End-to-end Tibetan Speech Synthesis Based on Phones and Semi-syllables

Guanyu Li*, Lisai Luo , Chunwei Gong and Shiliang Lv
Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education,
Northwest Minzu University, Lanzhou, Gansu 730000, China
*E-mail: guanyu-li@163.com Tel: +86-13809316272

***Abstract*—Due to the 2D architecture of Tibetan characters, it is not convenient to treat the letters sequences as the input of the end-to-end speech synthesis system. The experiments are conducted based on phones and semi-syllables sequences respectively. In training and testing, the text is segmented into a sequence of syllables first, then syllables are transformed into phones and semi-syllables as the input sequence of the model. The results demonstrate the encoding and decoding alignment effect of Tibetan speech synthesis based on phones is better than that based on semi-syllables. In addition, the Highway network in the architecture plays a key role in the convergence of the model.**

I. INTRODUCTION

Over the last two years, the end-to-end method has been used in speech synthesis more widely. Traditional statistical parametric text-to-speech (TTS) pipelines are complex. It is common for it to have a text frontend extracting various linguistic features, a duration model, an acoustic feature prediction model and a complex signal-processing-based vocoder et al. The complexity leads to more errors. Different from traditional parametric speech synthesis system, the end-to-end based text-to-speech model synthesizes speech directly from characters. Given<text, audio> pairs, the model can be trained completely from scratch with random initialization. An end-to-end model could allow us to train on huge amounts of rich, expressive yet often noisy data found in the real world [1].

Tibetan language is a key member in the family of minor languages in China. It belongs to the Sino-Tibetan language family, the Tibeto-Burman subgroup. With a long history and profound culture, Tibetan is one of the oldest ethnic groups in China and South Asia. However, there is little research on Tibetan speech synthesis technology compared with that of mandarin Chinese and other languages. Tibetan phonetic technology is still in the early stage of development because of the lack of absolute standards, data, literature and human resources. To Tibetan speech synthesis, it's more difficult to construct a text frontend. So it's an efficient way to improve the quality of Tibetan synthesis by use of end-to-end method. Because of the peculiarity of spelling of Tibetan language, phones and semi-syllables are treated as the modelling units respectively. End-to-end Lhasa Tibetan system based on phones and semi-syllables are implemented. The work is based on the state-of-the-art speech synthesis tool Tacotron proposed by Google in 2017[1].

II. END-TO-END SPEECH SYNTHESIS

The model architecture of Tacotron consists of an encoder, a decoder and a post-processing net. The encoder and decoder are connected by the attention mechanism. Traditionally, the inputs of Tacotron are text characters. The outputs are the corresponding original spectrum graph parameters. Finally, a post-processing network is added, and the Griffin-Lim reconstruction algorithm generates the corresponding audio with the generated spectrum graph parameter sequence. The overall architecture of the Tacotron model is shown in Fig 1. The left part is the encoding module, the right part is the decoding module, and the upper right part is the model post-processing network. There is a two-layer pre-net under CBHG both in encoding and decoding module.

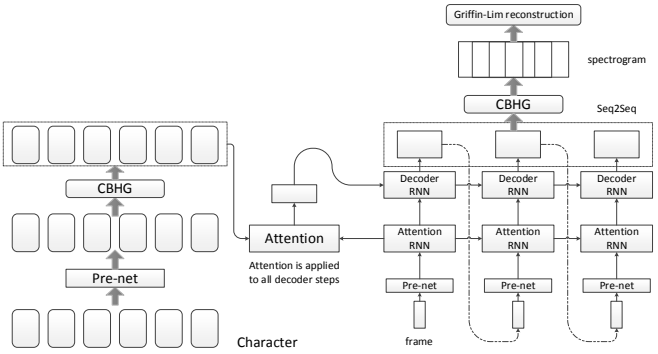

Fig 1 Architecture of Tacotron

A CBHG module consists of a one-dimensional convolution filter banks, Highway Networks and Bi-directional gated recurrent unit (BiGRU). It can extract valuable features from the inputs to improve the generalization ability of the model further. The CBHG is an effective module in feature expression of extracted sequences. Each convolution layer is followed by a maximum pooling operation, which can reduce training time. Outputs of convolution layer are added with the original input sequence via residual connections. The convolution outputs are fed

into a multi-layer highway network to extract high-level features, and then highway networks outputs are fed into Bi-directional GRU to the extract the sequential features both in forward and backward context. The architecture of CBHG module is shown in Fig 2.

