Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU33578

Speech Emotion Recognition: Method Evolution, Current Challenges and Future Trends

Zeyang Wang

Aulin College, Northeast Forestry University, Harbin, China
zeyangwang@nefu.edu.cn

Abstract. Speech emotion recognition (SER) is an important research direction in the field of affective computing and human-computer interaction. It is mainly used to automatically recognize human emotion from human speech. With the rapid development of artificial intelligence large language models and deep learning technology, SER has attracted more attention due to its wide application in the fields of intelligent AI assistants, customer service systems, virtual telephones, and medical monitoring. This paper reviews the latest progress of SER. Firstly, the basic framework of SER is introduced, including speech preprocessing, feature extraction, representation learning, and emotion classification. Then, according to the technology evolution, this paper summarizes the existing methods, including traditional machine learning methods, deep learning models, Transformer-based architecture, self-supervised learning, and multimodal emotion recognition. In addition, the data sets commonly used in SER research are reviewed. Finally, the current challenges and future research directions are discussed. Future research will focus on large-scale SSL models, cross-language and cross-cultural emotion recognition, and the combination of SER and large language models.

Keywords: Speech emotion recognition, Self-supervised learning, Multimodal fusion, Deep learning

1. Introduction

Speech emotion recognition (SER) is a key research direction in the field of affective computing and human-computer interaction, which aims to automatically recognize and interpret the emotion in human language. Accurate recognition of emotions in speech is very important to enhance the ability of human-computer interaction so that the machine can make a more humanized response. SER is widely used in the fields of intelligent AI assistants, customer service systems, virtual telephones, and medical monitoring.

SER can not only promote the development of human-computer interaction ability but also help to understand and identify human emotional states, which is important for improving user experience and interaction quality. The progress of SER technology promotes the development of affective computing, helps intelligent systems better understand and respond to users' emotional needs, and improves user experience. SER has experienced a significant paradigm shift from traditional feature engineering to deep representation learning in recent years. Traditional methods

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU33578

rely on manually designed prosody, spectrum and sound quality features, but this process is time-consuming and hinders the universality of the model. The progress of deep learning enables the system to automatically learn hierarchical representation in a data-driven manner, which effectively improves the recognition performance [1]. Transformer architecture shows advantages in modeling the local and global information of speech emotion. The CNN-Transformer network proposed by Tang et al. improves the efficiency and accuracy of SER [2].

Although some progress has been made, SER still faces many challenges, including limited data sets, cross-corpus generalization, subjectivity of emotional state, and complexity of emotional expression. This paper reviews the research status of SER and discusses the technology evolution, common data sets, and current challenges. This paper also discusses the current emerging trends, in which the integration of SER with large language models and multimodal systems is particularly critical.

2. Speech emotion recognition framework

SER aims to infer the emotional state of the speaker from the original speech signal. Its recognition process is shown in Figure 1, including speech signal preprocessing, feature extraction, model training and classification, and emotion classification. Emotion modeling methods are mainly divided into discrete category model and dimensional model. The discrete model is based on the basic emotion categories (such as sadness, happiness, fear, anger, disgust and surprise), while the dimensional model interprets emotion as a continuous value, mainly including valence, arousal, and dominance. The five key steps of speech signal processing process are signal framing, windowing, voice activity detection, normalization, and noise reduction [3].

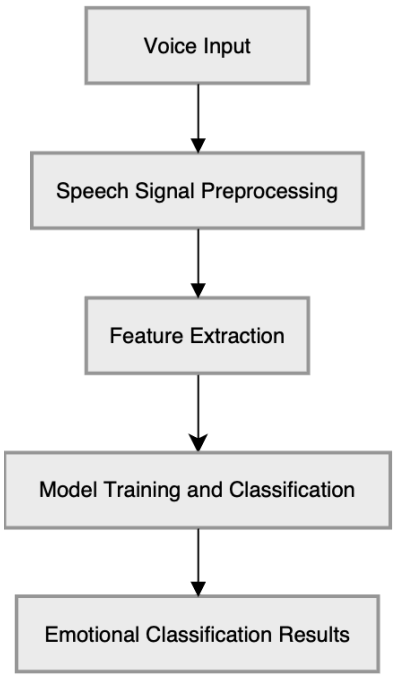

Figure 1. Speech emotion recognition process (picture credit: original)

