carbon emissions, and financial impacts from the incorporation of CHP and CCHP systems in existing UK hotel buildings," *Energies*, vol. 11, no. 5, p. 1219, 2018.

[25] G. S. Piperagkas, A. G. Anastasiadis, and N. D. Hatziargyriou, "Stochastic PSO-based heat and power dispatch under environmental constraints incorporating CHP and wind power units," *Electric Power Systems Research*, vol. 81, no. 1, pp. 209–218, 2011.

[26] N. Ghadimi, M. Afkousi-Paqaleh, and A. Nouri, "PSO based fuzzy stochastic long-term model for deployment of distributed energy resources in distribution systems with several objectives," *IEEE Systems Journal*, vol. 7, no. 4, pp. 786–796, 2013.

[27] Z. Aidong, H. Sipeng, X. Qingshan et al., "A day-ahead optimal economic dispatch schedule for building CCHP system based on centralized energy storage infrastructure," *Elektronika Ir Elektrotechnika*, vol. 24, no. 4, pp. 53–58, 2018.

[28] T. Bingmann, A. Eberle, and P. Sanders, "Engineering parallel string sorting," *Algorithmica*, vol. 77, no. 1, pp. 235–286, 2017.

[29] Z. Marszałek, M. Woźniak, and D. Połap, "Fully flexible parallel merge sort for multicore architectures," *Complexity*, vol. 2018, Article ID 8679579, 19 pages, 2018.

[30] X. Zhang, Y. Tian, R. Cheng et al., "An efficient approach to nondominated sorting for evolutionary multiobjective optimization," *IEEE Transactions on Evolutionary Computation*, vol. 19, no. 2, pp. 201–213, 2014.

[31] Y. Qin, X. U. Er-shu, and Y. Yong-ping, "Pollutant emission reduction analysis of distributed energy resource," in *Proceeding of the The 2nd International Conference on Bioinformatics and Biomedical Engineering (iCBBE 2008)*, Shanghai, China, May 2008.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 8273173, 12 pages
https://doi.org/10.1155/2020/8273173

Research Article

Study on Optimal Generative Network for Synthesizing Brain Tumor-Segmented MR Images

Hyunhee Lee,[1] **Jaechoon Jo**,[2] **and Heuiseok Lim**[1]

[1]*Department of Computer Science and Engineering, Korea University, Seoul 02841, Republic of Korea*
[2]*Division of Computer Engineering, College of Information Technology, Hanshin University, Osan-si 18101, Republic of Korea*

Correspondence should be addressed to Heuiseok Lim; limhseok@korea.ac.kr

Received 24 February 2020; Accepted 13 April 2020; Published 20 May 2020

Guest Editor: Sanghyuk Lee

Due to institutional and privacy issues, medical imaging researches are confronted with serious data scarcity. Image synthesis using generative adversarial networks provides a generic solution to the lack of medical imaging data. We synthesize high-quality brain tumor-segmented MR images, which consists of two tasks: synthesis and segmentation. We performed experiments with two different generative networks, the first using the ResNet model, which has significant advantages of style transfer, and the second, the U-Net model, one of the most powerful models for segmentation. We compare the performance of each model and propose a more robust model for synthesizing brain tumor-segmented MR images. Although ResNet produced better-quality images than did U-Net for the same samples, it used a great deal of memory and took much longer to train. U-Net, meanwhile, segmented the brain tumors more accurately than did ResNet.

1. Introduction

General characteristics of medical imaging data are as follows. It is difficult to obtain a large volume of data, and it is more difficult to acquire labelled data necessary for supervised learning. As shown in Figure 1, since the picture archiving and communicating system (PACS) was introduced in hospitals, vast amounts of multimedia data in the medical imaging field have been stored. However, due to various institutional and privacy issues, external institutions have difficulty gaining access to such data. Additionally, utilizing the accumulated data for learning requires preprocessing the data, which consequentially takes considerable time and effort.

In addition, medical imaging data have the following characteristics. A typical chest X-ray image contains 2,000 pixels horizontally and 2,500 vertically, which results in a total of five million pixels. Meanwhile, the lesions usually occupy a relatively small part of the whole image. Magnetic resonance imaging (MRI) scans provide more detailed information about the inner organs such as the brain, skeletal system, and other organ systems than do computerized tomography (CT) scans. Although MRI has many advantages, it has some disadvantages such as prolonged acquisition time (about 45 min), high costs, and limiting patient factors such as claustrophobia or metal devices in their bodies [3]. Because MRI scanners use strong magnetic fields and magnetic field gradients, MRI scans could be dangerous especially for a patient with nonremovable metal inside the body [4], and therefore the acquired images could be blurred and abnormal. CT scans are combinations of X-ray images taken from different angles, and they are fast, painless, and noninvasive. However, they do expose the patient to radiation, albeit at a relatively low dose. To minimize radiation exposure, CT scans produce low-dose images, and as a result, they unfortunately tend to be severely degraded by excessive noise and streak artifacts. For these reasons, the medical imaging field contains a small number of available medical data.

Generative adversarial networks (GANs) [5] provide a generic solution to the lack of medical imaging data. As shown in Figure 2, they can be applied to diverse tasks such as image synthesis [7, 9], segmentation [6, 11, 12], reconstruction, and classification [8]. Figure 3 shows the statistics

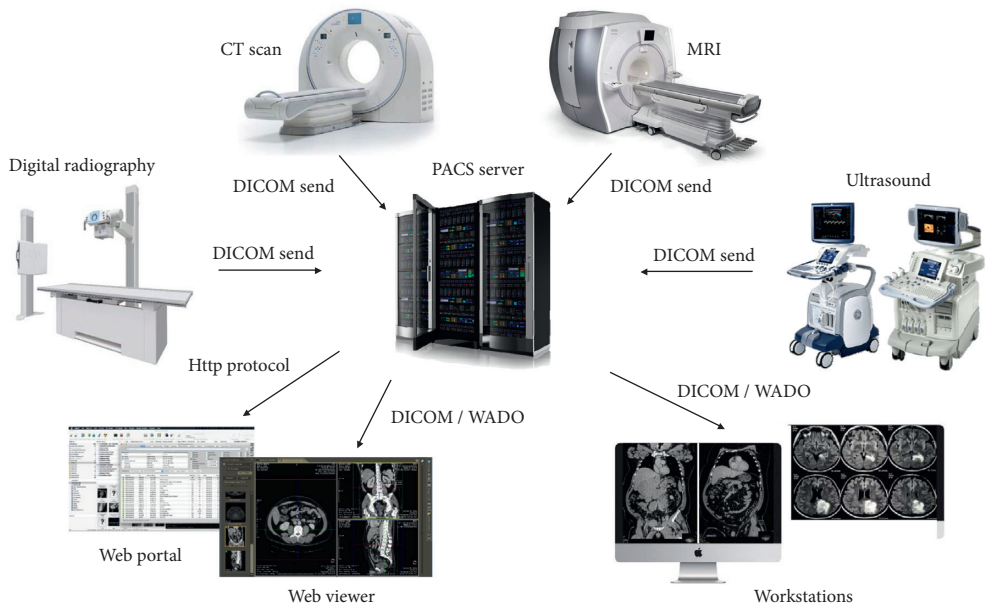


Figure 1: The architecture of PACS [1]. It is a medical imaging technology that provides economical storage and convenient access to images from multiple modalities such as MR, CT, and X-ray [2].

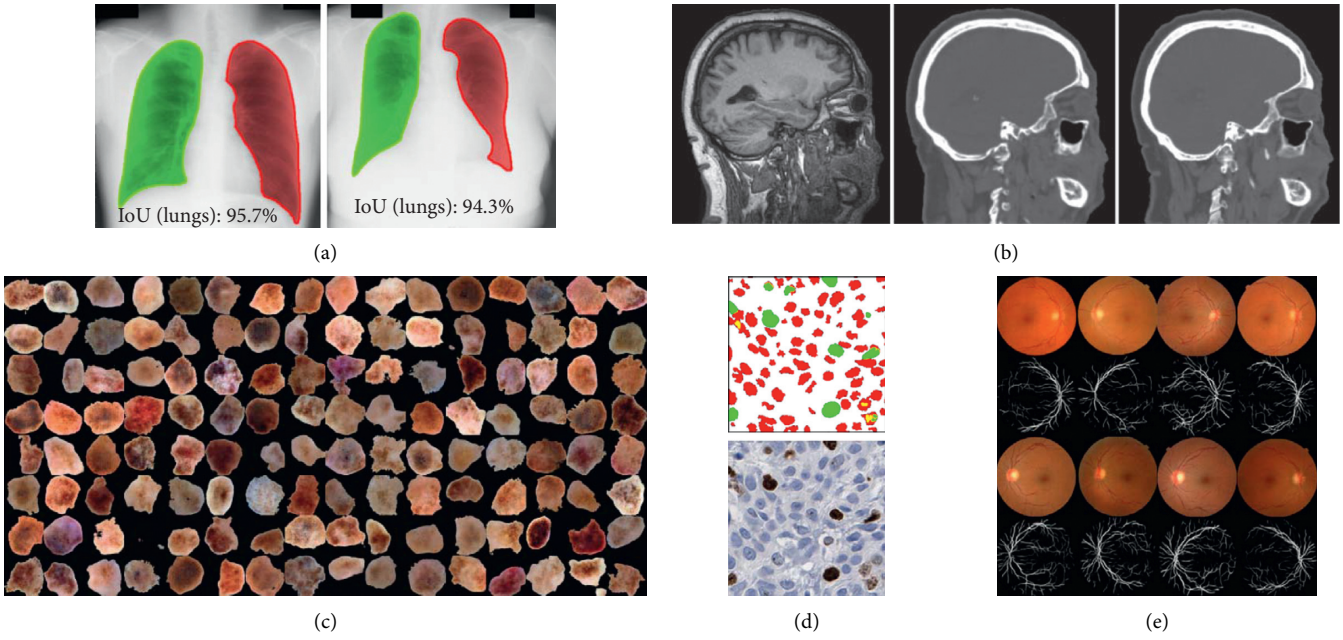


Figure 2: Example applications using GANs. (a) Organ (lung and heart) segmentation on chest X-ray of adult [6]. (b) Input MR image, synthesized CT image, and real CT image [7]. (c) Randomly generated skin lesions from random noise (a mixture of malignant and benign) [8]. (d) Ki67 synthetic image by using the segmentation/annotation [9]. (e) Generated retinal fundus images and vessel maps [10].

of GAN-related papers categorized by tasks and imaging modalities. These statistics are based on the databases of PubMed, arXiv, proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), SPIE Medical Imaging, IEEE International Symposium on Biomedical Imaging (ISBI), and International conference on Medical Imaging with Deep Learning (MIDL) [13]. The cutoff date of the search was set as July 30, 2018, and the number of GAN-related papers increased significantly in 2017 and 2018. As Figure 3 shows, about 70% of these papers studies on image synthesis and segmentation, and MR is the most-studied imaging modality in the GAN-related publications.

In this paper, we use a CycleGAN [14] which has the significant performance in medical imaging so as to synthesize brain tumor-segmented MR images. Generating brain tumor-segmented MR images consists of two tasks, namely, synthesis and segmentation. One entails image-to-image

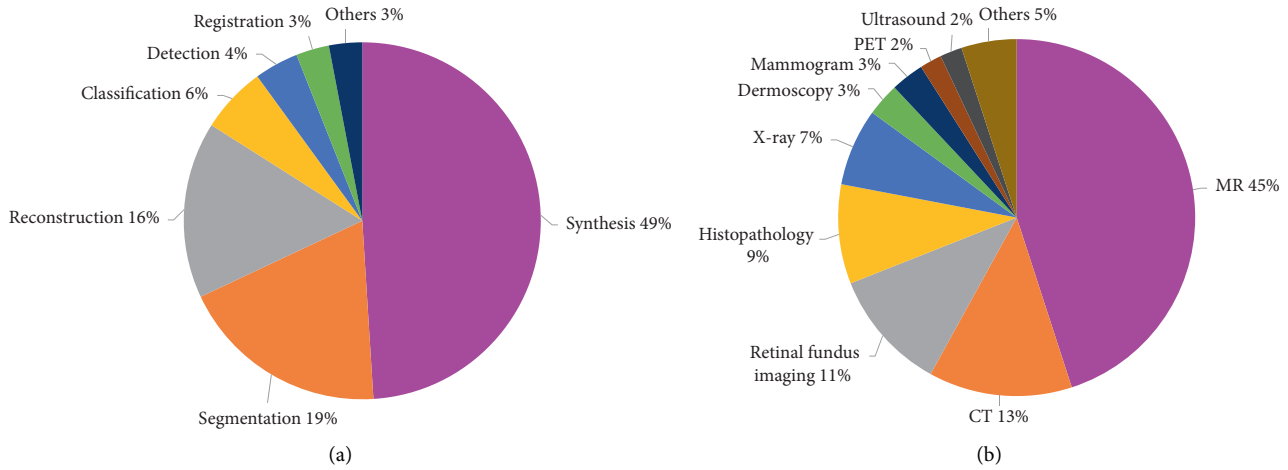


Figure 3: GAN-related publications published on or before July 30, 2018, categorized according to (a) canonical tasks and (b) imaging modalities [13].

translation, which synthesizes a novel image by combining the content of one image with the style of another. The other involves locating and marking lesions on the images. To perform these two tasks, we conduct two experiments using two different generative networks. In the first experiment in this paper, we use a ResNet [15] model, which has significant advantages in style transfer. In the second, we use a U-Net [16] model, one of the most commonly used segmentation techniques. We compare the performance of each model and propose a more robust model for synthesizing brain tumor-segmented MR images, which could lead to high-quality multimedia data augmentation in the medical imaging field.

2. Related Work

In a study on brain CT image synthesis from MR images by Wolterink et al. [7], training using unpaired images was even better than using paired images [13]. That is, the performance of CycleGAN is greater than that of Pix2pix [17] in cross-modality synthesis of medical images. In addition, as shown in Table 1, the use of CycleGAN in the recent publications for medical image synthesis is increasing. In this section, we discuss CycleGAN, ResNet, and U-Net used in our experiments.

2.1. CycleGAN. The problem that Pix2pix or CycleGAN has to solve is the interpretation of images from one domain to another. In Pix2pix, there must be data pairs corresponding to both domains, whereas CycleGAN can solve this problem without these data pairs [14]. CycleGAN is a GAN using two generators and two discriminators. We call one generator G, and it converts images from the X domain to the Y. The other generator is called F, and it converts images from the Y domain to the X. Each generator has a corresponding discriminator that attempts to tell apart its synthesized images from real ones. CycleGAN is not just a simple mapping technique. It considers the returning mapping and puts a constraint on coming back to its original state. As shown in Figure 4, not only mapping from X to Y but also mapping back to original X from Y should be defined, and this applies to the opposite mapping as well. The reason for doing this is X and Y are unpaired domains. When X goes to Y, it is checked that it looks like Y, and the actual constraint is to keep the original shape when it returns to X again. That is, the shape of X does not change much, only its style is changed to that of Y, and therefore it looks as if only the style has been transferred.

Table 1: Published works on cross-modality image synthesis.

Modality	Publication	Method	Remarks
CT ⟶ MR	Jin et al. [18]	CycleGAN	Brain
MR ↔ CT	Zang et al. [19]	CycleGAN*	(3D) Cardiac
MR ↔ CT	Hiasa et al. [20]	CycleGAN*	Musculoskeletal
MR ↔ CT	Wolterink et al. [7]	CycleGAN	Brain
T1 ↔ T2 MR	Dar et al. [21]	CycleGAN	Brain
T1 ↔ T2 MR	Welander et al. [22]	CycleGAN	Brain

In the third column, *following the method denotes some modifications either on the architecture of the network or on the employed losses.

There are two components in the CycleGAN objective function, an adversarial loss and a cycle-consistency loss. Both are essential for successful results. The adversarial loss alone is not sufficient to produce high-quality images, which leaves the model underconstrained [14]. In other words, it forces the generated output to be of the appropriate domain but does not force the input and output to be recognizably the same. The cycle-consistency loss addresses this underconstrained problem. The full objective function by putting these loss terms together and by weighting the cycle-consistency loss with a hyper parameter λ is defined as equation (1). Pix2pix can rely heavily on L1 loss; therefore, adversarial loss may play a supplementary role and it may be better to subtract it. In contrast, the CycleGAN does not learn at all except for the adversarial loss.

$$\mathcal{L}_{\text{full}} = \mathcal{L}_{\text{adv}} + \lambda \mathcal{L}_{\text{cyc}}. \quad (1)$$

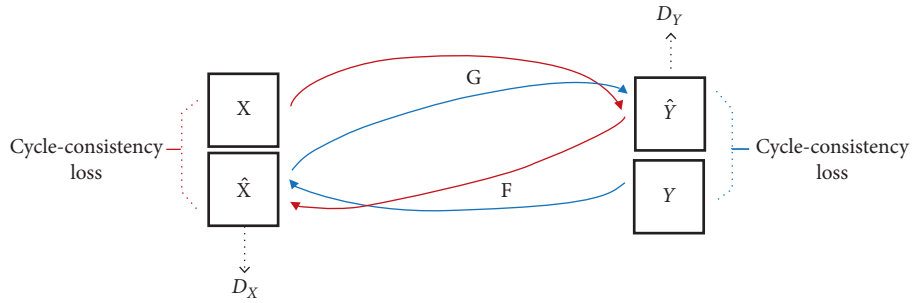


Figure 4: CycleGAN with mapping functions G: X $\longrightarrow$ Y and F: Y $\longrightarrow$ X, and associated adversarial discriminators D_Y and D_X. Forward cycle-consistency loss (red): X $\longrightarrow$ G(X) $\longrightarrow$ F(G(X)) $\approx \hat{X}$, and backward cycle-consistency loss (blue): Y $\longrightarrow$ F(Y) $\longrightarrow$ G(F(Y)) $\approx \hat{Y}$.

2.2. ResNet. The general CNN network in Figure 5(a) receives the input x and yields the output $H(x)$ through the two weighted layers. This output is the input to the next layer. Figure 5(b) shows the architecture of the ResNet, and it uses a *shortcut connection* that connects the input of the layer directly to the output [15]. It is simple network, but its performance is significantly high. As can be seen in Figure 5(b), the output $H(x)$ has been changed to $F(x) + x$, and consequently $H(x) = F(x) + x$ is derived as $F(x) = H(x) - x$. Learning $F(x)$ can be seen as learning residual, and therefore it is called ResNet. It adds the results from the weight layer and the previous results and uses ReLU. ResNet learns in the direction that $F(x)$ becomes zero. In addition, since x is directly connected to *shortcut connection*, there is no increase in computation, and it is possible to select which layer to include. For example, a fully connected layer as well as a convolution layer can be added [15].

2.3. U-Net. Segmentation is the process of partitioning an image into different meaningful segments [23]. In medical imaging, these segments often correspond to different tissue classes, organs, pathologies, or other biologically relevant structures [24]. In the past, there were few medical images, and therefore experts could segment images directly. However, more needs for the automation of the segmentation have arisen as the volume of the medical images has increased exponentially. Analysing medical images can often be difficult and time consuming, and therefore deep neural networks can help doctors make more rapid and more accurate diagnoses [25].

U-Net is one of the most preferred models when segmenting images. Figure 6(a) shows the general encoding and decoding process, and Figure 6(b) shows a U-Net model with a *skip connection* added to the encoder-decoder structure. If the image size is reduced (down sampling) and then reraised (up sampling), sophisticated pixel information disappears. This is a big problem for image segmentation which requires dense prediction on a pixel-by-pixel basis. The *skip connection*, which passes important information directly from the encoder to the decoder, results in a much clearer image at the decoder section, allowing for more accurate prediction.

3. Materials and Methods

3.1. Materials. The first CycleGAN used the U-Net model. The advantage of *skip connection* is that it has much more detail, but the disadvantage is that the performance is not good when the two contents are similar. On the other hand, the last CycleGAN used ResNet model, which is good for image quality but has the disadvantage of using a great deal of memory.

In this paper, we synthesize high-quality brain tumor-segmented MR images. It consists of two tasks, namely, synthesis and segmentation, and therefore we conduct experiments from two perspectives. One is to perform image-to-image translation, which synthesizes a novel image with the style of another image. The other is to locate and mark tumors in brain MR image. Therefore, we perform two experiments with two different generative networks. In the first experiment, we use a ResNet model, which has significant advantages in style transfer. In the second experiment, we use a U-Net model, which is one of the most commonly used segmentation techniques. In this paper, we compare the performance of each model and propose a more powerful model for synthesizing brain tumor-segmented MR images.

Figure 7 shows the datasets of source and target domains used in our experiments. Figure 7(a) represents the brain lesion images in the source domain, and Figure 7(b) is the segmentation mask images of the brain lesion in the target domain. We used a training set of 765 unaligned image pairs and a test set of 92 unaligned image pairs in our experiments. The size of each image is on average 30 to 50 KB. The environments for our experiments are as follows: Ubuntu 16.04.4 LTS for OS, GeForce GTX 1080Ti and CUDA Toolkit for high-performance GPU-accelerated application, and TensorFlow library for deep learning framework.

3.2. Architecture of Our Discriminative Model. The configuration of the discriminative model in our experiments is shown in Figure 8. It consists of four convolution layers, and we use leaky ReLU as an activation function for each layer. In the first step, we extract the features from the image, and in the last, we decide which specific category these features belong to. For that, we add a final convolution layer that

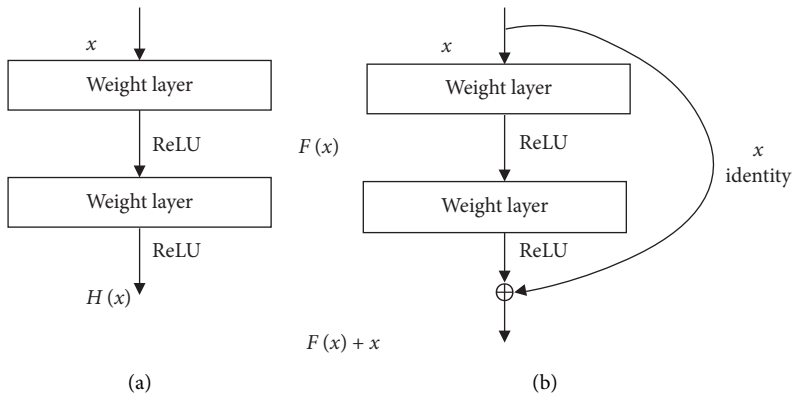


Figure 5: The architecture of the networks: (a) CNN and (b) ResNet. The output in CNN is the input of the next layer. The input in ResNet connects directly to the output of the layer [15].

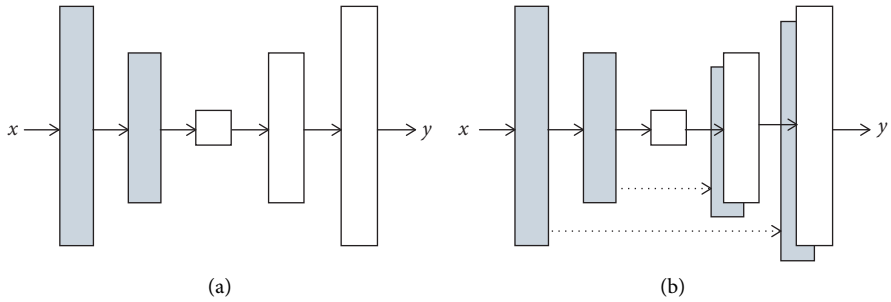


Figure 6: The architecture of the networks: (a) Encoder-decoder and (b) U-Net. U-Net is a model with a *skip connection* added to the encoder-decoder structure [17].

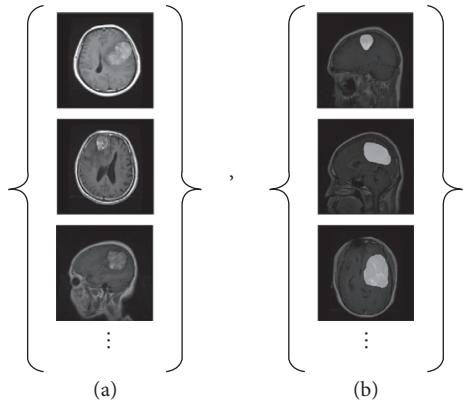


Figure 7: Our unpaired datasets: (a) brain tumor; (b) segmentation mask.

produces a one-dimensional output. Both ResNet and U-Net generative models use this model as a discriminator in our experiments.

3.3. Architecture of Our Generative Model Using ResNet. The generator has the job of taking an input image and performing the transformation to produce the target image. The architecture of our generative model using ResNet can be viewed in Figure 9. First, the encoding process consists of three convolution layers, and ReLU is used as an activation function for each layer. In the transformation process, nine residual blocks are constructed, and each block consists of convolution layer-ReLU-convolution layer. The decoding process consists of two deconvolution layers, and each uses ReLU as an activation function. In the last decoding step, we add a final convolution layer.

3.4. Architecture of Our Generative Model Using U-Net. As shown in Figure 10, the encoder-decoder structure of our generative model using U-Net is as follows. First, the encoding process consists of eight convolution layers, and leaky ReLU is used as an activation function for each layer.

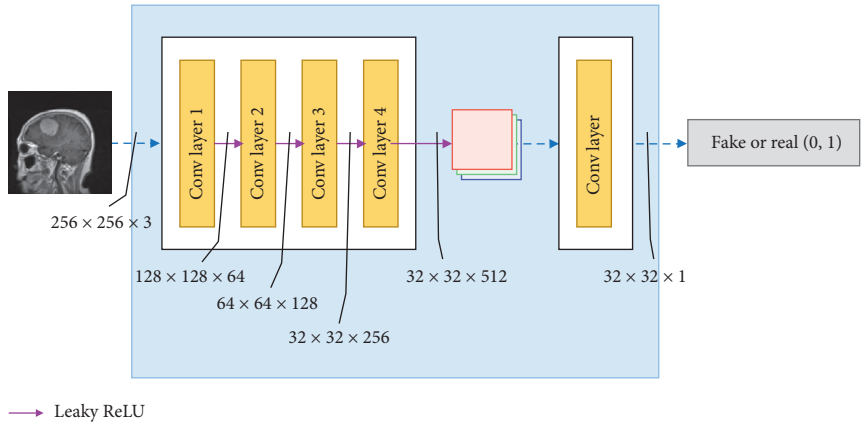


Figure 8: The architecture of our discriminative model.

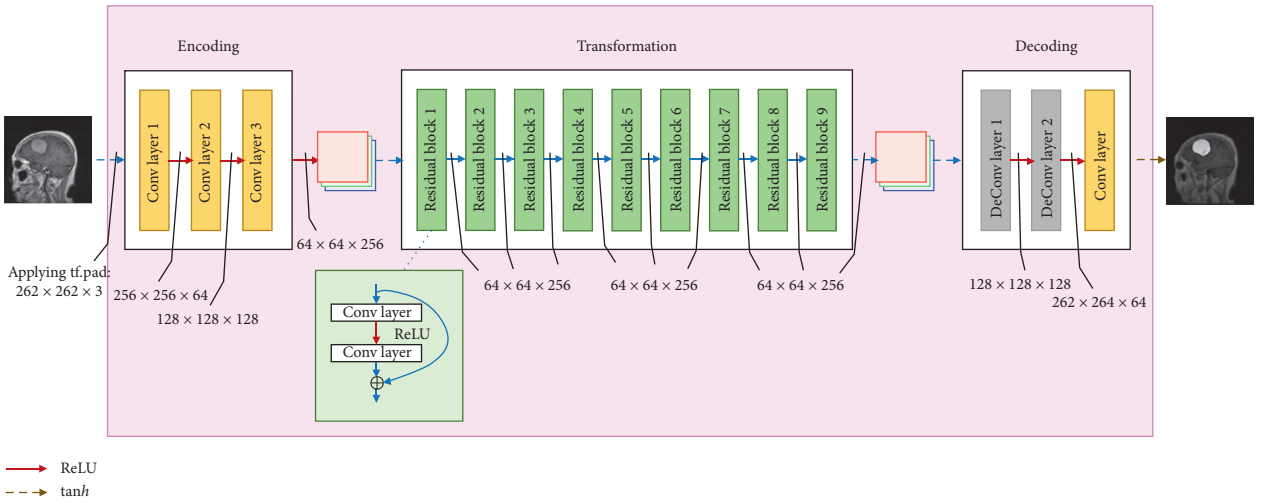


Figure 9: The architecture of our generative model using ResNet. It consists of encoding, transformation, and decoding processes.

The decoding process consists of eight deconvolution layers. ReLU is used as an activation function for each layer, and 50% dropout is performed in the first to third decoding processes. We also use *concat* function in the decoding process to perform a *skip connection* that passes important information directly from the encoder to the decoder.

3.5. Methods for Synthesizing Brain Tumor-Segmented MR Images. The models in our work includes forward and backward cycles, just like the CycleGAN model proposed by Zhu et al. [14]. With these cycles, the novel synthesized image can only obtain the style of the target image while retaining the shape of the original image.

As shown in Figure 11, the architecture of our forward and backward cycles is composed of two generators ($\text{Gen}_{A\longrightarrow B}$ and $\text{Gen}_{B\longrightarrow A}$) and two discriminators (Dis_B and Dis_A). Forward process is as follows: $A \longrightarrow \text{Gen}_{A\longrightarrow B}(A) \longrightarrow \text{Gen}_{B\longrightarrow A}(\text{Gen}_{A\longrightarrow B}(A)) \approx \widehat{A}$. More specifically, our forward process can be explained in three steps. First, generator $\text{Gen}_{A\longrightarrow B}$ is trained to translate an input *brain tumor* domain (A) into a *segmentation mask* domain ($\widehat{B}$). Second, Dis_B is trained to discriminate the generated image $\widehat{B}(\text{Gen}_{A\longrightarrow B}(A) \approx \widehat{B})$ from the real image B. Third, $\text{Gen}_{B\longrightarrow A}$ is trained to translate the generated image $\widehat{B}$ into the brain tumor MR image $\widehat{A}(\text{Gen}_{B\longrightarrow A}(\widehat{B}) \approx \widehat{A})$. Likewise, backward process is as follows: $B \longrightarrow \text{Gen}_{B\longrightarrow A}(B) \longrightarrow \text{Gen}_{A\longrightarrow B}(\text{Gen}_{B\longrightarrow A}(B)) \approx \widehat{B}$. More specifically, our backward process can be also defined in three steps. First, $\text{Gen}_{B\longrightarrow A}$ is trained to translate an input *segmentation mask* domain (B) into a *brain tumor* domain ($\widehat{A}$). Second, Dis_A is trained to discriminate the generated image $\widehat{A}(\text{Gen}_{B\longrightarrow A}(B) \approx \widehat{A})$ from the real image A. Third, $\text{Gen}_{A\longrightarrow B}$ is trained to translate the generated image $\widehat{A}$ into the segmentation mask MR image $\widehat{B}(\text{Gen}_{A\longrightarrow B}(\widehat{A}) \approx \widehat{B})$.

The goal of the discriminator is to distinguish the novel image generated by the two generators from the real one, and therefore the discriminative neural network is trained to minimize the final classification error. On the other hand, the goal of the generator is to fool the discriminator, and therefore the generative neural network is trained to maximize the final classification error. Both networks attempt to beat each other, and this competition between them makes them evolve with respect to their respective goals. Hence, the adversarial loss function that discriminator Dis_B aims to

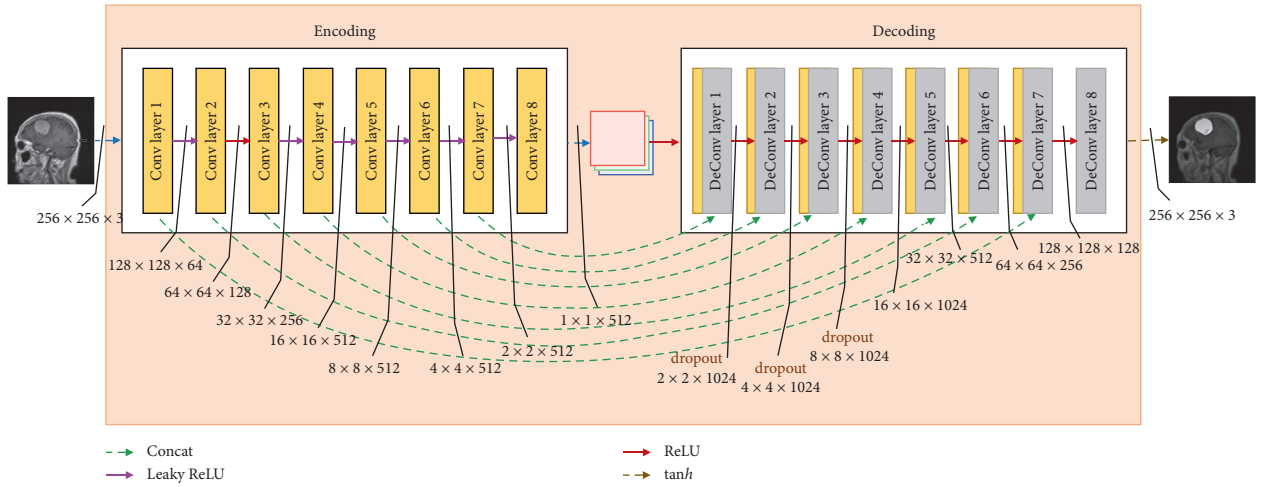


FIGURE 10: The architecture of our generative model using U-Net. We use *concat* in the decoding process to perform a *skip connection* that passes important information directly from the encoder to the decoder.

minimize and generator $\text{Gen}_{A\longrightarrow B}$ aims to maximize is defined as

$$\mathscr{L}_B = \left(1 - \text{Dis}_B(B)\right)^2 + \text{Dis}_B\left(\text{Gen}_{A\longrightarrow B}(A)\right)^2. \tag{2}$$

Next, the adversarial loss function that discriminator Dis_A aims to minimize and generator $\text{Gen}_{B\longrightarrow A}$ aims to maximize is defined as

$$\mathscr{L}_A = \left(1 - \text{Dis}_A(A)\right)^2 + \text{Dis}_A\left(\text{Gen}_{B\longrightarrow A}(B)\right)^2. \tag{3}$$

Adversarial loss alone cannot guarantee that the learned function can map an input domain A to a target domain B. To regularize the model and to transform source distribution into target and then back again, we should introduce the constraint of cycle-consistency into the model. An additional loss term, cycle-consistency loss, is defined as

$$\begin{aligned}\mathscr{L}_{\text{cyc}} = &\left\|\text{Gen}_{B\longrightarrow A}\left(\text{Gen}_{A\longrightarrow B}(A)\right) - A\right\|_1 \\ &+ \left\|\text{Gen}_{A\longrightarrow B}\left(\text{Gen}_{B\longrightarrow A}(B)\right) - B\right\|_1.\end{aligned} \tag{4}$$

Our full objective function by putting these loss terms together and by weighting the cycle-consistency loss with a hyper parameter λ is defined as

$$\mathscr{L}_{\text{full}} = \mathscr{L}_B + \mathscr{L}_A + \lambda\mathscr{L}_{\text{cyc}}, \tag{5}$$

where we set λ to 10, which is the optimal value as introduced in CycleGAN.

3.6. Evaluation. We basically use metrics like mean-squared error and mean absolute error to evaluate the performance of the ResNet model as well as that of the U-Net model. We also evaluate the performance of the discriminator Dis_B in each model using the following metrics:

$$\mathscr{L}\left(\text{Dis}_{\text{real}B}\right) = \frac{1}{M}\sum_{i=1}^{M}\left(1 - \text{Dis}_B(B)\right)^2, \tag{6}$$

which is the discriminator's loss for the real image B:

$$\mathscr{L}\left(\text{Dis}_{\text{fake}B}\right) = \frac{1}{M}\sum_{i=1}^{M}\left(\text{Dis}_B\left(\text{Gen}_{A\longrightarrow B}(A)\right)\right)^2, \tag{7}$$

which is the discriminator's loss for the fake image B synthesized by a generator $\text{Gen}_{A\longrightarrow B}$, and

$$\mathscr{L}\left(\text{Dis}_B\right) = \mathscr{L}\left(\text{Dis}_{\text{real}B}\right) + \mathscr{L}\left(\text{Dis}_{\text{fake}B}\right), \tag{8}$$

which is the sum of the discriminator's losses for both $\text{Dis}_{\text{real}B}$ and $\text{Dis}_{\text{fake}B}$.

We also evaluate the performance of the generator $\text{Gen}_{A\longrightarrow B}$ in each model using the following metrics:

$$\begin{aligned}\mathscr{L}(\text{Cycle}) = &\lambda\frac{1}{M}\sum_{i=1}^{M}\left|B(i) - \text{Gen}_{A\longrightarrow B}(A(i))\right| \\ &+ \lambda\frac{1}{M}\sum_{i=1}^{M}\left|A(i) - \text{Gen}_{B\longrightarrow A}(B(i))\right|,\end{aligned} \tag{9}$$

which is the forward and backward cycle-consistency loss, where the hyper parameter λ is set to 10, and

$$\mathscr{L}\left(\text{Gen}_{A\longrightarrow B}\right) = \mathscr{L}(\text{Cycle}) + \mathscr{L}\left(\text{Dis}_{\text{fake}B}\right), \tag{10}$$

which is the generator's loss for the fake image B synthesized by the generator $\text{Gen}_{A\longrightarrow B}$, and vice versa for Dis_A and $\text{Gen}_{B\longrightarrow A}$. In addition to the above-metioned evaluation metrics, we use human perception to judge the visual quality of samples. We evaluate the quality of the generated images and whether they can segment the brain tumors well.

4. Results and Discussion

Our training process is as follows. We used 20 epochs, 100 steps, and Adam as an optimizer with initial learning rate of 0.0002 and Adam's momentum term of 0.5. The training took 28.6 hours for the ResNet model and 6.9 hours for the U-Net model. That is, the training of the ResNet took four times as long as that of the U-Net.

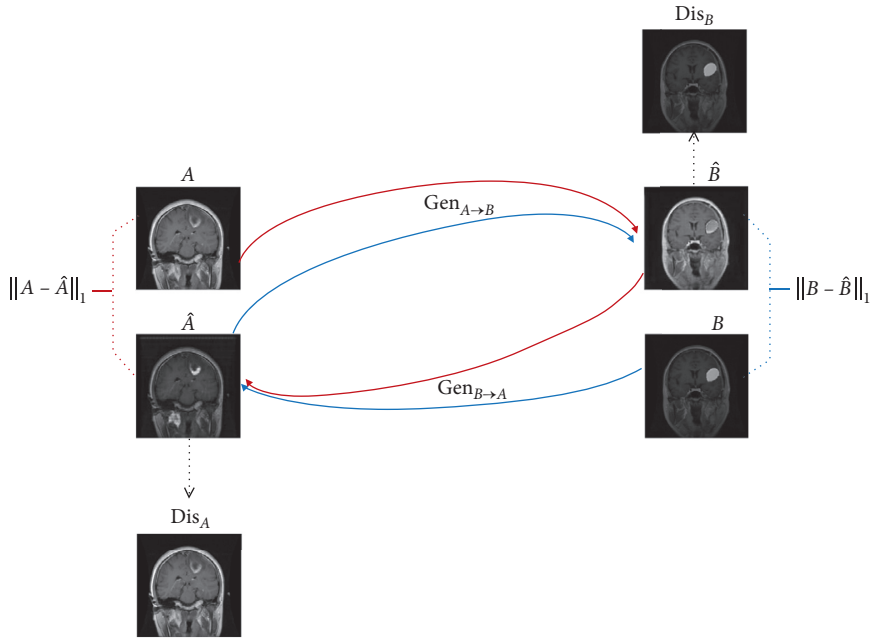
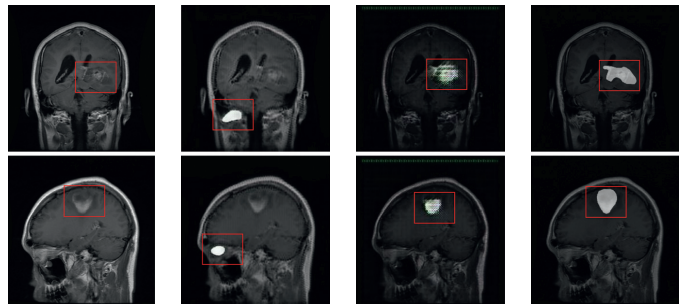


Figure 11: The architecture of our forward and backward cycles, composed of two generators ($\text{Gen}_{A\longrightarrow B}$ and $\text{Gen}_{B\longrightarrow A}$) and two discriminators (Dis_B and Dis_A). Forward cycle process: $A \longrightarrow \text{Gen}_{A\longrightarrow B}(A) \longrightarrow \text{Gen}_{B\longrightarrow A}(\text{Gen}_{A\longrightarrow B}(A)) \approx \widehat{A}$. Backward cycle process: $B \longrightarrow \text{Gen}_{B\longrightarrow A}(B) \longrightarrow \text{Gen}_{A\longrightarrow B}(\text{Gen}_{B\longrightarrow A}(B)) \approx \widehat{B}$. Dis_B is trained to discriminate the generated image $\widehat{B}$ from the real image B. Dis_A is trained to discriminate the generated image $\widehat{A}$ from the real image A.

