Does the Inclusion of Other Modalities Enhance the Performance of Speech Emotion Recognition Systems?

Junchen Liu[1], Jesin James[1], Karan Nathwani[2]

[1]The University of Auckland, [2]India Institute of Technology Jammu

jliu522@aucklanduni.ac.nz, jesin.james@auckland.ac.nz, karan.nathwani@iitjammu.ac.in

Abstract

The pursuit of natural human-computer interaction has driven the advancement of emotion recognition technology. Speech emotion recognition (SER) has gained widespread attention due to its high applicability. Recently, some researchers have been interested in developing multi-modal emotion recognition (MER) systems that integrate speech with text and video modalities to enhance robustness and accuracy. We analyse the performance of these systems using the IEMOCAP and RAVDESS datasets, highlighting the impact of different modality combinations on emotion recognition accuracy. This paper aims to guide future research in optimising MER by leveraging the complementary advantages of various modalities.

Index Terms: speech emotion recognition, multi-modal emotion recognition, impact of different modality combinations

1. Introduction

Emotion recognition (ER) technology could improve the feasibility of human-computer interaction in real-world applications; it enables computers and other intelligent devices to understand and analyse the emotional state of users, thereby providing personalised and humanised services. However, due to the diversity, complexity and subjectivity of human emotional expression, the implementation of ER technology faces numerous challenges and difficulties.

Speech emotion recognition (SER) systems have received significant attention due to their potential applications in various fields, such as mental health monitoring and customer service enhancement. In recent years, there has been increasing enthusiasm about incorporating text or video modalities into SER models to develop multi-modal emotion recognition (MER) systems. MER systems concurrently process information from multiple modalities, which may increase the stability of the system [1]. For example, when the speech modality is unable to effectively predict the emotional states due to the environmental noise, the intervention of video or text modality can provide valuable emotional information to the MER systems. In previous research on MER tasks, rarely studies have identified which modality contributes most to emotion recognition. However, it is crucial to investigate the impact of each modality on recognition accuracy. These explorations can enhance the understanding of the complementarity and contributions of each modality, thereby simplifying model design by removing unnecessary modalities, reducing system complexity and improving computational efficiency.

Adopting deep learning techniques, especially deep neural networks (DNN), has become a prevalent trend in ER research. DNN, characterised by deep network structures and a substantial number of parameters, can automatically learn complex representations and extract relevant features from input data, capturing subtle emotional cues that may be missed by hand-crafted features [2].

In summary, this paper's primary objective is to explore different modalities' contributions to the performance of MER systems and answer the following research questions:

1. What are the advantages and disadvantages of each modality (speech, video, text) in detecting various emotional states?
2. How does combining different modalities affect the overall performance of MER systems?

2. Literature Review

This section will elaborate on the datasets and the methods of the state-of-the-art ER systems.

2.1. Datasets

Researchers in the field of SER and MER often select English datasets such as the interactive emotional dyadic motion capture (IEMOCAP) [3] and ryerson audio-visual database of emotional speech and song (RAVDESS) [4] as their primary resources for model construction. These datasets are favoured due to the rich emotional content and multiple modalities, offering robust support for developing effective ER systems. The details of these datasets are shown in Table 1.

Table 1. *IEMOCAP and RAVDESS dataset's information. S, V, T, H, F represents speech, video, text, hand movement and facial expression respectively.*

Dataset	Modalities	Utterances	Emotions
IEMOCAP [3]	S, V, T, H, F	7529	9
RAVDESS [4]	S, V	1440	8

Specifically, the IEMOCAP dataset records the performance data of 10 actors (5 female, 5 male). It has 9 emotions, such as neutral (1708), happy (595), sad (1084), anger (1103), surprise (107), fear (40), disgust (2), frustrated (1849) and excitement (1041). In total, the database contains approximately twelve hours of data. Due to the fact that IEMOCAP is an unbalanced dataset, some researchers tend to use only four categories to construct balanced datasets; they combine happiness and excitement as the happy category, and the other three classes are anger, sadness and neutral.

