Building Mongolian TTS Front-End with Encoder-Decoder Model by Using Bridge Method and Multi-view Features

Rui Liu, Feilong Bao(✉), and Guanglai Gao

College of Computer Science, Inner Mongolia Key Laboratory of Mongolian Information Processing Technology, Inner Mongolia University, Hohhot 010021, China
liurui_imu@163.com, csfeilong@imu.edu.cn

Abstract. In the context of text-to-speech systems (TTS), a front-end is a critical step for extracting linguistic features from given input text. In this paper, we propose a Mongolian TTS front-end which joint training Grapheme-to-Phoneme conversion (G2P) and phrase break prediction (PB). We use a bidirectional long short-term memory (LSTM) network as the encoder side, and build two decoders for G2P and PB that share the same encoder. Meanwhile, we put the source input features and encoder hidden states together into the Decoder, aim to shorten the distance between the source and target sequence and learn the alignment information better. More importantly, to obtain a robust representation for Mongolian words, which are agglutinative in nature and lacks sufficient training corpus, we design specific multi-view input features for it. Our subjective and objective experiments have demonstrated the effectiveness of this proposal.

Keywords: Text-to-speech · Front-end · Phrase break · Grapheme-to-Phoneme · Mongolian

1 Introduction

A text-to-speech system (TTS) consists of two components. One is a front-end, which takes a given text as its input and returns a phoneme sequence annotated with prosody information of the text. The other is a back-end, which converts the output of a front-end into speech. For a front-end, the vital part is Grapheme-to-Phoneme conversion (G2P) and phrase break prediction (PB), as the intelligibility and naturalness depend on their correctness. To estimate the correct phoneme sequence of a sentence, we need to recognize words and determine their phoneme sequences. Furthermore, to split an utterance into prosodic units which can be easily understood by people, we need to identify the prosody phrase boundaries of sentence.

For English and Mandarin TTS, there have been attempts at solving this problem. G2P can be treated as a sequence prediction problem. A typical approach to G2P involves using joint sequence model [1]. Recently, there has been

T. Gedeon et al. (Eds.): ICONIP 2019, CCIS 1143, pp. 642–651, 2019.
https://doi.org/10.1007/978-3-030-36802-9_68

some work using long short-term memory (LSTM) networks and encoder-decoder approach for G2P problem [3,8]. PB can be treated as a sequence labeling task. Typically PB methods usually use maximum entropy Markov models (MEMMs) [9], conditional random fields (CRFs) [10], and recurrent neural networks (RNNs) [11]. But in these above works, G2P and PB are usually processed separately. However, for Mongolian TTS, the research on G2P or PB is at its initial stage. There are many works which have made great contributions [12], but the performance is less than satisfactory.

In this work, we investigate how G2P and PB can be jointly modeled while benefiting from the strong modeling capacity of the encoder-decoder models and built a Mongolian TTS front-end. We use a bidirectional recurrent neural network (RNN) as the encoder side. For decoder side, we build two decoders for G2P and PB based on unidirectional RNN separately. These two decoders share the same encoder parameter. Learning from the attention mechanism in encoder-decoder model [13], we further utilizes attention to the G2P Decoder and alignment-based PB Decoder. Such attention provides additional information to the G2P and PB. To shorten the distance between the source and target sequence and learn the alignment information better, we use bridge method inspired by the machine translation [7], in which we put the source input features and encoder hidden states together into the two Decoders.

In addition, all these methods mentioned for high resource language taking the word embeddings as input. It is hard to work with scripts of Mongolian languages, in which the necessary linguistic resources are not readily available and lack sufficient training corpus. Thus we take a multi-view approach to learning word-level representations as the encoder input for Mongolian, leverages the agglutinative property. To obtain a robust representation for Mongolian word, we first identify the sequence of morpheme (stem&suffix) automatically and encode them to a morphological representation, which captures the morphological information of the word. Then we extract acoustic features of each word. At last, the morphological vector, acoustic features and word embeddings are comprised together to a multi-view input features for each Mongolian word.

