Exploiting Morphological and Phonological Features to Improve Prosodic Phrasing for Mongolian Speech Synthesis

Rui Liu, *Member, IEEE*, Berrak Sisman, *Member, IEEE*, Feilong Bao, Jichen Yang, *Senior Member, IEEE*, Guanglai Gao, and Haizhou Li, *Fellow, IEEE*

***Abstract*—Prosodic phrasing is an important factor that affects naturalness and intelligibility in text-to-speech synthesis. Studies show that deep learning techniques improve prosodic phrasing when large text and speech corpus are available. However, for low-resource languages, such as Mongolian, prosodic phrasing remains a challenge for various reasons. First, the database suitable for system training is limited. Second, word composition knowledge that is prosody-informing has not been used in prosodic phrase modeling. To address these problems, in this article, we propose a feature augmentation method in conjunction with a self-attention neural classifier. We augment input text with morphological and phonological decompositions of words to enhance the text encoder. We study the use of self-attention classifier, that makes use of global context of a sentence, as a decoder for phrase break prediction. Both objective and subjective evaluations validate the effectiveness of the proposed phrase break prediction framework, that consistently improves voice quality in a Mongolian text-to-speech synthesis system.**

***Index Terms*—Mongolian speech synthesis, prosodic phrasing, phrase break prediction, self-attention, morphological and phonological features.**

Manuscript received January 18, 2020; revised June 14, 2020 and September 22, 2020; accepted November 13, 2020. Date of publication November 25, 2020; date of current version December 14, 2020. This work was supported in part by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award AISG-GC-2019-002) and (AISG Award AISG-100E-2018-006), and its National Robotics Programme under Grant 192 25 00054, and in part by RIE2020 Advanced Manufacturing and Engineering Programmatic Grants A1687b0033, and A18A2b0046. The work of Rui Liu, Feilong Bao, and Guanglai Gao was supported in part by the National Key Research and Development Project (2018YFE0122900), in part by the National Natural Science Foundation of China (61773224, 62066033), in part by the National Natural Science Foundation of Inner Mongolia (2018MS06006), in part by Achievements Transformation Project of Inner Mongolia Autonomous Region (CGZH2018125), and in part by Applied Technology Research and Development Foundation of Inner Mongolia Autonomous Region (2019GG372, 2020GG0046). The work of Berrak Sisman was supported in part by SUTD Startup Grant Artificial Intelligence for Human Voice Conversion (SRG ISTD 2020 158) and SUTD AI Grant titled 'The Understanding and Synthesis of Expressive Speech by AI'. The associate editor coordinating the review of this manuscript and approving it for publication was Jianhua Tao. *(Corresponding authors: Feilong Bao; Jichen Yang.)*

Rui Liu is with the Department of Computer Science, Inner Mongolia University, Hohhot 010021, China, and also with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583, Singapore (e-mail: liurui_imu@163.com).

Berrak Sisman is with the Singapore University of Technology and Design, Singapore 487372, Singapore (e-mail: berraksisman@u.nus.edu).

Feilong Bao and Guanglai Gao are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583, Singapore (e-mail: csfeilong@imu.edu.cn; csggl@imu.edu.cn).

Jichen Yang is with the Department of Electrical and Computer Engineering, National University of Singapore, 117583, Singapore (e-mail: NisonYoung@163.com).

Haizhou Li is with the Department of Electrical and Computer Engineering, National University of Singapore, 117583 Singapore, Singapore, and also with the Kriston AI Lab,, China (e-mail: haizhou.li@nus.edu.sg).

Digital Object Identifier 10.1109/TASLP.2020.3040523

I. Introduction

ACCURATE prosodic phrasing improves text-to-speech (TTS) synthesis [1]–[5], that can be achieved by phrase break prediction [6]–[8]. Prosodic phrasing breaks a long utterance into prosodic units according to syntactic structure or intonational properties, that improves the naturalness and intelligibility of speech. More importantly, in speech synthesis, phrase breaking is often the first step in generating a prosody pattern, such as intonation prediction and duration modeling [9]–[11]. Any errors made in the phrase breaking are propagated to other downstream prosodic models, resulting in unnatural speech [12]–[14]. Nonetheless, some newly developed speech synthesis systems, such as Tacotron [15]–[21], WaveNet-based approaches [22]–[28], and Deep Voice [29] have not specifically modeled prosodic cues from input text. Therefore, they cannot explicitly control prosodic phrasing [30]. It remains a challenging research problem in speech synthesis [31] to identify prosodic phrase breaks from input text, that is the focus of this paper.

Statistical modeling approaches to prosodic phrasing include maximum entropy models [32], [33], hidden Markov models [34], and conditional random fields (CRF) [35], [36], which are trained with a large set of labeled data. Such methods build the models based on linguistic features, for example, part of speech (POS) and length of word [37], [38], that are discrete representations of words and their syntax. Deep learning approaches [39]–[41], such as deep neural networks (DNNs), recurrent neural networks (RNNs), bidirectional long short-term memory (BiLSTM), and representation learning [42]–[44], open the opportunities to represent the linguistic features in a continuous space, and to discover useful features from unlabeled data. For example, word embeddings are commonly used [45]–[53] to represent words and their syntax, which are used as input to predict prosodic breaks.

In [45], Watts *et al.* propose to use continuous-valued word embeddings, that summarize the distributional characteristics of

word types, as surrogates of POS features. In [48], Vadapalli and Gangashetty propose to use an RNN to predict phrase breaking of a word embedding sequence. These methods benefit from the context modeling ability of RNN. Despite the success, RNN modeling has its limitation. Due to its autoregressive nature, RNN faces considerable challenges when processing long sequences [54]. Self-attention [55] provides an alternative solution, which is effective in various sequential modeling tasks, such as machine translation [55], semantic role labeling [56], speech recognition [57], [58] and speech synthesis [59], [60], and recently prosodic phrasing [61].

The above prior studies, that use RNN or self-attention to model prosodic phrasing, achieve good performance in English and Mandarin applications. However, these models do not work well for highly agglutinative, low-resource, and morphologically rich languages, such as Mongolian, because there is a lack of training data for low-resource agglutinative languages [62]. As a result, the models produce poor prosodic phrasing for Mongolian sentences that generally contain a substantial number of out-of-vocabulary (OOV) words.

We note that morphological and phonological decompositions of words, referred to as subwords, are prosody-informing [63]. By using subword embeddings in addition to whole word embeddings as input, we address two problems. First, there are much fewer subwords than agglutinative whole words, therefore, we expect that subword embeddings are more reliable than whole word embeddings; Second, subwords are prosody-informing that directly contribute to phrase break prediction. We propose to use continuous-valued embeddings to represent lexical words, morphological subwords, and phonological subwords, and use the sequence of embeddings as the input for phrase break prediction. Furthermore, we include an attention layer in the text encoders that learns to weight the relative contributions of the various embeddings.

While there have been studies on the use of prosody-related linguistic features [64] and subwords, such as character, stem, and suffix [53], [65], to predict phrase breaks, this work is different from the prior work in many ways,

- For the first time, we propose the use of morphological and phonological decompositions of words to augment the input text for Mongolian prosodic phrasing.
- Unlike the prior work, where multiple linguistic features are concatenated directly, we propose an attention layer that learns to weight word and subword embeddings in the text encoder.
- Phrase break prediction is based on temporal information beyond adjacent words. We adopt a self-attention neural classifier, which handles long range dependency of words better than RNN [55].

This work is an extension to our previous work [66] with several novel contributions,

- A novel strategy is proposed to incorporate word, morpheme, syllable, and phoneme embeddings into phrase break prediction model.
- An attention layer is studied to weight various embeddings in the text encoders, that enhances word embeddings for phrase break prediction.

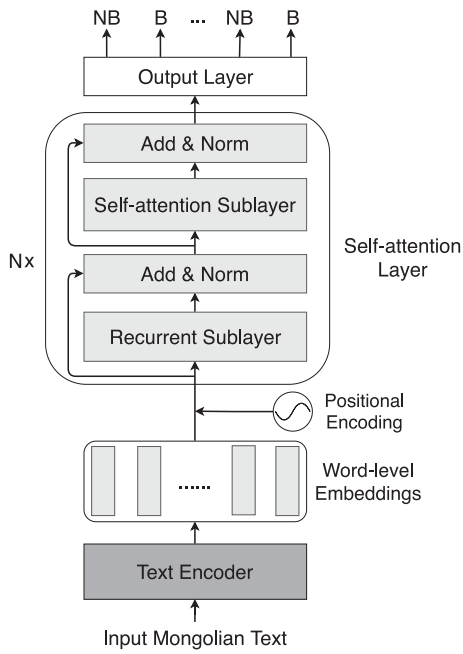


Fig. 1. Illustration of a text encoder and a self-attention neural classifier for phrase break prediction. The text encoder extracts word embeddings from input text; The neural classifier consists of N self-attention blocks that produce context-dependent intermediate representations, which are taken by the output layer for decision making.

