Journal of Computational Science 82 (2024) 102397

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

# Exploring the integration of IoT and Generative AI in English language education: Smart tools for personalized learning experiences

Wanjin Dong [a,b], Daohua Pan [c], Soonbae Kim [b,*]

[a] Nanyang Medical College, Nanyang, China
[b] Chungbuk National University, Chungbuk, Republic of Korea
[c] Heilongjiang Vocational College for Nationalities, Harbin 150066, China

ARTICLE INFO

*Keywords:*
IoT
Generative AI
Smart tools
Adaptive learning environment

ABSTRACT

English language education is undergoing a transformative shift, propelled by advancements in technology. This research explores the integration of the Internet of Things (IoT) and Generative Artificial Intelligence (Generative AI) in the context of English language education, with a focus on developing a personalized oral assessment method. The proposed method leverages real-time data collection from IoT devices and Generative AI's language generation capabilities to create a dynamic and adaptive learning environment. The study addresses historical challenges in traditional teaching methodologies, emphasizing the need for AI approaches. The research objectives encompass a comprehensive exploration of the historical context, challenges, and existing technological interventions in English language education. A novel, technology-driven oral assessment method is designed, implemented, and rigorously evaluated using datasets such as Librispeech and L2Arctic. The ablation study investigates the impact of training dataset proportions and model learning rates on the method's performance. Results from the study highlight the importance of maintaining a balance in dataset proportions, selecting an optimal learning rate, and considering model depth in achieving optimal performance.

## 1. Introduction

English language education stands at the forefront of pedagogical evolution, facing challenges inherent in traditional teaching methodologies that may not adequately address the diverse and evolving needs of learners. As the global landscape of education undergoes dynamic transformations, there is an increasing recognition of the imperative to adapt and personalize pedagogical approaches. This research embarks on a transformative journey by exploring the integration of two cutting-edge technologies: the IoT [1] and Generative AI [2]. This convergence holds the potential to redefine English language education, offering nuanced and personalized learning experiences that dynamically adapt to the unique linguistic, cognitive, and cultural requirements of each learner.

The historical context of English language education reveals a persistent struggle to align teaching methods with the individualized nature of language acquisition. Traditional approaches, marked by standardized curricula and assessments, often fall short in accommodating the varied linguistic proficiency, cultural backgrounds, and learning preferences among students. This incongruity has not only led to disparities in educational outcomes but has also underscored the need for innovative, adaptive, and culturally sensitive pedagogical solutions. In the contemporary educational landscape, where globalization and multiculturalism are increasingly prominent, the limitations of traditional methods become more pronounced. English language learners span diverse demographics, each bringing a unique set of challenges and opportunities. From language learners with different native languages to those with varying degrees of exposure to technology, the spectrum of diversity requires a pedagogical shift that can dynamically adjust to individual needs.

Against this backdrop, the integration of IoT and Generative AI emerges as a promising frontier [3]. By leveraging the real-time data collection capabilities of IoT and the language generation prowess of Generative AI, this research seeks to address the multifaceted challenges faced by English language educators. The significance of this exploration extends beyond the immediate concerns of language proficiency to encompass adaptability, inclusivity, and cultural sensitivity. Generative AI, unlike traditional AI, can create new data samples rather than simply mapping input data to output labels. While traditional AI excels at tasks like image recognition and language translation, generative AI models,

* Corresponding author.
*E-mail addresses:* donglele@chungbuk.ac.kr (W. Dong), pandaohua@ftcl.hit.edu.cn (D. Pan), pearlpoet@chungbuk.ac.kr (S. Kim).

https://doi.org/10.1016/j.jocs.2024.102397
Received 15 January 2024; Received in revised form 19 June 2024; Accepted 26 July 2024
Available online 4 August 2024

such as GANs and VAEs, learn to represent the underlying structure of data and generate new, realistic samples. This capability has led to creative applications in art generation, music composition, and storytelling, but also raises ethical concerns regarding the misuse of generated content, such as deepfakes, and its impact on society and culture.

IoT facilitates the collection of granular data on student interactions with language learning materials, offering insights into individual learning patterns, preferences, and areas of struggle. This data-driven approach can enable educators to tailor instructional content in real-time, ensuring that it aligns with the specific needs and cultural contexts of each learner. Generative AI, with its ability to simulate human-like language production, holds promise in crafting personalized educational content and assessments. The potential to generate language exercises, scenarios, and assessments that resonate with individual learners can bridge the gap between standardized curricula and the diverse linguistic backgrounds of students. The significance of this research lies not only in its potential to enhance language proficiency but also in fostering a more inclusive and culturally responsive English language education. By embracing technology to personalize learning experiences, this study aims to contribute to a paradigm shift that prioritizes adaptability, inclusivity, and cultural relevance in the pursuit of effective language education.

Building on the recognition of the challenges faced by traditional English language teaching methods, this study sets forth a comprehensive set of objectives aimed at exploring the integration of IoT and Generative AI: 1. Investigating the historical context and challenges of traditional English language teaching methods, emphasizing the need for adaptive and culturally sensitive pedagogical approaches. 2. Examining the potential of IoT in real-time data collection for obtaining granular insights into individual learning patterns, preferences, and cultural contexts. 3. Scrutinizing the language generation capabilities of Generative AI, evaluating its potential for crafting personalized educational content and assessments that are culturally sensitive and adaptive. 4. Designing and implementing a novel, technology-driven, and culturally responsive English oral assessment method that incorporates insights from IoT and Generative AI. 5. Evaluating the effectiveness of the proposed method in enhancing student engagement, language acquisition, and overall learning outcomes in diverse linguistic and cultural contexts. Specifically, we use deep learning algorithms and Internet of Things devices to implement oral language assessment to correct the deficiencies in student learning in English education.

To comprehensively address the multifaceted nature of the research topic, this paper adopts a structured organization. Following the detailed introduction, the literature review section provides a nuanced understanding of the historical context, challenges, and existing technological interventions in English language education. The methodology section meticulously outlines the research design, data collection methods, and analytical approaches employed to achieve the outlined objectives. The presentation of results is thorough, with detailed analysis and interpretation, paving the way for a robust discussion that situates the findings within the broader context of existing literature. The conclusion synthesizes key insights, emphasizes contributions to the field, and propels the discourse forward by suggesting avenues for future research. Through this meticulously crafted organization, the paper aims to provide a comprehensive and insightful contribution to the ongoing dialogue on the integration of IoT and Generative AI in English language education in diverse and multicultural settings. This paper aims to combine deep learning generative AI technology and IoT technology to solve unique problems in the English pronunciation education environment, and improve teaching quality and experience by developing new technical means.

