Interspeech 2018
2-6 September 2018, Hyderabad

Improving Mongolian Phrase Break Prediction by Using Syllable and Morphological Embeddings with BiLSTM Model

Rui Liu, Feilong Bao ✉, Guanglai Gao, Hui Zhang, Yonghe Wang

College of Computer Science, Inner Mongolia University,
Inner Mongolia Key Laboratory of Mongolian Information Processing Technology,
Hohhot 010021, China

liurui_imu@163.com, csfeilong@imu.edu.cn

Abstract

In the speech synthesis systems, the phrase break (PB) prediction is the first and most important step. Recently, the state-of-the-art PB prediction systems mainly rely on word embeddings. However this method is not fully applicable to Mongolian language, because its word embeddings are inadequate trained, owing to the lack of resources. In this paper, we introduce a bidirectional Long Short Term Memory (BiLSTM) model which combined word embeddings with syllable and morphological embedding representations to provide richer and multi-view information which leverages the agglutinative property. Experimental results show the proposed method outperforms compared systems which only used the word embeddings. In addition, further analysis shows that it is quite robust to the Out-of-Vocabulary (OOV) problem owe to the refined word embedding. The proposed method achieves the state-of-the-art performance in the Mongolian PB prediction.
Index Terms: Mongolian, Syllable, Morphological, Phrase Break Prediction, BiLSTM

1. Introduction

Phrase break (PB) prediction is a crucial step in speech synthesis [1, 2]. It breaks long utterances into meaningful units of information and makes the speech more understandable. More importantly, in the context of speech synthesis, phrase breaks are often the first step for other models of prosody, such as intonation prediction and duration modeling [3, 4, 5]. Any errors made in the initial phrasing step are propagated to other downstream prosody models. Ultimately resulting in synthetic speech that is unnatural and difficult to understand.

Traditional PB prediction methods use machine learning models like Hidden Markov Models (HMMs) [6] or Conditional Random Fields (CRFs) [7, 8] which trained with large sets of labeled training data. Work in this area has traditionally involved linguistic features - for example, part-of-speech (POS), length of word etc [9, 10]. However, the linguistic features are discrete linguistic representations of words, which don't take into account the distributional behavior of words. Recent developments in neural architecture and representation learning have opened the door to models that can discover useful features automatically from the unlabelled data. With this development, word embedding [11] was proposed to learn distributed representation of word, which encodes a word as a real-valued low-dimensional vector. There are many works applying the word embedding techniques to Natural Language Processing (NLP) tasks, such as question answering, machine translation and so on [12, 13, 14]. Related ideas have been successfully applied to statistics parameter based and unit selection based speech synthesis system [15, 16]. Furthermore, for PB prediction task, some systems which do not rely on the linguistic feature are developed [17, 18, 19, 20, 21]. In [19], the authors obtain continuous-valued word embedding features that summarize the distributional characteristics of word types as surrogates of POS features. In [20], researchers utilize deep neural networks (DNNs) and recurrent neural networks (RNNs) to model PB by using word embeddings. Some further work can be found in [22] and [23]. In [22], authors obtain useful character embedding features to prediction PB in Chinese. In [23], a character-enhanced word embedding model and a multi-prototype character embedding model are proposed for Mandarin PB prediction.

All the methods mentioned has made great contributions, while they are not directly applicable to highly agglutinative languages such as Mongolian, Korean and Japanese for two reasons. First, sufficient training corpus is necessary for these methods to achieve such great performance, while the Mongolian training corpus is not very abundant; Second, such embeddings learned from these methods is unaware of the morphology of words. Mongolian is agglutinative in its morphology, words mainly contain different morphemes to determine the meaning of the word [24, 25, 26] hence increasing the vocabulary size for word embedding training and bring a considerably great challenge to train entire word-level distributed representation. Specifically, many suffixes can be in addition to word-stem to generate many new words. Its suffixes often serve as a positive signal which implies the POS of the word. It's like that the word implied by the suffix '-ly' is an adverb in English. For example:  ,  ,  ,  , . These words share the same word-stem "  " (Latin: "sandali", means: "chair"). In addition, a sequence of syllables forms a Mongolian word, and the composition of 2 or 3 characters forms a syllable. A single syllable possess a semantic meaning similar to morpheme. For instance, representation of "qihirag-tv", "qihiju", "qihitai" are constructed by the same syllables "qi" and "hi".