Table 2: Our discriminator's and generator's losses for each generative model: ResNet, and U-Net.

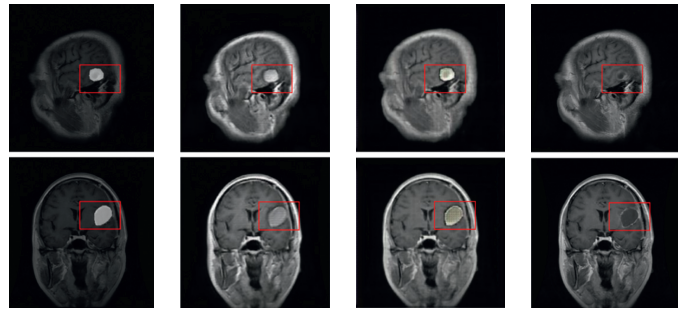
Loss	ResNet	U-Net	Remarks
Dis	0.430	4.890e-3	Sum of discriminator's losses for A, B
Dis_A	0.206	2.358e-3	Sum of discriminator's losses for both real A and fake A
$\text{Dis}_{\text{fake}A}$	0.282	8.767e-4	Discriminator's loss for fake A
$\text{Dis}_{\text{real}A}$	0.131	3.840e-3	Discriminator's loss for real A
Dis_B	0.223	2.532e-3	Sum of discriminator's losses for both real B and fake B
$\text{Dis}_{\text{fake}B}$	0.252	3.000e-3	Discriminator's loss for fake B
$\text{Dis}_{\text{real}B}$	0.195	1.027e-3	Discriminator's loss for real B
Gen	1.338	2.572	Sum of generator's losses for $A \longrightarrow B$, $B \longrightarrow A$
$\text{Gen}_{A\longrightarrow B}$	1.185	1.591	Generator's loss for $A \longrightarrow B$
$\text{Gen}_{B\longrightarrow A}$	1.125	1.549	Generator's loss for $B \longrightarrow A$

A is the brain tumor domain and B is the segmentation mask domain.



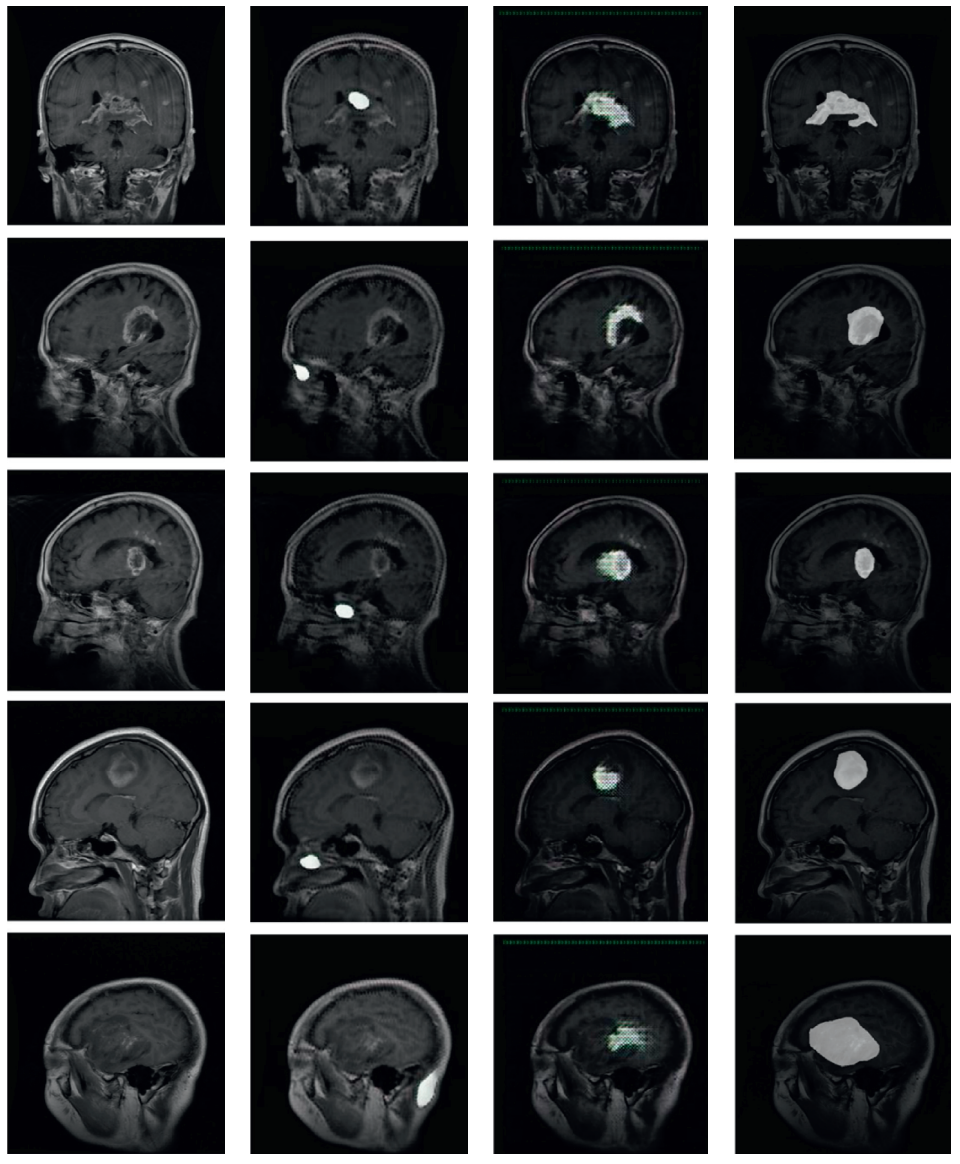
(a) (b) (c) (d)

Figure 12: Images generated by each generative model, ResNet, and U-Net, for the same samples ($A \longrightarrow B$). The U-Net model locates and marks tumors in the brain MR image, whereas ResNet does not segment the exact location of the brain lesions: (a) input, (b) ResNet, (c) U-Net, and (d) ground truth.



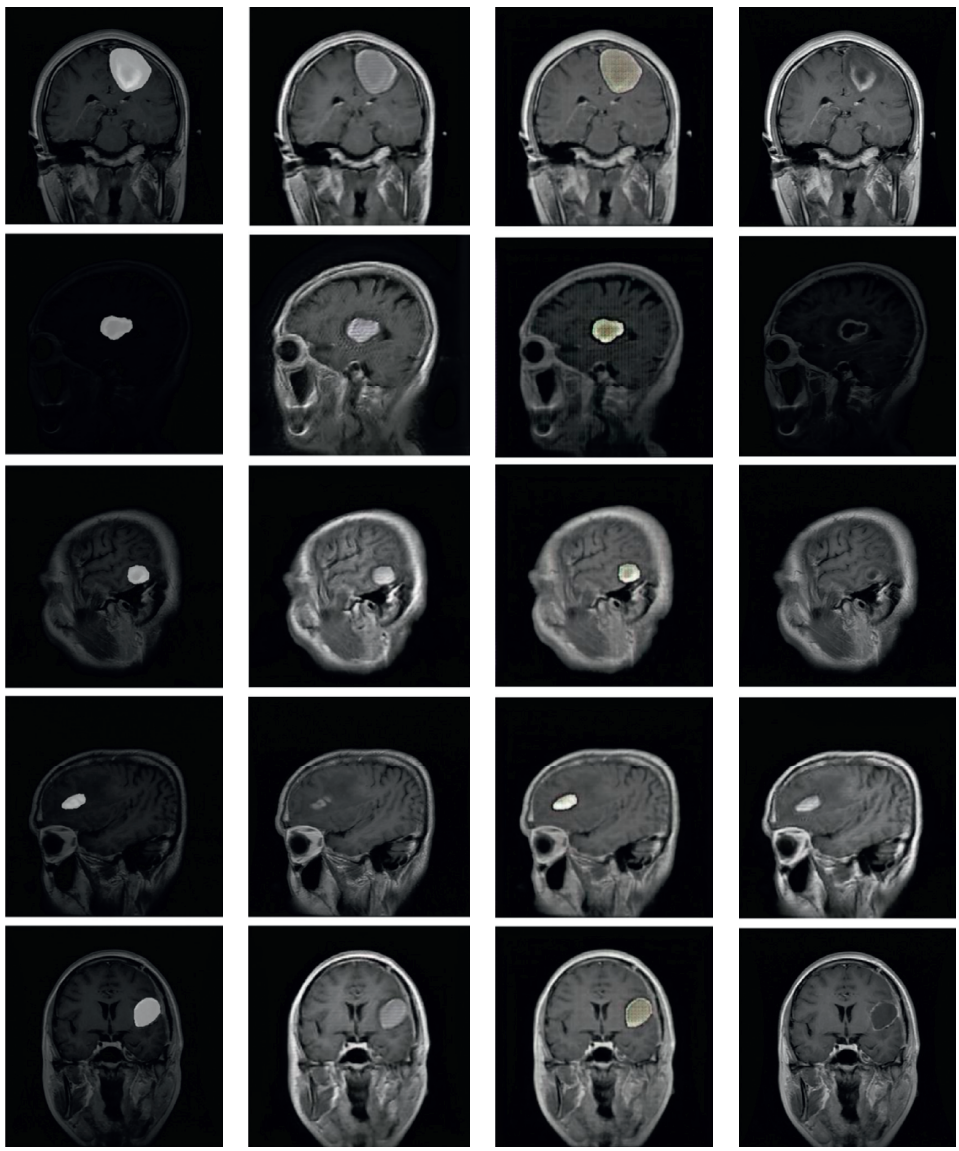
(a) (b) (c) (d)

FIGURE 13: Images generated by each generative model, ResNet, and U-Net, for the same samples ($B \longrightarrow A$). Both networks produced high-quality images. However, the product's quality of ResNet is better than that of U -Net, (a) input, (b) ResNet, (c) U-Net, and (d) ground truth.



(a) (b) (c) (d)

FIGURE 14: Images generated by each generative model, ResNet, and U-Net, for the same samples ($A \longrightarrow B$). U-Net locates and marks tumors in the brain MR image. In contrast, ResNet does not segment the exact location of the brain lesions: (a) input, (b) ResNet, (c) U-Net, and (d) ground truth.



(a) (b) (c) (d)

Figure 15: Images generated by each generative model, ResNet, and U-Net, for the same samples ($B \longrightarrow A$). Both networks produced high-quality images. However, the product's quality of ResNet is better than that of U-Net: (a) input, (b) ResNet, (c) U-Net, and (d) ground truth.

Table 2 shows all losses of the discriminator and the generator in each training process of the ResNet and the U-Net. Full loss means the sum of all losses not only from the *brain tumor* domain to the *segmentation mask* domain but also from the *segmentation mask* domain to the *brain tumor* domain. As shown in Table 2, the full loss of the discriminator is 0.430 for the ResNet and $4.890e-3$ for the U-Net. It indicates that the performance of the discriminator is significantly higher when using U-Net than using ResNet. In contrast, the full loss of the generator is 1.338 for the ResNet and 2.572 for the U-Net. That is, the performance of the generator is slightly higher when using ResNet than using U-Net.

Additionally, Figure 12 ($A \longrightarrow B$) and Figure 13 ($B \longrightarrow A$) show the novel images generated by the ResNet and the U-Net model for the same samples. As shown in Figure 12 ($A \longrightarrow B$), U-Net, which is a more robust model for segmentation, located and marked tumors in the brain MR image and produced brain tumor-segmented MR image synthesis similar to ground truth. ResNet, in contrast, did not segment the exact location of the brain lesion. Figure 13 ($B \longrightarrow A$) shows the novel brain images generated from the segmentation mask domain. Both networks produced high-quality images, whereas the synthesis-quality of ResNet was higher than that of U-Net (see Figures 14 and 15 for more details about the novel images generated by the ResNet and the U-Net model for the same samples).

5. Conclusions

In our work, we augmented brain tumor-segmented MR images, which consists of two tasks: synthesis and

segmentation. Therefore, we conducted two experiments, one to perform image-to-image translation, namely, image style transfer, and the other to locate and mark tumors in brain MR image, that is to say image segmentation. We performed experiments with two different generative networks, the first one using the ResNet model, which has great advantages in style transfer, and the second one, the U-Net model, one of the highly robust models for segmentation.

The performance comparison between the ResNet and the U-Net generative model is as follows. When the generator used ResNet, its training loss was slightly less than that of U-Net, and moreover it produced better-quality images than U-Net. However, it was a memory-intensive network and took much longer to train, and it did not segment the brain tumors better than U-Net. On the other hand, when the generator was with U-Net, the discriminator performed the better discrimination of whether the generated image was real or fake. Additionally, for the same samples, U-Net segmented the brain tumors more accurately than did ResNet, i.e., the segmented images generated from the brain tumor domain marked the exact location of the brain lesions.

The generative networks proposed in our paper will enable the synthesis of not only brain tumor-segmented images, but also medical images in Figure 2, as well as the novel images of segmenting tumors from the breast, uterus, and other organs, depending on the intended application. In future work, we will apply a network that combines the advantages of two networks. If we merge two models, it will be possible to generate a high-quality synthetic image with accurate segmentation. We will also increase the number of epochs and adjust hyper parameters such as initial learning rate. High-quality multimedia data augmentation using GANs has a direct impact on radiology workflow and patient care improvement. Although promising results have been reported, the adoption of GANs in medical imaging is still in its infancy and there are no clinically adopted breakthrough applications yet. Therefore, more studies and more diverse attempts are needed.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by the Ministry of Science and ICT (MSIT), Korea, under the Information Technology Research Center (ITRC) support program (IITP-2020-2018-0-01405) supervised by the Institute for Information and Communications Technology Planning and Evaluation (IITP). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2017M3C4A7068189).

References

[1] O. Ratib, N. Roduit, D. Nidup, G. De Geer, A. Rosset, and A. Geissbuhler, "PACS for Bhutan: a cost effective open source architecture for emerging countries," *Insights Into Imaging*, vol. 7, no. 5, pp. 747–753, 2016.

[2] R. H. Choplin, J. M. Boehme II, and C. D. Maynard, "Picture archiving and communication systems: an overview," *Radiographics*, vol. 12, no. 1, pp. 127–129, 1992.

[3] S. Lütje, J. W. J. de Rooy, S. Croockewit, E. Koedam, W. J. G. Oyen, and R. A. Raymakers, "Role of radiography, MRI and FDG-PET/CT in diagnosing, staging and therapeutical evaluation of patients with multiple myeloma," *Annals of Hematology*, vol. 88, no. 12, pp. 1161–1168, 2009.

[4] F. G. Shellock and J. V. Crues, "MR procedures: biologic effects, safety, and patient care," *Radiology*, vol. 232, no. 3, pp. 635–652, 2004.

[5] I. Goodfellow, J. Pouget-Abadie, and M. Mirza, "Generative adversarial nets," in *Proceedings of the Paper presented at the Advances in Neural Information Processing Systems*, Montreal, Canada, December 2014.

[6] W. Dai, J. Doyle, X. Liang, Scan: structure correcting adversarial network for chest X-rays organ segmentation, arXiv preprint arXiv:1703.08770 1, 2017, 2.

[7] J. M. Wolterink, A. M. Dinkla, and M. H. F. Savenije, "Deep MR to CT synthesis using unpaired data," in *Proceedings of the Paper presented at the International Workshop on Simulation and Synthesis in Medical Imaging*, Québec, Canada, September 2017.

[8] X. Yi, E. Walia, P. Babyn, Unsupervised and semi-supervised learning with categorical generative adversarial networks assisted by wasserstein distance for dermoscopy image classification, arXiv preprint arXiv:1804.03700, 2018.

[9] C. Senaras, M. K. K. Niazi, and B. Sahiner, "Optimized generation of high-resolution phantom images using cgan: application to quantification of Ki67 breast cancer images," *PLoS One*, vol. 13, no. 5, 2018.

[10] P. Costa, A. Galdran, and M. I. Meyer, "End-to-end adversarial retinal image synthesis," *IEEE Transactions on Medical Imaging*, vol. 37, no. 3, pp. 781–791, 2017.

[11] F. Pollastri, F. Bolelli, R. Paredes, and C. Grana, "Augmenting data with gans to segment melanoma skin lesions," *Multimedia Tools and Applications*, pp. 1–18, 2019.

[12] M. Rezaei, H. Yang, and C. Meinel, "Recurrent generative adversarial network for learning imbalanced medical image semantic segmentation," *Multimedia Tools and Applications*, pp. 1–20, 2019.

[13] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: a review," *Medical Image Analysis*, vol. 58, p. 101552, 2019.

[14] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the Paper presented at the IEEE International Conference on Computer Vision*, Venice, Italy, October 2017.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, June 2016.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in

Proceedings of the Paper presented at the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, October 2015.

[17] P. Isola, J. Y. Zhu, and T. Zhou, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, July 2017.

[18] C.-B. Jin, H. Kim, M. Liu et al., "Deep CT to MR synthesis using paired and unpaired data," *Sensors*, vol. 19, no. 10, p. 2361, 2019.

[19] Z. Zhang, L. Yang, and Y. Zheng, "Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network," in *Proceedings of the Paper Presented at the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake, UT, USA, June 2018.

[20] Y. Hiasa, Y. Otake, and M. Takao, "Cross-modality image synthesis from unpaired data using cyclegan," in *Proceedings of the Paper presented at the International workshop on simulation and synthesis in medical imaging*, Shenzhen, China, October 2018.

[21] S. U. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Cukur, "Image synthesis in multi-contrast mri with conditional generative adversarial networks," *IEEE Transactions on Medical Imaging*, vol. 38, no. 10, pp. 2375–2388, 2019.

[22] P. Welander, S. Karlsson, A. Eklund, Generative adversarial networks for image-to-image translation on multi-contrast MR images-a comparison of cyclegan and unit, arXiv preprint arXiv:1806.07777, 2018.

[23] H. Ji, D. Hooshyar, K. Kim, and H. Lim, "A semantic-based video scene segmentation using a deep neural network," *Journal of Information Science*, vol. 45, no. 6, pp. 833–844, 2019.

[24] M. Forouzanfar, N. Forghani, and M. Teshnehlab, "Parameter optimization of improved fuzzy C-means clustering algorithm for brain mr image segmentation," *Engineering Applications of Artificial Intelligence*, vol. 23, no. 2, pp. 160–168, 2010.

[25] A. So, D. Hooshyar, K. Park, and H. Lim, "Early diagnosis of dementia from clinical data by machine learning techniques," *Applied Sciences*, vol. 7, no. 7, p. 651, 2017.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 1958149, 9 pages
https://doi.org/10.1155/2020/1958149

Research Article

Validation of Text Data Preprocessing Using a Neural Network Model

HoSung Woo,[1] **JaMee Kim**,[2] **and WonGyu Lee**[3]

[1]*Department of Computer Science and Engineering, Graduate School, Korea University, Seoul 02841, Republic of Korea*
[2]*Major of Computer Science Education, Graduate School of Education, Korea University, Seoul 02841, Republic of Korea*
[3]*Department of Computer Science and Engineering, College of Informatics, Korea University, Seoul 02841, Republic of Korea*

Correspondence should be addressed to WonGyu Lee; lee@inc.korea.ac.kr

Received 24 March 2020; Accepted 1 May 2020; Published 14 May 2020

Guest Editor: Sanghyuk Lee

Many artificial intelligence studies focus on designing new neural network models or optimizing hyperparameters to improve model accuracy. To develop a reliable model, appropriate data are required, and data preprocessing is an essential part of acquiring the data. Although various studies regard data preprocessing as part of the data exploration process, those studies lack awareness about the need for separate technologies and solutions for preprocessing. Therefore, this study evaluated combinations of preprocessing types in a text-processing neural network model. Better performance was observed when two preprocessing types were used than when three or more preprocessing types were used for data purification. More specifically, using lemmatization and punctuation splitting together, lemmatization and lowering together, and lowering and punctuation splitting together showed positive effects on accuracy. This study is significant because the results allow better decisions to be made about the selection of the preprocessing types in various research fields, including neural network research.

1. Introduction

Recently, attempts have been made to increase work efficiency through studies using similarities between sentences. Examples include document classification, plagiarism detection, document summarization, paraphrasing, and automatic question-and-answer systems using a similarity measurement model between sentences [1].

Studying the similarity between sentences requires a deep understanding of the semantic and structural information of the language. Previous studies have extracted and used features in sentences [2], but the feature-extraction process was complicated, and the performance was irregular depending on the extracted features. Therefore, attempts have been made to learn a language model that computes probability distributions without extracting features. The linguistic model, which was based on statistical theory, used the conditional probability of a single word (unigram) or a sequence of multiple words (n-gram). In addition, a method has been proposed that combines a word-embedding method, in which information about the meaning or structure of a word is expressed in terms of a real-time multidimensional vector and a deep belief network structure that uses a prelearning method [3].

To improve the prediction accuracy of a high-performance neural-network-based sentence model or a natural-language-based study, confidence in the data should be the highest priority. The dataset of the public database is already purified. Data for research studies should be processed through a filtering step, in which the researcher himself conducts the preprocessing. Therefore, it is necessary to investigate the data preprocessing features that should be selected for machine learning [4] as well as the effects of various preprocessing tasks on the performance of classification models [5–7].

Data integration, refinement, reduction, discretization, feature selection, and data conversion can be used for data preprocessing. Recently, studies have been conducted in which various evaluation methods or preprocessing steps are performed automatically to select appropriate data features

[8]. However, most studies on machine learning to date do not include data preprocessing [9–14]. Furthermore, even in the studies where preprocessing was mentioned, only some parts of various processes, such as word normalization and elimination, were presented [15, 16].

The purpose of this study is to analyze the effect of text data preprocessing on the sentence model. If previous studies were aimed at improving the performance of the model through preprocessing, this study focuses on the effect of combinations of data preprocessing types on performance. Various preprocessing methods were compared and analyzed, such as the setting method for using the preprocessing technique or performance analysis where learning is performed according to the order of complexity of sentences.

The Materials and Methods section describes the different types of text data preprocessing. It also describes the research methods used and explains the sentence model, preprocessing type, and dataset. In the Results and Discussion section, the results of using combinations of different preprocessing types are discussed, and finally, the conclusions of this study are presented in the last section.

2. Materials and Methods

2.1. Text Data Preprocessing. The quality of the data plays an important role in the performance of the algorithm. If data are not preprocessed, the algorithm may behave unexpectedly due to inconsistent data, and performance may be affected.

Existing data preprocessing studies have been mainly conducted in the field of data mining. There have been studies that process web data to format them into an analytical form. These studies did not explain the effect of data preprocessing on the algorithm as a method included in the process of preparing data for analysis [17–19]. There is also a study that analyzed the effect of data preprocessing on predictive ability, limited to numerical data in neural network models [4, 20]. Feature selection, outlier data removal, dimension reduction, etc., were conducted; however, it is difficult to understand their effects on text data.

The sentence model uses word-based text data that include plural words, special characters, and numerals. Therefore, preprocessing for analysis is divided into transformation, in which the original form is transformed to a word-based form, and elimination, in which the words that are considered unnecessary for semantic interpretation are eliminated. The text preprocessing technique is shown in Table 1.

There are three types of normalization: lowering, which converts uppercase letters to lowercase letters, stemming, and lemmatization.

Stemming, which is a normalization technique that reduces the complexity of data, removes affixes and separates stems from words with modified word forms. Lemmatization is a technique that converts words used in various forms into dictionary forms [21, 22]. Table 2 compares stemming and lemmatization techniques.

In stemming, words with different roots are mapped to the same stem. Therefore, it is mainly used in search engines. Lemmatization extracts the original form of a word as the word is converted to a basic form. Therefore, lemmatization does not change the meaning of words [23–25].

As an example of punctuation, the method of removing "-" from the word "brute-force" and obtaining the two words "brute" and "force" is called splitting. Furthermore, if you get the word "bruteforce," it is called merging. Splitting is the same as tokenization, which divides sentences into words.

Elimination assumes that all words that make up a sentence, paragraph, or document do not have the same significance. In other words, according to this method, a word with a low frequency of occurrence or a word with a high frequency of occurrence in a document but with low semantic information, such as a stop word, a one-syllable character or a special character, is deleted.

2.2. Research Methods. This study was conducted to analyze the effects of data preprocessing on sentence models and not to examine fine-tuning or performance improvement. This section describes the setting of the preprocessing study, structure of the sentence model, and datasets used in the study. The procedure used in this study is shown in Figure 1.

This study, which aims to analyze the performance of combining types of text data preprocessing in a sentence model, can be divided into a typical preprocessing type and a preprocessing type that is developed according to the needs of the study. In the typical data preprocessing step, lowering, lemmatization, punctuation splitting and merging, and special character elimination were used by considering the preservation and accuracy of the part-of-speech information. However, splitting and merging, which transform based on punctuation, were also used.

This study considered that preprocessing steps such as normalization and punctuation could semantically damage the meaning of a sentence by making modifications. For example, technical terms can be important in a paragraph or sentence and can help readers to understand the meaning. Because terminology can consist of a single word or multiple words, the segmentation of all words may not reflect the essential meaning of the terminology. To analyze the method of using technical terms in the sentence model, this study developed a module that can identify technical terms composed of complex words and process them as multiple single words. In addition, an entropy-based sorting module was developed to check the effect of sentence complexity on accuracy.

This study applied various combinations of typical preprocessing types, and the techniques developed in this study were analyzed separately. Table 3 shows the preprocessing types used in the analysis.

The accuracy of the model was measured five times for each of the 25 preprocessing types, for a total of 125 measurements.

2.3. Preprocessing Techniques. The preprocessing techniques developed in this study are the sorting module, which sorts

Table 1: Text preprocessing technique.

Technique			Feature
Normalization	Lowering	Conversion to lowercase	Pros: search accuracy can be improved Cons: proper nouns composed of capital letters can be incorrectly classified as general nouns
	Stemming	Conversion to stems	Pros: time efficiency can be improved by reducing the size of the text Cons: dilution of meaning can affect accuracy
	Lemmatization	Conversion to headwords	Pros: part-of-speech information is converted into a preserved form, and search accuracy can be improved Cons: conversion time is long
Punctuation	Splitting	Word splitting	Pros: meaning can be preserved Cons: different rules should be applied depending on the purpose, and the rules are complicated
	Merging	Word merging	

Table 2: Stemming vs. lemmatization.

Word	Stemming	Lemmatization
Innovation	Innovat	Innovation
Innovations	Innovat	Innovation
Innovate	Innovat	Innovate
Innovates	Innovat	Innovate
Innovative	Innovat	Innovative

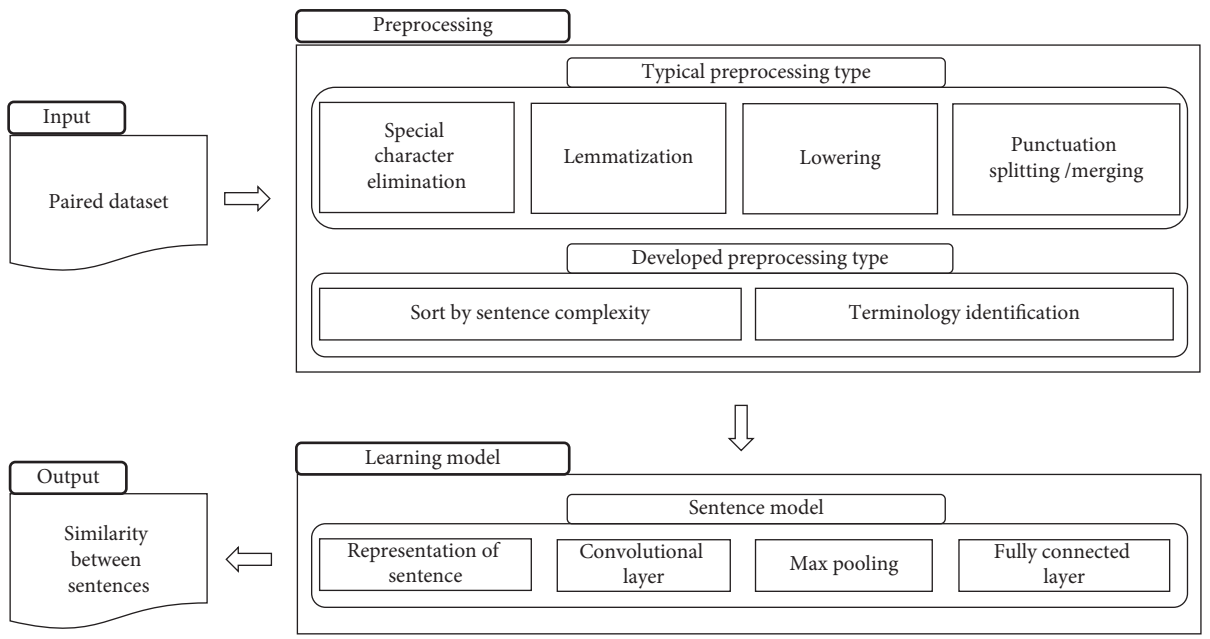


Figure 1: Structure of the similarity measurement model between sentences.

Table 3: Preprocessing types.

Type	Number
Preprocessing is not applied	1
Only one type of preprocessing is applied (a combination of traditional techniques + two proposed techniques)	8
Two types of preprocessing are combined	9
Three types of preprocessing are combined	5
Four types of preprocessing are combined	2
Total	25

the sentences according to an order of complexity of sentences, and the terminology-identification module. The details are described as follows.

2.3.1. Complexity Sorting Module. The characteristics of the data used in machine learning are an important factor in determining the efficiency of learning [26]. Documents are composed of various sentences, and each sentence has a different length, parts of speech, and complexity. Determining entropy for information complexity is a general-purpose technology that is commonly used for signal or video compression. The entropy of a sentence is calculated according to the distribution of syllables, and the calculated entropy is used to define the sentence complexity. This study also analyzed the relationship between the complexity, which is a characteristic of a sentence, and the accuracy of the model. In other words, based on the entropy, a sorting module that classified sentences according to their complexity was developed and confirmed.

The process of model development progressed as follows. Information entropy is an expected value (average) of information in data (as explained below), and when the expected value is high, it can be expressed as "much information." In other words, "much information" in a sentence indicates that the sentence is complicated on the surface. For a random variable for an event, $P(X)$, the information entropy, $H(X)$, is defined as follows:

$$H(p) = -\sum_{x \in X} p(x) \log p(x). \quad (1)$$

The operation algorithm of the complexity sorting module is shown in Figure 2.

In Figure 2, D^m indicates a set of sentences in the corpus and counts each sentence read from D^m according to the ASCII code value. In other words, it calculates the number of ASCII codes in a sentence. Then, the entropy is calculated based on the ASCII code value of a sentence. Finally, it returns the sentences sorted in the ascending or descending order based on the calculated entropy.

2.3.2. Module to Identify Technical Terms Composed of Complex Nouns. Technical term identification is a time-consuming and costly task that can be divided into statistical and rule-based methods. Statistical methods can have high portability because they are not affected by domain restrictions [27]. However, the low accuracy of the identified terms and the inclusion of noise pose difficulties in semantic interpretation. The rule-based method analyzes many terms and processes them through morphemes such as prefixes and suffixes. Although this method can have a low portability because the rules are manually defined and supplemented for each specific field, the accuracy of the identified terms can be high.

In this study, an algorithm to apply the rule-based method and achieve high accuracy was developed. To extract the rules, 1,540 morphemes in the technical term corpus published by the Japan Information Processing Society in 2018 were analyzed. The analysis found that the number of parts-of-speech among the technical terms was 82, and these were composed of single words or combinations of morphemes. The technical term identification algorithm we developed is shown in Figure 3.

First, morphemes are analyzed for each word in the sentence. For analysis, NLTK's pos_tag module was used. Second, technical terms were identified from the learning data using the extracted rules. From the most common composition to the least common composition of the part of speech, there were 485 singular nouns, 396 adjectives + singular nouns, and 257 singular nouns + singular nouns, etc. Third, a search engine (Wikipedia API) was used to verify the identified technical terms. When a search result for a technical term exists in the search engine, the term is converted into an identified sentence. For example, in the sentence "People in a car_race," "car race" is identified to convert it into the sentence "People in a car_race." Fourth, the sentence in which technical terms have been processed is newly stored in Transformed_D. Learning of the sentence model was conducted using the dataset in Transformed_D.

2.4. Sentence Model. The model used to measure the similarity between sentences consists of an encoder/decoder method, which can process two sentences. Siamese networks include two identical subnetwork components to handle each of the two inputs. In other words, Siamese networks can be used as a method to measure the similarity between two sentences. This section describes the structure and performance of the Siamese networks' convolutional neural network- (CNN-) based sentence model for the study.

2.4.1. Structure of the Sentence Model. To measure the similarity between sentences, the CNN model was implemented based on Siamese networks [9]. The model developed by Kim et al. [9] for emotion analysis and question classification is a CNN structure using one layer, but the model proposed in this study is composed of two layers. In addition, hyperparameters for filter size and feature map size were properly tuned. The proposed model is shown in Figure 4.

In Figure 4, n is the number of words in the sentence, k is the dimension of the word vector, and h is the filter window size. Based on the CNN, sentences $X(i)$ and $X(j)$ are sequentially processed through the convolutional, pooling, and fully connected layers, and feature vectors are produced as outputs. For hyperparameters, the filter sizes were set to 2, 3, and 4, and they were set to have 50 feature maps. The dropout was set to 0.5. To compare two sentences, $X(i)$ and $X(j)$, the distance can be expressed as follows, by outputting a feature vector that is the encoding result of the same neural network structure.

If $x(i)$ and $x(j)$ are semantically similar, $\| f(X(i)) - f(X(j)) \|^2$ is small.

If $x(i)$ and $x(j)$ are semantically different, $\| f(X(i)) - f(X(j)) \|^2$ is large.

If the characteristics of the two sentences are well expressed, the distance between the vectors is small. Otherwise, the distance is large. In this study, the Manhattan

```
Input : D^m = {Sen_[0] , ..., Sen_[m]} # Set of sentences
Initialization : E^m = {0, ..., Sen_[m]} # Set of entropy complexity values
Output : Sorted_D^m = {Sen_[0] , ..., Sen_[m]} # Set of sentences sorted by entropy complexity

for i = 0, ..., N - m do
  Size = len (Sen_[i]) * 1.0
  for j = 0, ..., 128 do
     result + = Sen_i.count (chr (j)) / size * log (Sen_[i].count (chr (j)) / size, 2)
  E(i) = result * -1.0

E = dictionary (E)
Sorted_D = sort (D, E)

return Sorted_D
```

Figure 2: Complexity sorting algorithm.

```
Input : D^m = {Sen_[0] , ..., Sen_[m]} # Set of sentences
Initialization : Tagged_Sen^m = {{0, ... Sen_[0,n]}, ..., {0, ... Sen_[m,n]}} # Set of entropy
complexity values
Output : Transformed_D^m = {TSen_[0] , ..., TSen_[m]} # Set of sentences sorted by entropy
complexity
for i = 0, ..., N - m do
  Tagged_Sen_i = pos_tag (Sen_[i])
  Sentence = ‘’
  for j = 0, ..., len (Sen_[i,j]) do
    Terminology = Rule_based_Identifier (Sen_[I,j], Size_of_Terminology)
    Exists = Search_engine_API (Terminology)
    if Exists is True
      Sentence = Sentence + Terminology
      j = j + Size_of_Terminology
    else if Exists is not True
      Sentence = Sentence + Sen_I,j

  Transformed_D_i.Append (Sentence)

return Transformed_D
```

No	The morpheme of the technical term	N
1	Singular noun	485
2	Adjective + singular noun	396
3	Singular noun + singular noun	257
4	Adjective	75
5	Adjective + singular noun + singular noun	53
6	Present participle	30
7	Present participle + singular noun	23
8	Singular noun + singular noun + singular noun	23
9	Plural noun + singular noun	17
10	Adjective + plural noun	16
11	Past participle + singular noun	11
12	Past tense verb + singular noun	11
13	Singular noun + preposition + singular noun	11

Figure 3: Technical term identification algorithm.

distance, which is a similarity function and has excellent performance, was applied (Jonas Mueller, 2016). This study was developed using Python 3.6.8 in the Linux 16.04 operating system environment.

2.4.2. Comparison of the Performance of Sentence Models. In the present study, a CNN-based model (Figure 4), Siamese LSTM model [28], and transformer model [29] were implemented for model selection. For the learning

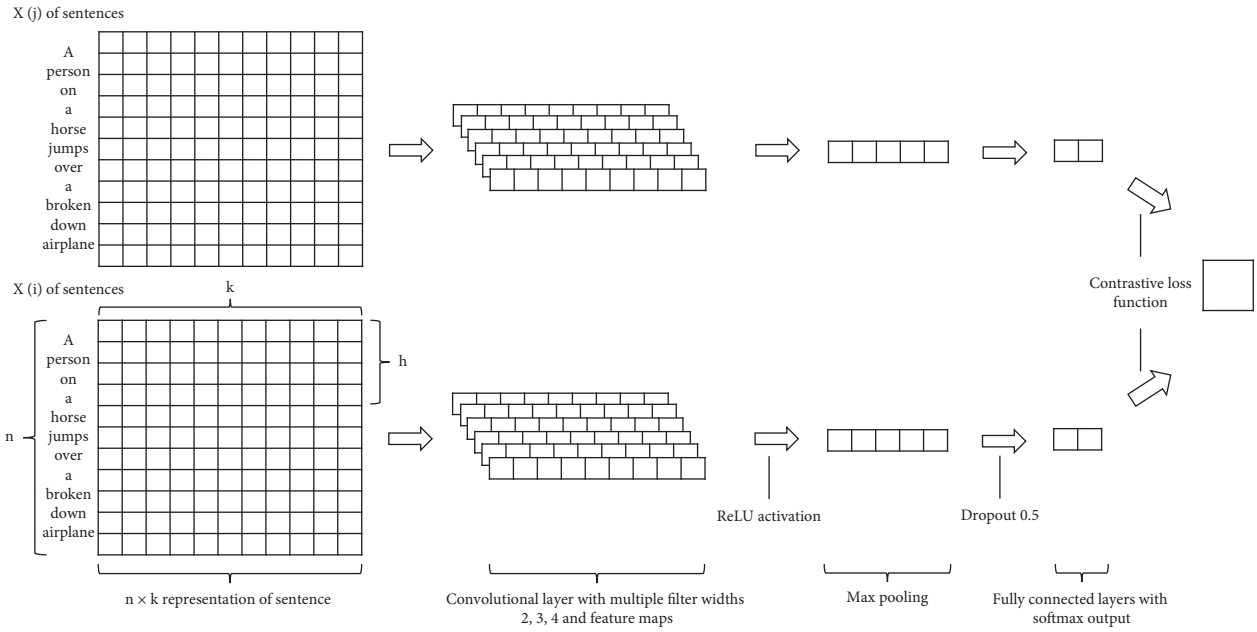


Figure 4: Siamese network based on the CNN model.

parameters of each model, the epoch was set to 10, batch size to 512, and learning rate to 0.001, and learning time and accuracy were as shown in Table 4.