DOI: 10.54254/2755-2721/2026.GU33578

3. Technology evolution and method classification

The development of SER method is shown in Figure 2. It is an evolution from handcrafted features to multimodal collaborative modeling. This process is not a linear replacement, but is driven by the improvement of computing power, data set expansion and self-supervised learning (SSL) breakthrough. Early SER systems highly relied on manually designed acoustic features. After extracting features through the signal preprocessing stage, they combined linear classifiers (such as Bayesian network, SVM) or nonlinear classifiers (such as GMM, HMM) to complete emotion classification [4]. With the rise of deep learning technology, SER method has entered the stage of deep learning stage, realizing the extension from shallow representation (such as PCA, LDA) to deep architecture. In supervised learning, CNN, as a data-driven filter bank, can effectively capture the spectral envelope. RNN (especially LSTM/GRU) can capture the dynamic changes of emotion in time series data. CNN-LSTM hybrid model has been proved to be better than a single architecture in extracting robust emotion features. At the same time, in order to meet the challenge of the scarcity of labeled data, the research is extended to unsupervised and semi-supervised learning, using autoencoder (AE), variational autoencoder (VAE) and generative adversarial networks (Gans) to learn the internal structure from unlabeled data [1]. The introduction of multimodal emotion recognition (MER) has further promoted the development of SER and improved the recognition robustness by integrating text, speech and facial cues [5].

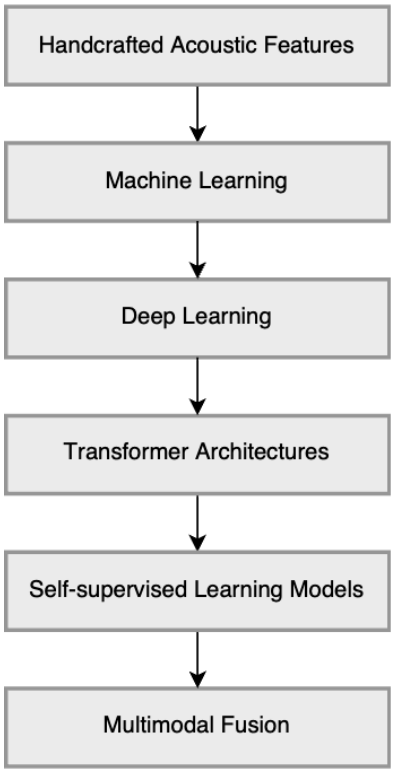

Figure 2. Evolution of speech emotion recognition methods (picture credit: original)

3.1. Handcrafted feature engineering and traditional machine learning methods

In traditional SER research, the construction of acoustic feature system mainly depends on handcrafted feature engineering. Acoustic features widely used in existing studies can be

DOI: 10.54254/2755-2721/2026.GU33578

summarized as prosodic features (such as fundamental frequency, energy, duration), spectral features (such as Mel frequency cepstral coefficient, linear predictive cepstral coefficient), voice quality features (such as jitter, shimmer, harmonic noise ratio) and Teager energy operator (Teo) related features [3]. Manual feature design for complex SER tasks requires a lot of human efforts to convert speech signals into meaningful attributes, which is labor-intensive and time-consuming, hindering the generalization ability of the model and slowing down the pace of innovation [1]. With the development of machine learning, in the field of SER, classical classifiers such as support vector machine (SVM), Gaussian mixture model (GMM) and hidden Markov model (HMM) are often combined with manually designed acoustic features (such as prosodic and spectral features) to process high-dimensional handcrafted features. The results show that the handcrafted features can effectively capture the emotion related acoustic characteristics and significantly improve the performance of the model; However, these methods are sensitive to noise. On the IEMOCAP dataset, there is a significant confusion between happiness and neutral emotions, which is attributed to the non-representative features and noise learned by the model during the training process; On the EMOVO dataset, the misclassification between "happiness" and "surprise" emotions reveals the overlapping problem of acoustic feature space and the insufficient capture of discriminant features [6,7].

