International Journal for Research Trends in Social Science & Humanities
Volume 4, Issue 2
Mar-Apr 2026, PP 1-11

Multimodal Emotion Detection using voice and text for mental well-being.

S.Kanimozhi [1] , G.Pooja Sri [2] , N.Suganya[3] , R.Vishnu Priya[4]
[1]Assistant Professor Department of Computer Science and Engineering Kongunadu College of Engineering and Technology Tamilnadu,India
[2 3 4] Department of Computer Science and Engineering Kongunadu College of Engineering and Technology Tamilnadu,India

Abstract- Mental health monitoring has become increasingly significant due to the growing prevalence of stress, anxiety, and emotional disorders in modern society. Multimodal Emotion Detection using Voice and Text aims to enhance emotion recognition accuracy by analyzing multiple forms of human communication simultaneously. The proposed system integrates speech signals and textual data to detect emotional states such as happiness, sadness, anger, fear, and neutrality. Voice inputs are processed by extracting acoustic features including pitch, tone, speech rate, and intensity, while textual inputs are analyzed using Natural Language Processing (NLP) techniques to identify semantic meaning and sentiment patterns. Advanced machine learning and deep learning algorithms are employed to perform multimodal feature fusion and classify emotions more effectively than single-modal approaches. The framework includes stages such as data acquisition, preprocessing, feature extraction, multimodal fusion, and emotion classification. By accurately identifying emotional conditions, the system supports mental well-being monitoring and helps in the early detection of stress or negative emotional states. This technology can be applied in healthcare systems, intelligent virtual assistants, counseling platforms, and educational environments to provide timely emotional insights, personalized support, and improved human–computer interaction.

Keywords: Multimodal Emotion Detection, Voice Analysis, Text Analysis, Natural Language Processing, Machine Learning, Deep Learning, Sentiment Analysis, Mental Well-being Monitoring.

I. Introduction

Mental health has become a critical concern in modern society due to the increasing levels of stress, anxiety, depression, and emotional imbalance experienced by individuals. Rapid technological growth, social pressures, and lifestyle changes have significantly affected psychological well-being. Early identification of emotional distress plays an important role in providing timely support and preventing serious mental health issues. In recent years, advancements in Artificial Intelligence (AI), Machine Learning (ML), and Natural Language Processing (NLP) have enabled the development of intelligent systems capable of recognizing human emotions through computational methods. Emotion detection systems can assist in monitoring mental well-being and improving human–computer interaction.

Traditional emotion detection methods mainly rely on a single modality such as facial expressions, speech signals, or textual content. However, human emotions are complex and are often expressed through multiple communication channels simultaneously. Relying on a single source of information may lead to inaccurate or incomplete emotional interpretation. To overcome these limitations, multimodal emotion detection systems have been introduced, which combine multiple data sources such as voice, text,

facial expressions, and physiological signals. Among these, voice and text are widely available and provide valuable emotional cues in everyday communication.

Voice-based emotion recognition analyzes acoustic characteristics of speech, including pitch, tone, intensity, speech rate, and rhythm. These features often reveal the emotional state of a person even when the spoken words remain neutral. On the other hand, text-based emotion detection focuses on the semantic meaning and sentiment conveyed through language. Natural Language Processing techniques such as sentiment analysis, tokenization, and word embeddings help in identifying emotional patterns in written or spoken transcripts. By integrating both voice and textual information, a multimodal approach can significantly improve emotion recognition accuracy and reliability.

The proposed Multimodal Emotion Detection System utilizes both speech signals and textual inputs to detect emotional states such as happiness, sadness, anger, fear, and neutrality. The system involves several stages including data collection, preprocessing, feature extraction, multimodal fusion, and emotion classification using machine learning or deep learning models. The integration of multiple modalities enables the system to capture richer emotional information and produce more accurate predictions. Such systems have significant applications in mental health monitoring, virtual assistants, online counseling platforms, educational environments, and healthcare support systems. They can help identify emotional distress early and provide appropriate guidance or intervention. By leveraging advanced AI technologies, multimodal emotion detection can contribute to enhancing mental well-being, improving communication between humans and machines, and supporting the development of more empathetic and intelligent digital systems.

