Hindawi
Complexity
Volume 2021, Article ID 6682871, 8 pages
https://doi.org/10.1155/2021/6682871

Research Article

End-to-End Speech Synthesis for Tibetan Multidialect

Xiaona Xu, Li Yang, Yue Zhao, and Hui Wang

School of Information Engineering, Minzu University of China, Beijing 100081, China

Correspondence should be addressed to Li Yang; 654577893@qq.com and Yue Zhao; zhaoyueso@muc.edu.cn

Received 30 October 2020; Revised 27 December 2020; Accepted 12 January 2021; Published 27 January 2021

Academic Editor: Ning Cai

The research on Tibetan speech synthesis technology has been mainly focusing on single dialect, and thus there is a lack of research on Tibetan multidialect speech synthesis technology. This paper presents an end-to-end Tibetan multidialect speech synthesis model to realize a speech synthesis system which can be used to synthesize different Tibetan dialects. Firstly, Wylie transliteration scheme is used to convert the Tibetan text into the corresponding Latin letters, which effectively reduces the size of training corpus and the workload of front-end text processing. Secondly, a shared feature prediction network with a cyclic sequence-to-sequence structure is built, which maps the Latin transliteration vector of Tibetan character to Mel spectrograms and learns the relevant features of multidialect speech data. Thirdly, two dialect-specific WaveNet vocoders are combined with the feature prediction network, which synthesizes the Mel spectrum of Lhasa-Ü-Tsang and Amdo pastoral dialect into time-domain waveform, respectively. The model avoids using a large number of Tibetan dialect expertise for processing some time-consuming tasks, such as phonetic analysis and phonological annotation. Additionally, it can directly synthesize Lhasa-Ü-Tsang and Amdo pastoral speech on the existing text annotation. The experimental results show that the synthesized speech of Lhasa-Ü-Tsang and Amdo pastoral dialect based on our proposed method has better clarity and naturalness than the Tibetan monolingual model.

1. Introduction

Speech synthesis, also known as text-to-speech (TTS) technology, mainly solves the problem of converting text information into audible sound information. Up to now, speech synthesis technology has become one of the most commonly used methods of human-computer interaction. It is gradually replacing traditional human-computer interaction methods, making human-computer interaction more convenient and faster. With the continuous development of speech synthesis technology, multilingual speech synthesis technology has become a research interest for researchers. This technology can realize the synthesis of different languages in a unified speech synthesis system [1–3].

There are lots of ethnic minorities in China. Many ethnic minorities have their own languages and scripts. Tibetan is one of the minority languages; it can be divided into three major dialects: Ü-Tsang dialect, Amdo dialect, and Kham dialect, which are mainly used in Tibet, Qinghai, Sichuan, Gansu, and Yunnan. All dialects use Tibetan characters as written text, but there are some differences in the pronunciation of each dialect, so it is difficult for the people who use different dialects to communicate with each other. There have been some research studies on Lhasa-Ü-Tsang dialect speech synthesis technology [4–12]. The end-to-end method [12] has more training advantages than the statistical parameter method, and the synthesis effect is better. There are few existing research studies on the speech synthesis of Amdo dialect, and only the work [13] applied the statistical parameter speech synthesis (SPSS) based on the hidden Markov model (HMM) for Tibetan Amdo dialect.

For the multilingual speech synthesis, the research works mainly use unit-selection concatenative synthesis technique, SPSS based on HMM, and deep learning technology. The unit-selection concatenative synthesis technique mainly includes selecting an unit scale, constructing a corpus and designing an algorithm of unit selection and splicing. This method relies on a large-scale corpus [14, 15]. Additionally, the synthesis effect is unstable and the connection of the splicing unit may have discontinuities. SPSS technology usually requires a complex text front-end to extract various

linguistic features from raw text, a duration model, an acoustic model, which is used to learn the transformation between linguistic features and acoustic features, and a complex signal-processing-based vocoder to reconstruct waveform from the predicted acoustic features. The work [16] proposes a framework for estimating HMM on data containing both multiple speakers and multiple languages, aiming to transfer a voice from one language to others. The works [2, 17, 18] propose a method to realize HMM-based cross-lingual SPSS using speaker adaptive training. For speech synthesis technology based on deep learning, the work [19] realizes a deep neural network- (DNN-) based Mandarin-Tibetan bilingual speech synthesis. The experimental results show that synthesized Tibetan speech is better than the HMM-based Mandarin-Tibetan cross-lingual speech synthesis. The work [20] trains the acoustic models with DNN, hybrid long short-term memory (LSTM), and hybrid bidirectional long short-term memory (BLSTM) and implements a deep learning-based Mandarin-Tibetan cross-lingual speech synthesis under a unique framework. Experiments demonstrated that the hybrid BLSTM-based cross-lingual speech synthesis framework was better than the Tibetan monolingual framework. Additionally, there are some research studies which reveal that multilingual speech synthesis using the end-to-end method gains a good performance. The work [21] presents an end-to-end multilingual speech synthesis model using a Unicode encoding "byte" input representation to train a model which outputs the corresponding audio of English, Spanish, or Mandarin. The work [22] proposes a multispeaker, multilingual TTS synthesis model based on Tacotron which is able to produce high-quality speech in multiple languages.