To evaluate how the performance was affected using combinations of various types of preprocessing, a study was conducted using a CNN-based model, which had low accuracy but the fastest learning time.

Table 4: Learning time and accuracy by the model.

Architectures	Learning time	Accuracy
Siamese LSTM model	2 : 14 : 02	79.66
Transformer model	1 : 54 : 52	79.61
CNN-based model	25 : 48	77.67

2.5. Dataset. For model training, the Stanford Natural Language Inference (SNLI) corpus was used. The SNLI corpus is a dataset that expresses the logical relationship between two sentences, and it is used in research for various inferences [30, 31]. The corpus consists of 367,369 instances, of which 257,158 (70%) are training sets, 36,737 (10%) are validation sets, and 73,474 (20%) are testing sets. The structure of the SNLI corpus is shown in Table 5.

This structure includes two sentences and a corresponding label that is based on the similarity between the sentences. For example, "A person on a horse jumps over a broken down airplane" and "A person is at a diner, ordering an omelette" in the first line are not semantically/logically equivalent, and therefore the label is 0 (False).

3. Results and Discussion

Table 6 shows the results when the models with different combinations of preprocessing types were sorted by accuracy.

The variance of the average value of the results by the preprocessing type ranged from a minimum of 0.05 to a maximum of 0.34. Based on analysis of the results, we determined that a combination of two preprocessing techniques showed good performance. If only two preprocessing techniques can be used, lemmatization and punctuation splitting (No. 1) are good candidates. This combination showed the highest accuracy with 79.09%, which is 0.73% higher than the accuracy without preprocessing (No. 21). Furthermore, it was found that the use of the normalization techniques, lemmatization and lowering (No. 2) together or the use of lowering and punctuation splitting or merging (Nos. 3 and 4) also increased accuracy. It should be noted that, in any of these combinations, lemmatization, lowering, and punctuation splitting were used. Lemmatization and lowering are techniques that can improve accuracy by normalizing different words.

If only one technique was used (Nos. 9, 10, 11, 15, and 18), then the accuracy was higher than that without preprocessing. Among these, the accuracies associated with lemmatization (No. 9), lowering (No. 10), and punctuation splitting (No. 11) were similar and ranged from 78.876% to 78.842%. This implies that splitting a word into two words based on punctuation, such as during normalization or punctuation splitting, can extend the length of a sentence and have a positive effect on accuracy.

Characteristic features include the use of special character elimination and punctuation merging. When special character elimination was combined with lemmatization and lowering, the accuracy was increased (Nos. 5, 7, and 8). However, if lemmatization or lowering was used separately, the accuracy decreased (Nos. 12, 13, 16, 19, 20, and 22). In addition, except for the case where it was combined with lowering (No. 3), the accuracy decreased when punctuation merging was used. This contrasts with the fact that

Table 5: Structure of the SNLI corpus.

Premise	Hypothesis	Label
A person on a horse jumps over a broken down airplane	A person is at a diner, ordering an omelette	0
A person on a horse jumps over a broken down airplane	A person is outdoor on a horse	1
Children smiling and waving at the camera	There are children present	1
Children smiling and waving at the camera	The kids are frowning	0
A boy is jumping on the skateboard in the middle of a red bridge	The boy skates down the sidewalk	0
A boy is jumping on the skateboard in the middle of a red bridge	The boy does a skateboarding trick	1
Two blond women are hugging one another	The women are sleeping	0
Two blond women are hugging one another	There are women showing affection	1

Table 6: Accuracy based on the preprocessing type.

	No.	Type	Accuracy					
			1st	2nd	3rd	4th	5th	Mean (SD.)
Traditional method	1	[2] + [4]	79.17	79.01	79.31	78.93	79.03	79.090 (0.15)
	2	[2] + [3]	79	79.14	79.25	78.73	79.09	79.042 (0.20)
	3	[3] + [5]	79.15	78.98	79.17	78.89	78.97	79.032 (0.12)
	4	[3] + [4]	78.96	79.12	78.82	79.13	79.09	79.024 (0.13)
	5	[1] + [2] + [3] + [4]	79	78.96	79.12	78.93	79.01	79.004 (0.07)
	6	[1] + [4]	78.94	78.59	79.33	79.15	78.97	78.996 (0.28)
	7	[1] + [2] + [3] + [5]	78.8	78.86	79.09	78.79	78.98	78.904 (0.13)
	8	[1] + [2] + [3]	78.54	79.22	78.93	78.95	78.87	78.902 (0.24)
	9	[3]	79.02	78.97	78.75	78.73	78.91	78.876 (0.13)
	10	[4]	78.94	79	78.62	78.77	78.89	78.844 (0.15)
	11	[2]	78.83	78.78	78.83	78.87	78.9	78.842 (0.05)
	12	[1] + [3] + [4]	78.44	78.94	78.8	79.15	78.85	78.836 (0.26)
	13	[1] + [3] + [5]	78.41	78.8	78.67	79.34	78.8	78.804 (0.34)
	14	[2] + [5]	78.79	78.77	78.85	78.68	78.78	78.774 (0.06)
	15	[5]	78.34	78.67	78.86	78.94	78.7	78.702 (0.23)
	16	[1] + [2] + [5]	78.78	78.25	78.88	78.75	78.67	78.666 (0.24)
	17	[1] + [5]	78.51	78.38	78.72	78.96	78.59	78.632 (0.22)
	18	[1]	78.7	78.28	78.46	78.51	78.51	78.492 (0.15)
	19	[1] + [2]	78.22	78.61	78.27	78.56	78.36	78.404 (0.17)
	20	[1] + [2] + [4]	78.26	78.45	78.35	78.17	78.31	78.308 (0.10)
	21	[0]	78.42	78.36	78.28	78.11	78.34	78.302 (0.12)
	22	[1] + [3]	78.24	78.16	78.07	78.36	78.3	78.226 (0.11)
Developed method	23	[7-2]	78.35	78.68	79.1	78.65	78.34	78.624 (0.31)
	24	[7-1]	78.39	78.15	78.19	78.37	78.33	78.286 (0.11)
	25	[6]	73	72.96	72.43	72.59	72.8	72.756 (0.24)

[0]: no applied preprocessing; [1]: special character elimination; [2]: lemmatization; [3]: lowering; [4]: punctuation splitting; [5]: punctuation merging; [6]: essential terminology preprocessing; [7-1]: ascending order based on entropy complexity; [7-2]: descending order based on entropy complexity.

punctuation splitting leads to high accuracy. The results are similar to those obtained for the technical term identification preprocessing technique that was developed in this study for comparison.

In No. 25, technical terms composed of complex nouns were recognized as one word for learning, and the associated accuracy was 72.76%. The accuracy was lower than that without preprocessing by 5.54%, and it was also relatively low compared with other preprocessing combinations. As suggested by the algorithm, there are many technical terms that contain more than two words, such as in the forms (adjective-singular noun), (singular noun-singular noun), and (adjective-singular noun-singular noun). As a result, the meanings of these terms cannot be correctly interpreted. In other words, because the technical terms consisting of two or more segments were processed as one word and the length of the sentence was shortened, the accuracy decreased.

When sentences were sorted according to their entropy complexity (Nos. 23 and 24), the accuracies were 78.624% and 78.286% for the descending and ascending order, respectively. This represents a difference of +0.322% and −0.016% compared with data that were not preprocessed. Therefore, ordering sentences according to their complexity may not affect the accuracy.

4. Conclusions

In neural network research, data, algorithms, and parallel hardware are essential elements. Even with good algorithms and high-performance hardware, studies cannot be conducted

if the quality of data is low or no data are available. Despite its importance, many existing neural network studies do not provide any information about data preprocessing.

This study analyzed the effect of preprocessing through text data preprocessing of sentence models. To this end, experiments were conducted to evaluate combinations of typical data preprocessing types. Furthermore, the effects of two new techniques on the accuracy of the model were analyzed: preprocessing of technical terms composed of compound words and determining the learning order based on data complexity.

Based on the results of this study, the following conclusions can be drawn. First, when only two preprocessing techniques are used, we recommended using lemmatization and punctuation splitting, lemmatization and lowering, or lowering and punctuation splitting. Second, when only one preprocessing technique is used, it is better to use lemmatization, lowering, or punctuation splitting. Third, to improve accuracy, it is generally not recommended to use a preprocessing type that shortens the lengths of sentences. Fourth, the use of special character elimination and normalization techniques does not contribute to improving the accuracy. Fifth, setting the learning order according to sentence complexity does not contribute to improving the accuracy.

Building predictive or sentence models from refined data can help improve the performance of the model. The accuracy of the preprocessing of text data in this study suggested a certain combination of preprocessing types could improve performance when various models are established. Consequently, this study is significant in that it allows better decision-making about which preprocessing type should be selected according to the purpose of the study or the type of the construction model.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (no. 2019R1H1A1079885).

References

[1] D.-K. Lee, K.-J. Oh, and H.-J. Choi, "Measuring the syntactic similarity between Korean sentences using RNN," *The Korean Institute of Information Scientists and Engineers*, vol. 6, pp. 792–794, 2016.

[2] Y. Bengio, *Neural Probabilistic Language Models," Innovations in Machine Learning*, Springer, Berlin, Germany, 2006.

[3] T. Mikolov, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 3111–3119, Lake Tahoe, NV, USA, December 2013.

[4] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

[5] S. F. Crone, S. Lessmann, and R. Stahlbock, "The impact of preprocessing on data mining: an evaluation of classifier sensitivity in direct marketing," *European Journal of Operational Research*, vol. 173, no. 3, pp. 781–800, 2006.

[6] C. A. Gonçalves, R. Camacho, and E. C. Oliveira, "The impact of pre-processing on the classification of MEDLINE documents," pattern recognition in information systems," in *Proceedings of the 10th International Workshop on Pattern Recognition in Information Systems, PRIS 2010*, Funchal, Madeira, Portugal, June 2010.

[7] A. K. Uysal and S. Gunal, "The impact of preprocessing on text classification," *Information Processing & Management*, vol. 50, no. 1, pp. 104–112, 2014.

[8] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: a new perspective," *Neurocomputing*, vol. 300, pp. 70–79, 2018.

[9] Y. Kim, "Convolutional Neural Networks for Sentence Classification," 2014, http://arxiv.org/abs/1408.5882.

[10] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proceedings. Of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, September 2015.

[11] T. Lei, R. Barzilay, and T. Jaakkola, "Molding cnns for text: non-linear, non-consecutive convolutions," 2015, http://arxiv.org/abs/1508.04112.

[12] H. Chen, "Neural sentiment classification with user and product attention," *Proceedings. Of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, TX, USA, November 2016.

[13] Y. Xiao and K. Cho, "Efficient character level document classification by combining convolution and recurrent layers," 2016, http://arxiv.org/abs/1602.00367.

[14] K. Kowsari, "Hdltex: hierarchical deep learning for text classification," 2017, http://arxiv.org/abs/1709.08267.

[15] Z. Yang, "Hierarchical attention networks for document classification," in *Proceedings. Of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*, Human Language Technologies, San Diego, CA, USA, June 2016.

[16] S. Lai, "Recurrent convolutional neural networks for text classification," *AAAI*, vol. 333, 2015.

[17] T. Kuzar and P. Navrat, "Preprocessing of slovak blog articles for clustering," in *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Toronto, ON, USA, September 2010.

[18] J.-S. Lee, "A study on the data mining preprocessing tool for efficient database marketing," *Journal of Digital Convergence*, vol. 12, no. 11, pp. 257–264, 2014.

[19] S. K. Dwivedi and B. Rawat, "A review paper on data preprocessing: a critical phase in web usage mining process," in *Proceedings of the 2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*, Greater Noida, Delhi, India, October 2015.

[20] H.-S. Shin, Y. Jin, and C.-S. Park, "Influence of data preprocessing on a machine learning model," *Architectural Institute of Korea*, vol. 37, no. 1, pp. 491-492, 2017.

[21] M. Kasthuri and Dr. S. Britto Ramesh Kumar, "A framework for language independent stemmer using dynamic programming," *International Journal of Applied Engineering Research*, vol. 10, pp. 39000–39004, 2013.

[22] K. Abainia, S. Ouamour, and H. Sayoud, "A novel robust Arabic light stemmer," *Journal of Experimental & Theoretical Artificial Intelligence*, no. 1–17, 2016.

[23] P. Han, S. Shen, D. Wang, and Y. Liu, "The influence of word normalization in english document clustering," in *Computer Science and Automation Engineering (CSAE), 2012 IEEE International Conference*, Zhangjiajie, China, May 2012.

[24] M. Attia and J. Van Genabith, "A jellyfish dictionary for Arabic," in *Proceedings of the Electronic lexicography in the 21st century: thinking outside the paper: proceedings of the eLex 2013 conference*, Tallinn, Estonia, Europe, October 2013.

[25] R. J. Prathibha and M. C. Padma, "Design of rule based lemmatizer for kannada inflectional words," in *Proceedings of the Emerging Research in Electronics, Computer Science and Technology (ICERECT), 2015 International Conference*, Mandya, India, December 2015.

[26] S. Samsani, "An RST based efficient preprocessing technique for handling inconsistent data," in *Proceedings of the 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pp. 1–8, Chennai, India, December 2016.

[27] P. Pantel and D. Lin, "A statistical corpus-based term extractor," in *Proceedings of the Biennial Conference of the Canadian Society on Computational Studies of Intelligence*, pp. 36–46, Ottawa, Canada, June 2001.

[28] J. Mueller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity," in *Proceeding AAAI'16 Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 2786–2792, Quebec, Canada, May 2000.

[29] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need,," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 6000–6010, Cambridge, MA, USA, December 2015.

[30] A. Talman and S. Chatzikyriakidis, "Testing the generalization power of neural network models across NLI benchmarks," in *Proceedings of the Second BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, Brussels, Belgium, November 2018.

[31] A. Conneaum, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from Natural Language inference data," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 670–680, Copenhagen, Denmark, September 2018.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 2637403, 10 pages
https://doi.org/10.1155/2020/2637403

Research Article

Consensus Formation Control and Obstacle Avoidance of Multiagent Systems with Directed Topology

Lichao Wang,[1,2] **Xing Li**,[1,2] **Bingyou Liu**,[1,2] **and Zhengzheng Zhang**[1,2]

[1]*College of Electrical Engineering, Anhui Polytechnic University, Wuhu, Anhui, China*
[2]*Key Laboratory of Advanced Perception and Intelligent Control of High-End Equipment, Ministry of Education, Wuhu, Anhui, China*

Correspondence should be addressed to Bingyou Liu; lby009@mail.ustc.edu.cn and Zhengzheng Zhang; 2180210110@stu.ahpu.edu.cn

Received 13 December 2019; Revised 10 April 2020; Accepted 22 April 2020; Published 13 May 2020

Guest Editor: Sanghyuk Lee

This study addresses the problems of formation control and obstacle avoidance for a class of second-order multiagent systems with directed topology. Formation and velocity control laws are designed to solve the formation tracking problem. A new obstacle avoidance control law is also proposed to avoid obstacles. Then, the consensus control protocol consists of the formation, velocity, and obstacle avoidance control laws. The convergence of the proposed control protocol is analyzed by a redesigned Lyapunov function. Finally, the effectiveness of theoretical results is illustrated by simulation examples. The simulation results show that the formation tracking problem of the given multiagent systems can be realized and obstacles can be avoided under the proposed control protocol.

1. Introduction

The cooperative control problem of multiagent systems has been attracted outstanding attention in the past few years due to its widespread applications in multisensor systems [1], mobile robot systems [2, 3], unmanned aerial vehicle systems [4, 5], power distribution networks [6], and so forth. The basic problem of cooperative control is consensus, in which the objective of consensus is to design an appropriate control protocol, such that the output of all agents can achieve synchronization or track a desired trajectory.

Currently, the consensus control problems of multiagent systems have been extensively addressed in existing papers. Many control protocols have been designed to achieve the consensus control of multiagent systems. In [7], the distributed linear control protocol for the linear multiagent systems with limited interaction ranges was designed. Iterative learning control protocols were proposed in [8–10] to solve the consensus tracking problem of nonlinear multiagent systems. In [11], the consensus problem of nonlinear multiagent systems with directed topology and communication constraints was investigated, in which each agent communicated only with its neighbors. Moreover, in [12], a consensus protocol with the local state information was proposed to solve the event-triggered control problem of general linear multiagent systems. The finite-time consensus tracking control problem of multiagent systems with uncertain nonlinear dynamics and error constraints was investigated in [13], in which the nonsingular fast sliding mode control technique was used.

It is not difficult to see from the abovementioned papers that the research on the consensus control problem of multiagent systems has achieved rich results. However, these papers do not further analyze the formation control problem of multiagent systems. As an important research direction, the formation control problem has played an important role in many fields, such as the formation control of spacecraft [14], multiple aerial vehicles [15], multiple quadrotors [16], and mobile robots [17–19]. The formation control of multiagent systems has been discussed in some studies. In [20], the optimal formation problem of first-order multiagent systems with fixed communication topology was considered.

In [21], the formation tracking problem with distributed observer was addressed, in which the distributed formation tracking control protocol was constructed. The control protocol with communication time-varying delay was presented in [22]. Furthermore, the formation control strategy with position estimation [23] and the distributed formation iterative learning control protocol [24] were also addressed. Meanwhile, the consensus control protocols for the multiagent systems were proposed to solve the problem of collision avoidance [25–27]. It should be pointed out that the results on the formation control or the control of collision avoidance are discussed separately. To the best of the authors' knowledge, however, it should be paid attention to prevent the collision with obstacles while solving the formation problem of multiagent systems. However, the problem has received minimal attention in the existing literature.

Inspired by the abovementioned facts, this study investigates the formation control and obstacle avoidance for a class of second-order multiagent systems with directed topology. The main contributions of this work are as follows: (i) the formation control law and velocity control law are designed to solve the formation tracking problem of given multiagent systems with directed topology. Furthermore, a new obstacle avoidance strategy is proposed to guarantee that all agents avoid obstacles. (ii) By comparing with the control protocols proposed in [24–26], the current consensus control protocol consists of the designed formation control law, velocity consensus control law, and obstacle avoidance control law. The purpose of this is to solve the formation control and obstacle avoidance problems of the given multiagent systems at the same time. (iii) To prove the convergence of the proposed control protocol, a new Lyapunov function is structured in this paper. Finally, two simulation examples are provided to illustrate the effectiveness of the proposed control protocol.

The remainder of this paper is organized as follows. Graph theory is introduced in Section 2, and the problem formulation is given in Section 3. In Section 4, the control protocol design and convergence analysis are discussed. Next, the simulation examples are provided to illustrate the effectiveness of theoretical analysis in Section 4. Finally, conclusions are drawn in Section 5.

2. Graph Theory

Let $\mathcal{G} = (\mathcal{V}, \mathrm{E}, \mathcal{A})$ denote a directed graph with a set of nodes $\mathcal{V} = \{v_1, \ldots, v_n\}$ and a set of directed edges $\mathcal{E} = \{(i, j), i, j \in \mathcal{V}, \text{ and } i \neq j\}$. The weighted adjacency matrix is $\mathcal{A} = [a_{ij}] \in \mathbb{R}^{n\times n}$, where $a_{ij} > 0$ if and only if $(i, j) \in \mathcal{E}$; otherwise, $a_{ij} = 0$. Agent j is called the neighbor of i if agent i receives the information from agent j. The set of neighbors of agent i is defined as $\mathcal{N}_i = \{v_j : (v_j, v_i) \in \mathcal{E}\}$. The Laplacian matrix $\mathcal{L}$ is denoted by $\mathcal{L} = \mathcal{D} - \mathcal{A}$, where $\mathcal{D} = \mathrm{diag}\{d_1, \ldots, d_n\}$ with $d_i = \sum_{j\in\mathcal{N}_i} a_{ij}$. The graph $\mathcal{G}$ is connected if there exists a path between any two vertices.

In this paper, the multiagent systems with n agents are considered. Hence, the exchange information among agents can be modeled as the directed graph n with n nodes. According to the related knowledge of the graph theory $\mathcal{G}$, we can theoretically analyze the control problem of multiagent systems. In addition, in order to achieve the desired formation shape, the distance between agents should be set. Hence, in this paper, let the matrix h is defined as the desired formation shape of given multiagent systems. Here, the matrix $h = [h_1, \ldots, h_n]$ and $h_i = [h_{i1}, \ldots, h_{in}]^{\mathrm{T}}$ with h_{ij} being the desired distance between agent i and agent j.

3. Problem Formulation

In this paper, a class of second-order multiagent systems with n agents is studied, and the ith agent's dynamics are described as

$$\begin{cases} \dot{x}_i(t) = v_i(t), \\ \dot{v}_i(t) = u_i(t), \end{cases} \tag{1}$$

where $x_i(t) \in \mathbb{R}$, $v_i(t) \in \mathbb{R}$, and $u_i(t) \in \mathbb{R}$ $(i = 1, \ldots, n)$ are the position, velocity, and control input of agent i, respectively.

To facilitate the following discussion, the time variable t will be ignored if there is no ambiguity. In addition, some definitions and lemmas are given as follows.

Definition 1. Function $\Theta(x_{ij})$ for x_{ij} is a nonnegative function if the following properties are satisfied at the same time:

(1) $\Theta(x_{ij}) \longrightarrow \infty$ for $x_{ij} \longrightarrow 0$
(2) $\Theta(x_{ij}) \longrightarrow \infty$ for $x_{ij} \longrightarrow \infty$
(3) $\Theta(x_{ij}) \longrightarrow 0$ for $x_{ij} \longrightarrow h_{ij}$

where x_{ij} is the Euclidean distance from agent i to j, i.e., $x_{ij} = \|x_i - x_j\|$.

Lemma 1. *For a given multiagent system, let $\nabla_{x_i}\sigma(x_{ij})$ be the gradient function of a continuous differentiable function $\sigma(x_{ij})$; then, the following property is held:*

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\nabla_{x_i}\sigma\left(x_{ij}\right)\left(\dot{x}_i - \dot{x}_j\right) = \sum_{i=1}^{n} v_i \sum_{j=1,j\neq i}^{n}\nabla_{x_i}\sigma\left(x_{ij}\right). \tag{2}$$

Proof. For a given multiagent system, let $z_{ij} = x_i - x_j$; then, one has

$$\dot{z}_{ij} = \dot{x}_i - \dot{x}_j = v_i - v_j. \tag{3}$$

Furthermore, we can obtain

$$\nabla_{x_i}\sigma\left(x_{ij}\right) = \frac{\partial\sigma\left(x_{ij}\right)}{\partial x_i} = \frac{\partial\sigma\left(x_{ij}\right)}{\partial z_{ij}}\frac{\partial z_{ij}}{\partial x_i} = \frac{\partial\sigma\left(x_{ij}\right)}{\partial z_{ij}}, \tag{4}$$

$$\nabla_{x_j}\sigma\left(x_{ij}\right) = \frac{\partial\sigma\left(x_{ij}\right)}{\partial x_j} = \frac{\partial\sigma\left(x_{ij}\right)}{\partial z_{ij}}\frac{\partial z_{ij}}{\partial x_j} = -\frac{\partial\sigma\left(x_{ij}\right)}{\partial z_{ij}}. \tag{5}$$

For the given multiagent systems, it is easy to obtain that $x_{ij} = \|x_i - x_j\| = \|x_j - x_j\| = x_{ji}$, where x_{ji} represents the Euclidean distance from agent j to i, and then, we have

$\sigma(x_{ij}) = \sigma(x_{ij})$. Hence, on the basis of equations (4) and (5), one obtains

$$\nabla_{x_i}\sigma\left(x_{ij}\right) = -\nabla_{x_j}\sigma\left(x_{ij}\right). \tag{6}$$

Considering equation (6), we obtain

$$\begin{aligned}
\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\nabla_{x_i}\sigma\left(x_{ij}\right)\left(\dot{x}_i-\dot{x}_j\right) &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\nabla_{x_i}\sigma\left(x_{ij}\right)\dot{x}_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\nabla_{x_i}\sigma\left(x_{ij}\right)\dot{x}_j \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\nabla_{x_i}\sigma\left(x_{ij}\right)\dot{x}_i - \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1,i\neq j}^{n}\nabla_{x_j}\sigma\left(x_{ij}\right)\dot{x}_i \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\nabla_{x_i}\sigma\left(x_{ij}\right)\dot{x}_i + \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1,i\neq j}^{n}\nabla_{x_i}\sigma\left(x_{ij}\right)\dot{x}_i \\
&= \left(\frac{1}{2}\sum_{i=1}^{n}\sum_{i=1,i\neq j}^{n}\dot{x}_i + \frac{1}{2}\sum_{j=1}^{n}\sum_{j=1,j\neq i}^{n}\dot{x}_i\right)\nabla_{x_i}\sigma\left(x_{ij}\right) = \sum_{i=1}^{n}\sum_{i=1,i\neq j}^{n}\nabla_{x_i}\sigma\left(x_{ij}\right)\dot{x}_i,
\end{aligned} \tag{7}$$

and it is obtained that

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\nabla_{x_i}\sigma\left(x_{ij}\right)\left(\dot{x}_i-\dot{x}_j\right) = \sum_{i=1}^{n}v_i\sum_{j=1,j\neq i}^{n}\nabla_{x_i}\sigma\left(x_{ij}\right). \tag{8}$$

The proof is completed.

The control objective of this study is to design a suitable control protocol $u_i(t)$ $(i = 1, \ldots, n)$, such that the output of all agents can achieve desired formation shape without obstacles, that is, $\lim_{t\longrightarrow\infty}(x_i(t) - x_j(t)) = h_{ij}$ and $\lim_{t\longrightarrow\infty}v_i(t) = v_j(t)$, where $i \neq j \in \{1, \ldots, n\}$ and $h_{ij} \in h$. Meanwhile, it is also guaranteed that the desired formation shape can be maintained after avoiding obstacles. □

4. Control Protocol Design and Convergence Analysis

4.1. Control Protocol Design. On the basis of Definition 1, the nonnegative function $\Theta(x_{ij})$ is designed as

$$\Theta\left(x_{ij}\right) = \alpha\ln\left[\frac{h_{ij}}{2\left\|x_i-x_j\right\|} + \frac{\left\|x_i-x_j\right\|}{2h_{ij}}\right], \quad (i \neq j \in \{1, 2, \ldots, n\}), \tag{9}$$

where $\alpha > 0$ is the formation control coefficient. The curve of function $\Theta(x_{ij})$ is shown in Figure 1.

Figure 1 presents that $\Theta(x_{ij}) \geq 0$ regardless of how to change the distance between agents i and j. That is to say, $\Theta(x_{ij}) \longrightarrow +\infty$ for $\|x_i - x_j\| \longrightarrow 0$ and $\|x_i - x_j\| \longrightarrow +\infty$, and $\Theta(x_{ij}) = 0$ for $\|x_i - x_j\| = h_{ij}$, where $i \neq j \in \{1, \ldots, n\}$. This condition implies that agents i and j can hold the desired distance. Therefore, the formation control law of the given multiagent systems can be designed as

$$u_{ia} = -\sum_{j=1,j\neq i}^{n}\nabla_{x_i}\Theta\left(x_{ij}\right). \tag{10}$$

To maintain the velocity consensus, a velocity control law is designed as

$$u_{ib} = \sum_{j=1,j\neq i}^{n}a_{ij}\Psi\left(v_i - v_j\right), \tag{11}$$

where $\Psi(v_i - v_j)$ is called velocity adjust function and represented as

$$\Psi\left(v_i - v_j\right) = \frac{\beta\left(v_j - v_i\right)}{\sqrt{1+\left(v_i - v_j\right)^2}}, \quad (i \neq j \in \{1, 2, \ldots, n\}), \tag{12}$$

where $\beta > 0$ is the velocity control coefficient. On the basis of equation (12), the function $\Psi(v_i - v_j)$ will be equal to zero as $v_i = v_j$ for $i \neq j \in \{1, \ldots, n\}$.

When obstacles exist in the environment, to prevent the collision with obstacles, a new control strategy $U_{O_{bi}}$ for agent i is regarded as

$$U_{O_{bi}} = \begin{cases} 0, & \left\|x_i - O_b\right\| > R, \\ \gamma\left(\frac{R^2-r^2}{\left\|x_i-O_b\right\|^2-r^2} - 1\right)^2, & r < \left\|x_i - O_b\right\| \leq R, \\ \text{not define}, & \left\|x_i - O_b\right\| \leq r, \end{cases} \tag{13}$$

where $\gamma > 0$ is the avoidance control coefficient; x_i is the position of agent i; O_b, r, and R represent the center, radius, and maximum detection radius of an obstacle, respectively; and $\|x_i - O_b\|$ is the Euclidean distance between agent i and an obstacle.

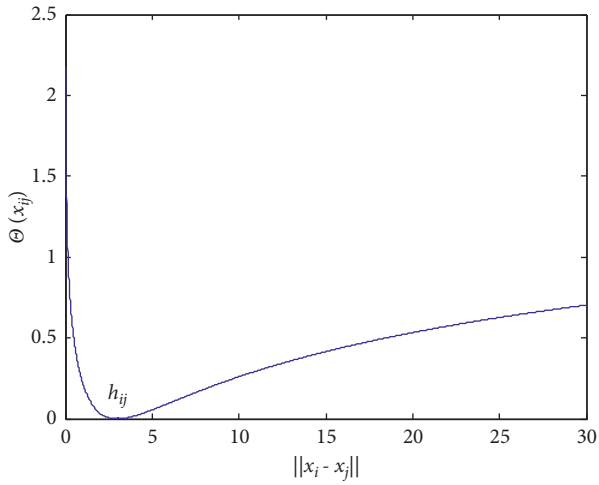


Figure 1: The curve of function $\Theta(x_{ij})$.

By taking gradient of function $U_{O_{bi}}$,

$$\nabla_{x_i} U_{O_{bi}} = \begin{cases} 0, & \|x_i - O_b\| > R, \\ \gamma \dfrac{-(R^2 - r^2)\left(R^2 - \|x_i - O_b\|^2\right)}{\left(\|x_i - O_b\|^2 - r^2\right)^3} (x_i - O_b), & r < \|x_i - O_b\| \le R, \\ \text{not define}, & \|x_i - O_b\| \le r. \end{cases} \tag{14}$$

By checking equations (13) and (14), we have $\lim_{\|x_i - O_b\| \longrightarrow R^-} U_{O_{bi}} = \lim_{\|x_i - O_b\| \longrightarrow R^+} U_{O_{bi}} = 0$ from equation (13), which implies that $U_{O_{bi}}$ is continuous at $\|x_i - O_b\| = R$, and $\lim_{\|x_i - O_b\| \longrightarrow R^-} \nabla_{x_i} U_{O_{bi}} = \lim_{\|x_i - O_b\| \longrightarrow R^+} \nabla_{x_i} U_{O_{bi}} = 0$ from equation (14), which indicates that $\nabla_{x_i} U_{O_{bi}}$ is also continuous at $\|x_i - O_b\| = R$.

On the basis of the above analysis, the new obstacle avoidance control law can be designed as

$$u_{ic} = -\nabla_{x_i} U_{O_{bi}}. \tag{15}$$

Hence, considering the formation control law u_{ia}, velocity control law u_{ib}, and obstacle avoidance control law u_{ic}, the consensus control protocol for agent i can be selected as $u_i = u_{ia} + u_{ib} + u_{ic}$, that is,

$$\begin{aligned} u_i = &-\alpha \sum_{j=1, j\neq i}^{n} \nabla_{x_i} \ln\left[\frac{h_{ij}}{2\|x_i - x_j\|} + \frac{\|x_i - x_j\|}{2h_{ij}}\right] \\ &+ \beta \sum_{j=1, j\neq i}^{n} a_{ij} \frac{(v_j - v_i)}{\sqrt{1 + (v_i - v_j)^2}} - \nabla_{x_i} U_{O_{bi}}. \end{aligned} \tag{16}$$

4.2. Convergence Analysis. The main results of this work are shown in Theorem 1.

Theorem 1. *Consider the multiagent system (1) with the directed communication topology and the consensus control protocol (16), and assume that multiple obstacle surroundings being considered or not; then, all agents can achieve the desired formation shape and maintain the velocity unchanged, i.e.,* $\lim_{t\longrightarrow\infty}(x_i(t) - x_j(t)) = h_{ij}$ *and* $\lim_{t\longrightarrow\infty} v_i(t) = v_j(t)$ *with* $i \neq j \in \{1, \ldots, n\}$ *and* $h_{ij} \in h$. *Meanwhile, the desired formation shape will still be held after avoiding obstacles.*

Proof. The Lyapunov function candidate is considered:

$$V(x, v) = \frac{1}{2}\sum_{i=1}^{n} v_i^2 + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} \Theta\left(x_{ij}\right) + \sum_{i=1}^{n} U_{O_{bi}}. \tag{17}$$

On the basis of equation (17), $(1/2)\sum_{i=1}^{n} v_i^2$ and $(1/2)\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} \Theta(x_{ij})$ are continuously differentiable in x and v. As shown in equation (13), $\sum_{i=1}^{n} U_{O_{bi}}$ is also a differentiable function. Equation (17) presents that $(1/2)\sum_{i=1}^{n} v_i^2 \ge 0$. Figure 1 indicates that the second item in equation (17) $(1/2)\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} \Theta(x_{ij}) \ge 0$ is easily obtained. Furthermore, function $U_{O_{bi}} \ge 0$ is obtained from the description of equation (13). Then, the third item in equation (17) $\sum_{i=1}^{n} U_{O_{bi}} \ge 0$ can be acquired. Equation (17) is hence an effective Lyapunov function, and $V(x, v) \ge 0$.

By taking the derivative of function $V(x, v)$ along with x and v, we have

$$\begin{aligned}\dot{V}(x,v) &= \sum_{i=1}^{n} v_i u_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} \nabla_{x_i}\Theta\left(x_{ij}\right)\left(\dot{x}_i - \dot{x}_j\right) + \sum_{i=1}^{n} v_i \nabla_{x_i} U_{O_{bi}} \\ &= -\sum_{i=1}^{n} v_i \sum_{j=1,j\neq i}^{n} \nabla_{x_i}\Theta\left(x_{ij}\right) + \sum_{i=1}^{n} v_i \sum_{j=1,j\neq i}^{n} a_{ij}\Psi\left(v_i - v_j\right) \\ &\quad - \sum_{i=1}^{n} v_i \nabla_{x_i} U_{O_{bi}} \\ &\quad + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} \nabla_{x_i}\Theta\left(x_{ij}\right)\left(\dot{x}_i - \dot{x}_j\right) + \sum_{i=1}^{n} v_i \nabla_{x_i} U_{O_{bi}} \\ &= -\sum_{i=1}^{n} v_i \sum_{j=1,j\neq i}^{n} \nabla_{x_i}\Theta\left(x_{ij}\right) + \sum_{i=1}^{n} v_i \sum_{j=1,j\neq i}^{n} a_{ij}\Psi\left(v_i - v_j\right) \\ &\quad + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} \nabla_{x_i}\Theta\left(x_{ij}\right)\left(\dot{x}_i - \dot{x}_j\right),\end{aligned} \tag{18}$$

where the control protocol (16) is applied.

Lemma 1 is considered, and we obtain

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} \nabla_{x_i}\Theta\left(x_{ij}\right)\left(\dot{x}_i - \dot{x}_j\right) = \sum_{i=1}^{n} v_i \sum_{j=1,j\neq i}^{n} \nabla_{x_i}\Theta\left(x_{ij}\right). \tag{19}$$

Substituting equation (19) into (18) results in

$$\dot{V}(x,v) = \sum_{i=1}^{n} v_i \sum_{j=1,j\neq i}^{n} a_{ij}\Psi\left(v_i - v_j\right) = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_{ij}\left(v_i - v_j\right)\Psi\left(v_i - v_j\right). \tag{20}$$

Substituting equation (12) into equation (20) yields

$$\begin{aligned}\dot{V}(x,v) &= \beta\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_{ij}\frac{\left(v_i - v_j\right)\left(v_j - v_i\right)}{\sqrt{1+\left(v_i - v_j\right)^2}} \\ &= -\beta\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_{ij}\frac{\left(v_i - v_j\right)^2}{\sqrt{1+\left(v_i - v_j\right)^2}} \leq 0.\end{aligned} \tag{21}$$

If

$$\Gamma \triangleq \left\{(x,v):\ \dot{V}(x,v) = 0\right\} = \left\{(x,v):\ \beta\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_{ij}\frac{\left(v_i - v_j\right)\left(v_j - v_i\right)}{\sqrt{1+\left(v_i - v_j\right)^2}} = 0,\ i = 1,\ldots,n-1\right\}, \tag{22}$$

then we have $\Gamma \triangleq \{(x,v):\ v_1 = \cdots = v_n\}$.

The preceding analysis implies that $\dot{V}(x,v) = 0$ if and only if $v_1 = \cdots = v_n$. Moreover, let $v_1 = \cdots = v_n = d$, and obtain $\dot{x}_i - \dot{x}_j = v_i - v_j = d - d = 0$. Thus, the consensus problem of the multiagent system (1) can be achieved.

Considering equation (9), let

$$\begin{aligned}&\nabla_{x_i}\left(\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\Theta\left(x_{ij}\right)\right) \\ &= \nabla_{x_i}\left\{\frac{\alpha}{2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\ln\left[\frac{h_{ij}}{2\left\|x_i - x_j\right\|} + \frac{\left\|x_i - x_j\right\|}{2h_{ij}}\right]\right\} = 0,\end{aligned} \tag{23}$$

where $\|x_i - x_j\| = h_{ij}$ is the equilibrium point of the multiagent system (1), and equation (17) has the minimum value under the equilibrium point.

The analysis of equation (23) shows that each agent can maintain the desired distance h_{ij}, that is, $\|x_i - x_j\| = h_{ij}$, which indicates that the desired formation is achieved. Function $V(x,v)$ is a bounded function due to $\dot{V}(x,v) \leq 0$. Therefore, the agents can avoid obstacles under the consensus control law (16). The proof is completed. □

5. Simulation Analysis

A class of second-order multiagent systems with four agents is considered. The desired formation shape is defined as a square. The length of the square is set as 4. The number of obstacles is set as 3. The dynamics of the four agents are described as equation (1), that is,

$$\begin{cases}\dot{x}_i(t) = v_i(t), \\ \dot{v}_i(t) = u_i(t),\end{cases} \tag{24}$$

and Figure 2 presents the directed topology among agents.