Ii. Related Works

S. Yoon, S. Byun, and K. Jung, "Multimodal Speech Emotion Recognition Using Audio and Text."
This paper presents a multimodal emotion recognition framework that integrates speech signals and textual transcripts for improved emotion classification. Acoustic features such as pitch, tone, and energy are extracted from speech signals, while textual data is analyzed using natural language processing techniques. Deep neural networks are used to combine both modalities through a fusion mechanism. The proposed model was evaluated using benchmark emotion datasets and achieved higher accuracy compared to single-modality approaches. The results demonstrate that integrating voice and text information enhances the performance of emotion recognition systems and improves applications in human–computer interaction and intelligent conversational systems.

[2] S. W. Byun, J. H. Kim, and S. P. Lee, "Multi-Modal Emotion Recognition Using Speech Features and Text Embedding."
This study proposes a multimodal emotion recognition system that combines speech acoustic features and text embeddings for detecting emotional states. Mel-frequency cepstral coefficients and other spectral speech features are extracted from audio data, while text embeddings capture semantic meaning from speech transcripts. Deep learning architectures such as LSTM networks are used to process these features. The outputs are fused to improve classification performance. Experimental evaluation

shows that the multimodal model significantly outperforms unimodal systems. The research highlights the importance of combining linguistic and acoustic information for accurate emotion detection in affective computing applications.

[3] M. R. Makiuchi, K. Uto, and K. Shinoda, "Multimodal Emotion Recognition with High-Level Speech and Text Features."
This research focuses on emotion recognition using high-level representations extracted from speech and text data. Speech signals are processed using advanced acoustic feature extraction techniques, while textual information is analyzed using transformer-based language models. The outputs from both modalities are combined using score-level fusion for emotion classification. Experiments were conducted using the IEMOCAP dataset, demonstrating improved performance compared with unimodal approaches. The study emphasizes the effectiveness of multimodal learning and deep neural architectures for capturing complex emotional cues in speech and language communication.

[4] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning Alignment for Multimodal Emotion Recognition from Speech."
The authors propose an attention-based alignment framework for multimodal emotion recognition. The system processes speech signals and textual transcripts simultaneously and learns the temporal alignment between acoustic and textual features. By capturing the relationships between spoken words and emotional tone, the model improves emotion classification performance. The approach was evaluated on the IEMOCAP dataset and achieved better results than traditional fusion methods. This work demonstrates that learning cross-modal alignment can effectively capture emotional patterns and enhance the accuracy of emotion recognition systems.

[5] Z. J. Chuang and C. H. Wu, "Multi-Modal Emotion Recognition from Speech and Text."
This paper investigates emotion recognition using both speech and textual modalities. The authors analyze prosodic speech features such as pitch, intensity, and speech rate along with linguistic features extracted from speech transcripts. Machine learning algorithms are used to classify emotional states. Experimental results show that combining acoustic and textual features significantly improves emotion recognition performance compared with using a single modality. This research provides an early foundation for multimodal affective computing systems and demonstrates the importance of integrating different communication channels in emotion detection.

[6] S. Ghosh, U. Tyagi, S. Ramaneswaran, H. Srivastava, and D. Manocha, "MMER: Multimodal Multi-Task Learning for Speech Emotion Recognition."
This study introduces a multimodal multi-task learning framework for speech emotion recognition. The system incorporates speech and textual features and employs cross-modal attention mechanisms to capture interactions between modalities. In addition to emotion classification, the model performs auxiliary tasks to improve representation learning. Experimental results on benchmark datasets demonstrate that the proposed approach achieves state-of-the-art performance. The study highlights the effectiveness of multi-task learning and multimodal fusion techniques for improving emotion recognition accuracy in conversational AI systems.

[7] Y. Lee, S. Yoon, and K. Jung, "Multimodal Speech Emotion Recognition Using Cross Attention with Aligned Audio and Text."
This research proposes a cross-attention-based model for multimodal emotion recognition using aligned speech and textual data. The system divides audio and text inputs into synchronized segments and applies attention mechanisms to learn relationships between them. This interaction enables the model to identify emotional cues more effectively. The proposed approach was tested on emotion recognition datasets and demonstrated improved accuracy compared with conventional multimodal models. The study emphasizes the importance of cross-modal attention in capturing contextual emotional information.