- A self-attention neural classifier is proposed in conjunction with various text encoders through a comprehensive comparative study.

The remainder of this paper is organized as follows: in Section II, we study a basic self-attention framework for phrase break prediction. In Section III, we present the proposed enhanced text encoders that benefit from morphological and phonological information. We report the experiments in Section IV. Finally, Section V concludes the study.

II. Basic Self-Attention Model for Prosodic Phrasing

We first introduce a basic self-attention model for Mongolian prosodic phrase prediction, that represents the overall phrase break prediction framework. The prosodic phrasing model takes a sequence of words as input and generates their prosodic phrase break labels. The network structure is illustrated in Figure 1, which includes (1) a text encoder that encodes input text into a sequence of word embeddings; (2) a self-attention neural classifier, consisting of a self-attention layer and an output layer, that predicts the phrase break labels.

A. Text Encoder

The text encoder is designed to map input text into a sequence of word embeddings ω. There are many ways to implement the text encoder. An easy way is to use pre-trained word embedding, where a word embedding can be retrieved through table lookup. A word embedding network can be trained [42]–[44] from a text corpus. In this work, the skip-gram model [42], which performs

well for low-frequency words [43], is adopted to train the word embedding network.

Specifically, given a text corpus, the skip-gram model defines the embedding vector of each word via the matrix W^e and defines the context vector by means of the output matrix W^c. Given an input word w_I, let us label the corresponding row of W^e as vector v_{w_I}, which serves as the word embedding ω after skip-gram model is completed, and its corresponding column of W^c as v'_{w_I} (context vector). The output layer applies softmax to compute the probability of predicting the output word w_O given w_I. Therefore,

$$p(w_O|w_I) = \frac{\exp(v'^{\top}_{w_O} v_{w_I})}{\sum_{n=1}^{V} \exp(v'^{\top}_{w_n} v_{w_I})} \tag{1}$$

where O and I denote the index of output and input words in the vocabulary, and V is the vocabulary size.

The above learning process yields word embeddings that are distributed in a continuous semantic space; the position information of the words, which is crucial to the global context [55], is not fully modeled. To encode the position of each input word, sine and cosine functions of different frequencies [55] are adopted to represent the position encoding π of a word:

$$\pi(t, 2i) = \sin(t/10000^{2i/d}) \tag{2}$$

$$\pi(t, 2i+1) = \cos(t/10000^{2i/d}) \tag{3}$$

where t is the word position and i denotes the element in a d-dimensional encoding. As π vector has the same dimension as ω vector, they can be added to form a position-aware word embedding

$$x = \pi + \omega \tag{4}$$

Finally, the input sentence is encoded into a sequence of position-aware word embeddings $\mathbf{x} = \{x_1, x_2, ..., x_T\}$.

B. Self-Attention Layer

The main objective of the self-attention layer is to capture long-range dependency between word pairs in a sentence by using the attention weights. The layer consists of N identical self-attention blocks that contain a recurrent sublayer and a self-attention sublayer. The output layer is designed to make classification decision. The self-attention layer and output layer form a neural classifier for phrase break prediction.

1) Recurrent Sublayer: A recurrent sublayer is introduced to strengthen the sequential modeling. We implement the recurrent sublayer with a BiLSTM. Given a sequence of input embedding vectors $\mathbf{x} = \{x_1, x_2, ..., x_T\}$ for a sentence, two LSTMs process the inputs in opposite directions [67]. We then use the last hidden states ($\overrightarrow{s_t}$ and $\overleftarrow{s_t}$) from each of the LSTM components, combine them via the sum operation. The resulting hidden state, s_t forms a sequence, $\mathbf{s} = \{s_1, s_2, ..., s_T\}$, to represent the input sentence.

$$\overleftarrow{s_t} = \mathrm{LSTM}(x_t, \overleftarrow{s_{t+1}}) \tag{5}$$

$$\overrightarrow{s_t} = \mathrm{LSTM}(x_t, \overrightarrow{s_{t-1}}) \tag{6}$$

$$s_t = \overrightarrow{s_t} + \overleftarrow{s_t} \tag{7}$$

We provide the details of the forward LSTM next,

$$i_t = \sigma(W_{xi} x_t + W_{si} s_{t-1} + W_{ci} c_{t-1} + b_i) \tag{8}$$

$$c_t = (1 - i_t) \odot c_{t-1} + i_t \odot \tanh(W_{xc} x_t + W_{sc} s_{t-1} + b_c) \tag{9}$$

$$o_t = \sigma(W_{xo} x_t + W_{so} s_{t-1} + W_{co} c_t + b_o) \tag{10}$$

$$\overrightarrow{s_t} = o_t \odot \tanh(c_t) \tag{11}$$

where $\odot$ indicates element-wise product and σ indicates element-wise sigmoid function. x_t is the input vector and s_t is the hidden unit vector. W_{xi}, W_{xc}, W_{xo} are the different weight matrices for input x_t; W_{si}, W_{sc}, W_{so} denote the different weight matrices for hidden state h_t , W_{ci}, W_{co} are the different weight matrices for cell state c_t, and b_i, b_c, b_o denote the bias vectors.

2) Self-Attention Sublayer: The key component of the self-attention sublayer is the multihead self-attention [55], which consists of h attention heads, each of which learns a distinct attention function from different representation subspaces to attend at different positions in the sequence. Specifically, given a hidden state sequence $\mathbf{s}$, that is generated from the recurrent sublayer for an input sentence of T words, the multihead attention mechanism first maps $\mathbf{s} \in \mathbb{R}^{t \times d}$ to h different query, key and value matrices via linear projection. Formally, for the i^{th} head, we denote the queries, keys and values by $Q \in \mathbb{R}^{t \times d/h}$, $K \in \mathbb{R}^{t \times d/h}$ and $V \in \mathbb{R}^{t \times d/h}$ respectively. Then, scaled dot-product attention [55] is used to compute the context vectors:

$$\begin{aligned} M_i &= \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V) \\ &= \mathrm{softmax}(\frac{(QW_i^Q)(KW_i^K)^{\top}}{\sqrt{d}}) VW_i^V \end{aligned} \tag{12}$$

where W_i^Q, W_i^K, and W_i^V represent the learned linear maps that correspond to Q, K, and V, respectively.

Finally, the output $\mathbf{y} = \{y_1, y_2, ...y_T\}$ is computed as follows:

$$\mathbf{y} = MW \tag{13}$$

$$M = \mathrm{Concat}(M_1, ..., M_h) \tag{14}$$

where $M \in \mathbb{R}^{t \times d}$ and $W \in \mathbb{R}^{d \times d}$. The $\mathrm{Concat}(\cdot)$ function means that all the vectors produced by multiple parallel heads are concatenated to form a single vector.

3) Residual Connection & Layer Normalization: To facilitate the training, we employ a residual connection [68] around each of the two sublayers in Fig. 1, and apply layer normalization [69] after the residual connection to stabilize the activation of the deep neural network [55].

C. Output Layer

The output layer makes final decision for phrase break prediction. We study two implementations, namely, a $\mathrm{softmax}$ layer and a CRF layer.

1) Softmax Layer: The $\mathrm{softmax}$ layer computes a normalized probability distribution over all possible K phrase break labels for each word:

$$p(l_t = k|y_t) = \frac{\exp(w_k^o y_t)}{\sum_{\tilde{k}=1}^{K} \exp(w_{\tilde{k}}^o y_t)} \tag{15}$$

where $p(l_t = k|y_t)$ is the probability of the t^{th} word taking a phrase break k, and w_k^o is the k^{th} row of output weight matrix W^o.

To optimize this model, we minimize the categorical cross entropy, which is equivalent to minimizing the negative log-probability of the correct labels. For a sentence of T words, we have,

$$E = -\sum_{t=1}^{\mathrm{T}} \log(p(l_t = k_{ref}|y_t)) \tag{16}$$

where k_{ref} is the reference label.

2) CRF Layer: For sequence labeling, it is effective to consider the correlation between taggers and to jointly decode the best output for a given token. Following [53], we can also use a CRF as the output layer to condition each prediction on the previously predicted label, thereby allowing the network to look for the optimal path among all possible sequences. During training, we optimize the model by maximizing the CRF score $c(\mathbf{l})$ for the correct label sequence $\mathbf{l} = \{l_1, l_2, ..., l_T\}$ of a sentence of T words, while minimizing those for all other label sequences:

$$E = -c(\mathbf{l}) + \log \sum_{\tilde{\mathbf{l}} \in \Lambda} \exp(c(\tilde{\mathbf{l}})) \tag{17}$$

where Λ represents all possible label sequences for the sentence.