RAVDESS dataset consists of 1440 utterances by 24 professional actors (12 female, 12 male), encompassing eight emotion types: calm, happy, sad, angry, fearful, surprised, disgusted, and

neutral. RAVDESS is the balanced dataset, with 192 utterances for each emotion except for neutral, which has 96 utterances.

2.2. Speech Emotion Recognition

Acoustic information, such as tone, pitch, and frequency, contains rich emotional cues that can precisely express human emotional states. This information in speech forms the foundation for the application of SER systems in various fields, including intelligent customer service, personal assistants, and market research. However, the performance of SER systems is often affected by the environmental noise [5]. To address this issue, researchers are dedicated to developing SER systems that exhibit robust noise resistance, ensuring reliable and accurate emotion detection in diverse and noisy environments.

Convolutional neural networks (CNN) have gradually become the primary model for SER tasks. This is because CNN models can capture local and multi-level features with translation invariance in speech signals; these features will help improve the robustness and noise resistance of SER systems [6]. Most researchers choose to extract speech features from raw speech waveforms [7] and mel-spectrograms [8], or transfer the speech signal into the mel-frequency cepstral coefficients (MFCC) [9] as the input.

Additionally, some researchers illustrate that the transformer model based on DNN is suitable for processing long-time sequences of speech data. This is because the transformer model possesses a self-attention mechanism, enabling it to dynamically focus on the most relevant information at various positions in the input sequence, thereby improving the model's understanding of speech data and filtering out noise [10]. For example, Chen et al. [11] utilises Wav2Vec-2.0 to extract speech features from the raw speech waveform, and authors in [12] evaluate the effectiveness of different transformer models in SER, such as HuBERT, Wav2vec-2.0, and WavLM. In addition, the transformer-based model designed explicitly for log-mel spectrograms, as outlined in [13], has demonstrated high recognition accuracy in handling SER tasks.

2.3. Multi-modal Emotion Recognition

This section primarily introduces the feature extraction methods for text and video modalities in MER systems, as well as the techniques for feature fusion. The feature extraction method for the speech modality remains consistent with the approach described in the previous section and will not be reiterated here.

Text data has the characteristics of accessible collection, storage, and processing. Meanwhile, different languages and cultures have specific ways of expressing emotions, text-based ER systems can utilise specific emotional vocabulary and phrases in these languages to improve the accuracy of emotion recognition. However, it is worth noting that polysemy in words can reduce the effectiveness of text-based ER systems [14]. Therefore, developing advanced text feature extraction methods that capture contextual information is essential for better understanding the meaning of the text.

The GloVe word vector model has been proven effective in text feature extraction. Rajan et al. [15] proposed a method that combines GloVe with the bidirectional gated recurrent unit (BiGRU) for text feature extraction, aiming to capture contextual dependencies in the text and enhance the model's ability to process long sequences. Additionally, embeddings from language models V2 (ELMoV2) [16], built on bidirectional long short-term memory (BiLSTM), provide context-aware word embeddings, meaning that the representation of each word dynamically changes based on its context within the sentence. This allows for better handling of polysemy and synonyms, enhancing the model's understanding capabilities. Currently, using transformer models for text feature extraction has garnered widespread attention. Transformer models, such as bidirectional encoder representations from transformers (BERT) [17] and robustly optimized BERT pretraining approach (RoBERTa) [18], possess self-attention mechanisms that can capture global dependencies between words in a text. These models also provide efficient parallel processing capabilities and strong bidirectional context understanding. These advantages make transformer models excel in text feature extraction tasks.

Facial images extracted from videos are commonly used as inputs for video feature extraction. Facial expressions are a direct and natural way of expressing emotions, different emotional states, such as happiness, sadness and anger, are usually manifested through changes in facial muscles, which are easily captured and recognised. Additionally, some microscopic details that indicate changes in emotional states, such as eye movements or the corners of the mouth turning up or down, can only be captured through facial expression. Notably, most research on video-based ER is affected by the issue of facial occlusion [19].