[8] J. Sun, "Research and Application Analysis of Multimodal Emotion Recognition Methods Based on Speech, Text, and Facial Expressions."
This paper provides a comprehensive analysis of multimodal emotion recognition techniques that integrate speech, textual, and facial expression features. The study reviews different feature extraction methods, machine learning models, and fusion strategies used in emotion detection. The research highlights the advantages of multimodal systems in capturing complex emotional information compared with unimodal approaches. Applications of emotion recognition in healthcare, psychology, education, and human–computer interaction are also discussed. The findings indicate that combining multiple modalities improves robustness and accuracy in emotion detection systems.

[9] N. Joshi and R. K. Khare, "A Novel Multimodal Approach for Emotion Recognition Using Text, Speech, and Facial Expression Data."
This study presents a multimodal emotion recognition framework that integrates speech, textual, and facial expression features. Each modality undergoes independent feature extraction and processing using deep learning techniques. An attention-based fusion mechanism combines the outputs from the three modalities to improve emotion classification accuracy. Experimental results show that the proposed system performs better than unimodal emotion recognition systems. The research highlights the potential of multimodal approaches for enhancing emotional intelligence in artificial intelligence systems and interactive applications.

S. Seetha, M. S. V., P. M. Pradyumna, and S. Anthony, "Real-Time Multimodal Emotion Recognition."
This paper proposes a real-time emotion recognition system that integrates speech, text, and facial expressions. The system uses convolutional neural networks for facial feature extraction, recurrent neural networks for speech analysis, and natural language processing techniques for text emotion detection. The outputs are combined to produce accurate emotion predictions. The proposed framework is implemented in a real-time environment and evaluated for performance. The results indicate that multimodal emotion detection improves reliability and accuracy, making it suitable for applications such as mental health monitoring, intelligent assistants, and interactive learning systems.

III. Proposed Method

The proposed system aims to develop a Multimodal Emotion Detection model using voice and text data to assist in monitoring mental well-being. Human emotions are expressed through various communication channels, and analyzing a single source of information may lead to incomplete or inaccurate emotional understanding. To address this limitation, the proposed approach combines speech signals and textual information to enhance the accuracy and reliability of emotion recognition. The system is designed to detect emotional states such as happiness, sadness, anger, fear, and neutral mood by examining both vocal patterns and language usage.

The proposed framework is organized into several key stages, including data collection, preprocessing, feature extraction, multimodal integration, and emotion classification. In the data collection stage, the system gathers voice recordings and textual inputs from users through microphones, chat platforms, or recorded conversations. During preprocessing, the collected data is cleaned and prepared for analysis. Noise reduction and normalization techniques are applied to speech signals, while text data is processed by removing stop words, punctuation, and irrelevant characters.

In the feature extraction phase, meaningful emotional features are derived from both modalities. For voice data, acoustic characteristics such as pitch, tone, speech rate, Mel-frequency cepstral coefficients (MFCC), and energy levels are extracted to represent the emotional tone of the speaker. For text data, Natural Language Processing (NLP) techniques including tokenization, sentiment analysis, and word embedding methods are used to capture semantic and contextual information related to emotions.

Following feature extraction, the system performs multimodal fusion, where the features obtained from voice and text are combined to create a comprehensive emotional representation. Machine learning and deep learning models such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks are used to analyze the integrated features and classify the emotional state.

The final output of the system identifies the user's emotional condition and provides insights related to mental well-being monitoring. This approach can be applied in healthcare support systems, virtual assistants, counseling platforms, and educational environments to promote better emotional understanding and improve human–computer interaction.

