Mongolian emotional speech synthesis based on transfer learning and emotional embedding

Aihong Huang
School of Computer Science, Inner Mongolia University
National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian
Provincial Key Laboratory of Mongolian Information Processing Technology
Hohhot, China
e-mail:huangaihong1996@163.com

Feilong Bao
School of Computer Science, Inner Mongolia University
National & Local Joint Engineering, Research Center of Intelligent Information Processing, Technology for Mongolian
Provincial Key Laboratory of Mongolian Information Processing Technology
Hohhot, China
e-mail:csfeilong@imu.edu.cn

Guanglai Gao
School of Computer Science, Inner Mongolia University
National & Local Joint Engineering, Research Center of, Intelligent Information Processing, Technology for Mongolian
Provincial Key Laboratory of Mongolian Information Processing Technology
Hohhot, China
e-mail:csggl@imu.edu.cn

Yu Shan
School of Computer Science, Inner Mongolia University
National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian
Provincial Key Laboratory of Mongolian Information Processing Technology
Hohhot, China
e-mail:529117810@qq.com

Rui Liu
School of Computer Science, Inner Mongolia University
National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian
Provincial Key Laboratory of Mongolian Information Processing Technology
Hohhot, China
e-mail:liurui_imu@163.com

Abstract*—In recent years, end-to-end speech synthesis based on attention has achieved better performance than traditional speech synthesis models, and the technology of end-to-end Mongolian speech synthesis has reached the application standard. However, due to the sparse training corpus, the research on Mongolian emotional speech synthesis is still far from perfect. In response to these problems, we established a Mongolian emotional corpus and constructed an emotionally controllable Mongolian speech synthesis system for the first time. Through combining transfer learning and emotional embedding, the Mongolian emotional speech synthesis system with 8 kinds of emotions (happy, angry, sadness, surprise, fear, disgust, boredom and neutral) has been achieved. We proposed the method that emotional labels are used as the input of the emotional embedding layer to generate emotional vectors, which are spliced with the output vectors of the bidirectional LSTM layer, so that the text representation vectors contain information about emotional category, thereby synthesize a variety of different emotional voices. Experiments show that our method can synthesize high-quality Mongolian emotional speech.

***Keywords*—Traditional Mongolian, Emotional speech synthesis, End-to-End, Transfer learning, Emotional embedding**

I. INTRODUCTION

With the application and development of deep learning technology, the stability of acoustic models and speed has been greatly improved in recent years, and the quality of audio synthesized is closer to the human voice. However, the technology of speech synthesis has been facing with a series of new challenges recently. For example, in applications such as audiobooks and voice assistants, the voice synthesized not only requires clear pronunciation, but also requires rich expressiveness, including emotion, speaking speed, stress, pauses, and so on. Currently, the technology of speech synthesis still cannot meet all the requirements of actual scenarios. In order to synthesize more neutral and expressive speech as much as possible, the technology of emotional speech synthesis has become an important research subject.

After years of research, the technology of Mongolian speech synthesis has gone through multiple stages and has been able to synthesis better neutral Mongolian speech now. The method of Mongolian speech synthesis based on a deep neural network (DNN) [1] was proposed years ago. The use of the DNN model has improved further the performance of synthesized speech comprehensively. In recent years, end-to-end (E2E) approaches [2–7] have pushed the frontiers of speech synthesis with advantages of less feature engineering, unified acoustic-duration modeling and more natural-sounding speech, which can directly establish the relationship between text and speech has been widely used in the process of speech synthesis method. Then, researchers have applied end-to-end speech synthesis technology to Mongolian speech synthesis tasks[8].In addition, the method of modifying the Mongolian front-end processing process and using a high-quality vocoder [9] has been proposed, which has improved the sound quality of synthesized speech and reduced the word error rate of it effectively. However, different from languages with very large user bases such as Chinese and English, Mongolian emotional speech synthesis technology has not been implemented due to the huge vocabulary of Mongolian and the sparse training corpus.

