INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Speech Emotion Recognition using Decomposed Speech via Multi-task Learning

Jia-Hao Hsu, Chung-Hsien Wu, and Yu-Hung Wei

Department of Computer Science and Information Engineering
National Cheng Kung University, Taiwan
jiahaoxuu@gmail.com, chunghsienwu@gmail.com, cool.daniel1234@gmail.com

Abstract

In speech emotion recognition, most recent studies used powerful models to obtain robust features without considering the disentangled components, which contain diverse emotion-rich information helpful for speech emotion recognition. In this study, an autoencoder is used as the speech decomposition model to obtain the disentangled components, including content, timbre, pitch, and rhythm features, which are regarded as emotion-rich features, for speech emotion recognition. The mechanism of multi-task training is then used to train the tasks of speech emotion recognition, speaker recognition, speech recognition, and spectral reconstruction at the same time, while exploiting commonalities and differences across tasks. The model proposed in this study achieved an accuracy of 77.50% on the four-classes emotion recognition task of IEMOCAP. Experiments showed that the proposed methods can effectively improve speech emotion recognition performance, outperforming the SOTA approach.

Index Terms: speech emotion recognition, decomposed speech, multi-task learning

1. Introduction

Human-computer interaction has gradually become a daily need, and how to make these machines respond more like humans is also one of the ongoing studies [1-4]. Emotion recognition technology is one of the ways to humanize these machines. It includes various signal modalities, such as speech signals, semantic words, facial expressions, and other physiological signals. Among them, speech is a kind of signal that is relatively easy to obtain in daily life. Speech provides rich emotional information as well as the semantic information [5]. Despite the breakthroughs in speech processing, there are still some limitations in speech emotion recognition (SER). Especially, extracting discriminating emotion features has always been a challenging task.

In recent years, due to the development of deep learning technology, more and more studies have begun to use deep models to extract emotion features [6, 7]. These studies used neural networks to obtain the relationship between sounds and emotional targets and achieved satisfactory performance [6, 7]. The well-known Wav2vec-2.0 model [8] as the feature extraction model in automatic speech recognition (ASR) has been widely used to extract emotion features [7, 9] and achieved good results. As human speech conveys a rich stream of information, the decomposed components, including content, rhythm, timbre and pitch, could be separately employed for speech emotion recognition, removing interference introduced by irrelevant components. In the disentangled components, the language content is highly related to semantic representation of emotion, timbre is closely connected with the speaker's identity, while pitch and rhythm express the emotion of the speaker. Therefore, this study uses the decomposed speech components to obtain discriminating emotion features. In addition, multi-task learning is a training method to improve the model effect, and it has been widely used in many fields [10]. Existing studies [11, 12] indicated that multi-task learning considering both ASR and SER results on Wav2vec-2.0 can further improve the performance of SER. Taking the advantage of multi-task learning, this study trains the tasks of speech emotion recognition, speaker recognition, speech recognition, and spectral reconstruction at the same time to obtain a better performance of speech emotion recognition.

The contributions of this study are mainly divided into two parts. First, the mainstream speech feature extraction model Wav2vec-2.0 is used as the system backbone. Under this model, this study adopts the ASR, spectral reconstruction and speaker recognition task to enhance Wav2vec-2.0. Second, a speech decomposition model is considered to extract the speech emotion features. Compared with the existing decomposition model [13], this study includes the speaker recognition task, so that the autoencoder can consider different timbre between speakers. The experiments showed that the method proposed in this study can achieve better performance.

2. Proposed methods

The proposed system architecture is shown in Figure 1. It is divided into data pre-processing, feature extraction, and downstream task modelling.