Here are the main contributions of the paper:

- We presents a generative AI-based system for spoken language assessment using Transformer models. This system accurately evaluates students' English speaking abilities and provides personalized feedback to enhance learning outcomes.
- We develops a method for generating diverse spoken English materials using generative models. This enriches students' learning resources, promoting more comprehensive oral training and practical application skills.
- We demonstrates the potential of generative AI technology in the educational field. By developing intelligent and personalized learning tools and assessment systems, it advances the modernization and innovation of English language education.

This paper is organized as follows. Section II briefly introduces the current research status of oral assessment algorithms. Section III describes the proposed method in detail, including the data collection methods, model design, and evaluation analysis methods used in this paper. Section IV outlines the experimental design, encompassing data collection, evaluation metrics, experimental environment setup, and comparative analyses. The last section concludes the full paper.

## 2. Related work

In this section, we introduce the current research status of oral assessment algorithms, including the basic knowledge of oral assessment algorithms and commonly used algorithms. For the oral assessment system, its typical teaching process includes target text, students, audio, oral assessment and error feedback. The system first provides the target text to be evaluated, and then the learner attempts to read the target text aloud. For example, the target text to be evaluated is "like" (its phoneme is "L I KE"). By accurately detecting pronunciation errors and providing evaluation feedback (for example, pointing out that the pronunciation of "I" is wrong), the oral assessment system can guide learners to correct the reading corresponding to the target text to standard pronunciation and improve their oral reading level.

With the development of speech technology, spoken language evaluation algorithms based on different speech recognition models have also made corresponding progress. This section first introduces the main speech recognition models, and then describes different spoken language evaluation algorithms. For a speech feature $X$, the task of identifying its corresponding text $Y$ can be described by a maximum posterior probability optimization problem. Hidden Markov Model (HMM) [4] maps the phonemes corresponding to the text sequence $Y$ to its HMM state, and describes the relationship between audio features and HMM state through Gaussian Mixture Models (GMM) probabilistic relationship, and then the probabilistic optimization solution can be solved. After GMM, Deep Neural Network (DNN) has demonstrated stronger modeling capabilities and gradually replaced GMM [5,6]. Considering the high performance of DNN and the cumbersome training process of HMM, researchers have begun to focus on how to use deep models to complete the entire recognition process. One of the most successful related works is the CTC (Connectionist Temporal Classification) loss function [7].

For the speech recognition task $(Y|X)$, the model based on the CTC loss function converts it into the form of probability sum. Among them, the text sequence $Y$ is first aligned to the path of the audio feature sequence $X$. By exhaustively enumerating all paths $H$ and finding their probability sum, you can get $P\ (Y\ |X)$ and optimize it. The CTC loss function introduces a filler $\phi$ for path alignment. The aforementioned HMM-based methods output text by aligning input audio features to HMM states. The CTC loss function aligns the features to the intermediate variable $H$ through fillers and self-repetition, thereby implicitly establishing the HMM state. Since the CTC loss actually completes the alignment indirectly through the HMM state, this process still uses the conditional independence assumption of the HMM. Applying such assumptions to continuous speech may result in poor speech recognition.

With the introduction of the Transformer structure [8] in the field of natural language processing, models based on the Attention mechanism

have begun to be widely used in sequence-to-sequence tasks in various fields, including computer vision [9] and speech synthesis related to audio features [10–12], speech conversion [13], and speech recognition [14], etc. Among them, the task of speech recognition can be regarded as converting from speech sequence to text sequence, aligning audio features and text information through the Attention mechanism, and then directly solving $P(Y|X)$.

Typical sequence-to-sequence models related to speech recognition include Listen, Attend and Spell (LAS) [15], and Transformer. Among them, LAS uses a bidirectional Long Short-Term Memory [16] (LSTM) network to model the temporal information of speech features and text features, and aligns the modeled audio features and text features to complete recognition. However, for Recurrent Neural Networks (RNN) such as LSTM, it is often difficult to capture long correlation features due to the nature of sequence modeling. In addition, such RNN models also suffer from low computational efficiency. The proposal of the Transformer structure partially solves the above problems. Transformer models show good performance in sequence-to-sequence modeling and achieve better performance in speech recognition tasks [17,18]. This article also mainly chooses Transformer as the basic framework of each spoken language evaluation model.

From the perspective of language learning, if a pronunciation obviously deviates from the standard pronunciation corresponding to the target text, the pronunciation is a mispronunciation. Based on this idea, early oral language evaluation algorithms adopted a scheme that compared students' pronunciation characteristics with standard teacher pronunciation characteristics [19–21]. This type of algorithm first extracts the features corresponding to standard pronunciation and the time position of each phoneme, then extracts the features corresponding to student pronunciation, and uses the Dynamic Time Warping (DTW) algorithm [22] to align the two features. Based on the distance between each phoneme unit of the standard pronunciation and the student's pronunciation characteristics, the corresponding pronunciation quality score can be calculated. The focus of such comparison-based algorithms is how to extract more robust features for alignment. In addition, the algorithm needs to provide standard speech for comparison reference. Therefore, this algorithm is suitable for oral teaching with fixed assessment content (such as pre-set courses). It is difficult to deploy when the assessment content is not fixed (for example, students want to set their own text to read).

With the development of speech recognition technology, speech modeling has become more mature, and the features extracted from audio have become more robust. Among them, the acoustic model based on HMM has been widely used, and spoken language evaluation algorithms based on this acoustic model have begun to emerge. Among them, the most classic algorithm is Goodness of Pronunciation (GOP) [23]. The original version of the GOP algorithm is based on the GMM-HMM model. The algorithm first builds an HMM acoustic model based on standard speech training, aligns the audio features to the HMM state corresponding to the target text, and combines the probability of each phoneme as the evaluation score. After exceeding a certain threshold, the phoneme is judged to be correct; otherwise, the phoneme is judged to be wrong. For a certain target phoneme $m$, given an audio feature $X$, its GOP score is defined as the logarithmic posterior probability of that phoneme. This type of HMM-based GOP spoken language evaluation algorithm requires a lot of engineering processing, including building a pronunciation dictionary, splitting words into phonemes, forming a three-phoneme model, and building a corresponding HMM for training, etc. In order to improve these processes, domain experts need to spend a lot of effort adjusting the parameters of the models at each stage. This undoubtedly makes the deployment of the spoken language evaluation algorithm more complex and limits the application of this technology. This type of algorithm is also gradually being replaced by models that perform speech recognition directly based on phonemes.