Taking into account that traditional methods require a lot of professional knowledge for phoneme analysis, tone, and prosody labelling, the work is time-consuming and costly, and the modules are usually trained separately, which will lead to the effect of error stacking [23] while the end-to-end speech synthesis system can automatically learn alignments and mapping from linguistic features to acoustic features. These systems can be trained on < text, audio > pairs without complex language-dependent text front-end. Inspired by above works, this paper proposes to use an end-to-end method to implement speech synthesis in Lhasa-Ü-Tsang and Amdo pastoral dialect, using a single sequence-to-sequence (seq2seq) architecture with attention mechanism as the shared feature prediction network for Tibetan multi-dialect and introducing two dialect-specific WaveNet networks to realize the generation of time-domain waveforms.

There are some similarities between this work and works [12, 24]. The WaveNet model is used in these works. However, in our work and [12], the WaveNet model is used for the generation of waveform sample with the input of predicted Mel spectrogram for speech synthesis. In the work [24] about speech recognition, the WaveNet model is used for the generation of text sequence and the input is MFCC features. The work [12] achieved the speech synthesis for Tibetan Lhasa-Ü-Tsang by using end-to-end model. In this paper, we improved the model of the work [12] to implement multidialect speech synthesis.

Our contributions can be summarized as follows. (1) We propose an end-to-end Tibetan multidialect speech synthesis model, which unifies all the modules into one model and realizes the speech synthesis for different Tibetan dialects using one speech synthesis system. (2) Joint learning is used to train the shared feature prediction network by learning the relevant features of multidialect speech data, and it is helpful to improve the speech synthesis performance of different dialects. (3) We use Wylie transliteration scheme to convert the Tibetan text into the corresponding Latin letters, which is used as the training units of the model. It effectively reduces the size of training corpus, reduces the workload of front-end text processing, and improves the modelling efficiency.

The rest of this paper is organized as follows. Section 2 introduces the end-to-end Tibetan multidialect speech synthesis model. The experiments are presented in detail in Section 3 and the results are discussed as well. Finally, we describe our conclusions in Section 4.

2. Model Architecture

The end-to-end speech synthesis model is mainly composed of two parts: the first part contains a seq2seq feature prediction network containing attention mechanism and the second part contains two dialect-specific WaveNet vocoders based on Mel spectrogram. The model adopts a synthesis method from text to intermediate representation and intermediate representation to speech waveform. The encoder and decoder implement the conversion from text to intermediate representation, and the WaveNet vocoders restore the intermediate representation into waveform samples. Figure 1 shows the end-to-end Tibetan multidialect speech synthesis model.

2.1. Front-End Processing. Although Tibetan pronunciation has evolved over thousands of years, the orthography of written language remains unchanged. It led to Tibetan spelling becoming very complicated. Tibetan sentence is written from left to right and consists of a series of single syllable. Single syllable is also called Tibetan character. The punctuation mark "་"means the "soundproof symbol" between the syllables, and the single hanging symbol "།" is used at the end of a phrase or sentence. Figure 2 shows a Tibetan sentence.

Each syllable in Tibetan has a root, which is the central consonant of the syllable. A vowel label can be added above or below the root to indicate different vowels. Sometimes, there is a superscript at the top of the root, one or two subscripts at the bottom, and a prescript at the front, indicating that the initials of the syllable are compound consonants. The sequence of connection of compound consonants is prescript, superscript, root, and subscript. Sometimes, there is one or two postscripts after the root, which means that the syllable has one or two consonant endings. The structure of Tibetan syllables is shown in Figure 3.

