Tibetan speech synthesis based on an improved neural network

Yuntao Ding[1,3,4,*] , *Rangzhuoma* Cai [1,2,3,4], and *Baojia* Gong [1,3,4]

[1]College of Computer Science and Technology, Qinghai Normal University, Qinghai Xining 810016, China
[2]School of Computer Science and Technology, Southwest Minzu University, Sichuan Chengdu 610041, China
[3]Tibetan Information Processing and Machine Translation Key Laboratory of Qinghai Province,Qinghai Xining 810008, China
[4]Key Laboratory of Tibetan Information Processing, Ministry of Education, Qinghai Xining 810008, China

Abstract. Nowadays, Tibetan speech synthesis based on neural network has become the mainstream synthesis method. Among them, the griffin-lim vocoder is widely used in Tibetan speech synthesis because of its relatively simple synthesis.Aiming at the problem of low fidelity of griffin-lim vocoder, this paper uses WaveNet vocoder instead of griffin-lim for Tibetan speech synthesis.This paper first uses convolution operation and attention mechanism to extract sequence features.And then uses linear projection and feature amplification module to predict mel spectrogram.Finally,use WaveNet vocoder to synthesize speech waveform. Experimental data shows that our model has a better performance in Tibetan speech synthesis.

1 Introduction

The speech synthesis method based on neural network greatly reduces the error rate of speech synthesis because the neural network unit has independent learning and back propagation capabilities, and the synthesized speech is closer to the human voice. Therefore, the method of speech synthesis based on neural network has become the mainstream method of speech synthesis in the world [1, 2, 3, 4].

As an important part of Chinese information processing, Tibetan speech synthesis is also the key and difficulty of Tibetan intelligent human-computer interaction. Although it started late, it has gradually from the wave-splicing-based Tibetan speech synthesis [5] and the statistical parameter-based Tibetan speech synthesis [6] into Tibetan speech synthesis based on neural network[7,8].In 2019, the literature[7] first proposed speech synthesis based on neural networks, which brought Tibetan speech synthesis into a new era.

Based on the literature [7], this paper proposes a Tibetan speech synthesis method based on improved neural network.By constructing an improved neural network,using WaveNet

* Corresponding author: 2498775654@qq.com

vocoder [9] to synthesize Tibetan speech.Subjective and objective experiments show that our model has a better performance in Tibetan speech synthesis.

2 Improved neural network structure

Due to Ando Tibetan has no tonal characteristics [10], and there are similar pronunciations in the 30 consonants, such as ཅ and ཇ, ཀ and ཁ, etc.In order to better distinguish similar pronunciations and make the synthesized Tibetan language more natural, this paper proposes an improved neural networks for Tibetan speech synthesis. The structure is mainly composed of three parts: sequence feature extraction module, spectrum prediction module and waveform synthesis module. Among them, the sequence feature extraction module extracts sequence feature information by performing a convolution operation on the preprocessed Tibetan word vector and assigning attention weight to it. The spectrum prediction module predicts the spectrum characteristics by performing nonlinear transformation on the characteristic information and using linear projection and convolution operations. The waveform synthesis module uses the self-return characteristics of WaveNet vocoder to recover the phase information, and then synthesize the speech waveform. The specific model components of this paper are shown in Figure 1 below:

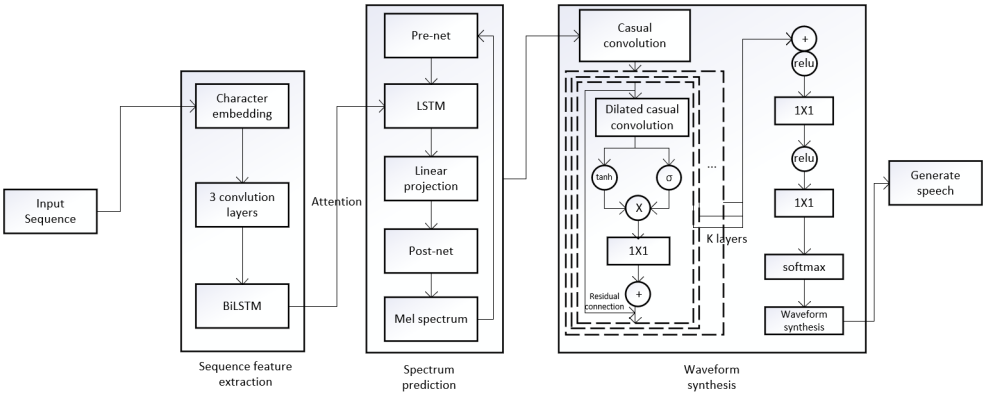

Fig. 1. Improved neural network.

2.1 Sequence feature extraction

Sequence features are indispensable to the speech synthesis process. Therefore, this paper first uses character embedding to preprocess the sequence, then uses 3 convolution layers to initially extract sequence features, and finally uses attention mechanism to assign corresponding weights to sequence features to complete sequence feature extraction.