On the basis of Figure 2, adjacency matrix $\mathscr{A}$ and desired distance h are as follows:

$$\begin{aligned}\mathscr{A} &= \begin{bmatrix}0 & 0 & 0 & 1\\ 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\end{bmatrix}, \\ h &= \begin{bmatrix}0 & 4 & 4\sqrt{2} & 4\\ 4 & 0 & 4 & 4\sqrt{2}\\ 4\sqrt{2} & 4 & 0 & 4\\ 4 & 4\sqrt{2} & 4 & 0\end{bmatrix}.\end{aligned} \tag{25}$$

The four agents' initial positions are given as $(2.5, 1.7)$, $(1.3, 0.8)$, $(1.9, 2.4)$, and $(0.6, 1.5)$, and the initial velocities are given as $(0.5, 0.3)$, $(-0.5, 0.3)$, $(0.5, -0.3)$, and $(-0.5, -0.3)$, respectively. The positions of the three obstacles are set as $(26, 16)$, $(40, 30)$, and $(60, 25)$. The radius and detection radius of the three obstacles are $(2.5, 5.5)$, $(1.0, 2.5)$, and $(2.0, 4.0)$, respectively. The simulation time

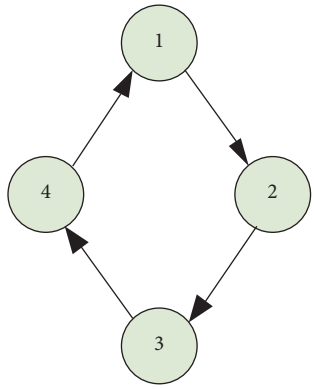


Figure 2: Communication topology with four agents.

$t = 30$ s, and the other parameters are given as $\alpha = 2.5$, $\beta = 1.5$, and $\gamma = 4.0$.

Two simulation examples are provided below to illustrate the validity of theoretical analysis.

Example A. Consensus analysis without obstacles.

The formation tracking problem without obstacles is considered in this example. Figures 3–7 display the formation tracking results.

Figures 3 and 4 exhibit the tracking results of position and velocity without obstacles, respectively. By virtue of the results of Figures 3 and 4, it can be easily found that the consensus formation tracking problem of the multiagent system (1) with the designed control protocol (16) can be achieved. Figure 5 shows the desired formation shape, which indicates that the desired formation tracking problem can also be solved under the proposed control protocol (16). Figure 6 depicts the desired distance among agents. This is completely consistent with the theoretical results, which further illustrate that the control protocol designed in this paper is effective. In addition, the control input curves of the four agents are given in Figure 7.

Overall, the results of Figures 3–7 show that the control protocol designed in this paper is effective. Although the existence of obstacles is not considered, the four agents can still achieve consensus, and at the same time, they can achieve and maintain the desired formation shape.

Example B. Consensus analysis with multiple obstacles.

In this example, three obstacles are considered in the process of achieving formation tracking. The initial position and velocity of each agent are the same as those in Example A. Figures 8–12 display the formation tracking results.

Figures 8 and 9 show the tracking results of position and velocity with three obstacles, respectively. Although three obstacles are considered in the process of achieving formation tracking control, the consensus can still be solved under the consensus control protocol (16). The tracking results of position and velocity of the four agents have changed due to the existence of obstacles. Figure 10 presents

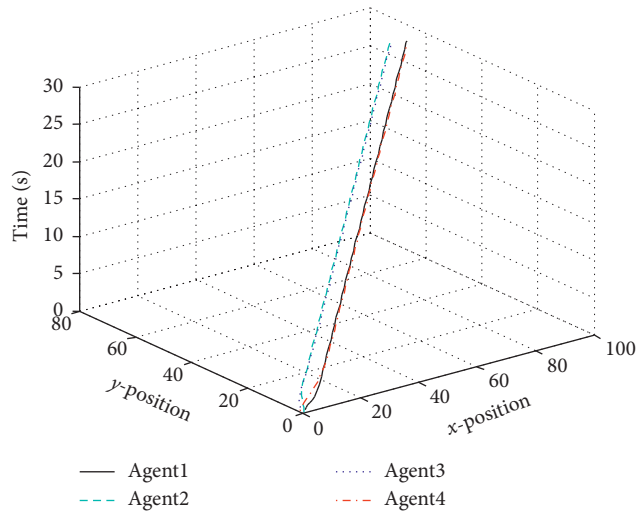


Figure 3: Position tracking curves without obstacles.

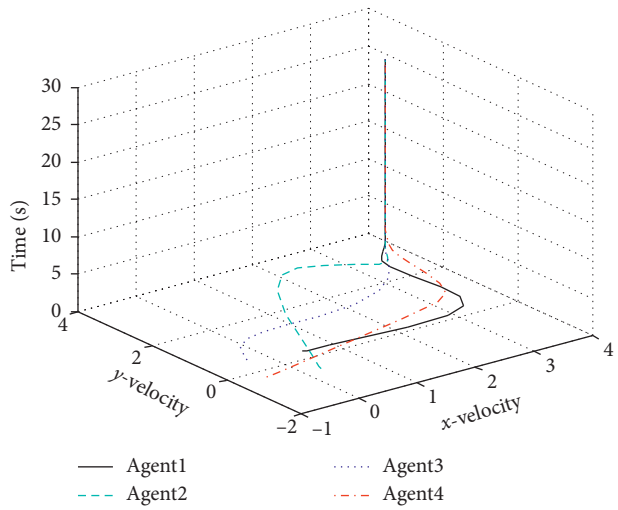


Figure 4: Velocity tracking curves without obstacles.

the desired formation shape, which implies that the formation can be maintained after avoiding obstacles. As can be seen from Figure 10, when an obstacle appears during the operation of the agent, the agent will bypass the obstacle under the action of the designed control protocol. After circumventing obstacles, the agents will continue to maintain the desired formation shape under the control protocol. This also illustrates the effectiveness of the control protocol designed in this paper from another angle. Figure 11 exhibits the desired distance of the four agents. The results in Figures 10 and 11 illustrate the effectiveness of the theoretical results. Figure 12 displays control input curves.

In general, the consensus control protocol (16) designed in this study not only can achieve the desired formation control but also can avoid obstacles and maintain the formation shape after avoiding the obstacles.

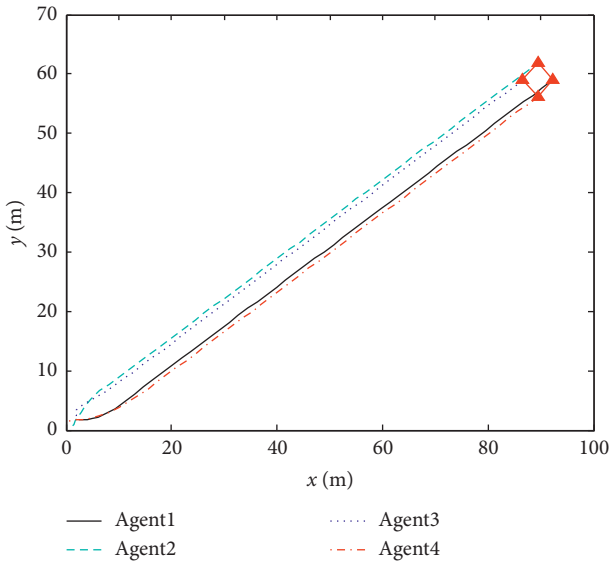


Figure 5: Desired formation shape without obstacles.

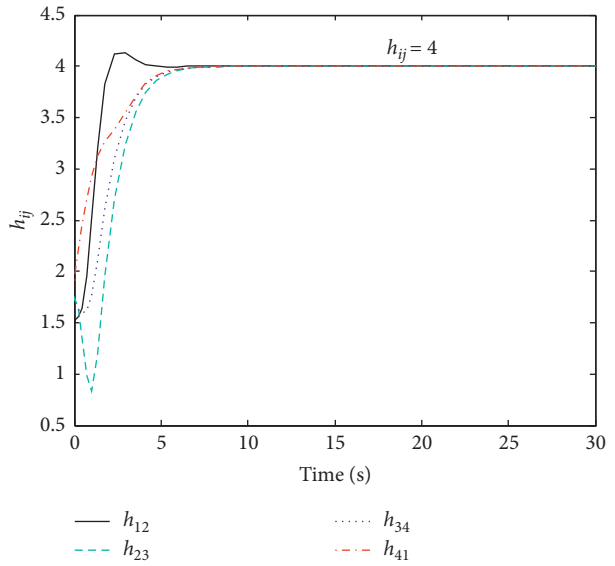


Figure 6: Desired distance without obstacles.

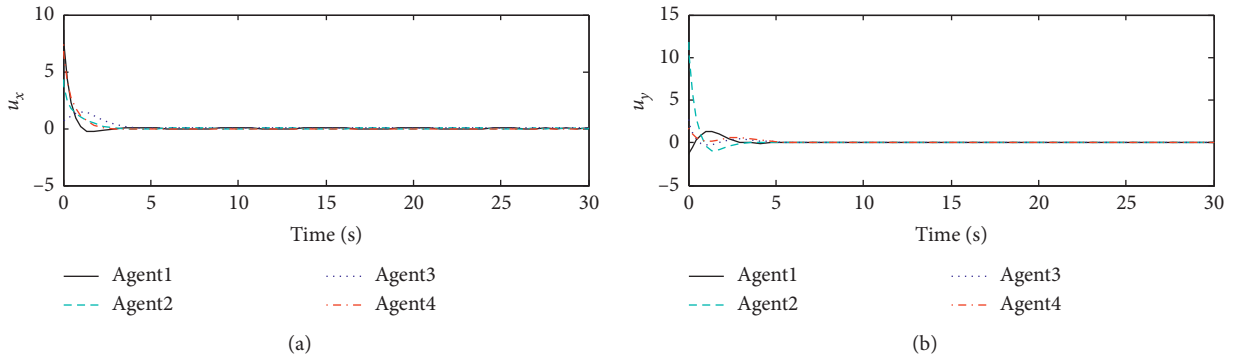


Figure 7: Control input curves.

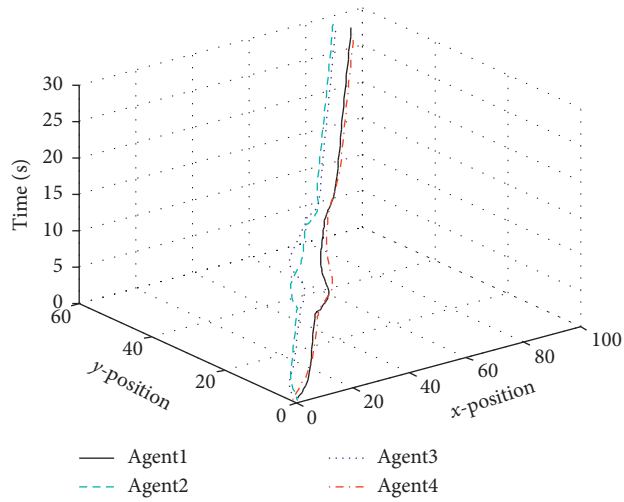


Figure 8: Position tracking curves with obstacles.

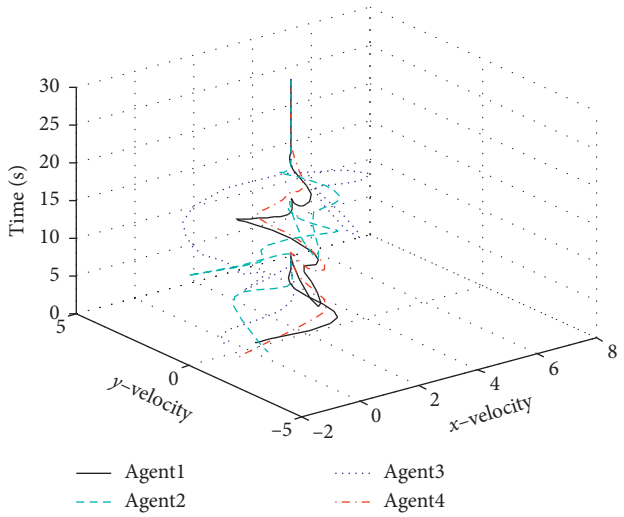


Figure 9: Velocity tracking curves with obstacles.

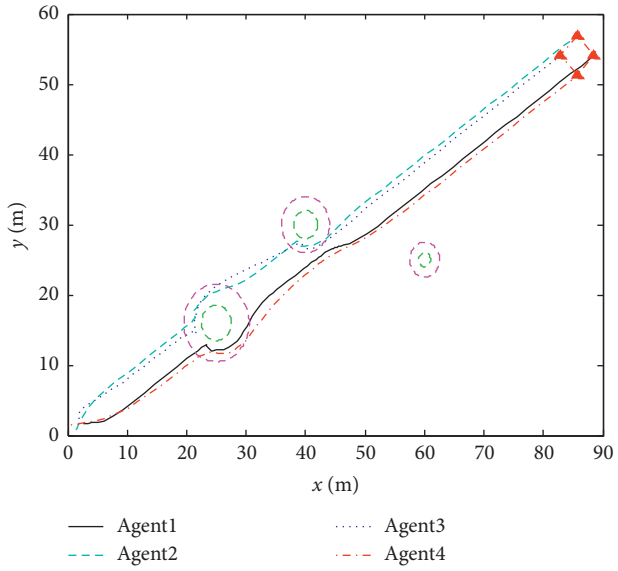


Figure 10: Desired formation shape with obstacles.

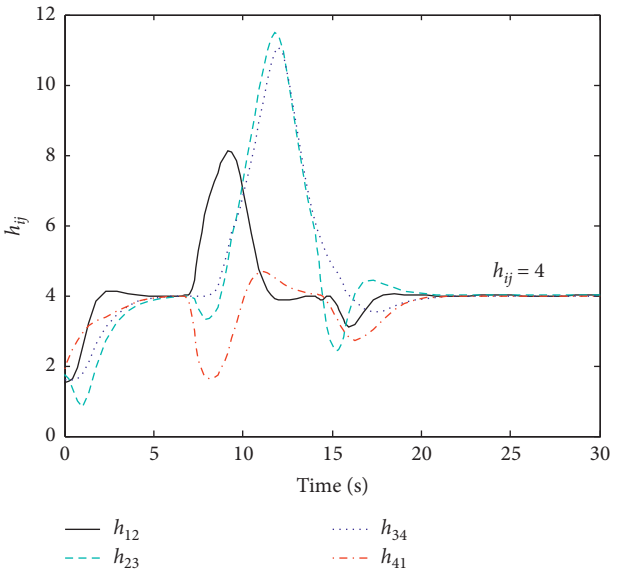


Figure 11: Desired distance with obstacles.

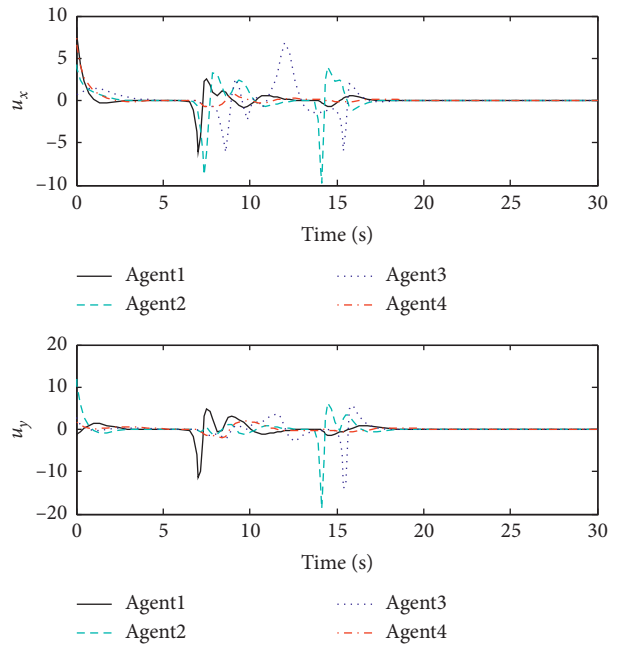


Figure 12: Control input curves.

6. Conclusions

The consensus problems of formation control and obstacle avoidance for a class of second-order multiagent systems with directed topology were considered in this study. The designed control protocol consisted of the formation, velocity consensus, and obstacle avoidance control laws. A designed Lyapunov function was applied to analyze the convergence of the designed consensus control protocol. Under the given directed topology, the formation control problem of the multiagent systems without obstacles was solved by using the designed consensus control protocol. At the same time, the desired formation control can be achieved and maintained despite the existence of obstacles in the environment.

Data Availability

The figure data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Lichao Wang and Xing Li contributed equally to this work.

Acknowledgments

This work was partially funded by the Anhui Provincial Natural Science Foundation (no. 1808085MF182), Anhui Natural Science Foundation (no. 1808085MF182), Anhui Provincial Key Research and Development Plan Project-Special Scientific and Technological Cooperation with Foreign Countries (no. 1804b06020368), National Natural Science Foundation of Anhui Polytechnic University (no. Xjky02201906), and Youth Research Fund of Anhui Polytechnic University (no. 2017YQ03).

References

[1] F. Rastgar and M. Rahmani, "Consensus-based distributed robust filtering for multisensor systems with stochastic uncertainties," *IEEE Sensors Journal*, vol. 18, no. 18, pp. 7611–7618, 2018.

[2] Y. Wang, D. Wang, S. Yang et al., "A practical leader–follower tracking control scheme for multiple nonholonomic mobile robots in unknown obstacle environments," *IEEE Transactions on Control Systems Technology*, vol. 27, no. 4, pp. 1684–1693, 2019.

[3] P. Glotfelter, J. Cortes, and M. Egerstedt, "Nonsmooth barrier functions with applications to multi-robot systems," *IEEE Control Systems Letters*, vol. 1, no. 2, pp. 310–315, 2017.

[4] L. He, P. Bai, X. Liang, J. Zhang, and W. Wang, "Feedback formation control of UAV swarm with multiple implicit leaders," *Aerospace Science and Technology*, vol. 72, pp. 327–334, 2018.

[5] X. Zhang and H. Duan, "Altitude consensus based 3D flocking control for fixed-wing unmanned aerial vehicle swarm trajectory tracking," *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, vol. 230, no. 14, pp. 2628–2638, 2016.

[6] J. S. Dominguez, A. J. Cerqueira Junior, D. S. Dominguez et al., "Using a multi-agent system for monitoring indicators of quality of service in power distribution networks," *IEEE Latin America Transactions*, vol. 13, no. 4, pp. 1048–1054, 2015.

[7] X. Ai, S. Song, and K. You, "Second-order consensus of multi-agent systems under limited interaction ranges," *Automatica*, vol. 68, pp. 329–333, 2016.

[8] X. Deng, X. Sun, R. Liu, and S. Liu, "Consensus control of leader-following nonlinear multi-agent systems with distributed adaptive iterative learning control," *International*

Journal of Systems Science, vol. 49, no. 16, pp. 3247–3260, 2018.

[9] X. Deng, X. Sun, S. Liu et al., "Consensus learning control for leader-following nonlinear multiagent systems with control delay," *Wireless Communications and Mobile Computing*, vol. 2019, Article ID 9858504, 13 pages, 2019.

[10] X. Deng, X. Sun, and S. Liu, "Iterative learning control for leader-following consensus of nonlinear multi-agent systems with packet dropout," *International Journal of Control, Automation and Systems*, vol. 17, no. 8, pp. 2135–2144, 2019.

[11] H. Li and H. Su, "Second-order consensus in multi-agent systems with directed topologies and communication constraints," *Neurocomputing*, vol. 173, pp. 942–952, 2016.

[12] Y. Liu and Y. Jia, "Event-triggered consensus control for uncertain multi-agent systems with external disturbance," *International Journal of Systems Science*, vol. 50, no. 1, pp. 130–140, 2019.

[13] X. Li, X. Luo, J. Wang, and X. Guan, "Finite-time consensus of nonlinear multi-agent system with prescribed performance," *Nonlinear Dynamics*, vol. 91, no. 4, pp. 2397–2409, 2018.

[14] T. Chen, D. Zhang, and X. Shao, "Control parameters design of spacecraft formation flying via modified biogeography-based optimization," *Aerospace Systems*, vol. 3, no. 1, pp. 1–8, 2020.

[15] G. Sun, R. Zhou, K. Xu et al., "Cooperative formation control of multiple aerial vehicles based on guidance route in a complex task environment," *Chinese Journal of Aeronautics*, vol. 33, no. 2, pp. 701–720, 2020.

[16] H. Liu, T. Ma, F. L. Lewis, and Y. Wan, "Robust formation control for multiple quadrotors with nonlinearities and disturbances," *IEEE Transactions on Cybernetics*, vol. 50, no. 4, pp. 1362–1371, 2020.

[17] Z. Yang, S. Zhu, C. Chen, G. Feng, and X. Guan, "Leader-Follower formation control of nonholonomic mobile robots with bearing-only measurements," *Journal of the Franklin Institute*, vol. 357, no. 3, pp. 1628–1643, 2020.

[18] A. Wasik, P. Lima, and A. Martinoli, "A robust localization system for multi-robot formations based on an extension of a Gaussianmixture probability hypothesis density filter," *Autonomous Robots*, vol. 44, no. 3-4, pp. 395–414, 2020.

[19] H. Du, G. Wen, Y. Cheng, Y. He, and R. Jia, "Distributed finite-time cooperative control of multiple high-order nonholonomic mobile robots," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 12, pp. 2998–3006, 2017.

[20] C.-B. Yu, Y.-Q. Wang, and J.-L. Shao, "Optimization of formation for multi-agent systems based on LQR," *Frontiers of Information Technology & Electronic Engineering*, vol. 17, no. 2, pp. 96–109, 2016.

[21] W. Zhao, W. Yu, and H. Zhang, "Observer-based formation tracking control for leader-follower multi-agent systems," *IET Control Theory & Applications*, vol. 13, no. 2, pp. 239–247, 2019.

[22] H. Chu, Y. Cai, and W. Zhang, "Consensus tracking for multi-agent systems with directed graph via distributed adaptive protocol," *Neurocomputing*, vol. 166, pp. 8–13, 2015.

[23] Y. Xia, X. Na, Z. Sun, and J. Chen, "Formation control and collision avoidance for multi-agent systems based on position estimation," *ISA Transactions*, vol. 61, pp. 287–296, 2016.

[24] D. Meng and Y. Jia, "Formation control for multi-agent systems through an iterative learning design approach," *International Journal of Robust and Nonlinear Control*, vol. 24, no. 2, pp. 340–361, 2014.

[25] J. Wang and M. Xin, "Optimal consensus algorithm integrated with obstacle avoidance," *International Journal of Systems Science*, vol. 44, no. 1, pp. 166–177, 2013.

[26] Y. Chen and J. Sun, "Distributed optimal control for multi-agent systems with obstacle avoidance," *Neurocomputing*, vol. 173, pp. 2014–2021, 2016.

[27] J. Alonso-Mora, T. Naegeli, R. Siegwart, and P. Beardsley, "Collision avoidance for aerial vehicles in multi-agent scenarios," *Autonomous Robots*, vol. 39, no. 1, pp. 101–121, 2015.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 2534238, 12 pages
https://doi.org/10.1155/2020/2534238

Research Article

Graph Theory-Based Mathematical Calculation Modeling for Temperature Distribution of LED Lights' Convective Cooled Heat Sinks under Moisture Environment

Bei-xuan Lyu,[1] **Yu-ren Chen**,[2] **and Yong Li**[1]

[1]*Department of Mechanical and Electrical Engineering, Hulunbeier Vocational Technical College, Hulunbeier, China*
[2]*Department of Electronic Information Materials, School of Materials Science and Engineering, Shanghai University, Shanghai, China*

Correspondence should be addressed to Bei-xuan Lyu; shurobots@126.com

Received 11 December 2019; Revised 12 March 2020; Accepted 6 April 2020; Published 27 April 2020

Guest Editor: Sanghyuk Lee

In this paper, a mathematical model based on graph theory is proposed to calculate the heat distribution of LED lights' convective cooled heat sink. First, the heat and mass transfer process of a single fin under moisture environment is analyzed. Then, the heat transfer process is characterized by a digraph, defining fins and joints of a heat sink as edges and vertices in graph theory. Finally, the whole heat transfer process is described by two criteria achieved based on graph theory. Therefore, the temperature-heat calculation equations of the whole heat sink are deduced. The accuracy of this model is verified by testing the junction temperature of different LED chips mounted on the same heat sink under moisture environment, and the relative errors between the calculated value and the experimental data are all within 5%, and it is also concluded from the model that heat sinks with an identical heat digraph but different types have close cooling performance and are verified by two typical heat sinks, cylindrical heat sink and rectangular plate-fin heat sink, under the same conditions. The mathematical model based on group theory developed in this paper combined with computer technology is convenient for the performance analysis among a large number of heat sink fin arrangement schemes.

1. Introduction

Nowadays, light-emitting diode (LED) has met an extensive application in every aspect of our life. According to some comprehensive studies' conclusions [1–3], high LED junction temperature will not reduce the service life of LEDs significantly yet has a negative impact on the optical performance, revealing the great significance of reasonable heat sink design to LED lanterns. Now, extensive research studies have been conducted well and widely, concerning heat sinks of LED lights to improve their cooling performance.

Mathematical models to analyze thermal characteristics of heat sinks are developed by various scholars in recent years, such as Sun [4], who established a model to predict the peak temperature of a LED bulb by analyzing the performance of its heat sink with effective heat conductivity, and according to simulations and experimental results in [4], Sun [5] extended the simplified mathematical model to estimate the heat flow of a LED bulb in the spherical coordinate system further and verified the validity of the model by numerical simulation with the explicit finite-difference method (EFDM), considering the actual working conditions. Jang et al. [6] developed a correlation model to predict the Nusselt number, which was validated experimentally, thus evaluating the influence of orientation effect on the heat dissipation capacity of an inclined cylindrical heat sink. Also, the model showed that the cooling performance escalated, accompanied with the increasing angle of inclination of the heat sink. Subsequently, they [7] supplemented the experiment to characterize the relationship between thermal resistance of a plate-fin heat sink and its installation angles on the basis of [6], which showed that the thermal resistance

decreased first and then increased, according to the increased angle. Jang further detailed the analytical model in [8] to predict the thermal resistance of a plate-fin heat sink and correlated the installation angle with its essential design parameters at last. Lv and Xiong [9] deduced a temperature distribution model of a heat sink and verified its correctness by experiment and numerical simulation. Besides, they [10] established a mathematic model of LED heat sinks in greenhouses by the condition that the heat conducted into the fin heat equals the total heat from the air convection and the condensation of wet air.

Besides, many optimization methods were applied to balance the influence of various factors, e.g., mass, size, and number of fins to achieve optimal performance. Khan [11] optimized cylindrical pin-fin heat sinks by using the entropy generation minimization (EGM) method, concerning the influence of heat transfer and pressure drop. Ahmed [12] analyzed the influence of orientation effect on heat transfer performance of heat sinks and optimized heat sinks by numerical calculation and simulation. Lv and Liu maximized the capacity of a heat sink to dissipate heat by using a topology optimization method in [13]. By using the particle swarm optimization (PSO) algorithm, Lampio and Karvinen [14] optimized cooled heat sinks. Bar-Cohen [15] studied the total coefficient to measure the cooling capacity of heat sinks and optimized their performance by the least-energy method.

Lately, researchers also studied the cooling performance of heat sinks, which vary in different contractures, forms, and geometric parameters, by analyzing their thermal and flow characteristics, thus derived and innovated novel heat sinks. Schmid et al. [16] analyzed the influence of varying the inter-fin base length on the natural convective heat sinks by the numerical parametric study, and they found that inter-fin base length determined the optimal fin spacing after the comparison of heat sinks with two different fin orientations with the identical base area at the same condition. Yu et al. [17] found that it was impossible to meet the demand of achieving the optimal performance with a lightweight at the same time after performing parametric studies of a radial heat sink. Jeong et al. [18] proposed a modified plate-fin heat sink with openings and investigated the effects of its all design parameters on heat dissipation performance, and the novel structure can reduce 30.5% thermal resistance compared with the conventional no-opening form.

In a word, existing research mainly focuses on the study of novel heat sink development and optimization and improvement of heat sinks and the calculation model foundation in cooling performance. Graph theory is the study of graphs, which can analyze relationships among objects quantitatively and has a wide application in various fields, e.g., networking, algorithm, decision making, and electronics [19]. Especially in mechanism synthesis, a large number of mechanisms with similar performance can be automatically generated by using graph theory in computers [20]. Enlightened by this, the combination of graph theory and computer technology in order to evaluate the optimal design among a large number of heat sinks' fin arrangement swiftly will play a positive role in the development of LED cooling technology. In this study, based on our previous research work [9, 10] and the equivalent thermal admittance proposed in [21], we developed a mathematical model based on graph theory to evaluate the heat performance and thus laid the foundation for the follow-up development.

2. Mathematical Model Based on Graph Theory

2.1. Assumptions and Simplification. To carry out a steady-state analysis of a typical plate-fin exposed to moisture environment, as shown in Figure 1, the following essential assumptions, also known as the Murray–Gardner assumptions, are made to simplify this analysis:

(i) The fin material is isotropic, and its thermal conductivity remains constant in all directions

(ii) The thermal resistance of the condensed film is negligible

(iii) The latent heat of condensation of water vapor is unchanged

(iv) Compared with the heat flowing through the side of the fin, the heat flowing through the outermost part of the fin is neglected

(v) The influence of air pressure drops caused by airflow is neglected

(vi) The effect of heat radiation is ignored, and the heat sink surface is diffuse and gray

(vii) The flow is three dimensional and laminar

2.2. Fundamentals of Heat and Mass Transfer. A typical plate-fin and its terminology and coordinate system are shown in Figure 1. The origin of the length coordinate is set at the tip of the fin, and the positive sense is in a direction from the tip towards the base. In this process, mass transfer accompanied by the heat transfer and the heat conducted into the fin from LED chips equals the total energy taken away by the air convection and the condensation of wet air. Therefore, considering Fourier's law of heat conduction, Newton's law of cooling, and the law of mass transfer in wet air, these physical phenomena at any position x along the length coordinate can be expressed in the following form:

$$(q + dq) + 2 \cdot (H + \delta) \cdot h \cdot (T - T_a) \cdot dx + 2 \cdot (H + \delta) \cdot h_D \cdot i_{fg} \cdot (\omega_a - \omega) \cdot dx - q = 0. \tag{1}$$

According to the Chilton–Colburn analogy [22], the relationship between heat transfer and mass transfer coefficients can be expressed by the following equation:

$$\frac{h}{h_D} = c_p \cdot Le^{0.48} = c_p \cdot \left(\frac{\alpha}{D}\right)^{0.48}. \tag{2}$$

Hence, if $H >> \delta$ and the previous assumptions permit, the equation can be written in the following form:

$$\frac{d^2\theta}{dx^2} - \left(\frac{2 \cdot h}{k \cdot \delta}\right) \cdot \left[\theta + \frac{i_{fg}}{c_p Le^{0.48}} \cdot (\omega - \omega_a)\right] = 0, \tag{3}$$

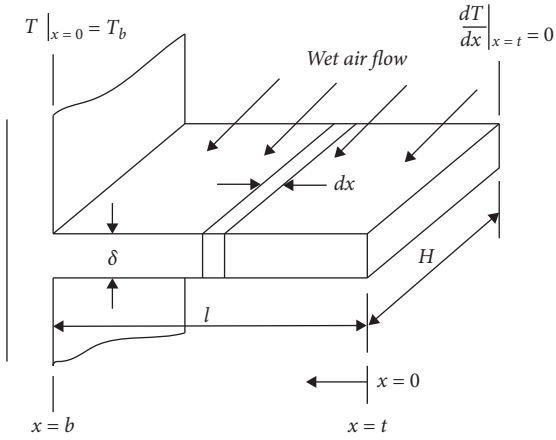


Figure 1: A fin of typical rectangular profile and its terminology and coordinate system.

where θ is the temperature excess between the fin and the surrounding environment. Considering the boundary conditions $\theta_{(x=b)} = \theta_b$ and $q_{(x=b)} = q_b$, we have

$$\begin{bmatrix} e^{mx} & e^{-mx} & -N\cdot(\omega-\omega_a) \\ kAme^{mH} & -kAme^{-mH} & 0 \end{bmatrix} \cdot \begin{bmatrix} C_1 \\ C_2 \\ 1 \end{bmatrix} = \begin{bmatrix} \theta_b \\ q_b \end{bmatrix}, \tag{4}$$

where

$$C_1 = \frac{k\delta Lm\cdot[\theta_b + N\cdot(\omega-\omega_a)]\cdot e^{-mH} + q_b\cdot e^{-mH}}{2k\delta Lm}, \tag{5}$$

$$C_2 = \frac{k\delta Lm\cdot[\theta_b + N\cdot(\omega-\omega_a)]\cdot e^{mH} - q_b\cdot e^{mH}}{2k\delta Lm}. \tag{6}$$

Equation (9) suggests the relationship between the temperature excess and the heat flow at position x, and it can be rearranged as

$$\begin{bmatrix} \theta_t \\ q_t \end{bmatrix} = \begin{bmatrix} \cosh(mH) & -\frac{1}{k\delta Lm}\cdot\sinh(mH) \\ -k\delta Lm\cdot\sinh(mH) & \cosh(mH) \end{bmatrix} \cdot \begin{bmatrix} \theta_b \\ q_b \end{bmatrix} + \begin{bmatrix} \left[\frac{1}{k\delta Hm}\cdot\cosh(ml) - 1\right]\cdot N(\omega-\omega_a) \\ -\sinh(ml)\cdot N(\omega-\omega_a) \end{bmatrix}. \tag{7}$$

2.3. Mathematical Modeling Based on Graph Theory

2.3.1. The Graph Model of Heat Flow Analysis. In graph theory, the graph is defined as a pair set $G = (V, E_\psi)$ and can satisfy the following conditions:

(1) V is the nonempty set

(2) $E \subseteq [V]^2$; thus, the elements of E are 2-element subsets of V

(3) Function ψ: $E \longrightarrow V \times V$

Here, V denotes vertex-set and E represents edge-set, and if all the elements of E_ψ are ordered pairs, the graph is called digraph; if a graph in which the edges are not ordered, it is called an undirected graph. A graph (V, E_ψ) with vertex-set $V = \{x_1, x_2, \ldots, x_n\}$ and edge-set $E = \{a_1, a_2, \ldots, a_m\}$ can be specified by the adjacency matrix, which is the n × n matrix A:

$$A = \left(a_{ij}\right)_{n\times n}, \quad a_{ij} = \mu\left(x_i, x_j\right), \tag{8}$$

where μ(xi, xj) is the number of edges connecting vertices i and j. Furthermore, a graph can also be represented by the $n \times m$ matrix M called incidence matrix:

$$M = \left(m_x(a)\right), \quad x \in V,\ a \in E. \tag{9}$$

For the commonly used plate-fin heat sinks, they can be regarded as a composition of fins with joints arranged by different geometric requirements, and the joint is the junction connecting several fins, e.g., the W-shaped and K-shaped heat sinks shown in Figure 2, respectively, are all composed of four fins, but with different numbers of joints. In order to simplify the procedure to get graph mappings from specific heat sinks, fins and joints of a heat sink as its elementary geometric components can be regarded as edges and vertices in graph theory, respectively. The undirected graph $G_H(V, E)$, in which the subscript H represents the heat sink, characterizes the geometric characteristics of the heat sink, in which V represents the joint set and E represents the fin set, and its adjacency matrix A_H of this undirected graph, which is the $n \times n$ matrix, can represent the actual geometric relationship of this heat sink.

When using the heat sink to transfer heat from LED chips to the surrounding environment, a heat transfer network is formed and can be characterized by a digraph $G_F(V', E')$, in which subscript F represents the heat flow, and this digraph is established on the basis of undirected graph G_H by adding the following conditions:

(1) The joint around the heat source often is regarded as the start vertex of the digraph

(2) Adding an extra vertex as the end vertex in the digraph, considering the influence of the surrounding environment

(3) The edge-set E' is founded on the basis of making all edges in set E of the graph $G_H(V, E)$ directional, and adding directed edges from all vertices except the start one towards end vertex

(4) Referring to the conservation of energy, constructing another directed edge starting from the end vertex, namely, representing the surrounding environment towards the start vertex, thus forming a closed loop of the digraph

If the undirected graph mapping geometric characteristics of the heat sink has these conditions $G_H(V, E)$, $|V| = n$, and $|E| = m$, we can easily find that $|V'| = n + 1$ and $|E'| = m + n$ for the digraph $G_F(V', E')$, reflecting the actual heat exchange. Therefore, the heat transfer from LED chips to the

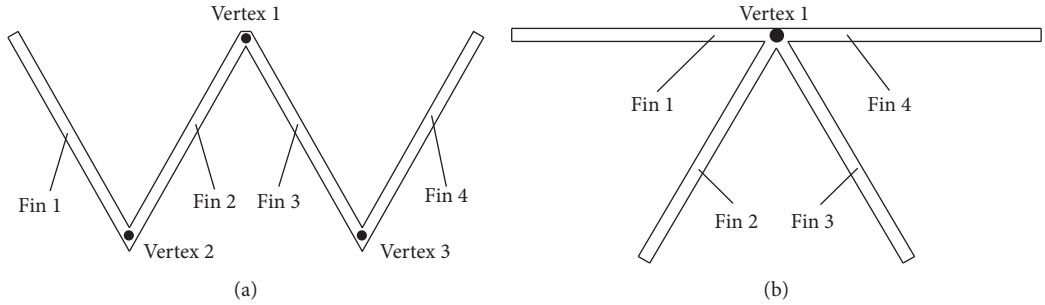


Figure 2: Two types of fin arrays with identical fin numbers but different joint numbers. (a) W-shaped fin array with four fins and three vertices. (b) K-shaped fin array with four fins and one vertex.

surrounding environment, inclusive of heat conduction and thermal convection, can be expressed by the digraph $G_F(V', E')$, namely, the association between multiple vertices and edges, and the whole heat flow can also be expressed by the incidence matrix M_F by $(n+1) \times (m+n)$ and the elements a_{ij} in M_F are "1," "−1," or "0" in accordance with the scheme:

$$a_{ij} = \begin{cases} 1, & \text{if the heat flow } j \text{ leaves vertex } i, \\ -1, & \text{if the heat flow } j \text{ enters vertex } i, \\ 0, & \text{if the heat flow } j \text{ does } n\text{ot touch vertex } i. \end{cases} \tag{10}$$

The heat generated by LED chips flows into the environment through several different fin branches; thus, there are multiple heat flow loops during the whole heat transfer process. In order to give an accurate expression, a matrix CF by $(n-1) \times (m+n)$ named as the heat flow loop matrix is introduced. The elements c_{ij} in heat flow loop matrix C_F are "1" or "0" in accordance with the scheme:

$$c_{ij} = \begin{cases} 1, & \text{if the heat flow } i \text{ inclusive of edges } j, \\ 0, & \text{if the heat flow } i \text{ exclusive of edges } j. \end{cases} \tag{11}$$

Therefore, the geometric characteristics of a heat sink composed by n vertices and m edges, namely, $|V| = n$, $|E| = m$, can be represented by an undirected graph $G_H(V, E)$, and the adjacency matrix A_H is used to describe the geometric relationship of edges by n vertices. After LED chips are mounted on the heat sink, a digraph $G_F(V', E')$ established by the undirected graph $G_H(V, E)$ is used to describe the heat generated by LED chips flowing towards the surrounding environment through the heat sink and for the digraph $G_F(V', E')$, the incidence matrix MF by $(n+1) \times (m+n)$ and heat flow loop matrix CF by $(n-1) \times (m+n)$ are used to reflect the actual heat exchange. The specific calculation of LED temperature excess by matrix M_F and C_F is elaborated in the following section.