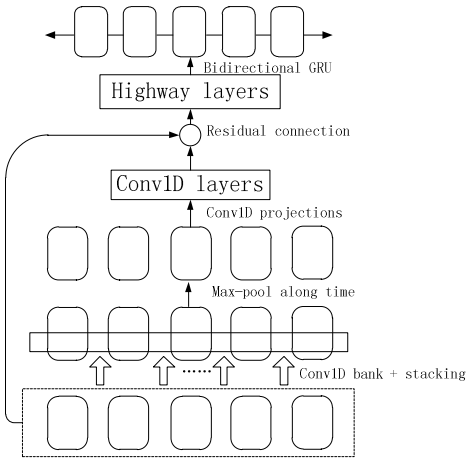

Fig 2 Architecture of CBHG

The goal of the encoder is to extract a robust sequential representation of the text. The encoder takes the sequence of characters as input, which are represented as one-hot vector sequence. Then a set of nonlinear transformations are done by pre-net layers. A dropout bottleneck layer is used in the architecture as a pre-training network, which help speed up convergence and improve generalization. The CBHG module converts the output of pre-net into the final representation of the encoder by use of the attention module. Encoder based on CBHG not only reduces overfitting, but also has fewer pronunciation errors than standard multilayer cyclic neural network encoders.

In decoding module, attention-based decoder is used at each decoding time step. The input to the decoder's neural network is formed by connecting the context vector and the output of the attention mechanism neural network. A simple fully connected output layer is used to predict the target of the decoder and multiple non-overlapping output frames are predicted in each decoder step.

III. FEATURES OF TIBETAN LANGUAGE

There are 3 Tibetan dialect areas in China: U-Tsang, Amdo and Kham. People in the three areas use the same written form, but pronounce very differently. In U-Tsang, the most popular dialect is the Lhasa Tibetan. Moreover, Lhasa city is the capital of Tibet Autonomous Region. So Lhasa Tibetan is chosen as the research object.

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. In the Tibetan script, syllables and words are written from left to right, and are separated by the same delimiter " · " (called ཙེག (/tsheg/) in Tibetan) [2].

Tibetan syllables (characters) are 2-dimensional in structure. Each syllable involves one and only one radical consonant letter, and other consonant letters could be appended to the radical consonant optionally as superscript, subscript, prescript, postscript and post-postscript to form a syllable (Fig 3). A syllable must contain a vowel sign, but a vowel sign corresponds to a sound /a/ can be omitted. In general, the vowel signs ི, ུ, ེ, ོ sound /i/, /u/, /e/, /o/ respectively, but exceptions also exist, as their pronunciations can be changed following some regular rules. Note that in all the dialects of Tibetan, two syllables may be pronounced the same but each syllable has only a single pronunciation. In other words, there are many homophones but no polyphones in Tibetan. The radical consonant, the prescript, subscript 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.

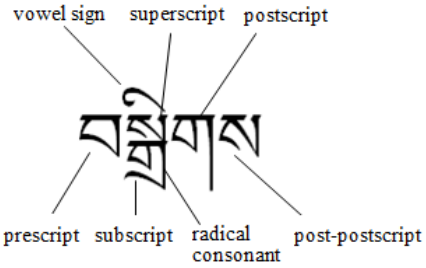

Fig. 3 Constitution of syllable.

IV. EXPERIMENT AND ANALYSIS

Compared with the traditional statistical parameter speech synthesis technology, the end-to-end mechanism does not need to analyze text, acoustics and prosody. In theory, text sequences and spectral parameter sequences can be treated as inputs in end-to-end speech synthesis mechanism directly.

Due to the 2-D structure of syllables in Tibetan, it is not convenient to treat sequence of letters as input of the model. The number of possible syllables in Tibetan is huge, so syllables are not suitable to be the inputs of the models either. There is no polyphone in Tibetan. Text can be conveniently transformed into phones or semi-syllables sequence by use of pronunciation lexicon. In addition, Tibetan is a phonetic language, phones and semi-syllables sequences can not only depict the pronunciation but also the spelling of the words. So phones and semi-syllables are chosen as the modeling units to form the input sequences.

