End-to-End Speech Synthesis Method for Lhasa-Tibetan Multi-speaker

Xiaona Xu[1], Wanyue Ma[1], Zhengjia Dan[1], Huilin Ma[1](✉), Tianyi Liu[2], and Yue Zhao[1]

[1] Minzu University of China, Beijing 100081, China
ma.huilin@163.com
[2] Department of Computer Science and Engineering, Santa Clara University,
500 El Camino Real, Santa Clara, CA 95053, USA

Abstract. Tibetan text-to-speech generally focuses on a single speaker or a single dialect, and there is a lack of research on Tibetan multi-speaker speech synthesis. This paper explores the speech synthesis methods based on an end-to-end model for Lhasa-Tibetan multi-speaker. We propose to convert Tibetan characters into Latin letters to improve the effect of model learning. We compare the end-to-end model using the speaker ID embedded into the spectrogram feature prediction network against using some WaveNet vocoders trained on specific speaker data. Referring to the results of objective and subjective experiments, our method has better speech quality than the model using some WaveNet vocoders trained on specific speaker data.

Keywords: End-to-end speech synthesis · Multi-speaker · Lhasa-Tibetan dialect

1 Introduction

Speech synthesis technology converts text into speech, which generates a corresponding speech waveform for a given target text. There are various speech synthesis applications in man-machine interaction, assistive technology, media and entertainment. With the development of speech synthesis technology, multi-speaker speech synthesis has attracted more research interest of researchers on synthesizing different speakers' speech in a unified speech synthesis model [1–3].

As one of the Chinese minority languages, Tibetan is mainly divided into three major dialects: Ü-Tsang, Kham and Amdo. Although the characters of all dialects are the same, due to their different pronunciations, people in different regions have difficulties with communication. In addition, because of the lack of technical standards and language resources in Tibetan, its research is still in the developing stage.

For the research of Tibetan speech synthesis technology, there are generally three methods: statistical parameter modeling, speech synthesis unit selection, waveform splicing, and deep learning. However, the Tibetan speech corpus is still in the stage of construction and development. At the same time, traditional speech synthesis methods usually require researchers to have more linguistic knowledge, which is not conducive to

X. Sun et al. (Eds.): ICAIS 2022, CCIS 1586, pp. 340–350, 2022.
https://doi.org/10.1007/978-3-031-06767-9_28

the study of Tibetan speech synthesis [4–7]. Therefore, the effect of speech synthesis based on deep learning is better than that based on statistical parameters and waveform splicing.

Some works have shown that the end-to-end speech synthesis has better performance than traditional methods. The work of [8] showed that the end-to-end method takes a shorter time than the statistical parameter method. The works of [9–12] found that the performance of the TTS (text-to-speech) system has been further improved through using end-to-end neural network frameworks. Tacotron2, an advanced TTS system, consists of an encoder-decoder architecture and a neural WaveNet vocoder [13, 14]. Since the synthesized voice by the Tacotron series is the closest to natural speech, it is widely used and researched.

For the research on Tibetan speech synthesis, the work of [5] proposed a speech synthesis method which is a breakthrough in traditional recurrent neural network and convolutional neural network, using the Seq2Seq model and Griffin-lim as a vocoder for Tibetan speech synthesis, and obtained a clearer and more natural synthesized speech. Based on the Seq2Seq model, the work of [12] proposed to use WaveNet as a vocoder to conduct experiments. The experimental results show that the synthesized speech by WaveNet has better clarity and naturalness.

Combining the advantages of Tacotron2 and WaveNet, this paper proposes an end-to-end speech synthesis method for Tibetan multi-speaker. The main method is to insert the speaker ID in the Tibetan text, and Tibetan characters are transcribed by the corresponding Latin alphabets. The preprocessed text will be input to a sequence-to-sequence feature prediction network, which outputs a predicted Mel spectrogram. And then WaveNet is trained by the target speaker's speech data and the Mel spectrogram.

The contributions of this paper are as follows: (1) Proposing an end-to-end speech synthesis method for Lhasa-Tibetan multi-speaker. All modules are merged into one system. Our method can be used to synthesize different speaker voices. (2) Transliterating Tibetan characters into Latin letters using the Wiley transliteration. This process will effectively enhance the learning effect of the model under the limited Tibetan speech data. (3) The speaker ID is embedded in the model to synthesize speech for the specific speaker.