2.1. Pre-processing

For different feature extraction models, the audio should be firstly pre-processed to fit the corresponding input type. In the speech decomposition model, the input speech is converted into spectrogram and pitch contour. The short-time Fourier transform (STFT) is used to obtain the spectrogram. The speech information contained in the pitch contour is the prosody information consisting of pitch, speaking style and speed. This study uses the Robust Algorithm for Pitch Tracking (RAPT) [14] to obtain pitch contour. Referring to the SpeechSplit model [13], the extracted spectrogram and pitch contour are perturbed using Random Resampling (RR) [15] and fed to different extraction models. Resampling cuts the sound information into paragraphs of equal length and the sound paragraphs are used to disturb the speech speed. In the training process, as the length of the speech signal is perturbed, the model can focus on learning the context. This part is described in more detail in Section 2.2.

2.2. Feature Extraction

This study uses Wav2vec-2.0-base model to extract speech features. Wav2vec-2.0 is pre-trained on a large amount of unlabeled data by restoring masked frames and applying

10.21437/Interspeech.2023-396

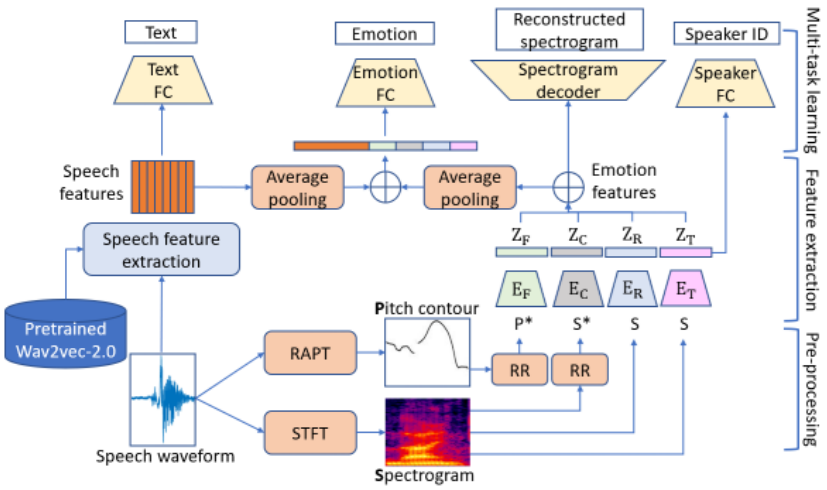
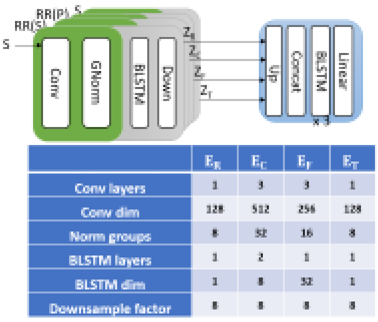


Figure 1: *The overall architecture of the proposed system.*

contrastive learning. Like large pre-trained language models, Wav2vec-2.0 has considerable understanding of the relationship between speech frames. The input of Wav2vec-2.0 is the entire signal. The architecture includes a convolutional-layer-based encoder (7-layer convolutional layer) and a 12-layer transformer to finally output a 768-dimensional vector.

SpeechSplit is a model capable of free transformation of speaker characteristics, pitch, and prosody. The model utilizes three tuned encoders and a decoder that reconstructs the spectrogram. The input of SpeechSplit is the spectrogram, resampled spectrogram, and resampled pitch contour. In Figure 1, E_R (Encoder R) is responsible for extracting the features containing rhythm, E_C (Encoder C) is responsible for the features of content, and E_F (Encoder F) is responsible for the features of pitch. The input of E_R is the raw spectrogram containing the information of prosody, content, pitch, and timbre, which can be regarded as comprehensive prosodic features. The input of E_C is the resampled spectrogram, which destroys the prosody and speech speed information. The input of E_F is the resampled pitch contour, whose speech speed is destroyed so that E_F only focuses on the characteristics of the pitch. Compared with SpeechSplit, this study has one more encoder, E_T (Encoder T). The purpose of the additional encoder is to encode the features containing the speaker's timbre. As shown in Figure 2, the encoders of the speech decomposition model include convolutional layers, normalization layers, and a bi-LSTM. The four disentangled features Z_R, Z_C, Z_F, and Z_T, which represent rhythm, content, pitch, and timbre respectively, are concatenated and then used as the input of the decoder to reconstruct the spectrogram. The decoder consists of a bi-LSTM and a linear layer.