2.3.2. Calculation of Thermal Admittance Matrix by Graph Theory. The relationship between the temperature excess and the heat flow at both ends discussed in Section 2.2 can be analyzed by the application of the graph model to form the heat flow digraph. Therefore, after the linear transformation of equations (12), (13), and (14), the relation between heat flow and temperature excess may be written in the matrix form as

$$\begin{bmatrix} q_b \\ q_a \end{bmatrix} = \begin{bmatrix} k\delta Lm \cdot \coth(mb) & -k\delta Lm \cdot \operatorname{csch}(mb) \\ k\delta Lm \cdot \operatorname{csch}(mb) & -k\delta Lm \cdot \coth(mb) \end{bmatrix} \cdot \begin{bmatrix} \theta_b \\ \theta_a \end{bmatrix} + \begin{bmatrix} \left[\frac{1}{k\delta Hm} \cdot \cosh(ml) - 1\right] \cdot N(\omega - \omega_a) \\ -\sinh(ml) \cdot N(\omega - \omega_a) \end{bmatrix}, \tag{12}$$

and according to the definition of the equivalent thermal admittance Y, a parameter to measure heat transfer performance of the fin was proposed in [21], and its definition formula is as follows :

$$Y = \frac{q}{\theta} = \frac{1}{R}. \tag{13}$$

Figure 3 shows a typical thermal transmission case of a single flat-plate fin from the fin base to the fin tip, namely, from point A to point B, representing thermal energy flows from high temperature to low temperature by heat conduction. As mentioned in Section 2.3.1, the heat flows from vertex A towards vertex B, and the two vertices are all connecting the vertex C which is representing the surrounding environment, thus forming two directional edges all towards the vertex C, but started from vertex A and vertex B, respectively. We can give the following:

$$\begin{cases} Q_A = q_1 + q_2, \\ Q_B = q_1 - q_3, \end{cases} \tag{14}$$

wherein heat flow q_1 and q_3 represent the heat quantity by convective heat transfer and q_2 represents the heat conducted through the fin. Referring to equation (13), each heat flow equals a product of admittance and a temperature excess:

$$\begin{cases} q_1 = Y_1(\theta_A - \theta_B), \\ q_2 = Y_2\theta_A, \\ q_3 = Y_3\theta_B, \end{cases} \tag{15}$$

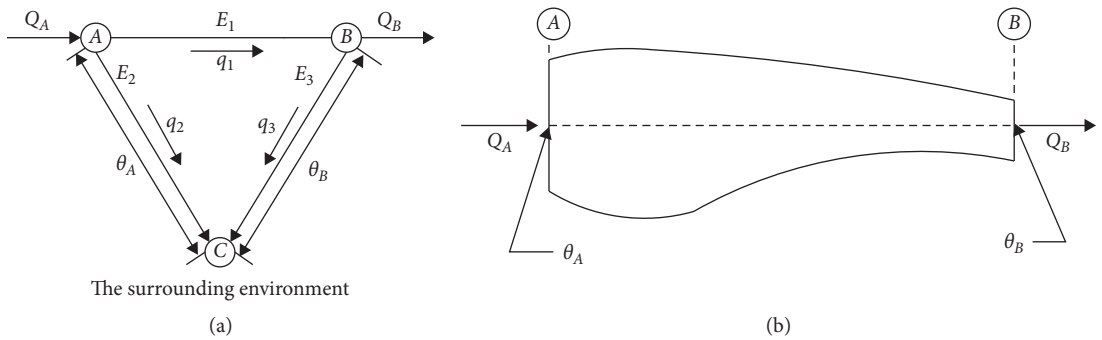


FIGURE 3: (a) The actual thermal transmission of a fin and its equivalent thermal digraph. (b) In the thermal digraph, the positive direction is set from high temperature towards low temperature.

wherein Y_2 and Y_3 represent the convection thermal admittance of edge 2 and edge 3 and Y_1 is measuring the conductive performance of edge 1. Considering the one-dimensional Fourier equation, Newton's law of cooling, and the law of mass transfer in wet air, we have

$$\begin{cases} q_2 = k\delta H \cdot \dfrac{\theta_A - \theta_B}{l}, \\ q_1 = 2 \cdot (H + \delta) \cdot \left[h \cdot \theta_a + h_D \cdot i_{fg} \cdot (\omega_a - \omega)\right], \\ q_3 = 2 \cdot (H + \delta) \cdot \left[h \cdot \theta_b + h_D \cdot i_{fg} \cdot (\omega_a - \omega)\right]. \end{cases} \tag{16}$$

Referring to equation (15), we obtain

$$\begin{cases} Y_1 = \dfrac{k\delta H}{l}, \\ Y_2 = 2 \cdot (H + \delta) \cdot h, \\ Y_3 = 2 \cdot (H + \delta) \cdot h. \end{cases} \tag{17}$$

As the basic geometric composite unit of the whole heat sink, each fin has limited capacity to transfer heat into the environment, and its undissipated heat is transferred to other fins by heat conduction. Therefore, heat will be spread on the whole heat sink by conductive transfer, namely, through the fins connected to form the heat sink, in order to exchange the heat into the environment by convective transfer. Obviously, heat conduction does not consume heat in this process, and according to the definitions mentioned above, the convective heat transferred into the environment by a heat sink with n vertices can be calculated as follows:

$$Q = \sum_{i=1}^{n-1} Y_i \cdot \left(T_{\text{LED}} - T_{\text{Ambient}}\right), \tag{18}$$

wherein Y_i represents the convective thermal admittance.

Referring to the conservation of energy and relevant contents mentioned above, the whole heat transfer process of the LED lights by using heat sinks can be described by the following criteria:

(i) The algebraic sum of the heat flow at each vertex in the graph is zero

(ii) The algebraic sum of temperature difference of each edge along any flow in the graph is zero

They can also be expressed by the following equations:

$$M_F \cdot \widetilde{q} = O, \tag{19}$$

and

$$C_F \cdot \widetilde{\theta} = O, \tag{20}$$

wherein $\widetilde{\mathbf{q}}$ is the heat flux column vector by $(m+n) \times 1$, $\widetilde{\boldsymbol{\theta}}$ is the temperature excess column by $(n-1) \times 1$, respectively, and O is the zero vector. By associating equations (18), (19), and (20), the temperature calculation equations of the whole heat sink is deduced. Therefore, the incidence matrix M_F and the heat flow loop matrix C_F, all derived from matrix A_H, are the key to finish the temperature calculation.

2.3.3. Identical Digraph of Different Heat Sinks. In graph theory, isomorphic digraphs, as the two shown in Figure 4, have the exact same pattern, and the sole difference is the label of their vertices and edges. In general, if digraphs G and H are isomorphic, it can be written G ≅H. Besides, there are bijections θ: $V(D) \longrightarrow V(H)$ and φ: $E_\psi(D) \longrightarrow E_{\psi'}(H)$, thus making $a \in E_\psi(D)$ have $\psi(a) = (x, y)$ and $\psi'(\varphi(a)) = (\theta(x)$ and $\theta(y)) \in E_{\psi'}(H)$; such a pair of mappings is called an isomorphism between G and H.

Hence, based on the isomorphic digraphs concept, identical digraphs to characterize heat transfer from LED chips to the surrounding environment through heat sinks, even with different types, must satisfy the following conditions:

(1) The two digraphs are isomorphic, and they have the same digraph pattern and directions of each directed edges

(2) The thermal admittance of one edge in a digraph and the mapping edge in its digraph pair is equal

3. Results and Discussion

3.1. Case 1: Verification under Moisture Environment

3.1.1. Experimental Apparatus and Procedure. The test heat sink is made by grinding extruded aluminum alloy 6063 to

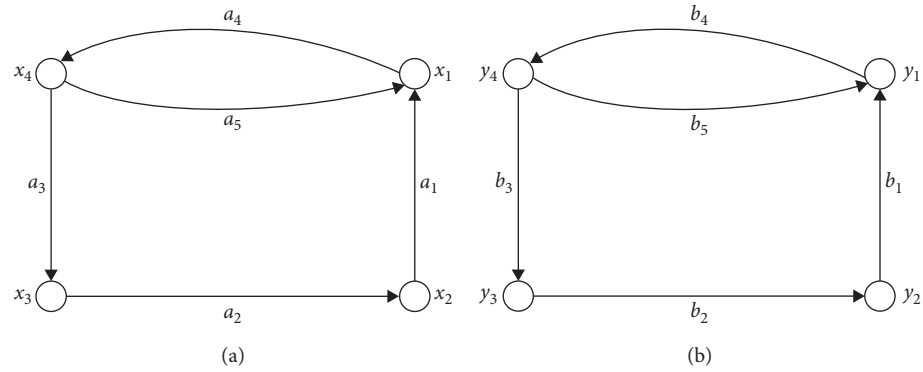


Figure 4: Two digraphs (a, b) represented by the same pattern.

0.8 μm roughness by CNC milling machine, thus reducing radiation emissivity, therefore, increasing the accuracy of the measurement. The main component's materials and thermal performance parameters are shown in Table 1 and the detailed dimensions are shown in Figure 5. In order to verify the mathematical model, the junction temperature of 20 W, 25 W, 30 W, 35 W, and 40 W LED chips mounted on the same heat sink was measured at 50%–100% relative humidity with T-type thermocouple and the specific experimental apparatus is shown in Figure 6. Therefore, we place the fins of the heat sink upright, namely, parallel to the gravity direction, to avoid the influence of the orientation effect. The temperature of the experimental environment is controlled at 20°C, and the data acquisition started 1 hour after the LED lamp is lighted when the LED chip reaches a stable level.

3.1.2. Calculation Based on Graph Theory. The heat sink shown in Figure 5 presents a typical symmetrical structure; thus, we reduce computation by analyzing the heat flow of one side; therefore, the amount of heat generated by the LED chip is half. We achieve the digraph shown in Figure 7(b) which is mapping the actual heat transfer through the heat sink. In this digraph, the vertex 1 marked in red circle is the heat source, also representing the start vertex in the digraph, and the vertices in blue circles represent the actual joints of this heat sink. Besides vertex 9, namely, the vertex in gray in Figure 7(b) is the extra vertex acting as the end vertex mentioned in Section 2.2, thus considering the influence of the surrounding environment. Black lines in solid connecting vertices, which are mapping the actual geometrical structure of the heat sink, indicate the heat transfer by conduction through the heat sink pointed from high to low temperature, and black dash lines denote all vertices in blue circles which transfer heat from heat sink towards the environment by thermal convection. In addition, the red dash line, namely, the E_{15} in Figure 7(b) denotes an illusory heat flow from the environment towards the heat source, considering the conservation of energy and relevant contents mentioned in Section 2.3.2, thus forming closed loops for calculation.

Hence, the heat generated by LED chips, thus starts from the vertex 1, along seven flow loops, all inclusive of heat conduction and thermal convection, denoted by solid lines and dash lines, respectively, in Figure 7(b) to finish the whole thermal circulation, and the flow matrixes of the seven flow loops are as follows:

$$\begin{aligned} C_1 &= \{e_1, e_2, e_8, e_{15}\}, \\ C_2 &= \{e_1, e_9, e_{15}\}, \\ C_3 &= \{e_3, e_{10}, e_{15}\}, \\ C_4 &= \{e_3, e_4, e_{11}, e_{15}\}, \\ C_5 &= \{e_3, e_4, e_5, e_{12}, e_{15}\}, \\ C_6 &= \{e_6, e_7, e_{14}, e_{15}\}, \\ C_7 &= \{e_6, e_{13}, e_{15}\}, \end{aligned} \tag{21}$$

The matrix A_H by $n \times n$ can be easily achieved. Besides, by analyzing the heat flow, we achieved the incidence matrix M_F and the heat flow loop matrix C_F, and all the corresponding matrixes for calculation are shown in Table 2. Referring to equations (19) and (20), we got curves between relative humidity and junction temperature of 20 W, 25 W, 30 W, 35 W, and 40 W LED by calculation and conducted experiments for comparison.

3.1.3. Results and Analysis. Figure 8 shows the calculational and experimental relation curves between relative humidity and junction temperature of 20W, 25W, 30W, 35W, and 40W LED chips. All curves present the same change trend, the junction temperature of all kinds increase with the increasing ambient humidity, and the changing trend of calculation curves is basically consistent with that of experimental curves, namely, the junction temperature of LED increasing with the increase of humidity. Comparing the difference between calculational and experimental values, all the temperature differences between the measured and calculated values are all within 5°C, namely, the relative errors between the calculated value and the experimental data are all within 5%. Therefore, the correctness of the temperature distribution calculation model proposed in this paper is in good agreement with the experimental measurements.

Table 1: Main component's materials and thermal performance parameters.

Components	Materials	Density/kg·m^{-3}	Conductivity/W·m^{-1} K^{-1}	Specific heat/J·kg^{-1} K^{-1}	Thermal emissivity
Heat sink	Polished aluminum alloy 6063	2690	202	900	0.1
LED chip	Semiconductor material	6150	130	417	—

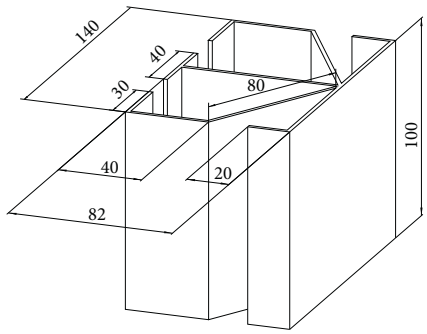


Figure 5: The heat sink and its specific dimensions.

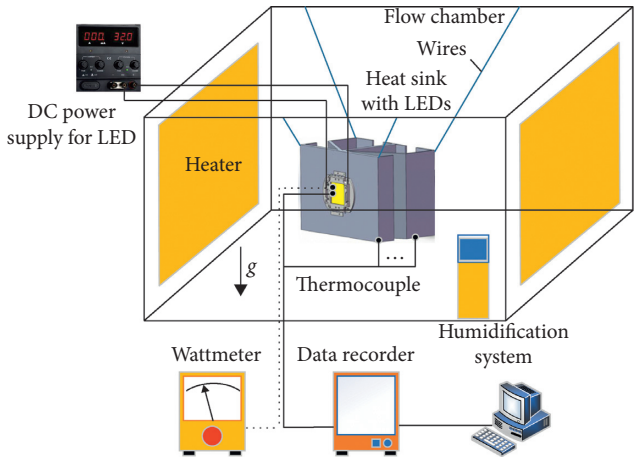
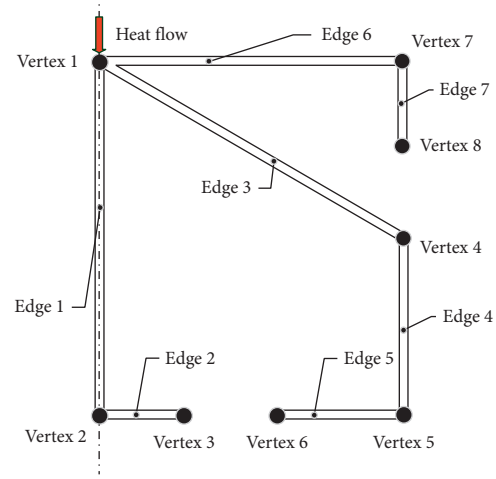
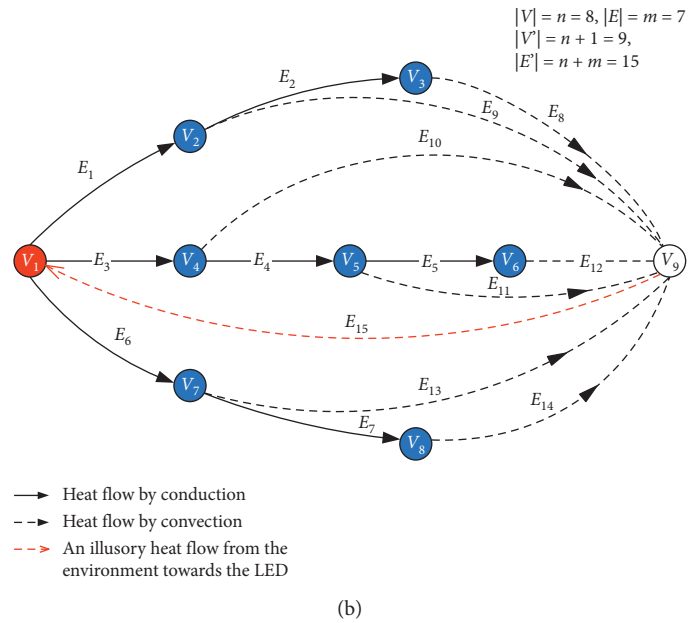


Figure 6: Simplified model of the experimental apparatus.



(a)



(b)

Figure 7: Vertices and edges of the heat sink (a) and the digraph of the heat flow (b).

3.2. Case 2: Verification of Close Cooling Performance of Heat Sinks with Identical Heat Digraphs but Different Types. As shown in Figure 8, a cylindrical heat sink and a rectangular plate-fin heat sink are used in this case, and for the rectangular plate-fin heat sink, we treat its base as a joint such as the cylindrical one, thus making the two heat sinks sharing a similar heat flow digraph. Referring to equation (15) in Section 2.3.2, the equivalent thermal admittance Y of each edge is equal, considering all fins of the two heat sinks having identical geometry size under the same working condition. Thus, it can be easily deduced from the model mentioned above that heat sinks sharing an identical heat digraph but different types have close cooling performance. In order to verify this conclusion, we tested the junction temperature of 20°W LED chips, respectively, mounted on two typical heat sinks, cylindrical heat sink and rectangular plate-fin heat sink, all with 6 fins of the same size, apparently belonging to two different geometrical structures but sharing with an identical heat sink graph, as shown in Figures 8 and 9, respectively, using the same experimental method and heat sink materials, as mentioned in Section 3.1.

We achieve the digraph shown in Figure 10 which maps the actual heat transfer through the heat sink, the heat along with six flow loops expressed by equation (22), to finish the whole thermal circulation, and corresponding matrixes for calculation are shown in Table 3:

Table 2: The matrix representation of the heat flow graph.

Each matrix	The specific matrix
The adjacency matrix A_H by $n \times n$ (8×8)	$\begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$
The incidence matrix M_F by $(n+1) \times (m+n)$ (9×15)	$\begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 1 \end{pmatrix}$
The flow matrix CF by $(n-1) \times (\mathrm{m}+\mathrm{n})$ (7×15)	$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}$

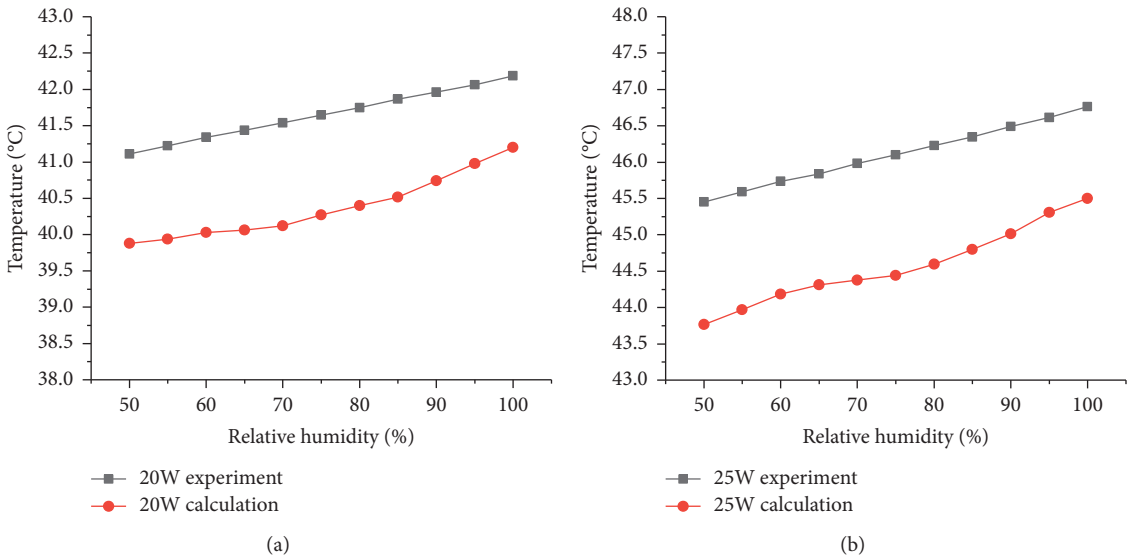


Figure 8: Continued.

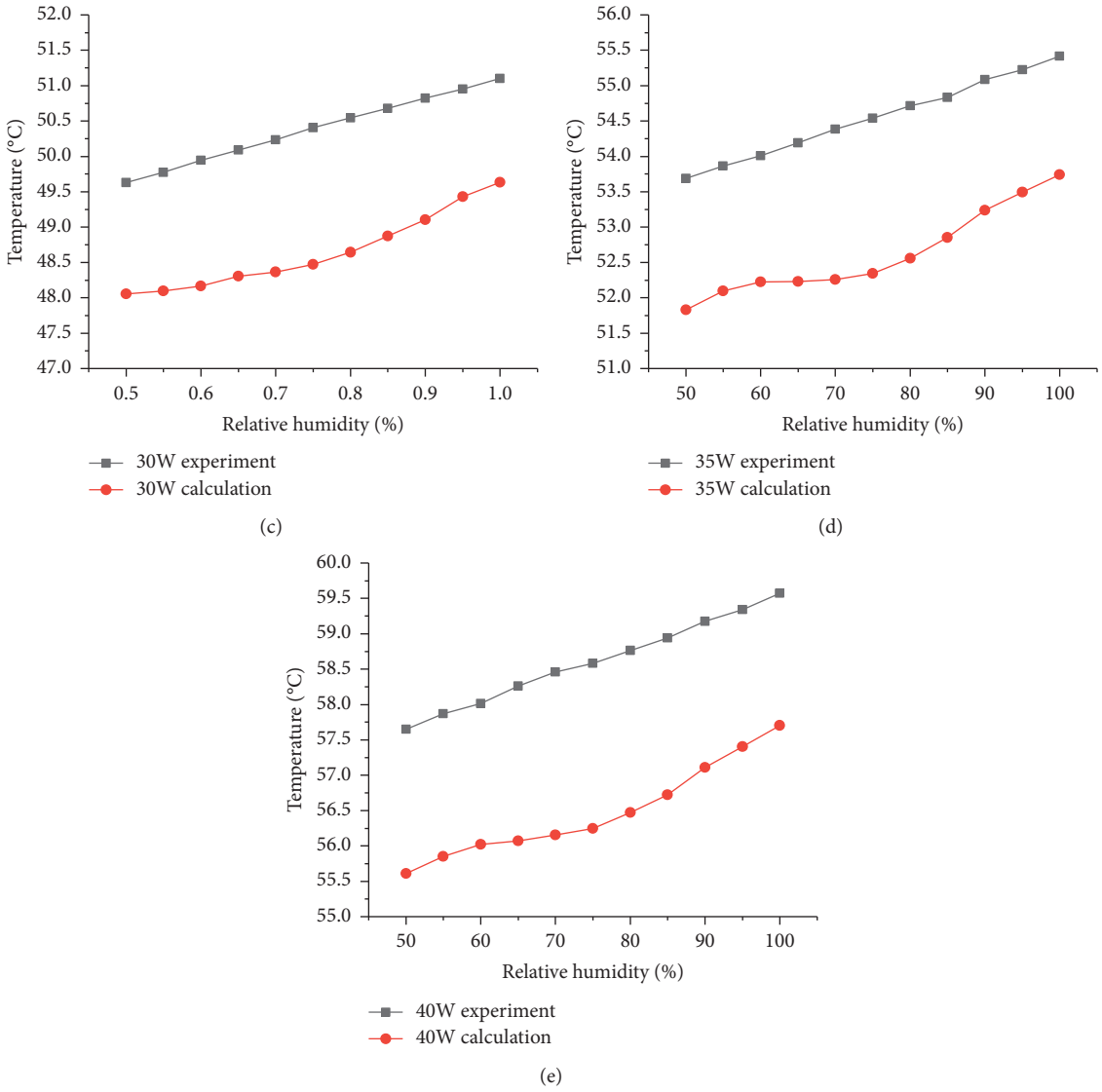


Figure 8: Calculational and experimental relation curves between relative humidity and junction temperature of 20W, 25W, 30W, 35W, and 40W LED chips.

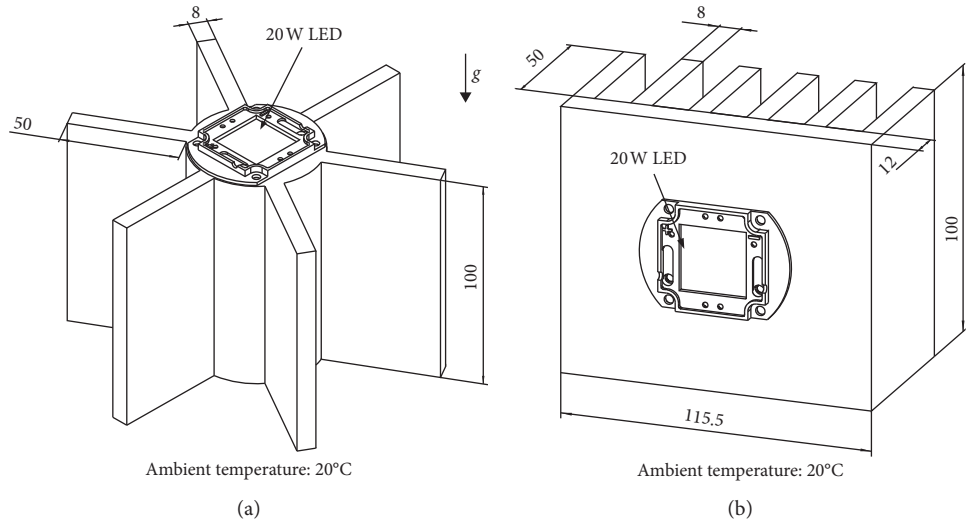


Figure 9: (a) Cylindrical heat sink and (b) rectangular plate-fin heat sink.

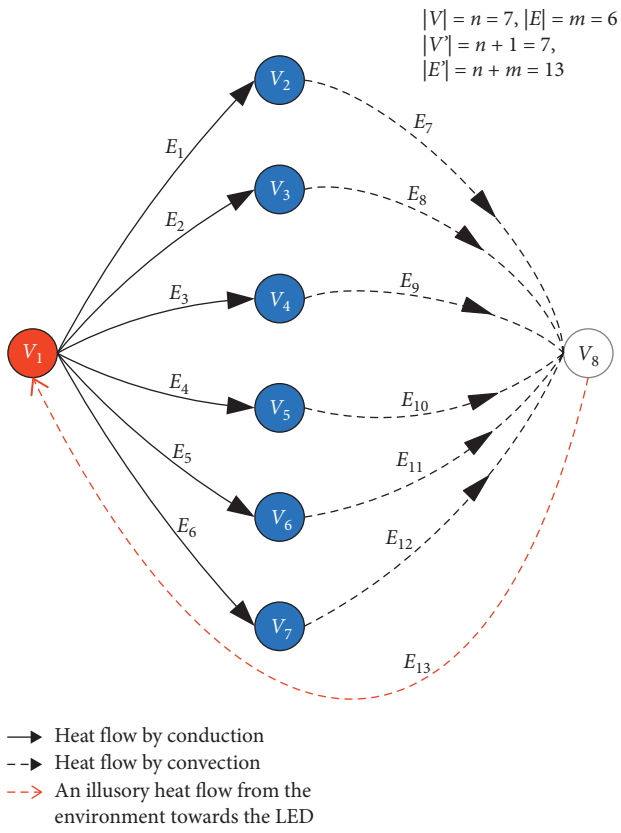


Figure 10: The digraph of the two heat sinks.

Table 3: The matrix representation of the heat flow graph.

Each matrix	The specific matrix
The adjacency matrix A_H by $n \times n$ (7×7)	$\begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$
The incidence matrix M_F by $(n+1) \times (m+n)$ (8×13)	$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & -1 & 1 \end{pmatrix}$
The flow matrix C_F by $(n-1) \times (m+n)$ (6×13)	$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$

Table 4: Experimental temperature excesses of a 20°W LED chip mounted on a cylindrical heat sink and a rectangular plate-fin heat sink against calculation values at 20°C ambient temperature.

Ambient relative humidity (%)	Calculation (°C)	Experimental temperature excesses on a rectangular heat sink (°C)	Relative errors between experiment and calculation for the rectangular heat sink (%)	Experimental temperature excesses on a cylindrical heat sink (°C)	Relative errors between experiment and calculation for the cylindrical heat sink (%)
50	22.8	24.10	5.39	22.60	0.88
55	23.1	24.23	4.66	22.71	1.72
60	23.3	24.49	4.86	22.95	1.53
65	23.5	24.87	5.50	23.30	0.87
70	24.1	25.35	4.95	23.76	1.43
75	24.8	25.97	4.49	24.35	1.85
80	25.4	26.70	4.86	25.05	1.42
85	26.3	27.56	4.58	25.85	1.76
90	27.2	28.54	4.68	26.75	1.67
95	28.3	29.62	4.47	27.76	1.94
100	29.5	30.37	2.86	28.88	2.14

$$\begin{aligned} C_1 &= \{e_1, e_7, e_{13}\}, \\ C_2 &= \{e_2, e_8, e_{13}\}, \\ C_3 &= \{e_3, e_9, e_{13}\}, \\ C_4 &= \{e_4, e_{10}, e_{13}\}, \\ C_5 &= \{e_5, e_{11}, e_{13}\}, \\ C_6 &= \{e_6, e_{12}, e_{13}\}. \end{aligned} \tag{22}$$

Table 4 shows the experimental and calculated temperature excesses between the environment and the junction of 20°W LED mounted on a cylindrical heat sink and rectangular plate-fin heat sink in the humidity range of 50%–100%. It suggests that the junction temperature of LED increases with the increase of humidity, which is consistent with the conclusion in Section 3.1.3 and the conclusions of other papers [10]. Besides, the two heat sinks with identical digraph and thermal admittance matrix performed closely, thus verifying the correctness of heat sinks with an identical heat digraph but different types having close cooling performance.

4. Conclusion

This study developed a mathematical model based on graph theory, in association with computer technology, in order to lay a foundation to evaluate the optimal design among a large number of heat sinks' fin arrangement swiftly. After analyzing the heat and mass transfer process of a single fin under moisture environment, the whole heat dissipation process transferred from LED chips into the air through a heat sink can be represented by a digraph and described by two criteria considering the conservation of energy. On the basis of previous work, the concept of improved thermal admittance considering the effect of humidity is proposed; therefore, the temperature-heat calculation equations of the whole heat sink are deduced.

Based on experimental results, the following statements are concluded: the junction temperature of 20 W, 25 W, 30W, 35 W, and 40 W LED chips mounted on the same heat sink was measured at 50%–100% relative humidity to compare with the calculation results, and the relative errors between the calculated value and the experimental data are all within 5%, thus verifying the calculation model we constructed under moisture environment.

The experimental and calculated temperature excesses between the environment and the junction of 20°W LED mounted on a cylindrical heat sink and rectangular plate-fin heat sink are in the humidity range of 50%–100%, thus verifying the correctness of heat sinks with an identical heat digraph but different types having close cooling performance.

Abbreviations

$T_{ambient}$: Ambient temperature (°C)
h: Surface heat transfer coefficient ($W \cdot m^{-2} K^{-1}$)
Nu: Nussle number
Pr: Prandtl number
l: Characteristic length (m)
v: Kinematic viscosity ($m^2 \cdot s^{-1}$)
k: Thermal conductivity ($W \cdot m^{-1} K^{-1}$)
g: Acceleration of gravity ($m \cdot s^{-2}$)
q: Thermal power (W)
θ: Temperature excess between the fin and the surrounding environment (°C)
R: Thermal resistance ($K \cdot W^{-1}$)
L: Length of the fin (m)
ω: Relative humidity (%)
H: Height of the fin (m)
δ: The thickness of the fin (m)
Gr: Grashof number
β: Volume expansion coefficient (1/K)
Le: Lewis number
α: Thermal diffusivity ($m^2 \cdot s^{-1}$)
I_{fg}: Latent heat of moisture ($J \cdot kg^{-1}$)
c_p: Specific heat of moisture ($J \cdot kg^{-1} K^{-1}$)
h_d: Mass transfer coefficient ($kg \cdot m^{-2} \cdot s^{-1}$).

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This research was supported by the Research Program of Science and Technology at Universities of Inner Mongolia Autonomous Region (NJZY20241).

References

[1] M.-H. Chang, D. Das, P. V. Varde, and M. Pecht, "Light emitting diodes reliability review," *Microelectronics Reliability*, vol. 52, no. 5, pp. 762–782, 2012.

[2] B. Sun, X. Jiang, K.-C. Yung, J. Fan, and M. G. Pecht, "A review of prognostic techniques for high-power white LEDs," *IEEE Transactions on Power Electronics*, vol. 32, no. 8, pp. 6338–6362, 2017.

[3] X. Qian, J. Zou, M. Shi et al., "Development of optical-thermal coupled model for phosphor-converted LEDs," *Frontiers of Optoelectronics*, vol. 12, no. 3, pp. 249–267, 2019.

[4] L. Sun, J. Zhu, and H. Wong, "Simulation and evaluation of the peak temperature in LED light bulb heatsink," *Microelectronics Reliability*, vol. 61, pp. 140–144, 2016.

[5] J. X. Zhu and L. X. Sun, "Mathematical model and computation of heat distribution for LED heat sink," *European Physical Journal Plus*, vol. 131, no. 5, p. 179, 2016.

[6] D. Jang, S.-J. Park, S.-J. Yook, and K.-S. Lee, "The orientation effect for cylindrical heat sinks with application to LED light bulbs," *International Journal of Heat and Mass Transfer*, vol. 71, pp. 496–502, 2014.

[7] D. Jang, D. R. Kim, and K.-S. Lee, "Correlation of cross-cut cylindrical heat sink to improve the orientation effect of LED light bulbs," *International Journal of Heat and Mass Transfer*, vol. 84, pp. 821–826, 2015.

[8] D. Jang, S.-J. Yook, and K.-S. Lee, "Optimum design of a radial heat sink with a fin-height profile for high-power LED lighting applications," *Applied Energy*, vol. 116, no. 3, pp. 260–268, 2014.

[9] B. Lv and F. Xiong, "Mathematical calculation model for temperature distribution of LED lamp heat sinks," *Bandaoti Guangdian/Semiconductor Optoelectronics*, vol. 39, no. 2, pp. 229–233, 2018.

[10] B.-X. Lyu, Y.-R. Chen, and F. Xiong, "Mathematical calculation model and its verification for temperature distribution of LED lighting's heatsinks for plant growth in the summer greenhouse," *Chinese Journal of Luminescence*, vol. 39, no. 8, pp. 1115–1122, 2018.

[11] W. A. Khan, J. R. Culham, and M. M. Yovanovich, "Optimization of pin-fin heat sinks using entropy generation minimization," *IEEE Transactions on Components and Packaging Technologies*, vol. 28, no. 2, pp. 247–254, 2005.

[12] H. E. Ahmed, "Optimization of thermal design of ribbed flat-plate fin heat sink," *Applied Thermal Engineering*, vol. 102, pp. 1422–1432, 2016.

[13] Y. Lv and S. Liu, "Topology optimization and heat dissipation performance analysis of a micro-channel heat sink," *Meccanica*, vol. 53, no. 15, pp. 3693–3708, 2018.

[14] K. Lampio and R. Karvinen, "Optimization of convectively cooled heat sinks," *Microelectronics Reliability*, vol. 79, pp. 473–479, 2017.

[15] A. Bar-Cohen, R. Bahadur, and M. Iyengar, "Least-energy optimization of air-cooled heat sinks for sustainability-theory, geometry and material selection," *Energy*, vol. 31, no. 5, pp. 579–619, 2006.

[16] G. Schmid, L.-H. Yang, T.-H. Yang, and S.-L. Chen, "Influence of inter-fin base length on thermal performance of free hanging horizontal base heat sinks," *Applied Thermal Engineering*, vol. 108, pp. 1226–1236, 2016.

[17] S.-H. Yu, K.-S. Lee, and S.-J. Yook, "Optimum design of a radial heat sink under natural convection," *International Journal of Heat and Mass Transfer*, vol. 54, no. 11-12, pp. 2499–2505, 2011.

[18] W. J. Min, S. W. Jeon, and Y. Kim, "Optimal thermal design of a horizontal fin heat sink with a modified-opening model mounted on an LED module," *Applied Thermal Engineering*, vol. 91, pp. 105–115, 2015.

[19] F. Harary, *Graph Theory*, Addison-Wesley Publishing Co, Boston, MA. USA, 2019.

[20] Y. Lu, Y. Wang, Y. Lu, and N. Ye, "Derivation of contracted graphs with ternary/quaternary links for type synthesis of parallel mechanisms by characteristic strings," *Robotica*, vol. 33, no. 3, pp. 548–562, 2015.

[21] A. D. Kraus and A. Bar-Cohen, *Design and Analysis of Heat Sinks*, Wiley, New York, NY, USA, 1995.

[22] M. H. Sharqawy and S. M. Zubair, "Efficiency and optimization of a straight rectangular fin with combined heat and mass transfer," *Heat Transfer Engineering*, vol. 29, no. 12, pp. 1018–1026, 2008.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 1730946, 11 pages
https://doi.org/10.1155/2020/1730946

Review Article

Intelligent Operation of Wheel Loader Based on Electrohydraulic Proportional Control

Bing-wei Cao, Xin-hui Liu, Wei Chen, Peng Tan, and Ping-fang Niu

School of Mechanical and Aerospace Engineering, Jilin University, Changchun 130022, China

Correspondence should be addressed to Wei Chen; chenwei_1979@jlu.edu.cn

Received 23 November 2019; Accepted 30 March 2020; Published 25 April 2020

Guest Editor: Mohamed Nayel

In this paper, the wheel loader with electrohydraulic proportional control technology is used as the carrier. For the first time, the dual-angle sensor is used for intelligent operation, which allows the wheel loader working device to be precisely controlled. First, the theoretical analysis of the electrohydraulic proportional control technology on the wheel loader studied in this paper is carried out. Next, according to the feedback of the boom and bucket angle sensor signals, the electrohydraulic proportional control technology is used to initially realize the boom memory and bucket automatic levelling function of the wheel loader working device. Finally, the data acquisition equipment is connected to provide experimental verification, although the test results did not achieve precise control of the working device. After analysis, the detected problems were solved by constructing a neural network algorithm model, which successfully realizes the intelligent and precise operation of the wheel loader, reducing unnecessary energy loss.

1. Introduction

As the main type of groundwork construction equipment in a project, the wheel loader has the advantages of fast operational speed, high efficiency, good manoeuvrability, and light operation [1]. The application of electrohydraulic proportional control technology further enhances the intelligence of the wheel loader operation, reduces the driver's operational intensity, and provides an application platform for other advanced technologies [2–4].

Electrohydraulic proportional control technology has been successfully applied in wheel loader power, transmission, hydraulic, cooling, and other systems. In recent years, there have been many studies on electrohydraulic proportional control [5–8]. Bing et al. used the electrohydraulic load sensitive method of variable pressure margin to control the displacement of variable pump and realized the energy-saving control of the combined control system of pump and valve [9]. Yongling studied the defects of poor stability and low control precision, in real operational conditions, of the electromechanical liquid proportional control system. The electrohydraulic proportional control method, based on PWM technology, is proposed [10]. Based on the working principle of electrohydraulic proportional directional valve, Zhang et al. tested the electrohydraulic proportional position control approach. They then built a valve control cylinder test bench, realizing precise control of hydraulic cylinder position [11]. Fang et al. started with the classic PID control method, along with fuzzy control approach, and adopted the intelligent fuzzy-PID collaborative control strategy, achieving higher system control precision [12]. Although the electrohydraulic proportional control technology has been successfully applied to the wheel loader, it has not been found that the angle sensor is used as a feedback signal to realize the boom memory and the bucket automatic levelling function of the wheel loader. The realization of the above functions can assist the driver to complete the operation of the wheel loader, that is, to realize intelligent operation; this is also the innovation of this paper. The application of the above proportional control technology provides a platform for implementing the function in this paper.

The wheel loader, studied in this paper, successfully applied the electrohydraulic proportional control

technology to realize the control of the working hydraulic system. However, when the boom memory and the bucket automatic levelling function were realized, the control precision proved to be poor and did not reach the expected target. Therefore, this paper successfully applied the neural network algorithm to the intelligent operation of the wheel loader and achieved the precise control of the working device.