2 Method

2.1 Text Preprocessing

Although Tibetan has a very long history, the orthography of written language is still unchanged. The work of [15] has detailed the structure of Tibetan sentences: Tibetan letters are composed of 30 consonants and 4 vowels, and each sentence consists of a sequence of single syllables. The writing order is from left to right. Figure 1 is an example of a Tibetan sentence.

Each syllable in Tibetan has a root character, which is the central consonant of the syllable. A vowel tag can be above or below the root character to indicate different syllables. Sometimes there is a superscript character at the top of the root character, one or two subscript characters at the bottom, and a prescript character in front, indicating that the initials of the syllable are compound consonants. Sometimes there are one or

འཇའ་ཚོན་གྱི་ཚུལ་གྱིས་ཁྱེད་ཀྱི་བར་སྣང་ལ་སྤྲོས་ཆོག་པར་ཞུ།

Fig. 1. A Tibetan sentence.

two postscripts after the root character, which means that the syllable ends with one or two consonants. The structure of Tibetan syllables is shown in Fig. 2.

Speech synthesis units affect the effect of synthesis. A Tibetan syllable can have 20,450 spelling schemes. The work of [16] explored the problem of selecting Tibetan synthesis unit and proposed a Tibetan speech synthesis method that combines components, combination components, characters, words, and phonemes. If Tibetan monosyllables are used as the primitives of speech synthesis, a large amount of corpus needs to be built, which is not easy to implement. The Tibetan speech synthesis system in [17] uses Tibetan consonants and vowels as the primitives of speech synthesis, but this often requires researchers to have a lot of knowledge of Tibetan linguistics. The works of [18, 19] used phonemes as speech synthesis units, but it is necessary to build a phoneme dictionary, which increases the workload of front-end analysis. In this paper, 26 Latin letters are used as speech synthesis primitives and Tibetan text can be transcribed in Latin letters by the Willy transliteration scheme. This approach will effectively reduce the required training data, the workload of text processing, and improve the synthesis efficiency.

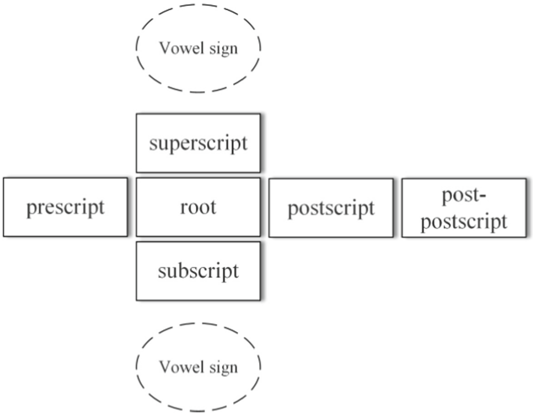

Fig. 2. The structure of Tibetan syllables.

Figure 3 shows the converted Tibetan sentences obtained by using the Willy transliteration scheme.

ja’ tshon gyi tshul gyis khyed kyi bar snang la spros chog par zhu

Fig. 3. Example of Willie Transliteration.

2.2 Spectrogram Feature Prediction Network

The Tacotron2 model was designed and developed by Google in 2017. Since its synthesis quality is close to the human voice, it is used by many researchers in the speech synthesis field. An end-to-end speech synthesis system based on the Tacotron2 model is designed in this paper. The first part of the model is the spectrogram feature prediction network with an attention mechanism, which is mainly composed of encoder, attention mechanism and decoder. We use the Tibetan text transcribed by the Latin alphabet and inserted by speaker IDs as the input sequence of the spectrogram feature prediction network. After encoding and decoding operations, the second part of the model is WaveNet vocoder, which is used to synthesize the target speaker's speech waveform.

The encoder consists of a character embedding layer, a 3-layer convolutional neural network, and a Long Short Term Memory Network (LSTM). The decoder is composed of Pre-Net, LSTM, linear projection and Post-net. The new sequence is obtained by the input sequence through the encoder, and then the Mel spectrogram is obtained through the attention mechanism and the decoder. The encoder and decoder implement the conversion from text to context vector, and the WaveNet vocoder converts the context vector into waveform samples. Figure 4 below shows the model of end-to-end speech synthesis for Lhasa-Tibetan multi-speaker [20, 21].

Feature Extraction of Mel Spectrogram. In order to simulate the human ear auditory system and obtain a better speech synthesis effect, the spectrogram feature prediction network will output the Mel spectrograms. Mel spectrogram is the addition of Mel filter function to the general speech spectrogram, so that the frequency of the sound is within the range of human hearing.