	E_R	E_C	E_F	E_T
Conv layers	1	3	3	1
Conv dim	128	512	256	128
Norm groups	8	32	16	8
BLSTM layers	1	2	1	1
BLSTM dim	1	8	32	1
Downsample factor	8	8	8	8

Figure 2: *The architecture of the decomposition model*

2.3. Multi-task learning

The tasks for multi-task learning in this study consists of speech recognition, emotion recognition, spectral reconstruction, and speaker recognition. The speech feature extraction model Wav2vec-2.0 and the four speech decomposition encoders are fine-tuned by the four tasks. The training loss of the speech recognition task is shown in Equation (1) based on the CTC loss function to calculate the result obtained from the Wav2vec-2.0 and the fully connected layer (FC). Where N represents the total number of samples, and T represents the time domain length. a_{it} represents one of the alignments of the correct answer for sample X_i, and p_{it} represents the probability of that alignment. The training loss of the emotion recognition task is shown in Equation (2), the cross-entropy loss function. The emotion prediction is obtained by concatenating the output of Wav2vec-2.0 and the decomposed speech features followed by feeding them to the FC. Where p_{ie} presents the predicted emotion profile in emotion class e of sample i, and y_{ie} presents the label of sample i. The training loss of the spectral reconstruction task is the mean square error (MSE), as shown in Equation (3). The MSE is calculated from the spectrogram reconstructed by the decoder and the original spectrogram of the speech. The loss of speaker recognition task is also a cross-entropy loss as shown in Equation (4). The features obtained by the E_T encoder are fed to the FC to obtain the predicted speaker ID, which is then used for loss calculation with the real speaker. Where U presents the number of speakers. The spectral reconstruction task and speaker recognition task are first used to pre-train four speech decomposition encoders. Then the multiple task learning is adopted to fine-tune the whole system. The fine-tuning loss of the system is shown in Equation (5). Where α means the weight of the task.

$$L_{CTC} = -\frac{1}{N}\sum\prod_{t=1}^{T} p_{it}(a_{it}|X_i) \qquad (1)$$

$$L_E = -\frac{1}{N}\sum_{i=1}^{N}\sum_{e=1}^{E} y_{ie}\log(p_{ie}) \qquad (2)$$

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \qquad (3)$$

$$L_S = -\frac{1}{N}\sum_{i=1}^{N}\sum_{s=1}^{U} y_{is} \log{(p_{is})} \tag{4}$$

$$L_{mul} = \alpha_S L_S + \alpha_{MSE} L_{MSE} + \alpha_{CTC} L_{CTC} + \alpha_E L_E \tag{5}$$

3. Experiments

This section presents the experimental setup and experimental results of this study. The experimental setup includes the evaluated dataset and evaluation metrics. In the experimental results, this study analyzed the training loss of the pre-training model. Through ablation study, this study analyzed the improvement of SER by the method proposed in this study. Finally, the proposed method was compared with other existing methods.

3.1. Experimental Setup

The IEMOCAP dataset was adopted to evaluate the proposed approaches in this study. The IEMOCAP dataset was recorded by a total of 10 actors. The database is divided into scripted scenarios and improvisations according to the recording method. The emotional classes include a total of 10 emotions. Referring to the current mainstream evaluation, this study used four relatively balanced emotion classes consisting of angry, sadness, happiness, and neutral for evaluation. The data distribution of the four emotion classes is shown in Table 1.