2.2 Spectrum prediction

Considering mel spectrogram close to human auditory system, and as the lower-layer acoustic characteristics of the audio signal, it is more direct in speech synthesis [11]. Therefore, this paper chooses mel spectrogram as the spectrum feature to achieve spectrum prediction.

In this paper, an autoregressive neural network is used to achieve multi-frame prediction of the spectrum. The main steps are as follows:

1) Predict a frame of spectrum vector through linear projection of the Sequence feature matrix;

2) Pass the spectrum vector into the post-net to amplify useful spectral feature information;

3) Pass the spectrum vector into the pre-net to achieve nonlinear transformation;

4) Combine spectrum matrix and sequence feature matrix as a new sequence feature matrix;

After completing step 4), return to step 1) and repeat the steps until the mel spectrogram prediction is complete.

2.3 Waveform synthesis

Compared with the griffin-lim vocoder to achieve Tibetan speech synthesis, the speech waveform is smoother and close to the original sound waveform. Therefore, this paper uses WaveNet vocoder in waveform synthesis.

The internal structure of WaveNet is composed of causal convolution and one-dimensional convolution layer and dilated convolution layer and various gated activation functions (tanh, sigmoid, relu). Among them, causal convolution and dilated convolution are important components in WaveNet. Causal convolution ensures the timing of spectrum information, and dilated convolution can improve the receptive field of spectrum convolution. The following briefly introduces the specific process of WaveNet.

First, it generates a new spectrum matrix from the predicted spectrum matrix through causal convolution, and then passes the spectrum matrix through dilated convolution and a series of gated activation functions to effectively make the neural network perform coarse-grained convolution. Secondly, use its own autoregressive characteristics to recover the lost phase information. Finally, the posterior probability of sampling points is output through the softmax function [12]. Among them, the autoregressive characteristic is to predict the t sampling point through previous t-1 sampling points, and its formula (1) is as follows:

$$p(x) = \prod_{t=1}^{T} p(\boldsymbol{x}_t | \boldsymbol{x}_1, \dots, \boldsymbol{x}_{t-1}) \tag{1}$$

Therefore, by repeating the above formula, the waveform generated in the past is sampled, and the phase information is reconstructed through the WaveNet structure.

3 Experiments

The experimental corpus uses the corpus of the Key Laboratory of Tibetan Information Processing of Qinghai Normal University, which contains 2400 sentences of professional Tibetan female voices, the sampling rate is 16000hz, the sampling accuracy is 16bits, Use the Hamming window, the frame length is 50ms, the frame shift is 12.5ms, and the gradient descent optimizer uses the Adam optimizer. Set beta1 to 0.9, beta2 to 0.999, and epsilon to 1e-6. The WaveNet layer descent rate is set to 0.05, the exponential moving average decay rate is set to 0.9999, and the number of training steps is set to 100000 steps. The dimension of the word vector in sequence feature extraction is set to 512 dimensions, the size of the convolution kernel of the one-dimensional convolution layer is set to 5*1, stride is 1, and the padding is 2. The post-net layer uses 5 layers of one-dimensional convolutional layers and uses residual connection, and the first four layers are activated by relu.

3.1 Objective experiment

Figure 2 below depicts an image data comparison between the number of training steps and the percentage error. The model fit can be measured objectively through the training/error graph. It can be clearly seen that as the number of training steps increases, the percentage error is also continuously reduced.

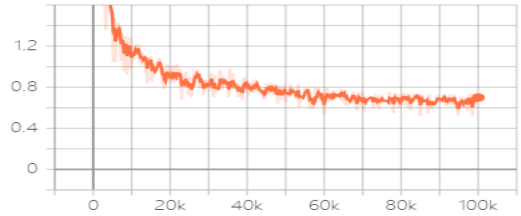

Fig. 2. Training steps/percentage error graph.

In order to evaluate the quality of the model, this paper compares the baseline model[7]to synthesize the sentence "འཇོག་པར་མགོ་བའི་ཡོ་བྱད་མང", the mel spectrogram is shown in Figure 3,4,5.

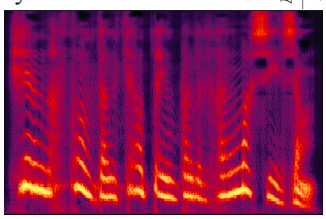

Fig. 3. Baseline model.

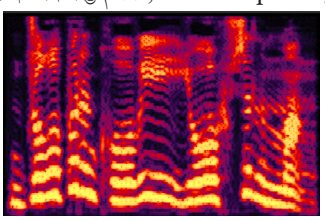

Fig. 4. Our model.

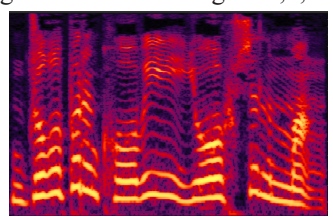

Fig. 5. Original sound.

