Two-Stage Multi-lingual Speech Emotion Recognition for Multi-lingual Emotional Speech Synthesis

Xin Huang[1], Zuqiang Zeng[1], Chenjing Sun[1], and Jichen Yang[2(✉)]

[1] School of Electronics and Information Engineering,South China Normal University, Foshan 528225, China

[2] School of Cyber Security, Guangdong Polytechnic Normal University, Guangzhou 510665, China

nisonyoung@163.com

Abstract. In multi-lingual emotional speech synthesis, it is difficult to incorporate suitable emotional expressions in the synthesis process due to the differences between the emotional expressions of different linguals. In order to extract better emotional expressions of different linguals to assist the multi-lingual emotional speech synthesis, this paper conducts research on multi-lingual speech emotion recognition. In the current study of multi-lingual speech emotion recognition (SER), the combining method (TCM) and multi-task method (TMM) are the popular methods. However, good performance can't be obtained, the reason is that TCM doesn't consider the emotional difference of different linguals and it is not easy to train the good emotion recognition model and good language recognition model at the same time for TMM. In order to settle the issue, a two-stage multi-lingual SER method is proposed in this paper, wherein language recognition is to recognize the language type at the first stage, and then emotion recognition is applied at the second stage. In addition, wav2vec 2.0 is used as the input while ResNet18 is selected as the model for language recognition and emotion recognition respectively. The experimental results show that the proposed method can work on multi-lingual SER, meanwhile, the proposed method performs better than TCM and TMM.

Keywords: Speech emotion recognition · Multi-lingual · Emotional speech synthesis

1 Introduction

In recent years, the performance of text-to-speech (TTS) systems in terms of quality and naturalness of synthesized speech has improved significantly [1,2]. With globalization, bilinguals and polyglots are becoming a common trend in today's world, which makes speech communication more complex. In response to this trend, the performance of speech analysis tools such as emotional speech

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2025
Published by Springer Nature Switzerland AG 2025. All Rights Reserved
J. Cai et al. (Eds.): SPNCE 2023, LNICST 525, pp. 202–211, 2025.
https://doi.org/10.1007/978-3-031-73699-5_14

synthesis needs to be further improved in terms of multi-lingual. However, due to the large differences between the emotional expressions of different linguals, the challenge of multi-lingual emotional speech synthesis is that it is difficult to incorporate emotional expressions suitable for various linguals in the synthesis process. A well-performing multi-lingual speech emotion recognition (SER) system can extract the emotion expressions from different linguals, which enables the multi-lingual emotional speech synthesis system to incorporate the emotion expressions containing the matching emotions during the speech synthesis, and thus improves the performance of the multi-lingual emotional speech synthesis system. Therefore, this paper is aimed at multi-lingual SER.

SER has been an important research topic in the field of speech signal processing. The goal of SER is to accurately recognize the emotion type for the input speech utterance under the trained model, where the model is trained by using training data and corresponding emotion labels. Many methods have been proposed for single-lingual SER. For example, the accuracy of emotion recognition can be improved by optimizing the structure of a single model after feature extraction of the audio information [3]. Optimized neural networks can also be used to recognize emotions in a single language [4]. In the multi-lingual emotion recognition task, different features of speech information are used to recognize and classify. For example, combining features such as formant peaks, intensity and pitch can improve the accuracy of emotion recognition [5].

To date, there have been two methods for multi-lingual SER: the combining method (TCM) [6] and the multi-task method (TMM) [7,8]. TCM trains the model by combing different-lingual training data with the same emotion as a type, which borrows the method of single-lingual SER for multi-lingual SER. While TMM regards the multi-lingual SER as two tasks, one is language recognition and the other is emotion recognition, in other words, the language recognition and emotion recognition models are trained at the same time.

Since TCM does not take into account the emotion differences of different linguals in multi-lingual SER, it usually fails to obtain good performance as in [6]. Theoretically speaking, TMM can obtain good performance if the emotion recognition model and the language recognition model are successfully trained in the training stage. However, it is not easy to train a good emotion recognition model and a good language recognition model at the same time, for example, it is very difficult to assign suitable weights to the loss function for different tasks.