Recently, methods for extracting video features mainly revolve around CNN models. For example, [20] and [21] utilises the ability of 3-dimensional CNN (3D CNN) models to simultaneously capture spatial and temporal features to process a series of consecutive frames, thereby understanding the changes in facial expressions and movements in videos. On the basis of the 3D CNN model, [22] adds long short-term memory (LSTM) to gain spatial and temporal information, thereby better understanding long-term emotional changes in the video. For 2-dimensional CNN (2D CNN) models, The residual block of the residual convolutional network (ResNet) enables it to simultaneously extract features from low-level to high-level, forming a hierarchical representation, [23] illustrates that this model can capture subtle emotional changes in facial images. In addition, [24] proposed a spatio-temporal convolutional neural framework to extract features from face images. Combining a deep spatial network and a deep temporal network makes it possible to simultaneously capture spatial and temporal features in images, thereby generating comprehensive feature representations.

Feature fusion plays a crucial role in MER systems. An effective feature fusion method can utilise the complementarity between different modalities to integrate the most relevant features and ignore redundant or noisy information, thereby improving the model's recognition accuracy and generalisation ability. In recent years, researchers have tended to use attention mechanism-based feature fusion methods to integrate features from different modalities. For example, [18] uses cross-modal attention to fuse speech and text features. Cross-modal attention allows the model to adaptively learn interactions and dependencies between modalities. By dynamically assigning weights to features based on their contributions, cross-modal attention effectively integrates information from multiple modalities. [15] compared the impact of using cross-modal attention and self-attention as feature fusion methods on the accuracy of the MER system that combines video, text, and speech.

Table 2 shows the accuracy of state-of-the-art MER systems, which contain bi-modal and tri-modal ER systems.

Table 2. *Accuracy for MER system in IEMOCAP and RAVDESS datasets. S, V, T represents the speech, video and text modality respectively.*

Model	Modalities	Dataset	emotions	UA
33 speech features + ELMo v2 [16]	S, T	IEMOCAP	4	0.745
MSRFG [18]	S, T	IEMOCAP	6	0.716
(2D CNN + RNN) + 3D CNN [20]	S, V	IEMOCAP	3	0.717
(2D CNN + GRU) + 3D CNN [21]	S, V	IEMOCAP	4	0.764
MCWSA-CMHA [23]	S, V, T	IEMOCAP	4	0.863
1D CNN + ResNet + GloVe [15]	S, V, T	IEMOCAP	7	0.642
RDesBert [17]	S, V, T	IEMOCAP	7	0.792
MRPN [25]	S, V	RAVDESS	8	0.914
DSN + DTN + 1DCNN [24]	S, V	RAVDESS	8	0.949
RDes [17]	S, V	RAVDESS	8	0.980

3. Methodology

This section primarily elaborates on the methodology for SER and speech-based MER systems. The model selection is based on one of the state-of-the-art MER systems, [17] proposes an approach to improving the accuracy and robustness of the MER system. By integrating advanced feature extraction techniques, including BERT, ResNet and DenseNet, with a novel feature fusion method, highlighting the importance of feature recalibration through squeeze-and-excitation (SE) blocks, which improves model generalisation ability and performance. This MER system addresses critical challenges in understanding and processing complex human emotions. The architecture of the ER systems is shown in Figure 1.

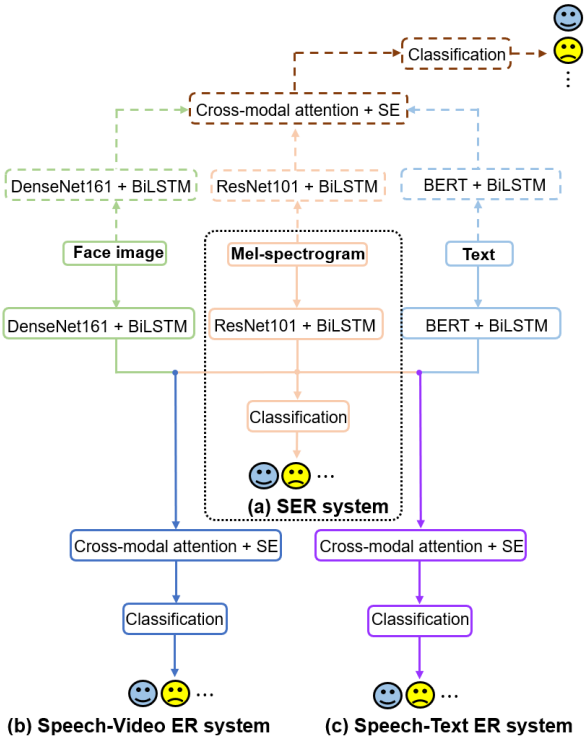

Figure 1: *Architecture of speech, speech-video, speech-text and speech-video-text ER systems.*

