Text-to-Speech for Low-Resource Agglutinative Language With Morphology-Aware Language Model Pre-Training

Rui Liu, *Member, IEEE*, Yifan Hu, Haolin Zuo, Zhaojie Luo, *Member, IEEE*, Longbiao Wang, *Member, IEEE*, and Guanglai Gao

***Abstract*—Text-to-Speech (TTS) aims to convert the input text to a human-like voice. With the development of deep learning, encoder-decoder based TTS models perform superior performance, in terms of naturalness, in mainstream languages such as Chinese, English, etc. Note that the linguistic information learning capability of the text encoder is the key. However, for TTS of low-resource agglutinative languages, the scale of the <text, speech> paired data is limited. Therefore, how to extract rich linguistic information from small-scale text data to enhance the naturalness of the synthesized speech, is an urgent issue that needs to be addressed. In this paper, we first collect a large unsupervised text data for BERT-like language model pre-training, and then adopt the trained language model to extract deep linguistic information for the input text of the TTS model to improve the naturalness of the final synthesized speech. It should be emphasized that in order to fully exploit the prosody-related linguistic information in agglutinative languages, we incorporated morphological information into the language model training and constructed a morphology-aware masking based BERT model (MAM-BERT). Experimental results based on various advanced TTS models validate the effectiveness of our approach. Further comparison of the various data scales also validates the effectiveness of our approach in low-resource scenarios.**

***Index Terms*—Text-to-speech (TTS), agglutinative, morphology, language modeling, pre-training.**

I. Introduction

SPEECH synthesis or Text-to-Speech (TTS) [1] seeks to generate human-like speech from input text, which attracts broad interest in the audio and speech processing community. TTS also is a key component in various applications such as providing navigation directions in smartphones and cars, and interactive interfaces such as smart assistants [2]. Typical TTS models [3], [4], [5], [6] consist of a complex pipeline, including a natural language parser, duration model, acoustic model, and vocoder. Each of these modules is individually trained and then stitched together to form the final TTS system. However, errors generated by the previous modules accumulate and are passed on to the subsequent modules, thus affecting the final performance [7].

With the development of deep learning techniques, end-to-end (E2E) TTS systems have become viable and serve as the mainstream approach [8]. In general, the overall framework is composed of a text encoder, acoustic decoder, and the vocoder. The text encoder aims to extract the linguistic knowledge from the input text automatically. Acoustic decoder seeks to learn the mapping between linguistic knowledge and the acoustic feature of speech. The vocoder then transforms the acoustic feature into the speech waveform. Among them, the Tacotron 1/2 [9], [10], Transformer TTS [11] and FastSpeech 1/2(s) [12], [13] series models are some typical representatives. For example, the main structure of the Tacotron2 model is the Recurrent neural Network (RNN) based encoder and decoder. Transformer TTS adopted the self-attention mechanism [14] to improve the encoder and decoder, in order to learn the global context and accelerate the training. FastSpeech 2(s) model also consists of self-attention based encoder and decoder, and employs a non-autoregressive decoding strategy to achieve parallel inference. Coupled with the addition of a neural network-based vocoder, such as WaveNet [15], WaveRNN [16], WaveGAN [17], HiFi-GAN [18], etc., these models exhibit effect that are indistinguishable from those of real people. All the above works have excellent performance in speech naturalness, and it is worth mentioning that the linguistic information mining ability of the text encoder is crucial to the naturalness performance. However, training these neural TTS models requires large amounts of text labeled speech corpora for high-quality speech generation. For example, LJSpeech [19], a public single speaker corpus for TTS, consists of more than 20 hours of delicately recorded speech with transcriptions. With such natural advantages as large-scale <text, speech> paired data, TTS in mainstream languages has made great progress. We find that the TTS model still suffers

Manuscript received 23 July 2023; revised 8 November 2023; accepted 19 December 2023. Date of publication 1 January 2024; date of current version 18 January 2024. The work of Rui Liu was supported in part by the Young Scientists Fund of the National Natural Science Foundation of China under Grant 62206136 and in part by Guangdong Provincial Key Laboratory of Human Digital Twin under Grant 2022B1212010004, and in part by "One Zone, Two Bases" Supercomputing Capability Construction Project (Inner Mongolia University), under Grant 21300-231510. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zhizheng Wu. *(Corresponding author: Zhaojie Luo.)*

Rui Liu, Yifan Hu, Haolin Zuo, and Guanglai Gao are with the Department of Computer Science, Inner Mongolia University, Hohhot 010021, China (e-mail: liurui_imu@163.com; hyfwalker@163.com; zuohaolin_0613@163.com; csggl@imu.edu.cn).

Zhaojie Luo is with SANKEN, Osaka University, Osaka 567-0047, Japan (e-mail: luozhaojie@hotmail.com).

Longbiao Wang is with the Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin 300072, China (e-mail: longbiao_wang@tju.edu.cn).

Digital Object Identifier 10.1109/TASLP.2023.3348762

from the low-resource and morphologically-rich scenarios of agglutinative language, including Turkish, Kazakh, Mongolian, etc., TTS tasks.

The TTS research of low-resource agglutinative languages suffers from *insufficient linguistic knowledge encoding* issue. Specifically, large-scale <text, speech> paired data for low-resource agglutinative languages is difficult to obtain. Collecting high-quality alignment data is time-consuming and laborious, thus limiting adequate learning of model parameters. Therefore, the text encoder has difficulty learning rich linguistic information based on text encoding information extracted from data of limited size, which ultimately limits the naturalness performance of synthesized speech. On the one hand, we found inspiration for solving the above issue from the field of natural language processing (NLP) fortunately. 1) It is relatively easy to collect unlabeled large-scale text data, which can be collected from social media, news websites, and other domains on various topics. 2) Self-supervised language model training based on large-scale unlabeled text can learn a wealth of linguistic knowledge [20]. On the other hand, agglutinative language has a complex morphological structure, in which words can be decomposed into root and suffixes [21]. It has been demonstrated that such morphology information is relevant for the prosody naturalness of speech [21]. Therefore, whether it is possible to collect large-scale unlabeled texts of low-resource sticky languages for pre-training of self-supervised language models, and then transfer the linguistic knowledge to the text encoder to enhance the text encoder's ability to mine linguistic information, thus improving the naturalness of the synthesized speech, is the focus of our research.

In this paper, we propose a new TTS framework for low-resource agglutinative language with morphology-aware language model pretraining. Specifically, we first pre-train a BERT-like language model with a novel **M**orphology-**A**ware **M**asking strategy, denoted as **MAM-BERT**. Unlike the traditional random masking strategy, the morphology-aware masking strategy fully considers the morphology information within the text when modeling the language model, richer morphology information can be learned for the input words, thus facilitating the learning of subsequent prosody naturalness of TTS. After that, the trained language model is treated as an auxiliary linguistic encoder to extract the linguistic information from the input text. At last, the outputs from the auxiliary linguistic encoder and original text encoder are fused together and sent to the TTS decoder to produce synthesized speech with a high degree of prosody naturalness. We take the Mongolian language as the research object. A series of subjective and objective experiments, based on various advanced E2E TTS models (such as Tacotron2, Transformer TTS, FastSpeech2, etc.), have demonstrated the effectiveness of our approach. Further comparison of the various data scales also validates the effectiveness of our approach in low-resource scenarios.

The main contributions of this study are summarized as follows: 1) We try to transfer linguistic knowledge from pre-trained language models in unlabeled text data to TTS model to improve the naturalness of synthesized speech for low-resource agglutinative languages; 2) We propose a new TTS framework for low-resource agglutinative language with a novel morphology-aware masking strategy based language model, that serves as an auxiliary linguistic encoder; 3) Experimental results on the Mongolian language, a typical representative of agglutinative language, with various advanced TTS models, validate our method.

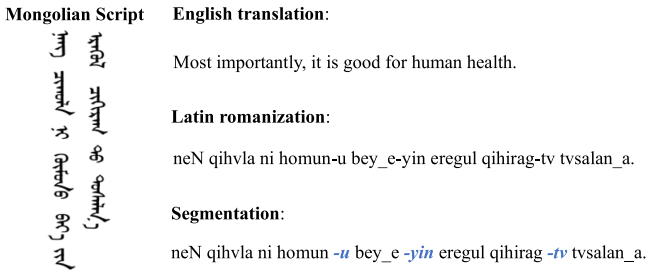


Fig. 1. Example of Mongolian words of sentence "neN qihvla ni homun-u bey_e-yin eregul qihirag-tv tvsalan_a". There are 3 NNBS suffixes, as shown in the blue part in the figure, in a sentence with only 8 words.

In the rest of this paper, we first briefly introduce the related works in Section II. The methodology of the proposed method is introduced in Section III. After that, we present the experimental setup in Section IV, which includes the dataset, the baseline, and the implementation details. We show all the experiment results and conduct in-depth analyses in Section V. Finally, we conclude this paper and discuss future work in Section VI.

II. Related Works

A. Characteristic of Low-Resource Agglutinative Language

Low-resource agglutinative languages have a complex morphological structure, in which words can be decomposed into the root and various suffixes [22]. Turkish, Japanese, Korean, and Mongolian are some typical representatives. In the following, we introduce the morphological characteristics of the agglutinative language using Mongolian as an example.