Neural networks have infinite approximation ability to nonlinear models, while they have been widely used as tools for building nonlinear models [13–15]. By adjusting the weight of the connected network, any input can be made to obtain the desired output [16–18]. Xu and Mai designed and utilized the supervised neural network learning algorithm to deal with the problem of low precision and system instability during the operation of hydraulic material testing machine. The study results show that the neural network control optimized electrohydraulic position servo system demonstrates better stability and robustness [19]. Ni et al. combined the BP network with the traditional PID controller to monitor the actual value and set and error values of the slider at different times in real time while using the online self-learning ability of the neural network to realize the optimal combination of PID parameters. Accurate slider displacement control and simulation curves also prove that BP neural network is very effective in improving hydraulic servo control [20]. Gao and Han, in order to solve the coupling problem of hydraulic four-legged robot joints, established a single-legged mechanism for the hydraulic four-legged robot. In this work, based on the multivariable decoupling theory, the neural network (NN) model reference decoupling controller is designed [21]. The successful application of the neural network algorithm described above in the hydraulic system provides a prerequisite for the intelligent operation of the wheel loader working device.

The wheel loader, studied in this paper, uses electrohydraulic proportional control technology as an application platform. For the first time, the dual-angle sensor is used, combined with the typical V-type working cycle of the wheel loader, to realize the boom memory and bucket automatic levelling function of the working device. In addition, theoretical analysis of the electrohydraulic proportional control technology is used in the wheel loader. Signals, such as cylinder pressure, sensor angle, handle voltage, and engine speed are acquired and recorded, through the connected data acquisition device. The data curve shows that the working device is not accurately controlled, while the actual stopping angle of the boom is always greater than the memory angle of the hydraulic system controller. After analyzing the reasons, the neural network algorithm is applied to optimize, followed by a test verification of the results. The experiment showed that the neural network algorithm can be used for intelligent and precise control of the wheel loader working device, reducing unnecessary energy loss.

2. Electrohydraulic Proportional Control Technology

Figure 1 shows the evolution of the control mode of the wheel loader working device. Figure 1(a) shows the most primitive form of control. The spool movement is controlled by a flexible shaft. Figure 1(b) shows the movement of the spool controlled by the pilot oil pressure. The handle is equivalent to a pressure reducing valve, and the pilot pressure corresponds to the displacement of the spool. Figure 1(c) shows the control of spool movement through an electromagnetic proportional pressure reducing valve, currently the most advanced type of control. The wheel loader designed in this paper also adopts this control method.

The principle of electrohydraulic proportional control technology applied to this wheel loader is as follows: the control handle is used as a signal generator to supply the required controlled voltage to the controller. Then, it performs a corresponding operation and supplies output current to the electromagnetic proportional pressure reducing valve, according to the front-end signal and completes control of the hydraulic cylinder of the working device after the main reversing valve. Figure 2 shows the schematic diagram of the electrohydraulic proportional control hydraulic system of the working device.

The control handle is divided into a boom handle and a bucket handle. The handle is single axis and is equipped with an electromagnetic stop function. The structure is a spring-returned, dual-sensor device, providing an output voltage of 0.5~4.5 V. The electrical characteristics of the handle are shown in Figure 3 [22, 23].

$$U = K_l \cdot \theta, \tag{1}$$

where $K_l$is the handle scale factor and θ is the handle action angle, with a range of $-20^\circ \le \theta \le 20^\circ$.

After the analogy voltage signal output, by the handle, the voltage signal is input to the hydraulic system controller, and the controller outputs the corresponding current signal to the electromagnetic proportional pressure reducing valve according to the voltage signal, where the output is I [24]:

$$I = K_a U, \tag{2}$$

where K_a is the voltage-current gain of the controller, which is set as 0.215.

The main control valve uses electrohydraulic remote control, while the proportional pressure reducing valve is integrated with the main valve. After receiving the current signal, the electromagnetic proportional pressure reducing valve on the main valve realizes the control of the direction and flow of the main control oil passage and then controls the movement direction and speed of the wheel loader working device. The turn-on current is 300 mA, and the final current is 500 mA. The main control valve uses an internal oil circuit in parallel form, which can realize the joint operation of the boom and the bucket.

Since the conventional wheel loader does not have a hydraulic system controller, unable to accept the signal from the angle sensor, the wheel loader operation needs to be

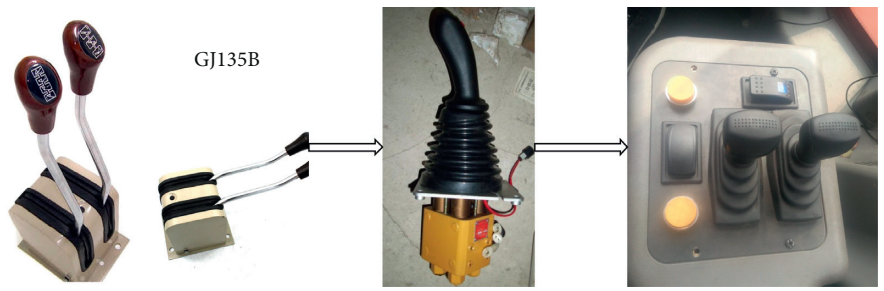


Figure 1: The evolution of control mode of the wheel loader. (a) Flexible shaft. (b) Hydraulic control. (c) Electrohydraulic proportional control.

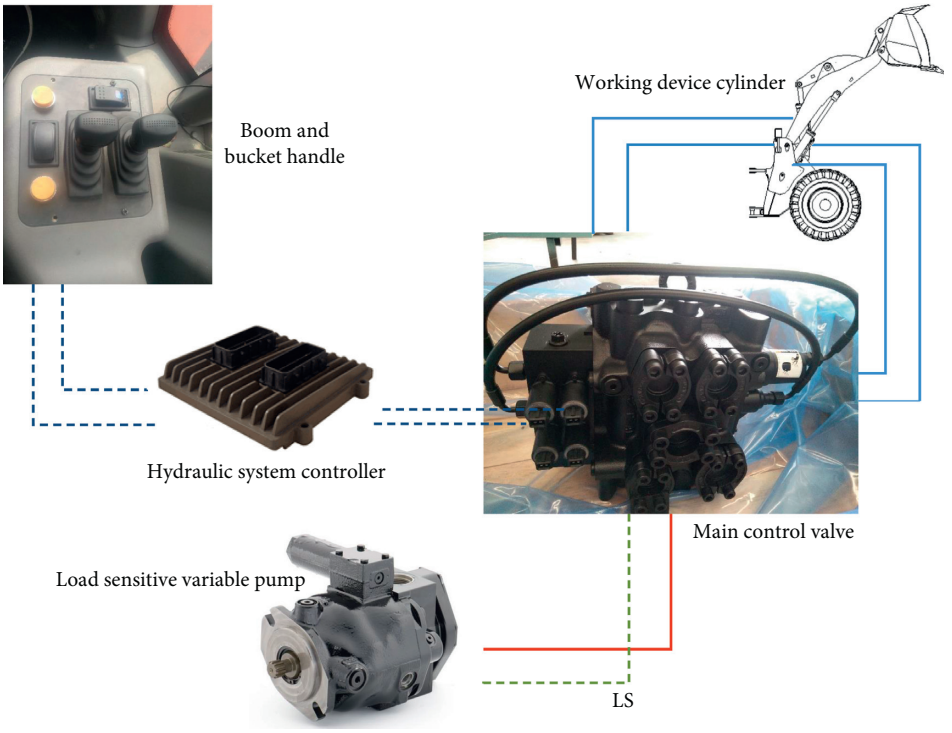


Figure 2: Electrohydraulic proportional control hydraulic system.

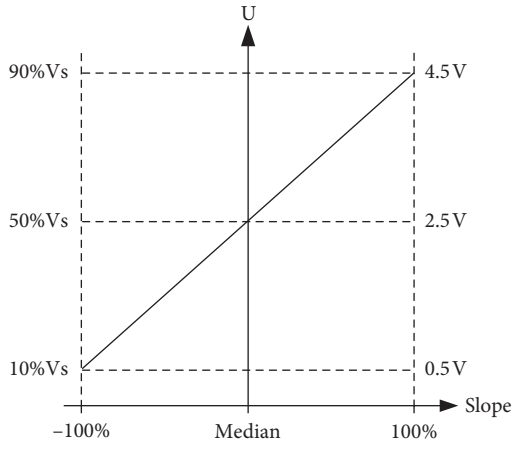


Figure 3: Handle electrical characteristics.

manually controlled by the driver, so the boom memory and the bucket automatic levelling function cannot be realized. The most used switch is the limit switch, but the limit switch can only remember one position. Real-time monitoring of the angle of the loader's working device can be achieved by installing an angle sensor in the appropriate position, combined with a hydraulic system controller. The following describes the intelligent working mechanism of the wheel loader.

3. Intelligent Operation

The V-type working mode of the wheel loader has the characteristics of wide adaptability and short working cycle time [25]. The working cycle of the V-type working mode is shown in Figure 4 [26, 27]. The intelligent operation of the so-called

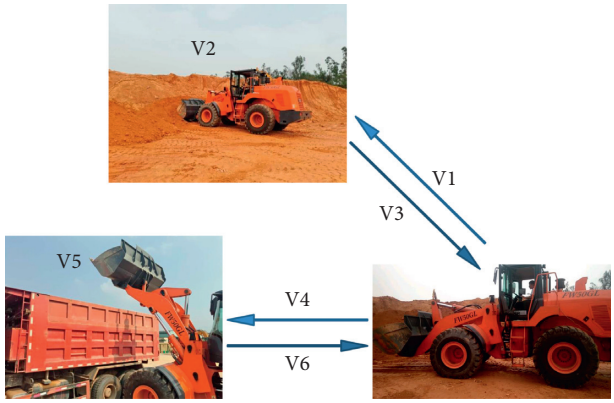


Figure 4: Wheel loader V-type work cycle.

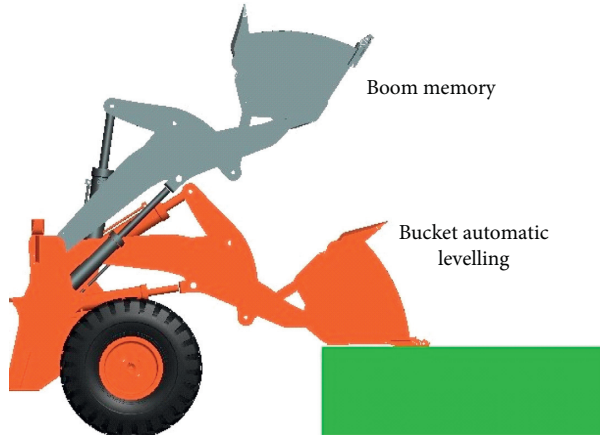


Figure 5: Function diagram.

wheel loader working device is to assist the driver to complete related operations through the hydraulic system controller, thereby reducing the labour intensity of the driver. The following describes the intelligent operation of the working device in detail with the V-shaped cycle of the wheel loader.

During the V3 and V4, the driver needs to operate the boom handle to raise the working device to the discharge height. When the discharge position is reached, the driver operates the bucket handle to perform the unloading operation. After unloading, the bucket needs to be recovered to prevent it from colliding with the truck box and be ready for the next shovel operation. During the V6 and V1, the driver operates the boom handle to lower the boom. When the boom is lowered, the bucket needs to be adjusted to a flat position, which is beneficial for the shovel loading operation.

During a V-type cycle operation, since the type of the truck does not have a large gap, the height reached by the wheel loader is constant each time. In order to facilitate the next shovel loading operation, the bucket should be levelled, when it falls on the ground. Therefore, in the entire V-shaped cycle, the driver needs to frequently operate the boom and the bucket handle to adjust the posture of the working device [25]. In order to reduce the driver's operation involvement, this research group and a company worked together to develop a 5*t* wheel loader, as a prototype. The function diagram of the boom memory and the bucket automatic levelling is shown in Figure 5. For the first time, two angle sensors are used to measure the angle of the boom and bucket in real time to realize the boom memory and bucket automatic levelling function of the wheel loader working device. The functional flowchart is shown in Figure 6. The handle outputs a voltage signal to the hydraulic system controller, which ranges from 0.5 volts to 4.5 volts. In the hydraulic system controller, the voltage signal is converted to the percentage of handle opening. 0.5~2.5 corresponds to 0~100%, and 2.5~4.5 corresponds to −100~0%. The control current of the proportional solenoid's valve is 300~730 mA. It is one-to-one correspondence between the percentage of signal and control current signal. Therefore, the handle opening signal is converted into current signal in the hydraulic controller.

As the wheel loader performs the loading operation, after the boom is lifted to the discharge height, the driver presses the boom memory switch, so that the hydraulic system controller saves the angle of the boom angle sensor. At this boom angle, there is a unique bucket angle value, corresponding to the automatic levelling function of the bucket. The automatic levelling angle of the bucket corresponding to different boom heights is different, but the two angle values are one-to-one correspondence. The aforementioned angle is input to the hydraulic system controller, as data in a table, derived from the ADAMS model of the working device. Next, the driver operates the bucket handle to unload the material, and the bucket is required to retract after unloading. Currently, the bucket handle is in the suction state under the action of electromagnet. When the angle monitored by the bucket angle sensor corresponds to the angle in the data table, the electromagnet of the bucket handle is powered off. The handle is in the middle position. The bucket is levelled, when the working device is placed on the ground. After the hydraulic system controller has memorized the boom angle, next time the boom is lifted, the electromagnet of the boom handle is in the suction state. As the boom memory angle is reached, the electromagnet is powered off and the handle is in the middle position. At this time, the height of the working device is the same as the height set by the driver. In the above operation, under the action of the hydraulic system controller, the boom memory of the wheel loader and the automatic levelling function of the bucket are realized, which reduces the operation intensity of the driver.

According to the above analysis, the combination of the angle sensor and the electrohydraulic proportional control technology can assist the driver to complete the boom raising and the bucket levelling operation. The function described above is the intelligent operation of the wheel loader; the following is an experiment to verify the intelligent operation of the wheel loader.

4. Test Verification

This section describes the verification of the loading boom memory and the bucket automatic levelling functions, by connecting the data acquisition device. The test data to be collected include the boom and bucket handle voltage

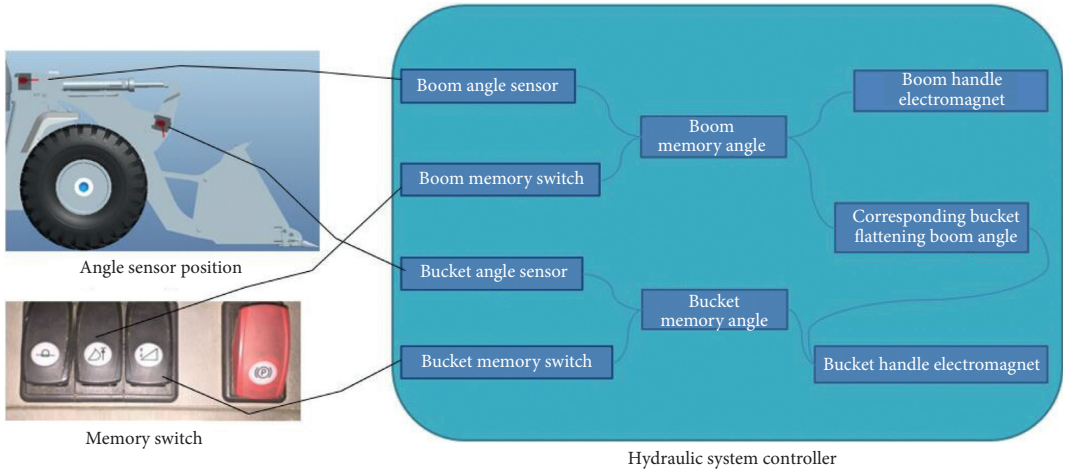


Figure 6: Function flowchart.

signals; the boom and bucket pilot pressure signals; the boom and bucket angle sensor signal; and engine speed signal. The test equipment is shown in Figure 7.

After the memory boom angle is 45°, the wheel loader is started, and the boom height memory function is realized by the hydraulic system controller. The engine speed ranges from 800 r/min to 2200 r/min, while the angle value, during the stop position of the boom, is recorded every 200 r/min, as shown in Table 1. The actual stop angle curve of the boom, during the verification, is shown in Figure 8.

Figure 8 shows that the actual stop angle of the aforementioned boom becomes larger, as the engine speed increases, while at the maximum speed, it reaches 51.58°, which far exceeds the value (45°) memorized by the hydraulic system controller. The same is true for the bucket angle sensor feedback signal. The above phenomenon causes the actual stop angle of the boom to be higher than the memory angle. The bucket recovery angle is also greater than the memory angle of the hydraulic system controller. Real-life boom memory and bucket automatic levelling cannot be realized. Since the actual stop angle is always greater than the memory angle, this will result in loss of certain amount of power. This angle difference is also meaningless because the memory angle has reached the driver's needs.

During the test, it was also found that as the memory angle in the hydraulic system controller increases, the actual stop angle of the boom also increases. The reciprocating boom memory function is performed at different angles, and the actual stopping angle of the boom at different speeds is measured as shown in Table 2.

The above tests were carried out under no-load conditions, and it was impossible to carry out the no-load operation when the wheel loader was working. Therefore, the load factor must also be considered. The boom memory function tests with loads of $3t$ and $5t$ were also performed, while the actual stop angle of the boom was obtained, as shown in Table 3.

As can be seen from Tables 1–3, the data in the table are the actual stopping angle of the boom obtained according to different engine speeds and different memory angles. Comparing Tables 1–3, the actual stop angle of the boom when there is load is not much different from the actual stop angle of the boom when there is no load, so the influence of the load on the memory function of the boom can be ignored.

The reasons for the above phenomena are analyzed below. The boom handle voltage signal, pilot pressure signal, and boom angle signal are recorded, as shown in Figure 9. The illustrated graph shows that when the hydraulic system controller cuts off the voltage signal of the boom handle, its value fluctuates, due to the spring inside the handle. As the voltage oscillates, the pilot pressure is not cut off in time, so the main control valve is not closed in time. This is the main reason that causes the boom stop angle to be greater than the hydraulic system controller angle value. As the rotational speed increases, the flow rate, provided by the variable pump, continues to rise, causing the flow, through the main valve, to increase per time unit, leading to the difference increase, as the engine speed rises.

In theory, the dead zone area of the boom handle and the bucket handle should be increased in the hydraulic system controller. However, if the value is too large, the driver's operating experience will be affected, which is not conductive to microaction. Therefore, an optimization algorithm should be considered to realize the function of the boom memory and the bucket automatic levelling.

5. Algorithm Optimization

Neural network algorithms have successfully solved many practical problems that are difficult to solve in an intelligent way, in the fields of pattern recognition, automatic control, predictive evaluation, etc [28–31]. In this case, a large amount of data has been obtained from the actual test and input into the neural network algorithm model, while data other than the test data can be predicted, thereby realizing intelligent and precise control of the wheel loader working device. Through experiments on different engine speeds and memory angles, the actual stopping angle of the boom can be obtained. Speed and memory angle are used as the input of

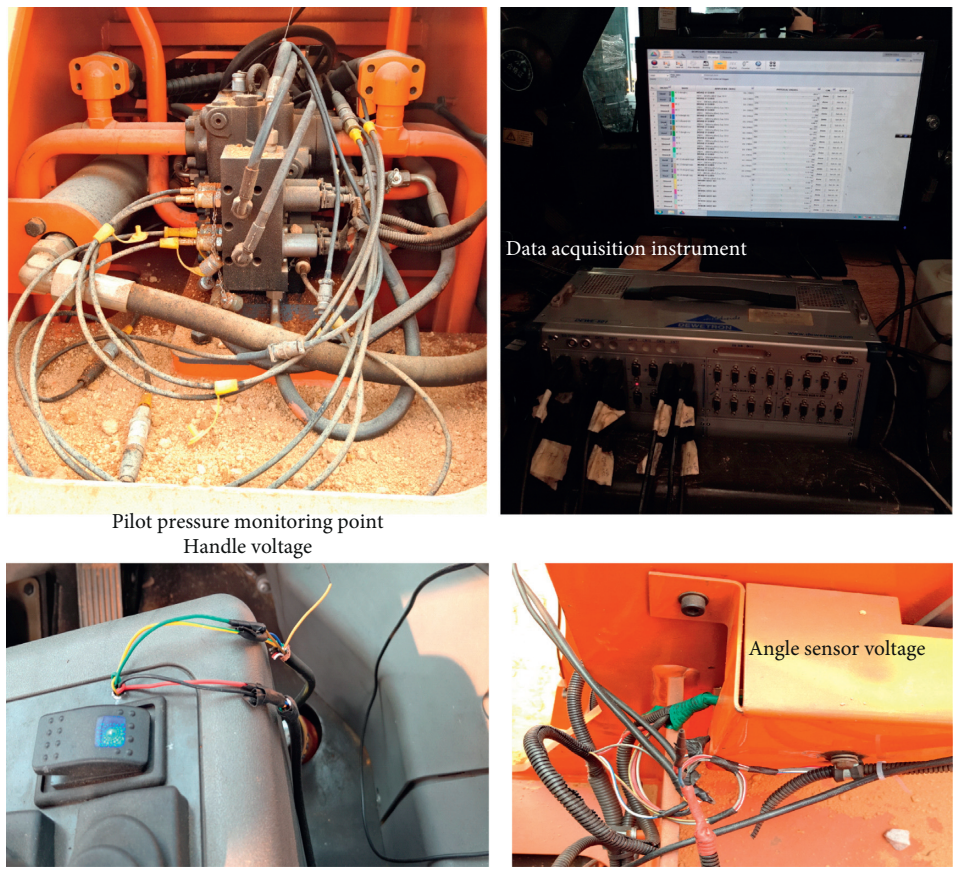


Figure 7: Test equipment.

Table 1: Actual stop angle at memory boom angle of 45°.

Engine speed	800 r/min	1000 r/min	1200 r/min	1400 r/min	1600 r/min	1800 r/min	2000 r/min	2200 r/min
Memory angle 45°	47.42	48.15	48.66	49.17	49.73	50.11	50.56	51.38

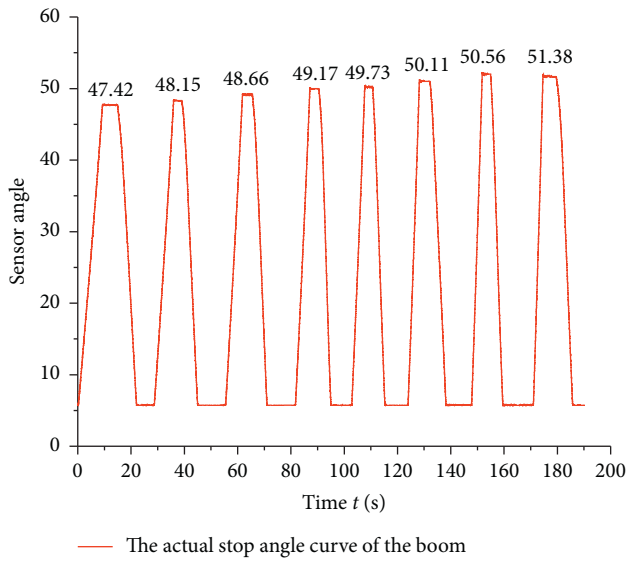


Figure 8: Actual stop angle of the boom.

Table 2: Verifying different boom memory angles.

	30°	35°	40°	50°	60°	75°
800 r/min	32.94	38.02	42.92	53.17	63.4525	79.2525
1000 r/min	33.32	39.04	43.42	54.13	64.085	80.3725
1200 r/min	34.08	39.65	43.92	54.91	65.015	81.4425
1400 r/min	34.74	40.31	44.76	55.13	65.7425	82.4125
1600 r/min	35.42	40.87	45.31	55.92	66.4075	83.275
1800 r/min	35.88	41.03	46.04	56.34	67.175	84.17
2000 r/min	36.21	41.65	46.58	57.21	68.06	85.09
2200 r/min	36.54	42.15	46.76	57.87	68.94	85.09

the neural network model, and the difference between the actual stop angle and the memory angle is used as the output of the neural network model. The combination of speed and memory angle is used as input vector. There are 56 groups, of which 48 groups are used as the training set and the remaining 8 groups are used as the test set. The number of hidden layer is 16, and the learning rate is 0.12. The following steps will analyze the selection of the above parameters. The neural network model prediction data are input to the hydraulic system controller as shown in Figure 10.

Table 3: Verification of the boom memory function at different loads.

	$3t$~45°	$5t$~45°	$3t$~60°	$5t$~60°	$3t$~75°	$5t$~75°
800 r/min	47.67	47.42	63.27	63.08	79.01	79.64
1000 r/min	48.33	48.13	63.5	63.79	79.45	80.92
1200 r/min	49.16	48.63	65.02	64.61	81.12	81.42
1400 r/min	49.73	49.74	65.63	65.6	82.12	82.52
1600 r/min	50.26	50.56	66.23	66.19	82.95	83.14
1800 r/min	51	51.36	67.16	66.85	83.51	84.08
2000 r/min	51.5	51.58	68.63	67.61	84.62	85.09
2200 r/min	51.79	51.8	69.89	68.11	85.09	85.12

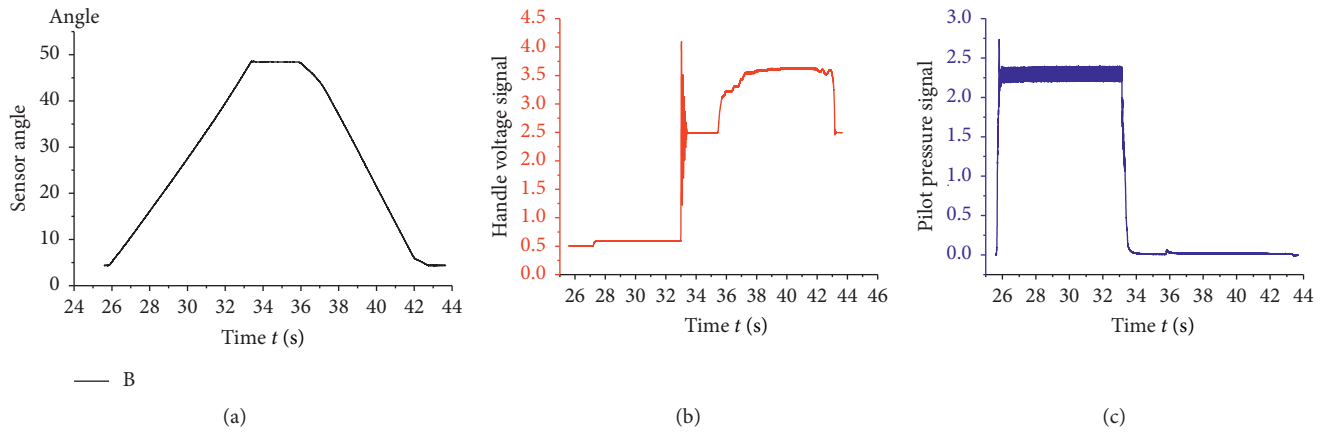


Figure 9: Verification curve.

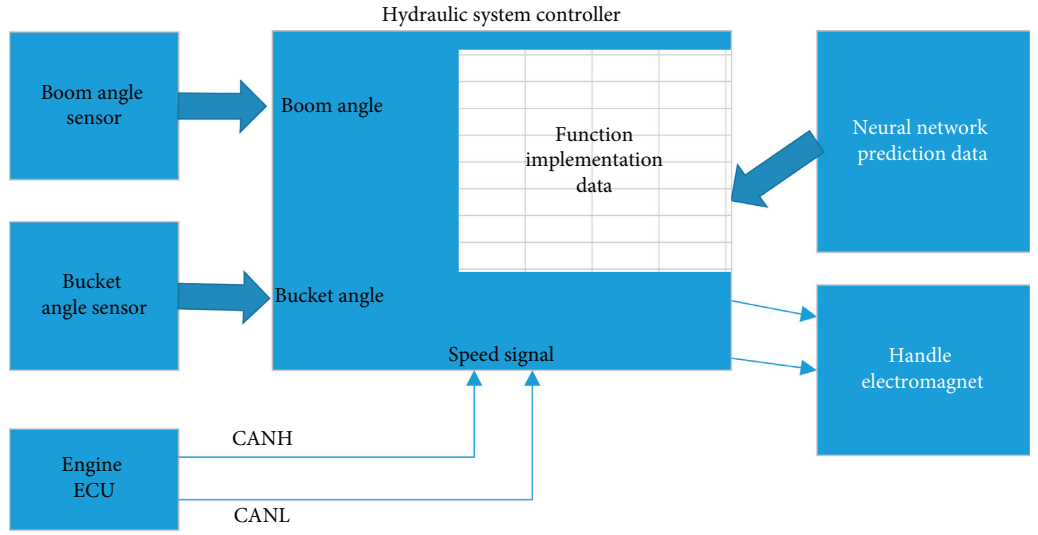


Figure 10: Hydraulic system controller data model.

The error (different speed and memory value) can be predicted by the neural network prediction model, and the above error data are input into the hydraulic system controller in the form of a table. When the driver presses the memory switch, the engine speed can be monitored via the CAN line. According to the above setting conditions, the hydraulic system controller can accurately control the actual stopping angle of the boom by subtracting the error value from the stored angle value.

A detailed description of the neural network algorithm model is as follows:

Step 1: importing training data [32]. By loading the boom memory test function, the difference between the actual stop angle of the boom and the memory angle of the hydraulic controller can be obtained, as shown in Table 4. Table 4 shows the difference between the boom actual stop angle and the memory angle with different speeds and different memory angles, which is used as the output of the neural network model. The engine speed and the memory values are taken as inputs, while the difference between the memory value and the boom actual value is taken as output.

Step 2: initialization processing. After creating the training set and test set, the data are normalized by using the mapminmax function (normalization function in MATLAB, preprocessing the data) [33].

Step 3: selection of number of hidden nodes and learning rate [34]. The above parameters are determined by calculating the influence of different hidden nodes and learning rates on the training error. Figure 11 shows the effect of different hidden nodes on the training error, while Figure 12 shows the effect of learning rate on the training error. The number of hidden nodes selected is 16 and the selected learning rate is 0.12.

Step 4: building a neural network [35]. The number of trainings is 1000, and the accuracy is 0.000000001; the prediction is performed after training, while the anti-normalization process is also realized. The curve of the training error vs the number of training times is shown in Figure 13.

As can be seen from Figure 13, after the neural network model was trained 25 times, the error value was minimized.

The above operation concludes the construction of the neural network model, while the prediction of data, other than the test, follows, along with the required test and verification. The verification curve is shown in Figure 14, when the memory angle is 45°, while Figure 15 illustrates the same curve, when the memory angle is 60°.

The above test curves show specifically that at memory angle of 45°, the actual stop angle of the boom is 45.29°, at 800 r/min, and 44.90°, at 1100 r/min, while at memory angle of 60°, the actual stop angle of the boom is 60.54°, at 1100 r/min, and 59.32°, at 1900 r/min.

The engine speed is changed during the loader operation. The engine speed test is used to verify the correctness of the neural network model. The obtained test curve is shown in Figure 16.

As can be seen from the Figure 16, the actual boom angle is 59.03° (the biggest difference from the memory angle) and the maximum error rate is $((60 - 59.03)/60) \times 100\% = 1.61\%$, which meets the requirements of the boom memory function. Considering the measurement error of the sensor, the neural network model can meet the requirements of the boom memory function.

Table 4: Input and output data.

	30°	35°	40°	45°	50°	60°	75°
800 r/min	2.94	3.02	2.92	2.625	3.17	3.4525	4.2525
1000 r/min	3.32	4.04	3.42	3.425	4.13	4.085	5.3725
1200 r/min	4.08	4.65	3.92	4.095	4.91	5.015	6.4425
1400 r/min	4.74	5.31	4.76	4.87	5.13	5.7425	7.4125
1600 r/min	5.42	5.87	5.31	5.4825	5.92	6.4075	8.275
1800 r/min	5.88	6.03	6.04	6.2125	6.34	7.175	9.17
2000 r/min	6.21	6.65	6.58	6.7025	7.21	8.06	10.09
2200 r/min	6.54	7.15	6.76	7.1	7.87	8.94	10.09

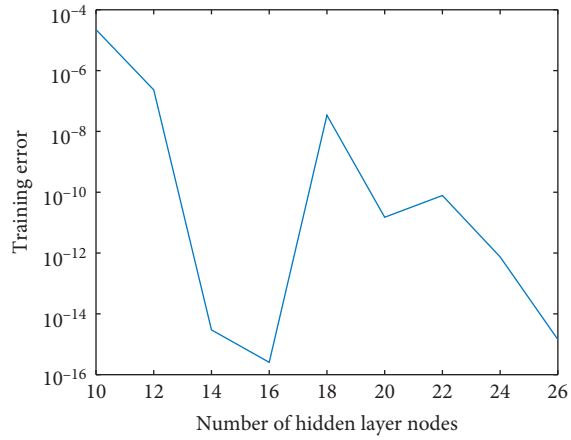


Figure 11: Effect of different hidden nodes on training error.

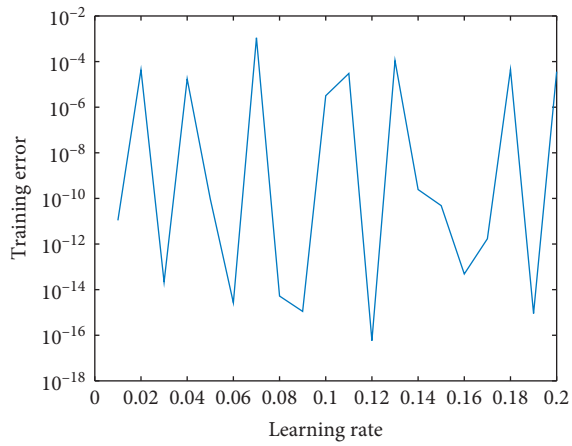


Figure 12: Effect of learning rate on training error.

Taking the engine speed of 800 r/min and the memory angle of 45° as an example, when the accurate control is not achieved, the actual stopping angle of the boom can reach 47.42°. However, after the error is predicted by the neural network algorithm, the actual stopping angle of the boom is 45.29°. The angle is reduced by 2.13°, which means that the boom will move less for a distance each time, which saves the energy.

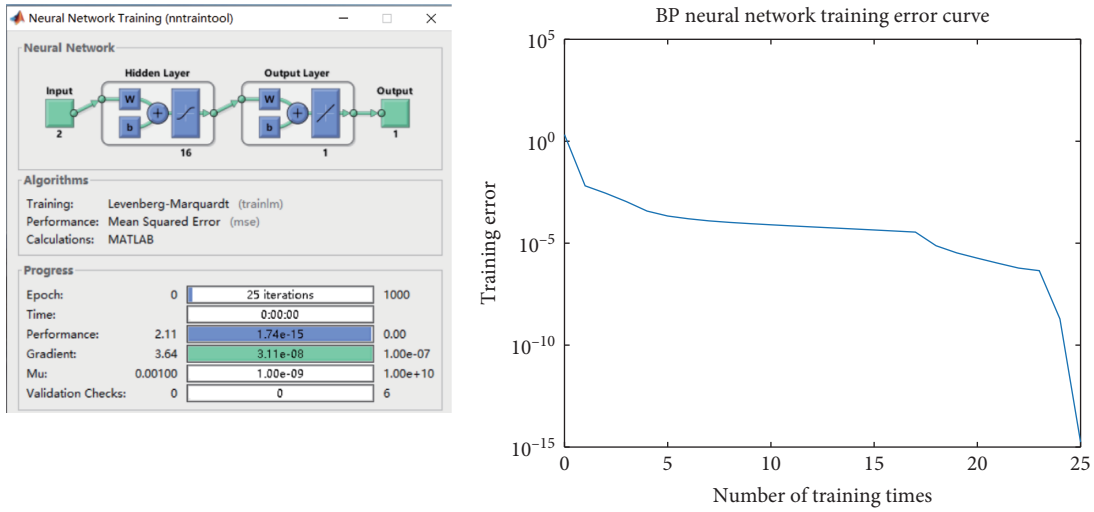


Figure 13: Training error vs number of training times.

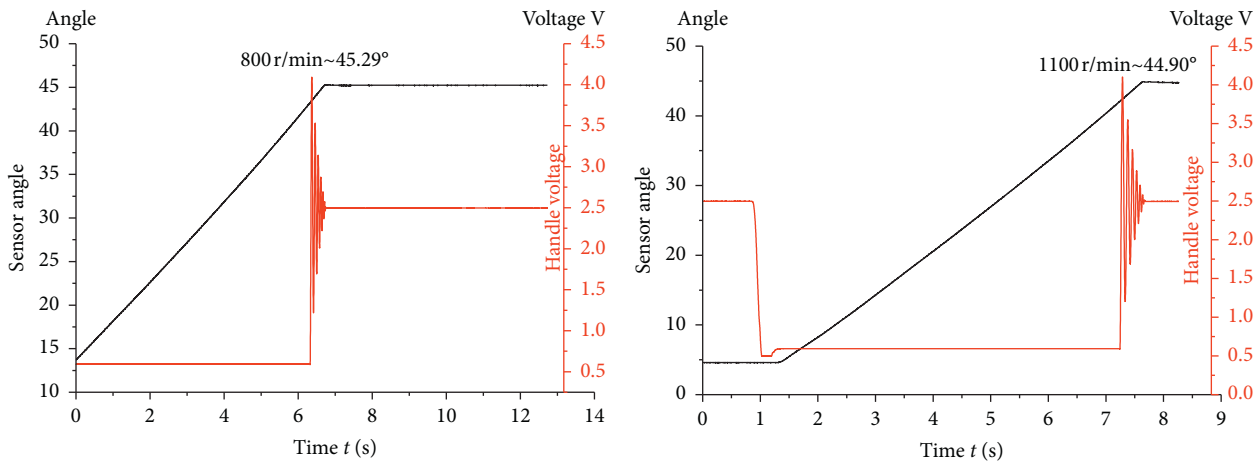


Figure 14: Verification curve at memory angle of 45°.

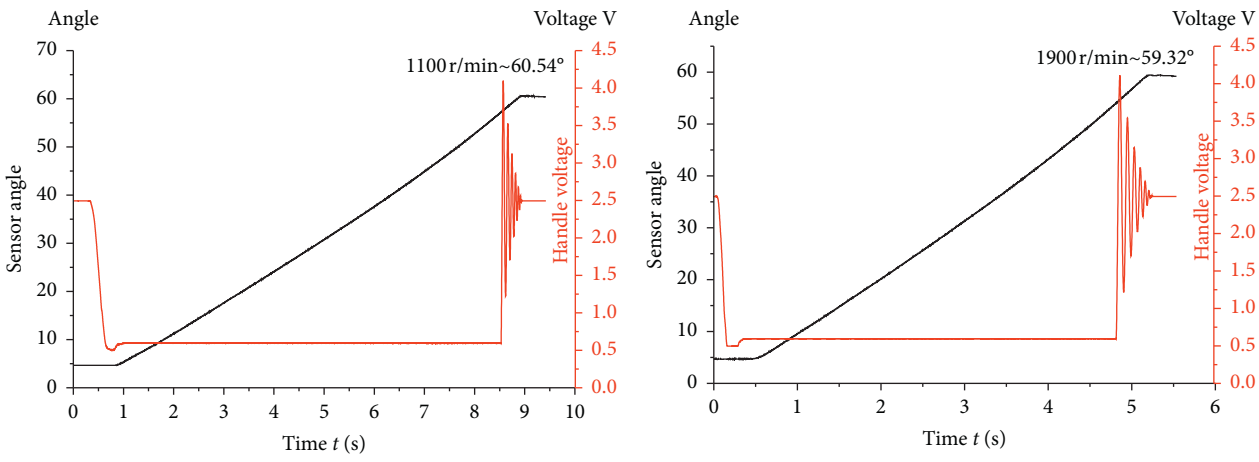


Figure 15: Verification curve at memory angle of 60°.