III. Enhanced Text Encoder for Word Embedding

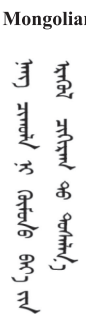
The basic self-attention model attends to words in all positions in an input sentence, thus captures long-range dependency for phrase break prediction. As the input sentence is encoded as a sequence of word embeddings, the performance of phrase break prediction depends on the appropriateness of word embeddings. Out-of-vocabulary problem in Mongolian presents a challenge to effective word embedding.

To address the problem, we propose to enhance word embeddings with morphologically and phonologically motivated subword embeddings. We believe that subword embeddings will complement whole word embeddings for agglutinative languages, such as Mongolian. One of the ways is to encode the input text by using lexical word, morpheme, syllable, and phoneme as input tokens.

Let's first discuss some characteristics of Mongolian language in Section III-A to set the stage for our study. We will then propose various text encoders designed for Mongolian in Sections III-B, III-C and III-D.

A. Mongolian Characteristics

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 [70]. Approximately 6 million people speak Mongolian around the world. Mongolian is one of the five major minority languages in China and is an official language of the Inner Mongolia Autonomous Region of China. Today, Mongolian is written in two different scripts: the classical Mongolian script used in China, and Cyrillic Mongolian used in Mongolia. In this work, we consider only the classical Mongolian script. An example of the classical Mongolian script and its Latin romanization are shown in Fig. 2. It is noted that classical Mongolian is written from top to bottom, left to right.



Mongolian

English translation:

Most importantly, it is good for human health.

Latin romanization:

neN qihvla ni homun-u bey_e-yin eregul qihirag-tv tvsalan_a.

Segmentation:

neN qihvla ni homun *-u* bey_e *-yin* eregul qihirag *-tv* tvsalan_a.

Phrase break label:

neN [NB] qihvla [NB] ni [B] homun [NB] -u [NB] bey_e [NB] -yin [B] eregul [NB] qihirag [NB] -tv [NB] tvsalan_a [B].

Fig. 2. The narrow non-breaking space (NNBS) suffixes within a Mongolian sentence. NNBS suffix segmented from a word is highlighted in italic. There are three pauses in the sentence, one of which is located at the NNBS suffix -yin.

As an agglutinative language, similar to Japanese, Korean, and Turkish, Mongolian has a complex morphological structure. Most Mongolian words can be decomposed into root, derivational suffixes and inflectional suffixes [71]–[73]. 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. The use of NNBS suffix is pervasive. For example, there are 3 NNBS suffixes in a sentence of only 8 words as illustrated in Fig. 2.

Many suffixes can be added to a word stem to generate new words. Suffixes often serve as a positive signal that implies the POS of a word. For example, in  ,  ,  ,  and  , the words share the same word stem "" (Latin romanization: "sandali"; English: "chair"). As a result, Mongolian has a large vocabulary of more than one million words, with only approximately thirty thousand unique stems [71]. The enormous vocabulary size leads to data sparsity in any word-based language modeling tasks. On the other hand, Mongolian is a low resource language which has less annotated text and speech data for word embedding training. We are interested in word embedding techniques that take good advantage of existing data resources for effective phrase break prediction.

In spoken Mongolian, syllables and phonemes are the basic unit of speech phonetically, while morphemes and characters are the basic unit in written form [74]. A Mongolian word consists of a sequence of syllables, each having several phonemes; a Mongolian word is written as a sequence of morphemes, each consisting of multiple characters. For instance, the Mongolian word "qihirag-tv" (English: "health"), "qihiju" (English: "put inside"), and "qihitai" (English: "wild donkey") are constructed with the same syllables "qi" and "hi". There is a general belief that Mongolian word, morpheme, syllable, and phoneme all

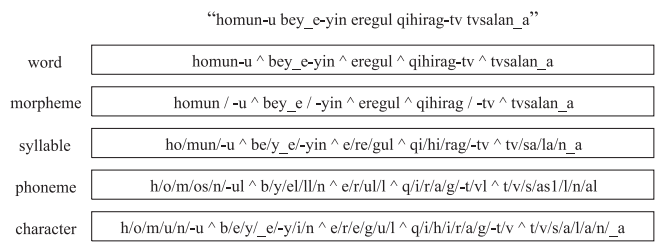


Fig. 3. Lexical analysis of a Mongolian sentence, where word, morpheme, syllable, phoneme, and character represent different level of basic unit. In this paper, word, morpheme, syllable, phoneme embeddings are used for phrase break prediction.

contribute to semantic-syntactic grouping of words in a sentence, therefore, prosodic phrasing. In Fig. 3, we use an exemplar to illustrate the decomposition of basic units and their relationship. We would like to study the use of the basic units as embedding tokens, as shown in Fig. 4, to enhance the word embeddings for phrase break prediction.

B. Morphologically Enhanced Text Encoder

Fig. 4(a) is the overall architecture of the morphologically enhanced text encoder. The input features for each word include two distinct units, word and morpheme. For each word, we first obtain the morpheme sequence using a rule-based morphological analyzer. Then, we use the BiLSTM embedding network to obtain the morphological embedding μ for each word. Finally, an attention layer is used to weight between word embedding ω and morphological embedding μ to form a joint embedding χ_m, which is then taken by a self-attention layer to decode the phrase break label.

First, the morpheme sequence, represented as a sequence of one-hot vectors, is processed by a BiLSTM using Equations 5 and 6. We then take the last hidden states from each of the LSTM components, concatenate them together, and pass the result through a separate nonlinear layer.

$$s = [\overleftarrow{s}\,;\overrightarrow{s}] \tag{18}$$

$$\mu = \tanh(W_m s) \tag{19}$$

where W_m is a weight matrix mapping the concatenated hidden states s from both LSTMs into a morphological embedding representation, denoted as μ.

We now have two feature representations for each word: ω is the word embedding as described in Section II-A, and μ is an intermediate representation dynamically built from the basic units in the t^{th} word of the input sentence. Following the idea in [75], instead of simply concatenating ω with μ, we concatenate the two embeddings via an attention layer:

$$w = \sigma(M_z^{(3)}\tanh(M_z^{(1)} \cdot \omega + M_z^{(2)} \cdot \mu)) \tag{20}$$

$$\chi_m = \mathrm{Concat}(w \cdot \omega, (1-w) \cdot \mu) \tag{21}$$

where $M_z^{(1)}$, $M_z^{(2)}$ and $M_z^{(3)}$ are the weight matrices for calculating w. $\sigma(\cdot)$ is the logistic function with values in the range [0, 1]. Vector w is of the same dimension as ω and μ and acts as the weight between the two vectors.

Finally, the enhanced word embeddings χ_m of morphological enhanced text encoder is combined with the position embedding π in the same way as Equation 4 to form a position-aware word embedding x, which serves as the input to the following self-attention layer as shown in Fig. 1.

C. Phonologically Enhanced Text Encoder

Fig. 4(b) is the overall architecture of the phonologically enhanced text encoder. The input features for each word include three distinct units, word and syllable and phoneme. We first transform Mongolian words into their phoneme and syllable sequence. Specifically, a phoneme sequence is generated by a rule-based Mongolian grapheme-to-phoneme conversion module. The syllable sequence is automatically obtained according to the Mongolian syllable construction rule. They are then mapped to a sequence of one-hot vectors, that are further processed by the BiLSTM network described in Section III-B using Equations 5 and 6.

Similar to that in morphologically enhanced text encoder, we concatenate the two last hidden vectors from two directions and pass the result through a separate nonlinear layer to generate the high-level phoneme embedding, denoted as χ_{phn}, and syllable embedding denoted as χ_{syl} using Equations 18 and 19.

We then concatenate χ_{phn} and χ_{syl} to obtain a word-level phonological embedding, denoted as χ_{ps}. We now have two feature representations for each word: ω is the word embedding as described in Section II-A, and χ_{ps} is its phonological embedding. These two embedding vectors are fused by an attention layer as described in Section III-B to produce an enhanced word embedding χ_p,

$$w' = \sigma(M_z^{(3)}\tanh(M_z^{(1)} \cdot \omega + M_z^{(2)} \cdot \chi_{ps})) \tag{22}$$

$$\chi_p = \mathrm{Concat}(w' \cdot \omega, (1-w') \cdot \chi_{ps}) \tag{23}$$

All parameters involved in the above formula are configured in the same way as in Section III-B. Finally, the enhanced word embeddings χ_p is combined with the position embedding π in the same way as Equation 4 to form a position-aware word embedding x, that serves as the input to the following self-attention layer as shown in Fig. 1.