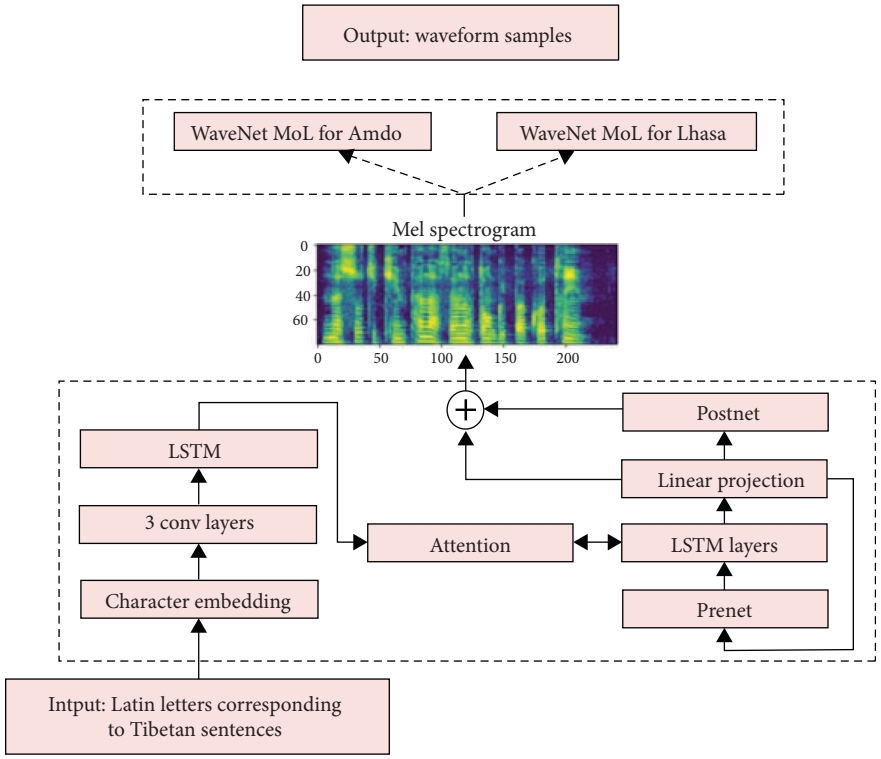

Figure 1: End-to-end Tibetan multidialect speech synthesis model.

དར་རས་ཀྱི་དོན་ཆུན་ཐོག་ནས་བཟོ་བའི་གལ་རིམ་ནི་བློས་འགེལ་ཚོག་པ་ཞིག་ཡིན་པ་དང།

Figure 2: A Tibetan sentence.

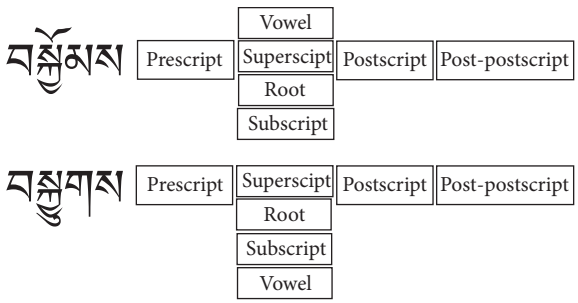

Figure 3: The structure of Tibetan syllables.

Due to the complexity of Tibetan spelling, a Tibetan syllable can have as many as 20450 possibilities. If a single syllable of a Tibetan character is used as the basic unit of speech synthesis, a large amount of speech data will need to be trained, and the corpus construction workload will be huge. The existing Tibetan speech synthesis system [10, 13] uses the initials and vowels of Tibetan characters as the input of the model, which requires a lot of professional knowledge of Tibetan linguistics and the front-end text processing. In this paper, we adopt the Wylie transliteration scheme, using only the basic 26 Latin letters, without adding letters and symbols, to convert the Tibetan text into the corresponding Latin letters. It effectively reduces the size of training corpus, reduces the workload of front-end text processing, and improves modelling efficiency. Figure 4 shows the converted Tibetan sentence obtained by using the Wylie transliteration scheme for the Tibetan sentence in Figure 2.

2.2. The Shared Feature Prediction Network. We use Lhasa-Ü-Tsang and Amdo dialect datasets to train the shared feature prediction network and capture the relevant features between two dialects speech data by joint learning. The shared feature prediction network is used to map the Latin transliteration vector of Tibetan character to Mel spectrograms. In this process, Lhasa-Ü-Tsang and Amdo dialect share the same feature prediction network. The shared feature prediction network consists of an encoder, an attention mechanism, and a decoder.

2.2.1. Encoder. The encoder module is used to extract the text sequence representation, including a character embedding layer, 3 convolutional layers, and a long short-term memory (LSTM) layer, as shown in the lower left part of Figure 1. Firstly, the input Tibetan characters are embedded into sentence vectors using character embedding layer and then input into 3 convolutional layers. These convolutional layers model longer-term context in the input character sequence, and the output of the final convolutional layer will

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

Figure 4: A Tibetan sentence after Wylie transliteration.

be used as the input of a single bidirectional LSTM layer to generate intermediate representations.