Table 1: *The data distribution of IEMOCAP.*

Emotion class	Count	Duration (mins)
Anger	1103	83.0
Sadness	1084	99.3
Happiness	1636	126.0
Neutral	1708	111.1
Total	5531	419.3

The hyperparameter of the pre-training of the speech decomposition model included batch size, learning rate and decomposed loss weight. The batch size was 16. The learning rate was 10^{-4}. The optimizer is Adam optimizer. The decomposed loss weight controlled the importance of the losses of spectral reconstruction and speaker recognition as shown in equation (7).

$$L_{decompose} = \omega \cdot L_S + (1-\omega) \cdot L_{MSE} \tag{7}$$

The pre-trained speech feature extraction model we used was the Wav2vec-2.0-base from the Huggingface Transformers [16], and we fine-tuned it with the speech decomposition model together using the IEMOCAP. The learning rate was 10^{-5}. The optimizer is Adam optimizer. We used 1 GPU (Titan RTX) with a batch size of 1 and trained the model for a total of 50 epochs.

For evaluation, ten-fold cross-validation method was adopted to conduct experiments. There are four evaluation metrics, spectrogram reconstruction MSE (Speech MSE), automatic speech recognition accuracy (ASR Acc.), speaker recognition accuracy (Speaker Acc.), and emotion recognition unweighted accuracy (SER Acc.). The Speech MSE is shown in Equation (3). The ASR accuracy is 1-(word error rate). Both Speaker accuracy and SER accuracy are calculated as the correct amount of predictions in all prediction data. And the SER accuracy is the main goal of this study.

3.2. Pre-training of speech decomposition model

The pre-training of the speech decomposition model has two outputs, the reconstructed spectrogram and the speaker ID. Two loss functions were calculated, namely the L_{MSE} and the L_S. Figure 3 presents the two losses for the pre-training of the speech decomposition model. We pre-trained the model for 1 million steps. The training of the two losses tended to be stable, and the reconstructed spectrogram and the speaker recognition results were relatively stable. The weight ω in equation (7) was adjusted in the interval 0 to 1 to obtain the best pre-trained speech decomposition model. Finally, we set it as 0.1 and it achieved the best L_S and L_{MSE}. This pre-trained model was used as an emotion feature extraction model for subsequent emotion recognition.

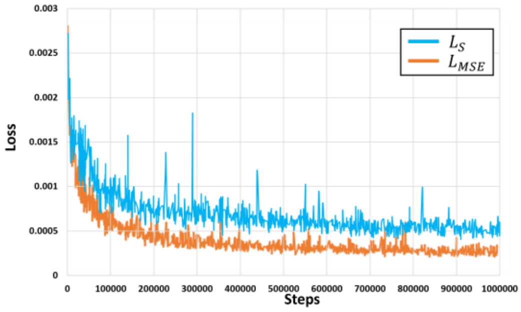


Figure 3: *Two losses for the pre-training of the speech decomposition model*

3.3. Evaluation of Speech Decomposition Features and Multi-task Learning

For spectral reconstruction task, Table 2 shows the ablation studies of the four disentangled components. The components were extracted as features and concatenated with the speech feature respectively. Only spectral reconstruction task, SER task and speaker recognition task (only for Z_T) were conducted to fine-tune the system here. The results indicated which components is more important and useful for SER.

Table 2: *Comparison of the disentangled components.*

Components	Speech MSE	SER Acc.
Z_F	8.2e-4	66.17%
Z_C	10.9e-4	65.45%
Z_R	7.7e-4	66.38%
Z_T	9.7e-4	65.71%
All	5.3e-4	76.06%

It can be seen that rhythm (Z_R) and pitch (Z_F) have higher importance for speech emotion recognition. Since the input of rhythm features is not disturbed, the spectrogram can be better restored. Since timbre (Z_T) was also fine-tuned on the speaker recognition task, the degree of spectral restoration is lower than that of rhythm and pitch. Although using content and timbre alone are not well at recognizing emotions, summing up all features can get the best results. This suggests that these disentangled components complement each other to improve emotion recognition.

The subsequent experiments verified whether multi-task learning can outperform single-task learning during model training, and the performance improvement that each task can bring. Table 3 shows the ablation studies of multi-task and methods in this study.