There is a general belief that it is easier to train one model than to train two models at the same time using the same training data. In this regard, in order to address the multi-lingual SER issues, a two-stage multi-lingual SER method is proposed in this paper. It consists of two stages: the first stage is to recognize language while the second stage is to recognize emotion. Compared with TMM, we can see that both the proposed method and TMM regard multi-lingual SER as two tasks.

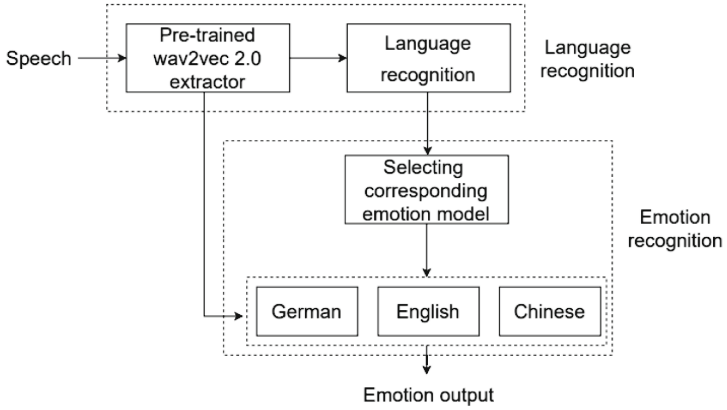

Fig. 1. The framework of the proposed two-stage multi-lingual SER.

2 Proposed Method

Figure 1 is the framework of the proposed two-stage multi-lingual SER method. From Fig. 1, it can be seen that the proposed two-stage multi-lingual SER consists of two parts: language recognition and emotion recognition, wherein language recognition is at the first stage while emotion recognition is at the second stage. Note that wav2vec 2.0 is used as the input for language recognition and emotion recognition because it is a well-known self-supervised representation and it has more useful information than some commonly used features.

In order to recognize the language type of the input speech signal, a language recognition model must be trained in advance. To do so, a three-class corresponding German, English and Chinese classifier is trained. Once the model training is finished, the language type of the input signal can be recognized under the model.

On the basis of language recognition, the corresponding emotion recognition model is selected to recognize emotion with wav2vec 2.0 of the input speech. To this end, three emotion recognition models (German, English and Chinese) are trained on their own training data and corresponding emotion labels respectively.

In the following, the two main models in the two-stage multi-lingual SER approach are described in detail, which are the pre-trained wav2vec 2.0 emotion representation extraction model and the recognition model used for both language and emotion recognition.

2.1 Pre-trained Wav2vec 2.0 Emotion Representation Extraction Model

As shown in Fig. 2, the pre-trained wav2vec 2.0 model consists of a multilayer convolutional feature encoder, a Transformer and a quantization module [9]. Where the multilayer convolutional feature encoder $f : X \rightarrow Z$ takes the raw

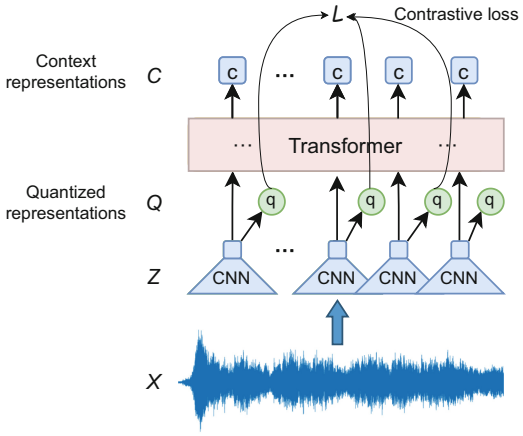

Fig. 2. The architecture of pre-trained W2V2 model.

audio X as input and outputs the latent speech representations $z_1, \ldots, z_T$ with time step T. The Transformer $g : Z \rightarrow C$ takes the latent speech representation Z as input, captures the global information of the sequence, and extracts the contextual representations $c_1, \ldots, c_T$ of the sequence. The quantization module $Z \rightarrow Q$ also takes the latent speech representation Z as input, discretizes it into q_t and outputs it. The outputs C and Q of the Transformer and the quantization module are used as inputs for the contrastive loss.