However, the Mongolian PB prediction research is at its initial stage compared with Chinese and English [27]. There are many works on Mongolian Text-to-Speech (TTS) which have made great contributions [28, 29], but the naturalness of synthetic speech is less than satisfactory especially without a good rhythm.

In this work, we leverage morphologic and syllable features to model Mongolian PB. We first use Bidirectional Long Short Term Memory (BiLSTM) networks to encode syllable and morphologic level information to capture the semantics of the word. Then we combine syllable, morphologic level representation and word level representation to an improved representation and

10.21437/Interspeech.2018-1706

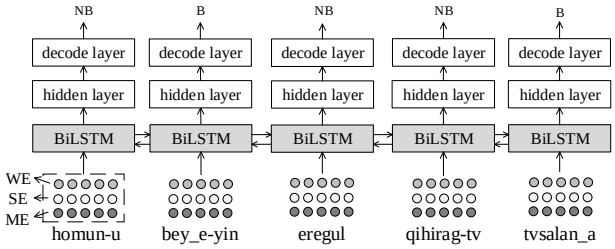

Figure 1: *Mongolian PB prediction system. The Word-, Syllable- and Morpheme-level embeddings (Figure 3) are concatenated as input; a BiLSTM produces context-dependent representations; the information is passed through a hidden layer and the output layer. The outputs are either probability distributions for softmax. (WE: word embedding; SE: Syllable embedding; ME: Morphological embedding)*

feed it in another BiLSTM to model context information of each Mongolian word and decode the corresponding right PB label.

Our experiments show the proposed approach achieves best performance. The syllable and morphologic level representation provides richer semantic information for word representation and play an important role to the neural network architecture. Moreover, this model is quite robust to the Out-of-Vocabulary (OOV) problem.

2. Proposed Model

Figure 1 shows the overall architecture of the Mongolian PB prediction model. The set of input features for each token is basically formed by three distinct components: the word embedding (WE) and two complementary information: syllable (SE) and morphological embeddings (ME). For each given token, we first obtain the word, syllable and morphological embeddings, then we concatenate these three embeddings to get a refined embedding and then fed it into a BiLSTM [30, 31] to decode the corresponding right PB label. We formulate each component of the model in the following subsections.

2.1. Input Features

2.1.1. Word Embedding

Word embedding are obtained from an unsupervised learning model that learns co-occurrence statistics of words from a Mongolian embedding corpus (Section 3.1), yielding word embeddings as distributional semantics [11]. Specifically, we use *Skip-Gram* model [11] to train the word embedding representation.

2.1.2. Syllable & Morphological Embedding

In the case of Mongolian, syllable and morpheme is a basic unit of sequence with short length compared to character level. Figure 2 highlights the difference between various embedding and the feature they capture. Syllable is the basic and smallest unit of speech. In Mongolian, words are made up of several syllables according to certain pronunciation rules. We transform each Mongolian word (Latin: 'qihirag-tv', means: health) into a sequence of syllables ('qi', 'hi', 'rag', '-tv') and then use the BiLSTM embedding method to create a syllabel-level feature representation for the word. On the other hand, morphologically, unlike many other languages, a

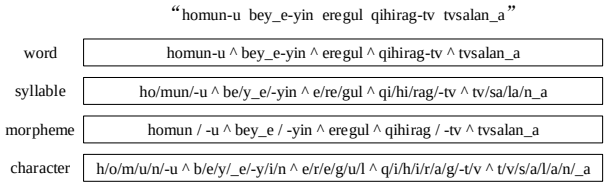

Figure 2: *Comparison of various embedding levels. In case of Mongolian, syllable and morpheme is a basic unit of sequence with short length compared to character level, which is effective to make the size of vocabulary smaller.*

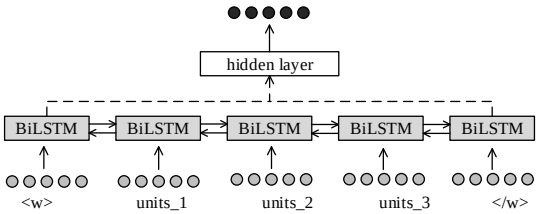

Figure 3: *Syllable or morpheme units are given as input (take three units for example); a BiLSTM produces context-dependent representations; the information is passed through a hidden layer. The proposed Syllable or Morphological embedding are generated.*