Attention Mechanism. Usually, a very long sequence passes through the encoder to get a fixed-length context vector, but that may cause that the encoder cannot represent the entire information of the sequence. At the same time, because the vector is continuously generated on the timeline, the subsequent vectors may cover the previous sequence. So more information will be lost in the subsequent decoding process, which is not conducive to the model learning. Therefore, the attention mechanism needs to be added to the system [22].

The attention mechanism aims to select some key information in a sequence to improve the efficiency of the neural network. It gives different weights for different information data. Attention mechanism makes it possible to focus on the encoding results related to the current in each step during decoding, thereby reducing the difficulty of learning the output representation of the encoder [23, 24]. Its working principle is shown in Fig. 5.

In the sequence-to-sequence model with the attention mechanism, the input n-frame sequence is encoded by the encoder into a context vector c, where a_n is the input n-th frame, and h_n is the output of the encoder. The formula (1) is as follows:

$$\begin{aligned} \mathrm{c} &= \mathrm{encoder}(a_1, a_2, a_3 \ldots a_n) \\ &= h_1, h_2, h_3 \ldots h_n \end{aligned} \tag{1}$$

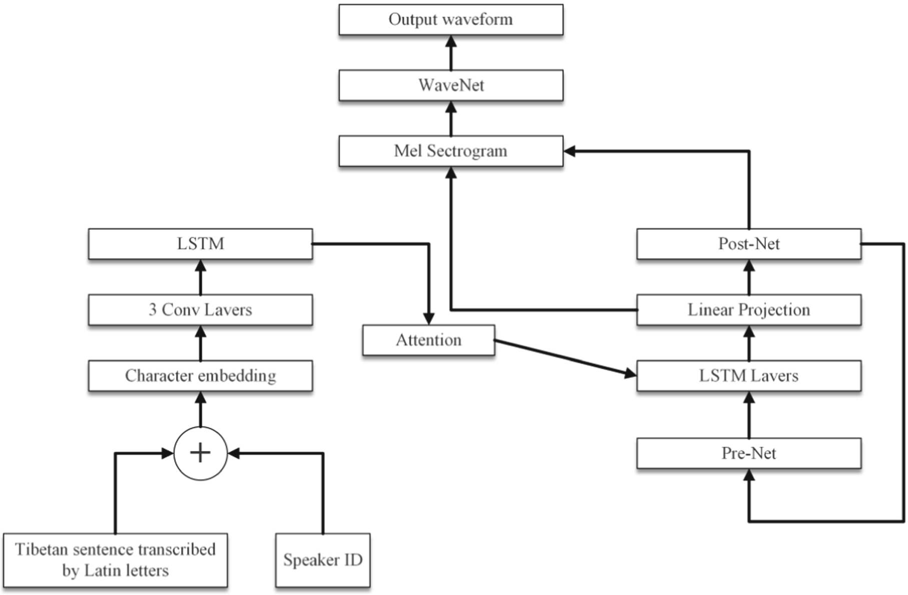

Fig. 4. The model of end-to-end speech synthesis for Lhasa-Tibetan multi-speaker.

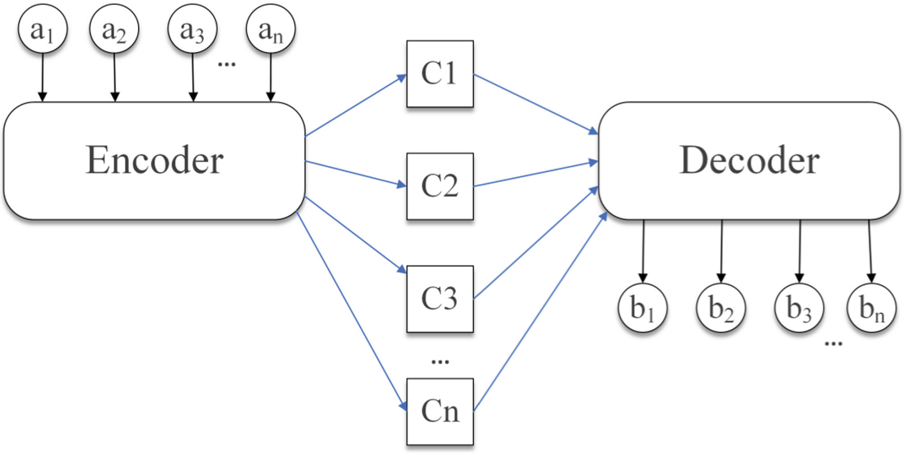

Fig. 5. Sequence-to-sequence model diagram with an attention mechanism.