2.2.2. Decoder. The decoder consists of a prenet layer, LSTM layers, and a linear projection layer, as shown in the lower right part of Figure 1. The decoder is an autoregressive recurrent neural network, which is used to predict the output spectrogram according to the encoded input sequence. The result of the previous prediction is first input to a prenet layer, and the output of the prenet layer and the output context vector of the attention mechanism network are concatenated and passed through 2 unidirectional LSTM layers. Then, the LSTM output and the attention context vector are concatenated and passed through a linear project layer to predict the target spectrogram frame. Finally, the predicted Mel spectrogram passes through a postnet, and the residual connection is made with the predicted spectrum to obtain the Mel spectrogram.

2.2.3. Attention Mechanism. The input sequence in the seq2seq structure will be compiled into a feature vector C of a certain dimension. The feature vector C always links the encoding and decoding stages of the encoder-decoder model. The encoder compresses the information of the entire sequence into a fixed-length vector. But with the continuous growth of the sequence, this will cause the feature vector to fail to fully represent the information of the entire sequence, and the latter input sequence will easily cover the first input sequence, which will cause the loss of many detailed information. To solve this problem, an attention mechanism is introduced. This mechanism will encode the encoder into different c_i, according to each time step of the sequence, that is, the original unified feature vector C will be replaced with a constantly changing c_i according to the current generated word. When decoding, combine each different c_i to decode the output so that when each output is generated, the information carried by the input sequence can be fully utilized, and the result will be more accurate.

The feature vector c_i is obtained by adding the hidden vector sequence $(h_1, h_2, \ldots, h_{T_x})$ during encoding according to the weight, as shown in equation (1). α_{ij} is the weight value, as in equation (2), which represents the matching degree between the jth input of the encoder and the ith output of the decoder.

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j, \tag{1}$$

$$\alpha_{ij} = \frac{\exp\left(e_{ij}\right)}{\sum_{k=1}^{T_x} \exp\left(e_{ik}\right)}. \tag{2}$$

2.3. WaveNet Vocoder. Tacotron [25, 26] launched by Google can convert phonetic characters or text data into frequency spectrum, and it also needs a vocoder to restore the frequency spectrum to waveforms to obtain synthesized speech. Tacotron's vocoder uses the Griffin–Lim algorithm for waveform reconstruction. The Griffin–Lim algorithm is an algorithm to reconstruct speech under the condition that only the amplitude spectrum is known and the phase spectrum is unknown. It is a relatively classic vocoder with simple and efficient algorithm. However, because the waveform generated by the Griffin–Lim vocoder is too smooth, the synthesized voice has a poor quality and sounds obviously "mechanical." WaveNet is a typical autoregressive generation model, which can improve the quality of synthetic speech. Therefore, this work uses the WaveNet model as a vocoder to cover the limitation of the Griffin–Lim algorithm. The sound waveform is a one-dimensional array in time domain, and the audio sampling points are usually relatively large. The waveform data at a sampling rate of 16 kHz will have 16000 elements per second, which requires a large amount of calculation using ordinary convolution. In this regard, WaveNet uses causal convolution, which can increase the receptive field of convolution. But causal convolution requires more convolutional layers, which is computationally complex and costly. Therefore, WaveNet has adopted the method of dilated causal convolution to expand the receptive field of convolution without significantly increasing the amount of calculation. The dilated convolution is shown in Figure 5. When the network generates the next element, it can use more previous element values. WaveNet is composed of stacked dilated causal convolutions and synthesizes speech by fitting the distribution of audio waveforms by the autoregressive method, that is, WaveNet predicts the next sampling point according to a number of input sampling points and synthesizes speech by predicting the value of the waveform at each time point waveform.

In the past, traditional acoustic and linguistic features were used as the input of the WaveNet model for speech synthesis. In this paper, we choose a low-level acoustic representation: Mel spectrogram, as the input of WaveNet for training. The Mel spectrogram emphasizes the details of low frequency, which is very important for the clarity of speech. And compared to the waveform samples, the phase of each frame in Mel spectrogram is unchanged; it is easier to train with the square error loss. We train WaveNet vocoders for Lhasa-Ü-Tsang dialect and Amdo pastoral dialect, and they can synthesize the corresponding Tibetan dialects with the corresponding WaveNet vocoder.