In order to solve the above-mentioned complex deployment problems of HMM-based algorithms, some spoken language evaluation algorithms began to use end-to-end speech recognition technology for spoken language evaluation [24–26]. This type of method uses a speech recognition model to obtain the phoneme actually read by the student, and obtains the error position as feedback based on the alignment result between the actual phoneme and the phoneme of the target text. In addition, there are some works that use artificial intelligence and Internet of Things technologies to solve some English education and emotion recognition problems. Ref. [27] used artificial intelligence images to model drone flight and provides an intelligent simulator for the education and training of drone pilots. Refs. [28,29] utilized convolutional neural networks for English education systems and emotion recognition processes, respectively. These algorithms have high requirements on the accuracy of speech recognition models. With the development of deep speech recognition technology, acoustic models can better extract high-level semantic features from original audio features. This type of algorithm has achieved better evaluation performance and has become the current mainstream spoken language evaluation solution.

## 3. Method

In this section, we describe the data collection methods, model design, and evaluation analysis methods used in this paper in detail. Firstly, we elaborate on the overall teaching scene design and introduce the overall environment, situation analysis methods, and interaction. Secondly, we introduce the speaking language evaluation method based on the generative AI model proposed in this article.

### 3.1. Teaching scene model design

The computer-aided oral English teaching algorithm removes the space and time constraints of traditional classroom teaching and can provide immediate feedback to students. Such feedback mechanism is particularly important for oral English learning. The final purpose of the oral evaluation algorithm is to improve the user's oral level, so good feedback should not only enable the user to judge their learning level, but also point out the location of their pronunciation errors, and further give the reason for the error and how to correct. However, most of the existing oral evaluation algorithms are discriminative, so they can only give text-based feedback for the user's current input. The feedback form of these discriminative oral evaluation algorithms is monotonous and difficult to understand. They cannot give more intuitive feedback of speech modes like oral English teachers, and it is difficult to help students improve their oral English level from the perspectives of perception and pronunciation.

The teaching interaction scenario is a special teaching experience method that aims to clearly demonstrate the task objectives. Including various teaching tasks to be completed by students, such as response messages given by the surrounding Internet of Things environment. These are all based on interactive scenarios. By delving into the details of the interaction between AI and students, we can better prompt students' problems in English learning and provide more accurate solutions to the problems. In the design process of English teaching scenarios, we construct a vivid story scenario based on data analysis of the relationship between teaching needs and student learning. The development of story scenarios requires the construction of corresponding interactive scenarios that meet teaching needs, solve student problems, and provide an effective interaction space between students and AI. After the teaching scene model is constructed, it is easy to recognize and construct intelligent teaching interaction scenes. Analyze and predict the target group through multiple elements, and use the collected data and scenario data to describe the problems existing in students' learning. The overall teaching interaction scene construction is shown in Fig. 1.

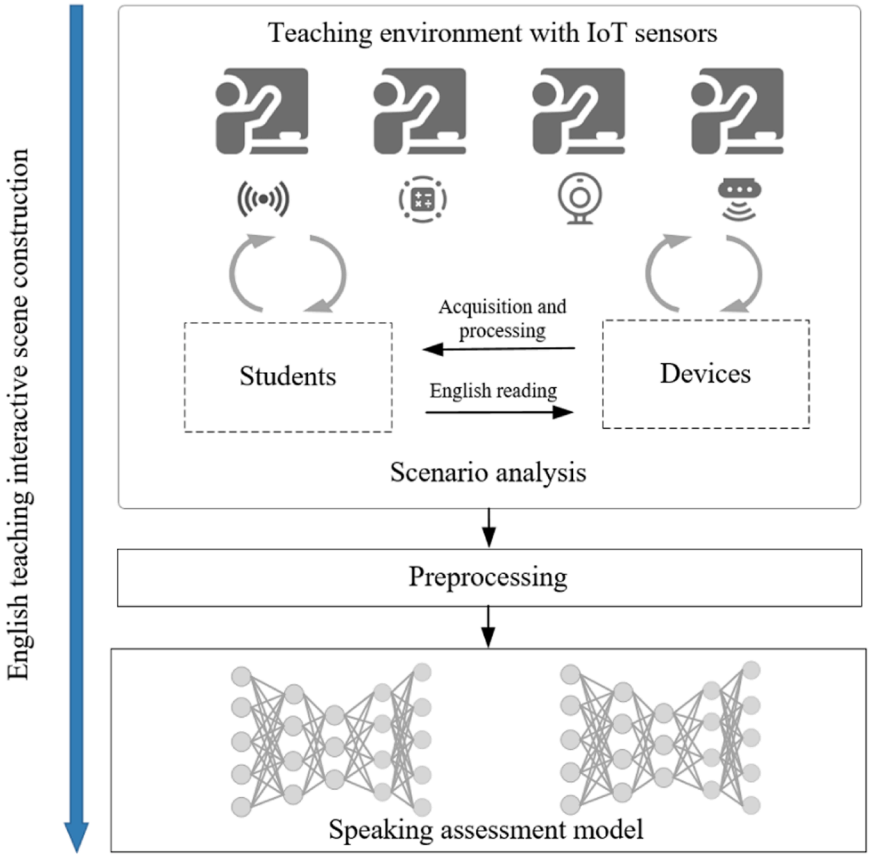


**Fig. 1.** The overall teaching interaction scene construction.

### 3.2. Speaking assessment model design

Spoken language assessment includes two parts: speech synthesis and pronunciation correction. Speech synthesis is a task that requires converting text sequences into audio feature sequences. From this perspective, the Tacotron [30] model based on the RNN structure uses an Encoder-RNN-Decoder structure similar to the previous LAS model to encode the text sequence through the Encoder. It is aligned with the input from the Decoder side by the Attention mechanism, predicts the corresponding spectrogram, and uses the Griffin-Lim algorithm [31] to convert it into a speech waveform. With the introduction of the Transformer model, sequence-to-sequence modeling performance has been greatly improved, and Transformer has also been applied to the task of speech synthesis. In addition, speech synthesis and speech recognition differ in how the model determines whether the output sequence is terminated. Speech recognition outputs a discrete sequence of text that can be stopped by predicting sentence terminators. For speech synthesis, the output is a continuous speech signal. These autoregressive deep speech synthesis models also additionally train a binary classification target of whether to terminate output.