The attention mechanism connecting the encoder and the decoder will combine each different c to select key information for decoding, so that the model will be more accurate. The output of the encoder is calculated by the formula (2), where h_j is the eigenvector output by the encoder and the a_{ij} is the weight.

$$\mathrm{c_i} = \sum_{j=1}^{T_i} a_{\mathrm{ij}} h_j \tag{2}$$

a_{ij} is calculated by the formula (3), and e_{ij} represents the matching degree between the j-th input of the encoder and the i-th character of the decoder, and the e_{ij} formula (4) is as follows.

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \tag{3}$$

$$\mathrm{e}_{ij} = a(s_{i-1}, h_j) \tag{4}$$

2.3 Vocoder

Google's end-to-end model Tacotron [25] first converts voice or text data into frequency spectrum and then uses a vocoder to obtain synthesized speech. The vocoder in Tacotron is the Griffin-lim algorithm for the time-domain waveform. The Griffin-lim algorithm reconstructs speech when only the amplitude spectrum is known but the phase spectrum is unknown. It can estimate the discarded phase information and use a short-time inverse fourier transform to convert the linear spectrogram into the time domain waveform. It is a simple and efficient vocoder. However, because the waveform generated by the Griffin-lim vocoder is too smooth, the synthesized speech quality is poor, and it sounds unnatural.

This work uses the WaveNet model as a vocoder to cover the limitation of the Griffin-lim algorithm. WaveNet is mainly composed of stacked dilated causal convolutions, and it uses causal convolution to increase the receptive field of convolution. WaveNet synthesizes speech by fitting the distribution of audio waveforms by the autoregressive method, which means after inputting predicted sampling points, these convolutional layers and functions of gating activation are passed, and WaveNet will predict the next sampling point according to the input sampling points, then finally synthesize speech by predicting the value of the waveform at each time step [8].

WaveNet usually used the traditional acoustic and linguistic features as the input of the model for speech synthesis. To improve the model's efficiency and reduce the loss of details, in this paper, we choose the Mel spectrogram as the input of WaveNet for training. Mel spectrogram emphasizes the details of low frequency, which is very vital for the accuracy of speech. And it is easier to train with the square error loss [26]. Therefore, in this work, we will train WaveNet vocoder for multi-speaker's speech synthesis.

3 Experiments Settings

The spectrogram feature prediction network is trained on a Tibetan speech data of 4 speakers. The speech data are about 9.97 h with 7419 text sentences. Speech data files are converted to a 16 kHz sampling rate, with 16 bit quantization.

The training data is shown in Table 1. For the training process, the batch size is 8, and after 50,000 iterations, the learning rate is reduced from 10^{-3} to 10^{-4}.

WaveNet vocoder is trained based on the data of multi-speaker. Each spectrogram frame is precisely aligned with the waveform sample.

Table 1. Details of experimental data.

Speaker	Number of text sentences	Time of the speech data (Hours)
Speaker-A	1827	3.175
Speaker-B	1932	2.43
Speaker-C	1818	2.89
Speaker-D	1842	1.475

4 Experimental Results

This paper adopts two experimental methods for the model evaluation: subjective and objective experiments. In the subjective experiment, we randomly selected 20 listeners. After listening to the synthesized speech, they will score it according to the scoring criteria in Table 2. Finally, the scores of all listeners are calculated as the average opinion score (MOS) of the synthesized speech. The higher the MOS, means the better the speech synthesis effect. In the objective experiment, the root mean square error (RMSE) of the time domain sequence is calculated to measure the difference between the synthesized speech and the reference speech. Among them, the smaller the RMSE, the closer the synthesized speech is to the reference and the better the speech synthesis effect.

4.1 Subjective Experiment

In the subjective experiment, we randomly selected 20 listeners. After listening to the synthesized speech, the original speech was used as a reference, and the synthesized speech was scored according to the scoring criteria in Table 2. In order to evaluate the effect of the model's synthesis speech, we compared the speech of using the speaker ID embedded into the spectrogram feature prediction network with using WaveNet vocoders trained on specific speaker data without embedding speaker ID in Tibetan text. Finally, the scores of all listeners are calculated as the average opinion score (MOS) of the synthesized speech. Table 3 shows the results.

From the experiment results, it can be seen that the synthesized speech effect obtained by our model is good. By the way, the synthesis quality of speaker-A is the highest, because the data is larger.