2.4. Training Process. Training process can be summarized into 2 steps: firstly, training the shared feature prediction network; secondly, training a dialect-specific WaveNet vocoder for Lhasa-Ü-Tsang dialect and Amdo pastoral dialect, respectively, based on the outputs generated by the network which was trained in step 1.

We trained the shared feature prediction network on the datasets of Lhasa-Ü-Tsang dialect and Amdo pastoral dialect. On a single GPU, we used the teacher-forcing method to train the feature prediction network, and the input of the

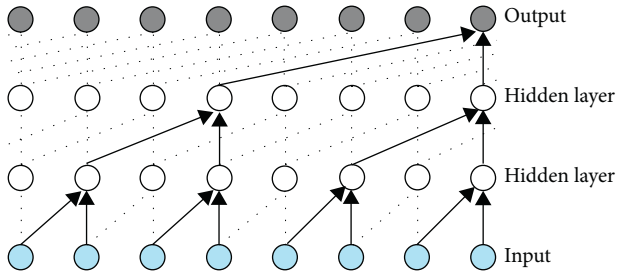

Figure 5: Dilated causal convolution [27].

decoder was the correct output, not the predicted output, with a batch size of 8. An Adam optimizer was used with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-6}$. The learning rate decreased from 10^{-3} to 10^{-4}, after 40000 iterations.

Then, the predicted outputs from the shared feature prediction network were aligned with the ground truth. We trained the WaveNet for Lhasa-Ü-Tsang dialect and Amdo pastoral dialect, respectively, by using the aligned predicted outputs. It means that these predicted data were generated in the teacher-forcing mode. Therefore, each spectrum frame is exactly aligned with a sample of the waveform. In the process of training the WaveNet network, we used an Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-6}$, and the learning rate was fixed at 10^{-3}.

3. Results and Analysis

3.1. Experimental Data. The training data consist of the Lhasa-Ü-Tsang dialect and Amdo pastoral dialect. The Lhasa-Ü-Tsang dialect speech data are about 1.43 hours with 2000 text sentences. The Amdo pastoral dialect speech data are about 2.68 hours with 2671 text sentences. Speech data files are converted to 16 kHz sampling rate, with 16 bit quantization accuracy.

3.2. Experimental Evaluation. In order to ensure the accuracy of the experimental results, we apply two methods, objective and subjective experiments, to evaluate the experimental results.

In objective experiment, the root mean square error (RMSE) of the time-domain sequences is calculated to measure the difference between the synthesized speech and the reference speech. The smaller the RMSE is, the closer the synthesized speech is to the reference and the better the effect of speech synthesis is. The formula of RMSE is shown in equation (3), where $x_{1,t}$ and $x_{2,t}$, respectively, represent the value of the time series of reference speech and synthesized speech at time t.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=1}^{n} \left(x_{1,t} - x_{2,t}\right)^2}{n}}. \tag{3}$$

For Lhasa-Ü-Tsang dialect and Amdo pastoral dialect, we randomly select 10 text sentences, use end-to-end Tibetan multi-dialect speech synthesis model for speech synthesis, and calculate the average RMSE to evaluate the closeness of the synthesized speech of the Lhasa-Ü-Tsang dialect and Amdo pastoral dialect to the reference speech. In order to evaluate the performance of the model, we compare it with the end-to-end Tibetan Lhasa-Ü-Tsang dialect speech synthesis model and end-to-end Tibetan Amdo pastoral dialect speech synthesis model. These two models were used to synthesize the same 10 text sentences, and the average RMSE was calculated. The results are shown in Table 1. For Lhasa-Ü-Tsang dialect, the RMSE of the multidialect speech synthesis model is 0.2126, which is less than the one of Lhasa-Ü-Tsang dialect speech synthesis model (0.2223). For Amdo pastoral dialect, the RMSE of the multidialect speech synthesis model is 0.1223, which is less than the one of Amdo pastoral dialect speech synthesis model (0.1253). It means that both Lhasa-Ü-Tsang dialect and Amdo pastoral dialect, which are synthesized by our model, are closer to their reference speech. The results show that our method has capability of the feature representation for both Lhasa-Ü-Tsang and Amdo pastoral dialect through the shared feature prediction network, so as to improve the multidialect speech synthesis performance against single dialect. Besides, the synthetic speech effect of Amdo pastoral dialect is better than that of Lhasa-Ü-Tsang dialect because the data scale of Amdo pastoral dialect is larger than that of Lhasa-Ü-Tsang dialect.