Mongolian word (Latin: 'qihirag-tv') is not just a concatenation of characters. It is constructed by the special agglutinative property. Mongolian words can be decomposed into a set of morphemes: one root and several suffixes ('qihirag', '-tv'). We use the BiLSTM embedding method, as same as the syllable embedding part, to get morphological embedding for words. The syllable and morphological embedding is used along with a word embedding to provide richer features for the word and attenuates the out-of-vocabulary (OOV) problem.

Figure 3 illustrate the BiLSTM embedding network architecture in detail. Each words in a Mongolian sentences is broken down into individual smaller unit: syllable and morpheme, these are then mapped to a sequence of embeddings (emb_1,...,emb_t), which are passed through a BiLSTM:

$$\overrightarrow{h_i^*} = LSTM(emb_i, \overrightarrow{h_{i-1}^*}) \quad (1)$$

$$\overleftarrow{h_i^*} = LSTM(emb_i, \overleftarrow{h_{i-1}^*}) \quad (2)$$

We then use the last hidden vectors from each of the LSTM components, concatenate them together, and pass the result through a separate non-linear layer.

$$h^* = [\overrightarrow{h_R^*}; \overleftarrow{h_l^*}] \qquad SE(ME) = \tanh(W_m h^*) \quad (3)$$

where W_m is a weight matrix mapping the concatenated hidden vectors from both LSTMs into a joint representation SE or ME, built from individual unit: syllables or morphemes .

We now have three alternative feature representation for each word - WE_t is an embedding learned on the word level as described in Section 2.1.1 , and SE_t or ME_t is a representation dynamically built from individual unit in the *t-th* word of the input Mongolian text. We concatenate the three vectors into a joint vector (WE^*) and use it as the new word-level representation for the PB prediction model:

$$WE^* = [WE; SE; ME] \quad (4)$$

2.2. BiLSTM Model

The joint embeddings ($W\not{E}^{*}$) which concatenate word, syllable and morphological embeddings are given as input to two LSTM components moving in opposite directions through the text, creating context-specific representations. The respective forward- and backward-conditioned representations are concatenated for each word position, resulting in representations that are conditioned on the whole sequence:

$$\overrightarrow{h_t} = LSTM(W\not{E}_t, \overrightarrow{h_{t-1}}) \quad (5)$$

$$\overleftarrow{h_t} = LSTM(W\not{E}_t, \overleftarrow{h_{t-1}}) \quad (6)$$

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \quad (7)$$

We include an extra narrow hidden layer on top of the LSTM, which allows the model to detect higher-level feature combinations, while constraining it to be small forces it to focus on more generalisable patterns:

$$d_t = \tanh(W_d h_t) \quad (8)$$

where W_d is a weight matrix between the layers, and the size of d_t is intentionally kept small.

Finally, to produce PB label predictions, we use a softmax layer. The softmax calculates a normalised probability distribution over all the possible labels of each word:

$$P(y_t = k|d_t) = \frac{e^{W_{o,k} d_t}}{\sum_{\tilde{k} \epsilon K} e^{W_{o,\tilde{k}} d_t}} \quad (9)$$

where $P(y_t = k|d_t)$ is the probability of the label of the t-th word (y_t) being k, K is the set of all possible labels, and $W_{o,k}$ is the k-th row of output weight matrix W_o. To optimise this model, we minimise categorical crossentropy, which is equivalent to minimising the negative log-probability of the correct labels:

$$E = -\sum_{T}^{t=1} \log(P(y_t|d_t)) \quad (10)$$

This approach assumes that the word-level, syllable-level and morpheme-level components learn somewhat disjoint information, and it is beneficial to give word embedding only as input to the Mongolian PB prediction system. It allows the model to take advantages of multi-view information from the syllable and morphological features in Mongolian.