The feature encoder processes the input signal into low-level features, which consists of a temporal convolution, a layer normalization and a GELU activation function. The raw waveform input to the encoder is normalized to zero mean and unit variance.

Transformer extracts contextual representations in low-level features, it uses a convolutional layer as relative positional embedding, and captures global information using self-attention mechanisms.

Quantization module discretizes the low-level features into a finite set of speech representations through product quantization. The product quantization operation makes the features more robust and less susceptible to a small number of perturbations by splitting the infinite spatial space of feature representations into finite codebooks.

2.2 Recognition Model

ResNet18 is used to train the language recognition model and emotion recognition model in this work. The reason behind this is that ResNet18 has good performance in the task of classification. Residual network (ResNet) was proposed by Microsoft Research in 2015 [10]. The problem solved by this network is that as the depth of the network becomes deeper and deeper, the problem of network 'degradation' becomes obvious, that is, as the number of layers of

the network increases, the accuracy of the model begins to saturate, and then rapidly degrades. ResNet describes the training degradation problem and new solutions in detail on the ImageNet dataset.

The overall structure of ResNet consists of a number of residual units and bottleneck residual units, each of which contains a number of residual blocks. The input of the whole network is a picture, which undergoes multi-layer convolution operation and outputs a vector with corresponding class probabilities for classification tasks. With the increase of network depth, ordinary networks show higher training errors, while residual networks appear to be very easy to optimize. As the number of layers of the network increases, the residual network can easily achieve better accuracy, and the accuracy will be higher than the previous network. ResNet was originally intended for image recognition work. Neural network also has a good effect on SER [11]. In this paper, we modify ResNet to make it applicable to SER and classification [12]. Therefore, ResNet also has a good effect on voice information processing.

3 Database Introduction

The German speech data used in this paper is EMO-DB, a German emotional speech database consisting of recordings from ten actors (five males and five females) simulating seven target emotions [13], which are anger, neutral, fear, boredom, happiness, sadness and disgust, respectively.

The Chinese and English speech data used in the experiment were taken from the emotion speech database (ESD) [14], which consists of parallel phrases recorded by 10 native speakers for Chinese and English, respectively. These phrases covered five emotion categories, which are angry, happy, neutral, sad and surprise, respectively. Here, we call the Chinese part and English part in ESD as ESD-Chi and ESD-Eng, respectively.

Considering angry, happy, sad and neutral types appear in EMO-DB, ESD-Chi and ESD-Eng respectively. Thus, the four types of emotion in the three databases are selected to evaluate the proposed two-stage multi-lingual SER. The utterance number of each emotion type in the training, test and eva subsets of the three databases are given in Table 1.

4 Experimental Setup

The experimental deployment in this study was conducted using the Pytorch platform. The weighted accuracy (WA) and the unweighted accuracy (UA) are used as the evaluation metrics. The pre-trained wav2vec 2.0 model XLSR-53 is used in the experiment because it was trained on multi-lingual data [15].

After wav2vec 2.0 is extracted by the pre-trained model, three-class language recognition model is trained based on ResNet18, in the same way, ResNet18 is also trained to train emotion recognition models using the training data of EMO-DB, ESD-Chi and ESD-Eng respectively.

Table 1. The utterance number of each emotion type in the training, test and eva subsets of the three databases.

Database	Subset	Angry	Happy	Sad	Neutral
ESD-Chi	Training	3000	3000	3000	3000
	Test	300	300	300	300
	Eva	200	200	200	200
ESD-Eng	Training	3000	3000	3000	3000
	Test	300	300	300	300
	Eva	200	200	200	200
EMO-DB	Training	84	49	42	49
	Test	12	7	6	7
	Eva	24	14	12	14

5 Experimental Results and Analysis

Table 2 reports the experimental results on the three databases using the proposed two-stage multi-lingual SER method in terms of accuracy, UA and WA. Accuracy is used to calculate every emotion recognition accuracy in every database while UA and WA are used to evaluate all emotion recognition accuracy in one database.