Figures 6 and 7, respectively, show the predicted Mel spectrogram and target Mel spectrogram output by the feature prediction network for Lhasa-Ü-Tsang dialect and Amdo pastoral dialect. It can be seen from the figures that the predicted mel spectrograms of Lhasa-Ü-Tsang dialect and Amdo pastoral dialect are both similar to the target Mel spectrograms.

In subjective experiment, the absolute category rating (ACR) measurement method was used to evaluate the synthesized speech of the Lhasa-Ü-Tsang and Amdo pastoral dialects mentioned above. In the ACR measurement, we selected 25 listeners. After listening to the synthesized speech, we used the original speech as a reference and scored the synthesized speech according to the grading standard in Table 2. After obtaining the scores given by all listeners, the mean opinion score (MOS) of the synthesized speech was calculated, and Table 3 shows the results. The MOS values of the synthesized speech in Lhasa-Ü-Tsang dialect and Amdo pastoral dialects are 3.95 and 4.18, respectively, which means that the synthesized speech has good clarity and naturalness.

3.3. Comparative Experiment. In order to verify the performance of the end-to-end Tibetan multidialect speech synthesis system, we have compared it with the "linear prediction amplitude spectrum + Griffin–Lim" and "Mel spectrogram + Griffin–Lim" speech synthesis system. The results of comparison experiment are shown in Table 4. According to Table 4, it can be seen that whether it is Lhasa-Ü-Tsang dialect or Amdo pastoral dialect, the MOS value of the synthesized speech of "Mel spectrogram + Griffin–Lim" speech synthesis system is higher than that of "linear prediction amplitude spectrum + Griffin–Lim" speech synthesis system. The results show that the Mel spectrogram is more effective as a predictive feature than the linear predictive

Table 1: Objective evaluation of the results.

Tibetan dialect	The RMSE of end-to-end Tibetan multidialect speech synthesis model	The RMSE of end-to-end Tibetan Lhasa-Ü-Tsang dialect speech synthesis model	The RMSE of end-to-end Tibetan Amdo pastoral dialect speech synthesis model
Lhasa-Ü-Tsang dialect	0.2126	0.2223	—
Amdo pastoral dialect	0.1223	—	0.1253

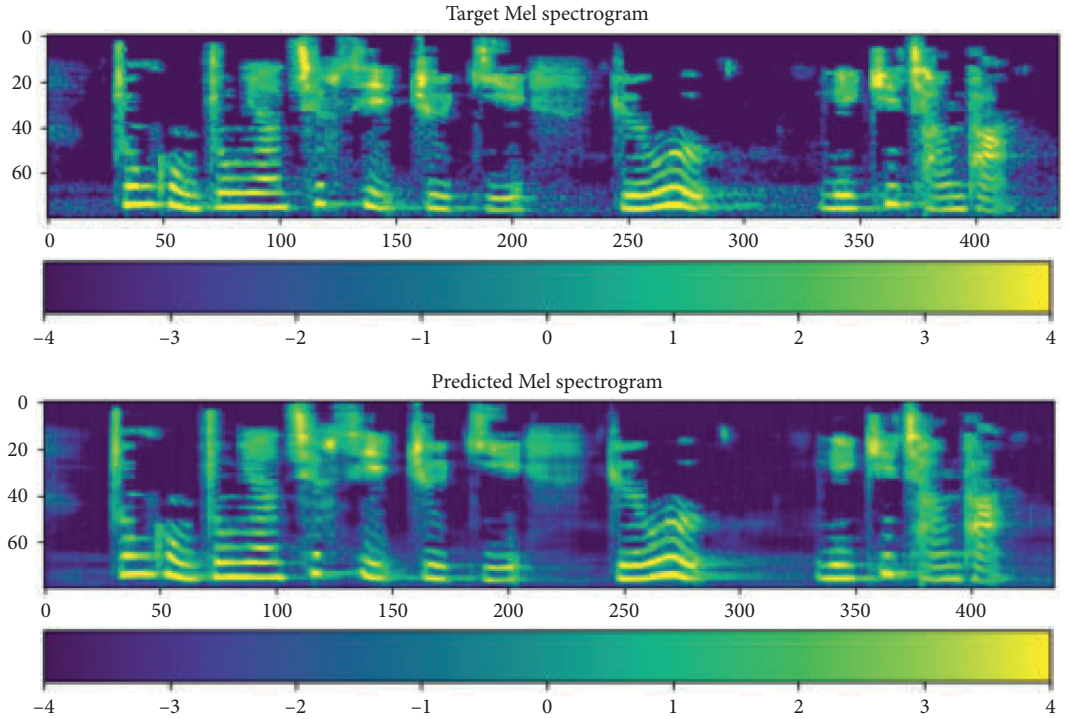

Figure 6: The comparison of the output Mel spectrogram and the target Mel spectrogram of Lhasa-Ü-Tsang dialect.