In this paper, we established a Mongolian emotional corpus and proposed for the first time that a Mongolian emotional speech synthesis model with eight emotions including happiness, anger, sadness, surprise, fear, disgust, boredom and neutral is realized by combining transfer learning and emotional tokens embedding in Tacotron2[3]. Meanwhile, fulfilled the emotionally controllable speech synthesis technology. Through subjective and objective evaluation, the effectiveness of the method is proved.

The rest of the paper is organized as follows: Section 2 introduces the related work of emotional speech synthesis technology, transfer learning and Mongolian speech synthesis technology. Section 3 gives the method of the Mongolian emotional controllable speech synthesis model in

detail. Section 4 shows the experiment and its results. Finally, section 5 summarizes this article and also puts forward the outlook for future work.

II. Related Work

In recent years, with the development of computing power, various methods based on neural network about speech synthesis have been extensively studied. Wang et al. have proposed the end-to-end models Tacotron[2] and Tacotron2[3]. Sercan et al. have invented Deep-Voice[4]. Vanden proposed WaveNet[5] which is an autoregressive vocoder based on extended convolution. Furthermore, Valin's LPCNet[6] vocoder improved the quality of speech synthesis largely. The end-to-end speech synthesis method can learn the mapping relationship between the input text sequence and the corresponding speech parameter sequence directly, which removes the complicated process of front-end text analysis, and simplifies the process of speech synthesis training and synthesis. Then, as an extension of Tacotron, the work of global style tokens(GST)[9,10] is proposed to learn interpretable style embeddings in an unsupervised way. On this basis, a semi-supervised training method based on GST [11] is proposed, which uses a small amount of emotionally annotated data for training, and then to obtain interpretable style tokens and control the synthesis of different emotional speech. Subsequently, emotional speech synthesis methods using emotional labels have been widely applied to end-to-end models [12-17]. Recently, Xie et al. [18,19] published an end-to-end model of fine-grained emotion control and prediction, which can control the emotion category and emotion strength of synthesized speech. The study of fine-grained promotes further the progress of emotion speech synthesis. At present, in emotional speech synthesis tasks, the main research is focused on the extended study based on the end-to-end speech synthesis model. However the above methods still do not solve the emotional speech synthesis of small languages such as Mongolian.

Transfer learning is an important tool in machine learning to solve the basic problem of insufficient training data. Network-based deep transfer learning refers to the reuse the partial network that pre-trained in the source domain, including its network structure and connection parameters, transfer it to be a part of deep neural network which used in target domain[20]. Firstly, pretraining strategy needed to be utilized when the large emotional speech corpus is not available. Transferring knowledge from massive, out-of-domain data to aid learning in the target domain have been shown to boost performance[21]. Then, there is no doubt that the large amount of data enables the end-to-end TTS model to produce high-quality speech. Unfortunately, there aren't any large, high-quality emotional Mongolian TTS corpus that are publicly available. In a nutshell, the end-to-end neural TTS framework do not work well for limited emotion speech corpus, because high quality speech datasets with emotional content needed for emotional TTS are quite difficult to collect [22].

Recently, Liu et al. proposed an end-to-end Mongolian acoustic modeling method based on knowledge distillation by improving the end-to-end acoustic model[23,24], which reduced the loss of synthetic speech caused by the exposure bias of the autoregressive model. In[25], researchers propose a feature augmentation method in conjunction with a self-attention neural classifier. They augment input text with morphological and phonological decompositions of words to enhance the text encoder, which improves largely voice quality in a Mongolian text-to-speech synthesis system. At present, the end-to-end Mongolian speech synthesis technology has reached an applicable standard. However, the research on Mongolian emotional speech synthesis is still not achieved. In order to solve this problem, this paper proposes a method of combining transfer learning and emotional embedding to realize the Mongolian emotional speech synthesis technology that can synthesize eight emotions.

III. Method

The process of Mongolian speech synthesis with controllable emotion includes text encoder, emotion encoder , attention-based decoder and pre-training based on transfer learning mainly. The structure of the attention-based decoder is the same as Tacotron-based model[2,3].