Although the above-mentioned autoregressive deep speech synthesis model has achieved great performance improvement compared to traditional methods, there are still the following problems that need to be solved: Slow inference speed and duration robustness issues. Due to the mechanism of autoregression, the network needs to predict when to terminate the output. Such predictions are not completely accurate, so early termination may result in incomplete output, or duplicate output may occur. Although existing speech synthesis models are already well able to generate corresponding speech based on text, this goal is still different from pronunciation correction. For single-speaker speech synthesis models (such as GoogleTTS) that fix the speaker's timbre, generation based solely on text completely ignores the student's own pronunciation. The pronunciation style generated by this type of model is relatively single and boring, and students can only follow the generated pronunciation mechanically, which limits its corresponding teaching effect.

In this paper, we design a spoken language evaluation method based on the transformer model, which can achieve the purpose of pronunciation correction. We combine the ideas of denoising autoencoder and speech synthesis model [32] to design a pronunciation correction method based on acoustic units. The network structure is shown in Fig. 2. We utilized an end-to-end speech-to-text (ASR) model to directly convert speech input into text format, which was then input into the Transformer model for processing. Specifically, we employed existing speech recognition models such as to transcribe the spoken data, followed by manual verification and correction to ensure the accuracy and quality of the transcription. Additionally, we performed preprocessing steps such as tokenization, cleaning, and formatting of the transcribed text for input into the Transformer model for training or evaluation.

Since the RNN structure is not used, in order to utilize the timing information of the sequence, Transformer explicitly adds position-related features to the sequence, that is, positional encoding Positional Encoding (PE) as follows:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_k}}}\right) \tag{1}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_k}}}\right) \tag{2}$$

where *pos* corresponds to the position of the sequence feature (i.e., time

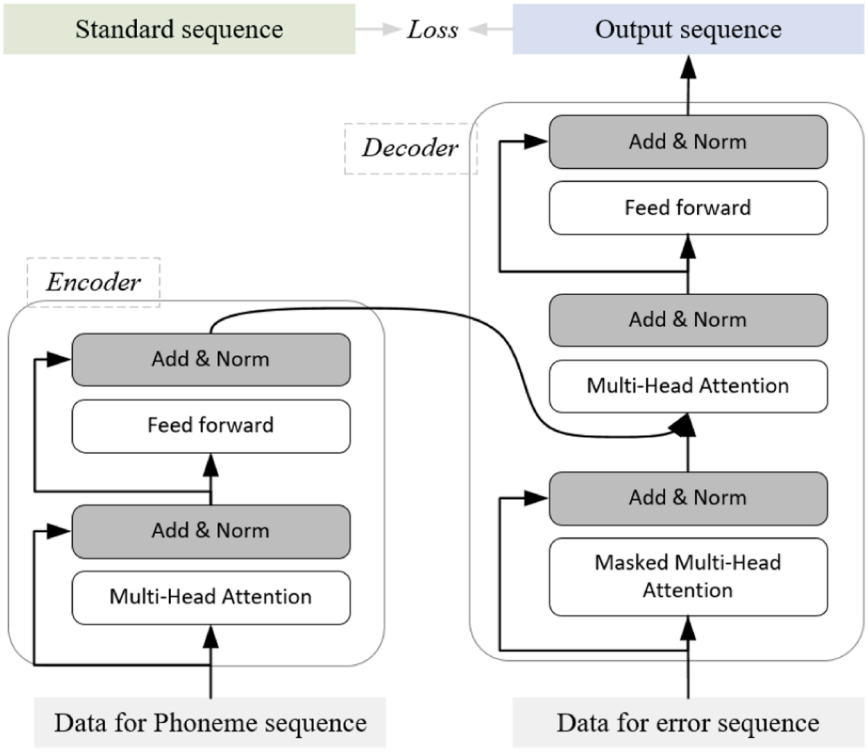


**Fig. 2.** The network structure based on Transformer model.

step), while $\nu$ corresponds to its channel dimension, and $d_k$ corresponds to the number of channels of the Attention layer. Since $Pe_{pos+k}$ can be represented by a linear function of $PE_{pos}$, by adding this $PE$ feature to the input features, the model can easily learn the time position corresponding to each sequence feature. The core of the Transformer model is its Multi-Head Attention (MHA):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

There are three inputs to the attention layer, the query matrix $Q$, the key matrix $K$ and the value matrix $V$. By calculating the correlation between $Q$ and $K$, the relevant information can be extracted from $V$ as output. The multi-head attention mechanism splices multiple Attention results on the channel as the output, that is

$$\text{MHA}(Q, K, V) = \text{concat}(hd_1, ..., hd_n)W^O \tag{4}$$

where can be calculated as follows:

$$hd_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{5}$$

where the parameter $n$ is the number of Attention heads and is set to 5, and $W^Q$, $W^K$, $W^V$, $W^O$ are respectively $Q$, $K$, $V$ and the projection matrix corresponding to the output. In the case of Self-Attention, $Q$, $K$ and $V$ are all the same. For example, in the speech recognition task, the Encoder part of Transformer uses the self-attention mechanism to extract high-level abstract features related to audio. For Cross-Attention, $K$ and $V$ are the same, but $Q$ is different. For example, the Decoder part of the Transformer in the speech recognition task sets the text feature to $Q$, which is used to query and output the speech features related to it.