Table 2. Experimental results on the three databases using the proposed two-stage multi-lingual SER method in terms of accuracy (%), UA (%) and WA (%).

Database	Accuracy				UA	WA
	Angry	Happy	Sad	Neutral		
ESD-Chi	94	98.5	96	99	96.87	96.87
ESD-Eng	96.5	84.5	87.5	92	90.12	90.12
EMO-DB	79.2	64.3	91.7	85.7	79.71	79.68
All	94.07	91	91.26	94.99	92.83	92.84

From Table 2, two conclusions can be drawn:

- The proposed two-stage multi-lingual SER on ESD-Chi performs better than that on ESD-Eng and that on EMO-DB in terms of UA and WA. In which, the result on EMO-DB gives the worst performance, the reason may be that there is not so much training data in EMO-DB.
- The proposed two-stage multi-lingual SER is able to recognize most of the emotion types in ESD-Chi in terms of accuracy while unable to recognize angry and happy types in EMO-DB, the reason is the same as mentioned above.

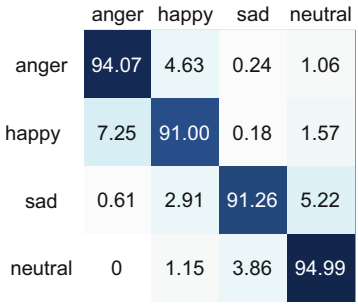

Fig. 3. Multi-lingual: Confusion matrix of the two-stage multi-lingual SER with UA of 92.83% and WA of 92.84% on multi-lingual data.

5.1 Confusion Matrix Analysis

Confusion matrix can clearly show the recognition effect of this paper's system on different emotions when inputting multi-lingual data. As shown in Fig. 3, the proposed system can achieve more than 90% recognition effect for each emotion, with the best recognition ability for anger and neutral emotions, and a slightly worse recognition ability for happy and sad emotions.

5.2 Comparison with Commonly Used Features

In this section, we would like to compare different inputs for the proposed two-stage multi-lingual SER. Commonly used features such as MFCC [16] and WavLM [17] are selected here. Table 3 gives the experimental results comparison between wav2vec 2.0 and commonly used features on the three databases in terms of UA and WA.

Table 3. Experimental results comparison between wav2vec 2.0 and commonly used features such as MFCC and WavLM in terms of UA (%) and WA (%).

Features	UA	WA
MFCC	78.26	78.28
WavLM	85.21	85.26
wav2vec2.0	92.83	92.84

As shown in Table 3, the proposed system with wav2vec 2.0 as the input significantly outperforms MFCC and WavLM in terms of UA and WA, which means that wav2vec 2.0 is more suitable for the task of multi-lingual SER. The reason may be that there is more emotion information in wav2vec 2.0 than that in MFCC and WavLM.

5.3 Comparison with State-of-the-Art Methods

To better validate the effectiveness of our proposed method, it is compared with state-of-the-art methods such as TCM and TMM. Wav2vec 2.0 is used as the input for both TCM and TMM, which is the same as the proposed method. Comparison results are given in Table 4.

From Table 4, it can be found that the proposed method performs better than TCM and TMM. Furthermore, TCM gives the worst performance. It means that the proposed method can correctly recognize more emotion types in multi-lingual SER. The reason is that TCM does not take into account the emotion differences between different linguals in multi-lingual SER. And it is not easy for TMM to train a good emotion recognition model and a good language recognition model at the same time.

Table 4. Comparison with state-of-the-art methods on the three databases in terms of UA (%) and WA (%).