In the SER system, the original speech signals are converted into mel-spectrogram images. These images serve as the input for the combination of ResNet101 and BiLSTM. The aim of ResNet is to extract features from the mel-spectrogram, and the bi-directionality of BiLSTM allows the network to obtain past and future dependencies in the input sequence. ResNet101 consists of 5 convolutional regions with repeated convolutional calculations, totalling 101 convolutional layers. The architecture of SER is depicted in Figure 1 (a).

The input for the video modality in the speech-video ER system is facial images cropped from the video. The details are shown in Figure 1 (b). This bi-modal ER system combines DenseNet and BiLSTM to extract features from facial images. DenseNet, through its dense connection structure, achieves efficient feature transmission and reuse, providing robust feature extraction and generalisation capabilities, and the BiLSTM can further capture temporal information in the video. DenseNet161 has 161 convolutional layers.

The speech-text ER system employs BERT as the feature extractor for the text modality, which is shown in Figure 1 (c). The BERT architecture is composed of three main components: input embedding, transformer encoder, and output layer. Among these items, the most crucial part is the series of transformer encoder layers, designed to capture the contextual information between words within a sentence, thereby improving the system's ability to understand text content.

In terms of feature fusion methods, cross-modal attention and SE block are combined to fuse features from different modalities. The SE block can selectively enhance relevant information within each modality's features by assigning different weights to channels, thereby indirectly reducing the impact of redundant information on the model. Additionally, by weighing different features, the SE block enhances feature distinctiveness, making the system sensitive to subtle emotional changes and improving its generalisation capability.

The classification module integrates a statistical pooling layer, fully connected layers, and softmax activation. The statistical pooling layer can effectively capture the temporal dynamics of the input features by calculating statistical measures such as mean and standard deviation. Additionally, the statistical pooling layer integrates the fused features through these statistical measures, creating a comprehensive representation. This representation reflects information from all modalities, enabling the model to fully understand and utilise the input data. The combination of fully connected layers and the softmax activation function is a commonly used method in classification tasks. The fully connected layers extract and integrate features, while

Table 3. *Comparison between SER and MER systems in IEMOCAP and RAVDESS dataset. S, V, T represents the speech, video and text modality respectively.*

Model	Modalities	Dataset	emotions	WA	UA
ResNet101	S	IEMOCAP	7	0.468	0.479
BERT	T	IEMOCAP	7	0.535	0.551
DenseNet161	V	IEMOCAP	7	0.687	0.701
RBert	S, T	IEMOCAP	7	0.552	0.607
DesBert	V, T	IEMOCAP	7	0.734	0.769
RDes	S, V	IEMOCAP	7	0.742	0.788
RDesBert	S, V, T	IEMOCAP	7	0.756	0.792
ResNet101	S	RAVDESS	8	0.742	0.752
DenseNet161	V	RAVDESS	8	0.938	0.944
RDes	S, V	RAVDESS	8	0.963	0.980

the softmax activation function converts these features into class probabilities, thereby achieving classification prediction.

4. Results

We compared the weighted accuracy (WA) and unweighted accuracy (UA) of speech, text, video, speech-text, speech-video, video-text, and speech-video-text ER systems on the RAVDESS and 7-class unbalanced IEMOCAP datasets. The results are shown in the Table 3.

From the table, it could be observed that when using the 7-class IEMOCAP dataset as input, the UA of the speech-based and text-based ER system is only 47.3% and 55.1%, respectively. After incorporating the text modality into the SER system, the emotion recognition accuracy reaches 55.2% (WA) and 60.7% (UA). Despite this improvement, it has not surpassed the recognition accuracy of the video-based ER system. In contrast, the UA of the speech-video and text-video ER systems increased by 18.1% and 16.2% compared to the speech-text ER system.

