

Speech Emotion Recognition Technology in Human-Computer Interaction

Jingming Wang ^a

Stony Brook Institute at Anhui University, Anhui University, Hefei, China

Keywords: Speech Emotion Recognition, Development History, Speech Feature Extraction.

Abstract: Speech Emotion Recognition (SER), as an important research direction in the field of human-computer interaction, enables computers to perceive and understand the user's emotional state, thereby improving the naturalness and intelligence of the interaction. This paper systematically reviews the development context and key technologies of speech emotion recognition. First, the development history of this field is reviewed and the main stages of its algorithm evolution are sorted out. Then, based on the overall process of speech emotion recognition, this paper focuses on the core link, the feature extraction stage, and deeply explores its key role in recognition performance, and systematically compares the differences between traditional methods and machine learning methods. In addition, this paper also deeply analyzes the core challenges faced by current research from the perspectives of features and models. Through a comprehensive review of existing research results, this paper aims to provide theoretical references and technical support for building a more efficient and robust speech emotion recognition system.


1 INTRODUCTION

With the rapid development of artificial intelligence technologies, human-computer interaction has gradually permeated various aspects of daily life. Voice interaction, in particular, has been widely applied in fields such as Siri, Xiao Ai, and smart home systems. By introducing speech emotion recognition technology into human-computer interaction systems, the focus of these systems can extend beyond merely understanding semantic information to also analyzing voiceprint signals and perceiving users' emotions. This makes interactions more humanized while improving both system intelligence and user experience.

Speech emotion recognition technology has a wide range of applications. In the field of intelligent customer service, it can replace manual quality inspection methods, providing a more efficient and cost-effective way to detect customer service staff's emotions and reduce conflicts with users (Zhang, 2023). In the power grid industry, SER can effectively monitor the emotional states of dispatchers, thereby significantly reducing human errors and preventing safety incidents (Luo, 2023).

However, in real life, speech emotions are characterized by diversity, hybridity, and uncertainty (Luo, Ran, Yang, & Dou, 2022), making emotion recognition quite challenging. Fortunately, with the continuous development of machine learning, its powerful data processing and feature learning capabilities have brought new opportunities for advancing speech emotion recognition (Lieskovská, Jakubec, Jarina, & Chmúlik, 2021).

This paper first reviews and summarizes the development history of SER, outlining the main stages of algorithm evolution. It then explains the workflow of emotion recognition, evaluates traditional methods and machine learning approaches from a feature perspective, and identifies the core challenges of each. Finally, feasible directions for future research are proposed.

^a <https://orcid.org/0009-0004-0654-2078>

2 THE OVERALL DEVELOPMENT OF SPEECH EMOTION RECOGNITION

Speech emotion recognition technology has developed rapidly over the past 40 years. Figure 1 illustrates the research progress in this field. In 1996, Daellert et al. conducted pioneering research in SER (Schuller, 2018). Early studies mainly relied on handcrafted features and traditional classification models. It was not until 2000, when Nicholson applied neural networks to this field, that machine learning models began to enter speech recognition research (Milton & Tamil Selvi, 2014). In 2005, Grimm et al. introduced a three-dimensional emotion description model to spontaneous SER (Elbarougy & Akagi, 2012). In 2006, Neiberg et al. applied Gaussian Mixture Models (GMM) to spontaneous SER (Neiberg, Elenius, & Laskowski, 2006).

In 2010, Eyben et al. developed OPEN-SMILE, a toolkit for extracting speech emotion features (Eyben, Wöllmer, & Schuller, 2010). By 2014, Mao et al.

introduced Convolutional Neural Networks (CNNs) to learn emotionally salient features for SER, marking the adoption of deep learning models in this field (Mao, Dong, Huang, & Zhan, 2014). Subsequently, deep learning models have continued to evolve. In 2016, Trigeorgis et al. proposed an end-to-end approach that combined CNNs with Long Short-Term Memory (LSTM) networks (Trigeorgis et al., 2016). In 2018, Schuller summarized the development, challenges, and future trends of CNNs and LSTMs in SER (Schuller, 2018).