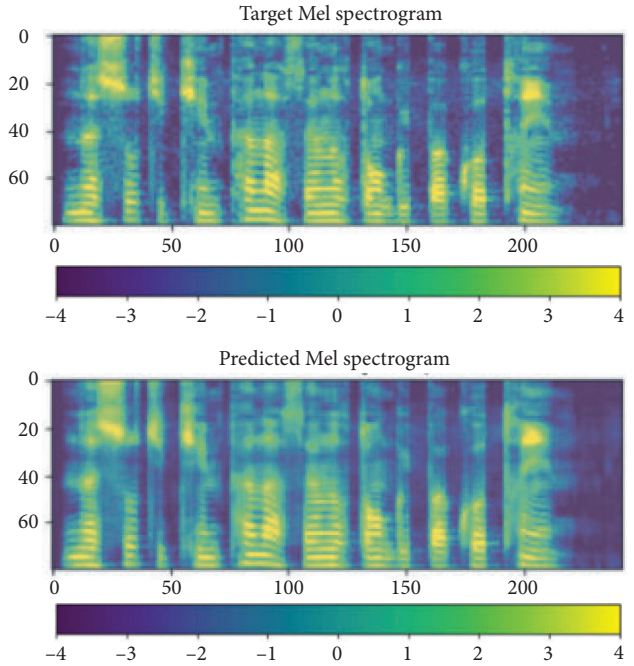

Figure 7: The comparison of the output Mel spectrogram and the target Mel spectrogram of Amdo pastoral dialect.

Table 2: Grading standards of ACR.

Grading value	Estimated quality
5	Very good
4	Good
3	Medium
2	Bad
1	Very bad

Table 3: The MOS comparison of speech synthesized by different synthesis primitive models.

Tibetan dialect	MOS
Lhasa-Ü-Tsang dialect	3.95
Amdo pastoral dialect	4.18

Table 4: The MOS comparison of speech synthesized by different models.

Model	MOS of Lhasa-Ü-Tsang dialect	MOS of Amdo pastoral dialect
Linear predictive amplitude spectrum + Griffin–Lim	3.30	3.52
Mel spectrogram + Griffin–Lim	3.55	3.70
Mel spectrogram + WaveNet	3.95	4.18

amplitude spectrum, and the quality of the generated speech is higher. The "Mel spectrogram + WaveNet" speech synthesis system outperforms the "Mel spectrogram + Griffin–Lim" speech synthesis system with the higher MOS value, which means that WaveNet has a better performance in recovering speech phase information and generating higher quality of the synthesis speech than the Griffin–Lim algorithm.

4. Conclusion

This paper builds an end-to-end Tibetan multidialect speech synthesis model, including a seq2seq feature prediction network, which maps the character vector to the Mel spectrogram, and a dialect-specific WaveNet vocoder for Lhasa-Ü-Tsang dialect and Amdo pastoral dialect, respectively, which synthesizes the Mel spectrogram into time-domain waveform. Our model can utilize dialect-specific WaveNet vocoders to synthesize corresponding Tibetan dialect. In the experiments, Wylie transcription scheme is used to convert Tibetan characters into Latin letters, which effectively reduces the number of composite primitives and the scale of training data. Both objective and subjective experimental results show that the synthesized speech of Lhasa-Ü-Tsang dialect and Amdo pastoral dialect has high qualities.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the National Natural Science Foundation of China under grant no. 61976236.

References

[1] Y. J. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," in *Proceedings of the Interspeech, 10th Annual Conference of the International Speech Communication Association*, Brighton, UK, September 2009.

[2] H. Liu, *Research on HMM-Based Cross-Lingual Speech Synthesi*, University of Science and Technology of China, Hefei, China, 2011.

[3] R. Sproat, *Multilingual Text-To-Speech Synthesis: The Bell Labs Approach*, Kluwer Academic Publishers, Amsterdam, Netherland, 1998.

[4] Z. M. Cairang, *Research on Tibetan Speech Synthesis Technology Based on Mixed Primitives*, Shanxi Normal University, Xi'an, China, 2016.

[5] L. Gao, Z. H. Yu, and W. S. Zheng, "Research on HMM-based Tibetan Lhasa speech synthesis technology," *Journal of Northwest University for Nationalities*, vol. 32, no. 2, pp. 30–35, 2011.

