IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 26, 2019, accepted September 3, 2019, date of publication September 9, 2019, date of current version October 8, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2940125

Lhasa-Tibetan Speech Synthesis Using End-to-End Model

YUE ZHAO, PANHUA HU, XIAONA XU, LICHENG WU, AND XIALI LI
School of Information Engineering, Minzu University of China, Beijing 100081, China

Corresponding author: Xiali Li (xiaer_li@163.com)

This work was supported in part by the Ministry of Education Research in the Humanities and Social Sciences Planning Fund under Grant 15YJAZH120, in part by the National Natural Science Foundation under Grant 61602539 and Grant 61873291, and in part by the MUC 111 Project and the First-Class University and First-Rate Discipline Funds of Minzu University of China.

ABSTRACT With the development of deep learning technology, speech synthesis based on deep neural networks has gradually become the mainstream method in the field of speech synthesis. In this paper, we explored the Tacotron2 model for Lhasa-Tibetan dialect speech synthesis by constructing a feature prediction network with a seq2seq structure which maps the character vector to Mel spectrum, and combining with the WaveNet model trained in a semi-supervised way to synthesize the Mel spectrum into a time domain waveform. The model avoids processing front-end text analysis that requires extensive prior knowledge in Lhasa-Tibetan dialect and reduces the need of a large amount of transcribed speech data. Experimental results show that the proposed method is effective and has higher clarity and naturalness than other related synthesis models for Lhasa-Tibetan dialect.

INDEX TERMS End-to-end model, Lhasa-Tibetan speech synthesis, WaveNet model, sequence-to-sequence structure.

I. INTRODUCTION

In recent years, with the rise of artificial intelligence, deep learning has been applied to fields including object recognition, signal processing, social network and others [1]–[6]. Recent applications of deep learning to speech synthesis (also known as Text-to-Speech (TTS)) have also advanced the performance contrasted to the exsiting hidden Markov model-based one. However, building speech synthesis system often requires extensive domain expertise, which is the bottleneck preventing rapid development of low-resource language speech synthesis system.

The traditional speech synthesis system usually includes two modules: front-end and back-end [7]. The front-end module mainly analyses the input text and extracts the linguistic information needed by the back-end module. For Tibetan synthesis system, the front-end module generally includes sub-modules such as text regularization, word segmentation, part-of-speech prediction, polysyllabic disambiguation and prosody prediction. The back-end module generates speech waveform by certain methods according to the results of front-end analysis. The back-end system is generally divided into speech synthesis based on statistical parameter modeling (or parameter synthesis) [8] and speech synthesis based on unit selection and waveform splicing (or splicing synthesis) [9]–[11].

Traditional speech synthesis systems are relatively complex systems. For example, front-end systems need a strong linguistic background, and the linguistic knowledge of different languages is also significantly different, so they need the support of experts in domain. The parameter system in the back-end module needs to have a certain understanding of the voice mechanism. Because of the information loss in modeling the conventional parameter system, the further improvement of the synthetic performance is limited. The splicing system of the same back-end system requires higher quality for speech database, and requires manual intervention to formulate a lot of selection rules and parameters. All these promote the emergence of end-to-end speech synthesis technology.

The end-to-end synthesis system inputs text or phonetic characters, and the system outputs speech waveform directly. End-to-end system reduces the requirement of linguistic knowledge, and can be easily replicated in different languages. End-to-end speech synthesis system also shows a strong and rich pronunciation style and prosodic expression.

The associate editor coordinating the review of this manuscript and approving it for publication was Guan Gui.

Tibetan is one of minority languages in China. Compared with some large languages such as Chinese and English, Tibetan lacks of linguistic knowledge and corpus data. At present, the works on Tibetan speech synthesis system mainly explored waveform splicing technology [12], [13], statistical parameter speech synthesis technology based on Hidden Markov model [14]–[18], and deep learning speech synthesis technology [19], [20]. Considering that the waveform splicing technology requires high storage capacity and the synthesis effect of statistical parameter technology and deep learning is unstable and low nature degree, in this paper, an end-to-end Lhasa-Tibetan speech synthesis method is presented, in which Sequence-to-Sequence Architecture [21]–[23] with attention mechanism is used as feature prediction network, then followed with a WaveNet network which is trained in a semi-supervised way. Differently from the works [19], [20] which used phonemes and Tibetan letters as the input of model respectively, our work adopts Wylie transliteration scheme to convert Tibetan characters into English letters as the input text.