The model still uses a Transformer-based architecture. During the training phase, we convert the audio in the standard dataset into a sequence of acoustic units $E_n$ and randomly replace them according to the misreading. That is, based on the original acoustic units $e_n$, sorted by their semantic distance, and randomly taking out a nearby new unit $e'$ as a replacement, a new sequence $E'$ is generated. Based on the input phoneme $P$, the model attempts to retain the correct sequence of acoustic units based on $E'$, modify the incorrect sequence, and output the modified sequence $E_r$. The acoustic units obtained through self-supervised model training can better model standard pronunciation and non-standard pronunciation. Therefore, the modified sequence $E'$ may also be converted into a pronunciation that does not match the text, thereby simulating misreading speech input during the inference process. After such a modification operation, with the phoneme $P$ and the modified sequence $E'$ as input, and the original standard sequence $E_n$ as the output target, the model can learn in a self-supervised manner. Therefore, the model only requires standard datasets for training and no longer requires paired corpora. We use the cross-entropy loss function to recover the original canonical sequence, calculated as follows.

$$L_{rct} = \text{CrossEntropy}(E_r, E_n) \tag{6}$$

In fact, the model proposed in this article can be regarded as a special Transformer-based speech synthesis model. The difference is that this model uses the acoustic unit sequence as the input of the Decoder, and directly modifies the original sequence in a feedforward manner to generate the final output. The original Transformer-based speech synthesis model uses spectrum as Teacher-Forcing input for autoregressive prediction. Since misreadings mainly occur between similar pronunciations, the difference between the original sequence and the final output sequence is not large, so this feedforward output method is reasonable. In fact, this idea has also appeared in speech recognition. For example, the MaskCTC model [33] uses a model based on the CTC loss function to predict an initial sequence to replace the Teacher-Forcing input. On this basis, the Attention model can correct the initial sequence to obtain a sequence with higher accuracy.

## 4. Experiments

The experimental phase of this research constitutes a crucial component in substantiating the proposed integration of the IoT and Generative AI within the context of English language education. This section outlines the experimental design, encompassing data collection, evaluation metrics, experimental environment setup, and comparative

analyses. Each facet of the experiment is meticulously crafted to shed light on the efficacy and potential of the proposed personalized English oral assessment method.

### 4.1. Implementation details

The software environment for our experiment is configured with Python 3.8 as the primary programming language, supported by PyTorch 1.9 as the machine learning frameworks for the development, training, and deployment of Generative AI models. The Generative AI models are deployed on hardware featuring an Intel Core i7–10700K CPU, NVIDIA GeForce RTX 3080 GPU, and 32GB DDR4 RAM to ensure efficient processing and adaptation based on the datasets. This comprehensive hardware and software configuration establishes a dynamic and technologically-enhanced experimental environment, aligning with the objectives of evaluating our proposed personalized English oral assessment method. The specific parameters of the Transformer model include: the number of encoding layers is 4, the number of decoding layers is 4, the attention dimension is 512, the forward vector dimension is 1024, and the number of attention heads is 4.

### 4.2. Dataset and metrics

In our experiment, we utilize the Librispeech [34] and L2Arctic [35] datasets. The Librispeech dataset serves as a foundational component in our experimental design, offering a rich and diverse collection of English speech data. The Librispeech dataset comprise over 1000 h of read English speech from a broad array of sources, of which 960 hours of 16 kHz are read speech by native English speakers. Librispeech spans a variety of accents, dialects, and linguistic contexts. This dataset is particularly valuable for evaluating the language comprehension and fluency aspects of the proposed personalized English oral assessment method. Its extensive coverage ensures that the model is exposed to a wide spectrum of linguistic nuances, contributing to the robustness and adaptability of the system. The L2Arctic dataset supplements our experimental framework by introducing a layer of linguistic diversity. This dataset focuses on the study of non-native English accents, providing recordings from speakers representing various first language backgrounds. These non-native accents present unique challenges in language comprehension, allowing us to assess the model's efficacy in understanding and adapting to diverse speech patterns. Incorporating the L2Arctic dataset enriches our research by simulating scenarios encountered in English language classrooms with learners from different linguistic backgrounds.

We use the Kaldi speech processing toolkit [36] to extract 80-dimensional Fbank features of audio files as model input. We utilized data augmentation techniques and transfer learning to address the scarcity and diversity issues of spoken language data. Specifically, we employed audio augmentation techniques and text augmentation techniques to augment the spoken language dataset, thereby increasing its diversity and richness. Additionally, we utilized pre-trained model parameters from other related tasks through transfer learning to enhance the model's performance and generalization ability on the spoken language assessment task. The summary of data set division is shown in Table 1. This experiment uses Librispeech as a dataset for speech recognition pre-training tasks. The 960 hours of 16 kHz native English speaker reading speech is divided into two parts: clean (460 hours) and other (500 hours) according to the difficulty of recognition. The clean part is further divided into 100-hour and 360-hour training, validation, and test sets. Since the spoken language evaluation task focuses on phoneme-level errors, this paper first uses the Montreal forced alignment tool [37] to convert the dataset from word-level to phoneme-level labels.

To rigorously assess the performance of our personalized English oral assessment method on the Librispeech and L2Arctic datasets, we employ a set of comprehensive evaluation metrics used in many classic evaluation tasks, including precision (Pre), recall (Rec), accuracy (Acc), specificity (Spe), and F1 score. These metrics offer a nuanced understanding of the model's proficiency in various linguistic and cultural dimensions. These evaluation metrics collectively provide a robust framework for quantifying the accuracy, precision, recall, and overall efficacy of the personalized English oral assessment method across diverse linguistic and cultural contexts represented by the Librispeech and L2Arctic datasets. The subsequent sections will delve into the experimental results, enabling a comprehensive analysis of the model's performance and its implications for personalized English language education.