Since 2021, an increasing number of studies have focused on incorporating Transformer models (Chen, Xing, Xu, Pang, & Du, 2023). In recent years, SER has been transitioning from traditional handcrafted features and machine learning classifiers toward deep learning and multimodal fusion approaches (Zhu, Sun, Wei, & Zhao, 2023). This includes integrating linguistic information with text, facial expressions, body movements, and other modalities, thereby making emotion recognition more accurate and efficient.

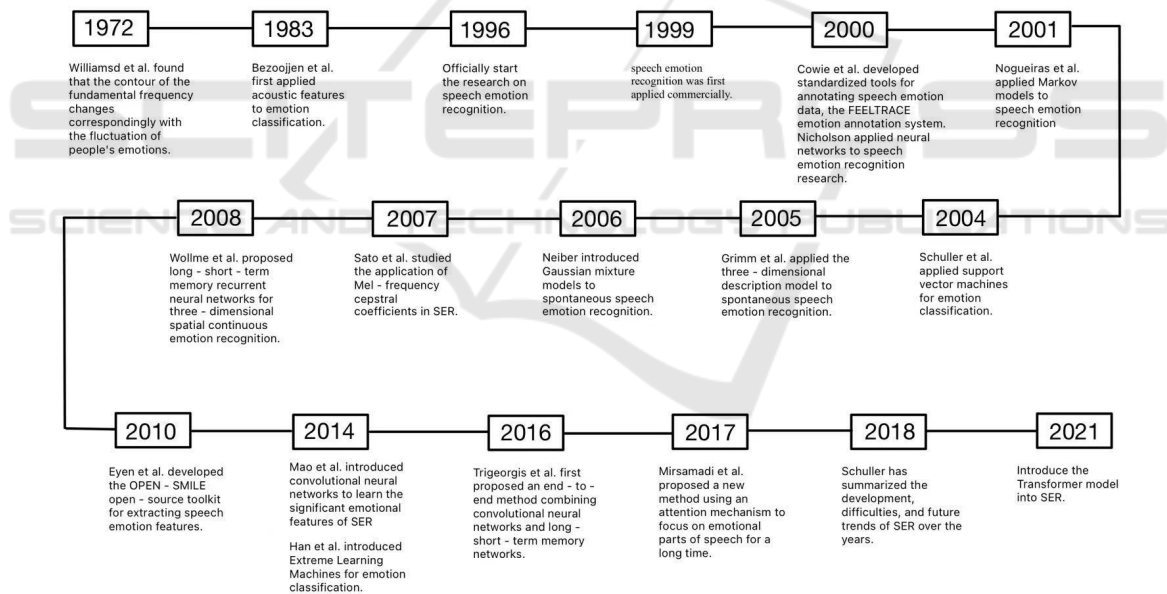


Figure 1: Schematic diagram of speech emotion recognition research development (Picture credit: Original).

3 EMOTION RECOGNITION METHODS

3.1 Overall Workflow

Speech Emotion Recognition (SER) technology involves using computers to analyze various

emotional information in preprocessed speech signals, extracting features that describe emotions, associating these features with specific emotional categories, and ultimately classifying the emotional states (Luo, Ran, Yang, & Dou, 2022). In the preprocessing stage, incomplete and unnecessary noise signals are removed. Subsequently, feature extraction is performed, typically extracting

traditional features such as prosodic features, wavelet features, spectral features, and cepstral features from the edges, segments, and utterances of speech signals, alongside automatically extracted features by deep learning models (Lieskovská, Jakubec, Jarina, & Chmulík, 2021; Ramyasree & Kumar, 2023). Several classic tools such as PRAAT, APARAT, OpenSMILE, and OpenEAR can be used for mining features from speech signals (Shukla & Jain, 2022). Finally, a classifier is employed to categorize the emotional features and build the SER model. Among these stages, feature extraction is a critical process that has a decisive impact on recognition performance.

3.2 Feature Extraction

3.2.1 Traditional Feature Extraction

Traditional handcrafted features include prosodic features, wavelet features, spectral features, and cepstral features. Prosodic features mainly cover aspects such as stress, pauses, and intonation, reflecting the rhythm and pitch variations in speech. Wavelet features are derived using wavelet transform techniques, which can effectively capture the local transformation characteristics of speech signals and offer certain advantages in analyzing non-stationary signals. Spectral features are obtained by transforming time-domain signals into the frequency domain through methods such as Fourier Transform, providing insights into the energy distribution of speech signals across different frequency bands. Cepstral features are further processed parameters based on spectral features, with the most representative being Mel-Frequency Cepstral Coefficients (MFCCs) and Perceptual Linear Prediction (PLP) coefficients. These features can reflect subtle adjustments in the spectrum during emotional changes and exhibit a degree of similarity to how humans perceive emotions.