Methods	UA	WA
TCM	89.66	89.67
TMM	91.37	91.38
The proposed method (Ours)	92.83	92.84

6 Conclusion

In order to improve the performance of the multi-lingual emotional speech synthesis system, this paper investigates its underlying multi-lingual SER. In order to address the issues of multi-lingual SER, a two-stage multi-lingual SER method is proposed, which consists of two stages, the first stage is to recognize the language and the second stage is to recognize emotion. In addition, wav2vec 2.0 is used as the input while ResNet18 is selected as the model for both language recognition and emotion recognition. The experimental results show that the proposed method can work on multi-lingual SER, meanwhile, the proposed method performs better than state-of-the-art methods such as TCM and TMM. This means that the multi-lingual emotional speech synthesis system can achieve better performance by using the emotional representations extracted from the multi-lingual SER system in this paper as inputs in the final synthesis phase.

Acknowledgments. This work was supported by NSFC(62001173, 62171188). The authors gratefully acknowledge the support of 2022 Guangdong Hong Kong-Macao Greater Bay Area Exchange Programs of South China Normal University (SCNU).

References

1. Wang, Y., Skerry-Ryan, R.J., Stanton, D., et al.: Tacotron: towards end-to-end speech synthesis. In: Proceedings of Interspeech 2017, pp. 4006–4010 (2017)
2. Lei, Y., Yang, S., Wang, X., Xie, L.: MsEmoTTS: multi-scale emotion transfer, prediction, and control for emotional speech synthesis. IEEE/ACM Trans. Audio Speech Lang. Process. **30**, 853–864 (2022)
3. Zayene, B., Jlassi, C., Arous, N.: 3D convolutional recurrent global neural network for speech emotion recognition. In: 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Sousse, Tunisia, pp. 1–5 (2020)
4. Kong, Q., Cao, Y., Iqbal, T., et al.: PANNs: large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 2880–2894 (2020)
5. Meftah, A., Alotaibi, Y., Selouani, S.-A.: Emotional speech recognition: a multilingual perspective. In: 2016 International Conference on Bio-engineering for Smart Technologies (BioSMART), Dubai, United Arab Emirates, pp. 1–4 (2016)
6. Yadav, A., Vishwakarma, D.K.: A multilingual framework of CNN and Bi-LSTM for emotion classification. In: 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, pp. 1–6 (2020)
7. Sharma, M.: Multi-lingual multi-task speech emotion recognition using wav2vec 2.0. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, pp. 6907–6911 (2022)
8. Yue, P., Qu, L., Zheng, S., Li, T.: Multi-task learning for speech emotion and emotion intensity recognition. In: 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Chiang Mai, Thailand, pp. 1232–1237 (2022)
9. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: a framework for self-supervised learning of speech representations. In: Advances in Neural Information Processing Systems, vol. 33, pp. 12449–12460 (2020)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 770–778 (2016)
11. Qayyum, A.B.A., Arefeen, A., Shahnaz, C.: Convolutional neural network (CNN) based speech-emotion recognition. In: 2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON), Dhaka, Bangladesh, pp. 122–125 (2019)
12. Zhang, Z., Zhang, X., Guo, M., et al.: A multilingual framework based on pre-training model for speech emotion recognition. In: 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan, pp. 750–755 (2021)
13. Burkhardt, F., Paeschke, A., Rolfes, M., et al.: A database of German emotional speech. In: Proceedings of Interspeech 2005, pp. 1517–1520 (2005)
14. Zhou, K., Sisman, B., Liu, R., Li, H.: Emotional voice conversion: theory, databases and ESD. Speech Commun. **137**, 1–18 (2022)
15. Conneau, A., Baevski, A., Collobert, R., et al.: Unsupervised cross-lingual representation learning for speech recognition. In: Proceedings of Interspeech 2021, pp. 2426–2430 (2021)

16. Latif, S., Rana, R., Khalifa, S., et al.: Survey of deep representation learning for speech emotion recognition. IEEE Trans. Affect. Comput. **14**(2), 1634–1654 (2021)
17. Chen, S., Wang, C., Chen, Z., et al.: WavLM: large-scale self-supervised pre-training for full stack speech processing. IEEE J. Sel. Top. Signal Process. **16**(6), 1505–1518 (2022)