[6] J. X. Zhang, *Research on Tibetan Lhasa Speech Synthesis Based on HMM*, Northwest University for Nationalities, Lanzhou, China, 2014.

[7] S. P. Xu, *Research on Speech Quality Evaluation for Tibetan Statistical Parametric Speech Synthesis*, Northwest Normal University, Lanzhou, China, 2015.

[8] X. J. Kong, *Research on Methods of Text Analysis for Tibetan Statistical Parametric Speech Synthesis*, Northwest Normal University, Lanzhou, China, 2017.

[9] Y. Zhou and D. C. Zhao, "Research on HMM-based Tibetan speech synthesis," *Computer Applications and Software*, vol. 32, no. 5, pp. 171–174, 2015.

[10] G. C. Du, Z. M. Cairang, Z. J. Nan et al., "Tibetan speech synthesis based on neural network," *Journal of Chinese Information Processing*, vol. 33, no. 2, pp. 75–80, 2019.

[11] L. S. Luo, G. Y. Li, C. W. Gong, and H. L. Ding, "End-to-end speech synthesis for Tibetan Lhasa dialect," *Journal of Physics: Conference Series*, vol. 1187, no. 5, 2019.

[12] Y. Zhao, P. Hu, X. Xu, L. Wu, and X. Li, "Lhasa-Tibetan speech synthesis using end-to-end model," *IEEE Access*, vol. 7, pp. 140305–140311, 2019.

[13] L. Su, *Research on the Speech Synthesis of Tibetan Amdo Dialect Based on HMM*, Northwest Normal University, Lanzhou, China, 2018.

[14] S. Quazza, L. Donetti, L. Moisa, and P. L. Salza, "Actor: a multilingual unit-selection speech synthesis syste," in *Proceedings of the 4th ISCA Workshop on Speech Synthesis*, Perth, Australia, 2001.

[15] F. Deprez, J. Odijk, and J. D. Moortel, "Introduction to multilingual corpus-based concatenative speech synthesis," in *Proceedings of the Interspeech, 8th Annual Conference of the International Speech Communication Association*, pp. 2129–2132, Antwerp, Belgium, August 2007.

[16] H. Zen, N. Braunschweiler, S. Buchholz et al., "Statistical parametric speech synthesis based on speaker and language

factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1713–1724, 2012.

[17] H. Y. Wang, *Research on Statistical Parametric Mandarin-Tibetan Cross-Lingual Speech Synthesis*, Northwest Normal University, Lanzhou, China, 2015.

[18] L. Z. Guo, *Research on Mandarin-Xingtai Dialect Cross-Lingual Speech Synthesis*, Northwest Normal University, Lanzhou, China, 2016.

[19] P. W. Wu, *Research on Mandarin-Tibetan Cross-Lingual Speech Synthesis*, Northwest Normal University, Lanzhou, China, 2018.

[20] W. Zhang, H. Yang, X. Bu, and L. Wang, "Deep learning for Mandarin-Tibetan cross-lingual speech synthesis," *IEEE Access*, vol. 7, pp. 167884–167894, 2019.

[21] B. Li, Y. Zhang, T. Sainath, Y. H. Wu, and W. Chan, "Bytes are all you need: end-to-end multilingual speech recognition and synthesis with bytes," in *Proceedings of the ICASSP*, Brighton, UK, May 2018.

[22] Y. Zhang, R. J. Weiss, H. Zen et al., "Learning to speak fluently in a foreign language: multilingual speech synthesis and cross-language voice cloning," 2019, https://arxiv.org/abs/1907.04448.

[23] Z. Y. Qiu, D. Qu, and L. H. Zhang, "End-to-end speech synthesis based on WaveNet," *Journal of Computer Applications*, vol. 39, no. 5, pp. 1325–1329, 2019.

[24] Y. Zhao, J. Yue, X. Xu, L. Wu, and X. Li, "End-to-end-based Tibetan multitask speech recognition," *IEEE Access*, vol. 7, pp. 162519–162529, 2019.

[25] R. Skerry-Ryan, E. Battenberg, Y. Xiao et al., "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proceedings of the International Conference on Machine Learning (ICML)*, Stockholm, Sweden, July 2018.

[26] Y. Wang, D. Stanton, Y. Zhang et al., "Style tokens: unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proceedings of the International Conference on Machine Learning (ICML)*, Stockholm, Sweden, July 2018.

[27] A. V. D. Oord, S. Dieleman, H. Zen et al., "WaveNet: a generative model for raw audio," 2016, https://arxiv.org/abs/1609.03499.