3.2. Deep learning methods

With the rapid development of artificial intelligence, deep learning has become the main paradigm of SER, which is much better than the traditional machine learning method. Unlike the traditional methods that rely on handcraft acoustic features, the deep learning model can automatically learn the hierarchical feature representation from the original or transformed speech signal directly. Convolutional neural network (CNN) is good at extracting local spectral features, such as capturing emotion related acoustic patterns from Mel-spectrograms, but it has weak modeling ability for the global temporal dependencies of speech, so it needs to stack additional temporal layers (such as LSTM) to take account of local and global. Recurrent neural network (RNN), including long short-term memory (LSTM) and gated recurrent unit (GRU), aims to simulate the temporal dependence in speech sequences, but it has the problems of gradient disappearance and explosion, which is difficult to capture the long-term dependence, and the serial computing efficiency is very low.

3.3. Transformer-based methods

In recent years Transformer-based architecture has become one of the main methods of SER, which is mainly manifested in the powerful ability of modeling remote dependencies and capturing global context information in speech signals. Different from the traditional recurrent neural network, Transformer uses self-attention mechanism to model the relationship between all-time steps at the same time. In the hybrid CNN-Transformer architecture, the convolutional neural network (CNN) module is specially used to efficiently extract local features of speech signals from the time-frequency perspective, while the transformer component is responsible for modeling long-distance sequence dependencies to form a complementary feature extraction mechanism. Researchers have proposed a Time-Shuffle Attention (T-Sa) mechanism and a Lightweight Convolution Transformer (LCT) module. T-Sa models the temporal dynamics of features through a bidirectional long short-term memory network (BiLSTM) to capture the changing patterns of tone and energy over time and uses Shuffle units to efficiently integrate the attention information of spatial and channel dimensions. LCT combines the large-kernel depthwise separable convolution and coordinate

DOI: 10.54254/2755-2721/2026.GU33578

attention to aggregate the dependent information along the temporal and frequency directions to explicitly enhance the representation ability of the speech energy duration cues. The experimental results show that the removal of LCT module leads to the weighted average accuracy (WA) and unweighted accuracy (UA) decreased by 4.52% and 3.79% respectively on the IEMOCAP and EmoDB data sets [2], which verifies the key role of global information modeling; The introduction of T-Sa module has increased WA and UA by 1.2% and 1.54% respectively [2], which fully proves the significant contribution of explicit modeling of speech dynamic changes to emotion recognition performance.

3.4. Self-supervised learning methods

Self-supervised learning (SSL) has recently become a powerful method in SER, which enables the model to learn meaningful information from large-scale unlabeled speech data. Unlike supervised learning that relies heavily on annotated emotional labels, SSL uses the inherent structure in the data to construct tasks, thus reducing the dependence on manual annotation. It first uses many unlabeled speech data, makes the model learn the essence of speech through pre tasks such as masked modeling, and then fine tune a small amount of emotional labeling data to obtain a high-precision and high generalization emotional recognition model. This not only solves the major limitation of traditional deep learning that relies on many labeled data but also makes it possible for SER to move from the laboratory to the real applications. In SER task, the transfer fine-tuning strategy of self-supervised learning model is very important for performance optimization. The effectiveness of fine-tuning strategy highly depends on the adjustment mechanism of Transformer layer and the domain characteristics of pretraining data. Experiments show that the fine-tuning process enables the model to implicitly capture the language information in the audio signal, which is the key mechanism to significantly improve its valence prediction performance [8].

3.5. Multimodal methods

With the rapid development of large-scale artificial intelligence system, multimodal large model has become a promising direction to promote SER. Unlike the traditional SER system which only relies on acoustic features, the multimodal large-scale model integrates information from multiple modes, such as speech, text and visual signals, to achieve a more comprehensive understanding of human emotions. In multimodal emotion recognition, the evolution of fusion mechanism has gradually developed from early simple feature concatenation to more refined multi-level fusion approaches. Although early fusion strategies such as feature concatenation (cLSTM-EF) can integrate multi-source information, they are easily affected by redundancy across modalities, resulting in low computational efficiency; Late fusion (cLSTM-LF) is difficult to capture deep interaction due to modal independent processing. In response, hybrid fusion paradigms (such as MMAN) significantly improve the quality of feature interaction by combining the advantages of early and late fusion. Specifically, the cLSTM-MMA mechanism designs three parallel directional multimodal attention modules, which use speech, vision and text as queries to dynamically calculate the cross-modal attention and self-attention scores, to achieve the selective enhancement of features in the time dimension, channel dimension and spatial dimension. This mechanism avoids the parameter expansion problem of simple concatenation. On the IEMOCAP data set, the accuracy of cLSTM-MMA is improved by 2% compared with cLSTM-EF, and the number of parameters is reduced by 40%, which verifies the efficiency of attention mechanism in modal interaction modeling [9].