Mongolian is the most widely spoken and best-known member of the Mongolic language family, which is a group of languages spoken in East-Central Asia [23]. Approximately 6 million people speak Mongolian around the world. Mongolian has two types of written script: in China, we have classical Mongolian; in Mongolia, which is a country near the Inner Mongolia Autonomous Region, we have Cyrillic Mongolian [24], [25]. In this work, we consider only classical Mongolian script. This script and its Latin letters are shown in Fig. 1, which is written from top to bottom, left to right. Note that most Mongolian words can be decomposed into root, derivational suffixes and inflectional suffixes [22]. The first two components together are called a word stem, which holds the major information contained in a word, and inflectional suffixes serve to discriminate words based on lexical meaning. For nouns, inflectional suffixes contain case suffixes, reflexive suffixes and plural suffixes. These three types of suffixes are attached to the stem through a narrow nonbreaking space (NNBS) (U+202F, Latin: "-"); therefore, we call such suffixes NNBS suffixes. NNBS suffix use is pervasive.

For example, there are 3 NNBS suffixes in a sentence with only 8 words in Fig. 1.

More importantly, the morphology information of the agglutinative language performs an explicit relationship with the prosody expression of its speech. For example, Liu et al. [21], [26], [27] first split the Mongolian word into subword units, which are stem and suffixes, then use the subword embedding to predict the phrase break label. Experimental results show that morphology information is a positive indicator for the phrase break of low resource agglutinative language, which is an important metric of speech prosody expression [21]. The above observations provide a solid basis for our work.

B. TTS for Low-Resource Languages

Data augmentation and transfer learning have been widely used in TTS for low-resource languages [28], [29], [30], or when the labeled dataset is insufficient [31].

In general, data augmentation [32] aims to create synthetic speech data for the target speaker by multi-speaker TTS or speaker voice conversion (VC) [33]. For example, the authors in [34], [35] applied the CopyCat VC model [36] to generate expressive synthetic speech in low-resource data settings to increase the amount of target speaker data. However, the data augmentation methods still need to train an additional TTS or VC model, bringing a new burden for low-resource languages.

In transfer learning, a neural network model is firstly pre-trained for indirectly related objectives, and then the pre-trained model is used for parameter initialization of fine-tuning or feature extraction [37], [38]. For low-resource TTS, researchers focus on cross-language or multilingual pre-training methods for transfer learning [39]. For example, Chen et al. [40] first pre-train a TTS model by leveraging data from English (source) language, and then try to adapt it to low-resource (target) languages (such as German and French). Furthermore, Marcel et al. [41] found that multilingual modeling can increase the naturalness of low-resource language speech, and showed that multilingual models can produce speech with a naturalness comparable to monolingual multi-speaker models. However, the space mismatch across languages [42], [43] may bring a huge challenge and limit the performance.

Unlike the aforementioned methods, this work does not need to train relevant TTS or VC models and does not receive the distress of language space mismatch issues. We use a large amount of text data from the target language for self-supervised language model training, which can be easily implemented for knowledge transfer.

C. TTS With Language Modeling Pre-Training

As mentioned before, the linguistic information mining ability of the text encoder is crucial to the naturalness performance of TTS. There are existing attempts to use the BERT-like language model in the text encoder of TTS [44], [45], [46], [47], [48]. For example, Tomiki et al. [49] and Xiao et al. [50] employed pre-trained BERT models with Tacotron2-based TTS models, and showed gains in overall speech naturalness. Kenter et al. [51] proposed an RNN-based prosody model that incorporates a BERT model, fine-tuned during training, to improve the prosody of synthesized speech. Zhang et al. [45] designed two context encoders, i.e., a sentence window context encoder and a paragraph-level context encoder, to integrate the contextual representations extracted from multiple sentences by BERT into Tacotron2 via an extra attention module.

All mentioned works have shown that BERT models can incorporate rich linguistic information in the text-only domain. However, the BERT model only implicitly captures the linguistic structure of the input text. In other words, these methods can not work well when applied directly to low-resource agglutinative languages. Note that the morphology information related to the prosody expression of agglutinative language is ignored. This work proposes a new TTS framework for low-resource agglutinative language with a novel morphology-aware language model pretraining, that has not been studied in previous works.

However, it should also be emphasized that the masking strategy at the word level [52], called whole word masking (WWM), and Byte-Pair Encoding (BPE) subword level [53], [54] in BERT-like model [54] is relatively similar to our approach, but our approach has significant differences. 1) WWM focuses only on word-level information, but it is not possible to learn information about the interior of words, thus limiting the learning of morphological information about the interior of words in agglutinative languages; 2) Although BPE splits words into subwords to expose the information inside the words, the splitting process of BPE is unsupervised, and the subword sequences do not have particularly explicit linguistic meanings, so it still has some limitations in the learning of morphological information in agglutinative languages. Our morphology-aware masking strategy focuses on whole word and subword simultaneously, and words with stems and suffixes are specifically treated to learn the rich morphological information of these words in order to enhance the prosody expressiveness of word representations.

III. The Proposed Method

We propose a novel TTS framework for low-resource agglutinative language with a novel morphology-aware language model pre-training. To face the limited <text, speech> paired data of low-resource agglutinative language, we collect large-scale text data and pre-train the MAM-BERT since large-scale text data without annotation is relatively easy to obtain. Then the pre-trained MAM-BERT language model is used to extract deep linguistic information, for the input text, to enhance the prosody naturalness of the TTS model. Note that all descriptions of our method and experiments are based on Mongolian, a typical agglutinative language.

A. TTS Framework Overview

The overall framework is illustrated in Fig. 2, which has two stages: 1) MAM-BERT Pre-training and 2) TTS Training. In Stage I, we pre-train a morphology-aware MAM-BERT language model that exploits morphology-aware linguistic knowledge of the agglutinative language. In Stage II, the pre-trained MAM-BERT of Stage I serves as a linguistic encoder to extract

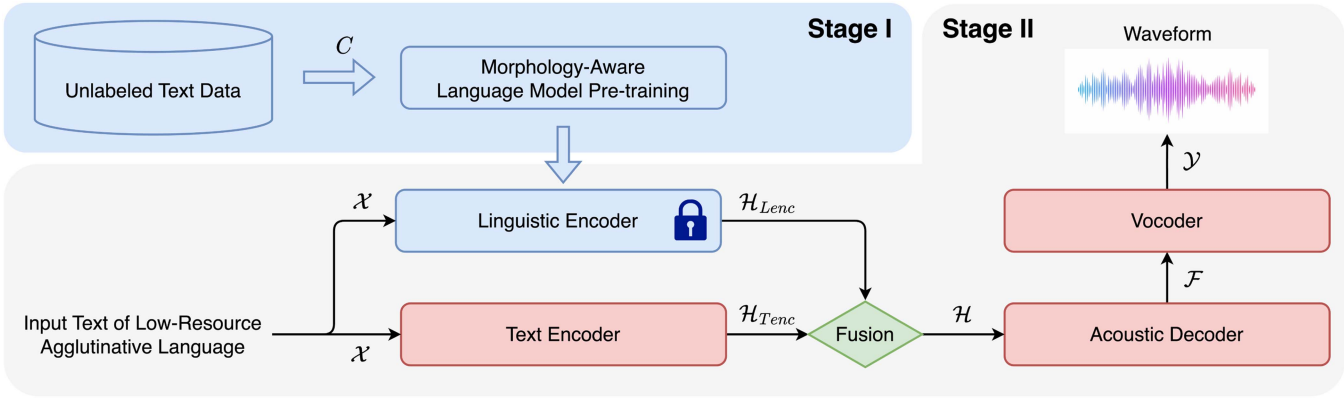


Fig. 2. Overview architecture of our TTS framework. The blue panel shows the workflow of the proposed morphology-aware language model pre-training (Stage I). The gray panel shows the pipeline of the TTS model training (Stage II), which consists of the linguistic encoder, text encoder, fusion operation, acoustic decoder, and vocoder. The blue lock means the parameters of the linguistic encoder are transferred from the pre-trained morphology-aware language model and do not update during TTS training.

prosody-related linguistic features for input text. The prosody-related linguistic feature and the text encoder output are fused to form the final decoder input. The acoustic decoder aims to predict the acoustic feature. The vocoder takes the acoustic feature as input to reconstruct the speech waveform.

Unlike the traditional encoder-decoder-based TTS framework, the proposed TTS framework includes a new linguistic encoder for agglutinative language. Such a scheme can make full use of unlabeled text data to alleviate the low resource problem of agglutinative language TTS. We will introduce the two-stage training strategy and the run-time inference next.

B. Stage I: MAM-BERT Pre-Training

Traditional BERT is a pre-trained language model trained on large amounts of unlabeled data using two objectives: 1) Masked Language Modeling (MLM) uses a random masking strategy to mask 15% of tokens in a sequence randomly, followed by a supervised prediction of the masked tokens; 2) Next Sentence Prediction (NSP) predicts in a binary fashion if two sentences follow each other. In this way, BERT produces powerful linguistic representations, applicable to a wide range of tasks, such as text classification, question answering, and machine translation. Fig. 3(a) shows the traditional BERT-like language model training process with a random masking strategy. Note that we omit the NSP details since some literature has concluded that NSP plays a negligible role [54] for language model training. Given Mongolian Latin text "neN qihvia ni homun-u bey_e-yin eregul qihirag-tv yvsalan_a." (English translation: Most importantly, it is good for human health.),[1] the random masking strategy masks three words randomly, that are "ni", "eregul" and "qihirag-tv". Then the Transformer network seeks to utilize the context information to reconstruct the masked words, thus learning the linguistic knowledge of the input text.