Our main contributes are: (i) presenting to train WaveNet vocoder in a semi-supervised way, which reduces the requirement of a large amount of transcribed samples. (ii) presenting that the synthesis units are English letters converted from Tibetan characters by Wylie transliteration scheme, which avoid using linguistic knowledge to construct lexicon, subword set and phones, as well as analyzing the language-specific text.

II. SPEECH SYNTHESIS MODEL STRUCTURE

The end-to-end speech synthesis model is divided into two parts. The first part is a feature prediction network based on seq2seq, which introduces attention mechanism, and is used to predict the frame sequence of Mel spectrogram from the input character sequence. The second part is to learn the vocoder that generates time-domain waveform samples based on the predicted Mel spectrogram of frame sequence.

A. ENCODER-DECODER FRAMEWORK

To solve the problem of sequence-to-sequence, an encoder-decoder framework has been proposed in [24]. Sequence-to-sequence problem mainly refers to the mapping problem from sequence to sequence. Sequence can be understood as a sequence of strings. When we give a sequence of strings, we hope to get another sequence of strings corresponding to it. This task is called sequence-to-sequence (seq2seq). The encoding is to transform the input sequence into a fixed length vector; decoding is to convert the fixed vector generated before into the output sequence. In this paper, the framework of encoder-decoder is applied to Lhasa-Tibetan speech synthesis. Its network structure is shown in Fig. 1.

For speech synthesis, the goal is to generate the target speech signal through the encoder-decoder framework given the input sentence. For the input sentence, it is a text sequence, and the output signal is a speech sequence, which is represented by *Input* and *Output* respectively. It can be written

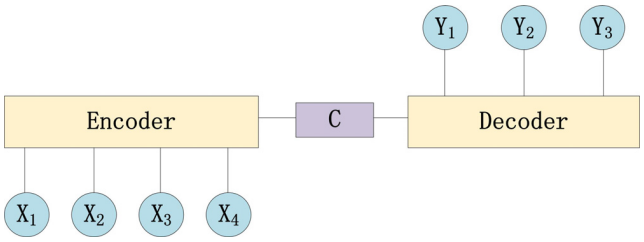


FIGURE 1. Encoder-decoder.

in the form as (1) and (2).

$$Input = (x_1, x_2, \ldots, x_m) \tag{1}$$

$$Output = (y_1, y_2, \ldots, y_n) \tag{2}$$

Encoder encodes input sentences and transforms them into intermediate semantics through non-linear transformation.

$$C = f(x_1, x_2, \ldots, x_m) \tag{3}$$

For the decoder, its task is to output the current speech signal according to the intermediate semantics of the input sentence and the historical information that has been generated before:

$$y_i = g(C, y_1, y_2, \ldots, y_{i-1}) \tag{4}$$

The advantage of this encoder-decoder architecture is that its input sequence length does not need to be consistent with the output sequence length.

B. ATTENTION MECHANISM

Although the encoder-decoder model solves the problem of sequence-to-sequence, it also has great limitations. The encoder-decoder model compresses the information of the whole sequence into a fixed length vector C. There are two drawbacks in this way. First, for a long sequence, the fixed-length semantic vector C can not fully represent the information of the whole sequence. Second, the information carried by the early input will be covered by the latter information input, which will lose more information and is not conducive to decoding.

To solve this problem, attention model is proposed. This model produces an "attention range" when it produces output, which means that the next output should focus on which parts of the input sequence, and then produce the next output according to the region of interest, so to repeat. The schematic diagram of the model is as Fig. 2.

