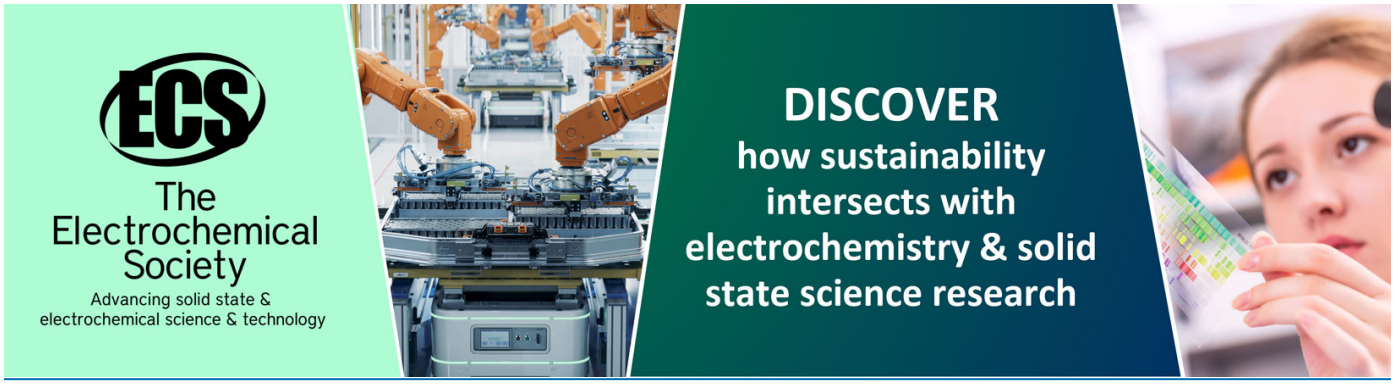
End-to-end Speech Synthesis for Tibetan Lhasa Dialect



IOP Conf. Series: Journal of Physics: Conf. Series **1187** (2019) 052061 doi:10.1088/1742-6596/1187/5/052061

End-to-end Speech Synthesis for Tibetan Lhasa Dialect

Lisai Luo[1], Guanyu Li[1*], Chunwei Gong[1], Hailan Ding[1]

[1]Key Laboratory of National language Intelligent Processing Gansu Province, Northwest Minzu University, Lanzhou, China

*guanyu-li@163.com

Abstract. Speech synthesis for Tibetan Lhasa dialect is implemented on the basis of an end-to-end novel speech synthesis framework, Tacotron. The training transcript has used the phoneme list transcribed from Tibetan characters, and feature parameters were extracted from the mel-spectrogram. Then the model is trained by the mapping of character to spectrum. Tibetan language is an important minority language of the Chinese nation, but there is little research on Tibetan language at present. The experimental results were compared with traditional speech synthesis methods, with the audio quality significantly better than that of the traditional GMM-HMM in both naturalness and rhythm. It provides a crucial reference for the later research methods of Tibetan language and promotes the development of Tibetan language research.

1. Introduction

Tibetan language is spoken by about 6 million people mainly distributed in 5 districts in China, including Tibet, autonomous region and Qinghai province, Sichuan Ganzi Tibetan autonomous prefecture, Aba Tibetan and qiang autonomous prefecture, as well as in Gannan Tibetan autonomous prefecture and Diqing Tibetan autonomous prefecture in Yunnan province. Some residents lived in Bhutan, India, Nepal and Pakistan also make Tibetan as their mother language. Tibetan is an influential language with long history. In recent years, more and more attention has been paid to Tibetan research and application to improve the level of informatization of Tibetan. There are 3 Tibetan dialect areas in China according to the characters of Tibetan dialects: Weizang, Kangba and Amdo. There are several dialects in each area, dialect in the same area is more similar to the dialects in other areas. As an important and influential dialect in Weizang, Lhasa dialect is chosen as the research object in this paper.