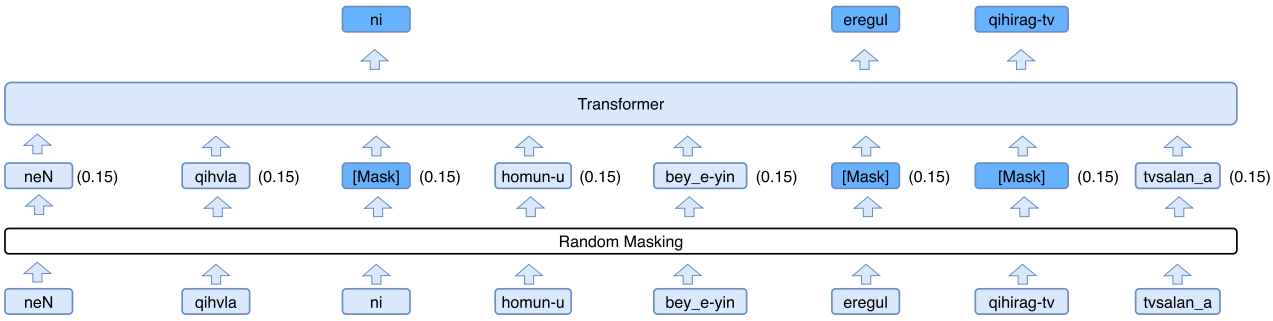
[1]We follow [21] to convert the Mongolian scripts to the Latin sequence for easy model training.

However, such a random masking strategy ignores the morphology knowledge of the agglutinative language. Therefore, the linguistic feature extracted from such a BERT model does not match the agglutinative language. To this end, we propose a novel MAM-BERT, which consists of three modules: 1) Morphology Analysis; 2) Morphology-Aware Masking; and 3) Transformer Network.

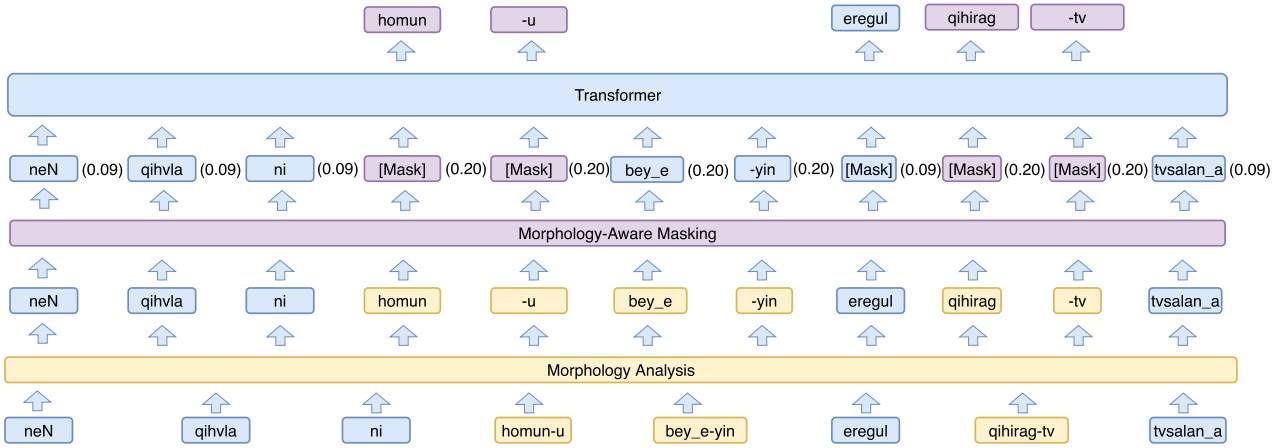
1) Morphology Analysis: Morphology analysis aims to convert the Mongolian word into subword units, according to the morphology characteristics as described in Section II-A, to exploit morphology knowledge into language model training. More importantly, the morphology structure of Mongolian text includes rich speech prosody information [21] which can enhance the prosody naturalness of synthesized speech.

We assume that the collected large-scale text data is C, given Mongolian Latin text $\mathcal{X} = \mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_i, \ldots, \mathcal{X}_T$ (T is the sequence length, $i \in [1, T]$, $\mathcal{X} \in C$), the morphology analysis module seeks to convert the word unit to subword unit according to the NNBS. As shown in Fig. 3(b), we assume that $\mathcal{X}$ is "neN qihvia ni homun-u bey_e-yin eregul qihirag-tv yvsalan_a.". Note that some words do not have suffixes, so the result is still the original word, but some words with suffixes are cut into two parts: the stem and the suffix. For example, "homun-u" was converted to "homun" and "-u", "bey_e-yin" was converted to "bey_e" and "-yin", and "qihirag-tv" was converted to "qihirag" and "-tv". Now the input word sequence is divided to three kinds of tokens: entire word, stem, and suffix. In this way, the modeling unit of the BERT language model is converted from a word sequence to a hybrid token sequence. After that, our method also adopts an MLM scheme to model the linguistics of input text. However, to further model the morphology information, we propose to conduct morphology-aware masking, instead of random masking, which will be discussed next.

2) Morphology-Aware Masking (MAM): It's worth mentioning that we use the WWM strategy as the key to our proposed morphology-aware masking method. Note that the hybrid tokens, which include the entire word, stem, and suffix, are treated



(a) Traditional BERT-like Language Model with Random Masking Strategy



(b) MAM-BERT with Morphology-Aware Masking Strategy

Fig. 3. Comparison about the (a) traditional BERT-like language model with random masking strategy and the proposed (b) MAM-BERT with morphology-aware masking strategy. Morphology analysis aims to convert the Mongolian word into subword unit according to the morphology characteristic. Morphology-aware masking also masks the input tokens with the guidance of morphology characteristics instead of the random mask. Both masking strategies are built on whole-word masking strategy. The floating point number in parentheses indicates the probability that the unit is masked.

as the modeling unit. We try to adjust the masking probability to learn the morphology knowledge.

Specifically, the hybrid token sequence of input text is denoted as $\hat{\mathcal{X}} = \hat{\mathcal{X}}_1, \hat{\mathcal{X}}_2, \ldots, \hat{\mathcal{X}}_i, \ldots, \hat{\mathcal{X}}_T$ (T is the sequence length, $i \in [1, T]$). Unlike traditional BERT, which uses a uniform probability (15%) to mask the tokens randomly, we assign higher probabilities to stem and suffix tokens.

We set the masking probability k of these tokens as:

$$P\left(\hat{\mathcal{X}}_j\right) = k, \tag{1}$$

where k is a hyperparameter in our method and $\hat{\mathcal{X}}_j$ belongs to the stem or suffix tokens.

To ensure we mask 15% of the tokens in total, we lower the masking probability of the entire word token using the following formula:

$$P\left(\hat{\mathcal{X}}_m\right) = \frac{\max(|\hat{\mathcal{X}}| \cdot 0.15 - |\hat{\mathcal{X}}_j| \cdot k, 0)}{|\hat{\mathcal{X}}| - |\hat{\mathcal{X}}_j|}, \forall \hat{\mathcal{X}}_m \notin \hat{\mathcal{X}}_j \tag{2}$$

where $|\cdot|$ represents the size of a set. Fig. 3(b) also shows the masking process. For instance, there are three stems and three suffixes, and five words in the text sequence. For the stems and suffixes, we use a masking probability of $k = 0.20$, thus the entire word probability is lowered from 15% to 9% to keep the sum of probabilities constant.

Although we adjust the masking probability, all types of tokens are still likely to be masked, but with different probabilities. As shown in Fig. 3(b), two stems ("homun" and "qihirag"), two suffixes ("-u" and "-tv"), and one entire word ("eregul") are selected and replaced with the special token [MASK]. Next, we need to predict the masked token.

3) Transformer Network: The rest of the masked tokens' prediction process is the same as the original BERT pre-training. The Transformer network utilizes the context information to predict the original tokens with the help of a self-attention mechanism. For the loss function, we just use the MLM loss function, that is the cross-entropy loss:

$$L_{mlm} = -\sum_{n=1}^{N} \log p\left(\hat{\mathcal{X}}_n\right) \tag{3}$$

where N is the total number of masked tokens and $p(\hat{\mathcal{X}}_n)$ is the predicted probability of the token $\hat{\mathcal{X}}_n$ over the vocabulary size.

After MAM-BERT pre-training, the trained MAM-BERT model can learn deep linguistic information in Mongolian text that is correlated with prosody expressions, so next, we will use the trained MAM-BERT model as an additional linguistic encoder for the TTS model to extract linguistic features for text sequences, to enhance the prosody naturalness of TTS synthesis.

Note that the last layer output $\mathcal{H}_{Lenc}$ of the MAM-BERT will be fed into the TTS model.

C. Stage II: TTS Training

The overall architecture of the TTS model includes a linguistic encoder, text encoder, acoustic decoder, and vocoder as illustrated in the gray panel of Fig. 2. Note that the Fusion operation was used to fuse the outputs from two encoders.

1) Linguistic Encoder, Text Encoder, and Fusion Operation: The linguistic encoder reads the input text $\mathcal{X}$ and converts it to the hybrid token sequence $\hat{\mathcal{X}}$ as mentioned before, then obtains the deep linguistic feature $\mathcal{H}_{Lenc}$ which contains the embeddings for all hybrid tokens $\hat{\mathcal{X}}_1, \hat{\mathcal{X}}_2, \ldots, \hat{\mathcal{X}}_i, \ldots, \hat{\mathcal{X}}_T$. As shown in Fig. 2, the blue lock of the linguistic encoder means the parameters of the linguistic encoder are transferred from the pre-trained morphology-aware language model and do not update during TTS training.