D. Morphologically-Phonologically Enhanced Text Encoder

By combining both morphologically enhanced and phonologically enhanced text encoders, we study the effect of the combined system in Fig. 4(c).

The input features for each word now consist of three distinct components: the word embedding ω and two pieces of complementary information, i.e., its morphological embedding μ and phonological embedding χ_{ps}. We first obtain the word, morphological, and phonological embeddings in the same way as discussed in Sections III-B and III-C. We then fuse the the three embeddings via an attention layer to obtain the morphologically and phonologically enhanced word embedding χ_{mp}.

$$w_2'' = 1 - \sigma(M_{z_1}^{(3)}\tanh(M_{z_1}^{(1)} \cdot \omega + M_{z_1}^{(2)} \cdot \chi_{ps})) \tag{24}$$

$$w_3'' = 1 - \sigma(M_{z_2}^{(3)}\tanh(M_{z_2}^{(1)} \cdot \omega + M_{z_2}^{(2)} \cdot \mu)) \tag{25}$$

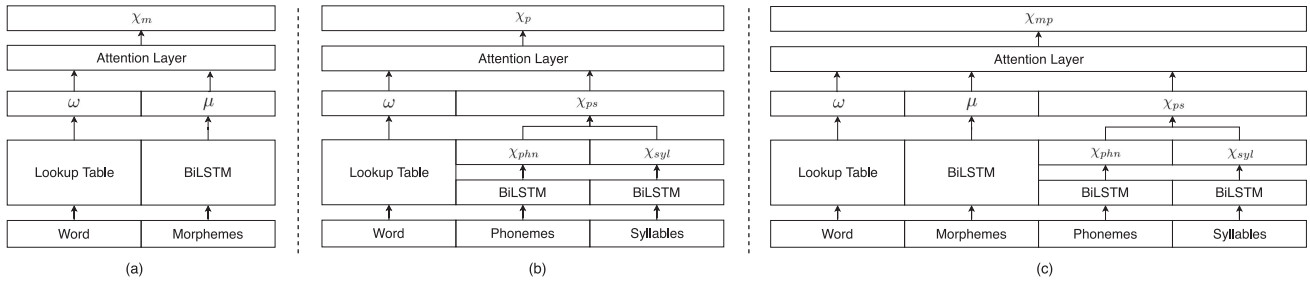


Fig. 4. The network architecture of enhanced text encoders. The left panel (a) is the morphologically enhanced text encoder. The middle panel (b) is the phonologically enhanced text encoder. The right panel (c) is the morphologically-phonologically enhanced text encoder.

$$w_1'' = 1 - w_2'' + w_3'' \tag{26}$$

$$\chi_{mp} = \text{Concat}(w_1'' \cdot \omega, w_2'' \cdot \chi_{ps}, w_3'' \cdot \mu) \tag{27}$$

where $M_{z_1}^{(1)}$, $M_{z_1}^{(2)}$, $M_{z_1}^{(3)}$ and $M_{z_2}^{(1)}$, $M_{z_2}^{(2)}$, $M_{z_2}^{(3)}$ are weight matrices for calculating w_1'', w_2'' and w_3'', and $\sigma(\cdot)$ is the logistic function with values in the range of [0, 1]. Finally, the enhanced word embeddings χ_{mp} is combined with the position embedding π in the same way as Equation 4 to form a position-aware word embedding x, that serves as the input to the following self-attention layer as shown in Fig. 1.

To summarize, we have formulated three text encoders that learn morpheme, syllable, and phoneme embeddings to enhance word embeddings with morphological and phonological features. We denote the proposed text encoders as M_Enc, P_Enc and MP_Enc. Next we evaluate their performance for phrase break prediction in Mongolian TTS.

IV. Experiments

We now evaluate the effectiveness of the proposed enhanced word embedding techniques in phrase break prediction, and their contributions to a DNN-based Mongolian TTS system [76].

A. Experimental Corpora

Mongolian speech data: We conducted experiments on a well-formulated Mongolian TTS database, that has rich phonetic and prosodic content [62], [76]. The database contains Mongolian daily expressions recorded by a female native speaker. Speech signal is recorded at 16 kHz sampling rate.

Prosodic prediction data: We use the text transcript of the Mongolian speech data as the training data of the phrase break prediction model. The transcript contains 59k sentences and more than 409k words, 1,065k syllables, 500k morphemes and 1,885k phonemes. 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 131k, and the average length of prosodic phrases is 3.5 words. To the best of our knowledge, this is the first database with prosodic phrase labels in Mongolian, and we will make this database publicly available in the near future. 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.

Mongolian text data: We pre-trained a skip-gram word embedding network on text data from mainstream Mongolian websites. After cleaning up, we obtained a text corpus of approximately 200 million words with a vocabulary size of 3 million. For phrase break prediction network training, all Mongolian words were Latin-cased before passing through the lookup table to retrieve their word embeddings.

B. Experimental Setup

To verify the effectiveness of our proposed text encoders, namely M_Enc, P_Enc and MP_Enc, we choose a prosodic prediction model with pre-trained word embeddings, denoted as W_Enc, as the baseline. We conducted three experiments to evaluate different phrase break prediction classifiers in combination with three proposed text encoders. The first experiment is designed to test a CRF classifier; the second experiment is designed to test a BiLSTM classifier; the third experiment is designed to test the proposed self-attention neural classifier. The default output layer is a softmax layer. In all the experiments, we use TensorFlow [77] to build the models.

1) Experiment 1 (CRF): We compare word embeddings generated by W_Enc, M_Enc, P_Enc and MP_Enc as the input to a CRF classifier, which predicts phrase break labels. The CRF classifier is a linear-chain CRF implemented with CRF++ toolkit.[1] It is a non-neural classifier solution. Bigram windows (previous two words, current words, and future two words) are used to learn the context information.

2) Experiment 2 (BiLSTM): We futher compare word embeddings generated by W_Enc, M_Enc, P_Enc and MP_Enc as the input to a BiLSTM classifier. The BiLSTM classifier has two hidden layers, each containing 160 memory blocks in each direction. The dropout method [78] is employed to regularize our model, and the dropout rate is set to 0.5.

3) Experiment 3 (Self-Attention): We finally compare word embeddings generated by W_Enc, M_Enc, P_Enc and MP_Enc as the input to the self-attention neural classifier illustrated in Section II, when predicting phrase breaks. The self-attention layer consists of 5 hidden layers with $h = 8$ heads. We also apply dropout to prevent the networks from overfitting. The dropout

[1]https://taku910.github.io/crfpp/

layers are added before the residual connections with a keep probability of 0.8.

For the text encoders, the size of LSTM-related layers is set to 200 in both directions. For M_Enc, the morphological embedding μ is of the same dimension as that of word embedding ω. Likewise, the phonological embedding χ_{ps} is of the same dimension as that of ω in P_Enc. For MP_Enc, the sum of μ and χ_{ps} dimension equals to that of ω. Specifically, the dimension of ω is set to 100. All tokens are initialized with 100-dimensional pre-trained vectors, as illustrated in Section II-A and updated during training. Both μ and χ_{ps} are randomly initialized with a uniform distribution in the interval $[-0{:}05; 0{:}05]$.

We set the learning rate to 1.0 and the batch size to 64. All parameters are optimized using the AdaDelta [79] algorithm. In every epoch, we calculate the performance on the training set: we stop training if the effectiveness does not increase for seven epochs. The above parameter choices are based on the model performance on the development data, which consists of 25% of the training data. Finally, the best model on the training stage is then evaluated on the test set.

For data preprocessing, all digits are replaced with the character "0" as digits don't carry semantic meaning [80]. Any words that occur only once in the training data are replaced by the generic OOV token for word embeddings but are still used in the phonological and morphological embedding components. For syllable liaison phenomenon, we restore the liaison words into two independent words, and then get their syllable sequences separately. According to statistics, the liaison phenomenon only occurs at about 3k words, and 90% of them at the end of sentences. For the 409k words of all data, 3k words only account for a small proportion, which we believe that it don't have a serious impact on the experiment results.

As the text in the databases has already been annotated with phrase break labels, the ground truth is available to compute the performance of our approaches. We report the performance of our approaches in terms of the precision, recall and F1 score, which is defined as the harmonic mean of precision and recall. The F1 score ranges from 0.0 to 100.0, with a higher value indicating better performance.

C. Objective Evaluation

We report three experiments for the proposed text encoders in combination with three different classifiers, i.e. CRF, BiLSTM and self-attention in Table I. It is observed that all enhanced text encoders consistently outperform W_Enc baseline across all CRF, BiLSTM and self-attention model.

