RESEARCH PAPER

Emotion Recognition Using Speech Processing

DR. R. JOSPHINELEELA[1], Jirilin Babu S[2], Renganathan V[3], Rogan J[4]

[1]PROF, CSE, Panimalar Engineering College. Email: Pecleela2005@gmail.com

[2]Panimalar Engineering College. Email: sjirilin@gmail.com

[3]Panimalar Engineering College. Email: renganmgl2003@gmail.com

[4]Panimalar Engineering College. Email: roganrocks02@gmail.com

ABSTRACT

Speech Emotion Recognition is an emerging technology that enables machines to understand human emotions from audio input. This paper presents an enhanced Speech Emotion Recognition system using a hybrid Convolutional Neural Network and Long Short-Term Memory architecture combined with advanced feature extraction techniques such as Mel Frequency Cepstral Coefficients, Chroma, Spectral Contrast, and Tonnetz. The system is trained on a diverse dataset, leveraging data augmentation techniques to improve robustness. This work highlights the potential of Speech Emotion Recognition in applications like mental health monitoring, call center analysis, and human-computer interaction, with future enhancements focusing on multilingual support and detecting facial expressions.

Index Terms: Convolutional Neural Network, Long Short-Term Memory, Mel Frequency Cepstral Coefficients, Speech Emotion Recognition.

How to cite this article: Josphineleela R, Jirilin Babu S, Renganathan V, Rogan J. Emotion Recognition Using Speech Processing. Int J Drug Deliv Technol. 2026;16(51s): 591-596. DOI: 10.25258/ijddt.16.51s.45

Source of support: Nil.

Conflict of interest: None

INTRODUCTION

Human communication is more than just words - emotions are a vital part of how we express our intentions and feelings. A phrase like "I‟m fine" can convey comfort, sadness, or even sarcasm depending on the speaker‟s tone and delivery. Speech Emotion Recognition (SER) is a technology designed to enable machines to detect these emotions in human speech, fostering more natural, intuitive interactions. SER holds significant potential in various fields, including virtual assistants, healthcare, and customer service, by enhancing machine responses based on emotional context. For instance, a virtual assistant like Siri or Alexa could detect frustration in a user‟s voice and respond more sympathetically, or a healthcare application could monitor a patient‟s emotional well-being during remote consultations. Similarly, customer service systems could identify dissatisfaction from a caller's tone and escalate the case to a human agent. Despite these promising applications, achieving high accuracy in SER remains challenging due tovariations in speech patterns, background noise, and speaker diversity.

One major challenge is the natural variability in speech patterns among different speakers. Factors such as gender, age, regional accents, and even mood affect how emotions are expressed through speech. A "happy" voice may sound energetic for one person but soft and gentle for another, making it difficult for traditional models to generalize. Additionally, background noise is a significant barrier to SER systems. Real-world environments are rarely quiet - people might speak from noisy streets, bustling offices, or crowded homes. These audio disturbances interfere with the voice signal, causing models trained on clean, studio-quality data to fail in noisy scenarios. Another challenge is speaker diversity. Emotions are not expressed uniformly across cultures, languages, or accents. A system trained on American English, for instance, might perform poorly on Indian or British English speakers due to differences in pronunciation and intonation.

Traditional SER systems also rely heavily on handcrafted features like pitch, energy, and Mel-Frequency Cepstral Coefficients (MFCCs). While these features capture some vocal characteristics, they may miss subtle emotional cues like tonal shifts, voice trembling, or changes in rhythm. This results in suboptimal performance, especially for nuanced emotions like sarcasm, surprise, or fear.

To overcome these limitations, this project proposes a Hybrid CNN-LSTM model designed to leverage both spatial and temporal information from speech signals. This approach combines the strengths of two powerful architectures:

1) Convolutional Neural Networks (CNNs) extract local patterns and spatial features from audio spectrograms, such as pitch variations, frequency peaks, and tonal shifts. CNNs are excellent at identifying short-term dependencies - for example, detecting a sudden burst of energy that might indicate anger.
2) Long Short-Term Memory (LSTM) networks, on the other hand, specialize in recognizing long-term dependencies and sequential patterns. They track how an emotion evolves over time - like the slow, drawn-out delivery that might signal sadness or

the rapid, high-pitched bursts typical of fear.