Objective experiment results show our proposed model achieves better performance than the conventional model. Subjective experiment results further show that this method is beneficial to improve the naturalness and the expression of the Mongolian synthesized speech.

2 Proposed Model

2.1 Joint Encoder-Decoder Model

Our joint Encoder-Decoder model includes one Encoder, which reads in the input Mongolian word sequence $x = (x_1, x_2, \ldots, x_T)$, and two Decoders, which generates Mongolian phoneme sequence $y = (p_1, p_2, \ldots, p_{T'})$ and the corresponding PB labels $y = (y_1, y_2, \ldots, y_T)$ simultaneously. The model structure is illustrated in Fig. 1.

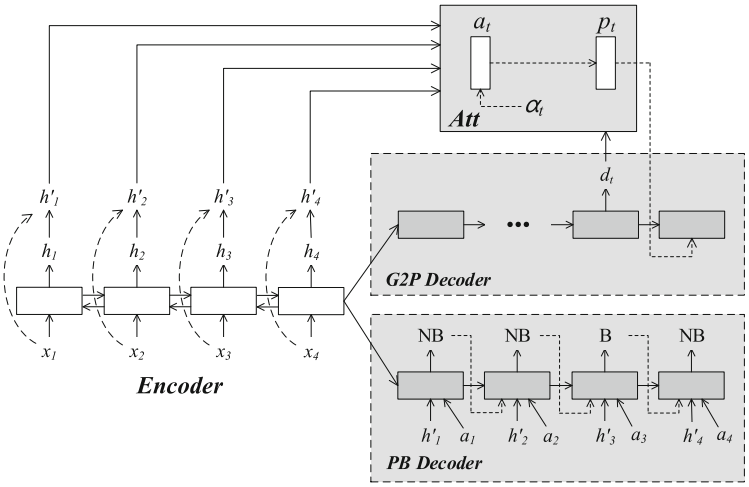

Fig. 1. Joint encoder-decoder model for G2P and PB.

Encoder Side Using Bridge Method. As shown in Fig. 1, we use bidirectional LSTM (BiLSTM) as basic component for the common encoder side on two decoders.

In conventional method, the source sequence $x = (x_1, x_2, \ldots, x_T)$ only provides source input information in generating the hidden state h_i once time, and then it is no longer used. To make the relation of source and target sequence more closely, we move source input features one step closer to the target output as illustrates in Fig. 1. After generating the final encoder hidden state h_i at each time step i , we concatenate h_i with its corresponding input features x_i as the bridge hidden vector $h_i^{'}$: $h_i^{'} = [h_i, x_i]$. In this bridge method, word-level input features as part of the encoder hidden state to form the attention information and consequently have a positive effect in Decoder stage. The first and last bridge states ($h_1^{'}$ and $h_T^{'}$) carries rich information of the entire source input. We use a linear combination of $h_1^{'}$ and $h_T^{'}$, with parameters learned during training, as initial hidden state for two Decoders.

Decoder Side for G2P and PB. The Decoders are modeled as unidirectional LSTM. In PB, at decode stage, the decoder state s_i is calculated as a function of the previous decoder state s_{i-1}, the previous predicted label y_{i-1}, the aligned encoder bridge hidden state $h_i^{'}$, and the attention vector a_i:

$$s_i = f(s_{i-1}, y_{i-1}, h_i^{'}, a_i) \tag{1}$$

where the attention vector a_i is computed as a weight sum of the bridge encoder states $h^{'} = (h_1^{'}, \ldots, h_T^{'})$:

$$e_{i,k} = g(s_{i-1}, h_k^{'}), \quad \alpha_{i,j} = \frac{\exp(e_{i,j})}{\sum_{k=1}^{T} \exp(e_{i,k})} \tag{2}$$

$$a_i = \sum_{j=1}^{T} \alpha_{i,j} h_j^{'} \tag{3}$$

g is a feed-forward neural network. It should be note that the encoder state h_i is the explicit aligned input at each decoding step in *PB Decoder*. The attention vector a_i provides abundant context information to the *PB Decoder*.