The text encoder also takes the $\mathcal{X}$ as input. However, for TTS training, the input text $\mathcal{X}$ is converted to the character or letter sequence $\mathcal{X}' = \mathcal{X}'_1, \mathcal{X}'_2, \ldots, \mathcal{X}'_{T'}$ since the modeling unit for TTS is character [9]. Note that T' is the character length for $\mathcal{X}'$, and longer than T absolutely. The output of the text encoder is the hidden feature $\mathcal{H}_{Tenc}$ which contains the embeddings for all characters.

The fusion operation seeks to combine $\mathcal{H}_{Lenc}$ and $\mathcal{H}_{Tenc}$ to a meaningful representation $\mathcal{H}$. To address the length mismatch, the token level feature $\mathcal{H}_{Lenc}$ was downsampled to the character level, according to the token-character mapping relationship, and concatenated with the character level feature $\mathcal{H}_{Tenc}$. Specifically, according to the characteristics of Mongolian words as illustrated in Section II-A, the Mongolian word can be cut into stems and suffixes based on NNBS suffixes, which are explicit markers (Latin "-") that exist inside a word, so the boundaries of stems and suffixes can be easily obtained. Section III-B mentioned that there are three kinds of tokens: word, stem and suffix. Assume that Num and Num' indicate the character number of word or subword (stem and suffix) token, for word token, we copy the learned word-level embedding Num times and concatenate it to the embedding of each character. For stems and suffixes, we copy the learned stem or suffix embeddings Num' times, and then concatenate them to the embeddings of each character within the stem or suffix. At last, the length of $\mathcal{H}$ is equal to that of $\mathcal{H}_{Tenc}$.

2) Acoustic Decoder and Vocoder: The acoustic decoder takes the $\mathcal{H}$ to predict the acoustic feature $\mathcal{F}$. The vocoder aims to convert it into the final speech waveform $\mathcal{Y}$.

Note that the specific structure of text encoder, acoustic decoder and vocoder can refer to Tacotron [9], [10], Transformer TTS [11], FastSpeech [12], [13] and other model implementations, so it is not expanded in detail here. The detailed structure information will be introduced in the experimental setup section.

D. Run-Time Inference

The inference stage only involves the TTS Model in the gray panel of Fig. 2. At run-time, the TTS model takes the text as input and generates natural speech as output. With the help of a linguistic encoder that is trained on large-scale text data, we can extract meaningful prosody-related linguistic features for the input text, thus synthesizing high-quality speech in terms of prosody naturalness even though the training data for TTS of agglutinative language is low-resource.

IV. Experiments

We now evaluate the effectiveness of the proposed TTS framework for low-resource agglutinative language with MAM-BERT pre-training.

A. Experimental Corpora

1) Large-Scale Mongolian Text Data: We collect the large-scale Mongolian text data from Mongolian Wikipedia[2] and News[3] websites. After cleaning up, it contains approximately 31 million utterances and 526 million words, with a vocabulary size of 200 million. For language model pre-training, all Mongolian words are Latin-cased with the help of a public translation engine.[4] As mentioned above, our morphology-aware masking focuses on stems or suffixes from a subword lexicon. According to statistics, the subword lexicon includes 520 million stems and 190 million suffixes.

2) Low-Scale Mongolian TTS Data: We conducted TTS experiments on a dataset of low-resource Mongolian TTS challenge with 2.3 hours, also known as NCMMSC2022-MTTSC.[5] The NCMMSC2022-MTTSC dataset was recorded by a professional female announcer whose native language is Mongolian. The entire recording process was recorded in a standard recording studio at Inner Mongolia University using Adobe Audition software. The data set includes 1000 utterances and consists of a training set (about 1.5 hours), a validation set (about 0.5 hours), and a test set (about 0.3 hours). All audios are sampled with 44.10 kHz. For TTS training, we resampled all audios to 16 kHz. All Mongolian words are also Latin-cased.

3) Mongolian Prosodic Prediction Data: We use the Mongolian prosodic prediction data reported in [21] as the prosodic prediction data. The transcript contains 59 k sentences and more than 409 k words. The prosodic phrase breaks of all the sentences were manually labeled by five native annotators who examined the text and listened to the speech. Each word is assigned to a phrase break label: "B" (break after a word) or "NB" (otherwise). The total number of prosodic phrase break labels is approximately 131 k, and the average length of prosodic phrases is 3.5 words. We divide the database into a training and test set in a ratio of 4 to 1, and we extract 25% from the training set as a development set to optimize model parameters.

B. Experimental Setup

1) Implementation Details of Our MAM-BERT: We use the large-scale Mongolian text data as pre-training data. The maximum length of the input sequence was set to 512. We ran Adam

[2]https://mn.wikipedia.org/wiki
[3]such as: https://montsame.mn/en/
[4]http://trans.mglip.com/
[5]http://mglip.com/challenge/NCMMSC2022-MTTSC/dataset.html

with learning rate of 1e-4, $\beta 1 = 0.9$, $\beta 2 = 0.999$, L2 weight decay of 0.01, learning rate warm-up over the first 10% of the total steps, and linear decay of the learning rate. We set a dropout probability of 0.1 for every layer. We denote the number of Transformer block layers as N_{Trans}, the size of hidden vectors as S_{hv}, and the number of self-attention heads as HN_{sa}. Following the practice of BERT, we set them as following settings: $N_{Trans} = 12$, $S_{hv} = 768$, $HN_{sa} = 12$, Number of parameters $\approx$ 100 M.

For morphology-aware masking, k is set to a value from 0.2 to 1.0 to determine the optimal masking probability. All models are trained with 300 k steps to ensure complete convergence. The codes are written in Python 3.6 using the PyTorch library 1.7.0. The GPU type is NVIDIA Tesla P100 with 24 GB GPU memory.

2) TTS Backbone Network: Since our approach focuses more on the additional linguistic encoder part of the TTS model, the selection of the TTS backbone network is very flexible. As mentioned in Section III-C, some of the leading encoder-decoder based end-to-end TTS models, such as Tacotron2 [10], Transformer TTS [11], FastSpeech2 [13] etc., are available.

For Tacotron2, the text encoder consists of a convolutional neural network (CNN) module [55] that has 3 convolutional layers, and a bidirectional LSTM (BLSTM) [56] layer. The attention-based acoustic decoder consists of five components: GMM attention mechanism [57], a 2-layer pre-net, 2 LSTM layers, a linear projection layer and a 5-convolution-layer post-net. We follow the parameter configuration of [10]. Please check [10] for more details.

For Transformer TTS, the text encoder consists of a multi-head self-attention and feed-forward network (FFN). The acoustic decoder includes masked multi-head self-attention, multi-head self-attention and FFN modules. We follow the parameter configuration of [11]. Please check [11] for more details.

For FastSpeech2, the feed-forward Transformer block, which is a stack of self-attention layer and 1D-convolution as in FastSpeech [20], forms the basic structure for the text encoder and acoustic decoder. Note that the additional pitch and energy predictors in the variance adaptor can provide more prosody variance information. The length regulator uses the character duration obtained by the attention map of the Tacotron2 teacher model. We follow the parameter configuration of [13]. Please check [13] for more details.

For all TTS models, we follow [25] and employ a pre-trained universal HiFi-GAN [18] vocoder for waveform generation. All models are trained with 200 k steps to ensure complete convergence. In the following comparative study, we will attach our pre-trained linguistic encoder to their text encoders and compare their performance.

V. Results and Discussion

In this section, we first conduct the prosodic prediction evaluation to determine the optimal sampling probability (k) and compare our MAM-BERT with the traditional BERT model. After that, we conduct the TTS evaluation to validate our proposed MAM-BERT with objective and subject tests.[6] A visualization study is also conducted to observe the performance intuitively. Furthermore, we validate the MAM-BERT in terms of low-resource effectiveness. In addition, we compare the subword-based MAM and the BPE subword mechanism.

A. Prosodic Prediction Evaluation

We first conduct the prosodic prediction evaluation with the Mongolian Prosodic Prediction Data to answer two questions: a) What is the optimal value of sampling probability (k); b) Whether MAM-BERT has mastered the prosody-related linguistic knowledge of the Mongolian language. Specifically, we add a linear layer after the MAM-BERT output layer to predict the binary break labels ("B" or "NB") and report the performance in terms of Precision (P), Recall (R), and F-score (F) which is defined as the harmonic mean of the P and R. F values range from 0 to 1, with a higher value indicating better performance.

1) Comparison of Various Sampling Probability (k): The results of the test set are reported in Table I. We can find that, for the MAM-BERT, as k increases from 0.2 to 0.6, the F metric gradually increases from 94.96 to 95.21. However, as k increases from 0.6 to 1.0, the F gradually decreases from 95.21 to 94.91. F achieves the highest performance when k is 0.6.

In a nutshell, setting k to 0.6 is more beneficial for our model to learn the rich linguistic information of the agglutinative language. In the following experiments, we will build our models with $k = 0.6$.