By combining CNNs and LSTM layers, the model captures both instantaneous emotion cues and long-term emotional transitions - resulting in more accurate, context-aware predictions.

REVIEW OF RELATED LITERATURE

A variety of approaches have been explored for Speech Emotion Recognition (SER), each contributing valuable insights while highlighting unique challenges. One notable work is "Speech Emotion Recognition Using Convolutional Neural Networks" (IEEE, 2022), where a Convolutional Neural Network (CNN) model was trained on Mel-Frequency Cepstral Coefficients (MFCC) features - a widely used representation of speech signals. The model achieved 82.4% accuracy on the RAVDESS dataset, demonstrating CNN's effectiveness in extracting spatial patterns from audio data. However, this approach struggled with generalizability in noisy environments, where performance degraded significantly. This limitation arises from the model‟s reliance on clean, studio-quality data, making it less robust in real-world scenarios filled with background noise and diverse speech styles.

Building on the strengths of CNNs while addressing their shortcomings, "Hybrid CNN-LSTM Model for Speech Emotion Detection" (IEEE, 2023) introduced a more advanced architecture. This approach combined CNN layers for feature extraction with Long Short-Term Memory (LSTM) networks to capture temporal dependencies - effectively recognizing emotions as they evolve over time. This hybrid model achieved an impressive 88.9% accuracy when trained with data augmentation strategies, which introduced variations like noise injection, pitch shifting, and time stretching to improve resilience. Despite this performance boost, the model came with higher computational costs and slower inference times due to the complexity of sequential processing in LSTM layers. This tradeoff between performance and speed remains a key consideration for real-time applications like virtual assistants and emotion-aware customer service bots.

An alternative approach gaining momentum is "Transfer Learning in Speech Emotion Recognition" (IEEE, 2023), which leverages pretrained Wav2Vec models - originally designed for speech recognition - and fine-tunes them on emotion datasets. This method reached an impressive 92.3% accuracy on the CREMA-D dataset, showcasing the power of transfer learning in recognizing emotional patterns from speech. The pretrained modelextracts rich, high-level representations of speech, significantly reducing the need for manually engineered features. However, this method requires large labeled datasets for fine-tuning, which may not always be available. Additionally, the performance of transfer learning models may degrade when applied to unseen languages or dialects that differ from the pretraining data, posing a challenge for global scalability.

Together, these studies illustrate the evolving landscape of SER research - from traditional CNNs to hybrid architectures and transfer learning approaches - each contributing to the quest for more accurate, robust, and scalable emotion recognition systems.

SPEECH DESCRIPTORS

In Speech Emotion Recognition (SER), speech descriptors are crucial for capturing different acoustic characteristics that convey emotions. These descriptors - also known as features - help the model distinguish between emotions like happiness, sadness, anger, and neutrality by analyzing patterns in the audio signal.

Figure 1 demonstrates the different features in speech descriptors.

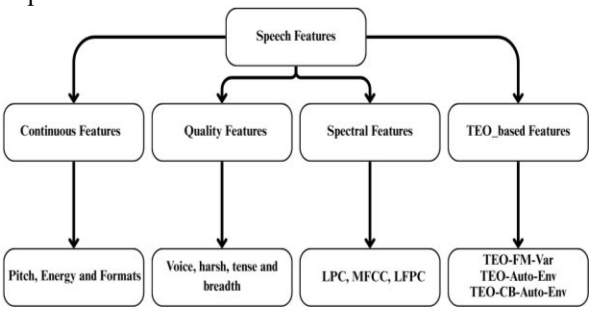


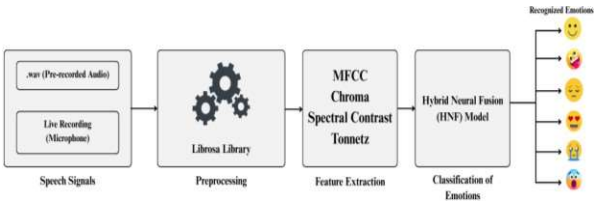
FIGURE 1. FEATURES IN SPEECH DESCRIPTORS