Research on speech recognition and speech synthesis in Tibetan Lhasa dialect has made progress in recent years. In this paper, speech synthesis of Lhasa Tibetan is implemented based on a novel end-to-end speech synthesis framework, Tacotron, proposed by Google in early 2017. A text-to-speech synthesis system typically consists of multiple stages, such as a text analysis frontend, an acoustic model and an audio synthesis module. Building these components often requires extensive domain expertise and may contain brittle design choices. However, Tacotron is an end-to-end generative text-to-speech model and it can synthesizes speech directly from characters [1].

The section 2 introduces the overall structure of Tacotron, and the section 3 introduces the data preparation and related work, as well as the design of corpus data. Then the experimental results are shown in the section 4. The section 5 is the summary and prospect.

2. Model architecture

This part mainly introduces the overall architecture of Tacotron model and the CBHG module structure used in encoder and decoder module.

2.1. Tacotron model architecture

The Tacotron framework is a relatively novel end-to-end TTS model. The model input is a text character, and the output is a spectrum diagram parameter. Finally, the corresponding audio is generated using the Griffin-Lim algorithm. Figure 1 depicts the model, which consists of an encoder, a decoder, two CBHG modules, and a post-processing network. The input (pre processed text and audio) is fed into an encoder, and generates attention features which are then used in every step of the decoder before generating spectrograms.

Another fascinating mechanism used in Tacotron is called attention, which is used in every step of coding and decoding. At present, the model based on attention mechanism has been widely used in machine translation, speech synthesis, speech recognition and computer vision. Attention mechanism is used in the speech synthesis model to realize end-to-end speech synthesis. Attention mechanism has great promotion in the sequence of learning tasks at decoder framework, through the model for the attention of in the code segment. The source data sequence weighted data transformation or introduction of Attention at the decoding end model, weighted changes of target data, can effectively improve the system performance.

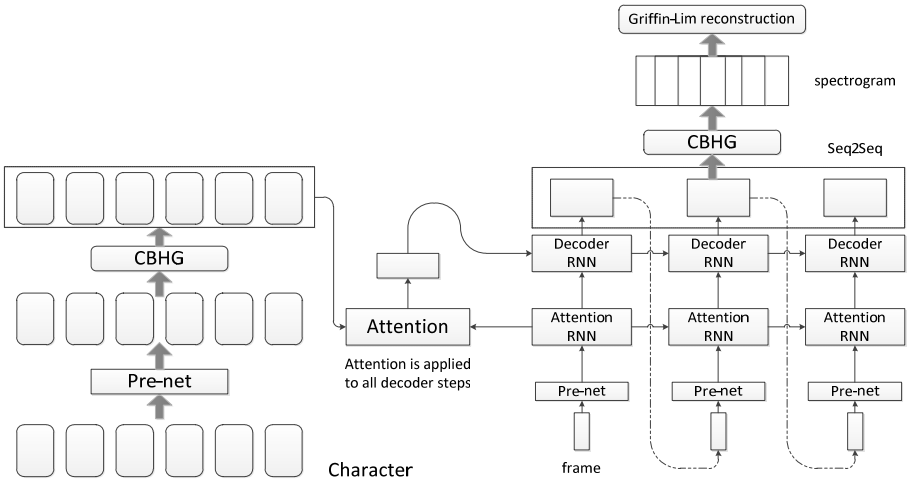


Figure 1. The Tacotron architecture adapted from paper [2].

2.2. CBHG module

CBHG consists of a bank of 1-D convolutional filters, followed by highway networks and a bidirectional GRU (Gated recurrent unit) recurrent neural network (RNN) [2]. Its function is extracting valuable features from the input, which is beneficial to improve the generalization ability of the model. Figure 2 depicts its structure.