2) Comparison of BERT and our MAM-BERT: To answer the second question about MAM-BERT, we compare the MAM-BERT model with the BERT baseline. Note that the BERT model doesn't adopt the morphology-aware masking strategy and achieves 93.84, 94.32 and 94.26 in terms of P, R and F respectively. We find that our MAM-BERT models with various k outperform BERT model consistently. The F score of the optimal MAM-BERT model with $k = 0.6$ is 0.95 higher than that of BERT model, which is remarkable. The above observations show that our method can grasp the deep linguistic information related to prosody very well, which is significantly better than the traditional BERT training method.

In the next section, we will incorporate the MAM-BERT model into the advanced TTS models to compare speech naturalness in terms of objective and subjective evaluations.

B. TTS Evaluation

To verify whether the linguistic features learned by MAM-BERT contribute to the naturalness of TTS, we concatenate together MAM-BERT as a pre-trained linguistic encoder and the TTS model as in Fig. 2. Specifically, we built nine TTS systems for comparative study.

- Tacotron2 (Mel + HiFi-GAN): The official Tacotron2 model was used to build the TTS framework. The pre-trained linguistic encoder was omitted.

[6]Speech Demo: https://ttslr.github.io/MAM-BERT/

TABLE I
COMPARISON OF THE PROSODIC PREDICTION ACCURACY FOR DIFFERENT SYSTEMS, IN TERMS OF PRECISION (P), RECALL (R), AND F-SCORE (F), WITH VARIOUS MASKING PROBABILITY (k)

System	Metrics	Masking Probability									
		k=0.2	k=0.3	k=0.4	k=0.5	k=0.6	k=0.7	k=0.8	k=0.9	k=1.0	Average
MAM-BERT	P	94.91	94.95	94.98	95.10	95.18	95.14	95.01	94.93	94.90	95.01
	R	94.98	95.03	95.09	95.19	95.35	95.19	95.13	95.01	94.97	95.10
	F	94.96	95.01	95.07	95.11	**95.21**	95.17	95.09	94.98	94.91	**95.06**
System	**Metrics**	**No MAM**									
BERT	P	93.84									
	R	94.32									
	F	94.26									

- Transformer TTS (Mel + HiFi-GAN): The official Transformer TTS model was used to build the TTS framework. The pre-trained linguistic encoder was omitted.
- FastSpeech2 (Mel + HiFi-GAN): The official FastSpeech2 model was used to build the TTS framework. The pre-trained linguistic encoder was omitted.
- BERT + Tacotron2 (Mel + HiFi-GAN): The pre-trained BERT was treated as the linguistic encoder to extract the linguistic feature for input text. The Tacotron2 model was used to build the encoder-decoder-based TTS framework. For speech generation, the HiFi-GAN vocoder was used to convert the predicted mel-spectrum features to the speech waveform.
- MAM-BERT + Tacotron2 (Mel + HiFi-GAN): The pre-trained MAM-BERT was treated as the linguistic encoder. The Tacotron2 acoustic model and HiFi-GAN vocoder were used to convert the text to the speech waveform.
- BERT + Transformer TTS (Mel + HiFi-GAN): The pre-trained BERT was treated as the linguistic encoder. The Transformer TTS acoustic model and HiFi-GAN vocoder were used to convert the text to the speech waveform.
- MAM-BERT + Transformer TTS (Mel + HiFi-GAN): The pre-trained MAM-BERT was treated as the linguistic encoder. The Transformer TTS acoustic model and HiFi-GAN vocoder were used to convert the text to the speech waveform.
- BERT + FastSpeech2 (Mel + HiFi-GAN): The pre-trained BERT was treated as the linguistic encoder. The FastSpeech2 acoustic model and HiFi-GAN vocoder were used to convert the text to the speech waveform.
- MAM-BERT + FastSpeech2 (Mel + HiFi-GAN): The pre-trained MAM-BERT was treated as the linguistic encoder. The FastSpeech2 acoustic model and HiFi-GAN vocoder were used to convert the text to the speech waveform.

Note that the duration target for FastSpeech2 was extracted by the pre-trained Tacotron2 model. In addition, we also add *Ground Truth (GT)* speech and the resynthesized speech, also termed as *GT (Mel + HiFi-GAN)*, as two reference baselines.

1) Objective Results: We use Mel Cepstral Distortion (MCD) [58] and Root Mean Squared Error (RMSE) [13] as the objective evaluation metrics to evaluate the performance in terms of mel-spectrum and pitch.

As the duration of the synthesized speech is usually different from that of the reference speech, we apply the dynamic time warping (DTW) algorithm [59] to obtain a frame-level alignment between the two to facilitate MCD calculation. We calculate an MCD between a reference speech and a synthesized speech of T frames as follows,

$$\mathrm{MCD} = \frac{10\sqrt{2}}{\ln 10}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{\mathcal{N}}\sqrt{\sum\nolimits_{k=1}^{\mathcal{N}}(y_{t,k}-\hat{y}_{t,k})^2}\right) \tag{4}$$

where $\mathcal{N}$ represents the dimension of the mel-spectrum, $y_{t,k}$ denotes the kth mel-spectrum component of tth frame for the reference speech, and $\hat{y}_{t,k}$ for that of the synthesized speech.

We use RMSE as the evaluation metrics for F0 modeling, which is calculated as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{T}\sum\nolimits_{t=1}^{T}\left(\mathrm{F0_t}-\widehat{\mathrm{F0_t}}\right)^2} \tag{5}$$

where $\mathrm{F0_t}$ and $\widehat{\mathrm{F0_t}}$ denote the reference and synthesized F0 at tth frame. Following the DTW alignment relationship, the F0 sequence extracted from the synthesized speech is also aligned toward ground truth. We note that the lower RMSE value suggests that the two F0 contours are more similar.

We report the MCD and RMSE values in the second and third columns of Table II. We just calculate the values for nine comparative systems except for two reference baselines, which are GT and GT (Mel + HiFi-GAN). We first compare the three TTS models without BERT models with the following TTS model armed with BERT/MAM-BERT. For example, the Tacotron2 model achieved 7.69 and 1.23 for MCD and RMSE, which are higher than all other models. Transformer TTS and FastSpeech2 also obtained relatively poor performance. It shows that pre-trained linguistic encoders obtained from BERT and MAM-BERT both boost the TTS performance. Compared with BERT and the MAM-BERT, our MAM-BERT performs more accurate mel-spectrum and pitch features that are close to the natural speech. For example, the MCD and RMSE values of system *MAM-BERT + Tacotron2 (Mel + HiFi-GAN) (ours)* are 7.35 and 1.01, which are 0.07 and 0.07 lower than that of *BERT + Tacotron2 (Mel + HiFi-GAN)*. Similarly, the MCD and RMSE values of system *MAM-BERT + Transformer TTS (Mel + HiFi-GAN) (ours)* are 7.30 and 0.95, which are 0.11 and 0.04 lower than that of *BERT + Transformer TTS (Mel + HiFi-GAN)*. *MAM-BERT + FastSpeech2 (Mel + HiFi-GAN) (ours)* achieved the best performance, that are 7.02 and 0.82, in terms of MCD and RMSE. In general, our MAM-BERT improves the speech naturalness of any advanced TTS model and is significantly better than the BERT model.

TABLE II
COMPARISON OF THE SPEECH NATURALNESS FOR DIFFERENT SYSTEMS IN TERMS OF MCD AND RMSE IN OBJECTIVE EXPERIMENTS, MOS AND BWS IN SUBJECTIVE EXPERIMENTS

System	Speech Naturalness				
	MCD (dB)	***RMSE*** (Hz)	***MOS***	**BWS Evaluation**	
				Best (%)	***Worst*** (%)
GT	NA	NA	4.40 ± 0.02	NA	NA
GT (Mel + HiFi-GAN)	NA	NA	4.38 ± 0.01	NA	NA
Tacotron2 (Mel + HiFi-GAN)	7.69	1.23	3.85 ± 0.02	0	25
Transformer TTS (Mel + HiFi-GAN)	7.61	1.19	3.87 ± 0.03	0	20
FastSpeech2 (Mel + HiFi-GAN)	7.63	1.17	3.91 ± 0.01	2	19
BERT + Tacotron2 (Mel + HiFi-GAN)	7.42	1.08	3.97 ± 0.02	8	11
MAM-BERT + Tacotron2 (Mel + HiFi-GAN) **(ours)**	7.35	1.01	4.18 ± 0.02	14	6
BERT + Transformer TTS (Mel + HiFi-GAN)	7.41	0.99	4.09 ± 0.02	6	7
MAM-BERT + Transformer TTS (Mel + HiFi-GAN) **(ours)**	7.30	0.95	4.22 ± 0.02	11	5
BERT + FastSpeech2 (Mel + HiFi-GAN)	7.11	0.88	4.13 ± 0.01	9	4
MAM-BERT + FastSpeech2 (Mel + HiFi-GAN) **(ours)**	**7.02**	**0.82**	**4.29 ± 0.01**	**50**	**3**

2) Subjective Results: We conduct the listening experiment by reporting the mean opinion score (MOS) scores [60] across all systems, and summarize in the fourth column of Table II. Each speech sample is rated on a scale of 1 to 5 with an interval of 0.5. "5" for excellent, "4" for good, "3" for fair, "2" for poor, and "1" for bad. We recruit 20 Mongolian listeners. Each listener listens to 200 speech samples. The listeners are instructed to pay attention to the naturalness of speech.