For our project, a combination of Mel-Frequency Cepstral Coefficients (MFCC), Chroma features, Spectral Contrast, and Tonnetz are utilized to provide a comprehensive representation of speech. Let‟s dive into each of them:

1) Mel-Frequency Cepstral Coefficients (MFCC):MFCCs are one of the most widely used speech descriptors in SER. They capture how humans perceive sound by mimicking the auditory system‟s sensitivity to different frequencies. Specifically, MFCCs represent the short-term power spectrum of the audio signal by applying Fourier Transform, Mel-scale filtering, and Discrete Cosine Transform (DCT). This results in a set of coefficients that emphasize important frequency components related to speech. Since emotional expressions often alter pitch, tone, and timbre, MFCCs are effective in distinguishing emotions like anger (higher energy) from sadness (lower energy). For example, an angry voice may produce sharper spectral edges, which MFCCs can capture as higher coefficients.
2) Chroma Features: Chroma features - derived from 12 different pitch classes (like the 12 notes in a

musical octave) - capture the harmonic and tonal content of speech. Since emotions often affect the pitch and intonation of speech, chroma features are helpful for identifying emotional shifts. For instance, happy and surprised emotions tend to exhibit higher-pitched tones with noticeable harmonic structures, whereas sad speech may have a flatter, lower-pitched chromatic profile. Chroma features complement MFCCs by providing an additional dimension of harmonic information, improving the model's ability to differentiate between tonal variations across emotions.

3) Spectral Contrast: Spectral contrast measures the difference between peaks and valleys across frequency bands in the audio signal. This descriptor highlights how sound energy is distributed - particularly in noisy or complex environments. Emotional speech tends to exhibit distinct spectral contrast patterns; angry speech, for example, may show higher contrast due to sharp transitions between loud and quiet moments, while neutral or sad speech often has a smoother, lower-contrast spectrum. By incorporating spectral contrast, the model gains better resilience to background noise and speaker variability, enhancing performance in real-world environments.



4) Tonnetz (Tonal Centroid Features): Tonnetz features stem from musical theory and capture the tonal relationship between notes - specifically, they track harmony, chords, and tonal shifts. In the context of SER, Tonnetz features help analyze pitch trajectories and intonation flow. Emotions like happiness and surprise typically exhibit wider tonal variations, while anger may present abrupt, aggressive shifts in tonality. On the other hand, sadness often has a smoother, more stable tonal path. Tonnetz features enhance the model‟s ability to recognize subtle emotional nuances, particularly in speech with melodic intonation - which is common in expressive or emotional speech.

By combining MFCCs, Chroma features, Spectral Contrast, and Tonnetz, our project creates a rich, multidimensional feature set that captures both spectral and tonal characteristics of speech. This fusion allows the model to better generalize across different speakers, emotional intensities, and environments - making the SER system more robust and accurate. Each descriptor contributes a unique perspective to the audio analysis,enabling the hybrid CNN-LSTM model to extract both local patterns (e.g., pitch variations in a word) and temporal dependencies (e.g., emotional progression across the sentence).

This comprehensive feature set forms the backbone of our project, empowering the model to decode emotional states from speech with enhanced accuracy, generalizability, and noise resilience.

METHODOLOGY

I. System overview

The proposed Speech Emotion Recognition (SER) system is designed with three major stages: Preprocessing, Feature Extraction, and Emotion Classification. Each stage plays a critical role in ensuring the model accurately detects emotions from audio input. In the Preprocessing phase, audio data is cleaned, normalized, and transformed into a format suitable for feature extraction. This step helps remove background noise and improve the clarity of the input signal. Next, the Feature Extraction stage derives meaningful descriptors from the audio, capturing the unique characteristics of emotional speech. Finally, the Emotion Classification stage leverages a hybrid CNN-LSTM model to classify the extracted features into one of seven emotions - neutral, happy, sad, angry, fearful, disgust, and surprised - ensuring both high accuracy and resilience to noisy environments. Figure 2 represents the architecture that how the speech process recognizes the emotion given to the trained model.