After introducing the attention mechanism, the formula for calculating the semantic encoding C_t is as equation (5).

$$C_t = \sum_{s=1}^{S} \alpha_t^{(s)} h_e^{(s)} \tag{5}$$

where $h_e^{(s)}$ is the feature vector of the encoder output and $\alpha_t^{(s)}$ is the weight. Weight vectors are differen in each prediction of the decoder at time t. The $\alpha_t^{(s)}$ calculation method is as (6).

$$\alpha_t^{(s)} = \frac{\exp(Score(h_{dt}, h_e^{(s)}))}{\sum_{s=1}^{S} \exp(Score(h_{dt}, h_e^{(s')}))} \tag{6}$$

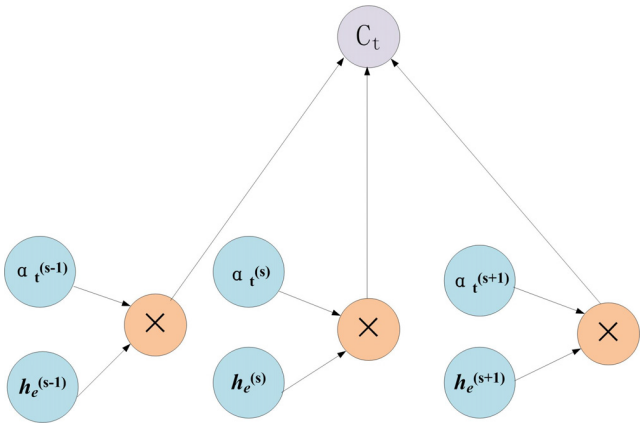


FIGURE 2. Attention model.

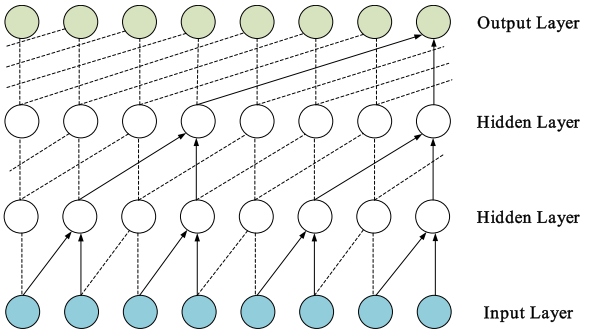


FIGURE 3. Dilated causal convolution.

It represents the matching degree between the s^{th} input of the encoder and the t^{th} character embedded output of the decoder. h_{dt} is the feature vector of the decoder at time t. Score(.) is a energy function, which is implemented by MLP in this work.

Compared with the previous encoder-decoder model, attention model is different in that it does not require the encoder to encode all input information into a fixed length vector. On the contrary, the encoder needs to encode the input into a sequence of vectors. After the encoder converts the input characters into the representation of vectors, the output of the encoder is given different weights for each prediction, and then the weighted sum is input to the decoder. In decoding, each step selects a subset with the largest weight from the vector sequence to decode. In this way, when producing each output, we can make full use of the information carried by the input sequence.

C. VOCODER

The encoder-decoder model converts characters into Mel spectrogram, and then Mel spectrogram is reverted to waveform by a vocoder. Some systems use Griffin-Lim algorithm to recover phase from Linear-Spectrum, and then use short-time Fourier transform to recover waveform. Using Griffin-Lim algorithm as vocoder is simple to implement, but it is slow and difficult to achieve real-time synthesis. Moreover, the waveform generated by Griffin-Lim is too smooth, with more voids and poor hearing. In this paper, WaveNet model is used as vocoder to remedy some shortcomings of Griffin-Lim algorithm.

The main component of WaveNet is the dilated causal convolution neural network. In Fig. 3, it is a schematic diagram of a dilated causal convolution neural network. Causal convolution can ensure the order of data output, that is, the prediction of model output at t time does not depend on any data in future. When modeling long sequences, causal convolution trains faster than recurrent neural networks because they have no cyclic connections. But causal convolution needs to enlarge the convolution kernel to enlarge the receptive field, and the computational cost is very high. WaveNet uses dilated convolution to increase the receptive field of causal convolution by several orders of magnitude without significantly increasing the computational effort [25].

Dilated convolution is a convolution method in which the convolution kernel jumps over data larger than itself. This method is similar to enlarging the convolution kernel by zero edge compensation. This method is efficient. Compared with normal convolution, dilated convolution can effectively make the network perform coarse-grained convolution operations and the output remains the same size as the input. As a special case, dilated convolution with dilated factor = 1 is the standard convolution. Stacked dilated convolution makes the network have a very large receptive field through only a few layers, while retaining the input resolution and computational efficiency.