A. Text encoder

In this paper, We use the Arabic numerals to mark different emotions. The corresponding relationship between emotions and labels as follows: neutral-0, happy-1, angry-2, sad-3, surprise-4, fear-5, disgust-6, bored-7.Due to the change of the model, the input is changed from the pair <voice, text> in the original model to <voice, text, emotion label> as the input of the model. In this way, text, speech and corresponding emotion categories can be obtained in the training process.

a) Convolutional network layer: In this layer, the model has learned appropriate parameters of convolution kernel to initialize the model. Compared with randomly initializing, the formula of the convolutional layer can be updated as:

$$F(x) = \sum_{i=1}^{n} \bar{f}(i)g(x-i) = \int_{1}^{n} \bar{f}(i)g(x-i)\,\boldsymbol{di} \qquad (1)$$

F(x) is the result of the convolution operation and *g(x)* is the vector to be convolved, furthermore, *n* is the length of vector. Use $\bar{f}(x)$ which is the parameters of pre-training convolution kernel instead of *f(x)* which is the parameters of initialized convolution kernel randomly.

b) Bidirectional LSTM layer: In this layer, the weight matrix among neurons has been obtained by pre-training. Therefore, the weight matrix from pre-training is used to training the new data in LSTM layer. The formula of the forget gate can be updated as：

$$f_t = \sigma\left(\overline{W_f} \cdot [h_{t-1}, x_t] + \overline{b_f}\right) \qquad (2)$$

The formula of input gate is updated to:

$$i_t = \sigma(\overline{W_i} \cdot [h_{t-1}, x_t] + \overline{b_i}) \qquad (3)$$

$$g_t = tanh\left(\overline{W_g} \cdot [h_{t-1}, x_t] + \overline{b_g}\right) \qquad (4)$$

The formula of output gates is updated as:

$$h_t = o_t * tanh\,(c_t) \qquad (5)$$
$$= tanh(\overline{W_o} \cdot [h_{t-1}, x_t] + \overline{b_o}) * tanh\,(c_t)$$

After through the bidirectional LSTM layer, the result formula is expressed as:

$$EncoderOutput(f) = \bigcup_{\substack{i=0 \\ j=0}}^{n} \left(h_{Li}, h_{Rj}\right) \qquad (6)$$

And, h_{Li} is the forward LSTM output, h_{Rj} is output for reverse LSTM layer.

B. Emotion encoder layer

In this paper, emotional labels are converted into multi-dimensional vectors and added to the model as the representation of emotional categories. Then, splice emotion vectors with the output of the encoder as the new input of the attention. As shown in Fig. 1, the new input information contains text and emotion categories. In the training step, for the initialization of emotion labels, we use the normal distribution with a standard deviation of 0.5.

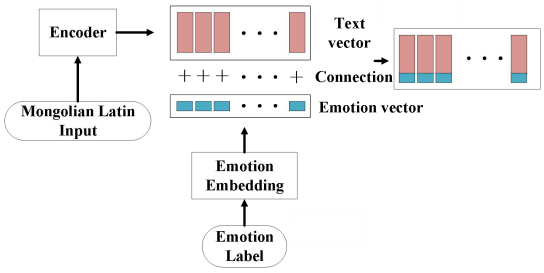

Fig. 1. The process of splicing Emotion vector

C. Pre-training based on transfer learning

Due to the lack of Mongolian emotional corpus at present, it is impossible to synthesize better-quality voice by training the annotated emotional corpus. So we propose a method based on transfer learning. The training process is shown in Fig. 2. The colored part in the figure indicates the model was initialized by pre-training and the white part indicates that it was initialized randomly.