**Table 1**
The summary of data set division on two datasets.

| Dataset | Split | Corpus | Person |
|---|---|---|---|
| Librispeech | Train | 261,256 | 5200 |
| | Val | 2607 | 40 |
| | test | 2183 | 40 |
| L2-Arctic | Train | 2500 | 20 |
| | Val | 200 | 15 |
| | test | 900 | 6 |

### 4.3. Performance comparison

In order to evaluate the performance of the model proposed in this article, we compared and analyzed the following five models. The first model is a mainstream phoneme-level automatic speech recognition (ASR) model based on HMM, denoted as HMM-ASR. The second model is a text prior-based phoneme-level speech recognition model (TC-ASR). The third type is the generative model proposed in this article, denoted as G-ASR. The remaining two methods are traditional speech synthesis models, including GoogleTTS and the multi-speaker version of GlowTTS. Table 2 shows the comparison results of different methods on two data sets. In order to evaluate the performance of the single-stage pronunciation correction model based on acoustic units, the output sequence of the model is first converted into an audio waveform by the Unit2Audio module. For the Transformer model used by the Uni2Audio module, we used 6 layers of Encoder and set its convolution kernel size to 7. The spectrum generated by all the above models is encoded by the same MelGAN vocoder [38]. First, we observe the F1 index in Table 2. The GoogleTTS and GlowTTS models based on speech synthesis can directly generate standard pronunciation based on the target text, so the F1 of such models is relatively high. On the standard LibriTTS data set, due to the loss of the Unit2Audio model itself, the Rec and Pre indicators are low due to re-synthesis. On the L2-Arctic data set, the model proposed in this article successfully converted L2 pronunciation into standard pronunciation and achieved optimal results, which is close to the performance of the speech synthesis model GlowTTS. Compared with other models, our method G-ASR achieves the best performance, which also reveals the effectiveness of the method proposed in this article. On the other hand, since GoogleTTS only uses text for generation and

**Table 2**
The comparison results of different methods on two datasets.

| Dataset | Model | Pre | Rec | Acc | Spe | F1 |
|---|---|---|---|---|---|---|
| LibriTTS | HMM-ASR | 0.534 | 0.545 | 0.864 | 0.604 | 0.539 |
| | TC-ASR | 0.647 | 0.658 | 0.883 | 0.724 | 0.652 |
| | GoogleTTS | 0.668 | 0.632 | 0.872 | 0.691 | 0.650 |
| | GlowTTS | 0.794 | 0.742 | 0.924 | 0.786 | 0.767 |
| | G-ASR (ours) | 0.735 | 0.761 | 0.945 | 0.815 | 0.748 |
| L2-Arctic | HMM-ASR | 0.486 | 0.438 | 0.761 | 0.568 | 0.461 |
| | TC-ASR | 0.517 | 0.494 | 0.748 | 0.583 | 0.505 |
| | GoogleTTS | 0.523 | 0.593 | 0.816 | 0.682 | 0.556 |
| | GlowTTS | 0.676 | 0.572 | 0.806 | 0.737 | 0.620 |
| | G-ASR (ours) | 0.634 | 0.583 | 0.864 | 0.784 | 0.607 |

ignores the audio input by students, its style loss is the most serious, so its F1 value is low and other related indicators are not as good as the method proposed in this article. GlowTTS can partially clone the timbre of the original speaker, so its F1 value increases. In addition, GlowTTS only uses X-vector to clone the speaker's timbre without retaining the correct pronunciation in the original input. The output audio rhythm has a greater change compared to the original input. In contrast, the model proposed in this article can complete pronunciation correction to varying degrees and can retain the correct part of the original pronunciation.

### 4.4. Ablation study

To comprehensively evaluate the impact of key components on the performance of our personalized English oral assessment method, we conducted an ablation study, systematically varying parameters related to training dataset proportions, model learning rates, and model depths. This study aims to elucidate the significance of these factors in shaping the model's effectiveness. The results of the ablation study provide a nuanced understanding of the personalized English oral assessment method's sensitivity to variations in training dataset proportions, model learning rates, and model depths. Through a thorough analysis of the experimental outcomes, we can identify key factors that significantly impact the model's efficacy and tailor the implementation to optimize performance in diverse educational settings.

(1) Training Dataset Proportions: In this experiment, we varied the proportions of the Librispeech and L2Arctic datasets during model training. Specifically, we trained the model using different ratios of native to non-native English speech data. By systematically altering the dataset proportions, we aim to assess how the diversity in linguistic backgrounds influences the model's adaptability and performance across varied learning scenarios. Figs. 3 and 4 respectively show the impact of different training ratios of the two data sets on model performance. It can be seen from the figure that increasing the proportion of the training data set can further improve the performance of spoken language evaluation, and the method proposed in this article can achieve the best F1 score when using each proportion of the training data set. When using all training sets, this method achieved an F1 score of more than 0.6 on the L2-Arctic data set and an F1 score of more than 0.7 on the LibriTTS data set. Even if only 20 % of the training data set is used, this method still obtains a better F1 score, which proves the effectiveness of this oral language evaluation training scheme based on the model proposed in this article.

(2) Model Learning Rates: The learning rate is a critical

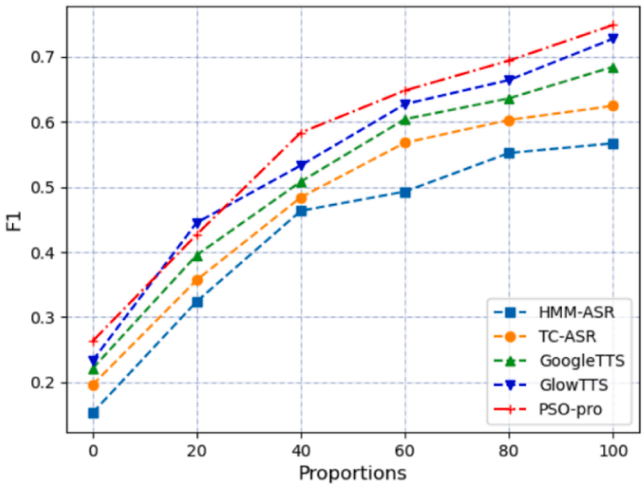


**Fig. 3.** The impact of different training data proportions on the F1 values for the LibriTTS dataset.

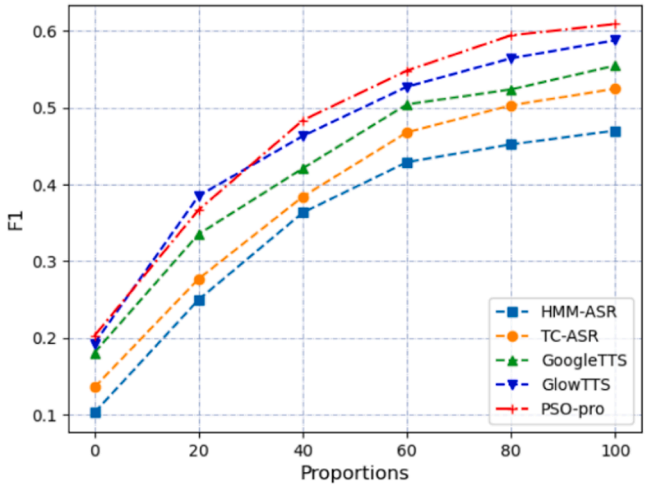


**Fig. 4.** The impact of different training data proportions on the F1 values for the L2-Arctic dataset.

hyperparameter influencing the speed and convergence of model training. We conducted experiments with varying learning rates to understand their impact on the personalized English oral assessment method's accuracy, precision, and recall. This analysis provides insights into the optimal learning rate that balances efficient convergence with model performance. Table 3 shows the comparison results of model performance under different learning rates. The experiment suggests that a learning rate of 0.001 performs optimally, striking a balance between convergence speed and model performance. Higher learning rates lead to faster convergence but may compromise precision, while lower rates result in slower convergence and reduced recall.