WaveNet consists of the stacked dilated causal convolution, which can be used to generate audio signals directly. By inputting the predicted audio information into the network and utilizing the autoregressive characteristics of WaveNet architecture, the speech waveform can be obtained after recovering the detailed phase information which losts in conventional vocoders. In the previous works, traditional MFCC features and linguistic features were used as input in speech synthesis using WaveNet. In this paper, Mel spectrogram is used as input of audio signal features. Compared with other linguistic and acoustic features used in conventional vocoders, Mel spectrogram is trained more conveniently.

For a high-quality WaveNet vocoder, a large training dataset is required. However, paired speech and text is limited for the reference speaker for Lhasa-Tibetan dialect, but unlabeled speech is easy to obtain. We can use the unlabeled data to pretrain WaveNet, then fine-tune the WaveNet using labeled data. The WaveNet vocoder can be trained in a semi-supervised way.

In the phase of WaveNet training, first, an initialization WaveNet is trained with untranscribed speech data by using ground truth Mel spectrogram, and then it is fine-tuned with the predicted Mel spectrogram of the transcribed data which is used to train the encoder-decoder. The learnt WaveNet vocoder is used for reconstructing time-domain waveform from Mel spectrogram.

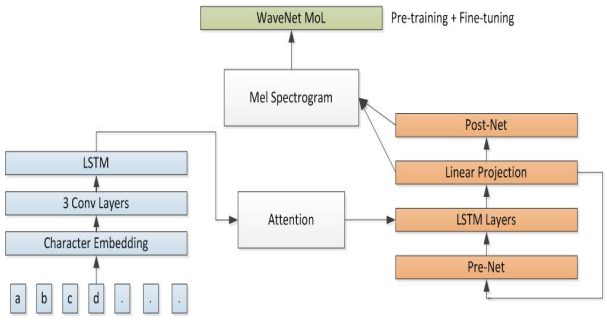


FIGURE 4. End-to-end speech synthesis model based on Tacotron 2.

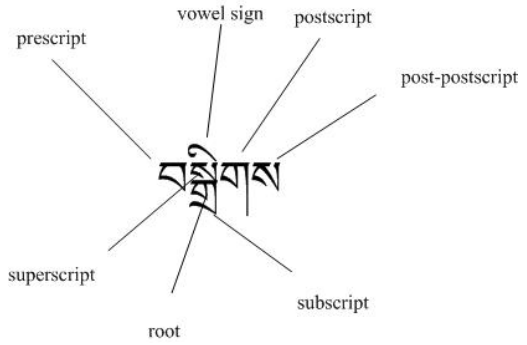


FIGURE 5. The structure of Tibetan syllable.

III. END-TO-END LHASA-TIBETAN SPEECH SYNTHESIS

The existing Tibetan speech synthesis based on deep neural network takes the phonemes of Lhasa-Tibetan dialect as the synthesis unit [20]. The front-end text processing is very heavy and requires extensive domain expertise of Tibetan linguistics to construct syllable lexicon, intials and finals, phoneme set. The process is time-consuming. However, one of the advantages of end-to-end model is to train Tibetan characters and synthesize speech signals directly. Considering the large number of Tibetan characters and the two-dimensional planar character structure, we used English letters converted from Tibetan characters as input text.

In this paper, the end-to-end speech synthesis model is based on Tacotron 2 [26], in which WaveNet model is used as the vocoder to solve the problems of Griffin-Lim algorithm in Tacotron producing unique artificial traces and low fidelity of synthetic speech. Fig. 4 is the end-to-end speech synthesis model used in this paper.

A. FRONT-END PROCESSING

Tibetan characters are written in Tibetan letters from left to right, but there is a vertical superposition in syllables (syllables are separated by delimiter "."), which is a two-dimensional planar character shown as Fig. 5.