We report the MOS results in the fourth column of Table II. The results show that two GT baselines achieved the optimal systems. Compared with Tacotron2, Transformer TTS, FastSpeech2, and those armed with the BERT/MAM-BERT, we found that BERT and MAM-BERT both improve the overall speech naturalness. For example, the MOS scores of Tacotron2, Transformer TTS, and FastSpeech2 are 3.85, 3.87, and 3.91 respectively. However, *BERT + Tacotron2 (Mel + HiFi-GAN)* obtained 3.97 and *MAM-BERT + Tacotron2 (Mel + HiFi-GAN)* achieved 4.18. Transformer TTS and FastSpeech2 systems armed with the BERT/MAM-BERT also further perform higher MOS scores. Note that the *MAM-BERT + FastSpeech2 (Mel + HiFi-GAN)* system obtained the optimal MOS score of 4.29. Then we compare the BERT and MAM-BERT models. The performance gap between them consistently validates our conjecture. In specific, *MAM-BERT + Tacotron2 (Mel + HiFi-GAN)* is higher than *BERT + Tacotron2 (Mel + HiFi-GAN)* with 0.21. *MAM-BERT + Transformer TTS (Mel + HiFi-GAN)* is higher than *BERT + Transformer TTS (Mel + HiFi-GAN)* with 0.13. *MAM-BERT + FastSpeech2 (Mel + HiFi-GAN)* is also higher than *BERT + Transformer TTS (Mel + HiFi-GAN)* with 0.16. Overall, our MAM-BERT can learn rich prosody-related linguistic information for Mongolian, thus enhancing the naturalness of the synthesized speech and synthesizing speech that better matches human speech perception.

We also conduct the second listening experiment through BWS [61], [62], which is an effective method to provide a ranking of a long list of listening samples [63]. In so doing, we randomly select 80 utterances from the test set. We also recruited 20 Mongolian listeners. For each utterance, nine speech samples (except for *GT* and *GT (Mel + HiFi-GAN)*) produced by these nine systems form a group. A listener picks the best and worst samples in terms of naturalness for each group. In other words, each listener listens to all 80 groups, 720 utterances in total. We report the results in the last two columns of Table II. It can be seen that out of all nine systems (excluding the two reference baselines), our *MAM-BERT + FastSpeech2 (Mel + HiFi-GAN)* system was very well received, with 50% of the testers feeling that it was the best-performing voice of the multiple systems, and only 3% felt that it performed worse than the other systems. This proves once again the effectiveness of the combination of the *MAM-BERT* method and the FastSpeech2 model.

We further conduct the AB preference test to assess the naturalness of the systems. Each audio is listened to by 20 subjects, each of which listens to 100 synthesized speech samples. Fig. 4 reports the naturalness evaluation results. Fig. 4(a) shows the results of comparing *BERT + Tacotron2 (Mel + HiFi-GAN)* and *MAM-BERT + Tacotron2 (Mel + HiFi-GAN)*, proving that *MAM-BERT + Tacotron2 (Mel + HiFi-GAN)* has a higher popularity of 46%. Fig. 4(b) shows the results of comparing *BERT + Transformer TTS (Mel + HiFi-GAN)* and *MAM-BERT + Transformer TTS (Mel + HiFi-GAN)*, proving that *MAM-BERT + Transformer TTS (Mel + HiFi-GAN)* has higher popularity of 50%. In similarity, Fig. 4(c) shows the results of comparing *BERT + FastSpeech2 (Mel + HiFi-GAN)* and *MAM-BERT + FastSpeech2 (Mel + HiFi-GAN)*, proving that *MAM-BERT + FastSpeech2 (Mel + HiFi-GAN)* has higher popularity of 55%. However, Fig. 4(d) shows the results of comparing GT and *MAM-BERT + FastSpeech2 (Mel + HiFi-GAN)*. This result is 22% for GT, 20% for *MAM-BERT + FastSpeech2 (Mel + HiFi-GAN)*, and 58% for *Neutral*, indicating that the speech synthesized by method *MAM-BERT + FastSpeech2 (Mel + HiFi-GAN)* does not differ much from GT and is difficult to be distinguished clearly by the listener.

C. Visualization Study

We now provide a case study to illustrate the prosodic phrase breaking behaviors. Fig. 5 shows a waveform plot of all comparative systems among *GT*, *FastSpeech2 (Mel + HiFi-GAN)* and

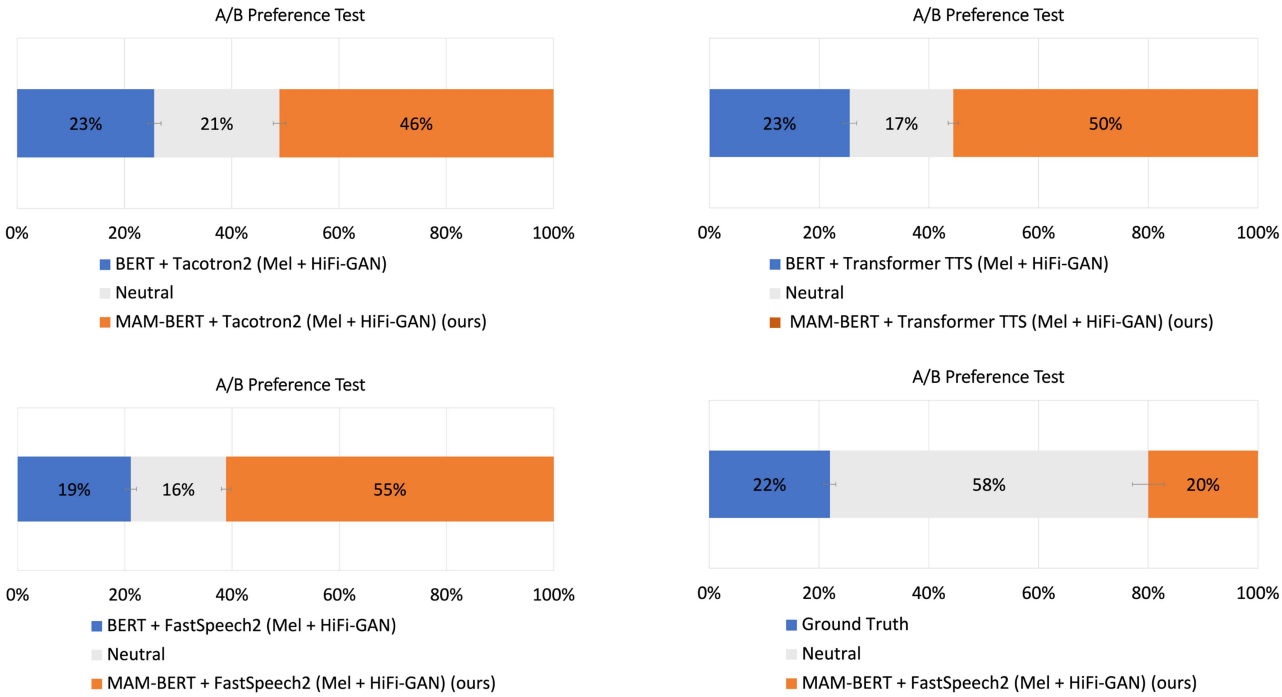


Fig. 4. Comparison of the speech naturalness for different systems in terms of the AB preference test.

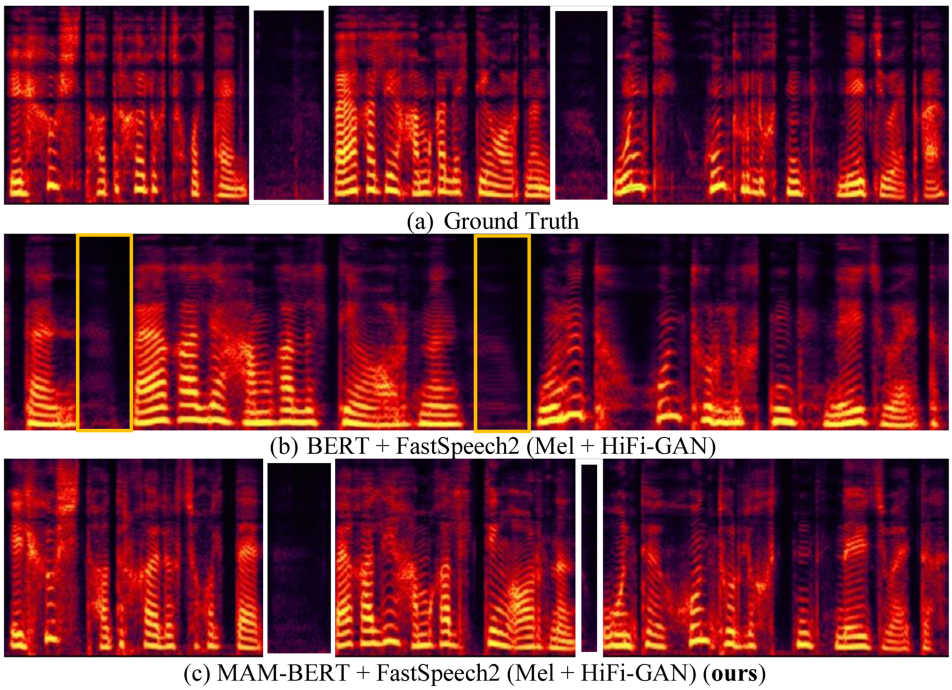


Fig. 5. Comparison of the mel-spectrum visualization for different systems. The white boxes mean the correct phrase breaks, while the yellow boxes indicate the wrong phrase breaks.