Table 3: *Comparison of the various tasks of multi-task learning.*

Methods	Speaker Acc.	Speech MSE	ASR Acc.	SER Acc.
M1	-	-	-	75.01%
M2 [11]	-	-	76.58%	75.40%
M3	98.4%	5.3e-4	-	76.09%
M4	-	5.7e-4	78.01%	75.67%
Ours	98.6%	5.5e-4	77.59%	**77.50**%

In Table 3, the first baseline method (M1) is a single-task emotion recognition model, which used only emotion recognition as the target to fine-tune the Wav2vec-2.0 model. The second method (M2) added the ASR task to the first baseline model. When the loss weights (α_E, α_{CTC}) were set as (0.9, 0.1), M2 could achieve best SER results. It can be found that adding ASR task can indeed improve the performance of speech emotion recognition, and the model understands emotional differences from the semantics of speech. The third method (M3) is the same as the best method in Table 2. It included the Wav2vec-2.0 model and the four decomposition models. In particular, the ASR task was not used in this method. Compared with M2, M3 included emotional features and achieved better performance. It indicated that adding speech decomposed feature improve more performance than adding ASR task. The fourth method (M4) consisted of the Wav2vec-2.0 speech feature extraction model and the emotion feature extraction model of the auto-encoder. This emotion feature extraction model did not decompose the speech input into four components, but directly encoded features and reconstructed the spectrogram with one encoder and one decoder. When the loss weights $(\alpha_E, \alpha_{MSE}, \alpha_{CTC})$ were set as (0.4, 0.4, 0.2), M4 could achieve best SER results. Compared with the first baseline model, M4 model considering emotion features improved the performance of emotion recognition. And the improvement brought by emotion features was higher than that of ASR. Compared with M2 model, adding spectral reconstruction make M4 achieve higher ASR accuracy. Finally, the system proposed in this study considered the multi-task results and used four encoders for the speech decomposition model. The best setting of loss weights $(\alpha_E, \alpha_{MSE}, \alpha_{CTC}, \alpha_S)$ were (0.45, 0.45, 0.05, 0.05). This method of subdividing the encoder allows the models to handle the corresponding tasks individually, and the extracted features can provide cleaner information. And it can be seen that the improvement of adding four disentangled components is higher than that of using an auto-encoder for spectral reconstruction. Due to the setting of loss weights, the fine-tuning of the system was more focused on the spectral reconstruction. That caused the ASR accuracy of our final method a little lower than that of M4. From this table, we can see that the method proposed in this study can indeed improve the performance of speech emotion recognition.

3.4. Comparison of existing studies

To verify if this study is comparable to the existing SER studies, the other studies on IEMOCAP corpus were selected, and the experimental methods and metric were consistent. These studies used the same amount of four classes of emotion data, the same 10-fold cross-validation setting (leave-one-speaker-out), and the same unweighted accuracy as the evaluation metric. The recognition result of this study was much higher than those of other studies. The use of a very powerful pre-trained speech model, Wav2vec-2.0 model, may also be responsible for this result. By decomposing the speech spectrogram, the obtained features of our system were much effective than that using a single autoencoder in [17], so it performed better in emotion recognition. In comparison with [18], it was helpful for this study to use an ASR task to increase text-related information. And the disentangled features derived from speech decomposition were also more reliable than statistical features in emotional expression. The result of "SER with MTL" was the result we implemented using the code they provided on GitHub. It was also the M2 method in Table 3 we have compared. With their excellent multi-task learning, this study added more emotion-related tasks to further improve the recognition performance.