First, we use a large amount of neutral speech corpus (whose emotion label is 0) that non-target speaker to train the model, which is called the pre-training model. The prior probability distribution of the emotional speech synthesis model and the neutral emotion vector has been obtained through pre-training. Then initialize parameters of the emotional controllable Mongolian TTS model by pre-training. Next, train the Mongolian emotional corpus, and finally get the emotional controllable Mongolian speech synthesis model, which can synthesizes different Mongolian emotional voices according to the content and emotional label.

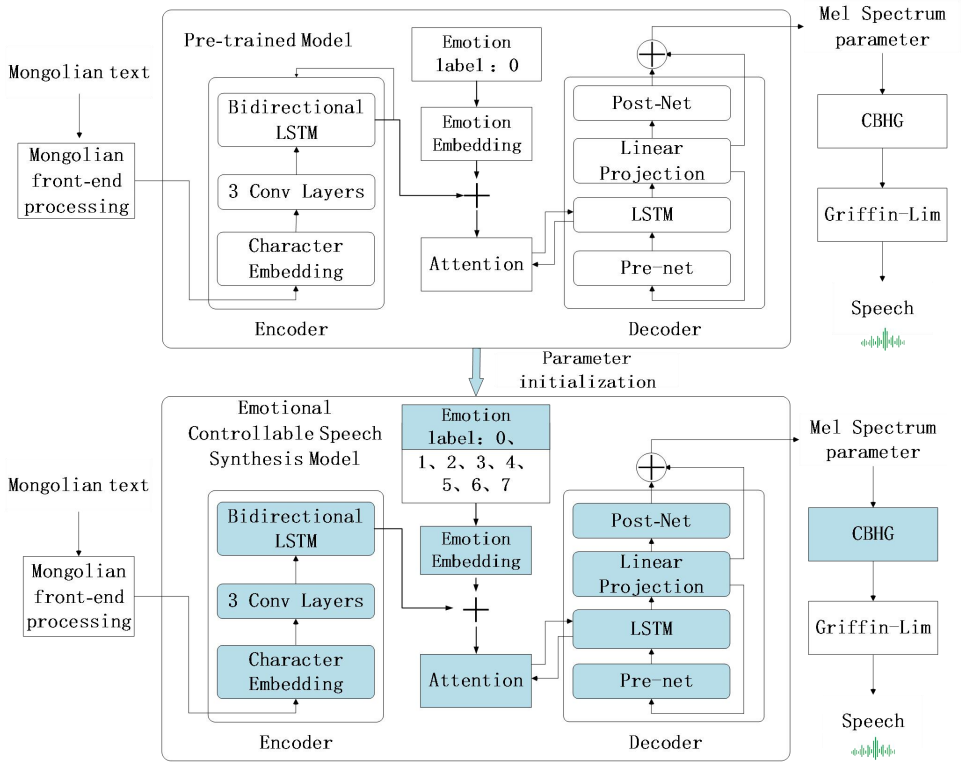

Fig. 2. The training process of pre-training combined with labeled Mongolian emotional speech synthesis model

For the process of acoustic modeling, after adding the way of emotional embedding, the formula of emotional speech synthesis modeling is expressed as:

$$\hat{\theta} = arg\,max\,p(Y|X,e,\theta) \qquad (7)$$

Given the text feature X and the emotional feature e, then search the acoustic model parameter θ and obtain the maximum probability speech feature Y. Finally $\hat{\theta}$ is the speech acoustic model.

IV. Experiment

A. Experimental setup

The neural corpus of the experimental pre-training model consists of 43,363 sentences about 58.7 hours, which were recorded by a Mongolian professional announcer. The Mongolian emotional corpus includes eight emotions such as happy, angry, sad, surprise, fear, disgust, bored and neutral, and there are 500 sentences for each emotion. The Mongolian emotional corpus is recorded by a professional Mongolian voice actress. Then, select 20 sentences from the 500 experimental data as the test set randomly, and the other sentences as the training set.

In the experiment, in order to discuss the influence of different dimensions of emotion vectors on the quality of the voice, we proposed using two-dimensional, four-dimensional, eight-dimensional, and sixteen-dimensional emotion vectors to represent emotion labels. The two-dimensional emotion vector model is represented as MET-2, and so on. Models with different dimensional emotion vectors are trained by same steps in the training process.