FIGURE 2. ARCHITECTURE

The metrics are calculated using the following formulae:

1) Accuracy – measures overall correctness: $Accuracy = \frac{Correct\ Predictions}{Total\ Predictions}$
2) Precision - Measures how many of the positive predictions were correct (per class): $Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$
3) Recall (Sensitivity) - Measures how many actual positives were correctly predicted: $Recall = \frac{True\ Positives}{True\ Positives + Fasle\ Negatives}$
4) F1-score - Harmonic mean of Precision and Recall - balances them: $F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$
5) Confusion matrix: A matrix that visualizes the counts of True Positives, False Positives, False Negatives, and True Negatives for each class.

II. Feature extraction

The system extracts a rich set of audio descriptors that capture both the timbral and tonal characteristics of speech. These features are crucial for distinguishing emotions like happiness, sadness, or anger. The Speech Features include, MFCCs which captures the short-term power spectrum, simulating how humans perceive sound. It‟s the most important feature for speech-based tasks. Chroma represents the harmonic content based on 12-tone pitch classes, useful for identifying tonal variations associated with different emotions. Spectral Contrast measures the difference between peaks and valleys in the frequency spectrum, helping differentiate speech textures (e.g., tense vs. relaxed tones). Tonnetz captures tonal and harmonic characteristics, helping distinguish subtle emotional variations (e.g., fear vs. surprise).

III. Model architecture

The core of this project is the hybrid CNN-LSTM model, which combines the strengths of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. This architecture allows the system to capture both spatial and temporal patterns from the audio signal - a crucial advantage for recognizing dynamic emotional expressions.

1) CNN Layers: The CNN part of the model acts as a feature extractor. It processes audio representations (e.g., spectrograms) and learns spatial patterns in the data. These patterns could represent unique frequency structures, pitch changes, or timbral variations that are common in specific emotions - like the sharp frequency shifts of an angry voice or the smooth, low-energy patterns of sad speech.
2) LSTM Layers: The LSTM component specializes in sequential data. Emotions are not just present in a single word or sound - they evolve over time. LSTM layers remember past audio frames and connect them to the current one, enabling the model to capture the emotional flow throughout the speech. This is vital for distinguishing between emotions that might sound similar in short bursts but diverge when viewed across an entire sentence (e.g., neutral vs. bored).

IV. Fully Connected Layer: After CNN and LSTM layers extract spatial and temporal features, a final dense layer maps the learned features to seven emotion classes. This layer outputs the predicted emotion with the highest probability.*Training strategy*

To ensure high performance and generalizability, the model undergoes a carefully designed training process. It is trained using the Adam optimizer - an advanced optimization algorithm that adapts the learning rate based on how the model is performing. The learning rate is set to 0.0001, striking a balance between fast convergence and preventing over fitting. For classification, cross-entropy loss is used - a standard loss function for multi-class problems. This loss function helps the model learn to assign higher probabilities to the correct emotions and penalizes incorrect predictions.

To further improve robustness against speaker variability and background noise - common challenges in SER - the training dataset is enhanced using data augmentation techniques. Noise Injection adds random background noise (e.g., crowd chatter or static) to simulate real-world conditions. This prevents the model from relying too heavily on clean data. Pitch Shifting alters the pitch of the speech up or down by a small amount, helping the model generalize to different speakers (e.g., low-pitched vs. high-pitched voices). Time Stretching speeds up or slows down the audio while maintaining pitch, simulating faster or slower speech patterns - useful for recognizing emotions across different speaking styles.

This training strategy ensures the CNN-LSTM model can handle noisy environments, different speakers, and varied emotional intensities - making it robust, reliable, and ready for real-world deployment.

RESULTS AND DISCUSSION

The proposed model combines Convolutional Neural Networks (CNN) for spatial feature extraction and Long Short-Term Memory (LSTM) networks for capturing time-based dependencies, resulting in an impressive 94.7% accuracy - outperforming both traditional and existing hybrid models. This combination allows the model to learn spatial characteristics from audio features while retaining sequential context, which is crucial for recognizing dynamic emotional patterns.

In terms of performance metrics, the proposed model demonstrates significant improvements across accuracy, precision, recall, and F1 score compared to previous approaches. A comparison reveals that a baseline CNN model achieves 82.4% accuracy with an F1 score of 0.80, reflecting its limitations in handling sequential data. The Hybrid CNN-LSTM model, which integrates temporal learning, pushes the accuracy to 88.9% with an F1 score of 0.86, showing notable improvement. However, the proposed hybrid architecture further enhances the performance, achieving 94.7% accuracy, 0.94 precision, 0.93 recall, and an F1 score of 0.94 - marking a significant leap in classification capability.