A Tibetan sentence is shown in Fig. 6, where the sign "|" is used as the end sign of a Tibetan sentence. Tibetan syllable may be used as synthesis unit for end-to-end model, but the number of single syllable used commonly in Tibetan language is about 5600 [27], so the bits of one-hot vector will be large, and a large amount of speech data will be needed to

ད་རེས་ཀྱི་དོན་རྐྱེན་ཐོག་ནས་བཟོ་པའི་གྲལ་རིམ་ནི་བློས་འགེལ་ཆོག་པ་ཞིག་ཡིན་པ་དང།

FIGURE 6. A Tibetan sentence.

da res kyi don rkyen thog nas bzo pa'i gral rim ni blos 'gel chog pa zhig yin pa dang

FIGURE 7. A Tibetan sentence after Wylie transliteration.

train. In the work [19], Tibetan letters were used as the input text, but it still need to analyze the spelling rules of Tibetan characters and transform the two-dimensional character into one-dimensional letter sequence.

In this work, we converted the Tibetan characters into English letters, shown as in Fig. 7, using Wylie transliteration scheme.

The advantage of this method is that the synthesis unit is very small, which results in computation efficiency and good performance.

B. FEATURE SELECTION

In this paper, the spectral features of speech are extracted as the input of WaveNet vocoder. The main spectral characteristics of speech are Mel spectrogram and linear spectrogram (the amplitude of short-time Fourier transform). Mel spectrogram is a non-linear transformation applied to the frequency axis of short-time Fourier transform. It is smoother than waveform sample by compressing and transforming the frequency range with fewer dimensions. Because each frame is phase invariant, it is easier to train with mean square error loss (MSE) using Mel frequency spectrogram.

IV. EXPERIMENTAL RESULTS

A. EXPERIMENTAL DATA

Lhasa-Tibetan speech data used in this paper is from https://pan.baidu.com/s/14CihgqjA4AFFH1QpSTjzZw. We used 4-hour transcribed data and 2-hour untranscribed data of a single speaker. The data set contains 4572 sentences. The average length of each sentence is 14 Tibetan single syllable with an average length of 5 seconds. Speech data files are converted to 22050 Hz sampling frequency, 16bit quantization accuracy.

B. EXPERIMENTAL SETTING

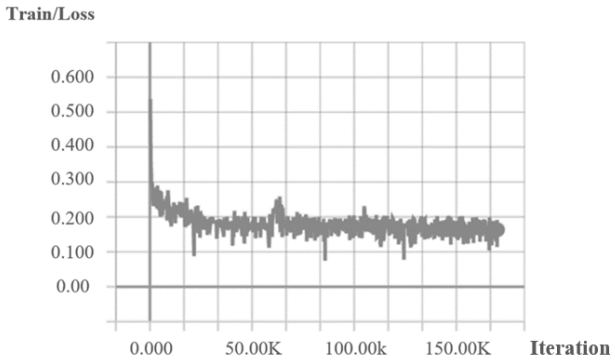
Some hyperparameters of our model based Tacotron 2 are provided in Table 1.

C. EXPERIMENTAL EVALUATION

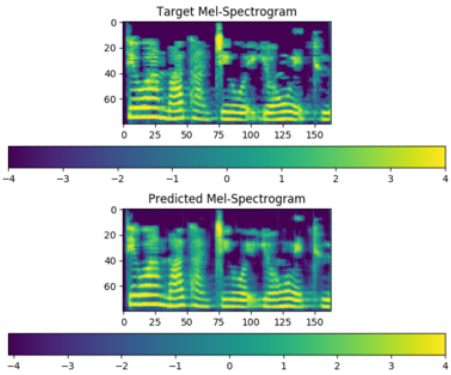
In order to evaluate the accuracy of the experimental results, subjective evaluation and objective evaluation are adopted in this paper. The fitting degree is objectively measured by the error loss function of model training. The smaller the loss value is, the better the fitting degree of the model is. Fig. 8(a) shows that the error of our model training is gradually minimized, which shows that the model fits well and converges quickly. At the same time, Fig. 8(b) and 8(c)

TABLE 1. Model hyperparameters in our model.