DOI: 10.54254/2755-2721/2026.GU33578

4. Commonly used datasets

In the research of SER, common data sets such as IEMOCAP and EmoDB are widely used, but there are significant limitations. The IEMOCAP dataset contains 12 hours of audio-visual data recorded by 10 professional actors, covering four emotions: anger, sadness, happiness and neutrality. The distribution of samples in each category is uneven (happiness accounts for 29.58%, anger accounts for 19.94%); The EmoDB data set contains only 535 German speech samples, and 10 actors recorded 7 emotions, which also has the problem of category imbalance (anger accounted for 23.74%, disgust only accounted for 8.60%) [10]. The main challenges faced by current research include low recognition rate, high computational cost, and low success rate when dealing with different languages and databases. Therefore, it is suggested that future work should test the performance of the system on more diversified databases (especially natural databases) to alleviate the problem of generalization.

5. Current research challenges

First, SER research faces a significant small data problem. Due to the small scale of most publicly available annotation data sets, which usually only contain dozens of speakers, the deep learning model is easy to over fit [11]. To alleviate data sparsity, the researchers proposed vocal tract length perturbation (VTLP) as a data enhancement technology. Experiments show that increasing the number of copies of enhanced data can effectively improve the accuracy of the model [12]. Secondly, the poor generalization ability is another key challenge. The traditional method based on handcrafted acoustic features requires a lot of manual efforts, which hinders the generalization ability of the model and slows down the pace of innovation [1]. In real scene applications, the performance of the model is often poor. About 2.58% of the annotations use natural language descriptions (such as "Slightly Angry, calm,"), but the existing SER classification model cannot handle such descriptions, resulting in the discarding of annotations containing valuable information such as emotional intensity [13]. Thirdly, the subjectivity of emotion also brings difficulties. Most studies still rely on discrete categorical emotion model, which lacks nuances and cannot capture the dynamics and mobility of human emotion over time [14]. Finally, emotional expression is inherently multimodal, as speech signals coexist with textual and linguistic components. Discarding language components may not be conducive to emotion recognition [15].

6. Future research directions

Firstly, emotional representation enhancement based on large-scale self-supervised model will continue to be an important trend. With the extensive application of wav2vec 2.0, HuBERT and other pretrained models in speech tasks, how to further exploit their potential emotion related information and improve their emotional expression ability combined with acoustic prior knowledge will be the key to subsequent research.

Secondly, cross-lingual and cross-cultural emotion recognition is the direction that SER must break through in its real application. Different languages, phonetic habits and cultural backgrounds will significantly affect emotional expression. For example, the performance of the multilingual pretrained model on the test set containing only English will decline, which reveals the deficiency of the current method in cross-lingual generalization ability [8]. Therefore, future research needs to pay more attention to the modeling of emotional features unrelated to language types, and combine domain adaptation, adversarial learning and other methods to enhance the stability and robustness of

DOI: 10.54254/2755-2721/2026.GU33578

the model in a multilingual environment, to promote the SER system from monolingual research to a more extensive global application scenario.

Thirdly, multimodal fusion will still be an important way to improve the performance of emotion recognition. Future research not only needs to further optimize the collaborative modeling of speech and text, but also can incorporate more modes such as expression, posture, physiological signals into a unified framework to build a more complete emotional understanding system.

In addition, the combination of SER and large language model (LLM) shows broad prospects. The large language model has strong semantic understanding, reasoning and knowledge transfer capabilities, and can be used to assist in emotional label generation, weakly supervised learning, data augmentation and emotional interpretation in human-computer interaction in the future.