In this paper, the pre-training model is trained 200K times based on Mongolian neutral corpus. When using Mongolian emotional corpus to train the model, select the number of iterations with low training error and stable error as the training times of the model.

B. *Objective evaluation*

In this part, we evaluate synthesized speech objectively by Mel Cepstral Distortion (MCD). MCD is mainly used to measure the difference between the Mel spectrum of two sequences objectively. Then we also show the Mel spectrograms of synthetic samples with the same text.

C. *Subjective evaluation*

In the subjective evaluation, 20 native Mongolian people were selected to evaluate the quality of speech synthesized by pre-training model and speech synthesized by non-pre-training model with mean opinion score (MOS) among different dimensions.

Because of the universality of emotional expression between different languages, mandarin speakers can also distinguish the different emotional voices. Therefore, 10 evaluators whose native language is Chinese and 20 evaluators whose native language is Mongolian are selected to judge the emotional categories of the synthesized speech.

D. *Result*

1) *The result of objectively evaluation:*

a) Analysis of MCD: The calculation result of MCD is shown in Fig. 3. It can be observed that in the emotions of surprise, disgust and bored, the value of MCD with 4-dimensional emotion vectors is higher than 8-dimensional and 16-dimensional. The values of MCD baesd 8-dimensional and 16-dimensional emotion vectors are kept at a low level.

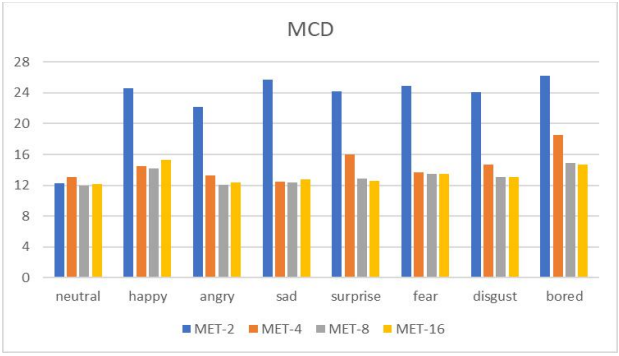

Fig. 3. MCD calculation results of emotion vector of different dimensions

b) Analysis of Mel spectrograms: Fig. 4 illustrates the Mel spectrograms of synthetic samples for eight emotions. Note that the input to the Mongolian emotional speech synthesis system is the same piece of text from the test set, but with different emotion category. We can clearly see different Mel spectrum patterns, showing varations in speaking rate, rhythm and pause duration.We can observe that the angry emotion(Fig. 4(c)) is characterized by quick speaking style and the Mel spectrograms is more bright than others.Second, compared with the neural(Fig. 4(a)) emotion, the bored(Fig. 4(h)) emotion is characterized by slow speaking rate with longer time.The Mel spectrum indicate that the way of Mongolian emotional speech synthesis we proposed can synthesize the voice with different emotions.