A. Data Preparation

By applying the pronunciation rules of Lhasa Tibetan, 6013 often-used syllables are transformed into phones lists represented by IPA marks as demonstrated in Tab 1. The lexicon is manually checked to ensure the quality. Then IPA marks are segmented into initials and finals, then the initials

and finals represented by IPA marks. Thus, lexicon dictionaries based on phones and semi-syllables of syllables are created respectively. In this experiment, the IPA symbols in lexicons are converted into Latin symbols which can be used conveniently in the programs on the basis of IPA2phone list. The list of IPA2phone is demonstrated in Tab 2.

Tab 1 Syllables List of Lhasa Tibetan

Tibetan Syllable	IPA
ཀ	ka
ཀང	kaŋ
ཀེན	kɛn
ཀབ	kap
ཀམ	kam
ཀེ	ke:
ཀར	kar

Tab 2 Phones and Latin Transformation of Lhasa Tibetan

IPA	Latin	IPA	Latin	IPA	Latin	IPA	Latin	IPA	Latin	IPA	Latin	IPA	Latin	IPA	Latin	IPA	Latin	IPA	Latin
c	c	kh	kh	ph	ph	tʂ	q	ɕ	x	tɕh	txh	e:	ew	o	o	yˀ	yb	ɛ	el
ch	ch	l	l	r	r	tʂh	qh	ʂ	ss	ˀ	ab	eˀ	eb	o:	ow	ø	f	ɛ:	elw
h	h	m	m	s	s	w	w	ʦ	ts	a:	aw	I	i	u	u	ø:	fw	ɛˀ	elb
j	j	n	n	t	t	ŋ	ng	ʦh	tsh	a	a	i:	iw	u:	uw	øˀ	fb	ẽ	eu
k	k	p	p	th	th	ɳ	nn	ʨ	tx	e	e	iˀ	ib	y:	yw	ĩ	il	ỹ	yu

To construct the speech database of Lhasa Tibetan. A maximum entropy method is used to select the most representative sentences. About 20,000 sentences were chosen from the original database with size of 100MB. After the collected text corpus is sorted, the speaker's voice collection is performed on the PC with the sampling rate of 16 KHZ, sampling width of 16-bit and monophonic.

20,635 sentences of audio signals read by 23 female speakers are used as training set. These speech corpora and the corresponding utterances are used as inputs in this experiment. Before training, sentences in text are segmented into syllables first. Then the syllable lists are transformed into lists of phones on the basis of phone lexicon. At the same time, the syllable lists are also transformed into lists of initials and finals on the basis of semi-syllable lexicon.

B. Experimental setting

In the experiment, the pre-net is a two-layered dropout DNN with 512 neurons and 256 neurons respectively, and the activation function of ReLU is used. The number of convolution channels is 256, and the sequence is convoluted through the convolution network. The convolution network is one-dimensional with the convolution kernels of size from 1 to k, and k is set to 16 in the encoding module and 8 in the decoding module. There is a maximum pooling operation after each convolution, and batch normalization is also used in the convolution process. The output of the convolution neural network is sent to Highway Networks neural Network, which is used to extract advanced features. The number of layers of Highway Network is 4. Output of Highway Network is sent to bi-directional GRU to extract sequence information. The number of layers of BiGRU is set as 2. The other hyper-parameters in the experiment are as follow: Sampling rate is 16khz. Fast Fourier point is 2048. Frame shift is 0.0125s. Frame length is 0.05s. Pre-emphasis is 0.97, and the number of Griffin-Lim iterations is 100.

C. Experiments and Results

Two experiments based on phones and semi-syllables are designed and implemented respectively. In the second experiment, Highway networks in the sequence-to-sequence structure are removed, and the final features are directly sent from the convolutional neural network to the two-way GRU without passing through the highway networks layer.

Data for training needs to be placed in specified directory and then be loaded and processed by data preprocessing module. To represent the sequence of phones or semi-syllables as a list index, that is, to digitally encode the phones or semi-syllables， the phones and semi-syllables are embedded as integers first.