In Expt 1, 85.55% of F1 is reported for the baseline system (W_Enc) with word embeddings only. We improve F1 to 86.10% with M_Enc, 86.93% with P_Enc, and 87.45% with MP_Enc. It is noted that word embedding alone (W_Enc) performs poorly due to inadequate training of word embedding for a large vocabulary on a small dataset. By incorporating phonological or morphological embeddings to enhance word embeddings, we achieve clear performance gain. MP_Enc combines phonological and morphological embeddings to achieve the best performance for Mongolian phrase break prediction.

TABLE I
SYSTEM PERFORMANCE (%) OF MONGOLIAN PHRASE BREAK PREDICTION WITH DIFFERENT TEXT ENCODERS. THE OPTIMAL MODEL IS SIGNIFICANTLY BETTER THAN ALL THE COMPARISON SYSTEMS WITH P-VALUE < 0.01

#	Model	Encoder	Precision	Recall	F1
Expt 1	CRF	W_Enc	85.11	85.46	85.55
		M_Enc	85.52	86.03	86.10
		P_Enc	86.25	86.83	86.93
		MP_Enc	87.14	87.28	87.45
Expt 2	BiLSTM	W_Enc	88.09	90.01	88.58
		M_Enc	88.21	90.13	90.21
		P_Enc	88.35	90.18	90.39
		MP_Enc	89.10	90.53	91.05
Expt 3	Self-Attention	W_Enc	90.03	89.92	90.07
		M_Enc	92.11	92.33	92.38
		P_Enc	92.29	92.45	92.56
		MP_Enc	93.22	93.43	**93.37**

TABLE II
SYSTEM PERFORMANCE (%) OF MONGOLIAN PHRASE BREAK PREDICTION WITHOUT AN ATTENTION LAYER INSIDE TEXT ENCODERS. THE CLASSIFIER FOLLOWS THE CONFIGURATION IN EXPT 3

Encoder	Without attention layer			p-value
	Precision	Recall	F1	
M_Enc	91.87	92.01	92.03	0.00211
P_Enc	91.71	91.83	91.85	0.00103
MP_Enc	92.68	92.80	92.83	0.00134

In Expt 2, we observe a similar trend as in experiment 1. It is worth noting that all results in experiment 2 consistently outperform their counterparts in experiment 1, confirming the effectiveness of contextual modeling of LSTM for phrase break prediction. The results further confirm that morphological and phonological information contributes substantially to performance gain.

In Expt 3, we observe a similar trend as in experiments 1 and 2. We are encouraged by the fact that self-attention neural classifier consistently outperforms both CRF and BiLSTM models. The F1 score of with MP_Enc is reported at 93.27% which is the highest in all experiments. The results confirm our intuition that structural information over long range in a sentence contributes to phrase break prediction, and self-attention layer effectively models such long range dependency.

All experiments confirm the effectiveness of morphological and phonological information in phrase break prediction. The performance is further improved by appropriate contextual models such as BiLSTM and self-attention. To better understand the role of the internal components in Expt 3, we conduct ablation tests in Section IV-D.

D. Ablation Tests

In this section, we analyze the main factors that affect the phrase break prediction performance in experiment 3.

1) Attention Layers inside Text Encoders: As discussed in Section III, an attention layer is adopted in the text encoder to dynamically weight various inputs, i.e. word, morpheme, syllable, and phoneme. Table II reports results from ablation

TABLE III
ABLATION TEST OF SYSTEM PERFORMANCE (%) FOR MONGOLIAN PHRASE BREAK PREDICTION. THE CLASSIFIER FOLLOWS THE CONFIGURATION IN EXPT 3 WITH MP_ENC. THE BEST RESULT OUTPERFORMS OTHERS WITH A P-VALUE < 0.01

Recurrent sublayer	Positional encoding	Depth	Precision	Recall	F1
yes	yes	2	92.07	91.98	92.08
yes	yes	3	93.10	92.99	93.11
yes	yes	4	93.05	93.11	93.28
yes	yes	5	93.22	93.43	**93.37**
yes	yes	6	92.17	92.19	92.35
no	yes	5	90.99	91.02	91.22
yes	no	5	85.24	85.72	85.89

TABLE IV
SYSTEM PERFORMANCE (%) FOR MONGOLIAN PHRASE BREAK PREDICTION WITH DIFFERENT OUTPUT LAYERS. THE CLASSIFIER FOLLOWS THE CONFIGURATION IN EXPT 3 WITH MP_ENC

Output layer	Precision	Recall	F1	p-value
Softmax	93.22	93.43	93.37	0.00102
CRF	92.37	92.51	92.82	0.00171

tests without an attention layer inside the text encoders. "Without attention layer" means that the multiple input embeddings are simply concatenated without attention weights.

We note that the performance in Table II is consistently lower than that in the last three rows of Table I, which highlights the importance of the attention layer for fusion of information, as lexical word, morphological unit, and phonological unit each has differentiated contributions to phrase break prediction.

2) Configuration of Recurrent Sublayer: In this experiment, we would like to study the effect of the depth of recurrent sublayer on the performance for phrase break prediction. Although the self-attention model has strong sequential modeling ability, it still requires nonlinear sublayers to enhance its expressive power. To demonstrate our idea, we conduct ablation experiments. Table III reports the test results.

It is noted that the performance increases as more layers are added, that saturates at 6 layers. We consider 5 layers are sufficient for prosodic phrase break modeling. The penultimate row shows the results of a 5-layer model without a recurrent sublayer, which has a lower performance than that of a 2-layer model with recurrent sublayer. The result suggests that the recurrent sublayers are essential components in the classifier.

The last row shows the results of the 5 layer-model without positional encoding. The significant decline in F1 score indicates that positional encodings are indispensable for our model: armed with the positional encoding, the model benefit from long range contextual information.

3) Choice of Output Layer: In this experiment, we study the effect of different output layers on the phrase break prediction performance. We compare CRF [53], which conditions each prediction on the previously predicted label, and a softmax layer. Table IV summarizes the experimental results for the best system (MP_Enc with self-attention classifier) in Expt 3, but with two different output layers, softmax or CRF.

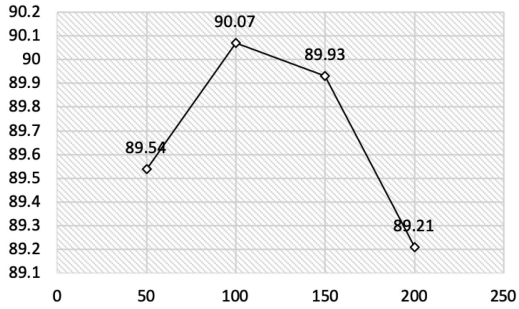


Fig. 5. Effect of varying the word embedding ω dimension on F1 Score (%) of the W_Enc system. All p-values are lower than 0.01.

Surprisingly, the softmax layer outperforms the CRF layer across all metrics. This result is in contrary to other prior empirical finding where the CRF output layer is likely to work better than the softmax output layer on other sequential labeling tasks, e.g., POS tagging [81] and name entity recognition (NER) [82]. We believe that this could be a result of two factors. First, the nature of the specific tasks. For NER, which gives each word an entity label e.g., time, location, organization, person and money, the distribution of different labels in a corpus reflect certain syntactic rules. Therefore, contextual entity labeling matters more when decoding the current label. By contrast, there are fewer phrase breaks, i.e. NB and B, than NER labels. Furthermore, the distribution of phrase break labels is highly unbalanced. The ratio between NB and B is roughly 85% to 15%. As a result, the phrase break labeling sequence may not be as informative as that in NER. Second, the nature of the specific model. Unlike traditional RNN-based sequence labeling models [66], [81], [82] that capture sequential context, self-attention sublayer in Fig. 1 connects two arbitrary words directly regardless of their distance [56], [61]. Furthermore, the recurrent sublayer in Fig. 1 well captures the long-range sequential dependency, therefore, the self-attention layer in Fig. 1 doesn't rely on an output layer to model the phrase break labeling sequence for decision making.

E. Analysis

To validate the usefulness of morphological and phonological cues, we would like look into how the proposed text encoders improve performance in out-of-vocabulary situations.

1) Morphological and Phonological Cues: We would like to examine if morphological and phonological cues are indeed informative. The morphological and phonological embeddings in the text encoders have a larger number of parameters than the word embedding baseline due to an increase of input dimension if all other hyperparameters are held constant. Such increase of model parameters may cause a fluctuation in performance.