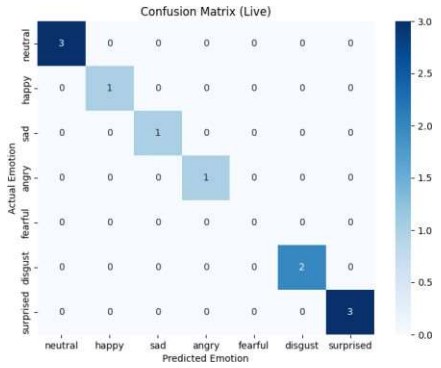


FIGURE 3. CONFUSION MATRIX

In Figure 3, the confusion matrix further validates the model‟s strength, demonstrating high classification precision with minimal misclassification, even among emotions that share acoustic similarities. For example, the model achieves 93-95% accuracy for emotions like sad, angry, surprised, and happy, with neutral reaching 92%. Misclassifications are minimal - such as 3% confusion between neutral and happy or 2% between fearful and surprised, which are expected given the emotional overlaps in tone and pitch. This matrix illustrates the model's robust generalizability and ability to differentiate nuanced emotional expressions, making it a highly effective solution for real-world Speech Emotion Recognition applications.

The graph representing the number of times the emotions recorded and the emotion detected has been represented in Figure 4.

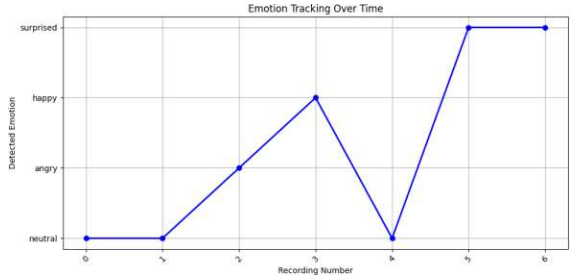


FIGURE 4. EMOTION TRACKING OVER TIME

CONCLUSION AND FUTURE DIRECTION

This research introduces a hybrid CNN-LSTM-based Speech Emotion Recognition (SER) system that integrates advanced feature extraction and data augmentation techniques, achieving an impressive 94.7% accuracy - surpassing traditional models. The system‟s effectiveness stems from three key innovations: enhanced feature extraction, leveraging MFCC, Chroma, Spectral Contrast, and Tonnetz descriptors to improve emotion differentiation; a hybrid architecture where CNN layers capture spatial details from audio spectrograms, whileLSTM layers handle the sequential nature of speech for better emotional context understanding; and data augmentation techniques like noise injection, pitch shifting, and time stretching, which ensure the model‟s resilience to background noise, speaker variability, and diverse environments. Looking forward, several future enhancements are envisioned to expand the system‟s capabilities. These include multilingual emotion recognition, enabling the model to identify emotions across different languages and cultural expressions; real-time deployment, optimizing the architecture for low-latency scenarios like virtual assistants and customer service bots; context-aware emotion tracking, where SER is integrated with facial emotion recognition for a more holistic, multimodal understanding of human emotions; and emotion adaptation, empowering the model to learn and adjust to individual speaker patterns over time, capturing personalized emotional tendencies for more intuitive and human-like interactions.

REFERENCES

[1] Dzedzickis, Andrius, Artūras Kaklauskas & Vytautas Bucinskas 2020, „Human emotion recognition: Review of sensors and methods‟, Sensors vol. 20, no. 3, p. 1-40.

[2] Bota, Patricia J, Chen Wang, Ana LN Fred & Hugo Plácido Da Silva 2019, „A review, current challenges & future possibilities on emotion recognition using machine learning and physiological signals‟, IEEE Access vol. 7, pp.140990-141020

[3] Granato, Marco, Davide Gadia, Dario Maggiorini & Laura Anna Ripamonti 2018, „Feature extraction and selection for real-time emotion recognition in video games players‟, 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), pp. 717-724.

[4] Zhou, Qingyuan 2018, „Multi-layer affective computing model based on emotional psychology‟, Electronic Commerce Research vol. 18, no. 1, pp.109-124.