3. Experiments and Results

3.1. Datasets

To verify the effectiveness of the proposed approach, we rely on a Mongolian TTS corpus which contains 59k sentences, more than 409k words, 1065k syllables and 500k morphemes. Each word in corpus was assigned to a PB label: "B" (means *"break after a word"*) or "NB" (means *"non-break"*). We divided the corpus into training and test set in a ratio of 4 to 1.

The word embedding train data were obtained from Mongolian mainstream websites. After deleting web page tags and too long sentences, its token size and vocabulary are about 200 million and 3 million respectively.

3.2. Setup

In the experiments, all digits were replaced with the arabic number "0". Any words that appeared only once in the training set were replaced by the common embedding representations of OOV words. But in syllable and morphological parts, we leave such words unchanged.

For the Mongolian datasets we used 300-dimensional pre-trained vectors as described in Section 2.1.1 and updated during training. We set both syllable and morpheme vector lengths to 150 and then do random initialization.

The LSTM layer size was set to 200 in both direction for all experiments. The size of hidden layer *d* is 50, and the combined representation *SE+ME* has the same length as the *WE*. We set learning rate as 1.0 and batch size as 64. All parameters were optimised using *AdaDelta* algorithm. The output layer's activation function is *Softmax*. At every epoch, we calculate the performance on the training set. We stop training if the effect does not increase seven epoches. The best model on training stage was then used for evaluation on the test set.

As the text in the databases has already been annotated with PB labels, a ground truth to compute the performance of our approaches is available. We report the performance of our approaches in terms of the 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 higher values indicating better performances.

3.3. Results

With this experiment, we wish to determine which type of embedding methods performs better. All Mongolian phrase break prediction systems are built at different embedding methods in our experiments.

- DNN ('WE' only): (vadapalli, 2016) only takes word embeddings as input. DNN model utilized to model the phrase break.
- LSTM ('WE' only): (vadapalli, 2016) only takes word embeddings as input in LSTM phrase break model.
- BiLSTM ('WE' only): only word embeddings are fed in BiLSTM model.
- BiLSTM + BiLSTM_CE (WE + CE): takes as input both word embeddings and character embeddings extracted from a BiLSTM embedding method. BiLSTM model takes concatenated vectors of word and character embeddings as input tokens.
- **(proposed)** BiLSTM + BiLSTM_SE + BiLSTM_ME (WE + SE + ME): We concatenate the word, syllable and morphological embeddings and use this as the new word-level representation for the BiLSTM PB prediction model.

Table 1 shows the performance of all above five systems. As can be seen, our experiment reaches its peak performance when we use proposed embeddings 'WE + SE + ME', which proves the effectiveness of proposed joint embedding. We found that incorporating the word- , syllable-, and morpheme-level information into the model improved performance on this task, indicating that capturing features regarding the above multi-view information is indeed useful in the Mongolian PB prediction system. 'WE + SE + ME' have shown competitive results compared to 'WE + CE'. A problem of 'WE + CE' is that character themselves have no semantic meanings so that model concentrate on only local syntactic features of words. while in 'WE + SE + ME', we select syllable and morpheme which have fine-granularity like a character but has its own meaning in Mongolian as a basic component of the representation of words.

Compared the 'WE + SE + ME', 'WE + CE', 'WE' systems, 'WE' system obtains the worst performance, the results indicate that adding extra information from Mongolian

Table 1: *System performance of Mongolian PB prediction with different Embedding methods.*

Embedding methods	Model	P	R	F
WE	DNN (vadapalli, 2016)	86.92	82.20	82.95
WE	LSTM (vadapalli, 2016)	87.12	85.41	86.26
WE	BiLSTM	88.73	90.24	88.58
WE + CE	BiLSTM + BiLSTM_CE	91.03	90.82	90.02
WE + SE + ME	BiLSTM + BiLSTM_SE + BiLSTM_ME	90.65	**91.49**	**90.20**

word's internal structure for word representation, to the original word embeddings, allow the model to learn useful patterns from sub-word units.

In addition, it can be seen that BiLSTM system achieve significant performance than DNN or LSTM system when using the word-level embeddings. While DNN are able to effectively capture dependencies across features, they lack the ability to capture long-term relations that are spread over time. On the other hand, PB prediction can be treated as a sequential labeling task that assigns boundary labels to words of an input sentence, BiLSTM are able to capture long-term temporal relations and thus are better for this task.