In general, SER research in the future will develop in the direction of stronger representation ability, higher cross domain generalization ability, deeper multimodal fusion and more intelligent human-computer cooperation. These studies will not only help to improve the performance of the model, but also further promote the practical application of SER in the fields of intelligent healthcare, educational companionship, affective computing and human-computer interaction.

7. Conclusion

This paper first introduces the basic framework of SER, then focuses on its development process and technology evolution, and finally expounds the current research challenges and future research trends. SER has made significant progress in the past few decades, from the traditional machine learning method based on handcrafted features to the advanced deep learning and self-supervised representation learning framework. Early methods, such as support vector machine (SVM) and hidden Markov model (HMM), rely heavily on feature engineering, while recent methods use end-to-end architectures, including convolutional neural network (CNN), recurrent neural network (RNN) and Transformer-based models, to automatically learn differentiated emotional representations from speech signals. In particular, the emergence of self-supervised learning method has greatly enhanced the ability of SER system to use large-scale unlabeled data, thus improving the robustness and generalization ability. Despite the progress made, SER still faces challenges such as limited data sets, subjective emotion annotation, and cross corpus generalization. Future research should focus on integrating large-scale pretraining models, improving cross-domain and cross-lingual generalization capabilities, and developing lightweight models for real-time applications. In general, SER has great potential in enhancing human-computer interaction and understanding human emotions more effectively.

References

[1] Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., & Schuller, B. (2021). Survey of deep representation learning for speech emotion recognition. IEEE Transactions on Affective Computing, 14(2), 1634-1654.

[2] Tang, X., Huang, J., Lin, Y., Dang, T., & Cheng, J. (2025). Speech emotion recognition via cnn-transformer and multidimensional attention mechanism. Speech Communication, 171, 103242.

[3] Hashem, A., Arif, M., & Alghamdi, M. (2023). Speech emotion recognition approaches: A systematic review. Speech Communication, 154, 102974.

[4] Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. IEEE access, 7, 117327-117345.

[5] Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., & Zong, Y. (2023). A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. Entropy, 25(10), 1440.

[6] Song, H., Zhang, L., Gao, M., Zhang, H., Hain, T., & Shan, L. (2025). MS-EmoBoost: a novel strategy for enhancing self-supervised speech emotion representations. Scientific Reports, 15(1), 21607.

DOI: 10.54254/2755-2721/2026.GU33578

[7] Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., & Ambikairajah, E. (2021). A comprehensive review of speech emotion recognition systems. IEEE access, 9, 47795-47814.
[8] Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., & Schuller, B. W. (2023). Dawn of the transformer era in speech emotion recognition: closing the valence gap. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(9), 10745-10759.
[9] Pan, Z., Luo, Z., Yang, J., & Li, H. (2020). Multi-modal attention for speech emotion recognition. arXiv preprint arXiv: 2009.04107.
[10] Anvarjon, T., Mustaqeem, & Kwon, S. (2020). Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features. Sensors, 20(18), 5212.
[11] Pepino, L., Riera, P., & Ferrer, L. (2021). Emotion recognition from speech using wav2vec 2.0 embeddings. arXiv preprint arXiv: 2104.03502.
[12] Xu, M., Zhang, F., Cui, X., & Zhang, W. (2021). Speech emotion recognition with multiscale area attention and data augmentation. ICASSP 2021-2021 IEEE International conference on acoustics, speech and signal processing (ICASSP) (pp. 6319-6323). Ieee.
[13] Wu, H., Chou, H. C., Chang, K. W., Goncalves, L., Du, J., Jang, J. S. R., ... & Lee, H. Y. (2024). Emo-superb: An in-depth look at speech emotion recognition. arXiv preprint arXiv: 2402.13018.
[14] Hazmoune, S., & Bougamouza, F. (2024). Using transformers for multimodal emotion recognition: Taxonomies and state of the art review. Engineering Applications of Artificial Intelligence, 133, 108339.
[15] Siriwardhana, S., Reis, A., Weerasekera, R., & Nanayakkara, S. (2020). Jointly fine-tuning" bert-like" self supervised models to improve multimodal speech emotion recognition. arXiv preprint arXiv: 2008.06682.