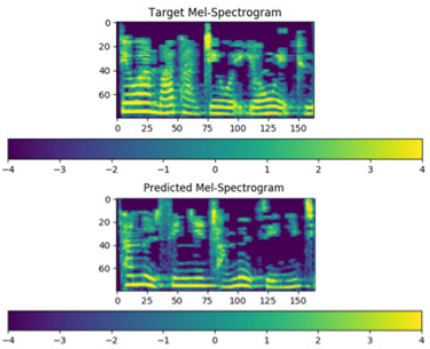
Audio parameter	
Window_size	50ms
Frame_shift	12.5ms
Num_mels	80
Num_freq	1025
Encoder parameter	
enc_conv_num_layers	3
enc_conv_kernel_size	5
enc_conv_channels	512
encoder_lstm_units	256
Attention parameter	
attention_dim	128
attention_filters	32
attention_kernel	31
attention_win_size	7
Decoder parameter	
prenet_layers	[256, 256]
decoder_layers	2
decoder_lstm_units	1024
max_iters	10000
Residual postnet	
postnet_num_layers	5
postnet_kernel_size	5
postnet_channels	512
CBHG mel->linear postnet	
cbhg_kernels	8
cbhg_conv_channels	128
cbhg_pool_size	2
cbhg_highwaynet_layers	4
cbhg_highway_units	128
cbhg_rnn_units	128
WaveNet	
layers	20
stacks	2
residual_channels	128
gate_channels	256
skip_out_channels	128
kernel_size	3
Tacotron Training	
tacotron_batch_size	32
tacotron_synthesis_batch_size	1
tacotron_decay_rate	0.5
tacotron_initial_learning_rate	1e-3
tacotron_final_learning_rate	1e-4



(a) The loss of our speech synthesis model



(b) The comparison of the output Mel spectrogram and the target Mel spectrogram for our model with English letters as input text



(c) The comparison of the output Mel spectrogram and the target Mel spectrogram for the model with single syllable of Tibetan character as input text

FIGURE 8. Feature prediction.

are the comparisons between the output Mel spectrogram and the target Mel spectrogram for our method and the one using single syllable of Tibetan characters as synthesis unit, respectively.

Subjectively, we conducted a Mean Opinion Score (MOS) test. In the MOS test, the listener scores 5 points on the clarity and naturalness of the synthesized speech after listening to each synthesized speech. The experimental results for our method with semi-supervised learning are shown in Table 2. It can be seen that the mean opinion scores of the clarity and naturalness of synthesized speech for our method

TABLE 2. MOS values of Lhasa-Tibetan synthetic speech based on our method with Semi-supervised learning.

Listener	Articulation	Naturalness
Listener1	3.9	4.0
Listener2	3.8	3.9
Listener3	4.0	4.1
Listener4	3.9	4.1
Listener5	4.1	4.2
MOS value	3.94	4.06

TABLE 3. MOS values of Lhasa-Tibetan synthetic speech based on our method without semi-supervised learning.

Listener	Articulation	Naturalness
Listener1	3.8	4.0
Listener2	3.8	4.0
Listener3	3.6	3.8
Listener4	3.7	3.8
Listener5	3.7	3.9
MOS value	3.72	3.9

TABLE 4. MOS values of Lhasa-Tibetan synthetic speech based on the method with single syllable of Tibetan character as input text.

Listener	Articulation	Naturalness
Listener1	3.3	3.5
Listener2	3.1	3.2
Listener3	3.4	3.5
Listener4	3.2	3.3
Listener5	3.0	3.2
MOS value	3.2	3.34

TABLE 5. MOS values of Lhasa-Tibetan synthetic speech based on linear predictive amplitude spectrum+Griffin-lim.

Listener	Articulation	Naturalness
Listener1	3.4	3.6
Listener2	3.3	3.5
Listener3	3.3	3.4
Listener4	3.5	3.6
Listener5	3.3	3.4
MOS value	3.36	3.5

TABLE 6. MOS values of Lhasa-Tibetan synthetic speech based on linear predictive amplitude spectrum+WaveNet.

Listener	Articulation	Naturalness
Listener1	3.7	3.8
Listener2	3.5	3.6
Listener3	3.7	3.8
Listener4	3.6	3.7
Listener5	3.6	3.8
MOS value	3.62	3.74

TABLE 7. MOS comparison of different feature for Lhasa-Tibetan speech synthesis.