For the speech-video-text ER system, the UA and WA increased by only 0.14% and 0.04% compared to the speech-video ER system. By analysing the experimental results shown in Table 3, it can be concluded that even though the accuracy of the text-based ER system is higher than that of the SER system, the contribution of the speech modality to the accuracy of the tri-modal ER system is more significant than that of the text modality. Therefore, the final contribution ranking is video, audio, and text. This result indicates that the video modality plays a more crucial role in providing emotional cues and enhancing overall recognition performance, while the addition of the text modality offers relatively minor improvements to the system's accuracy.

Since the RAVDESS dataset only provides speech and video modalities, it is impossible to determine the contribution of the text modality to recognition accuracy. However, the results presented in Table 3 indicate that the performance of the MER system, which combines video and speech, is superior to that of the SER system.

5. Discussion

Based on the results shown in Table 3, it can be observed that the recognition accuracy of the SER system for the 7-class IEMOCAP dataset is relatively low. Additionally, the accuracy gain achieved by incorporating the video modality into the SER system is much higher than that obtained by incorporating the text modality. The possible reason is that in an imbalanced dataset, the SER system may not thoroughly learn the features of the minority classes, leading the model to bias towards recognising the majority classes, thus affecting the overall recognition accuracy. Additionally, compared to video features, speech and text features are more dependent on temporal information, making them more susceptible to the effects of data imbalance.

Video modality provides rich and universal emotional expression information, which means that even if the amount of data for certain classes is relatively small, the system can still find cues to distinguish these classes from the limited features, making its performance relatively stable when facing imbalanced datasets. For example, compared to speech signals, certain emotions are more clearly and consistently expressed visually, such as smiles or frowns, which are consistent across different people and scenarios. Therefore, the model can learn these visual features easily, reducing its dependence on data balance.

The results of the MER system for the IEMOCAP dataset, as shown in Table 3, indicate that imbalanced datasets significantly impact the accuracy of emotion recognition. As the number of emotion categories increases, the imbalance becomes obvious, leading to a decrease in the recognition accuracy of the MER system. Notably, even with the 7-class imbalanced IEMOCAP dataset, the RDesBert system achieves an accuracy of over 79%. For the RAVDESS dataset, being a balanced dataset with frontal faces and no facial occlusions in the videos, MER systems typically achieve better results.

6. Conclusion

In conclusion, by leveraging the complementary strengths of different modalities, MER systems can achieve higher robustness and accuracy. Based on our research in the IEMOCAP dataset, incorporating video or text modalities into a SER system can improve the performance of emotion recognition. The video modality provides crucial emotional cues that significantly improve the system's ability to predict emotions precisely. While the text modality offers additional benefits, its contribution is comparatively less significant. Therefore, from the perspective of model complexity, it may be considered unnecessary to include the text modality if sufficient video and speech modality data are available. For future research, addressing challenges like data imbalance and improving feature fusion techniques are essential for the MER field.

7. References

[1] Zhao, S., Jia, G., Yang, J., Ding, G. and Keutzer, K., "Emotion Recognition from Multiple Modalities: Fundamentals and Methodologies", IEEE Signal Processing Magazine, 38(6): 59–73, 2021.

[2] Sarma, M., Ghahremani, P., Povey, D., Goel, N., Sarma, K. and Dehak, N., "Emotion Identification from Raw Speech Signals Using DNNs", in Interspeech, 3097–3101, 2018.

[3] Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J., Lee, S. and Narayanan, S., "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database", Language resources and evaluation, 42: 335-359, 2008.

[4] Livingstone, S. and Russo, F., "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English", Plos one, 13(5), 2018.

[5] Fahad, M., Ranjan, A., Yadav, J. and Deepak, A., "A Survey of Speech Emotion Recognition in Natural Environment", Digital Signal Processing, 110, 2021.

[6] Trinh, V., Dao, T., Le, X. and Castelli, E., "Emotional Speech Recognition Using Deep Neural Networks", Sensors, 22(4): 1414, 2022.