International Journal for Research Trends in Social Science & Humanities
Volume 4, Issue 2
Mar-Apr 2026, PP 1-11

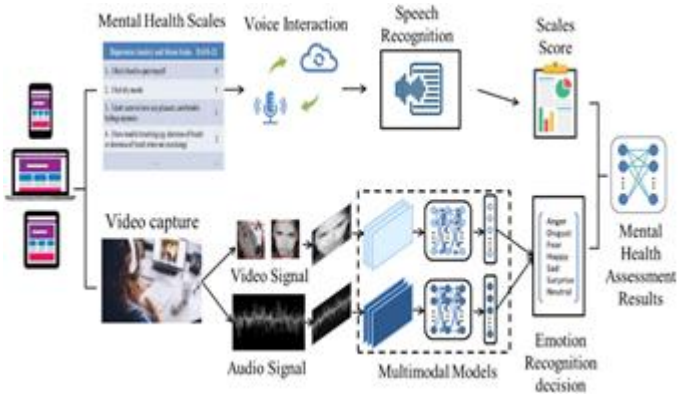

Figure.1.System Architecture

Data Acquisition Module

This module is responsible for collecting the input data required for the emotion detection system. The system gathers voice recordings and textual information from users through microphones, chat interfaces, or available datasets. These inputs represent real-time communication and contain valuable emotional cues embedded in both speech and language. The collected data serves as the primary source for further processing and analysis within the system.

Data Preprocessing Module

The preprocessing module prepares the collected data for efficient analysis. In the case of voice data, processes such as noise filtering, silence removal, and signal normalization are applied to enhance audio quality. For textual data, preprocessing involves removing stop words, punctuation marks, and irrelevant symbols. Additionally, tokenization and text normalization are performed to convert raw text into a structured and machine-readable format.

Feature Extraction Module

This module focuses on identifying meaningful emotional features from both modalities. For speech data, acoustic features such as pitch, tone, speech rate, energy levels, and Mel-Frequency Cepstral Coefficients (MFCC) are extracted to capture the emotional tone of the speaker. For textual inputs, Natural Language Processing (NLP) techniques including sentiment analysis, word embeddings, and contextual feature extraction are used to identify emotional patterns in language.

Emotion Classification and Fusion Module

In this module, the extracted features from voice and text are combined using multimodal fusion techniques. Machine learning or deep learning algorithms analyze the integrated features to classify emotional states such as happiness, sadness, anger, fear, or neutral mood, thereby supporting mental well-being monitoring.

International Journal for Research Trends in Social Science & Humanities
Volume 4, Issue 2
Mar-Apr 2026, PP 1-11

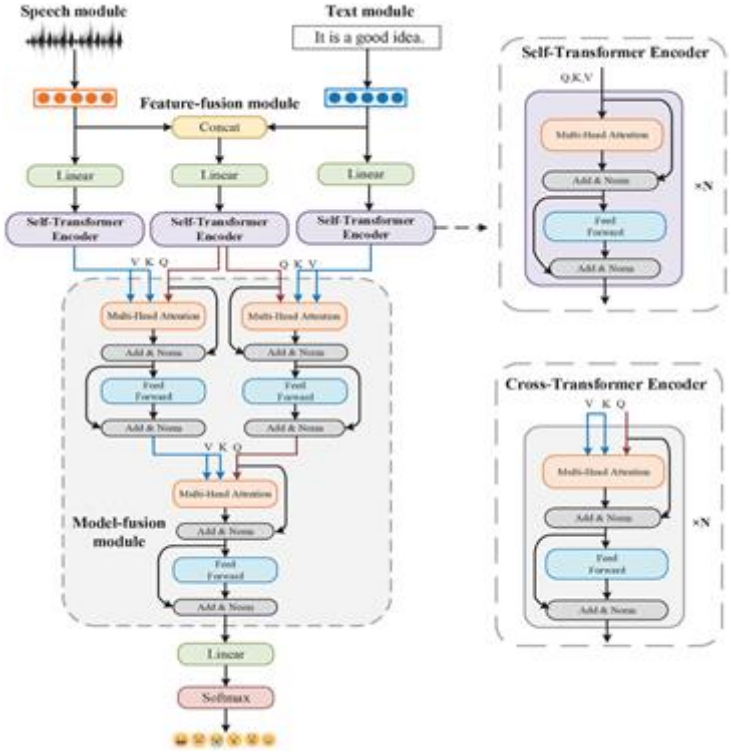

Figure.2. Methodology workflow of the Multimodal Emotion Detection Framework

Overall Working Flow of the Proposed System:
The proposed Multimodal Emotion Detection System using Voice and Text follows a structured workflow to accurately identify human emotions and support mental well-being monitoring. The workflow consists of several sequential stages, including data collection, preprocessing, feature extraction, multimodal fusion, emotion classification, and result generation. Each stage plays a crucial role in transforming raw input data into meaningful emotional insights.