Model	Aritculation	Naturalness
Linear predictive amplitude spectrum+Griffin-lim	3.36	3.5
Linear predictive amplitude spectrum+WaveNet	3.62	3.74
Mel spectrogram+WaveNet	3.72	3.9

with semi-supervised learning are 3.94 and 4.06 respectively, which are higher than the method without semi-supervised learning (shown in Table 3) and the one with single syllable of Tibetan characters as synthesis unit (shown in Table 4). WaveNet vocoder is benefited from using unlabeled speech data and effectively training in a semi-supervised way.

D. METHOD COMPARISON EXPERIMENT

In order to further verify the practicability of the end-to-end Lhasa-Tibetan speech synthesis system, we also compared the "Mel spectrum + WaveNet" model with the "linear prediction amplitude spectrum + Griffin-lim", "linear prediction amplitude spectrum + WaveNet" speech synthesis model. The MOS results of models are shown in Table 5, Table 6 for "linear prediction amplitude spectrum + Griffin-lim" and "linear prediction amplitude spectrum + WaveNet" respectively. The MOS comparison results of models are shown in Table 7.

As shown in Table 7, WaveNet performs better than Griffin-Lim in restoring Lhasa-Tibetan speech phase information and produces much higher quality of speech. It also shows that linear prediction amplitude spectrum has a competitive performance compared with Mel spectrogram for the feature prediction network. However, Mel spectrogram is more effective feature since it is a more compact representation.

V. CONCLUSION

In this paper, a Tibetan speech synthesis system based on end-to-end model is implemented by using attention-based encoding-decoding network and semi-supervised learning WaveNet. In the end-to-end model, the input is a series of English letters converted from Tibetan character sequences, which are trained by encoding-decoding network and output as acoustic spectrogram. Then the corresponding speech is generated using WaveNet network. Compared with the model without using semi-supervised learning way, the model using single syllable of Tibetan character and the model using Griffin-Lim algorithm as vocoder, the performance of our method are superior in clarity and naturalness.

REFERENCES

[1] S. Nie, M. Zheng, and Q. Ji, "The deep regression Bayesian network and its applications: Probabilistic deep learning for computer vision," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 101–111, Jan. 2018.

[2] G. Gui, H. Huang, Y. Song, and H. Sari, "Deep learning for an effective nonorthogonal multiple access scheme," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8440–8450, Sep. 2018.

[3] J. Wang, Y. Ding, S. Bian, Y. Peng, M. Liu, and G. Gui, "UL-CSI data driven deep learning for predicting DL-CSI in cellular FDD systems," *IEEE Access*, vol. 7, pp. 96105–96112, 2019.

[4] Y. Wang, M. Liu, J. Yang, and G. Gui, "Data-driven deep learning for automatic modulation recognition in cognitive radios," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 4074–4077, Apr. 2019.

[5] J. Sun, W. Shi, Z. Yang, J. Yang, and G. Gui, "Behavioral modeling and linearization of wideband RF power amplifiers using BiLSTM networks for 5G wireless systems," *IEEE Trans. Veh. Technol.*, to be published. doi: 10.1109/TVT.2019.2925562.

[6] H. Huang, Y. Song, J. Yang, G. Gui, and F. Adachi, "Deep-learning-based millimeter-wave massive MIMO for hybrid precoding," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 3027–3032, Mar. 2019.

[7] D. F. Zhang, G. Y. Li, and Y. D. Zhao, "Development review and research status of Speech Synthesis Technology," *Technol. Wind*, no. 22, p. 72, 2017.

[8] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[9] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Conf.*, May 1996, pp. 373–376.

[10] A. W. Black and P. A. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," in *Proc. EUROSPEECH*, 1997, pp. 601–604.

[11] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, Istanbul, Turkey, Jun. 2000, pp. 1315–1318.

[12] R. Z. M. Cai and Z. J. Cai, "Unit selection algorithm for corpus-based tibetan speech synthesis," *J. Chin. Inf. Process.*, vol. 31, no. 5, pp. 59–63, 2017.

[13] R. Z. M. Cai, "Research on tibetan speech synthesis technology based on mixed primitives," M.S. thesis, Comput. Sci. Technol. Specialty, Shanxi Normal Univ., 2016.