3.4. Analysis

This section provides an analysis to validate our chief claims and to elucidate some interesting aspects of proposed embedding representations for the Mongolian words.

Whether the proposed embeddings contains richer information?: Our proposed syllable and morphological embeddings increases the number of parameters in the BiLSTM PB model due to the increase in the input dimension if all other hyperparameters are held constant. To confirm that this did not have a material impact on the results, we ran an additional experiments. In the first, we trained a 'WE(BiLSTM)' system without the syllable and morphological embeddings but increased the word embedding dimension so that number of parameters was the same as in 'WE+SE+ME'. In this case, performance decreased (by 1.81% F) compared to the 'WE+SE+ME' model, indicating that solely increasing parameters does not improve performance. It also shows that our proposed embeddings provides richer representations then 'WE'.

How effective is the proposed method for alleviating OOV problem?: As described in Section 2.1.2, the proposed method can relieve the problem of out-of-vocabulary (OOV) words - if a token has never been seen before, then it's embedding representation will be enriched by syllable or morphological information instead of replaced by an unknown (UNK) tag representation. To validate the effectiveness of the proposed method in solving OOV problem, we report the performance with two open test sets by using 'WE+SE+ME' and 'WE(BiLSTM)' systems respectively in Table 2. 100 sentences were designed to form the open test set: 50 sentences without OOV words called 'Test-A', another 50 sentences with 30% OOV words called 'Test-B'. We observe that 'WE(BiLSTM)' system performs better on Test-A than Test-B. However, 'WE+SE+ME' system shows similar performance on both test sets and the performance of 'WE+SE+ME' system significantly outperforms 'WE(BiLSTM)' system. We conclude that the proposed method can attenuates the OOV problem.

Table 2: *Performance of Mongolian PB prediction on different open test datasets. Test-A does not contains any OOV words, Test-B contains 30% OOV words.*

System	TEST-A			TEST-B		
	P	R	F	P	R	F
WE(BiLSTM)	88.35	89.79	88.56	85.31	85.73	85.30
WE+SE+ME	90.52	90.89	90.14	90.46	91.12	90.09

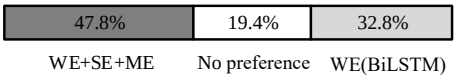

Figure 4: *The percentage preference of Subjective Evaluations.*

3.5. Preference Test

We further conducted an A/B preference test on the naturalness of the synthesised Mongolian speech. A set of 100 sentences were randomly selected from the test set and the PB labels were achieved by 'WE+SE+ME' and 'WE(BiLSTM)' systems. We carried out comparative evaluation through a DNN-based Mongolian TTS system [29]. A group of 10 subjects were asked to choose which one was better in terms of the naturalness of synthesis speech. The percentage preference is shown in Figure 4. We can clearly see that the proposed method can achieve better naturalness of synthesized Mongolian speech as compared with Unimproved word embedding.

4. Conclusion

In this paper, we investigated syllable and morphological-level model components for Mongolian PB prediction system, which allows the system to learn useful features from different viewpoint. In addition to a BiLSTM operating over entire words, a separate BiLSTM is used to construct additional representations from different individual smaller units: syllable, morpheme. We concatenated these three embeddings into a new representation which proved absorb richer and multi-view information than the original word embedding for Mongolian. In addition, further analysis shows that it is quite robust to the OOV problem owe to the refined word embedding. This work can also inspire other agglutinative language research.

5. Acknowledgements

This research was supports by the China national natural science foundation (No.61563040, No.61773224), Inner Mongolian nature science foundation (No.2016ZD06) and the Enhancing Comprehensive Strength Foundation of Inner Mongolia University (No.10000-16010109-23).

6. References

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

[2] H. Kim, T. Yoon, J. Cole, and M. Hasegawa-Johnson, "Acoustic differentiation of l- and l-l% in switchboard and radio news speech," in *Proceedings of Speech Prosody*, 2006.

[3] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, and J. Raiman, "Deep voice: Real-time neural text-to-speech," 2017.