MAM-BERT + FastSpeech2 (Mel + HiFi-GAN). It is clear that *MAM-BERT + FastSpeech2 (Mel + HiFi-GAN)* produces better prosodic phrase breaks than others. We have added orange boxes to indicate the natural prosodic phrase breaks. For example, in the Mongolian utterance, we have "*erhebxi tqdqrhailahv cihvlatai ni | baigali-yin hamagalalta-yin qrqn ni | onide homun torolhiten-d'u' negeju baidag.*" with the bars representing the natural breaks.

In sum, all the above experiments confirm that our MAM-BERT based linguistic encoder effectively addresses the poor naturalness issue due to the low-resource training data. The performance gain of MAM-BERT over BERT and the traditional

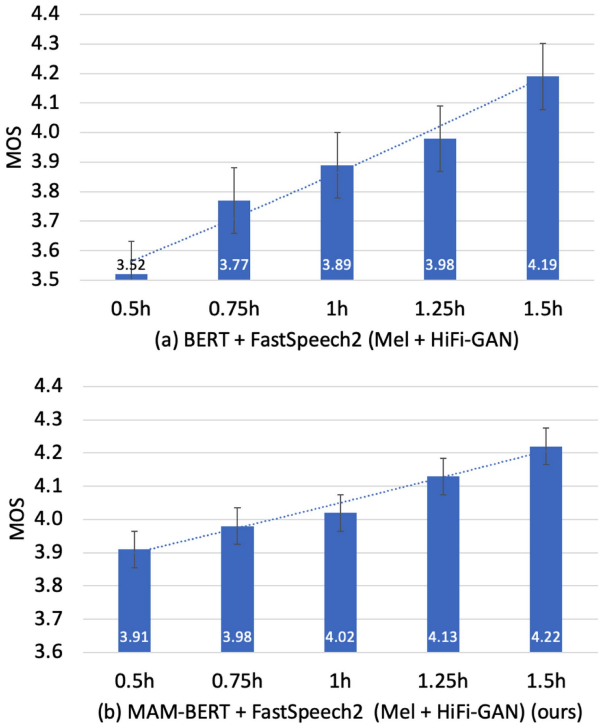


Fig. 6. Comparison of the low resource effectiveness for MAM-BERT and BERT systems. The number in the X-axis means the duration of available data of the NCMMSC2022-MTTSC training set.

TTS framework is attributed to the rich linguistic knowledge of the agglutinative language.

D. Low Resource Effectiveness Evaluation

To further validate the effectiveness of our approach in low-resource scenarios, we choose two systems, *MAM-BERT + FastSpeech2 (Mel + HiFi-GAN)* and *BERT + FastSpeech2 (Mel + HiFi-GAN)*, for comparison. We divide the training set of NCMMSC2022-MTTSC dataset into five equal parts and then construct the training data with different durations, including 0.5. 0.75, 1, 1.25, 1.5 hours, before re-training the model. At last, we carry out the MOS experiments with the same configurations as Section V-B2. The results are reported in Fig. 6.

We can find that our system is more friendly in the face of low-resource data. At 0.5 hours of data, the MOS score of baseline achieves 3.52, while ours is 3.91. After that, as the training data increases, although both MOS scores increase, it is obvious that our performance is better. In the end, we outperformed the baseline system on all 5 different durations of data. We also notice that the results of our system on 0.5 h data have surpassed the results of the baseline system on 1 h, and even approach the results of 1.25 hours. To summarize, our method is more effective in low-resource scenarios.

E. Comparison With BPE Baseline

To further validate our MAM-BERT, we add a new baseline called *BERT (BPE)* and reported the results in Section V. Note that the *BERT (BPE)* system takes BPE subword as input. The parameter setup of BPE is followed by [20]. For the fairness of the comparison, we determine whether there are stems and suffixes in the subwords obtained by BPE splitting, and if there are we increase their probability of being masked according to (2). After that, we compare the naturalness of the synthesized speech of the MAM-BERT and *BERT (BPE)* systems. We choose FastSpeech2 and HiFi-GAN as the backbone to generate the speech and follow the setup of Section V-B2. The Comparison MOS (CMOS) [64] results are reported in Table III.

TABLE III
CMOS RESULTS OF MAM-BERT AND *BERT (BPE)* BASELINE

Systems	CMOS
MAM-BERT + FastSpeech2 (Mel + HiFi-GAN) (ours)	0
BERT(BPE) + FastSpeech2 (Mel + HiFi-GAN)	-0.127

We can find that the synthesized speech from our MAM-BERT performs a higher degree of naturalness. We believe that adopting BPE subword sequences as a modeling unit fails to capture the rich morphological information within the words of low-resource agglutinative language. The naturalness of synthesized speech is inferior to that of our method because this morphological information is related to the expression of phonological rhythms [60].

VI. CONCLUSION

To address the low-resource issue of the $< text, speech >$ paired TTS data for agglutinative language. We have studied a novel TTS framework with a morphology-aware language model pre-training. We implement an improved BERT model, with the morphology-aware masking strategy, as the linguistic encoder to extract prosody-related linguistic features to enhance the prosody naturalness of synthesis speech. A series of experiments were conducted to validate the method. In further work, we will try to continue to improve the language model training method.

REFERENCES

[1] P. Taylor, *Text-to-Speech Synthesis*. New York, NY, USA: Cambridge Univ. Press, 2009.
[2] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.
[3] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proc. IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.
[4] H. Zen, A. W. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 7962–7966.
[5] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *Proc. ISCA Speech Synth. Workshop*, 2016, pp. 202–207.
[6] R. Liu, F. Bao, G. Gao, and Y. Wang, "Mongolian text-to-speech system based on deep neural network," in *Proc. Nat. Conf. Man- Mach. Speech Commun.*, 2017, pp. 99–108.
[7] W. Ping et al., "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," in *Proc. Int. Conf. Learn. Representations*, 2018.
[8] W. Wang, S. Xu, and B. Xu, "First step towards end-to-end parametric TTS synthesis: Generating spectral parameters with neural attention," in *Proc. Interspeech*, 2016, pp. 2243–2247.
[9] Y. Wang et al., "Tacotron: A fully end-to-end text-to-speech synthesis model," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[10] J. Shen et al., "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 4779–4783.

[11] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6706–6713.

[12] Y. Ren et al., "Fastspeech: Fast, robust and controllable text to speech," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 3171–3180.

[13] Y. Ren et al., "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. 9th Int. Conf. Learn. Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=piLPYqxtWuA

[14] A. Vaswani et al., "Attention is all you need," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[15] A. V. d. Oord et al., "WaveNet: A generative model for raw audio," in *Proc. 9th ISCA Speech Synth. Workshop*, 2016, pp. 125–125.

[16] N. Kalchbrenner et al., "Efficient neural audio synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2410–2419.

[17] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in *Proc. Int. Conf. Learn. Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=ByMVTsR5KQ

[18] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2020, pp. 17022–17033.

[19] K. Ito and L. Johnson, "The LJ speech dataset," 2017. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. naacL-HLT*, 2019, pp. 4171–4186.

[21] R. Liu, B. Sisman, F. Bao, J. Yang, G. Gao, and H. Li, "Exploiting morphological and phonological features to improve prosodic phrasing for Mongolian speech synthesis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 274–285, 2021.

[22] R. Liu, F. Bao, and G. Gao, "Building mongolian TTS front-end with encoder-decoder model by using bridge method and multi-view features," in *Proc. Neural Inf. Process.: 26th Int. Conf.*, 2019, pp. 642–651.

[23] J. Janhunen, "Mongolic languages," in *The Encyclopedia of Language & Linguistics*. Amsterdam, The Netherlands: Elsevier, 2006, pp. 231–234.

[24] Y. Hu, P. Yin, R. Liu, F. Bao, and G. Gao, "MnTTS: An open-source Mongolian text-to-speech synthesis dataset and accompanied baseline," in *Proc. IEEE Int. Conf. Asian Lang. Process.*, 2022, pp. 184–189.

[25] K. Liang, B. Liu, Y. Hu, R. Liu, F. Bao, and G. Gao, "MnTTS2: An open-source multi-speaker Mongolian text-to-speech synthesis dataset," in *Proc. Nat. Conf. Man- Mach. Speech Commun.*, 2023, pp. 318–329.

[26] R. Liu, F. Bao, G. Gao, H. Zhang, and Y. Wang, "A LSTM approach with sub-word embeddings for Mongolian phrase break prediction," in *Proc. 27th Int. Conf. Comput. Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018, pp. 2448–2455.

[27] R. Liu, F. Bao, G. Gao, H. Zhang, and Y. Wang, "Improving mongolian phrase break prediction by using syllable and morphological embeddings with bilstm model," in *Proc. Interspeech*, 2018, pp. 57–61, doi: 10.21437/Interspeech.2018-1706.

[28] Z. Byambadorj, R. Nishimura, A. Ayush, K. Ohta, and N. Kitaoka, "Multi-speaker TTS system for low-resource language using cross-lingual transfer learning and data augmentation," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2021, pp. 849–853.

[29] M. Kim, M. Jeong, B. J. Choi, S. Ahn, J. Y. Lee, and N. S. Kim, "Transfer learning framework for low-resource text-to-speech using a large-scale unlabeled speech corpus," in *Proc. Interspeech*, 2022, pp. 788–792.