A log file to save the model is created per 1000 iterations by API of storing and loading modules in TensorFlow. The checkpoint will be saved in a file when the program breaks abnormally. Then training procedure can continue from the checkpoint next time. The 5 newly generated models during training period are saved by default to avoid too much memory footprint.

During training, we need to observe log files, check the alignment status of encoding and decoding of attention, select the desired model for speech synthesis.

Fig. 4 is the log file which demonstrates the alignment status of the experiment based on phones by iterating for 555k times. Fig. 5 is the result of the experiment based on semi-syllable by iterating for 468k times. They are all the best alignment statuses in each experiment.

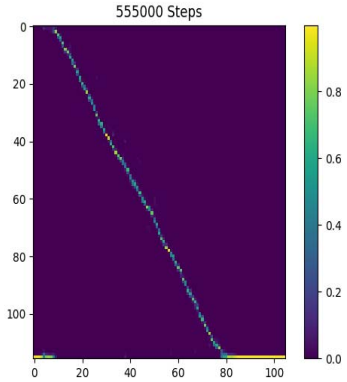

Fig. 4 iteration 555k times based on phones.

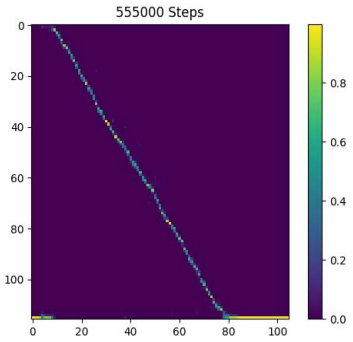

Fig. 5 iteration 468k times based on semi-syllables.

The final results of the experiment demonstrate that the alignment effect of Tibetan speech synthesis based on phones is better than that of Tibetan speech synthesis based on semi-syllables. MOS（Mean Opinion Score）is 4.0 in experiment based on phones, and 3.92 in experiment based on semi-syllables. In the Highway networks experiment, we observed the log file generated and found that the sequence-to-sequence model without Highway networks, and effect of convergence and alignment was much worse (Fig 6).

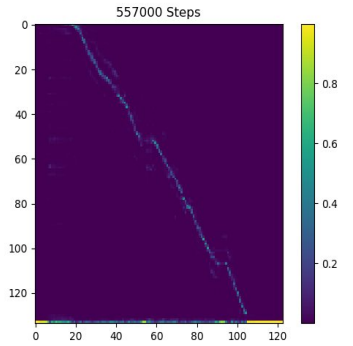

Fig. 6 iteration 557k times without Highway networks.

V. CONCLUSIONS

Due to the experiments results, phones are more suitable than semi-syllables to be used as inputs in Tibetan end-to-end speech synthesis. In the future research, it is hoped that the model can be adjusted to optimize the training speed. While optimizing the model, the recording quality of Tibetan language needs to be improved. At present, the training set consists of the recording files of only a few hundred sentences per speaker. There are a few accents in some speakers' pronunciation, so the data is not accurate enough. In future researches, it is hoped that it would be better to construct a recording database for an only speaker. Such a speech database can ensure the match between text and audio to improve the quality of synthesized speech.

ACKNOWLEDGMENT

Supported by the Natural Science Foundation of China under Project (Grant No. 61633013), Research Funds for the Central Universities (31920170145), Gansu Provincial First-class Discipline Program of Northwest Minzu University (NO.11080305). The authors would like to thank all the people who have helped us during the studying process of this work.

REFERENCES

[1] 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.

[2] Guanyu Li et al, *"Free Linguistic and Speech Resources for Tibetan"*, (APSIPA ASC 2017)

[3] 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.

[4] J. Shen, R Pang, et al. *Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions*. Accepted to ICASSP 2018.

[5] J. Engel, C. Resnick, A. Roberts et al. *Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders*. Accepted to ICML 2017.

[6] Madhavan, P. G. *"Recurrent neural network for time series prediction."* International Conference of the IEEE Engineering in Medicine & Biology Society IEEE, 2002.

[7] J. Sotelo et al., *"Char2wav: End-to-end speech synthesis"*, ICLR, 2017.

[8] Srivastava R K, Greff K, Schmidhuber J. *Highway Networks*. Computer Science, 2015.