G2P Decoder shares the same encoder hidden states with PB Decoder. The G2P Decoder predicts each phoneme p_t given the attention vector a_i and all of the previously predicted phonemes $p_1, p_2, \ldots, p_{t-1}$ in the following way:

$$s_t = g(\tilde{p}_{t-1}, s_{t-1}, a_i) \tag{4}$$

$$p(p_t|p_{<t}, x) = \text{softmax}(W_s d_t + b_s) \tag{5}$$

where s_{t-1} is the hidden state of the decoder LSTM and $\tilde{p}_{t-1}$ is the vector obtained by projecting the one hot vector corresponding to p_{t-1} using a phoneme embedding matrix E. The embedding matrix E is jointly learned with other parameters of the model. Similar to the PB Decoder, we use an attention vector a_i, a linear combination of all of the encoder bridge hidden states, at every decoder time step. But in G2P decoder, the alignment input from source word sequence is unknown. Therefore the attention vector a_i can be seen as a soft alignment between the source word sequence and target phoneme sequence. It allows the model to attend to different encoder states when decoding each output label. Motivated by [13], attention vector a_i at time step i is given by:

$$a_t = \sum_{i=1}^{T} \alpha_{i,t} h_i^{'} \tag{6}$$

and

$$\alpha_t = \text{softmax}(u_t) \tag{7}$$

$$u_{i,t} = v^T \tanh(W_1 h_i + W_2 d_t + b_a) \tag{8}$$

where the vector v, b_a and the matrices W_1, W_2 are parameters learned jointly with the rest of the model. The score $\alpha_{i,t}$ is a weight that represents the importance of the hidden bridge encoder state $h_i^{'}$ in generating the phoneme p_t.

Hence, Eq. (5) can be rewritten as:

$$p(p_t|p_{<t}, x) = \text{softmax}(W_s[a_t; d_t] + b_s) \tag{9}$$

2.2 Multi-view Input Features

To obtain more robust embedding representation for Mongolian words[1], we propose two novel word-level representations, which include **Morphological vector** and **Acoustic vector**, along with **Word embedding** as input features.

Morphological Vector (m_i). We use multi-layer stacked BiLSTM network to automatically learn the mapping from the sequences of morpheme, stem and suffix, to a word-level vector.

This operation is consistent with our previous work [5,6]. Each words in a Mongolian sentences is broken down into individual smaller unit: stem and suffix, these are then mapped to a sequence of embeddings, which are passed through multi-layer (stacked) BiLSTMs to obtain the morphological vector.

Acoustic Vector. For predicting the phrase pause boundaries in speech, we can incorporate acoustic features. Here we explore three types of features widely used in TTS [14]: **Duration parameter**, **Spectrum parameter** and **Excitation parameter**. All acoustic features are extracted according to the word boundaries obtain by force alignment with respect to the reference transcriptions by using the Speech Signal Processing Toolkit (SPTK)[2].

- **Duration parameter (t_i):** Word duration are strong cues to prosodic phrase boundaries. The word duration features d_i is computed as the actual word duration divided by the mean duration of the word, clipped to a maximum value of 6. The sample mean is used for frequent words (count $\geq$15). For infrequent words we estimate the mean as the sum over the sample means for their phoneme sequences.
- **Spectrum parameter (s_i):** The spectrum vector s_i consists of Mel-generalized cepstral coefficients (MGC) vector including the zeroth coefficients.
- **Excitation parameter (e_i):** The excitation vector e_i consists of log fundamental frequency (logF0).

Word Embedding (we_i). We use Skip-Gram [15] model to train the word embedding representation we_i for Mongolian.