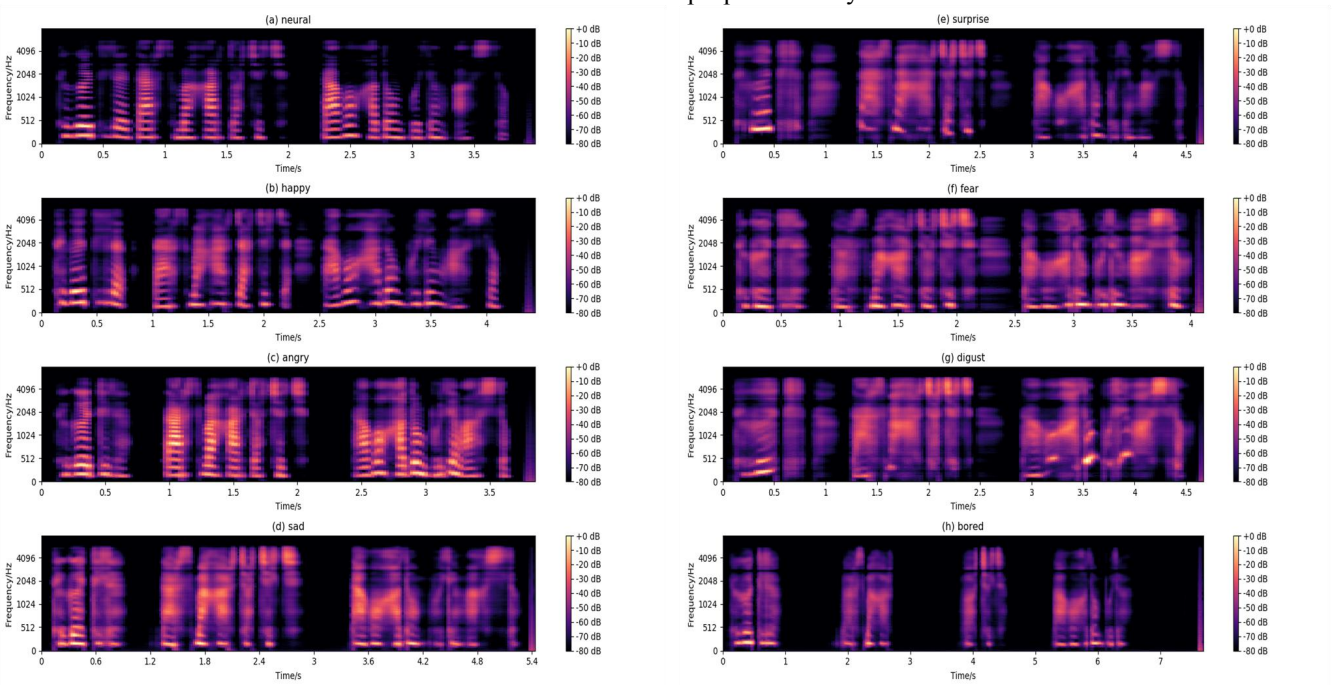

Fig. 4. Mel spectrograms of synthetic samples with the same text for eight emotions

2) The result of subjectively evaluation:

a) The results of MOS evaluation on model with pre-training: Results are shown in Table 1, which reflects that the average of MOS values are very low generally with MET-2. Furthermore, the values of MOS with MET-8 are higher than MET-4 and Met-16 overall. So, we can conclude that it is the most suitable for emotion lables represented with MET-8 in Mongolian emotional speech synthesis task.

TABLE I.

SENTIMENT VECTOR MOS VALUE SCORE TABLE OF DIFFERENT DIMENSIONS WITH PRE-TRAINING

emotion	MET-2(MOS)	MET-4(MOS)	MET-8(MOS)	MET-16(MOS)
neutral	3.65±0.09	3.63±0.10	3.63±0.10	3.64±0.10
happy	2.87±0.11	3.37±0.12	3.47±0.11	3.42±0.11
angry	2.93±0.10	3.51±0.11	3.55±0.13	3.54±0.12
sad	2.59±0.10	3.6±0.11	3.6±0.10	3.58±0.11
surprise	2.92±0.10	3.37±0.13	3.42±0.12	3.4±0.13
fear	2.94±0.11	3.53±0.11	3.57±0.10	3.57±0.10
disgust	2.55±0.14	3.4±0.10	3.61±0.10	3.6±0.11
bored	2.64±0.11	3.1±0.13	3.33±0.12	3.34±0.12

*b) The results of MOS evaluation on model trained on the labeled emotion corpus but without pretraining :*In this part we show the MOS evaluation on model trained on the labeled emotion corpus but without pretraining characterized by with MET-2, MET-4 MET-8 MET-16. Results are shown in Table 2. Compared with the result of pre-training, it is obviously that the values of MOS are lower universally,which can prove that pre-training baesd transfer learning can synthesis speech effectively although with sparse corpus.

TABLE II.