[7] Pandey, S., Shekhawat, H. and Prasanna, S., "Emotion Recognition from Raw Speech Using WaveNet", in Region 10 Conference (TENCON), IEEE, 1292–1297, 2019.

[8] Zhao, J., Mao, X. and Chen, L., "Speech Emotion Recognition Using Deep 1D & 2D CNN LSTM Networks", Biomedical Signal Processing and Control, 47: 312–323, 2019.

[9] Siadat, S., Voronkov, I. and Kharlamov, A., "Emotion Recognition from Persian Speech with 1D Convolution Neural Network", in Fourth International Conference Neurotechnologies and Neurointerfaces (CNN), IEEE, 152–157, 2022.

[10] Alonazi, B., Nauman, M., Jahangir, R., Malik, M., Alkhammash, E. and Elshewey, A., "Transformer-based Multilingual Speech Emotion Recognition Using Data Augmentation and Feature Fusion", Applied Sciences, 12 (18): 9188, 2022.

[11] Chen, L. and Rudnicky, A., "Exploring Wav2Vec 2.0 Fine Tuning for Improved Speech Emotion Recognition", in International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023.

[12] Kakouros, S., Stafylakis, T., Mošner, L. and Burget, L., "Speech-based Emotion Recognition with Self-supervised Models Using Attentive Channel-wise Correlations and Label Smoothing", in International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023.

[13] Lu, C., Lian, H., Zheng, W., Zong, Y., Zhao, Y. and Li, S., "Learning Local to Global Feature Aggregation for Speech Emotion Recognition", arXiv preprint arXiv:2306.01491, 2023.

[14] Peng, S., Cao, L., Zhou, Y., Ouyang, Z., Yang, A., Li, X., Jia, W. and Yu, S., "A Survey on Deep Learning for Textual Emotion Analysis in Social Networks", Digital Communications and Networks, 8(5): 745–762, 2022.

[15] Rajan, V., Brutti, A. and Cavallaro, A., "Is Cross-Attention Preferable to Self-Attention for Multi-Modal Emotion Recognition?" in International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 4693–4697, 2022.

[16] Singh, P., Srivastava, R., Rana, K. and Kumar, V., "A Multimodal Hierarchical Approach to Speech Emotion Recognition from Audio and Text", Knowledge-Based Systems, 229, 2021.

[17] Liu, J., James, J. and Nathwani, K., "Improved Multi-modal Emotion Recognition Using Squeeze-and-excitation Block in Cross-modal Attention", in Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2023.

[18] Wei, J., Hu, G., Tuan, L., Yang, X. and Zhu, W., "Multi-scale Receptive Field Graph Model for Emotion Recognition in Conversations", in International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023.

[19] Grahlow, M., Rupp, C. and Derntl, B., "The Impact of Face Masks on Emotion Recognition Performance and Perception of Threat", PLoS One, 17(2), 2022.

[20] Singh, M. and Fang, Y., "Emotion Recognition in Audio and Video Using Deep Neural Networks", arXiv preprint arXiv:2006.08129, 2020.

[21] Jia, N., and Zheng, C. and Sun, Wei., "A Multimodal Emotion Recognition Model Integrating Speech, Video and MoCAP", Multimedia Tools and Applications, 81(22): 32265–32286, 2022.

[22] Ren, M., Nie, W., Liu, A. and Su, Y., "Multi-modal Correlated Network for Emotion Recognition in Speech", Visual Informatics, 3(3): 150-155, 2019.

[23] Zheng, J., Zhang, S., Wang, Z., Wang, X. and Zeng, Z., "Multi-channel Weight-sharing Autoencoder based on Cascade Multi-head Attention for Multimodal Emotion Recognition", IEEE Transactions on Multimedia, 2022.

[24] Sharafi, M., Yazdchi, M., Rasti, R. and Nasimi, F., "A Novel Spatio-temporal Convolutional Neural Framework for Multimodal Emotion Recognition", Biomedical Signal Processing and Control, 78, 2022.

[25] Chang, X. and Skarbek, Władysław., "Multi-modal Residual Perceptron Network for Audio-video Emotion Recognition", Sensors, 21(16): 5452, 2021.