Therefore, our overall multi-view input feature vectors x_i designed for Mongolian is the concatenation of m_i, t_i, s_i, e_i and we_i in various combinations.

3 Experiments

3.1 Data

Mongolian Orthography dictionary as an experimental dataset in G2P. The whole dictionary, which contains about 40k items, is partitioned into training, validation and test set according to 8:1:1.

[1] There are two writing systems of Mongolian: Cyrillic Mongolian and traditional Mongolian. This paper only studies traditional Mongolian.

[2] http://sp-tk.sourceforge.net/.

Table 1. Ablation test with bridge method and multi-view features. **Bold** indicates the best model.

Bridge	*we*	*m*	Acoustic vector			WER (%)	F (%)
			t	*s*	*e*		
No	Yes	No	No	No	No	24.53	83.26
No	Yes	Yes	No	No	No	24.01	84.13
No	Yes	No	Yes	No	No	23.85	83.98
No	Yes	No	No	Yes	No	23.79	84.02
No	Yes	No	No	No	Yes	23.80	84.13
No	Yes	Yes	Yes	No	No	22.98	84.52
No	Yes	Yes	No	Yes	No	22.93	84.56
No	Yes	Yes	No	No	Yes	22.91	84.61
No	Yes	No	Yes	Yes	No	23.71	84.65
No	Yes	No	Yes	No	Yes	23.67	84.72
No	Yes	No	No	Yes	Yes	23.59	84.76
No	Yes	Yes	Yes	Yes	No	22.58	85.12
No	Yes	Yes	Yes	No	Yes	22.51	85.19
No	Yes	Yes	No	Yes	Yes	22.47	85.24
No	Yes	No	Yes	Yes	Yes	23.35	83.54
No	Yes	Yes	Yes	Yes	Yes	21.27	86.13
Yes	Yes	No	No	No	No	22.37	85.32
Yes	Yes	Yes	No	No	No	21.98	85.95
Yes	Yes	No	Yes	No	No	21.76	85.89
Yes	Yes	No	No	Yes	No	21.70	85.86
Yes	Yes	No	No	No	Yes	21.63	85.90
Yes	Yes	Yes	Yes	No	No	20.62	86.35
Yes	Yes	Yes	No	Yes	No	20.35	86.39
Yes	Yes	Yes	No	No	Yes	20.29	86.41
Yes	Yes	No	Yes	Yes	No	21.11	86.32
Yes	Yes	No	Yes	No	Yes	21.03	86.43
Yes	Yes	No	No	Yes	Yes	21.15	86.45
Yes	Yes	Yes	Yes	Yes	No	20.31	87.15
Yes	Yes	Yes	Yes	No	Yes	20.25	87.21
Yes	Yes	Yes	No	Yes	Yes	20.29	87.19
Yes	Yes	No	Yes	Yes	Yes	20.92	87.33
Yes	Yes	Yes	Yes	Yes	Yes	**19.46**	**88.50**

For evaluating the effectiveness of the PB model, we rely on a corpus corresponding to the Mongolian TTS database recorded by a professional native Mongolian female speaker. The corpus contains 59k sentences and more than 409k words. The speech data from the Mongolian TTS database, which was segmented according to the word boundaries obtained by forced alignment with respect to the reference transcriptions, are used to extract word-level acoustic feature.

The word embedding train data were crawled from mainstream websites in Mongolia. After cleaning web page tags and filtering longer sentences, its token size and vocabulary are about 200 million and 3 million respectively.

Table 2. Comparison with previous models and our independent model. Joint training model results on G2P and PB.

#	Model	WER (%)	F (%)
G2P	Joint sequence [1]	22.53	–
	BiLSTM [3]	22.01	–
	Encoder-decoder [8]	20.32	–
	Independent model (best model)	**19.46**	–
PB	CRF [10]	–	83.21
	LSTM [11]	–	85.16
	Independent model (best model)	–	**88.50**
G2P & PB	**Joint model**	**19.32**	**89.15**

3.2 Experiments Settings

For all experiments,we select the number of units in LSTM cell as 128. The default forget gate bias is set to 1. To prevent overfitting we use scheduled sampling with a linear decay on the decoder side. Parameters were optimised using Adam with default learning rate 1.0 and sentences were grouped into batch of size 64. We take the loss sum of two tasks as a joint loss. Performance on the training set was measured at every epoch and training was stopped if performance had not improved for 10 epoches.