Table 2. Subjective evaluation criteria.

Score	Sound quality
5	Very good
4	Good
3	Not bad
2	Bad
1	Very bad

Table 3. Subjective evaluation results.

Speaker	MOS of using the speaker ID embedded into the spectrogram feature prediction network	MOS of using WaveNet vocoders trained on specific speaker data without embedding speaker ID in Tibetan text
Speaker-A	4.16	3.68
Speaker-B	4.03	3.55
Speaker-C	4.12	3.24
Speaker-D	3.88	3.16

4.2 Objective Experiment

The objective experiment mainly calculates the root mean square error (RMSE) of the time-domain sequence. The RMSE formula is shown in formula (5), where $L_{1,t}$ and $L_{2,t}$ respectively represent the time series values of the reference speech and synthesized speech at a time t.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (L_{1,t} - L_{2,t})^2}{n}} \tag{5}$$

We randomly select 10 text sentences, using the end-to-end speech synthesis method for Lhasa-Tibetan multi-speaker, and calculate the average RMSE to evaluate the closeness of the synthesized speech and the reference speech. In order to evaluate the performance of the model, we compared it with the model using WaveNet vocoders trained on specific speaker data without embedding speaker ID in Tibetan text. These models are used to synthesize the same 10 text sentences and calculate the average value of RMSE. The results are shown in Table 4.

It can be seen from the Table 4 that the RMSE value of the speaker ID embedded into the spectrogram feature prediction network (0.183, 0.211, 0.220, 0.254) is less than the RMSE value of using WaveNet vocoder trained by speaker-A (0.235), speaker-B (0.240), speaker-C (0.277) and speaker-D (0.293). The speech synthesis quality obtained by speaker-A is relatively better, because the data is larger.

Table 4. Comparative evaluation results.

Speaker	RMSE value of using the speaker ID embedded into the spectrogram feature prediction network	RMSE value of using WaveNet vocoders trained on specific speaker data without embedding speaker ID in Tibetan text
A	0.183	0.235
B	0.211	0.240
C	0.220	0.277
D	0.254	0.293

Figure 6 is the Mel spectrogram of target speech and predicted spectrogram of speaker-A obtained by our method. Figure 7 is for speaker-B. It can be seen that both are very similar.

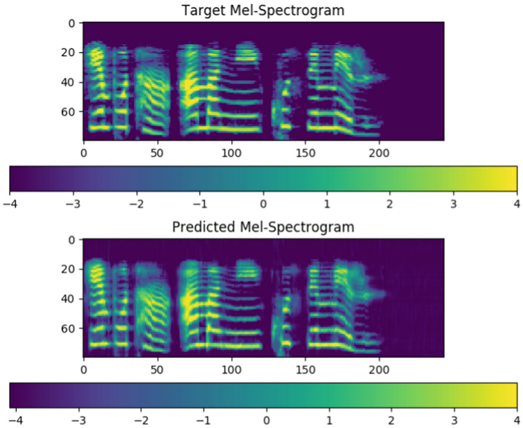

Fig. 6. The target speech Mel spectrogram and the predicted Mel spectrogram of Speaker-A.

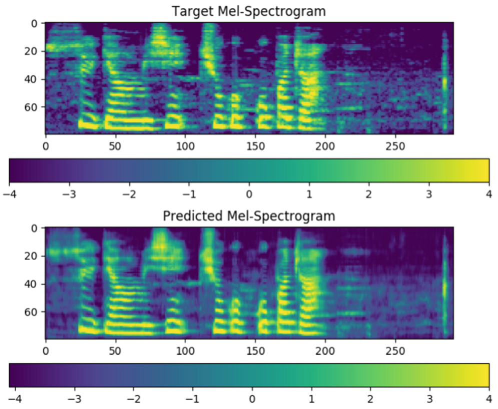

Fig. 7. The target speech Mel spectrogram and the predicted Mel spectrogram of Speaker-B.

5 Summary

This paper constructs an end-to-end speech synthesis method for Lhasa-Tibetan multi-speaker, including a spectrogram feature prediction network embedded speaker ID and a WaveNet vocoder which converted the Mel spectrograms to time-domain waveform. In the text processing, the Wiley transliteration scheme was used to convert Tibetan characters into Latin letters, which effectively reduced the scale of training data. Referring to the results of objective and subjective experiments, our method has better speech quality than the model using some WaveNet vocoders trained on specific speaker data.