SENTIMENT VECTOR MOS VALUE SCORE TABLE OF THE MODEL TRAINED ON THE LABELED EMOTION CORPUS BUT WITHOUT PRETRAINING

emotion	MET-2(MOS)	MET-4(MOS)	MET-8(MOS)	MET-16(MOS)
neutral	1.07±0.08	1.12±0.11	1.15±0.10	1.15±0.13
happy	0.86±0.07	0.93±0.09	0.95±0.11	0.93±0.11
angry	0.89±0.10	0.91±0.14	1.05±0.06	0.96±0.10
sad	0.75±0.14	0.83±0.11	0.86±0.13	0.86±0.11
surprise	0.76±0.13	0.8±0.13	0.83±0.11	0.82±0.14
fear	0.84±0.11	0.94±0.12	0.96±0.10	0.95±0.10
disgust	0.75±0.10	0.86±0.10	0.88±0.11	0.9±0.11
bored	0.86±0.11	0.89±0.10	0.93±0.12	0.91±0.13

c) The result of the accuracy of emotion recognition: We can observe from Table 3, a large number of synthesized emotion speeches cannot be recognized with 2-dimensional emotion vectors to represent emotion labels. It proves that 2-dimensional emotion vectors cannot distinguish all 8 emotions well. Using 4-dimensional emotion vectors can represent most emotions, but still has a high error rate in the recognition of surprise and bored. With 8-dimensional and 16-dimensional emotion vectors as the input of the model, all different emotion speech can be distinguished still has a high error rate.

TABLE III. PERCENTAGE TABLE OF EMOTION RECOGNITION

emotion	MET-2 (%)	MET-4 (%)	MET-8 (%)	MET-16 (%)
neutral	100	90	100	100
happy	40	100	100	100
angry	50	100	100	100
sad	50	100	100	100
surprise	10	70	95	95
fear	100	100	100	100
disgust	60	100	100	100
bored	0	70	100	100

V. CONCLUSION

In this paper, we established a Mongolian emotional corpus and achieved the Mongolian emotional speech synthesis model for the first time. For the problem of insufficient Mongolian emotional corpus, we propose the method of transfer learning to accomplish pre-training, which can reduce the demand for emotional corpus effectively. Then, it is proposed to use emotional labels as the input of the Mongolian speech synthesis model to achieve the Mongolian emotional speech synthesis system. And in order to discuss the influence of different dimensions of emotion vectors on the quality of the voice, we came to the conclusion that it is the most suitable for emotion labels represented by MET-8. These findings enhance the controllability of Mongolian end-to-end speech synthesis.

In future work, we will focus on how does the size of the pre-training corpus affects the model's performance. In addition, the research of synthesizing emotional speech with different strength based on fine-grained emotional speech synthesis is also be implemented.

ACKNOWLEDGMENT

This work has been supported by the National Key Research and Development Program (2018YFE0122900), the National Natural Science Foundation of China (61773224, 62066033), the Inner Mongolia Autonomous Region Applied Technology Research and Development Fund Project (2019GG372, 2020GG0046, 2021GG0158, 2020PT0002) funding, the Inner Mongolia Autonomous Region Achievement Transformation Project (2019CG028).

REFERENCES

[1] Liu R, Bao F, Gao G, et al. “Mongolian text-to-speech system based on deep neural network,” in National Conference on Man-Machine Speech Communication, NCMMSC, 2017, pp. 162-170.

[2] Y. Wang, R. Skerry-Ryan, D. Stanton, et al. “Tacotron: Towards End-to-End Speech Synthesis,” Interspeech, 2017, pp. 4006–4010.

[3] Shen J, Pang R, Weiss R 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), 2018, pp. 4779-4783.

[4] A. Gibiansky,W. Ping, K. Peng, et al. “Deep Voice 2: Multi-Speaker Neural Text-to-Speech,” Advances in Neural Information Processing Systems, (NIPS 2017), vol. 30, pp. 2962-2970.