In Morphological vector part, we have 3 layers stacked LSTMs, each with 512 units. We use an initial learning rate of 0.001 and reduce this learning rate by a multiplicative factor of 0.8. We use minibatch stochastic gradient descent (SGD) together with Adam using a minibatch size of 256. The embeddings for morpheme were set to 100. The network was trained 500 epochs.

Speech signals are sampled at 16 kHz, windowed by a 25-ms window shifted every 5-ms. The acoustic features vector contain 35 MGC, logF0 and word duration, totally 37 dimensions $(35 + 1 + 1 = 37)$. The word embeddings were initialised with pretrained vectors as illustrated in Sect. 2.2 and then fine-tuned

during model training. For the Mongolian datasets we used 100-dimensional word embeddings.

3.3 Evaluation Metrics

In PB, results are reported on the test set in terms of the F-score (F) which is defined as the harmonic mean of the *Precision* and *Recall*. In G2P, We report *word error rate (WER)*.

3.4 Independent Model Results

Table 1 report the results on the independent Encoder-Decoder model for G2P and PB. We first using the word embedding input features ($x_i = we_i$) to establish a strong baseline, on top of which we can add various features: m_i, t_i, s_i, e_i and use bridge method.

We note that adding any combination of features (individually or in sets) improves performance over the baseline. The proposed multi-view input features provide richer information than word embedding in Mongolian. Furthermore, morphological vector (m_i) and acoustic vector (t_i, s_i, e_i) both play a very good role in the performance for the two tasks. Specifically, in G2P, the introduction of acoustic vector contributes significantly in all metrics, compared to morphological vector. In PB, the contributions of acoustic vector and morphological vector are almost equally. We believe that the interesting phenomenon may owing to the nature of the specific tasks. For G2P, which gives every word a corresponding phoneme list. It is important to consider the pronunciation information when decoding the current word. However, the internal structure and pronunciation information is what we need to focus on in PB task for agglutinative language [2]. In Mongolian, stem and suffix serve to discriminate words based on syntactic meaning, and that these sub-word units can be used to model PB. The "$we_i + m_i + t_i + s_i + e_i$" model that uses all features has the remarkable performance over the we_i baseline.

We also notice that using bridge method, that concatenate the input features with hidden states, in conjunction with multi-view input features yields a significant improvement in F-Score. Lastly, we refer to the "$bridge\ method + we_i + m_i + t_i + s_i + e_i$" model as our "best model".

3.5 Joint Model Results

Table 2 shows our joint model performance on G2P and PB comparing to previous methods and the independent model under same Mongolian data. As shown in this table, the joint model using encoder-decoder architecture with bridge and multi-view features achieves the best performance. This indicates that G2P and PB can be closely linked together through joint training to achieve a joint performance improvement.

3.6 Front-End Based on Joint Model

To evaluate the naturalness of the synthesized Mongolian speech from the proposed front-end component, a subjective AB preference test was conducted.

In this evaluation, a set of 20 sentences were randomly selected from test set and the synthesised speech was generated through the DNN-based Mongolian TTS system [4] based on proposed front-end and the original front-end [4]. 20 subjects were asked to choose which one was better of paired synthesis speech. Figure 2 shows the subjective evaluation results. It can be seen from the figure that the proposed front-end component, with joint model for G2P and PB, obtain the higher quality Mongolian speech.

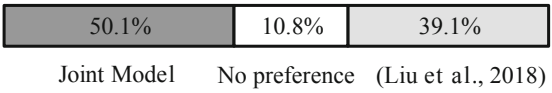

Fig. 2. Subjective evaluation results of the DNN-based Mongolian TTS by using proposed front-end and original front-end in [4].