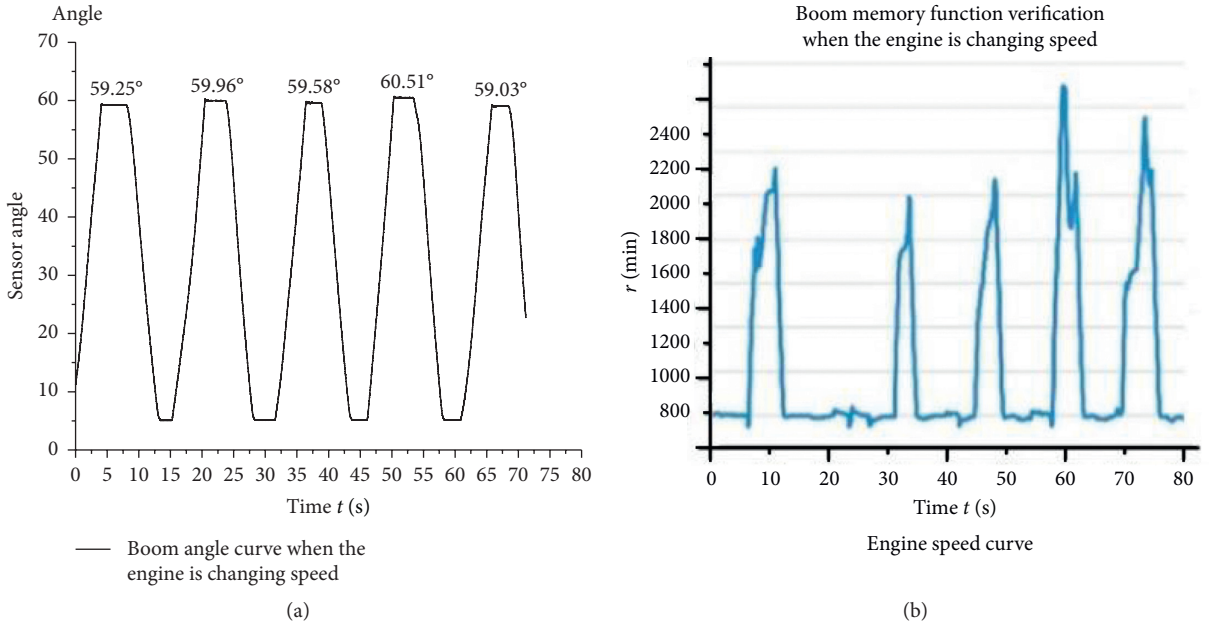


Figure 16: Verification curve at memory angle of 60°.

6. Conclusions

As an innovative point of this paper, the paper firstly proposes the combination of angle sensor and electrohydraulic proportional control technology and initially realizes the boom memory and bucket automatic levelling function of the wheel loader working device. However, subsequent tests have found that the actual stopping angle of the working device is always greater than the working angle of the hydraulic system controller. Then, the neural network algorithm is introduced, and the error value is predicted and input to the hydraulic system controller to achieve precise control of the above functions.

The research content of this paper can be summarized as follows:

(i) A typical V-type operating cycle of the wheel loader is introduced. In order to reduce the driver's operation intensity, the dual-angle sensor is used for real-time monitoring of the boom and bucket angle. The electrohydraulic proportional control technology is used as the carrier to feed the angle sensor signal to the hydraulic system controller. The electro-hydraulic proportional control technology is used as the carrier to feed the angle sensor signal to the hydraulic system controller to realize the boom memory and bucket levelling functions of the wheel loader working device.

(ii) After connecting the data acquisition equipment, the handle voltage, pilot pressure, cylinder pressure, angle sensor, and engine speed signal were collected. The curve analysis showed that the boom memory and the bucket level were not accurately controlled. According to the analysis of the test data, the actual stop angle of the boom increases, as the engine speed rises, while it increases with the memory value, but it is independent of the load.

(iii) The data acquisition instrument records real test data, which is used as the raw data of the neural network. After constructing the neural network algorithm model, the load manoeuvring arm memory and the bucket levelling function can be optimized to predict data beyond the test. The above data were imported into the hydraulic system controller, as the boom memory function was again verified. The maximum error rate was 1.61%. The result proves that the optimized model can achieve precise control of the wheel loader intelligent operation.

In the subsequent intelligent operation of the wheel loader, it may be considered to add a radar to measure the height of the truck. The height of the truck is sent to the hydraulic system controller as an input signal and becomes the memory angle of the boom. This eliminates the need to manually store the boom height through the memory switch, which further improves the intelligent operation level of the wheel loader.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was funded by the National Key Research and Development Program of China under grant no. 2018YFB2000900.

Supplementary Materials

Wheel loader data. (*Supplementary Materials*)

References

[1] X. Yan, J. Yang, and L. Quan, "Co-simulation and experiment of wheel loader during operation process," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 31, no. 16, pp. 102–109, 2015.

[2] X. Wang and S. M. Disney, "Mitigating variance amplification under stochastic lead-time: the proportional control approach," *European Journal of Operational Research*, vol. 256, no. 1, pp. 151–162, 2017.

[3] Y. Yang, L. Xu, and D. Ren, "Design of electro-hydraulic proportional control system for excavators," *Machine Tools and Hydraulics*, vol. 40, no. 22, pp. 82–84, 2012.

[4] J. Ye, "Application research of electro-hydraulic proportional control valve in construction machinery," *Development & Innovation of Machinery &electrical Products*, vol. 20, no. 4, pp. 3–5, 2007.

[5] Z. Huang, "Simple electro-proportional control of the A4VSO," *Fluid Power Transmission and Control*, vol. 4, pp. 26–29, 2016.

[6] F. Zhu, J. Lan, and B. Fu, "Simulation and analysis for electro-hydraulic proportional control system of 8 MN hydraulic press," *Chinese Hydraulics and Pneumatics*, vol. 5, pp. 79–82, 2018.

[7] lifu Yang, "Wheel loader working device electro-hydraulic proportional control system," *Construction Machinery*, vol. 13, no. 4, pp. 1–4, 2002.

[8] Y. J. Park, J.-Y. Oh, U.-K. Yun et al., "Design parameter analysis of the proportional control valve for wheel-wheel loader automatic transmission," *Journal of the Korean Society for Precision Engineering*, vol. 27, no. 5, pp. 27–35, 2010.

[9] X. Bing, R. Ding, and J. Zhang, "Experiment research on individual metering system of mobile machinery based on coordinate control of pump and valves," *Journal of Zhejiang University (Engineering Science)*, vol. 49, no. 1, pp. 93–101, 2015.

[10] K. Yongling, "Research and application of electro-hydraulic proportional control system based on PWM technology," *Coal Mine Electromechanical*, vol. 2, no. 3, pp. 28–32, 2017.

[11] Y. Zhang, L. Wang, and Y. Gai, "Research on valve-controlled cylinder experimental system based on electro-hydraulic proportional technology," *Disciplines Exploration*, vol. 35, no. 5, pp. 28–30, 2018.

[12] J. Fang, Y. Zhang, and Y. Fang, "Study of the control strategy of an electro-hydraulic control system," *Machine Tools & Hydraulics*, vol. 36, no. 2, pp. 114–116, 2003.

[13] D. Liu, "Rotational speed and simulation of wind turbine based on BP Neural network controller," *Instrument Technique and Sensor*, vol. 6, pp. 13–16, 2018.

[14] S. Huang, Z. Chen, Y. Jiao et al., "Fuzzy-neural networks PID control based on DSP for hydraulic servo system," *Chinese Hydraulics and Pneumatics*, vol. 9, pp. 15–25, 2015.

[15] H. Shan and G. Shen, "Damping neural network control of air suspension for mine vehicle," *Coal Mine Machinery*, vol. 35, no. 5, pp. 75–77, 2014.

[16] Y.-H. Lee and R. Kopp, "Application of Fuzzy Control for a hydraulic forging machine," *Fuzzy Sets and Systems*, vol. 118, no. 1, pp. 99–108, 2001.

[17] A. Ferhat bayram, "Application of an artificial neural network model to a na-k geothermometer," *Volcanology and Geothermal Research*, vol. 66, no. 112, pp. 75–81, 2008.

[18] J. Wongsaroj, K. Soga, and R. J. Mair, "Modeling of long-term ground response to tunnelling," *Journal of Geotechnical and Environmental Engineering*, vol. 57, no. 10, pp. 75–90, 2007.

[19] H. Xu and Y. Mai, "Adaptive PID control of hydraulic material testing machine system based on neural network," *Journal of Mechanical & Electrical Engineering*, vol. 35, no. 4, pp. 375–379, 2018.

[20] H. Ni, S. Wang, S. wang et al., "Improvement of hydraulic control system based on BP neural network PID control," *Forging & Stamping Technology*, vol. 40, no. 11, pp. 61–70, 2015.

[21] B. Gao and W. Han, "Neural network model reference decoupling control for single leg joint of hydraulic quadruped robot," *Assembly Automation*, vol. 38, no. 4, pp. 465–475, 2018.

[22] H. Huang, W. Gong, D. Zhao et al., "Wheel loader working device electronic positioning system," *Construction Machinery*, vol. 20, no. 4, pp. 26–32, 2002.

[23] T. Lou, H. Ding, J. Dong et al., "Design and simulation analysis of electro-hydraulic proportional closed loop drive systems in closed loop control o for coal mine hydraulic winch," *Chinese Hydraulics & Pneumatics*, vol. 6, pp. 63–68, 2019.

[24] N. Zhang, "Hydraulic leveling technology for electro-hydraulic proportional control based on fuzzy PID," *Shanxi Electronic Technology*, vol. 4, pp. 39–42, 2015.

[25] L. Xu, W. Li, C. Qin et al., "Experimental study on distribution characteristics of loader engine output power," *Machinery Design & Manufacture*, vol. 7, pp. 74–77, 2018.

[26] Z. Wang and S. Qin, "Analysis of dynamic characteristics of hydraulic torque converter applied to wheel loader," *Journal of South china University of Technology*, vol. 44, no. 7, pp. 41–46, 2016.

[27] L. Xu, R. Ge, and L. Chang, "Acquisition method of load time course of wheel loader hydraulic system under typical operating conditions," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 28, no. 6, pp. 57–62, 2012.

[28] J. Zhang, M. Zeng, and W. Liu, "Control strategy of cylinders synchronization motion of tensioner based on PID neural network integrated controller," *Machine Tool & Hydraulics*, vol. 38, no. 7, pp. 7–10, 2010.

[29] H. Mu, J. Han, and M. Li, "Research on the improved particle swarm optimization algorithm in the coal mill PID neural network control," *Journal of Jinling Institute of Technology*, vol. 35, no. 2, pp. 16–20, 2019.

[30] Q. Li and F. Ding, "Simulation and experimental study on stroke force characteristics of proportional electromagnet," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 15, no. 2, pp. 104–107, 2005.

[31] H. Pan, B. Li, and Y. Lei, "Method for evaluating the correct rate of neural network model prediction results," *Industrial Building*, vol. 14, no. 6, pp. 37–39, 2007.

[32] X. Luo and L. Li, "The research on communication base station's air-conditioner system fuzzy neural network control," *Journal of Dezhou University*, vol. 2, pp. 57–61, 2015.

[33] W. He and Y. Dong, "Adaptive fuzzy neural network control for a constrained robot using impedance learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 4, pp. 1174–1186, 2018.

[34] Y. Hu and B. Si, "A reinforcement learning neural network for robotic manipulator control," *Neural Computation*, vol. 30, no. 5, pp. 13–22, 2018.

[35] B. G. Kumar, S. R. Shankapal, A. S. Ravindran et al., "Fuzzy logic and neural network based induction control in a diesel engine," *Advances in Intelligent Systems and Computing*, vol. 270, pp. 77–91, 2014.

Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 6913813, 14 pages
https://doi.org/10.1155/2020/6913813

Research Article

Real-Time Low-Cost Speed Monitoring and Control of Three-Phase Induction Motor via a Voltage/Frequency Control Approach

Ali Hmidet [1] **and Olfa Boubaker** [2]

[1]*Tunis El Manar University, ISTMT, Tunis, Tunisia*
[2]*Carthage University, National Institute of Applied Sciences and Technology, Tunis, Tunisia*

Correspondence should be addressed to Ali Hmidet; hmidetali@yahoo.fr

Received 1 January 2020; Accepted 14 March 2020; Published 22 April 2020

Guest Editor: Mohamed Nayel

In this paper, a new design of a real-time low-cost speed monitoring and closed-loop control of the three-phase induction motor (IM) is proposed. The proposed solution is based on a voltage/frequency (V/F) control approach and a PI antiwindup regulator. It uses the Waijung Blockset which considerably alleviates the heaviness and the difficulty of the microcontroller's programming task incessantly crucial for the implementation and the management of such complex applications. Indeed, it automatically generates C codes for many types of microcontrollers like the STM32F4 family, also used in this application. Furthermore, it offers a cost-effective design reducing the system components and increasing its efficiency. To prove the efficiency of the suggested design, not only simulation results are carried out for a wide range of variations in load and reference speed but also experimental assessment. The real-time closed-loop control performances are proved using the aMG SQLite Data Server via the UART port board, whereas Waijung WebPage Designer (W2D) is used for the web monitoring task. Experimental results prove the accuracy and robustness of the proposed solution.

1. Introduction

During the last decades, induction motor (IM) drives are becoming more and more popular in industrial applications as well as in home appliances, thanks to their reliability, low cost, robustness, ease of maintenance, and simplicity of control [1–3].

The control methods of IM are mainly classified into two major classes: scalar control and vector control approaches [4–6]. Scalar control, popularly known as voltage/frequency (V/F) control, is considered as a simple approach based on the control of the supply voltage amplitude and the frequency. For indirect control of the IM speed, it uses a three-phase voltage source inverter (VSI) controlled by a pulse width modulation (PWM) technique. However, despite its simplicity of implementation, scalar control methods cannot achieve best performances during transients, which is considered as a major disadvantage [1, 7, 8].

Vector control approaches, also known as field-oriented control (FOC) approaches, allow not only the control of the amplitude and frequency of the voltage as it is the case for scalar control approaches but also the instantaneous position of the flux, the voltage vectors, and the current vectors [9, 10]. They are considered as mathematical model-based approaches. Especially in the transient regime, they guaranty better control performances compared to the scalar control approaches. Unfortunately, such control methods have complex algorithms and require a lot of computation time [11–13].

As the simplicity of the control algorithm is of major interest for users, the scalar control approach is still considered as the most used in industrial applications, especially when accuracy of the speed response in the transient regime is not mandatory, such as for ventilation and air conditioning applications and heating and pumping systems [13–15]. The objective of these methods is to control the IM

speed by maintaining constant stator flux. The magnitude of the stator flux is proportional to the ratio between the stator voltage and the frequency. However, if this ratio is kept constant, the flux remains constant. Also, by keeping V/F constant, the developed torque remains approximately constant. This method gives higher run-time efficiency [16, 17]. Therefore, most AC speed drives (ACSD) employ the constant V/F method for the speed control. Along with the wide range of speed control, this method also offers "soft start" capability [18, 19].

On the contrary, the great progress in microcontrollers and the power electronic components became an important factor in variable speed drive processing [20–22]. STM32F4 family of microcontrollers, for example, offers high-quality performances at the service of high-performance variable speed drives. Programming microcontrollers for such heavy and complex applications with conventional languages (such as assembler, *C*, or *C*++ languages) increase the scaling time of such applications. It requires, furthermore, the availability of an expert in computer programming with in-depth knowledge of processor architecture. In addition, in order to reduce operating costs and enhance the reliability and safe security of the proposed design, industrial applications should be monitored in real-time which allows perfect control and system supervision. Real-time monitoring has become consequently a major task for engineers and researchers in industrial applications such as pumping, mining industry, railways, and industrial drives. [23–25].

In this work, we propose a new design of a real-time low-cost speed monitoring and closed-loop control of the three-phase induction motor. The control system, operating according to the rule V/F constant, is developed using VSI based on a space vector pulse width modulation (SVPWM) technique and a three-phase IM loaded by a magnetic power breaker. Waijung is also designed to solve this problem. It is a Simulink Blockset that can be used as targets to easily and automatically generate C code from Matlab/Simulink simulation models for STM32F4 Discovery microcontroller. The real-time speed monitoring and control of the proposed algorithm are done in two different ways. The first is based on the COM/UART port and USB converter, and the second manner uses Waijung WebPage Designer (W2D). Hardware required for this application is grouped as aMG Labkit F4N, suitable for many industrial applications. To our best knowledge, such solutions have never been previously tested for ACSD despite the significance of the solution. Furthermore, performance evaluation of the proposed algorithm will be verified not only via computer simulation results but also via experimental assessment for a wide range of variations in load and reference speed.

This paper is structured as follows. Section 2 presents the mathematical model of the three-phase IM and the related drive system. Section 3 presents the designed scalar control approach for closed-loop control of the IM. In Section 4, the effectiveness of the proposed approach is proved via simulation results. Finally, Section 5 expands experimental results for different scenarios.

2. Process Mathematical Modeling

2.1. The Induction Motor. The electrical machine considered in this paper is a three-phase squirrel-cage (short-circuit rotor) asynchronous machine. The main electrical equations in the stationary reference frame can be written by the following form [7–9]:

$$\frac{d\overline{\varphi}_s}{dt} = \overline{v}_s - R_s\overline{i}_s, \tag{1}$$

$$\frac{d\overline{\varphi}_r}{dt} = j\omega_r\overline{\varphi}_r - R_r\overline{i}_r, \tag{2}$$

$$\overline{\varphi}_s = \ell_s\overline{i}_s + m\overline{\varphi}_r, \tag{3}$$

$$T_{em} - K_f\omega_r - T_r = \frac{j}{p}\frac{d\omega_r}{dt}. \tag{4}$$

In the previous set of equations, $\overline{v}_s$ is the stator voltage vector per phase and $\overline{\varphi}_s$ and $\overline{\varphi}_r$ represent the stator/rotor flux, respectively. The stator/rotor currents are denoted by $\overline{i}_s$ and $\overline{i}_r$. $R_s$and R_r are the stator/rotor resistances, respectively. ℓ_s and m are leakage inductance and the ratio between mutual inductance M and rotor inductance L_r, respectively. Electrical speed and the number of pole pairs are denoted by ω_r and p, respectively. Various expressions can be used to calculate electromagnetic torque T_{em}. The most used relation is described by equation (4), where K_f is the viscous coefficient, j is the moment of inertia, and T_r is the load torque.

2.2. The Induction Motor Drive. Figure 1 describes the association between a three-phase voltage converter and the IM. The power circuit of the converter is basically composed of three modules, namely, the three-phase rectifier, the AC/DC filter capacitor, and three-phase inverter, whereby D_1–D_6 are the three-phase rectifier diode circuit, C is the filter capacitor DC bus, and C_1–C_6 are the power switches. A three-phase source inverter, whose objective is to provide variable voltage and variable frequency output through pulse width modulation (PWM) control, drives the IM [26]. Several PWM techniques can be used to generate signal command for the voltage inverter [13, 15, 19]. In this work, we use the SVPWM technique.

2.3. Space Vector Pulse Width Modulation Technique. Considering a lossless three-phase inverter, the output voltage is obtained according to the DC bus voltage Vdc and the logic state of the three highest switches (C1, C2, C3). Thus, there are eight possible logical combinations of (C1, C2, C3) leading to six active voltage vectors and two null voltage vectors. The space vector of the output voltage of the inverter can be expressed by the following expression:

$$\overline{v}_k = \begin{cases} \sqrt{\frac{2}{3}}V_{dc}e^{j(k-1)(\pi/3)}, & \text{for } k = 1, \ldots, 6, \\ 0 & \text{for } k = 0, 7, \end{cases} \tag{5}$$

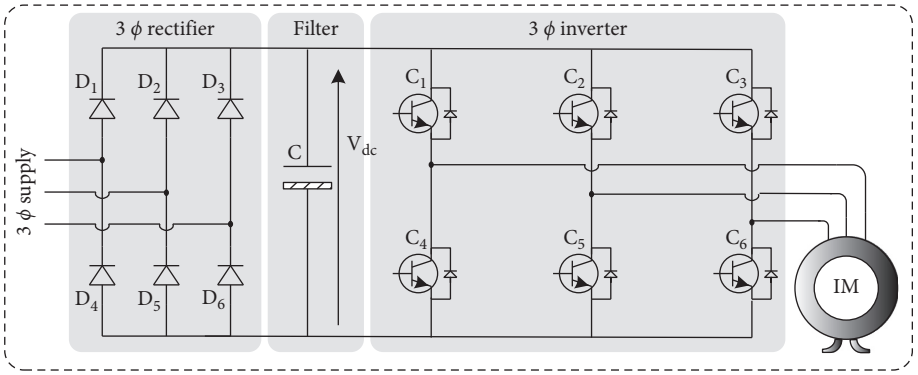


Figure 1: Association of power modules for the three-phase voltage converter-IM.

where k is an integer indicating the switching combinations [13, 26]. In the complex plane d-q, different voltage vectors delivered by the inverter and the associated combinations of the switches C1, C2, and C3 are given by Figure 2.

The goal of SVPWM is to produce a mean voltage vector during the PWM period (T_s) that is equal to the desired voltage vector $\overline{v}_{\text{ref}}$. This is done by applying the neighboring vectors $\overline{v}_k$ for a specified time (τ_k), $\overline{v}_{k+1}$for a specified time ($\overline{v}_{k+1}$), and the null vector $\overline{v}_0$ or $\overline{v}_7$ for the amount of time necessary τ_0. Thereby, to obtain an average value which equals to $\overline{v}_{ref}$ over the period T_s, we must have the following relation:

$$\frac{\tau_k \cdot \overline{v}_k + \tau_{k+1} \cdot \overline{v}_{k+1}}{T_s} = \overline{v}_{\text{ref}}. \tag{6}$$

The values of τ_k, τ_{k+1}, and τ_0 can be computed using the following equations:

$$\begin{cases} \tau_k = T_s \rho \sin\left(\left(\frac{\pi}{3}\right) - \varsigma\right), \\ \tau_{k+1} = T_s \rho \sin(\varsigma), \\ \tau_0 = T_s - \tau_k - \tau_{k+1}, \\ \rho = \dfrac{\sqrt{2}\, V_{ref}}{V_{dc}}, \end{cases} \tag{7}$$

where the coefficient ρ designates a voltage ratio.

To respect the condition of $\tau_k + \tau_{k+1} \leq T_s$, the module V_{ref} of the requested voltage $\overline{v}_{\text{ref}}$ should verify the following condition:

$$V_{\text{ref}} \leq \frac{V_{dc}}{\sqrt{2}}. \tag{8}$$

It should be noted here that the condition given by condition (6) related to the feasibility of the synthesis is established by considering a voltage vector in the Concordia reference frame, which implies that the magnitude V_{ref} used corresponds to either $\sqrt{3}$ times the rms value or $\sqrt{3/2}$ times the maximum value in the natural system.

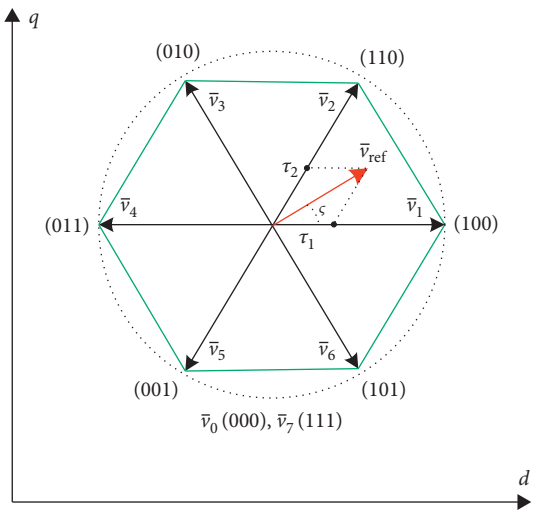


Figure 2: Inverter voltage vectors in Concordia's reference frame.

3. Scalar Control Design

The scalar control methodology focuses only on the steady state dynamic, allows to adjust the supply voltage as well as its frequency so that their ratio remains constant in order to avoid magnetic flux saturation, and affects the available torque of the machine [7, 8]. Starting from equation (1) and assuming that the voltage drop across the stator resistance is small in comparison with the stator voltage mainly at the low-slip regions, the stator flux Φ_s can be expressed as [12, 16]

$$\Phi_s \approx \frac{V_s}{2\Pi f_s}, \tag{9}$$

where V_s is the stator voltage magnitude and f_s is the supply frequency (Hz). The electromagnetic torque-slip characteristic in the stationary regime takes the following form [6]:

$$T_{em} = 3p \frac{R_r}{g} \frac{V_s^2}{\omega_s} \frac{1}{(R_r/g)^2 + X_{lr}^2}, \tag{10}$$

where g is the slip and R_r and X_{lr} represent the rotor resistance and total leakage reactance located in the rotor, respectively.

In the low-slip region (normal operating conditions), the above equation is simplified by

$$T_{em} = 3\frac{p}{R_r}\Phi_s^2 g\omega_s. \tag{11}$$

This implies that if the stator flux is maintained constant, the variation of the torque as a function of the slip is almost linear.

Frequency control is natural for adjustable speed drive applications. However, voltage is required to be proportional to frequency so that the stator flux remains constant if the stator resistance is neglected. The open-loop V/F control of an IM is one of the most common methods widely used in the industry due to its simplicity, low cost, and performance drive [27]. As the rotor speed $\omega_r$will be slightly less than the synchronous speed ω_s as a result of the slip speedω_g, the speed of the motor cannot be controlled accurately. Also, as the rotor speed is not measured in this schema, the slip pulsation cannot be maintained. Consequently, the operation in the unstable area of the torque-speed characteristics can take place. Equally, the stator currents can exceed greatly the nominal current under the effect of the point mentioned above and thus endangers the combination inverter machine. These drawbacks should have solutions by making an outer control loop in the IM drive, where the actual rotor speed ω_r should be compared with a reference value ω_{ref}, and hence, the error between these variables is generated [16–18]. This error is processed through a PI controller and a limiter to get the slip-speed command ω_g. The inverter frequency command ω_{sref} is generated by adding the slip command ω_g and the actual speed signal ω_r, and the frequency command generates the voltage command V_s or its voltage ratio ρ through a volt/hertz function generator. The resulting ω_{sref} and ρ are applied to the motor by means of an SVPWM-VSI.

In low-speed operation, the voltage drop in the stator resistance is not negligible compared to that in the leakage reactance. Inversely, in the operation at a speed greater than that corresponding to the nominal frequency, a weakening of the magnetic field is led. In order to avoid these anomalies, the V/F control structure should then consider the voltage drop in the resistance at the low speed to maintain a constant stator flux, and it must limit the voltage when the nominal frequency is reached. The input voltage is then adjusted according to the frequency required to a specific speed reference as shown by the following equation [12, 16, 21]:

$$V_s = \begin{cases} (V_{sn} - V_0)\dfrac{f}{f_n} + V_0, & \text{for } f < f_n, \\ V_{sn}, & \text{for } f \geq f_n. \end{cases} \tag{12}$$

A control scheme of the V/F and slip regulation-controlled IM is shown in Figure 3.

For a better performance of the conventional scalar control, an antiwindup PI regulator is used. The antiwindup PI controller is an improvement of the classical PI control. The use of antiwindup strategy is to prevent the controller from going into deep saturation and inspect the windup controller output when large set-point changes are made. This helps to prevent extreme overshoot during the startup of the IM and guarantees the desired performance independent of the operating conditions, i.e., reference variations and load torques [8, 20]. Figure 4 shows the block diagram of a PI controller where the difference between the output value and the input value of the saturation block is used as a feedback signal through a gain (1/Tt) to forward the input to the integrator.

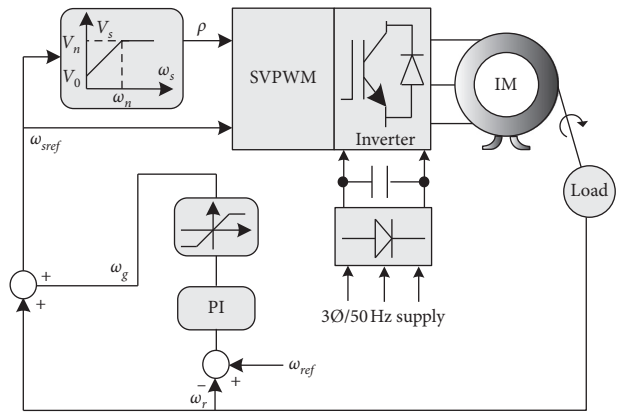


Figure 3: Closed-loop scalar control scheme.

4. Simulation Results

To analyze and verify the performances of the proposed solution, the control methodology shown by Figure 3 is carried out by using MATLAB/Simulink software for the technical specifications of the IM which is given by Table 1. Simulation results are conducted for various operating conditions. Two main scenarios are given here: speed response under crenel variations of load torque and speed response under crenel variations of reference speed. The sampling time used in all simulation is 100 μs.

4.1. Scenario 1: Speed Response under Crenel Load Torque Variations. In this test, a signal reference of 150 rad/s which is equal to nominal speed is used. A variation by crenels of the load torque is applied going from zero to the nominal torque: 0 at 5 s, 3 Nm at 10 s, 6.82 Nm at 20 s, 3 Nm at 30 s, and 0 at 40 s. Figures 5–9 show the obtained simulation results. These figures show that the proposed algorithm operates perfectly, and the antiwindup PI controller offers the best response. Figure 5 shows that the speed response is coincident with its reference. This is confirmed by Figure 6 showing a zoom on Figure 5. The speed response is practically perfect with no overshoot and an insignificant delay. Figures 7 and 8 show the response of the stator voltage. The voltage ratio varies as a function of the reference speed and proportional to the stator pulsation verifying that the flux is constant as the main property of scalar control. Figure 9 gives the stator and rotor pulsations and confirms that the slip pulsation (difference between stator and rotor pulsations) is proportional to the load torque as mentioned

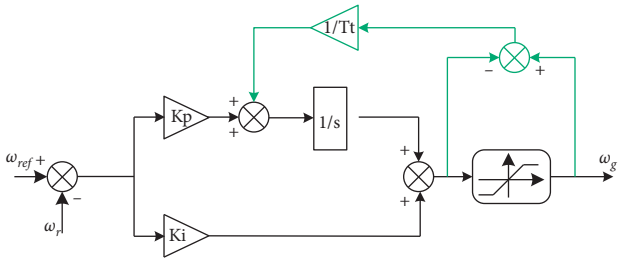


Figure 4: Speed regulation using the PI antiwindup controller.

Table 1: Motor data.

Parameter designation	Parameter value
Stator resistance	$R_s = 7.5\,\Omega$
Rotor resistance	$R_r = 6.5\,\Omega$
Principal stator inductance	$L_s = 354\,\text{mH}$
Principal rotor inductance	$L_r = 354\,\text{mH}$
Mutual inductance	$M = 340\,\text{mH}$
Number of pole pairs	$p = 2$
Rated frequency	50 Hz
Rated electrical speed	293.22 rad/s
Rated power	1 KW
Rated voltage	220/380 V
Rated power factor	0.78

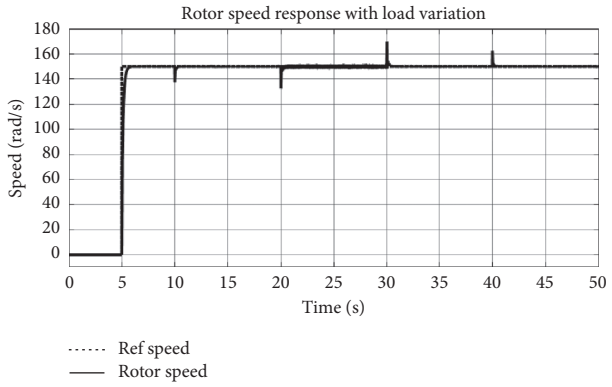


Figure 5: Speed response under load torque variation.

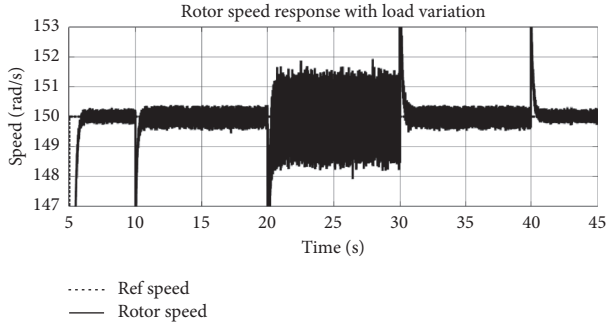


Figure 6: Steady state operation (zoom on Figure 5).

previously by equations (10) and (11). In Figure 9, the slip pulsation is between 0 (without load) and 30 rad/s (in nominal operating condition).

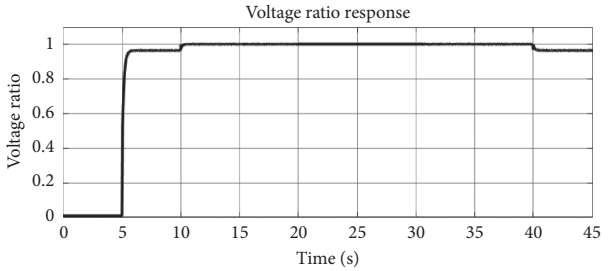


Figure 7: Voltage ratio response under load torque variation.

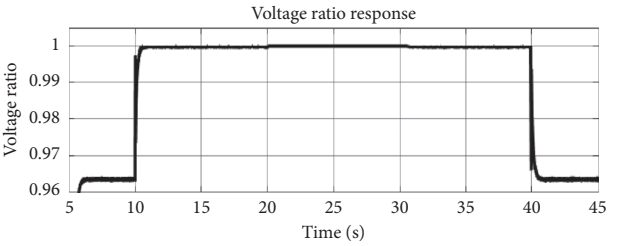


Figure 8: Zoom on Figure 7 (ratio 96.5% for IM without load, 98% for 3 Nm, and 100% for 6.82 Nm).

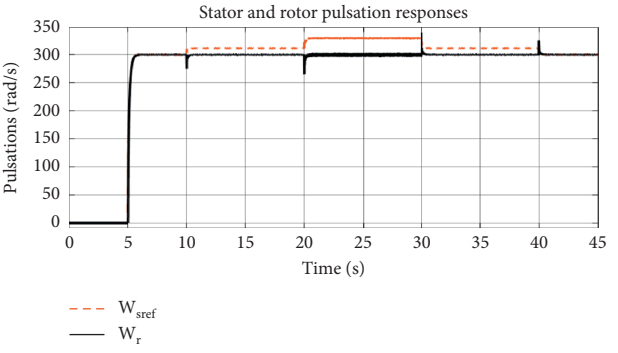


Figure 9: Stator and rotor pulsation responses under load torque variation.

4.2. Scenario 2: Speed Response under Reference Speed Variations. This simulation test performs the effects of the variations of the reference on the speed control. Figure 10 presents the rotor speed response under a reference speed crenel variation starting from 0, then 100, 120, 157, 150, and 80, and returns to 0 rad/s. In Figure 10, the rotor speed follows perfectly its reference. The IM is operated without load; thus, the stator and rotor pulsations are equal, so the slip is zero. Figure 11 shows the stator voltage represented by its ratio. It is easy to check here that the V/F constant is confirmed because the stator voltage-pulsation trajectories have the same form.

5. Experimental Validation

In this section, experimental results related to the real-time monitoring and control of the IM speed will be carried out via two communication techniques. For the first one, the communication is realized via the UART port, where for the second, the Waijung WebPage Designer (W2D) is exploited.

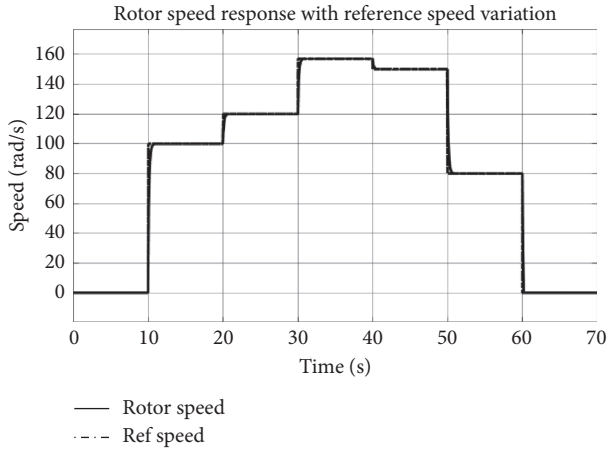


Figure 10: Speed response under reference speed variation.

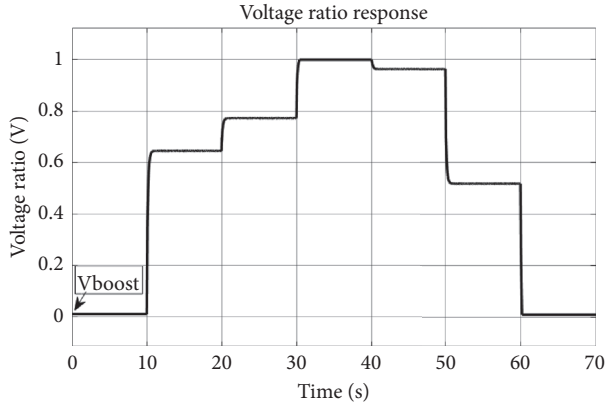


Figure 11: Voltage ratio response under reference speed variation.

Furthermore, for each technique, the experimental results are carried out for the two scenarios considered in simulation results.

The test bench [28] used for validation is built around an aMG labkit F4-based STM32F4 discovery. The aMG Labkit F4 [29] consists of low-cost high-performance DSP hardware with a variety of Plug-n-Play modules. This kit is supported by the Matlab/Simulink toolkit with Waijung Blockset and Waijung WebPage Designer (Drag and Drop Web-based Monitoring and Control) [30]. The photo of Figure 12 shows the main hardware components used in our application, including the following:

(i) aMG F4 Connect 2 as the shield to enable various add-on expansion boards cited below

(ii) aMG SQLite Database Server for the embedded database

(iii) aMG USB Converter-N2 (Converter-N Adapter) for real-time hardware in the loop simulation

(iv) aMG Ethernet INF for the LAN interface

(v) aMG CAN INF for the CAN Bus Interface

(vi) STM32F4 DISCOVERY Kit which is a low-cost development kit from STMicroelectronics

The overall test bench is shown in Figure 13. It includes in addition the following:

(i) A squirrel-cage IM with parameters given in Table 1.

(ii) A magnetic powder breaker as the load for the IM.

(iii) Omron incremental encoder having 360 pulses per revolution coupled to the machine shaft.

(iv) A Semikron converter based on the three-phase rectifier, a DC Bus filter, and the three-phase inverter. The inverter is used to drive the IM.

(v) A 2 channel, 20 MHz analog oscilloscope for visualization.

(vi) Sensors for current and voltage measurement, types LEM LA25-NP and LEM LV25-P, respectively.

The ST microcontroller TM32F4VG407 Discovery controller generates the main program. It ensures the measurement of the rotor speed via an incremental encoder, achieving the control algorithm (V/F) including a PI anti-windup regulator and the generation of the PWM signal with space vector modulation to activate six insulated-gate bipolar transistors (IGBTs) of a three-phase inverter. The experimental results are recorded with a sampling time of 100 μs.

5.1. Experimental Results via the UART Communication Port. The hardware modules used here are the aMG USB Converter-N and the aMG SQLite Data Server board. Two Simulink-generated codes are built to achieve the scalar control algorithm of the IM which are the target and the host codes. The target Simulink STM32F4-generated code shown in Figure 14 allows reading the incremental encoder information (rotor speed) for the PI regulator input. It forces the actual speed to reach its reference. It generates PWM signals using symmetrical SVPWM technique. Figure 14 shows that there are two parts: the Waijung Blockset (green color) and the developed code (light blue color). Waijung Blocksets used here are as follows:

(i) UART setup block for STM32F4 DISCOVERY + aMG F4 Connect 2 + aMG USB Converter-N2 setting use (setup UART3 Baudrate 5000000 Pin D8/D9 for Tx/Rx).