[14] L. Gao, H. Z. Yu, and W. S. Zheng, "Technology research on tibetan lhasa speech synthesis based on HMM," *J. Northwest Univ. Nationalities (Nature Sci.)*, vol. 32, no. 2, pp. 30–35, 2011.

[15] J. X. Zhang, "Research on tibetan lhasa speech synthesis based on HMM," M.S. thesis, Comput. Appl. Technol. Specialty, Northwest Univ. Nationalities, 2014.

[16] S. P. Xu, "Research on speech quality evaluation for tibetan statistical parametric speech synthesis," M.S. thesis, Circuits Syst. Specialty, Northwest Normal Univ., 2015.

[17] X. J. Kong, "Research on methods of text analysis for tibetan statistical parametric speech synthesis," M.S. thesis, Electron. Sci. Technol. Specialty, Northwest Normal Univ., 2017.

[18] Y. Zhou and D. C. Zhao, "Research on HMM-based tibetan speech synthesis," *Comput. Appl. Softw.*, vol. 32, no. 5, pp. 171–174, 2015.

[19] G. C. Du, R. Z. M. Cai, C. J. Nan, and T. B. Suan, "Neural network based tibetan speech synthesis," *J. Chin. Inf. Process.*, vol. 33, no. 2, pp. 75–80, 2019.

[20] L. Luo, G. Li, C. Gong, and H. Ding, "End-to-end speech synthesis for Tibetan Lhasa dialect," *J. Phys., Conf.*, vol. 1187, no. 5, 2019, Art. no. 052061.

[21] O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. Hinton, "Grammar as a Foreign language," 2014, *arXiv:1412.7449*. [Online]. Available: https://arxiv.org/abs/1412.7449

[22] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-Sequence models can directly translate foreign speech," 2017, *arXiv:1703.08581*. [Online]. Available: https://arxiv.org/abs/1703.08581

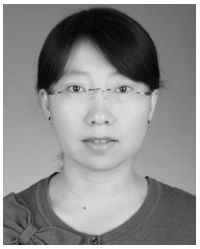
[23] L. Sutskever, O. Vinyals, and Q. V. Le, "Sequence-to-sequence learning with neural networks," 2014, *arXiv:1409.3215v3*. [Online]. Available: https://arxiv.org/abs/1409.3215

[24] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 4006–4010.

[25] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*. [Online]. Available: https://arxiv.org/abs/1609.03499

[26] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," 2018, *arXiv:1712.05884*. [Online]. Available: https://arxiv.org/abs/1712.05884

[27] L. D. Z. La, "Research on tibetan speech recognition technology," M.S. thesis, Chin. Ethnic Lang. Literature Specialty, Tibet Univ., 2015.



YUE ZHAO received the B.S. degree in automation from Northeastern University, Shenyang, China, in 1997, and the Ph.D. degree in control theory and control engineering from the University of Science and Technology Beijing, in 2006.

From 2009 to 2010, she has ever been a Visiting Scholar with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute. Since 2013, she has been a Professor of automation engineering with the Minzu University of China. She is the author of three books, more than 50 articles, and more than four inventions. Her research interests include probabilistic graphical models, speech signal processes and applications, computer vision, and embedded systems.

PANHUA HU was born in Fujian, in 1994. He received the bachelor's degree in computer science and technology from Shijiazhuang Railway University and the master's degree from the Minzu University of China. His current research interest includes speech synthesis based on deep learning.

XIAONA XU received the Ph.D. degree in control theory and control engineering from the University of Science and Technology Beijing, in 2008. She is currently a Lecturer with the Minzu University of China. Her research interests include pattern recognition, machine learning in general with focus on speech, and image data processing.

LICHENG WU received the Ph.D. degree in robotics from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 1995. He is currently a Full Professor and the Dean of the School of Information Engineering, Minzu University of China. His research interests include speech recognition, artificial robotics, and computer games.

XIALI LI was born in Henan, China, in 1979. She received the M.S. degree in computer science and technology from Xi'an Jiaotong University (XJTU), Xi'an, China, in 2004. From October 2008 to August 2009, she was a Visiting Scholar with the University of Edinburgh, U.K. Since 2013, she has been an Associate Professor with the Minzu University of China. She is the author of three books and more than 50 articles. Her research interests include computer games and artificial robotics.