Figure 3 is the baseline model, Figure 4 is our model, and Figure 5 is the original sound. It can be clearly seen that our model is clearer than the baseline model in terms of the sharpness of the formant. There are more detailed descriptions in the high-frequency area, and closer to the original sound.Therefore,our model has a better performance than the baseline model of Tibetan speech synthesis.

3.2 Subjective experiment

Through the subjective experiment comparison between our model and the baseline model, I found 5 random test sentences,according to the MOS scoring standard (0-5 points) and passed 10 different professional evaluators to score, and finally calculated the average score through these results. The results are as follows 3 shows:

Table 1. MOS score

Sentences	Our model	Baseline model
དབྱིན་ཇི་ནས་ཚན་རིག་པ་ཚེན་པོ་ཉིའུ་ཏུན་ཟེར་བ་ཞིག་ཡིན	4.2	3.6
ཉེན་ཁང་བོར་ལག་ཤེས་བཙོ་ལས་བྱེད་པ་དང	4.0	3.5
ཨ་ཅི་ལགས	3.9	3.7
འཇོག་པར་མགོ་བའི་ཡོ་བྱད་མང	4.0	3.4
སྐ་ལ་སྟོང་དང་ཡོད་ཚེན་ནེད	3.9	3.5
Average score	4.0	3.54

It can be seen from the Table 1 that our model has a better MOS score than the baseline model. The evaluators believe that the clarity and naturalness of our model are better than the baseline model.

4 Summary

Based on the literature [7], this paper uses three one-dimensional convolutional layers replace the complicated CBHG module to extract sequence features, and adds post-net to further process the mel spectrogram to make the prediction results more realistic, and finally uses WaveNet instead of griffin-lim vocoder synthesis Tibetan voice. Subjective and objective experiments show that our model has a better performance in Tibetan speech synthesis.

Compared with Chinese and English corpus, Tibetan corpus is very scarce, and we will consider expanding Tibetan corpus in the follow-up work. Considering that the training rate of WaveNet vocoder is very slow,we will consider improving the vocoder model that can be synthesized in parallel, such as WaveGlow [13].

This research was financially supported by the National Natural Science Foundation of China (61966031,61866032); Ministry of Education "Chunhui Program" (Z2016077, Z2012093); Qinghai Province Science and Technology Project (2019-SF-129); Qinghai Province Key Laboratory Project (2014-Z-Y32, 2015-Z-Y03); Key Laboratory of Tibetan Information Processing and Machine Translation (2013-Y-17).

References

1. P Wang,et al.Word embedding for recurrent neural network based TTS synthesis[C]//Proceedings of IEEE International Conference on Acoustics,2015:4879-4883.
2. Lipton Z C,Berkowitz J,Elkan C.A critical review of recurrent neural networks for sequence learning[J].arXiv preprint arXiv:1506.00019,2015.
3. Zen H,Sak H.Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis[C]//Proceedings of IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP),2015:4470-4474
4. Naihan Li, Shujie Liu, Yanqing Liu,et al. Neural Speech Synthesis with Transformer Network[J]. arXiv preprint arXiv:1809.08895v3,2019.
5. CAI Rangzhuoma, CAI Zhijie. Unit Selection Algorism for Corpus-based Tibetan Speech Synthesis[J]. JOURNAL OF CHINESE INFORMATION PROCESSING,2017,31(5):59-63.
6. Zhou Yan,Zhao Dongcai. RESEARCH ON HMM-BASED TIBETAN SPEECH SYNTHESIS[J].COMPUTER APPLICATIONS AND SOFTWARE,2015,32(5):171-174.
7. DOU Gecao,CAI Rangzhuoma,NAN Cuoji,SUAN Taiben. Neural Network Based Tibetan Speech Synthesis[J].JOURNAL OF CHINESE INFORMATION PROCESSING,2019,33(02):75-80.
8. DOU Gecao.Research on Neural Network Based Tibetan Speech Synthesis Technique[D].QingHai Normal University,2019.
9. Aäron van den Oord,Sander Dieleman,Heiga Zen,et al.WaveNet:a generative model for raw audio[J]. arXiv preprint arXiv:1609.03499v2,2016.
10. Hua Kan. Several Special Variations of the Consonants in Ando Tibetan [J]. Minority Languages of China,1983(03):43-46.
11. QIU Zeyu,QU Dan,ZHANG Lianhai. End-to-end speech synthesis based on WaveNet[J].JOURNAL OF COMPUTER APPLICATIONS,2019,39(05):1325-1329.

12. GUNDUZHAN E, MOMTAHAN K.Linear prediction based packet loss concealment algorithm for PCM coded speech[J].IEEE Transactions on Speech and Audio Processing,2001,9(8):778-785.
13. Ryan Prenger,Rafael Valle,Bryan Catanzaro.WaveGlow:A Flow-based Generative Network for Speech Synthesis[J].arXiv preprint arXiv:1811.00002v1,2018.