References

1. Fan, Y., Qian, Y., et al.: Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, pp. 4475–4479 (2015)
2. Cooper, E., et al.: Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, pp. 6184–6188 (2020)
3. Jia, Y., Zhang, Y., Weiss, R.J., et al.: Transfer learning from speaker verification to multispeaker text-to-speech synthesis. Adv. Neural. Inf. Process. Syst. **31**, 4485–4495 (2018)
4. Huang, C.: Frontiers of Tibetan studies in 2019. J. Tibet Nationalities Univ. (Philos. Soc. Sci. Ed.) **41**(5), 47–52 (2020)
5. Du, G.: Research on Tibetan speech synthesis technology based on neural network. M.S. dissertation, Qinghai Normal University (2019)
6. Luo, L.: Research and implementation of sequence-to-sequence Tibetan Lhasa speech synthesis. M.S. dissertation, Northwest University for Nationalities (2019)
7. Liu, F.: Research on key technologies of Tibetan speech synthesis system. J. Tibet Univ. (Nat. Sci. Ed.) **31**(2), 87–91 (2016)
8. Ling, Z., Wu, H.: Study on speech synthetic vocoder based on WaveNet. Artif. Intell. **1**, 83–91 (2018)
9. Luo, L., Li, G., et al.: End-to-end speech synthesis for Tibetan Lhasa dialect. J. Phys: Conf. Ser. **1187**(5), 052061 (2019)

10. Zhao, Y., Hu, P., et al.: Lhasa-Tibetan speech synthesis using end-to-end model. IEEE Access **7**, 140305–140311 (2019)
11. Li, G., Luo, L., et al.: End-to-end Tibetan speech synthesis based on phones and semi-syllables. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, pp. 1294–1297 (2019)
12. Ding, Y., Cai, R., Gong, B.: Tibetan speech synthesis based on an improved neural network. MATEC Web Conf. **336**(5), 0612 (2021)
13. Shen, J., et al.: Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, pp. 4779–4783 (2018)
14. Tobing, P., Wu, Y., et al.: An evaluation of voice conversion with neural network spectral mapping models and WaveNet vocoder. APSIPA Trans. Signal Inf. Process. **9**, E26 (2020)
15. Gongbao, C.: Research on Tibetan speech synthesis technology. M.S. dissertation, Qinghai University for Nationalities (2014)
16. Cairang, Z., Li, Y., Cai, Z.: Selection of Tibetan speech synthesis unit. J. Softw. **26**(6), 1409–1420 (2015)
17. Gongbao, C.: Research on Tibetan speech synthesis based on consonants and vowels. Inf. Comput. (Theoret. Ed.) **1**, 52–53 (2014)
18. Yang, H., Oura, K., Wang, H., Gan, Z., Tokuda, K.: Using speaker adaptive training to realize Mandarin-Tibetan cross-lingual speech synthesis. Multimedia Tools Appl. **74**(22), 9927–9942 (2014). https://doi.org/10.1007/s11042-014-2117-9
19. Li, M., Zhang, G., et al.: The phoneme automatic segmentation algorithms study of Tibetan lhasa words continuous speech stream. Advanced Materials Research, pp. 2051–2054 (2013)
20. Soonil, K.: 1D-CNN: speech emotion recognition system using a stacked network with dilated CNN features. J. Big Data **67**(3), 4039–4059 (2021)
21. Kalphana, I., Kesavamurthy, T.: Convolutional neural network auto encoder channel estimation algorithm in mimo-ofdm system. Comput. Syst. Sci. Eng. **41**(1), 171–185 (2022)
22. Prabhu, K., et al.: Facial expression recognition using enhanced convolution neural network with attention mechanism. Comput. Syst. Sci. Eng. **41**(1), 415–426 (2022)
23. Almars, A.M.: Attention-based Bi-LSTM model for Arabic depression classification. Comput. Mater. Continua **71**(2), 3091–3106 (2022)
24. Sun, J., Li, Y., Shen, Y., et al.: Joint self-attention based neural networks for semantic relation extraction. J. Inf. Hiding Privacy Prot. **1**(2), 69–75 (2019)
25. Skerry-Ryan, R., Battenberg, E., et al.: Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In: International Conference on Machine Learning (ICML), Stockholm, Sweden, pp. 4693–4702 (2018)
26. Tamamori, A., Hayashi, T., et al.: Speaker-dependent wavenet vocoder. In: Interspeech 2017, Stockholm, Sweden, pp. 1118–1122 (2017)