3.2.2 Machine Learning-Based Feature Extraction

The automatic learning capability of machine learning enables it to autonomously extract emotional features from speech signals, with different deep learning models capturing distinct types of features.

Features extracted using Convolutional Neural Networks (CNNs) primarily include local spectral features and hierarchical features. Compared with traditional spectral features, the spectral features extracted by CNN are higher in dimensionality, more

abstract, and harder to interpret. However, this extraction approach avoids the subjectivity of manual feature engineering, saves time, and offers strong adaptability to different types of speech data. Furthermore, it can comprehensively and meticulously describe variations in the spectrum. The hierarchical features extracted by CNN are progressively abstracted and summarized as the depth of the network layers increases. Each subsequent layer further processes the features from the previous layer, resulting in a more comprehensive and accurate representation of the intrinsic structure of speech information, thereby improving the recognition accuracy of the model.

Features extracted using Recurrent Neural Networks (RNNs) and their variants possess temporal dynamics and contextual dependency. Temporal dynamic features are extracted by combining the current feature with information from previous moments through the recurrent structure of RNNs, capturing the rising, falling, or steady trends of these features. This also allows the model to detect periodicity in speech signals and, through learning from previous cycles, extract related features more accurately. Context-dependent features are derived from speech information such as contextual pauses and speech rate, rather than isolated features at a single time point. RNNs and their variants can leverage these features to account for the coherence of speech, leading to a more accurate understanding of overall emotional states.

3.2.3 Advantages of Machine Learning in SER

With the advancement of deep learning, end-to-end deep SER has gained increasing attention, capable of directly using raw emotional speech signals or handcrafted features as input for deep learning models (Luo, Ran, Yang, & Dou, 2022). The integration of machine learning technology with SER brings numerous advantages. Firstly, the powerful self-learning ability of machine learning allows it to automatically extract features from large amounts of speech data, offering stronger adaptability and more representative features compared to traditional speech recognition methods (Lieskovská, Jakubec, Jarina, & Chmulík, 2021). Secondly, when dealing with imbalanced datasets, machine learning algorithms – such as convolutional recurrent neural networks (CRNNs) with variable-length inputs and focal loss – can adjust the contribution of different samples to the total loss, enabling the model to perform well even on minority samples (Liang, Li, &

Song, 2020). Additionally, the incorporation of attention mechanisms allows models like CNNs to compute attention weights, determining the importance of different parts of the speech, thus making emotion recognition more accurate (Zhu, Sun, Wei, & Zhao, 2023).

4 KEY CHALLENGES

4.1 Features

Compared with the automatically extracted features from deep learning methods, traditional features possess stronger interpretability and lower computational complexity. Most traditional features have explicit physical meanings, making them easily recognizable within speech signals. These intuitive features facilitate researchers' understanding and analysis. In contrast, features extracted through deep learning are more complex and represent deeper emotional characteristics. These features are typically difficult to detect using conventional approaches, but they exhibit stronger discriminative power and adaptability. When dealing with more complex speech information, deep learning-based features demonstrate superior expressive capabilities and are better suited for accommodating emotional expression differences across various regions and cultures.

Traditional feature extraction algorithms are relatively simple and computationally efficient. Under limited hardware resources, these algorithms can quickly extract features, making them suitable for scenarios where high precision in emotion recognition is not required. Conversely, features extracted by deep learning models require no manual design, and offer greater discriminative power and adaptability. Through automatic learning from large amounts of speech data, deep learning models can autonomously extract complex and deep emotional features, significantly enhancing recognition performance in challenging conditions.

4.2 Models

There are notable differences between traditional models and deep learning models in the field of SER. Traditional models, such as Hidden Markov Models (HMM) and Support Vector Machines (SVM), are relatively simple. These approaches typically offer stronger interpretability, demand less hardware, and are efficient when handling small-scale data training and recognition tasks. However, due to their simple

structure, the recognition accuracy of these models tends to degrade when dealing with complex speech signals.