(ii) Target setup block used to setup STM32F4 in a Simulink model.

(iii) UART Rx and Tx receive and send UART information from/to host program (via aMG USB connect 2 and USB converter N2).

(iv) Basic PWM uses time 1 to generate 3 active high PWM signals

(v) Encoder Read uses built-in STM32F4 timer to interpret encoder data. However, the target model allows to read encoder channels A and B using pin B4 and B5, respectively, and output encoder position and count as binary data packet via UART Setup pin D8.

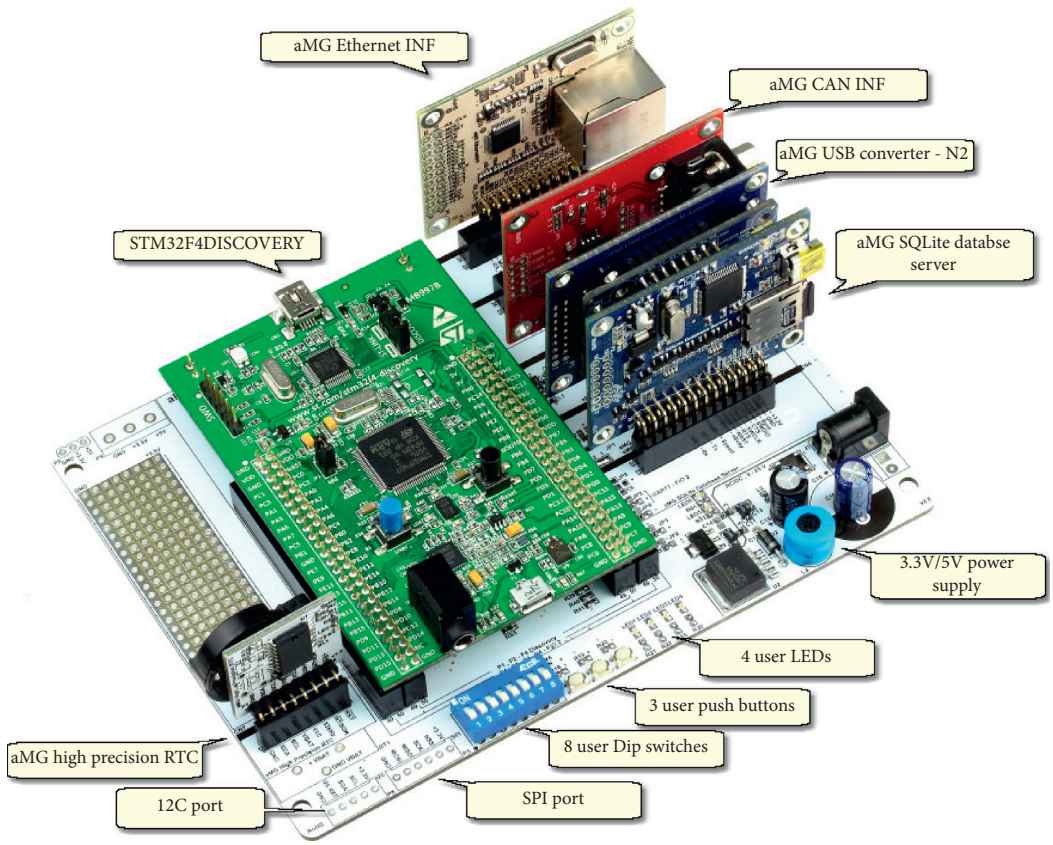


Figure 12: Low-cost high-performance DSP hardware with a variety of Plug-n-Play modules.

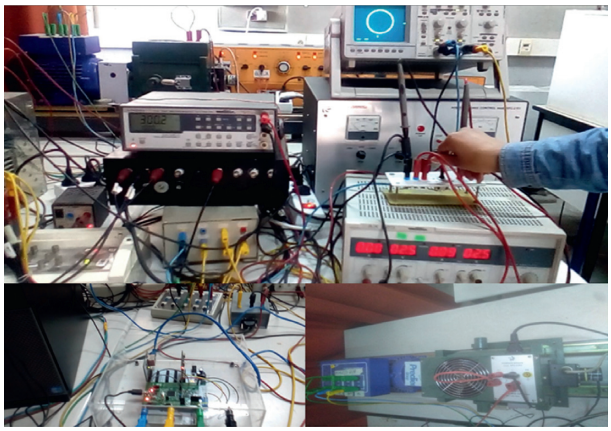


Figure 13: Experimental test bench.

Generated programs concern the incremental encoder, the symmetrical SVPWM, and the PI antiwindup controller. Figure 15 shows the encoder code developed to read the information from the incremental encoder and convert it into an angular velocity. Figure 16 shows the developed Simulink code to configure Timer 1 to generate 3 100 µs-period (10 KHz) active high PWM signals to pin E9, E11, and E13 with percentage of Ra, Rb, and Rc duty cycle, respectively. Figure 17 shows the PI controller Simulink-generated code to be used within the antiwindup PI controller.

The host Simulink STM32F4 code is shown in Figure 18. This model is configured around 3 Waijung Blocksets to send set-point changes of speed, and the received binary data packet is displayed in real time as display and plot within the COM/UART port and aMG Converter-N2 (USB-UART converter) hardware component channel *D*.

The performance and the accuracy of the designed real-time speed control of the IM are verified via two scenarios: the first scenario is done at step changing of the speed reference, where the second scenario is done at step changing of load torque under a constant reference speed.

5.1.1. Scenario 1: Reference Speed Change. This test is carried out to confirm the correctness and good performance of the antiwindup controller despite a large variation of speed reference. The reference speed variation crenels in rad/s apply from zero, then 100, 120, 157, 150, and 80, and finally zero. The IM is operated here without the load. Figure 19 proves that the sensed speed pursues its reference in a perfect way, and the time response for each variation does not exceed 2 s without overshot. This implies that the anti-windup controller operates perfectly, and the whole program is well designed for the high-performance IM speed drive.

Because the IM is without the load, the slip pulsation is null, and the stator and rotor pulsations are equal. This is confirmed by Figure 20 showing the response of the stator pulsation ω_s and the rotor speed noted here ω_r for the above

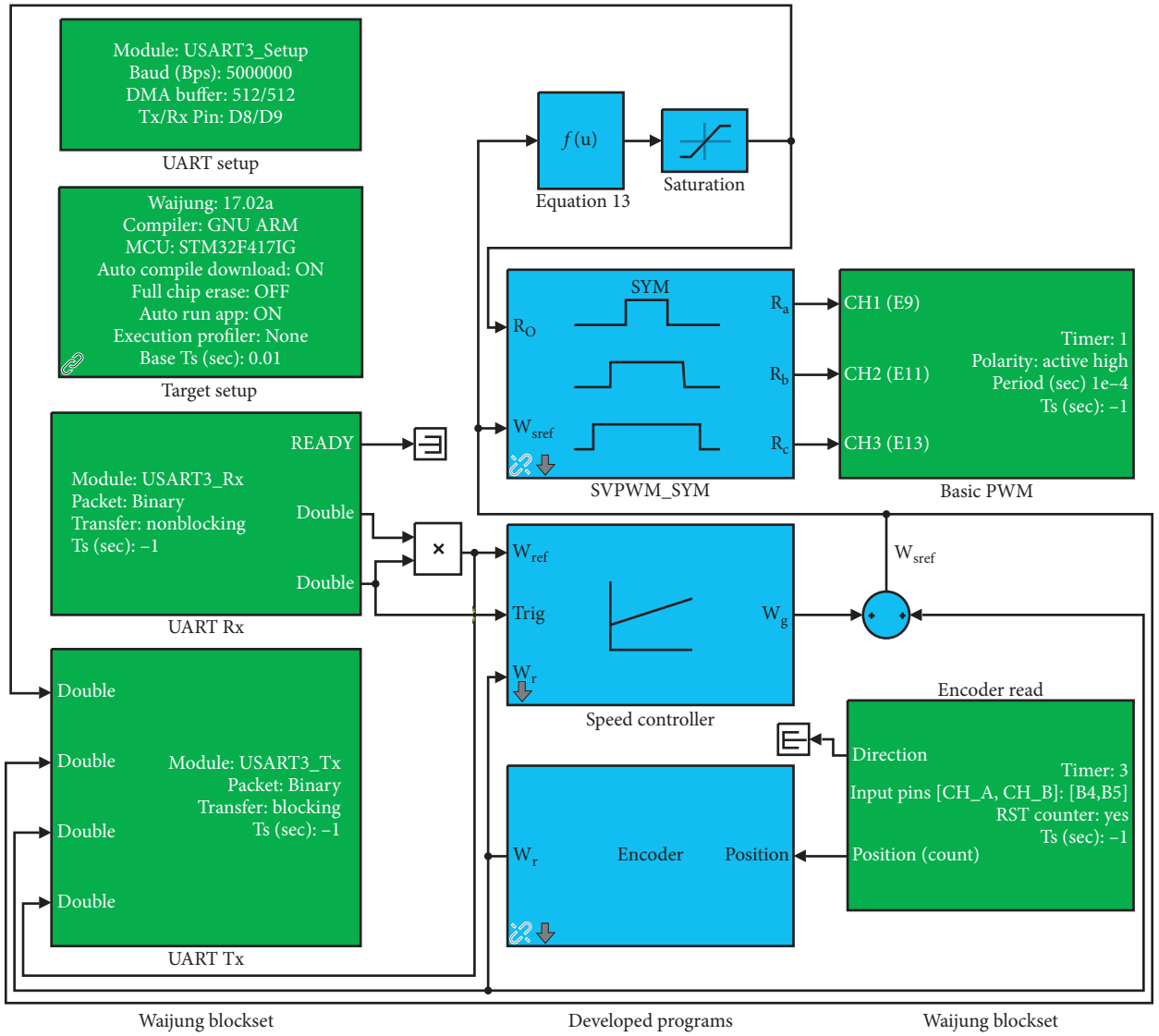


Figure 14: Target model for the implemented scalar control.

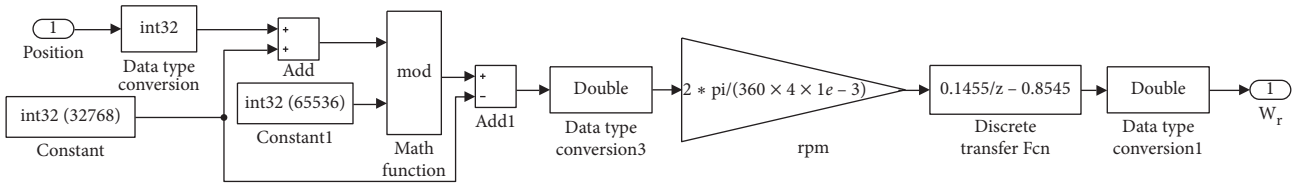


Figure 15: Simulink model for encoder target.

scenario. The stator pulsation varies parallel to the reference speed ω_{ref} and in proportionality with the stator voltage to maintain the V/F constant. This rule is confirmed by Figure 21 that the ratio voltage varies from 65% for reference speed of 100 rad/s, 78% for 120 rad/s, 100% for 157 rad/s (synchronous speed), 98% for 150 rad/s, and 52% for 80 rad/s.

5.1.2. Scenario 2: Load Torque Change. These series of tests employ a nominal reference speed of 150 rad/s when applying step variations of load torque. Figure 22 presents the response of the rotor speed when the IM is driven in a closed-loop scalar control. This figure demonstrates that, for a large variation of load torque of four crenels as in simulation tests (0, 3, 6.82, 3, and 0 Nm), the measured

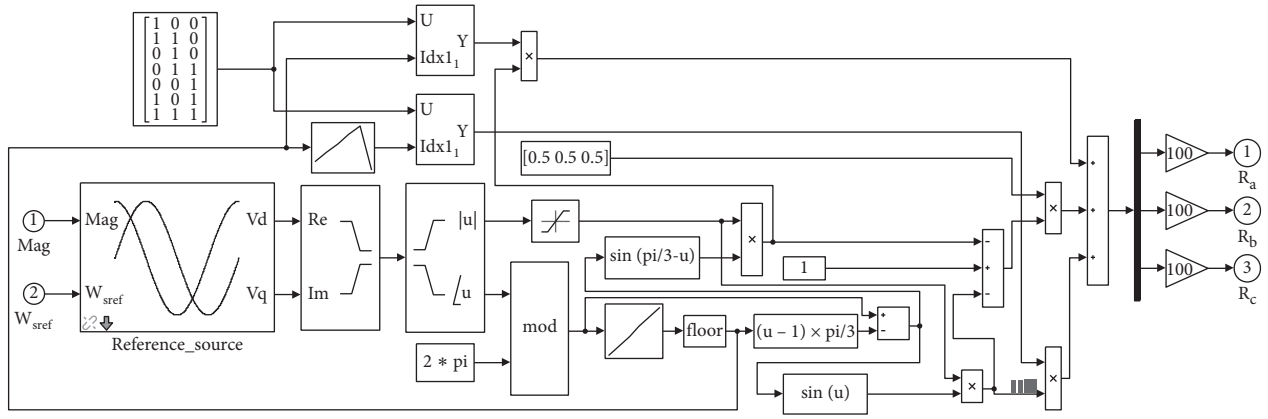


Figure 16: STM32F4-configured PWM.

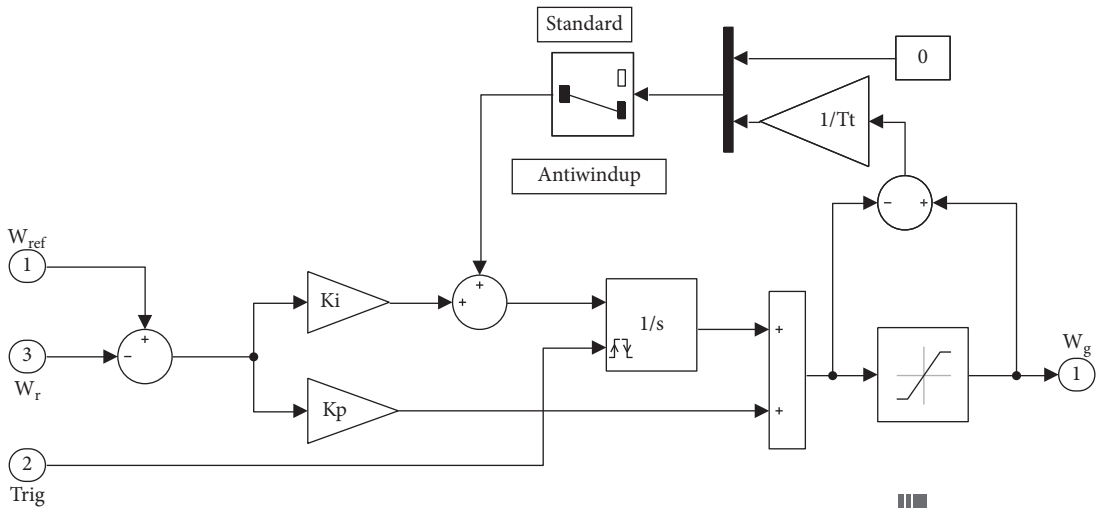


Figure 17: Antiwindup PI structure.

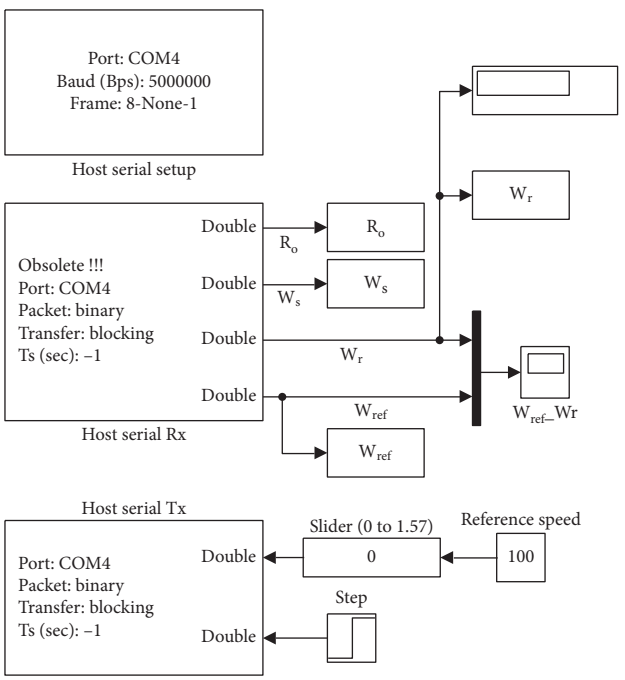


Figure 18: Host model for the implemented scalar control.

speed is kept constant close to the reference. Figure 23 presents a zoom off Figure 22 in steady state operation proving that the overshot does not exceed 2.6%, and the settling time is less than 2 s in nominal operation. Such performances prove that the antiwindup regulator PI operates correctly, and its parameters Ki and Kp are well designed. Besides, a fine consistency between the generated Simulink codes and the Waijung Blockset target for STM32F4 Discovery is proved. In addition, a better use of Waijung Blockset in electrical machines control applications.

The proprieties of the V/F control for the IM are well verified by Figures 24–27. Reverting to equations (11) and (12) that the electromagnetic torque is proportional to the square of the stator voltage and its frequency, these properties explain why the voltage ratio and the stator pulsation increase and decrease parallel to the load torque to keep rotor speed constant. The real-time response of the voltage ratio is shown in Figure 24. A zoom on this figure, Figure 25, proves that the ratio varies between 96.5% and 100% according to the load torque in respect to equation (12) (100% in the nominal operating regime). The reference stator pulsation used to compute the stator voltage,

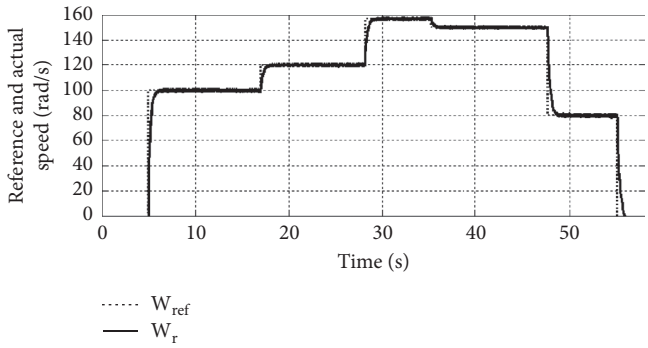


Figure 19: Speed response under the effect of reference speed changes.

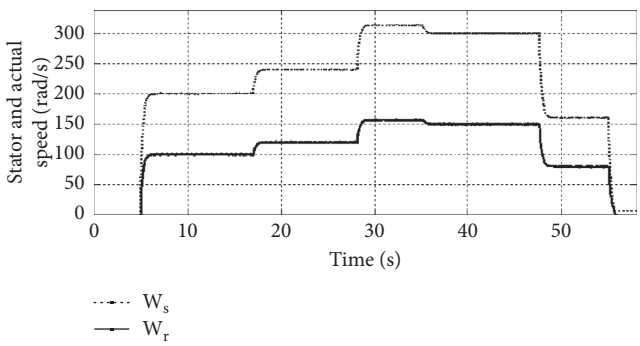


Figure 20: Stator and rotor pulsation responses under the effect of reference speed changes.

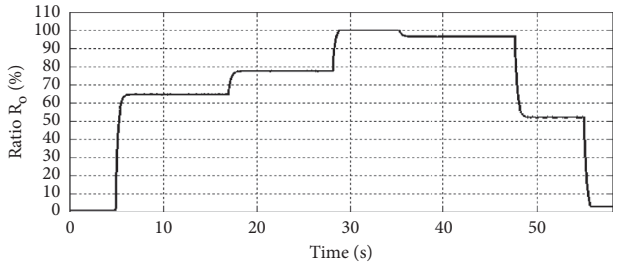


Figure 21: Voltage ratio response under the effect of reference speed changes.

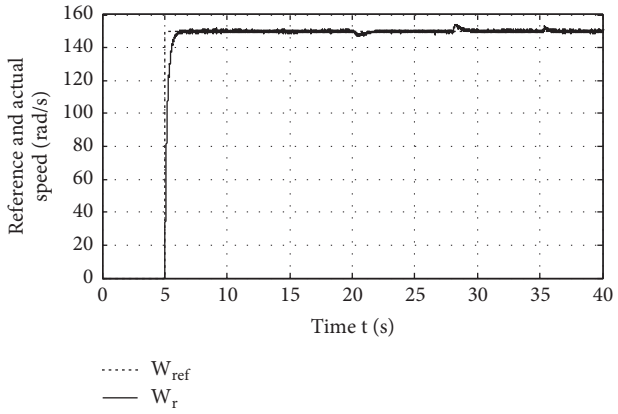


Figure 22: Speed response under load torque variation.

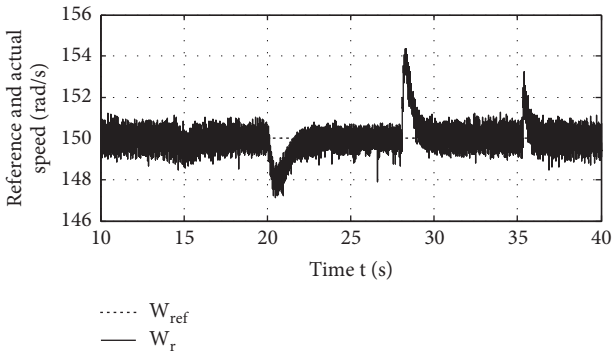


Figure 23: A zoom on measured speed.

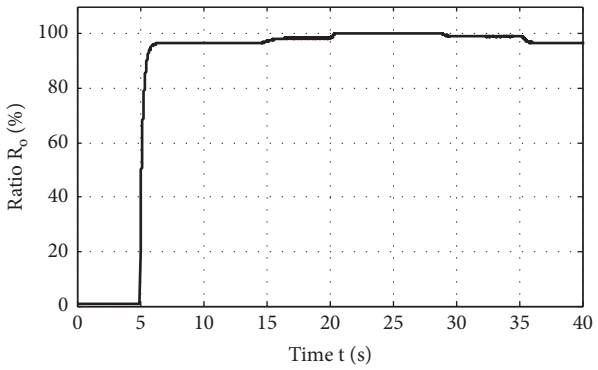


Figure 24: Voltage ratio response under load torque variation.

according to equation (12), is shown by Figure 26. A zoom on the last figure shown in Figure 27 denotes that such pulsation varies from 300 rad/s with no load torque to 331 rad/s in the nominal operating regime. The difference between the stator and rotor pulsation represents the slip pulsation, which is proportional to the torque (described by equation (11)) and proves the well regulation of the speed variable. It is obvious to observe that all obtained experimental results are incredibly close to simulation results. This finding proves the correctness and the accuracy of the real-time implementation of the proposed speed control hardware/software solution.

5.2. Experimental Results via Waijung WebPage Designer. Waijung WebPage Designer (W2D) is a set of Web tools dedicated to applications requiring monitoring and control on the Web in a simpler and faster way using the Waijung Blockset and Aimagin hardware. In other words, W2D is a compilation of web tools, such as HTML5, CSS, Javascript, JSON, jQuery, jQuery Mobile, SQLite, AJAX, and stream. The system also presents the automatic and manual control methods to stop or start the induction machine to avoid system failures. The hardware requirement for this application is cited as follows:

(i) STM32F4DISCOVERY

(ii) aMG F4 Connect 2 + Micro SD card

(iii) aMG Ethernet INF

(iv) aMG SQLite Database Server + Micro SD card

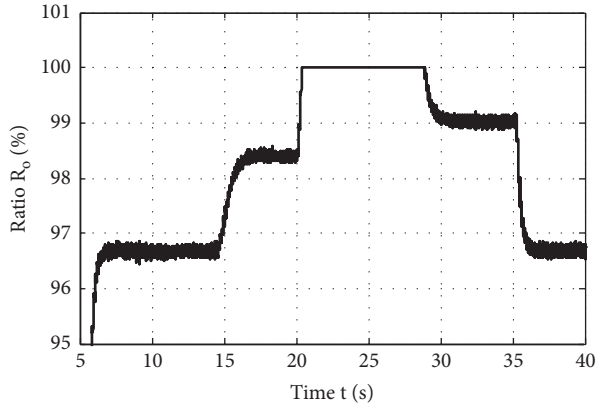


Figure 25: A zoom on voltage ratio.

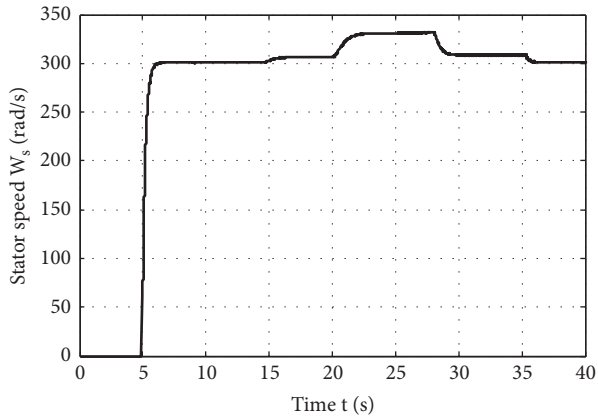


Figure 26: Stator pulsation response under load torque variation.

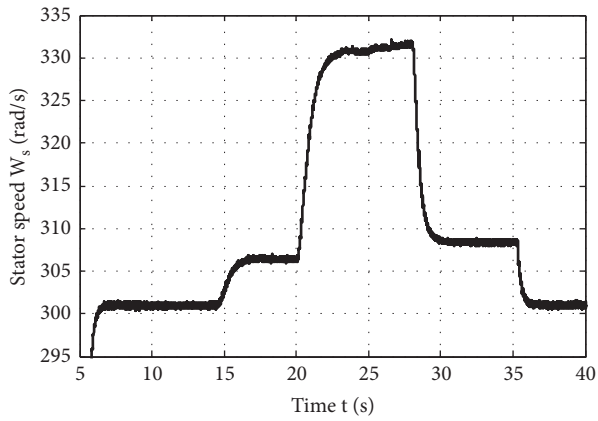


Figure 27: Zoom on stator pulsation.

The Micro SD card must be at least 4 GB so that it provides a data speed fast enough for communication with the MCU.

The Simulink/Matlab model used for web monitoring and control of the IM speed in closed-loop operation is shown in Figure 28. In addition to main previous programs (Encoder, SVPWM, PI controller and main Target Setup), it contains setup blocks for W2D. Inside the W2D Setup Subsystem (white color), it is grouped as five setup blocks.

(i) Webserver Setup composed by three blocks: Ethernet Link Setup, Ethernet Application Setup, and Http Server Setup

(ii) UART Setup uses UART Module 6 for Tx/Rx (to talk to aMG SQLite Database Server Module)

(iii) SQLite Database setup uses Port UART6 (interface with aMG SQLite Database Server)

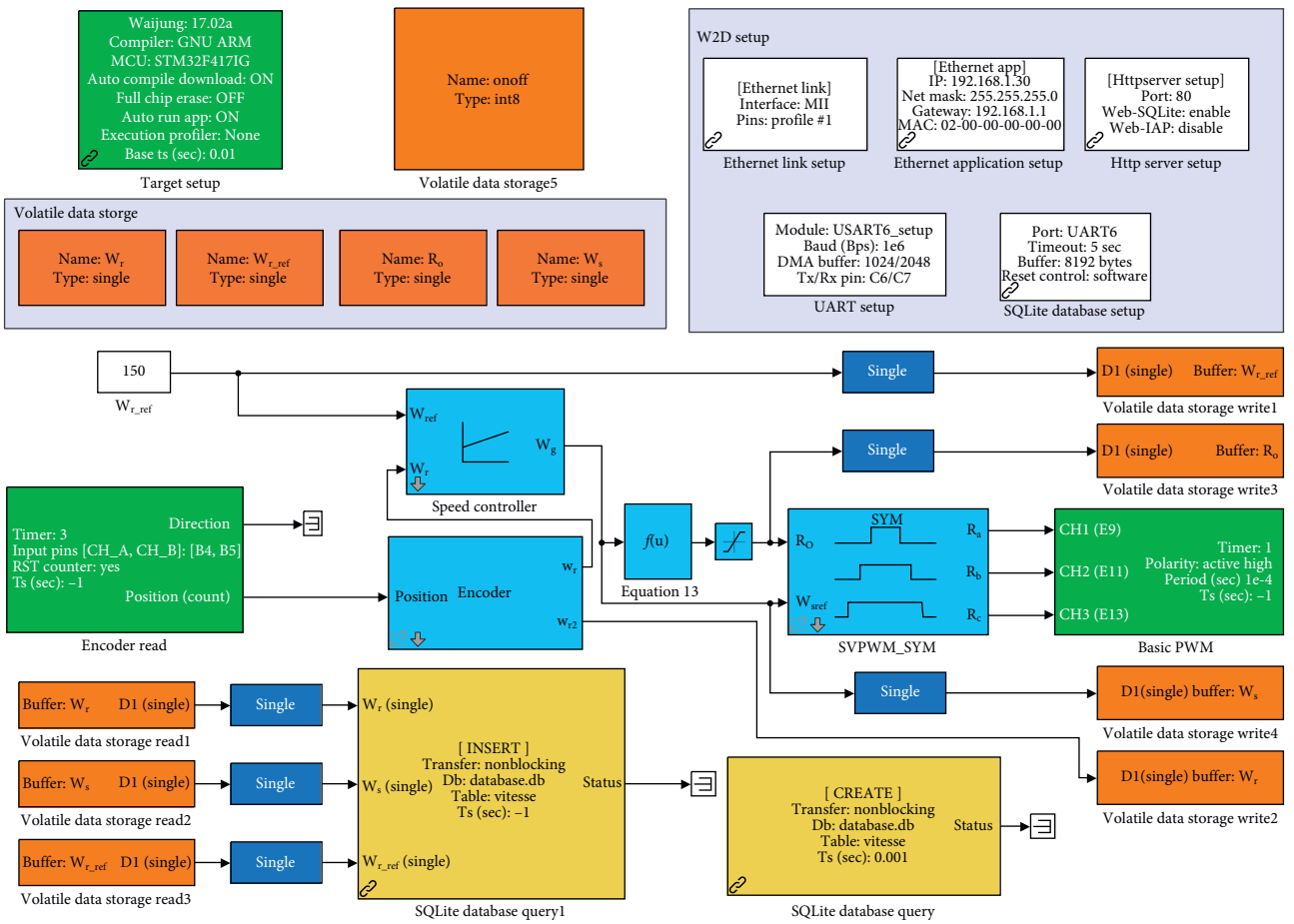


Figure 28: Simulink model of speed monitoring and control for IM with Waijung W2D.

Figure 29: Web-based monitoring and control.

The volatile data storage (orange color) is used to store data for another use of the model. Values stored in data storage are mapped for web access. Mapped variables are set to Read Only or Write/Read. The SQLite database query (yellow color) is configured for communication between Web Interface and Webserver System.

Using the Waijung WebPage Designer (W2D) tool, we created a web interface for monitoring values (rotor speed Wr, stator speed Ws, and voltage ratio RO) on the MCU webserver and displaying on web browser as text and digital values, Ws, Wr, and Ro. Figure 29 shows an illustrative example of the results on the web. The rotor speed value is in red color, and the voltage ratio and stator pulsation are presented in blue and yellow color, respectively. The purpose here is to reveal the scientific and technological feasibility of the web monitoring and control since the performance of the algorithm is verified in the previous part. We attached with this paper a sequence video summarizing these experimental results.

6. Conclusion

In this work, a technical approach for the monitoring and control of the IM speed driven under V/F control is introduced and carried out via simulation and experimental results using a Waijung Blockset and aMG hardware around an STM32F4 Discovery board.

The proposed control structure is strengthened by a solid antiwindup PI regulator and supported by the association of the STM32F4-Waijung Blockset which constitutes a perfect and promising platform for many industrial applications due to the rapid and the simple prototyping and the low-cost design. Please confirm that this is your intended meaning. Indeed, not only it eliminates the use of the voltage and the current sensors but also it removes use of hard programming languages. The estimated cost of the entire installation (STM32F4 and Aimagin hardware components) does not exceed 400 USD, while the estimated cost of a similar prototype using dSPACE 1104, for example, exceeds 12,000 USD.

Simulation and experimental tests carried out for speed reference changes and load torque prove the robustness of the proposed solution, where the PI controller admits a superior response even in severe operating regimes.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

Funding for this work was provided by the Research Laboratory ERCO (Energy, Robotics Control and Optimization), National Institute of Applied Sciences and Technology (INSAT) of Tunis, Tunisia.

References

[1] M. P. Sruthi, C. Nagamani, and G. Saravana Ilango, "An improved algorithm for direct computation of optimal voltage and frequency for induction motors," *Engineering Science and Technology, an International Journal*, vol. 20, no. 5, pp. 1439–1449, 2017.

[2] C. Liu and Y. Luo, "Overview of advanced control strategies for electric machines," *Chinese Journal of Electrical Engineering*, vol. 3, no. 2, pp. 53–61, 2017.

[3] M. A. Hannan, J. A. Ali, A. Mohamed, and A. Hussain, "Optimization techniques to enhance the performance of induction motor drives: a review," *Renewable and Sustainable Energy Reviews*, vol. 81, no. 2, pp. 1611–1626, 2018.

[4] K. Rakesh, D. Sukanta, and A. Kumar, "Comparative assessment of two different model reference adaptive system schemes for speed-sensorless control of induction motor drives," *IET Electric Power Applications*, vol. 10, no. 2, pp. 141–154, 2016.

[5] A. Khedher and M. Faouzi Mimouni, "Sensorless-adaptive DTC of double star induction motor," *Energy Conversion and Management*, vol. 51, no. 12, pp. 2878–2892, 2010.

[6] M. Jannati, S. A. Anbaran, and S. H. Asgari, "A review on variable speed control techniques for efficient control of single-phase induction motors: evolution, classification, comparison," *Renewable and Sustainable Energy Reviews*, vol. 75, pp. 1306–1319, 2017.

[7] S. Andrew, G. Shady, A. Matthew, and F. John, "Improved method for the scalar control of induction motor drives," *IET Electric Power Applications*, vol. 7, no. 6, pp. 487–498, 2013.

[8] M. P. Juan and V. D. Edilberto, "Implementation of V/F scalar control for speed regulation of a three-phase induction motor," in *Proceedings of the 2016 IEEE Andescon*, IEEE, Arequipa, Peru, October 2016.

[9] A. Hmidet, R. Dhifaoui, and O. Hasnaoui, "A new direct speed estimation and control of the induction machine benchmark: design and experimental validation," *Mathematical Problems in Engineering*, vol. 2018, Article ID 9215459, 10 pages, 2018.

[10] S. Tole, N. I. Nik Rumzi, and J. Auzani, "A review of direct torque control of induction motors for sustainable reliability and energy efficient drives," *Renewable and Sustainable Energy Reviews*, vol. 32, pp. 548–558, 2014.

[11] I. M. Alsofyani and N. Idris, "A review on sensorless techniques for sustainable reliability and efficient variable frequency drives of induction motors," *Renewable Sustainable Energy Review*, vol. 24, pp. 111–121, 2013.

[12] Y. Oguz and M. Dede, "Speed estimation of vector controlled squirrel cage asynchronous motor with artificial neural networks," *Energy Conversion and Management*, vol. 52, no. 1, pp. 675–686, 2011.

[13] B. Mohamed, L. Chrifi-A, N. V. Alessandro, M. Ch, and D. Said, "Online robust estimation of flux and load torque in induction motors," *The International Journal of Advanced Manufacturing Technology*, vol. 94, no. 5–8, pp. 2703–2713, 2018.

[14] B. Singh and S. Shukla, "Induction motor drive for PV water pumping with reduced sensors," *IET Power Electronics*, vol. 11, no. 12, pp. 1903–1913, 2018.

[15] I. Boldea, "Control issues in adjustable speed drives," *IEEE Industrial Electronics Magazine*, vol. 2, no. 3, pp. 32–50, 2008.

[16] T. H. dos Santos, G. Alessandro, S. A. O. da Silva, and S. Marcelo, "Scalar control of an induction motor using a neural sensorless technique," *Electric Power Systems Research Elsevier journal*, vol. 108, pp. 322–330, 2014.

[17] A. Pugachev, "Efficiency Increasing of Induction Motor Scalar Control Systems," in *Proceedings of the 2017 International Conference on Industrial Engineering*, Applications and manufacturing (ICIEAM), St. Petersburg, Russia, May 2017.

[18] G. R. Markadeh, E. Daryabeigi, C. Lucas, and M. A. Rahman, "Speed and flux control of induction motors using emotional intelligent controller," *IEEE Transactions on Industry Applications*, vol. 47, no. 3, pp. 1126–1135, 2011.

[19] K. S. Anup, M. K. Nallapaneni, P. Swapnajit, and K. Vinay Reddy, "Speed control of 3-phase induction motor fed through direct matrix converter using GSPWM technique with unity input power factor," in *Proceedings of the 2016 international conference on electrical power and energy systems (ICEPES)*, pp. 420–425, IEEE, Bhopal, India, Bhopal, India, December 2016.

[20] R. Maamouri, M. Boussak, and F. M'Sahli, "Sliding mode observer sensorless vector controlled induction motor drive with anti-windup PI speed controller," in *Proceedings of the XXII International Conference on Electrical Machines (ICEM)*, IEEE, Lausanne, Switzerland, September 2016.

[21] K. Sinan, O. Tolga, and O. Yuksel, "Design and implementation of Dspic33fj32mc204 microcontroller–based asynchronous motor voltage/frequency speed control circuit for the ventilation systems of vehicles," *Measurement and Control*, vol. 52, no. 7–8, pp. 1039–1047, 2019.

[22] D. Aziz and L. Ahmed, "Real-Time simulation and analysis of the induction machine performances operating at flux constant," *(IJACSA) International Journal Of Advanced Computer Science And Applications*, vol. 5, no. 4, pp. 9–64, 2014.

[23] C. Anurag, G. Deepam, L. S. Sudha, and A. Aparna, "Condition monitoring and fault diagnosis of induction motors: a review," *Archives of Computational Methods in Engineering*, vol. 26, pp. 1221–1238, 2019.

[24] R. Narayan Dash, "Condition monitoring of induction motors:—a review," in *Proceedings of the International Conference On Signal Processing, Communication, Power And Embedded System (SCOPES)*, IEEE, Paralakhemundi, India, October 2016.

[25] K. Adlen, M. Abderrezak, K. Ridha, and B. Mohamed, "Real-time safety monitoring in the induction motor using deep hierarchic long short-term memory," *The International Journal of Advanced Manufacturing Technology*, vol. 99, no. 9–12, pp. 2245–2255, 2018.

[26] S. Manivannan, S. Veerakumar, P. Karuppusamy, and A. Nandhakumar, "Performance analysis of three phase voltage source inverter fed induction motor drive with possible switching sequence execution in SVPWM," *The International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, vol. 3, no. 6, pp. 1081–1104, 2014.

[27] A. Oteafy and J. Chiasson, "A study of the lyapunov stability of an open-loop induction machine," *IEEE Transactions on Control Systems Technology*, vol. 18, no. 6, pp. 1469–1476, 2010.

[28] A. Hmidet and O. Hasnaoui, "Waijung blockset-STM32F4 environment for real time induction motor speed control," in *IEEE 5th international congress on information science and Technology (CiSt)*, IEEE, Marrakech, Morocco, October 2018.

[29] Aimagin product, https://www.aimagin.com.

[30] Waijung Web Page Designer (W2D), http://waijung.aimagin.com.