[30] A. Pine, D. Wells, N. Brinklow, P. Littell, and K. Richmond, "Requirements and motivations of low-resource speech synthesis for language revitalization," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (Volume 1: Long Papers)*, 2022, pp. 7346–7359.

[31] J. Xu et al., "LRSpeech: Extremely low-resource speech synthesis and recognition," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2020, pp. 2802–2812.

[32] M. Lajszczak et al., "Distribution augmentation for low-resource expressive text-to-speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 8307–8311.

[33] N. Tits, K. E. Haddad, and T. Dutoit, "Exploring transfer learning for low resource emotional TTS," in *Proc. SAI Intell. Syst. Conf.*, 2019, pp. 52–60.

[34] G. Huybrechts, T. Merritt, G. Comini, B. Perz, R. Shah, and J. Lorenzo-Trueba, "Low-resource expressive text-to-speech using data augmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6593–6597.

[35] R. Shah et al., "Non-autoregressive TTS with explicit duration modelling for low-resource highly expressive speech," in *Proc. ISCA Speech Synth. Workshop*, 2021, pp. 96–101.

[36] S. Karlapati, A. Moinet, A. Joly, V. Klimkov, D. Sáez-Trigueros, and T. Drugman, "Copycat: Many-to-many fine-grained prosody transfer for neural text-to-speech," in *Proc. Interspeech*, 2020, pp. 4387–4391.

[37] R. Liu, Q. Liu, H. Zhu, and H. Cao, "Multistage deep transfer learning for EmIoT-enabled human–computer interaction," *IEEE Internet Things J.*, vol. 9, no. 16, pp. 15128–15137, Aug. 2022.

[38] H. Zhang and Y. Lin, "Unsupervised learning for sequence-to-sequence text-to-speech for low-resource languages," in *Proc. Interspeech*, 2020, pp. 3161–3165.

[39] P. Do, M. Coler, J. Dijkstra, and E. Klabbers, "A systematic review and analysis of multilingual data strategies in text-to-speech for low-resource languages," in *Proc. Interspeech*, 2021, pp. 16–20.

[40] Y.-J. Chen, T. Tu, C.-c. Yeh, and H.-Y. Lee, "End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning," in *Proc. Interspeech*, 2019, pp. 2075–2079.

[41] M. d. Korte, J. Kim, and E. Klabbers, "Efficient neural speech synthesis for low-resource languages through multilingual modeling," in *Proc. Interspeech*, 2020, pp. 2967–2971.

[42] K. C. Sim and H. Li, "Context-sensitive probabilistic phone mapping model for cross-lingual speech recognition," in *Proc. Ninth Annu. Conf. Int. Speech Commun. Assoc.*, 2008, pp. 2715–2718.

[43] F. Lux and T. Vu, "Language-agnostic meta-learning for low-resource text-to-speech with articulatory features," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (Volume 1: Long Papers)*, 2022, pp. 6858–6868.

[44] B. Yang, J. Zhong, and S. Liu, "Pre-trained text representations for improving front-end text processing in mandarin text-to-speech synthesis," in *Proc. Interspeech*, 2019, pp. 4480–4484.

[45] Y.-J. Zhang and Z.-H. Ling, "Learning deep and wide contextual representations using BERT for statistical parametric speech synthesis," in *Proc. 5th Int. Conf. Digit. Signal Process.*, 2021, pp. 146–150.

[46] Y. Jia, H. Zen, J. Shen, Y. Zhang, and Y. Wu, "PnG BERT: Augmented BERT on phonemes and graphemes for neural TTS," in *Proc. Interspeech*, 2021, pp. 151–155.

[47] G. Zhang et al., "Mixed-phoneme BERT: Improving BERT with mixed phoneme and sup-phoneme representations for text to speech," in *Proc. Interspeech*, 2022, pp. 456–460.

[48] Y. A. Li, C. Han, X. Jiang, and N. Mesgarani, "Phoneme-level bert for enhanced prosody of text-to-speech with grapheme predictions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[49] T. Hayashi, S. Watanabe, T. Toda, K. Takeda, S. Toshniwal, and K. Livescu, "Pre-trained text embeddings for enhanced text-to-speech synthesis," in *Proc. Interspeech*, 2019, pp. 4430–4434.

[50] Y. Xiao, L. He, H. Ming, and F. K. Soong, "Improving prosody with linguistic and bert derived features in multi-speaker based Mandarin Chinese neural TTS," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6704–6708.

[51] T. Kenter, M. Sharma, and R. Clark, "Improving the prosody of RNN-based English text-to-speech synthesis by incorporating a BERT model," in *Proc. Interspeech*, 2020, pp. 4412–4416.

[52] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for Chinese BERT," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3504–3514, 2021.

[53] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics (Volume 1: Long Papers)*, 2016, pp. 1715–1725.

[54] Y. Liu et al., "RoBERTa: A robustly optimized bert pretraining approach," 2019, *arXiv:1907.11692*.

[55] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022.

[56] R. Liu, F. Bao, G. Gao, H. Zhang, and Y. Wang, "Phonologically aware BiLSTM model for Mongolian phrase break prediction with attention mechanism," in *Proc. Pacific Rim Int. Conf. Artif. Intell.*, 2018, pp. 217–231.

[57] E. Battenberg et al., "Location-relative attention mechanisms for robust long-form speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6194–6198.

[58] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. IEEE Pacific Rim Conf. Commun. Comput. Signal Process.*, 1993, pp. 125–128.

[59] M. Müller, "Dynamic time warping," in *Information Retrieval for Music and Motion*. Berlin, Germany: Springer, 2007, pp. 69–84.

[60] R. Liu, B. Sisman, G. Gao, and H. Li, "Expressive TTS training with frame and style reconstruction loss," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 1806–1818, 2021, doi: 10.1109/TASLP.2021.3076369.
[61] J. A. Lee, G. Soutar, and J. Louviere, "The best–worst scaling approach: An alternative to Schwartz's values survey," *J. Pers. Assessment*, vol. 90, no. 4, pp. 335–347, 2008.
[62] J. J. Louviere, T. N. Flynn, and A. A. J. Marley, *Best-Worst Scaling: Theory, Methods and Applications*. New York, NY, USA: Cambridge Univ. Press, 2015 .
[63] T. N. Flynn and A. A. Marley, "Best-worst scaling: Theory and methods," in *Handbook of Choice Modelling*. 2014.
[64] P. C. Loizou, "Speech quality assessment," in *Multimedia Analysis, Processing and Communications*. Berlin, Germany: Springer, 2011, pp. 623–654.

Rui Liu (Member, IEEE) received the bachelor's degree from the Taiyuan University of Technology, ShanXi, China, in 2014, and the Ph.D degree from Inner Mongolia University, Hohhot, China, in 2020. He is currently a Professor with Inner Mongolia University. From 2019 to 2020, he was an exchange Ph.D. Candidate with the Department of Electrical & Computer Engineering, National University of Singapore (NUS), Singapore, supported by China Scholarship Council (CSC). From 2020 to 2022, he was a Research Fellow with the Department of Electrical and Computer Engineering, NUS. His publications include top-tier NLP/ML/AI conferences and journals, such as IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, *Neural Networks*, AAAI, ICASSP, and INTERSPEECH. His research interests include audio, speech, and natural language processing.

Yifan Hu is currently a Ph.D. Student with Inner Mongolia University, Hohhot, China. His research focuses on conversational speech synthesis. He has published papers in top-tier international conferences, such as AAAI.

Haolin Zuo is currently a Ph.D. Student with Inner Mongolia University, Hohhot, China. His research focuses on multimodal emotion recognition. His work has been published in top-tier speech/multimedia international conferences, such as IEEE-ICASSP.

Zhaojie Luo (Member, IEEE) received the M.Eng. and Dr.Eng. degrees from Kobe University, Kobe, Japan, in 2017 and 2020, respectively. He is currently an Assistant Professor with Osaka University, Suita, Japan. From 2019 to 2020, he was a Researcher with the Department of Electrical & Computer Engineering, National University of Singapore, Singapore. His research interests include voice conversion, speech synthesis, facial expression recognition, multimodal emotion recognition, and statistical signal processing. He has published more than 20 papers in top-tier speech/multimedia journals and international conferences, such as IEEETRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING., IEEE TRANS. MULTIMEDIA, EURASIP JASMP, INTERSPEECH, SSW, ICME, and ICPR. He is a member of ISCA and ASJ, and is the Reviewer for many major referred journal and conference papers.

Longbiao Wang (Member, IEEE) received the Dr. Eng. degree from the Toyohashi University of Technology, Toyohashi, Japan, in 2008. He is currently a Professor, Director of Tianjin Key Laboratory of Cognitive Computing and Application, and the Vice Dean of School of Artificial Intelligence, Tianjin University, Tianjin, China. From 2008 to 2012, he was an Assistant Professor with the faculty of Engineering, Shizuoka University, Shizuoka, Japan. From 2012 to 2016, he was an Associate Professor with the Nagaoka University of Technology, Nagaoka, Japan. His research interests include robust speech recognition, speaker recognition, acoustic signal processing, and natural language processing.

Guanglai Gao received the B.S. degree from Inner Mongolia University, Hohhot, China, in 1985 and the M.S. degree from the National University of Defense Technology, Changsha, China, in 1988. He is currently a Professor with the Department of Computer Science, Inner Mongolia University. He was a Visiting Researcher with the University of Montreal, Montreal, OC, Canada. His research interests include artificial intelligence and pattern recognition.