Table 4: *The performance of the existing related studies.*

The existing studies	SER Acc.
ProgNet [19]	65.70%
Multi-task CNN [17]	65.60%
Multi-task AAE [17]	68.80%
Two-stream AE [18]	71.86%
SER with MTL [11]	75.40%
Ours	**77.50**%

4. Conclusions

This study proposed a new autoencoder-based model to decompose the speech signal into four speech components, namely rhythm, content, pitch, and timbre. We also combined the pre-trained Wav2vec-2.0 model for speech feature extraction in emotion recognition. This system considers both the power of the pre-trained speech feature extraction model and the importance of traditional emotion features. In particular, this study decomposes speech into four disentangled components as emotion features. And the mechanism of multi-task learning which included emotion recognition task, spectral reconstruction task, automatic speech recognition task and speaker recognition task is used to make the model consider the effect of various aspects to improve the performance of speech emotion recognition.

It can be seen from the experimental results that decomposing speech into emotion-rich features and training with multi-task learning can improve the performance even more. On the public benchmark, IEMOCAP, the performance proposed in this study achieved an unweighted accuracy of 77.50% in four-class speech emotion recognition. The method proposed in this study is indeed helpful for emotion recognition. We provided our code link, https://reurl.cc/XLWWae, which was built on Github.

5. References

[1] S. Kwon, "Att-Net: Enhanced emotion recognition system using lightweight self-attention module," Applied Soft Computing, vol. 102, p. 107101, 2021.

[2] E. Adamopoulou and L. Moussiades, "An overview of chatbot technology," in IFIP International Conference on Artificial Intelligence Applications and Innovations, 2020: Springer, pp. 373-383.

[3] C. Crolic, F. Thomaz, R. Hadi, and A.T. Stephen, "Blame the bot: anthropomorphism and anger in customer–chatbot interactions," Journal of Marketing, vol. 86, no. 1, pp. 132-148, 2022.

[4] J.-H. Hsu, M.-H. Su, C.-H. Wu, and Y.-H. Chen, "Speech emotion recognition considering nonverbal vocalization in affective conversations," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1675-1686, 2021.

[5] A.B. Ingale and D. Chaudhari, "Speech emotion recognition," International Journal of Soft Computing and Engineering (IJSCE), vol. 2, no. 1, pp. 235-238, 2012.

[6] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," arXiv preprint arXiv:1904.03833, 2019.

[7] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," arXiv preprint arXiv:2111.02735, 2021.

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," Advances in Neural Information Processing Systems, vol. 33, pp. 12449-12460, 2020.

[9] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," arXiv preprint arXiv:2104.03502, 2021.

[10] Y. Zhang and Q. Yang, "An overview of multi-task learning," National Science Review, vol. 5, no. 1, pp. 30-43, 2018.

[11] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech emotion recognition with multi-task learning," in Interspeech, 2021.

[12] J. Seo and B. Lee, "Multi-Task Conformer with Multi-Feature Combination for Speech Emotion Recognition," Symmetry, vol. 14, no. 7, p. 1428, 2022.

[13] K. Qian, Y. Zhang, S. Chang, M. Hasegawa-Johnson, and D. Cox, "Unsupervised speech decomposition via triple information bottleneck," in International Conference on Machine Learning, 2020: PMLR, pp. 7836-7846.

[14] D. Talkin and W.B. Kleijn, "A robust algorithm for pitch tracking (RAPT)," Speech coding and synthesis, vol. 495, p. 518, 1995.

[15] A. Polyak and L. Wolf, "Attention-based wavenet autoencoder for universal voice conversion," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019: IEEE, pp. 6800-6804.

[16] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, and M. Funtowicz, "Transformers: State-of-the-art natural language processing," in Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020, pp. 38-45.

[17] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, and B.W. Schuller, "Multi-task semi-supervised adversarial autoencoding for speech emotion recognition," IEEE Transactions on Affective computing, 2020.

[18] C. Zhang and L. Xue, "Two-stream Emotion-embedded Autoencoder for Speech Emotion Recognition," in 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 2021: IEEE, pp. 1-6.

[19] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E.M. Provost, "Progressive neural networks for transfer learning in emotion recognition," arXiv preprint arXiv:1706.03256, 2017.