[4] K. Prahallad, E. V. Raghavendra, and A. W. Black, "Learning speaker-specific phrase breaks for text-to-speech systems," in *Proceedings of 7th ISCA Speech Synthesis Workshop (SSW7), Kyoto, Japan*, 2010, pp. 148C–153.

[5] A. Parlikar and A. W. Black, "Minimum error rate training for phrasing in speech synthesis," in *Procedings of 8th ISCA Speech Synthesis Workshop (SSW8), Barcelona, Spain*, 2013, pp. 13C–16.

[6] R.Rabiner, Lawrence, and B.-H. Juang, "An introduction to hidden markov models," *IEEE ASSP Magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[7] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[8] Y. Qian, Z. Wu, X. Ma, and F. Soong, "Automatic prosody prediction and detection with conditional random field (crf) models," in *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*. IEEE, 2010, pp. 135–138.

[9] A. Parlikar and A. W. Black, "A grammar based approach to style specific phrase prediction," in *Proceedings of Interspeech, Florence, Italy*, 2011, pp. 2149C–2152.

[10] P. Taylor and A. W. Black, "Assigning phrase breaks from partof-speech sequences," in *Computer Speech and Language*, 1998, pp. 99C–117.

[11] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[12] A. Bordes, S. Chopra, and J. Weston, "Question answering with subgraph embeddings," *arXiv preprint arXiv:1406.3676*, 2014.

[13] C. Xiong, V. Zhong, and R. Socher, "Dynamic coattention networks for question answering," 2017.

[14] J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. M. Schwartz, and J. Makhoul, "Fast and robust neural network joint models for statistical machine translation," in *ACL*, 2014, pp. 1370–1380.

[15] P. Wang, Y. Qian, F. K. Soong, L. He, and H. Zhao, "Word embedding for recurrent neural network based tts synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 4879–4883.

[16] T. Merritt, J. Yamagishi, Z. Wu, O. Watts, and S. King, "Deep neural network context embeddings for model selection in rich-context hmm synthesis," 2015.

[17] A. Vadapalli and K. Prahallad, "Learning continuous-valued word representations for phrase break prediction," in *15th Annual Conference of the International Speech Communication Association (INTERSPEECH), Singapore*, 2014, pp. 41C–45.

[18] O. Watts, S. Gangireddy, J. Yamagishi, S. King, S. Renals, A. Stan, and M. Giurgiu, "Neural net word representations for phrase-break prediction without a part of speech tagger," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Florence*, 2014, pp. 2599C–2603.

[19] O. Watts, J. Yamagishi, and S. King, "Unsupervised continuous-valued word features for phrase-break prediction without a part-of-speech tagger," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

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

[21] 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," 2016.

[22] 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 *Automatic Speech Recognition and Understanding*, 2016, pp. 98–102.

[23] Y. Zheng, Y. Li, Z. Wen, X. Ding, and J. Tao, "Improving prosodic boundaries prediction for mandarin speech synthesis by using enhanced embedding feature and model fusion approach," in *INTERSPEECH*, 2016, pp. 3201–3205.

[24] G. Qing, "Mongolian syntax," *Mongolia people publishing house, Hohhot*, pp. 77–133, 1991.

[25] Temusurvn and Otegen, "Mongolian orthography dictionary," *Inner Mongolia People Publishing House, Hohhot*, pp. 77–133, 1999.

[26] F. Bao, G. Gao, X. Yan, and W. Wang, "Segmentation-based mongolian lvcsr approach," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 8136–8139.

[27] R. Liu, F. Bao, G. Gao, and H. Zhang, "Approach to predicition mongolian prosody phrase based on crf model," in *Proceedings of the National Conference on Man-Machine Speech Communication*, 2015.

[28] J. Zhao, G. Gao, F. D. Bao, and P. Mermelstein, "Research on hmm-based mongolian speech synthesis," *Computer Science*, no. 41, pp. 80–104, 2014.

[29] R. Liu, F. Bao, G. Gao, and Y. Wang, "Mongolian text-to-speech system based on deep neural network," *Man-Machine Speech Communication. NCMMSC 2017. Communications in Computer and Information Science, Springer*, vol. 807, 2018.

[30] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," vol. 28, no. 10, 2016, pp. 2222–2232.

[31] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," vol. 45, no. 11, 2002, pp. 2673–2681.