To confirm that this effect does not have a material impact on the results, we ran an additional experiment to study the effect of varying the word embedding dimension on the performance of the word embedding baseline W_Enc. Specifically, we varied the dimension of word embedding ω from 50 to 200 and evaluate the performance, in terms of F1 Score. Fig. 5 shows the performance of phrase break prediction as a function of the dimension of word

TABLE V
SYSTEM PERFORMANCE (%) FOR MONGOLIAN PHRASE BREAK PREDICTION ON TWO OPEN TEST SETS. IV SET DOES NOT CONTAIN ANY OUT-OF-VOCABULARY WORDS, WZHILE OOV SET ONLY CONSISTS OF OUT-OF-VOCABULARY WORDS. THE PROPOSED MODELS SIGNIFICANTLY OUTPERFORM THE BASELINE WITH P-VALUE < 0.01

Encoder	IV set			OOV set		
	Precision	Recall	F1	Precision	Recall	F1
W_Enc	89.76	90.14	90.11	85.01	87.54	85.32
M_Enc	91.85	92.05	92.23	85.95	86.62	86.46
P_Enc	91.99	92.16	92.31	87.96	88.32	88.02
MP_Enc	92.01	92.85	**92.88**	91.02	90.17	**90.38**

embedding ω, which peaks at 100. A higher dimension does not lead to better performance.

For a fair comparison, we compare the 200-dimension word embedding in W_Enc with the three enhanced word embeddings, namely χ_p, χ_m, and χ_{mp} obtained from the three systems M_Enc, P_Enc and MP_Enc which have the same number of parameters. We observe that the F1 score (90.07%) of W_Enc system is lower than M_Enc, P_Enc and MP_Enc, by 2.49%, 2.31% and 3.30%, respectively. The results again confirm that the morphological and phonological embeddings bring to the model additional cues for phrase break prediction.

Overall, the results suggest that solely increasing the embedding dimension does not lead to substantial improvements and that the use of meaningful linguistic units for representation is important.

2) Effect on Out-of-Vocabulary Words: As described in Section I, the proposed method seeks to relieve the out-of-vocabulary problem, i.e. if a word has never been seen before, instead of being labelled as (UNK), its word embedding is enhanced by phonological or morphological decompositions. To validate the idea in phrase break prediction task, we report the performance of W_Enc, M_Enc, P_Enc, and MP_Enc on two open test sets which have a contrastive out-of-vocabulary situations. A total of 600 sentences form the open test set: 300 sentences without out-of-vocabulary words, called "IV set", and another 300 sentences called "OOV set", including only out-of-vocabulary words.

As shown in Table V, the W_Enc system performs better on "IV set" than "OOV set". Additionally, although M_Enc and P_Enc also perform better on "IV set" than "OOV set", the gap between the two F1 scores is smaller than that of W_Enc. It is encouraging to observe that the MP_Enc system performs almost equally well on both test sets and significantly outperforms the W_Enc baseline. Once again, we confirm that MP_Enc system effectively attenuates the out-of-vocabulary problem in phrase break prediction.

To analyze the behavior of various text encoders, we use one example extracted from "OOV set" to compare their predicted phrase break labels in Table VI. The prediction errors are highlighted in gray. W_Enc text encoder doesn't provide sufficient information for some morphologically or phonologically rich words, such as "baigvlvmji-yin", "ogereqilelte-yin", and "batvlagsan". Therefore, the W_Enc model is unable to make informed decisions. We further observe that M_Enc and P_Enc both improve phrase break prediction for the sentence. For example, the words "baigvlvmji-yin" and "ogereqilelte-yin" were correctly predicted by virtue of their abundant internal word information. Finally, MP_Enc leverages both morphological and phonological information to eliminate all prediction errors.

TABLE VI
AN EXAMPLE OF PREDICTED PHRASE BREAK LABELS PRODUCED BY VARIOUS TEXT ENCODERS. THE LABELS NOT CORRECTLY PREDICTED FROM THE SOURCE MONGOLIAN TEXT (LATIN ROMANIZATION) ARE HIGHLIGHTED IN GRAY. ("*" DENOTES WORD BOUNDARY.)

Source Text:
toro-yin yabvdal-vn hwriyan-v baigvlvmji-yin ogereqilelte-yin tosul-i hinan batvlagsan yabvdal bwl
Morpheme Sequence:
toro -yin * yabvdal -vn * hwriyan -v * baigvlvmji -yin * ogereqilelte -yin * tosul -i * hinan * batvlagsan * yabvdal * bwl
Phoneme Sequence:
t o r Il n * y a b d as1 l il n * h o r ol n El * b ael g l as1 m j Il n * o g os r q l os l t Il n * t o s os l Il * h a n as1 n * b a t as1 l s as1 n * y a b d as1 l * b w l
Syllable Sequence:
to ro -yin * ya bv dal -vn * hw ri yan -v * bai gv lvm ji -yin * o ge re qi lel te -yin * to sul -i * hi nan * ba tv lag san * ya bv dal * bwl
Target Label:
NB NB NB B NB B NB NB NB B
W_Enc
toro-yin [NB] yabvdal-vn [NB] hwriyan-v [B] baigvlvmji-yin [NB] ogereqilelte-yin [B] tosul-i [B] hinan [B] batvlagsan [B] yabvdal [NB] bwl [B]
M_Enc
toro-yin [NB] yabvdal-vn [NB] hwriyan-v [NB] baigvlvmji-yin [B] ogereqilelte-yin [NB] tosul-i [B] hinan [B] batvlagsan [B] yabvdal [NB] bwl [B]
P_Enc
toro-yin [NB] yabvdal-vn [NB] hwriyan-v [NB] baigvlvmji-yin [B] ogereqilelte-yin [NB] tosul-i [B] hinan [B] batvlagsan [NB] yabvdal [NB] bwl [B]
MP_Enc
toro-yin [NB] yabvdal-vn [NB] hwriyan-v [NB] baigvlvmji-yin [B] ogereqilelte-yin [NB] tosul-i [B] hinan [NB] batvlagsan [NB] yabvdal [NB] bwl [B]

F. Subjective Evaluation

We conducted listening tests to further evaluate the contributions of phrase break prediction to speech synthesis quality. We compare four DNN-based Mongolian TTS systems [76], which differ only in terms of prosodic break inputs.

1) A/B Preference Test: We conducted an A/B preference test to compare the naturalness of synthesized speech. A set of 100 sentences were randomly selected from the test set for this listening test. We predicted their phrase break labels using W_Enc, M_Enc, P_Enc, and MP_Enc model with self-attention classifier. A group of 10 subjects were invited to perform the listening test. The preference percentages are reported in Fig. 6. The MP_Enc results are the most preferred ones across all three pairwise preference tests.

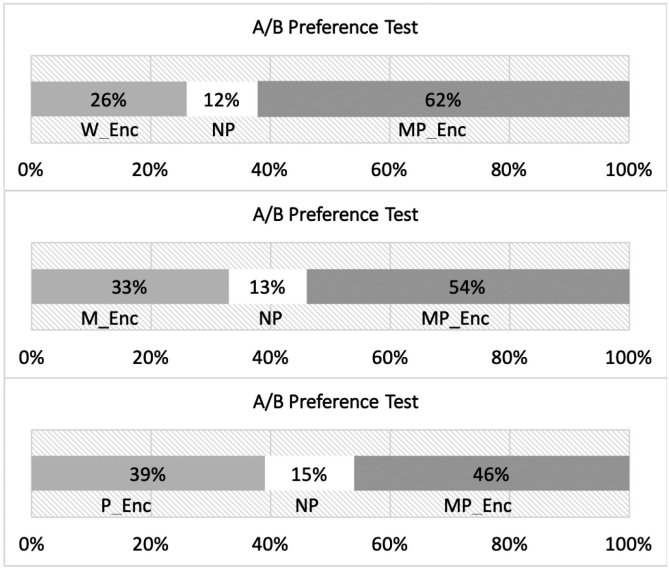


Fig. 6. The results of three pairwise A/B preference tests with a confidence level of 95%. The p-values are 0.00184, 0.00203 and 0.00126 respectively.

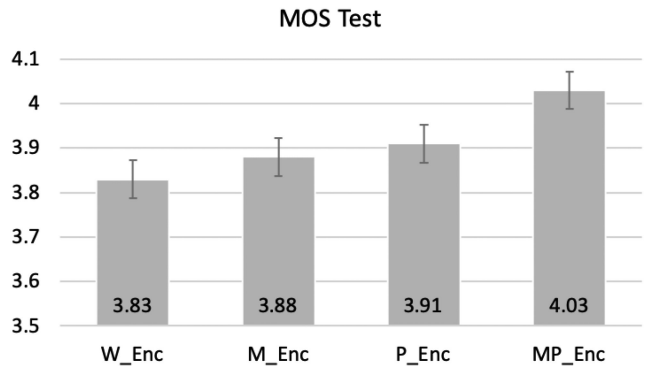


Fig. 7. MOS scores of speech quality with the 95% confidence intervals for four comparative systems with W_Enc, M_Enc, P_Enc, and MP_Enc text encoders.