(3) In addition, in order to depict the audio correction results more vividly, we display the spectrograms generated by each model, as shown in Fig. 5. Comparing the original input Fig. 5(a) and the HMM-ASR model synthesis (Fig. 5(b)), it can be seen that the model cannot retain the semantic and rhythm information of the original input. After completing the correction of the model TC-ASR model, the speaking style of the original input is retained (Fig. 5(c)). As shown in Fig. 5(d), models that only use text such as GoogleTTS completely discard the original speaker's information. In contrast, as shown in Fig. 5(e), the multi-speaker speech synthesis model GlowTTS can only partially clone the timbre of the original speaker and cannot retain information such as rhythm. As shown in Fig. 5(f), the proposed model can retain the semantic and rhythm information of the original input, and its spectrogram has basically not changed significantly.

## 5. Conclusion

In conclusion, this paper presented a pioneering exploration into the integration of the IoT and Generative AI to optimize English language education. The personalized oral assessment method developed in this study represented a significant stride towards addressing the limitations

**Table 3**
Comparison of model performance under different learning rates.

| Dataset | Learning rate | Pre | Rec | Acc | F1 |
|---|---|---|---|---|---|
| LibriTTS | 0.1 | 0.65 | 0.68 | 0.78 | 0.66 |
| | 0.01 | 0.67 | 0.74 | 0.86 | 0.70 |
| | 0.001 | 0.74 | 0.76 | 0.92 | 0.75 |
| | 0.0001 | 0.72 | 0.72 | 0.87 | 0.72 |
| L2-Arctic | 0.1 | 0.55 | 0.51 | 0.79 | 0.53 |
| | 0.01 | 0.58 | 0.56 | 0.82 | 0.57 |
| | 0.001 | 0.63 | 0.58 | 0.86 | 0.61 |
| | 0.0001 | 0.61 | 0.55 | 0.83 | 0.58 |

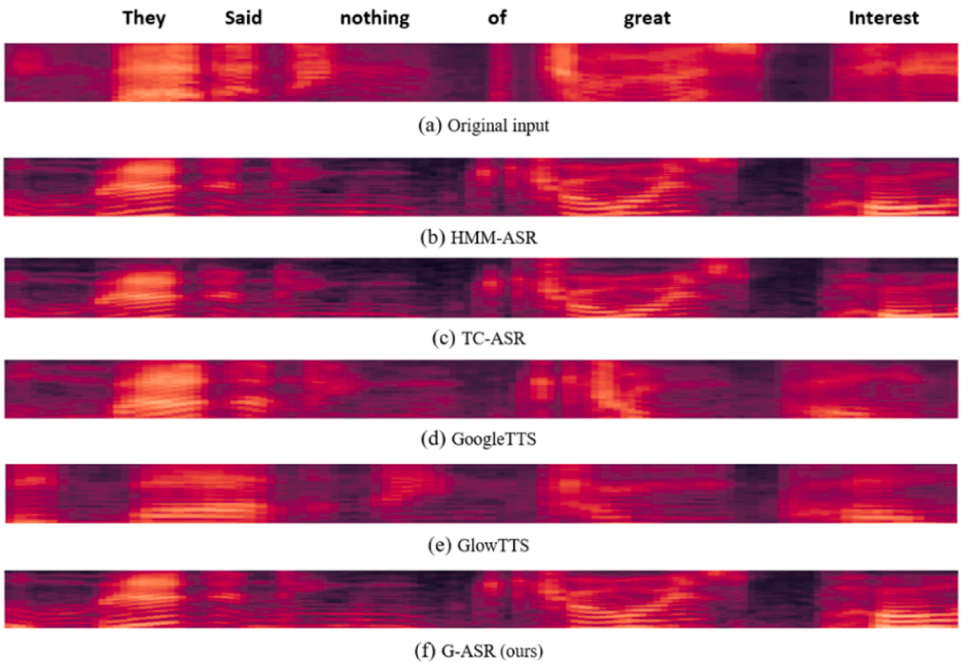


**Fig. 5.** Pronunciation spectrograms generated by different models.

of conventional teaching methodologies. By leveraging data collection from IoT devices and the linguistic diversity embedded in data, our method offered a dynamic study approach to language learning. The integration of IoT and Generative AI not only enriched language learning experiences but also fostered inclusivity by addressing the linguistic diversity encountered in modern classrooms. The findings of this research contribute to the ongoing discourse on technology-driven language education, offering a pathway for the development of adaptive, personalized, and culturally sensitive approaches. Furthermore, the findings contribute valuable insights to the ongoing discourse on the integration of IoT and Generative AI in English language education. This research aimed to advance the pedagogical landscape, fostering a more inclusive and effective approach to English language education in diverse and multicultural settings. The proposed method has several limitations: it depends heavily on the quality and quantity of training data and requires substantial computational resources. Additionally, there are potential difficulties in generalizing across different accents and dialects. Our future work will merge generative AI with English language education, particularly focusing on spoken language assessment. We aim to develop personalized learning tools and assessment systems, leveraging generative models to create diverse spoken English materials and provide accurate feedback on students' oral proficiency. This integration of technology with education will drive advancements in English language learning.