The process begins with the data collection stage, where the system gathers voice recordings and textual inputs from users. Voice data may be captured through microphones or recorded conversations, while text data can be obtained from chat messages, transcriptions, or written inputs. These inputs contain emotional information that reflects the user's psychological state.

Once the data is collected, it moves to the data preprocessing stage, where the raw inputs are cleaned and prepared for analysis. For speech signals, noise reduction, silence removal, and signal normalization are performed to improve audio quality. For textual data, preprocessing techniques such as removing stop words, punctuation, and irrelevant symbols are applied. Tokenization and text normalization help convert the raw text into a structured format suitable for computational analysis.

International Journal for Research Trends in Social Science & Humanities
Volume 4, Issue 2
Mar-Apr 2026, PP 1-11

The next stage is feature extraction, where meaningful emotional characteristics are obtained from both modalities. In voice data, acoustic features such as pitch, tone, speech rate, energy, and Mel-Frequency Cepstral Coefficients (MFCC) are extracted. For text data, Natural Language Processing techniques such as sentiment analysis, word embeddings, and contextual feature extraction are used to identify emotional patterns and semantic meaning.

Following feature extraction, the system performs multimodal fusion, where the features from voice and text are combined to create a unified emotional representation. This integration helps capture complementary emotional cues from both modalities.
Finally, the fused features are processed using machine learning or deep learning models such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), or Long Short-Term Memory (LSTM) networks for emotion classification. The system predicts emotional states such as happiness, sadness, anger, fear, or neutrality and generates the final output, which can be used to support mental well-being analysis and decision-making.

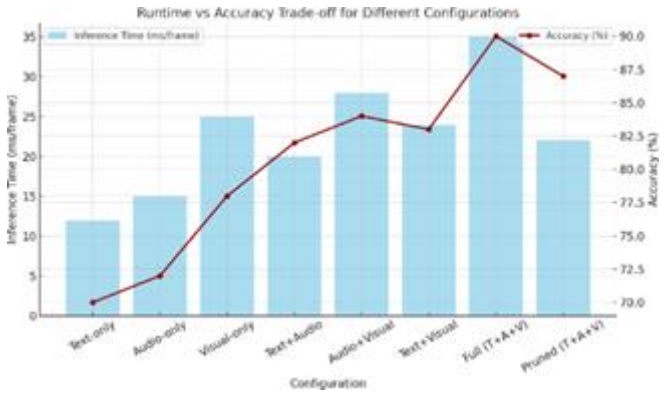

Figure.3.Performance Evaluation of Multimodal Emotion Detection Framework

$$MFCC(n) = \sum_{k=1}^{K} \log(S_k) \cos\left[\frac{\pi n}{K}\left(k - \frac{1}{2}\right)\right]$$

Multimodal Emotion Detection using voice and text plays a significant role in supporting mental well-being by providing a more comprehensive understanding of human emotions. Traditional emotion recognition systems often rely on a single source of data, which may lead to inaccurate interpretations. By integrating both vocal cues and textual information, the proposed system enhances the reliability and accuracy of emotion detection. Voice features such as tone, pitch, and intensity, along with sentiment analysis from textual input, help identify emotional states like happiness, sadness, anger, or stress more effectively.

International Journal for Research Trends in Social Science & Humanities
Volume 4, Issue 2
Mar-Apr 2026, PP 1-11

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}}$$

The Softmax function is commonly used in neural network classifiers to convert raw prediction scores into normalized probability values. In Equation (2), z_irepresents the output score for the i^themotion class generated by the neural network, while Nindicates the total number of emotion categories. The exponential function ensures that all predicted values are positive, and the denominator normalizes them so that the sum of probabilities equals one. This allows the model to interpret predictions as probabilities for different emotional states. In multimodal emotion detection systems, the Softmax function is typically applied at the final layer to determine the most probable emotional category based on fused voice and text features.