2) Mean Opinion Score Test: We further conducted a 5-point mean opinion score (MOS) test ("5" for excellent, "4" for good, "3" for fair, "2" for poor, and "1" for bad). A set of 100 sentences were randomly selected from the test set for MOS test. Fig. 7 compares the results. It is observed that M_Enc, P_Enc and MP_Enc systems outperform the W_Enc baseline, that shows the clear advantage of phonological and morphological features. As the four systems only differ in terms of phrase break predictions, the results suggest that better phrase break prediction leads to higher voice quality. Among the four systems, MP_Enc shows the best performance, with an MOS of 4.03.

V. Discussion and Conclusion

In this article, we investigate the use of morphological and phonological features for phrase break prediction in a Mongolian TTS system. We explore the way to encode input text in terms of word, morpheme, syllable, and phoneme through various text encoders. We show that a self-attention classifier effectively captures long range contextual information that improves phrase break prediction. The proposed phrase break prediction is particularly effective for agglutinative languages, as evidenced in the experiment where a large number of out-of-vocabulary Mongolian words are involved. The proposed framework does not require additional feature engineering specific to the task or language, nor additional training data. While data-driven represents the mainstream TTS solution, we believe that linguistically motivated features remain useful especially for low-resource languages.

References

[1] M. Shannon, H. Zen, and W. Byrne, "Autoregressive models for statical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 3, pp. 587–597, Mar. 2013.

[2] N. Obin and P. Lanchantin, "Symbolic modeling of prosody: From linguistics to statistics," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 3, pp. 588–599, Mar. 2015.

[3] R. Liu, B. Sisman, F. Bao, G. Gao, and H. Li, "Modeling prosodic phrasing with multi-task learning in tacotron-based TTS," *IEEE Signal Process. Lett.*, vol. 27, pp. 1470–1474, Aug. 2020.

[4] R. Liu, B. Sisman, and H. Li, "Graphspeech: Syntax-aware graph attention network for neural speech synthesis," 2020, *arXiv:2010.12423*.

[5] R. Liu, B. Sisman, G. Gao, and H. Li, "Expressive TTS training with frame and style reconstruction loss," 2020, *arXiv:2008.01490*.

[6] C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," *Journal Acoustical Soc. America*, vol. 91, no. 3, pp. 1707–1717, 1992.

[7] J. Yamagishi *et al.*, "Robust speaker-adaptative HMM-based text-to-speech synthesis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 6, pp. 1208–1230, Aug. 2009.

[8] H. Zen, M. J. F. Gales, Y. Nankaku, and K. Tokuda, "Product of experts for statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 794–805, Mar. 2012.

[9] 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, Mar. 2013.

[10] H. Ze, A. 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.

[11] A. Lacheret, N. Obin, and M. Avanzi, "Design and evaluation of shared prosodic annotation for spontaneous french speech: from expert knowledge to non-expert annotation," in *Proc. 4th Linguistic Annotation Workshop*, 2010, pp. 265–273.

[12] N. Obin, A. Lacheret, C. Veaux, X. Rodet, and A. C. Simon, "A method for automatic and dynamic estimation of discourse genre typology with prosodic features," in *Proc. INTERSPEECH*, 2008, pp. 1204–1207.

[13] N. Obin, V. Dellwo, A. Lacheret, and X. Rodet, "Expectations for discourse genre identification: A prosodic study," in *Proc. INTERSPEECH*, 2010, pp. 3070–3073.

[14] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," 2020, *arXiv:2010.14794*.

[15] Y. Wang *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proc. INTERSPEECH*, 2017, pp. 4006–4010.

[16] Y. Wang *et al.*, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. 35th Int. Conf. Mach. Learn.*, vol. 80, 2018, pp. 5180–5189.

[17] R. Skerry-Ryan *et al.*, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," *Proc. 35th Int. Conf. Mach. Learn.*, vol. 80, p. 4693–4702, 2018.

[18] D. Stanton, Y. Wang, and R. Skerry-Ryan, "Predicting expressive speaking style from text in end-to-end speech synthesis," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 595–602.

[19] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, 2019, pp. 5911–5915.

[20] R. Liu, B. Sisman, J. Li, F. Bao, G. Gao, and H. Li, "Teacher-student training for robust tacotron-based TTS," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, pp. 6274–6278.

[21] R. Liu, B. Sisman, F. Bao, G. Gao, and H. Li, "WaveTTS: Tacotron-based TTS with Joint Time-Frequency Domain Loss," in *Proc. Odyssey 2020: Speaker Lang. Recognit. Workshop*, 2020, pp. 245–251. [Online]. Available: http://dx.doi.org/10.21437/Odyssey.2020-35

[22] A. v. d. Oord *et al.*, "Wavenet: A generative model for raw audio," in *Proc. 9th ISCA Speech Synthesis Workshop*, 2016, p. 1.

[23] T. Okamoto, K. Tachibana, T. Toda, Y. Shiga, and H. Kawai, "An investigation of subband wavenet vocoder covering entire audible frequency range with limited acoustic features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, 2018, pp. 5654–5658.

[24] B. Sisman, M. Zhang, S. Sakti, H. Li, and S. Nakamura, "Adaptive wavenet vocoder for residual compensation in gan-based voice conversion," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 282–289.

[25] L. Liu, Z. Ling, Y. Jiang, M. Zhou, and L. Dai, "Wavenet vocoder with limited training data for voice conversion," in *Proc. INTERSPEECH*, 2018, pp. 1983–1987.

[26] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with wavenet-based waveform generation," in *Proc. INTERSPEECH*, 2017, pp. 1138–1142.

[27] B. Sisman, M. Zhang, and H. Li, "Group sparse representation with wavenet vocoder adaptation for spectrum and prosody conversion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 6, pp. 1085–1097, Jun. 2019.

[28] B. Sisman, M. Zhang, and H. Li, "A voice conversion framework with tandem feature sparse representation and speaker-adapted wavenet vocoder," in *Proc. INTERSPEECH*, 2018, pp. 1978–1982.

[29] A. Gibiansky *et al.*, "Deep voice 2: Multi-speaker neural text-to-speech," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2962–2970.

[30] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 1, no. 1, pp. 1–12, Dec. 2020.

[31] G. Beller, N. Obin, and X. Rodet, "Articulation degree as a prosodic dimension of expressive speech," in *Proc. 4th Int. Conf. Speech Prosody*, 2008, pp. 1–5.

[32] J. Li, G. Hu, and R. Wang, "Chinese prosody phrase break prediction based on maximum entropy model," in *Proc. INTERSPEECH*, 2004, pp. 729–732.

[33] V. K. R. Sridhar, S. Bangalore, and S. S. Narayanan, "Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 797–811, May 2008.

[34] L. R. Rabiner and B.-H. Juang, "An introduction to hidden Markov models," *IEEE ASSP Mag.*, vol. 3, no. 1, pp. 4–16, Jan. 1986.

[35] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn.*, vol. 80, 2001, pp. 282–289.

[36] Y. Qian, Z. Wu, X. Ma, and F. Soong, "Automatic prosody prediction and detection with conditional random field (CRF) models," in *Proc. 7th Int. Symp. Chin. Spoken Lang. Process.*, 2010, pp. 135–138.

[37] P. Taylor and A. W. Black, "Assigning phrase breaks from part-of-speech sequences," *Comput. Speech Lang.*, vol. 12, no. 2, pp. 99–117, 1998.

[38] A. Parlikar and A. W. Black, "A grammar based approach to style specific phrase prediction," in *Proc. INTERSPEECH*, 2011, pp. 2149–2152.

[39] K. Zhou, B. Sisman, and H. Li, "Transforming Spectrum and Prosody for Emotional Voice Conversion with Non-Parallel Training Data," in *Proc. Odyssey 2020: Speaker Lang. Recognit. Workshop*, 2020, pp. 230–237. [Online]. Available: http://dx.doi.org/10.21437/Odyssey.2020-33

[40] K. Zhou, B. Sisman, M. Zhang, and H. Li, "Converting anyone's emotion: Towards speaker-independent emotional voice conversion," in *Proc. INTERSPEECH*, 2020, pp. 3416–3420.

[41] K. E. Ak, A. A. Kassim, J. Hwee Lim, and J. Yew Tham, "Learning attribute representations with localization for flexible fashion search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7708–7717.

[42] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[43] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. 1st Int. Conf. Learn. Rep.*, 2013, pp. 1–12.

[44] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, 2014, pp. 1532–1543.

[45] O. Watts, J. Yamagishi, and S. King, "Unsupervised continuous-valued word features for phrase-break prediction without a part-of-speech tagger," in *Proc. INTERSPEECH*, 2011, pp. 2157–2160.