4 Conclusions

In this paper, we proposed a joint Encoder-Decoder model by using bridge method and multi-view feature for joint G2P and PB and built a Mongolian TTS front-end in a unified framework. The bridge method seeks to shorten the distance between source and target word-level features from the word sequence. In view of the limitation of necessary linguistic resources are not readily available in Mongolian, the word-level multi-view features combines three parts of information (morphological, acoustic and word) to form a robust representation for Mongolian word. In addition, joint training can better utilize the close connection between G2P and PB. The proposed model achieves better performance compared to conventional front-end component in Mongolian TTS.

Acknowledgments. This research was supports by the National Natural Science Foundation of China (No.61563040, No.61773224), Natural Science Foundation of Inner Mongolian (No.2018MS06006, No.2016ZD06).

References

1. Bisani, M., Ney, H.: Joint-sequence models for grapheme-to-phoneme conversion. Speech Commun. **50**, 434–51 (2008)
2. Vadapalli, A., Bhaskararao, P., Prahallad, K.: Significance of word-terminal syllables for prediction of phrase breaks in Text-to-Speech systems for Indian languages. In: 8th ISCA Tutorial and Research Workshop on Speech Synthesis (2013)

3. Rao, K., Peng, F., Sak, H., Beaufays, F.: Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. In: 40th IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4225–4229. IEEE Press (2015)
4. Liu, R., Bao, F., Gao, G., Wang, Y.: Mongolian text-to-speech system based on deep neural network. In: Tao, J., Zheng, T.F., Bao, C., Wang, D., Li, Y. (eds.) NCMMSC 2017. CCIS, vol. 807, pp. 99–108. Springer, Singapore (2018). https://doi.org/10.1007/978-981-10-8111-8_10
5. Liu, R., Bao, F., Gao, G.: A LSTM approach with sub-word embeddings for mongolian phrase break prediction. In: 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA, 20–26 August 2018, pp. 2448–2455 (2018)
6. Liu, R., Bao, F., Gao, G., Wang, Y.: Improving mongolian phrase break prediction by using syllable and morphological embeddings with BiLSTM model. In: 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2–6 September 2018, pp. 57–61 (2018)
7. Kuang, S., Li, J., Branco, A., Luo, W., Xiong, D.: Attention focusing for neural machine translation by bridging source and target embeddings. In: 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018 (2018)
8. Toshniwal, S., Livescu, K.: Jointly learning to align and convert graphemes to phonemes with neural attention models. In: Spoken Language Technology Workshop. IEEE Press (2017)
9. Yu, Z., Lee, G.G., Kim, B.: Using multiple linguistic features for Mandarin phrase break prediction in maximum-entropy classification framework. In: 5th INTERSPEECH 2004 - ICSLP, International Conference on Spoken Language Processing, Jeju Island, Korea, October 2004. DBLP (2004)
10. Qian, Y., Wu, Z., Ma, X., Soong, F.: Automatic prosody prediction and detection with Conditional Random Field (CRF) models. In: 7th International Symposium on Chinese Spoken Language Processing, pp. 135–138. IEEE Press (2010)
11. Vadapalli, A., Gangashetty, S.V.: An investigation of recurrent neural network architectures using word embeddings for phrase break prediction. In: 17th INTERSPEECH, pp. 2308–2312 (2016)
12. Liu, R., Bao, F., Gao, G., Wang, W.: Mongolian prosodic phrase prediction using suffix segmentation. In: 21th International Conference on Asian Language Processing, pp. 250–253. IEEE Press (2017)
13. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Proceedings of ICLR 2015, pp. 1–15 (2014)
14. Wang, X., Lorenzo-Trueba, J., Takaki, S., Juvela, L., Yamagishi, J.: A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis. In: 43th ICASSP (2018)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)