The CBHG module contains a one-dimension convolution kernels, followed by a Highway network and a bidirectional GRU. CBHG is a powerful module for extracting representations from sequences. The input sequence is first convolved with K sets of 1-D convolutional filters. And there is a maximum pooling operation after each convolution layer. During the encoding and decoding, reduces more training time because of without using the time-consuming RNN structure. Batch normalization is used for all convolution layers in this module. The output of convolution layer is fed into a multilayer highway network to extract high-level features. At the top layer of the model, a bidirectional GRU is stacked to extract sequence features bidirectionally. By training, it is found that using CBHG module can effectively improve the generalization ability of the model.

Highway Networks [4] is a new kind of neural network structure. The traditional neural network structure is much better in deep layer than in shallow layer. However, the deep neural network makes model training more and more difficult, while Highway can use simple SGD method to train extremely deep network, and with the increase of network depth, the network can be optimized even if the initialization variables remain unchanged. This is achieved through a gate mechanism that controls the flow of information through the neural network. Through this mechanism, the neural network can provide a pathway to allow information to pass through without loss.

The input x is converted into y output by the activation function H in the traditional neural network, and w is the weight in formula(1):

$$y = H(x, W_H) \tag{1}$$

Highway Networks of neural network, increased the two nonlinear transformation layer, one is T (transform gate) and a is C (carry gate), T represents the input information through convolutional or recurrent converted part of C said x retain part of the original input information in formula(2):

$$y = H(x, W_H) \bullet T(x, W_H) + C(x, W_H) \tag{2}$$

To simplify, replace C with 1 minus T in formula(3):

$$y = H(x, W_H) \bullet T(x, W_H) + x(1 - T(x, W_H)) \tag{3}$$

The dimensions of x, y, H and T must be consistent. Sub-sampling or zero-padding strategies can be adopted to make the dimensions consistent. Several formulas are compared, and formula (3) is more flexible than formula (1), and there is a special case, for example, when T= 0, y = x, the original input information is all retained without any change, when T = 1, Y = H, the original information is all converted, and the original information is no longer retained, just equivalent to an ordinary neural network. As shown in formula (4):

$$y = \begin{cases} x & if \quad T(x, W_H) = 0 \\ H(x, W_H) & if \quad T(x, W_H) = 1 \end{cases} \tag{4}$$

3. Experiments and analysis

Data for Lhasa dialect speech synthesis is prepared and preprocessed. Several experiments schemes are designed to testify the performance of end-to-end Lhasa dialect speech synthesis under various conditions.

3.1. Data preparation

A large amount of linguistic data is necessary to study end-to-end speech synthesis for Lhasa dialect. The speech corpora are recorded on PC in quiet environment and saved as PCM wave files, where the sampling rate is 16kHz, the sample size is 16 bits, and the vocal tract is mono type. All the speakers are in Lhasa dialect, and the recording style is reading. Transcripts in training are saved as files with suffix "mlf" or "txt". However, there are various ways to depict input sequences of transcript according to different languages. For example, in English, 26 letters with punctuation transcript can be used directly as the input sequence. The Korean language has its own set of alphabets, each of which can use Unicode code as its transcript character. For Mandarin, there are about more than 3,000 often-used characters, it is too complicated to enumerate all characters in the system. And Among which there are many homophones, so the Chinese pinyin is commonly used to transcript them. Tibetan scripts are written in alphabets.

From view of written form, there are 30 consonant letters and 4 vowel signs in Tibetan (note all dialects are the same in writing). Each syllable is a combination of several consonant letters and a vowel sign. Words are comprised of one or several syllables. Each syllable involves a radical consonant letter, and other consonant letters could be appended to the radical consonant as superscript, subscript, prescript, postscript and post-postscript to form a syllable. Constitution of syllable described in figure 3. The

radical consonant, the prescript and superscript consonants together form the initial part of a syllable, and the vowel sign, the postscript and post-postscript consonants altogether form the final part. The radical consonant, the prescript and superscript consonants together form the initial part of a syllable, and the vowel sign, the postscript and post-postscript consonants altogether form the final part.