[5] A. van den Oord, S. Dieleman, H. Zen, et al. “WaveNet: A generative model for raw audio,” in 9th ISCA Speech Synthesis Workshop, 2016 , pp. 125-125.

[6] Valin J M, Skoglund J. LPCNET. “Improving Neural Speech Synthesis through Linear Prediction,” in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 5891-5895.

[7] Li J, Zhang H, Liu R, et al. “End-to-end mongolian text-to-speech system,” 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2018, pp. 483-487.

[8] Liu Z, Bao F, Gao G. “Mongolian Grapheme to Phoneme Conversion by Using Hybrid Approach,” in Zhang M., Ng V., Zhao D., Li S., Zan H. (eds) Natural Language Processing and Chinese Computing. NLPCC 2018. Lecture Notes in Computer Science, vol. 11108. pp. 40-50.

[9] Skerry-Ryan R J, Battenberg E, Xiao Y, et al. “Towards end-to-end prosody transfer for expressive speech synthesis with tacotron,” in in Proceedings of the 35th International Conference on Machine Learning, PMLR , vol. 80,pp. 4693-4702, 2018.

[10] Wang Y, Stanton D, Zhang Y, et al. “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” in Proceedings of the 35th International Conference on Machine Learning, PMLR , vol. 80,pp. 5180-5189, 2018.

[11] Wu P, Ling Z, Liu L, et al. “End-to-end emotional speech synthesis using style tokens and semi-supervised training,” in 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2019, pp.623-627.

[12] Choi H, Park S, Park J, et al. “Multi-speaker emotional acoustic modeling for cnn-based speech synthesis,” in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 6950-6954.

[13] Y. Lee, A. Rabiee, and S. Lee, “Emotional End-to-End Neural Speech Synthesizer,” unpublished.

[14] Kwon O, Jang I, Ahn C H, et al. “An effective style token weight control technique for end-to-end emotional speech synthesis,” in IEEE Signal Processing Letters, vol. 26, no. 9, pp. 1383-1387, Sept. 2019.

[15] Tits N, El Haddad K, Dutoit T. “Exploring transfer learning for low resource emotional tts,” Intelligent Systems and Computing, vol. 1037,pp. 52-60, 2019.

[16] Um S Y, Oh S, Byun K, et al. “Emotional speech synthesis with rich and granularized control,” in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7254-7258.

[17] Li T, Yang S, Xue L, et al. “Controllable emotion transfer for end-to-end speech synthesis,” in 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2021, pp. 1-5.

[18] Zhu X, Yang S, Yang G, et al. “Controlling emotion strength with relative attribute for end-to-end speech synthesis,” in 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2019, pp. 192-199.

[19] Lei Y, Yang S, Xie L. “Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis,” in 2021 IEEE Spoken Language Technology Workshop (SLT), 2021, pp. 423-430.

[20] Tan C., Sun F., Kong T., Zhang W., Yang C., Liu C. “A Survey on Deep Transfer Learning,” Artificial Neural Networks and Machine Learning – ICANN 2018, vol. 11141, pp. 270-279.

[21] S. J. Pan and Q. Yang, “A survey on transfer learning,” in IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.

[22] E. Navas, I. Hernáez, and I. Luengo, “An objective and subjective study of the role of semantics and prosodic features in building corpora for emotional tts,” in IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 4, pp. 1117-1127, July 2006.

[23] Liu R, Bao F, Gao G. “Building mongolian tts front-end with encoder-decoder model by using bridge method and multi-view features,” in Gedeon T., Wong K., Lee M. (eds) Neural Information Processing. ICONIP 2019. Communications in Computer and Information Science, vol. 1143,pp. 642-651.

[24] Liu R, Sisman B, Li J, et al. “Teacher-student training for robust tacotron-based tts,” in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6274-6278.

[25] Rui Liu, Berrak Sisman, Feilong Bao* , et al. “Exploiting Morphological and Phonological Features to Improve Prosodic Phrasing for Mongolian Speech Synthesis,” in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp.274-285.