## CRediT authorship contribution statement

**Wanjin Dong:** Writing – original draft, Methodology, Conceptualization. **Daohua Pan:** Software, Investigation. **Soonbae Kim:** Writing – review & editing, Resources, Formal analysis, Conceptualization.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] S. Madakam, V. Lake, V. Lake, et al., Internet of Things (IoT): A literature review, J. Comput. Commun. 3 (05) (2015) 164.
[2] D. Baidoo-Anu, L.O. Ansah, Education in the era of generative artificial intelligence (AI): understanding the potential benefits of ChatGPT in promoting teaching and learning, J. AI 7 (1) (2023) 52–62.
[3] B. ChimbgaExploring the Ethical and Societal Concerns of Generative AI in Internet of Things (IoT) Environments, in: Southern African Conference for Artificial Intelligence Research. Cham: Springer Nature Switzerland, 2023: 44-56. .
[4] W.J. Yang, J.C. Lee, Y.C. Chang, et al., Hidden Markov model for Mandarin lexical tone recognition, IEEE Trans. Acoust., Speech, Signal Process. 36 (7) (1988) 988–992.
[5] A. Mohamed, G. HintonPhone recognition using restricted boltzmann machines, in: Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2010: 4354-4357. 2010.
[6] D. Povey, L. Burget, M. Agarwal, Subspace Gaussian mixture models for speech recognition, in: Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2010: 4330-4333. .
[7] A. Graves, S. Fernández, F. Gomez, Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, in: Proceedings of the 23rd International Conference on Machine Learning2006, , 369–376..
[8] A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need, Adv. Neural Inf. Process. Syst. (2017) 30.
[9] N. Carion, F. Massa, G. Synnaeve, End-to-end object detection with transformers, in: Proceedings of European conference on computer vision. Cham: Springer International Publishing, 2020: 213-229. .
[10] L. Dong, S. Xu, B. XuSpeech-transformer: a no-recurrence sequence-to-sequence model for speech recognition, in: Proceedings of the 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018: 5884-5888. .
[11] T. Okamoto, T. Toda, Y. Shiga, Transformer-based text-to-speech with weighted forced attention, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 6729-6733. .
[12] Y. Ren, Y. Ruan, X. Tan, et al., Fastspeech: Fast, robust and controllable text to speech, Adv. Neural Inf. Process. Syst. (2019) 32.
[13] R. Liu, X. Chen, X. WenVoice conversion with transformer network, in: Proceedings of ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 7759-7759. .
[14] Moritz N., Hori T., Le J. Streaming automatic speech recognition with the transformer model, in: Proceedings of ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 6074-6078.
[15] W. Chan, N. Jaitly, Q. Le, Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, in: Proceedings of the 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2016: 4960-4964. .
[16] L.S.T. Memory, Long short-term memory, Neural Comput. 9 (8) (2010) 1735–1780.
[17] Zhang Q., Lu H., Sak H., et al. Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss, in: Proceedings of

ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 7829-7833.
[18] Watanabe S., Hori T., Karita S., et al. Espnet: End-to-end speech processing toolkit. arXiv preprint arXiv:1804.00015, 2018.
[19] A. Lee, Y. Zhang, J. GlassMispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams, in: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013, pp. 8227–8231. .
[20] A. Lee, N.F. Chen, J. GlassLee A., Chen N.F., Glass J. Personalized mispronunciation detection and diagnosis based on unsupervised error pattern discovery, in: Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016: 6145–6149 2016.
[21] A. Lee, J. GlassA comparison-based approach to mispronunciation detection, in: Proceedings of the 2012 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2012: 382-387. .
[22] Berndt D.J., Clifford J. Using dynamic time warping to find patterns in time series, in: Proceedings of the 3rd international conference on knowledge discovery and data mining. 1994: 359-370.
[23] Witt S.M. Use of speech recognition in computer-assisted language learning. University of Cambridge, 2000.
[24] W.K. Leung, X. Liu, H. MengCNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis, in: Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 8132-8136. 2019.
[25] L. Zhang, Z. Zhao, C. Ma, et al., End-to-end automatic pronunciation error detection based on improved hybrid ctc/attention architecture, Sensors 20 (7) (2020) 1809.
[26] B.C. Yan, M.C. Wu, H.T. Hung, et al., An end-to-end mispronunciation detection system for L2 English speech leveraging novel anti-phone modeling, arXiv Prepr. arXiv:2005 (2020) 11950.
[27] J. Won, G. Hong, Research on Smart Construction Education Training Contents Using a Drone Simulator, J. Multimed. Inf. Syst. 9 (4) (2022) 345–354.
[28] Y. Zhang, J. Cao, Design of English teaching system using Artificial Intelligence, Comput. Electr. Eng. 102 (2022) 108115.
[29] Y.J. Choi, Y.W. Lee, B.G. Kim, Residual-based graph convolutional network for emotion recognition in conversation for smart Internet of Things, Big Data 9 (4) (2021) 279–288.
[30] Wang Y., Skerry-Ryan R.J., Stanton D., et al. Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135, 2017.
[31] D. Griffin, J. Lim, Signal estimation from modified short-time Fourier transform, IEEE Trans. Acoust. Speech, Signal Process. 32 (2) (1984) 236–243.
[32] N. Li, S. Liu, Y. Liu, et al., Neural speech synthesis with transformer network[C], Proc. AAAI Conf. Artif. Intell. 33 (01) (2019) 6706–6713.
[33] Y. Higuchi, S. Watanabe, N. Chen, et al., Mask CTC: Non-autoregressive end-to-end ASR with CTC and mask predict, arXiv Prepr. arXiv:2005. 08700 (2020).
[34] V. Panayotov, G. Chen, D. Povey, Librispeech: an asr corpus based on public domain audio books, in: Proceedings of the 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp. 5206-5210. 2015.
[35] Zhao G., Sonsaat S., Silpachai A., et al. L2-ARCTIC: A non-native English speech corpus//Interspeech. 2018: 2783-2787.
[36] Povey D., Ghoshal A., Boulianne G., et al. The Kaldi speech recognition toolkit [C]//IEEE 2011 workshop on automatic speech recognition and understanding. IEEE Signal Processing Society, 2011 (CONF).
[37] McAuliffe M., Socolof M., Mihuc S., et al. Montreal forced aligner: Trainable text-speech alignment using kaldi//Interspeech. 2017, 2017: 498-502.
[38] K. Kumar, R. Kumar, T. De Boissiere, et al., Melgan: Generative adversarial networks for conditional waveform synthesis, Adv. Neural Inf. Process. Syst. (2019) 32.