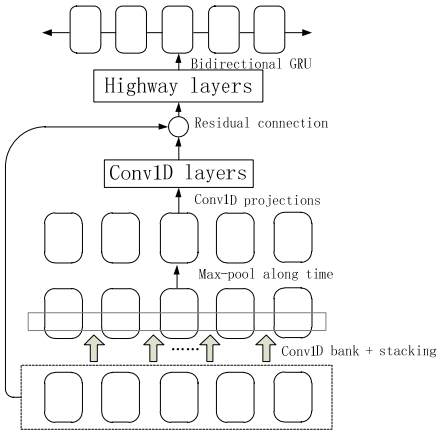


Figure 2. The CBHG module adapted from paper [2].

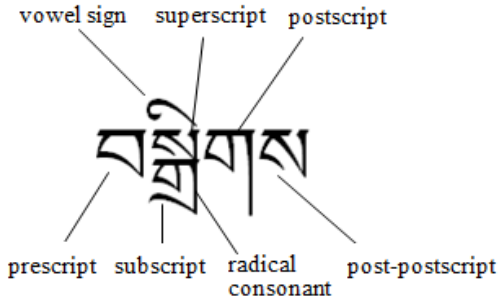


Figure 3. Constitution of syllable.

A syllable lexicon involves a set of syllables whose pronunciations are defined. There are more than 8,000 possible syllables in Tibetan, including syllables for foreign words. We construct the syllable lexicon by constructing a text corpus involving 420,000 sentences (including both written and spoken), and then selected the most frequent syllables from this corpus. After removing some syllables that are for transliterating Sanskrit words only, we obtained a syllable lexicon consisting of 6,013 syllables. By applying the pronunciation rules, these syllables were segmented into initials and finals, and the initials and finals were further split into phones [3]. All these syllables and their phone sequence forms were manually checked to ensure the quality. Therefore, a set of standard rules is given in the experiment, which converts the international phonetic symbols of Tibetan characters into a list of phonemes which can be written and adjusted conveniently. Sentences are transformed into list of phonemes on the basis of pronunciation dictionary. Then the list of phonemes is input to the model as input sequence.

For example, there is a sentence in the data set:

ཚངས་དུ་ཐུབའི་གདུལ་བྱ་ཐེག་དམན་གྱི་རིགས་ཅན་ལ་གསོ་བོར་དགོངས་ནས་གསུངས་པའི་ཐེ་སྐོང།

The sentence is transformed into a list of phonemes:

tjh eb th u tjh a w el t yw tjh a th e k m elu c h i r i k tj elu l a ts o ph ow k o ng n elb s u ng p el t e n eyb.

Sentences can also be transformed into lists of initials and finals. As mentioned above, all the texts are converted to the form of above phoneme transcript before loading the data, which is done by converting it to a file with the suffix "mlf" or "txt".

3.2. Experiments settings

Various experimental schemes are designed to testify the performance of Lhasa dialect under various conditions to find a best scheme to implement the final speech synthesis system.

- 8,000 sentences spoken by 21 male speakers are chosen to train the model.
- 13,000 sentences spoken by 21 male speakers and 13 female speakers are chosen to train the model.
- 20,000 sentences spoken by 23 female speakers are chosen to train the model.
- 5,000 sentences spoken by one male corpus were used for training at the beginning. After 352k iterations training, the female corpus was added to continue training this model. There are more than 5,000 sentences spoken by one male and 20,000 spoken by 23 female speakers.

4. Experimental Results

The part of the experimental results is shown here and only the third experiment is shown in this paper. We use synthetic spectrum to compare with original spectrum. Figure 4 is the original spectrum diagram, Figure 5, 6, 7 respectively are iteration of 795k, 860k and 1000k training times.