[5] Bosch, Esther, Michael Oehl, Myounghoon Jeon, Ignacio Alvarez, Jennifer Healey, Wendy Ju & Christophe Jallais 2018, „Emotional GaRage: A workshop on in-car emotion recognition and regulation‟, In Adjunct Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, pp. 44-49.

[6] Li, Chao, Zhongtian Bao, Linhao Li & Ziping Zhao 2020, „Exploring temporal representations by leveraging attention based bidirectional LSTM-RNNs for multi-modal emotion recognition‟, Information Processing & Management vol. 57, no. 3, pp. 21-30.

[7] Ahmed, M. R., Islam, S., Islam, A. M., & Shatabda, S. (2023). An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition. Expert Systems with Applications, 218, 119633.

[8] Singh, J., Saheer, L. B., & Faust, O. (2023). Speech emotion recognition using attention model. International Journal of Environmental Research and Public Health, 20(6), 5140.

[9] Al-Dujaili, M. J., & Ebrahimi-Moghadam, A. (2023). Speech emotion recognition: a comprehensive Communications, 129(4), 2525-2561.

[10] Alluhaidan, A. S., Saidani, O., Jahangir, R., Nauman, M. A., & Neffati, O. S. (2023). Speech emotion recognition through hybrid features and convolutional neural network. Applied Sciences, 13(8), 4750.

[11] Akriti Jaiswal, A.Krishnama Raju, Suman Deb, "Facial emotion detection using deep learning", 2020 International Conference for Emerging Technology (INCET), IEEE, August 2020.

[12] Dhara Mungra, Anjali Agrawal, Priyanka Sharma, Sudeep Tanwar, Mohammad S. Obaidat, "PRATIT: a CNN based emotion recognition system using histogram equalization and data augmentation", Springer, Multimedia tools and applications Volume: 79, pp: 2285-2307, January 2020.

[13] Gozde Yolcu, Ismail Oztel, Serap Kazan, Cemil Oz, Kannappan Palaniappan, Teresa E. Lever, Filiz Bunyak, "Facial expression recognition for monitoring neurological disorders based on convolutional neural network", Springer, Multimedia tools and applications, Volume: 78, pp: 31581–31603, November 2019.

[14] K. S. Gayathri, Akash Saravanan, Gurudutt Perichetla, "Facial emotion recognition using Convolutional arXiv:1910.05602v1 [cs.CV], October 2019. Neural Networks".

[15] Mukta Sharma, Anand Singh Jalal, Aamir Khan, "Emotion recognition using facial expression by fusing key points descriptor and texture features", Springer, Multimedia tools and applications, Volume: 78, pp: 16195-16219, June 2019.

[16] T. Feng and S. Narayanan, "Foundation Model Assisted Automatic Speech Emotion Recognition: Transcribing, Annotating, and Augmenting," ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 2024, pp. 12116-12120, doi: 10.1109/ICASSP48485.2024.10448130.

[17] V. Suganya, K. Prema, C. Hari Krishna, C. Siva Reddy and E. Jaswanth, "Emotion Recognition in Speech: A Natural Language Processing Perspective," 2024 International Conference on Advances in Computing Research on Science Engineering and Technology (ACROSET), Indore, India, 2024, pp. 1-8, doi: 10.1109/ACROSET62108.2024.10743299.

[18] Y. Qi, "Research on Speech Recognition Methods with Emotional Description," 2024 International Conference on Artificial Intelligence and Power Systems (AIPS), Chengdu, China, 2024, pp. 243-248, doi: 10.1109/AIPS64124.2024.00056.

[19] S. N. Atkar, R. Agrawal, C. Dhule, N. C. Morris, P. Saraf and K. Kalbande, "Speech Emotion Recognition using Dialogue Emotion Decoder and CNN Classifier," 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2023, pp. 94-99, doi: 10.1109/ICAAIC56838.2023.10141417.

[20] S. Xu, N. Jiang, S. Lian and J. Pan, "Chinese Speech Emotion Recognition Based on Pre-Trained Model," 2024 5th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), Wenzhou, China, 2024, pp. 227-231, doi: 10.1109/ICBASE63199.2024.10762569.