Overall, the proposed multimodal approach improves emotional recognition by combining complementary data sources, leading to more accurate and meaningful results. In the future, incorporating additional modalities such as facial expressions and physiological signals could further enhance system performance and contribute to better mental health awareness and support systems.

V. Future Work

Future enhancements of the Multimodal Emotion Detection system can focus on improving the accuracy, adaptability, and real-time performance of emotion recognition for mental well-being applications. One important direction is the integration of additional modalities such as facial expressions, physiological signals (heart rate or brain signals), and behavioral patterns. Combining these modalities with voice and text data can provide a more comprehensive understanding of human emotional states and improve the reliability of the system.

Another potential improvement is the use of advanced deep learning architectures such as transformer-based models and attention mechanisms to better capture complex relationships between speech and textual features. These techniques can enhance feature extraction and improve emotion classification accuracy. Additionally, the system can be trained on larger and more diverse datasets to support multilingual and cross-cultural emotion recognition, making the model more robust and widely applicable.

Future work can also focus on deploying the system in real-time environments such as mobile applications, virtual assistants, and healthcare monitoring platforms. Ensuring data privacy, security, and ethical use of emotional data will also be essential for practical implementation and user acceptance.

References

1. S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in Proc. IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 2018, pp. 112–118.
2. Y. Gu, S. Chen, and I. Marsic, "Deep multimodal learning for emotion recognition in spoken language," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5079–5083.
3. W. Y. Choi, K. Y. Song, and C. W. Lee, "Convolutional attention networks for multimodal emotion recognition from speech and text data," in Proc. ACL Workshop on Human Multimodal Language, 2018, pp. 28–34.
4. L. Cai, Y. Hu, J. Dong, and S. Zhou, "Audio-textual emotion recognition based on improved neural networks," Mathematical Problems in Engineering, vol. 2019, pp. 1–9, 2019.
5. H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," arXiv preprint arXiv:1909.05645, 2019.
6. W. Wu, C. Zhang, and P. C. Woodland, "Emotion recognition by fusing time synchronous and time asynchronous representations," arXiv preprint arXiv:2010.14102, 2020.
7. H. Li, W. Ding, Z. Wu, and Z. Liu, "Learning fine-grained cross modality excitement for speech emotion recognition," arXiv preprint arXiv:2010.12733, 2020.
8. T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3ER: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in Proc. AAAI Conf. Artificial Intelligence, 2020.
9. P. Kumar, V. Kaushik, and B. Raman, "Towards the explainability of multimodal speech emotion recognition," in Proc. Interspeech, 2021.
10. X. Zhang, M. J. Wang, and X. D. Guo, "Multimodal emotion recognition based on deep learning in speech, video and text," in Proc. IEEE Int. Conf. Signal and Image Processing (ICSIP), 2020.
11. A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L. P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset," in Proc. ACL, 2018.
12. S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," Information Fusion, vol. 37, pp. 98–125, 2017.
13. E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, A Practical Guide to Sentiment Analysis. Springer, 2017.
14. B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2013 computational paralinguistics challenge," in Proc. Interspeech, 2013.
15. S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," IEEE Signal Processing Letters, vol. 26, no. 7, pp. 1010–1014, 2019.
16. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL, 2019.
17. A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in Proc. IEEE ICASSP, 2013.

18. Y. Kim, “Convolutional neural networks for sentence classification,” in Proc. EMNLP, 2014.
19. S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
20. A. Vaswani et al., “Attention is all you need,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017.
21. D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, “ICON: Interactive conversational memory network for multimodal emotion detection,” in Proc. EMNLP, 2018.
22. S. Poria, E. Cambria, and A. Gelbukh, “Deep convolutional neural network textual features for multimodal sentiment analysis,” IEEE Intelligent Systems, vol. 31, no. 6, pp. 54–59, 2016.
23. A. Metallinou, A. Lee, and S. Narayanan, “Decision level combination of multiple modalities for emotion recognition,” in Proc. ICASSP, 2010.
24. B. Schuller, F. Eyben, and G. Rigoll, “Static and dynamic modelling for speech emotion recognition,” EURASIP Journal on Audio, Speech and Music Processing, 2009.
25. J. Jiaqi Sun, “Research and application analysis of multimodal emotion recognition methods based on speech, text, and facial ,2023.