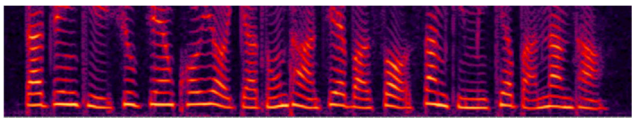


Figure 4. original spectrum diagram

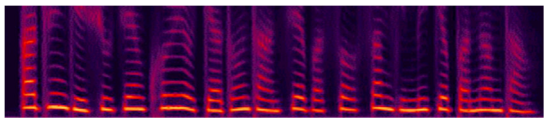


Figure 5. iteration 795k times

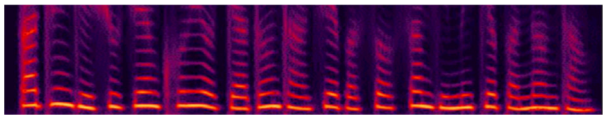


Figure 6. iteration 850k times

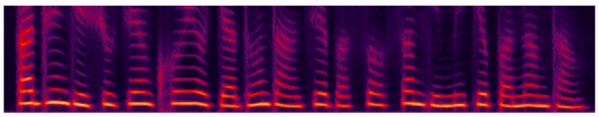


Figure 7. iteration 1000k times

Our experimental results show that more corpus and the same sex data training model will be more effective than less corpus and different sex, however, the generalization ability of different gender training is better. The result of End-to-end Speech Synthesis for Tibetan Lhasa Dialect is outperforms the traditional GMM-HMM system, and we will continue to optimize this model.

5. Conclusions

The above is the research progress up to now, and a good experimental result has been found through design and experiment. However, the structure of Tibetan character is two-dimension, its writing particularity restricts the flexibility of processing, so it must be transcript. Tacotron model structure is one of the classical structures in speech synthesis technology. The model will be adjusted to make the training speed more optimized in the next study. We will also try to use different research methods to further study speech recognition and speech synthesis, and apply it to Tibetan language and different languages. Further, the technology has achieved a breakthrough.

Acknowledgments

Supported by the Fundamental Research Funds for the Central Universities (Research on Unified Acoustic Models of Mandarin and Tibetan, 31920170145)

References

[1] K. Park, “A TensorFlow implementation of Tacotron: A fully end-to-end text-to-speech synthesis model,” 2017, Available at GitHub, https://github.com/ Kyubyong/tacotron(2018).
[2] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., & Jaitly, N., et al. (2017). Tacotron: towards end-to-end speech synthesis. 4006-4010.
[3] Guanyu Li et al, “Free Linguistic and Speech Resources for Tibetan”, (ASC 2017)
[4] R. Kumar Srivastava et al, “Highway Networks”, (ICML 2015)
[5] H. Tachibana et al., “Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention”, (ICASSP 2018)
[6] I. Goodfellow et al., Deep Learning, MIT Press, (2016)
[7] J. Sotelo et al., “Char2wav: End-to-end speech synthesis”, (ICLR, 2017)
[8] K. Tokuda et al., “Speech Synthesis Based on Hidden Markov Models”,(*Ipsj Magazine* ,2013)
[9] J. Lee, K. Cho, T. Hofmann. Fully CharacterLevel Neural Machine Translation without Explicit Segmentation. Transactions of the Association for Computational Linguistics, vol. 5, pp. 365–378, 2017.
[10] W. Ping, K. Peng, Andrew Gibiansky et al. Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning. Published as a conference paper at ICLR 2018.
[11] T. Le Paine, P. Khorrami et al. Fast Wavenet Generation Algorithm. In Technical Report 2016.

[12] J. Shen, R Pang, et al. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. Accepted to ICASSP 2018
[13] J. Engel, C. Resnick, A. Roberts et al. Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. Accepted to ICML 2017.
[14] A. van den Oord, S. Dieleman et al. WaveNet: A Generative Model for Raw Audio. in 2016.
[15] Qin Y, Song D, Chen H, et al. A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction[J]. 2017:2627-2633.
[16] Torfi A, Shirvani R A, Soleymani S, et al. Attention-Based Guided Structured Sparsity of Deep Neural Networks[J]. 2018.
[17] Hudson D A, Manning C D. Compositional Attention Networks for Machine Reasoning[J]. 2018.