Deep learning models effectively address this issue. With their powerful feature learning and fitting capabilities, machine learning and deep learning models achieve higher accuracy and better robustness in SER tasks. Nevertheless, deep learning models also have certain drawbacks, such as higher hardware requirements and a greater dependence on large-scale training datasets.

5 CONCLUSIONS

This paper focused on SER technology, first providing a comprehensive review of the 40-year development history of SER. Subsequently, the workflow of SER technology was explained, including preprocessing, feature extraction, feature-to-emotion mapping, and emotion classification. From a feature extraction perspective, this study extensively discussed the principles, advantages, and limitations of both traditional methods and machine learning approaches.

For traditional methods, due to their reliance on handcrafted features, these models offer higher interpretability and simpler structures, making them advantageous in scenarios with limited hardware resources and modest recognition accuracy requirements. In contrast, machine learning methods possess automatic feature extraction capabilities, enabling them to mine complex and deep emotional features from large volumes of speech data. These features exhibit greater adaptability and discriminative power, achieving better performance in complex speech environments and across diverse cultural contexts.

Future research can further explore fusion strategies for different deep learning models, combining the strengths of CNNs and RNNs to construct more powerful hybrid models. Such approaches can achieve collaborative optimization of local feature capture and long-term dependency processing, thereby enhancing model performance in complex SER tasks. Additionally, integrating multimodal information—such as text, facial expressions, and body movements—can facilitate the construction of multimodal fusion SER models, enabling a more comprehensive understanding of emotional expression and improving both recognition accuracy and system robustness.

REFERENCES

- Chen, W., Xing, X., Xu, X., Pang, J., & Du, L. (2023). SpeechFormer++: A hierarchical efficient framework for paralinguistic speech processing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, 775–788.
- Elbarougy, R., & Akagi, M. (2012). Speech emotion recognition system based on a dimensional approach using a three-layered model. In *Proceedings of the 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference* (pp. 1–9). IEEE.
- Eyben, F., Wöllmer, M., & Schuller, B. (2010). openSMILE: The Munich versatile and fast open-source audio feature extractor. *ACM*.
- Liang, Z., Li, X., & Song, W. (2020). Research on speech emotion recognition algorithm for unbalanced data set. *Journal of Intelligent & Fuzzy Systems*, 39(3), 2791–2796.
- Lieskovská, E., Jakubec, M., Jarina, R., & Chmúlik, M. (2021). A review on speech emotion recognition using deep learning and attention mechanism. *Electronics*, 10(10), 1163.
- Luo, D. (2023). Research on power grid dispatching operation safety early warning model based on speech emotion recognition (Master's thesis, Shaanxi University of Technology). CNKI.
- Luo, D., Ran, Q., Yang, C., & Dou, W. (2022). A review of speech emotion recognition. *Computer Engineering and Applications*, 58(21), 40–52.
- Mao, Q., Dong, M., Huang, Z., & Zhan, Y. (2014). Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia*, 16(8), 2203–2213.
- Milton, A., & Tamil Selvi, S. (2014). Class-specific multiple classifiers scheme to recognize emotions from speech signals. *Computer Speech & Language*, 28(3), 727–742.
- Neiberg, D., Elenius, K., & Laskowski, K. (2006). Emotion recognition in spontaneous speech using GMMs. *Interspeech 2006*.
- Ramyasree, K., & Kumar, C. S. (2023). Multi-attribute feature extraction and selection for emotion recognition from speech through machine learning. *Traitement du Signal*, 40(1), 265–275.
- Schuller, B. (2018). Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5), 90–99.
- Shukla, S., & Jain, M. (2022). Deep GANITRUS algorithm for speech emotion recognition. *Journal of Intelligent & Fuzzy Systems*, 43(5), 5353–5368.
- Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Zhang, M. (2023). Design and implementation of a customer service emotion monitoring system based on speech emotion recognition (Master's thesis, Southeast University). CNKI.
- Zhu, R., Sun, C., Wei, X., & Zhao, L. (2023). Speech emotion recognition using channel attention mechanism. In *2023 4th International Conference on Computer Engineering and Application (ICCEA)* (pp. 680–684). IEEE.