[46] A. Vadapalli and K. Prahallad, "Learning continuous-valued word representations for phrase break prediction," in *Proc. INTERSPEECH*, 2014, pp. 41–45.

[47] O. Watts *et al.*, "Neural net word representations for phrase-break prediction without a part of speech tagger," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, 2014, pp. 2599–2603.

[48] A. Vadapalli and S. V. Gangashetty, "An investigation of recurrent neural network architectures using word embeddings for phrase break prediction," in *Proc. INTERSPEECH*, 2016, pp. 2308–2312.

[49] Y. Zheng *et al.*, "Improving prosodic boundaries prediction for Mandarin speech synthesis by using enhanced embedding feature and model fusion approaches," in *Proc. INTERSPEECH*, 2016, pp. 3201–3205.

[50] A. Rendel, R. Fernandez, R. Hoory, and B. Ramabhadran, "Using continuous lexical embeddings to improve symbolic-prosody prediction in a text-to-speech front-end," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5655–5659.

[51] C. Ding, L. Xie, J. Yan, W. Zhang, and Y. Liu, "Automatic prosody prediction for chinese speech synthesis using BLSTM-RNN and embedding features," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 98–102.

[52] V. Klimkov *et al.*, "Phrase break prediction for long-form reading TTS: Exploiting text structure information," in *Proc. INTERSPEECH*, 2017, pp. 1064–1068.

[53] Y. Zheng *et al.*, "BLSTM-CRF based end-to-end prosodic boundary prediction with context sensitive embeddings in a text-to-speech front-end," in *Interspeech*, 2018, pp. 47–51.

[54] K. E. Ak, N. Xu, Z. Lin, and Y. Wang, "Incorporating reinforced adversarial learning in autoregressive image generation," in *Proc. 16th Eur. Comput. Vis. Conf.*, Springer, vol. 12366, 2020, pp. 18–34.

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

[56] Z. Tan, M. Wang, J. Xie, Y. Chen, and X. Shi, "Deep semantic role labeling with self-attention," in *Proc. 32th AAAI Conf. Artif. Intell.*, 2018, pp. 4929–4936.

[57] D. Povey, H. Hadian, P. Ghahremani, K. Li, and S. Khudanpur, "A time-restricted self-attention layer for ASR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, 2018, pp. 5874–5878.

[58] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, 2018, pp. 5884–5888.

[59] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 6706–6713.

[60] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 3171–3180.

[61] C. Lu, P. Zhang, and Y. Yan, "Self-attention based prosodic boundary prediction for Chinese speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, 2019, pp. 7035–7039.

[62] J. Li, H. Zhang, R. Liu, X. Zhang, and F. Bao, "End-to-end mongolian text-to-speech system," in *Proc. 11th Int. Symp. Chin. Spoken Lang. Process*, 2018, pp. 483–487.

[63] Z. Zhang, H. Zhao, K. Ling, J. Li, and G. Fu, "Effective subword segmentation for text comprehension," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 11, pp. 1664–1674, Nov. 2019.

[64] R. Sloan, S. S. Akhtar, B. Li, R. Shrivastava, A. Gravano, and J. Hirschberg, "Prosody prediction from syntactic, lexical, and word embedding features," in *Proc. 10th ISCA Speech Synthesis Workshop*, pp. 269–274.

[65] A. Vadapalli, P. Bhaskararao, and K. Prahallad, "Significance of word-terminal syllables for prediction of phrase breaks in text-to-speech systems for indian languages," in *Proc. 8th ISCA Workshop Speech Synthesis*, 2013, pp. 189–194.

[66] 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.

[67] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2222–2232, 2016.

[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[69] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
[70] J. Janhunen, "Mongolic languages," in *The Encyclopedia of Language & Linguistics*. Amsterdam, The Netherlands: Elsevier Scientific Publ. Co, 2006, pp. 231–234.
[71] F. Bao, G. Gao, X. Yan, and W. Wang, "Segmentation-based mongolian lvcsr approach," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process*, 2013, pp. 8136–8139.
[72] 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. Int. Conf. Neural Inf. Process*. Springer, 2019, pp. 642–651.
[73] 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*, 2018, pp. 2448–2455.
[74] 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*. Springer, 2018, pp. 217–231.
[75] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Rep.*, 2015, pp. 1–15.
[76] 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.*. Springer, 2017, pp. 99–108.
[77] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Syst. Des. Implementation*, 2016, pp. 265–283.
[78] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[79] M. D. Zeiler, "Adadelta: An adaptive learning rate method," 2012, *arXiv:1212.5701*.
[80] M. Rei, G. Crichton, and S. Pyysalo, "Attending to characters in neural sequence labeling models," in *Proc. 26th Int. Conf. Comput. Linguistics*, 2016, pp. 309–318.
[81] Y. Shao, C. Hardmeier, J. Tiedemann, and J. Nivre, "Character-based joint segmentation and POS tagging for Chinese using bidirectional RNN-CRF," in *Proc. 8th Int. Joint Conf. Natural Lang. Process.*, pp. 173–183.
[82] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proc. Conf. North American Chapter Assoc. Computat. Linguistics: Hum. Lang. Technol.*, 2016, pp. 260–270.

Rui Liu (Member, IEEE) received the B.S. degree in software engineering from the Taiyuan University of Technology, Taiyuan, China, in 2014. He is currently working toward the Ph.D. degree with the Inner Mongolia Key Laboratory of Mongolian Information Processing Technology, Inner Mongolia University, Hohhot, China. He has been with the Department of Electrical & Computer Engineering with the National University of Singapore. His research interests include prosody and acoustic modeling for speech synthesis, machine learning and natural language processing.

Berrak Sisman (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the National University of Singapore, Singapore, in 2020. She is currently working as an Assistant Professor with the Singapore University of Technology and Design (SUTD). She is also an Affiliated Researcher with the National University of Singapore. Prior to joining SUTD, she was a Postdoctoral Research Fellow with the National University of Singapore, and a Visiting Researcher with Columbia University, New York, United States. She was also a Visiting Scholar with The Centre for Speech Technology Research (CSTR), University of Edinburgh in 2019. She was attached to RIKEN Advanced Intelligence Project, Japan in 2018. Her research interests include speech information processing, machine learning and speech synthesis.

Feilong Bao received the Ph.D. degree in computer application technology from Inner Mongolia University, Hohhot, China, in 2013. He is currently an Associate Professor with the Department of Computer Science, Inner Mongolia University. His research interests include speech signal processing, natural language processing, speech synthesis, speech recognition, neural machine translation.

Jichen Yang (Senior Member, IEEE) received the Ph.D degree in communication and information system from the South China University of Technology (SCUT), Guangzhou, China, in 2010. He was a Postdoctoral Research Fellow from October 2011 to March 2016 in SCUT. Since April 2016, he has been a Postdoctoral Researcher Fellow initially with the Department of Human Language Technology, Institute for Infocomm Research (I^2R), A*STAR, Singapore and then with the Human Language Technology Lab, Department of Electrical and Computer Engineering, National University of Singapore, Singapore. His research interests mainly include anti-spoofing and forensics, speaker recognition and speech synthesis.

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

Haizhou Li (Fellow, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees from the South China University of Technology, Guangzhou, China in 1984, 1987, and 1990, respectively, all in electrical and electronic engineering. He is currently a Professor with the Department of Electrical and Computer Engineering, National University of Singapore (NUS). His research interests include automatic speech recognition, speaker and language recognition, and natural language processing. Prior to joining NUS, he taught in the University of Hong Kong (1988–1990) and South China University of Technology (1990–1994). He was a Visiting Professor with CRIN in France (1994–1995), Research Manager with the Apple-ISS Research Centre (1996–1998), Research Director in Lernout & Hauspie Asia Pacific (1999–2001), Vice President in InfoTalk Corp. Ltd. (2001–2003), and the Principal Scientist and Department Head of Human Language Technology in the Institute for Infocomm Research, Singapore (2003–2016). Dr Li served as the Editor-in-Chief of IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING (2015–2018), a Member of the Editorial Board of *Computer Speech and Language* (2012–2018). He was an Elected Member of IEEE Speech and Language Processing Technical Committee (2013–2015), the President of the International Speech Communication Association (2015–2017), the President of Asia Pacific Signal and Information Processing Association (2015–2016), and the President of Asian Federation of Natural Language Processing (2017–2018). He was the General Chair of ACL 2012, INTERSPEECH 2014 and ASRU 2019. Dr Li is a Fellow of the ISCA. He was the recipient of the National Infocomm Award 2002 and the President's Technology Award 2013 in Singapore. He was named one of the two Nokia Visiting Professors in 2009 by the Nokia Foundation, and Bremen Excellence Chair Professor in 2019.