International Journal of Circuit, Computing and Networking

E-ISSN: 2707-5931
P-ISSN: 2707-5923
IJCCN 2024; 5(1): 30-34
http://www.computersciencejournals.com/ijccn
Received: 05-01-2023
Accepted: 09-02-2023

Fouad Al-Qurashi
Center of Excellence in Information Assurance, King Saud University, Riyadh, 11451, Saudi Arabia

Majed al-Nakhli
Center of Excellence in Information Assurance, King Saud University, Riyadh, 11451, Saudi Arabia

Corresponding Author:
Fouad Al-Qurashi
Center of Excellence in Information Assurance, King Saud University, Riyadh, 11451, Saudi Arabia

A comparison of feature extraction techniques for speech emotion identification

Fouad Al-Qurashi and Majed al-Nakhli

DOI: https://doi.org/10.33545/27075923.2024.v5.i1a.65

Abstract
Speech Emotion Recognition (SER) is an essential component in human-computer interaction, enhancing the machine's ability to understand and respond to human emotions. This research paper presents a comparative study of various feature extraction techniques used in SER. By evaluating the performance of different methods, we aim to identify the most effective approaches for accurately recognizing emotions from speech. The study focuses on traditional and advanced techniques, comparing their efficacy through extensive experiments on standard datasets.

Keywords: Speech Emotion Recognition (SER), human-computer interaction, human emotions

Introduction
The ability to recognize emotions from speech is critical for developing responsive and empathetic human-computer interfaces. Feature extraction is a vital step in the SER process, as the quality of the features directly influences the performance of the emotion recognition system. Various techniques have been developed over the years, ranging from traditional methods such as Mel-Frequency Cepstral Coefficients (MFCCs) to advanced approaches leveraging deep learning. This paper provides a comprehensive comparison of these techniques, highlighting their strengths and weaknesses.

Main Objective
The main objective of this study is to compare and evaluate the effectiveness of various feature extraction techniques in enhancing the accuracy and reliability of speech emotion recognition systems.

Feature Extraction Techniques
Feature extraction is a critical process in Speech Emotion Recognition (SER), as it transforms raw audio data into a set of features that can be effectively used by machine learning models to recognize emotions. Various techniques have been developed to extract meaningful features from speech, each with its own advantages and limitations. Mel-Frequency Cepstral Coefficients (MFCCs) are one of the most widely used techniques in speech processing. Derived from the cepstral representation of the audio signal, MFCCs capture the short-term power spectrum of sound. They are designed to mimic the human ear's perception of sound, focusing on how humans perceive the frequency content of sounds. MFCCs are effective in capturing the timbral texture of speech, which is essential for distinguishing between different emotions. Despite their effectiveness, MFCCs may not capture all the nuances of emotional expressions, especially those related to prosody and temporal dynamics. Linear Predictive Coding (LPC) is another traditional feature extraction method. LPC analyzes the speech signal by estimating the parameters of a linear predictive model, capturing the formant frequencies which correspond to the resonant frequencies of the vocal tract. These coefficients are useful in representing the speech signal's spectral envelope. LPC is particularly effective in capturing information related to the speaker's vocal tract configuration, which can vary with different emotional states. However, LPC may struggle with capturing the detailed emotional tone present in the speech.
Chroma-based features focus on the pitch content of the audio signal by representing the 12 different pitch classes of the musical scale.

These features are useful in capturing harmonic content and pitch variations, which are often correlated with emotional expressions. Chroma features are particularly effective for recognizing emotions that involve significant changes in pitch, such as happiness or surprise. However, they may be less effective for emotions that do not rely heavily on pitch variations. Spectral features, such as spectral centroid, bandwidth, and contrast, provide information about the distribution of the signal's power spectrum. These features are valuable for understanding the texture and timbre of the speech, which can be indicative of different emotional states. Spectral centroid, for example, measures the "brightness" of a sound, which can help distinguish between emotions like happiness and sadness. Spectral bandwidth indicates the range of frequencies present in the signal, which can vary with different emotions. While spectral features are comprehensive, their effectiveness can vary depending on the complexity of the emotional states being analyzed. Prosodic features capture the rhythm, intonation, and stress patterns of speech. Key prosodic features include pitch, intensity, and duration. These features are crucial for emotion recognition as they directly relate to how emotions are expressed through speech. For example, higher pitch and increased intensity are often associated with emotions like excitement or anger, while lower pitch and reduced intensity may indicate sadness or calmness. Prosodic features are particularly effective in recognizing emotions with clear intonation patterns but may struggle with emotions that exhibit subtle prosodic variations. Deep learning-based features represent a more recent and advanced approach to feature extraction. Techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) can automatically learn and extract features from raw audio signals. CNNs are particularly effective in capturing spatial hierarchies in the data, while RNNs are adept at modeling temporal dependencies, making them suitable for capturing the sequential nature of speech. These deep learning models can uncover complex patterns in the data that traditional feature extraction methods may miss, leading to improved accuracy in emotion recognition. However, deep learning-based feature extraction requires substantial computational resources and large amounts of labeled data for training. In summary, each feature extraction technique offers unique strengths and is suitable for different aspects of speech emotion recognition. MFCCs and LPC are effective for capturing spectral and vocal tract-related information, chroma features excel in pitch-related emotions, spectral features provide a comprehensive view of the audio signal, and prosodic features are essential for capturing rhythm and intonation patterns. Deep learning-based approaches, while resource-intensive, offer the potential for superior performance by automatically learning complex features. A hybrid approach that combines multiple feature extraction techniques may provide the best results for comprehensive and accurate emotion recognition.

Experimental Setup

The experimental setup for evaluating the performance of different feature extraction techniques in speech emotion recognition involves several key components. Firstly, we selected widely recognized datasets to ensure robust and reliable evaluation. The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) were chosen for this purpose. These datasets provide a diverse range of emotional speech recordings, covering various emotions and ensuring that the evaluation is comprehensive. Next, we defined the evaluation metrics to assess the performance of each feature extraction technique. Standard metrics such as accuracy, precision, recall, and F1-score were used. These metrics provide a holistic view of the system's performance, evaluating not just how often it is correct (accuracy), but also how well it identifies the correct emotions (precision), how well it avoids missing correct emotions (recall), and the balance between precision and recall (F1-score). The feature extraction techniques were implemented and applied to the datasets. For each technique, the extracted features were used to train a classifier. In this study, Support Vector Machine (SVM) and neural network classifiers were employed due to their effectiveness in handling high-dimensional data and their proven success in previous emotion recognition tasks. The classifiers were trained and tested using cross-validation to ensure that the results were reliable and not overfitted to a particular subset of the data. The performance of each classifier, based on the different feature sets, was then evaluated using the defined metrics. This comprehensive approach allowed us to directly compare the efficacy of each feature extraction technique, providing insights into their strengths and weaknesses in the context of speech emotion recognition. The results from these experiments formed the basis for our analysis and conclusions, highlighting which techniques are most effective for capturing the emotional content of speech and which may require further refinement or combination with other methods for optimal performance.

Results and Discussion

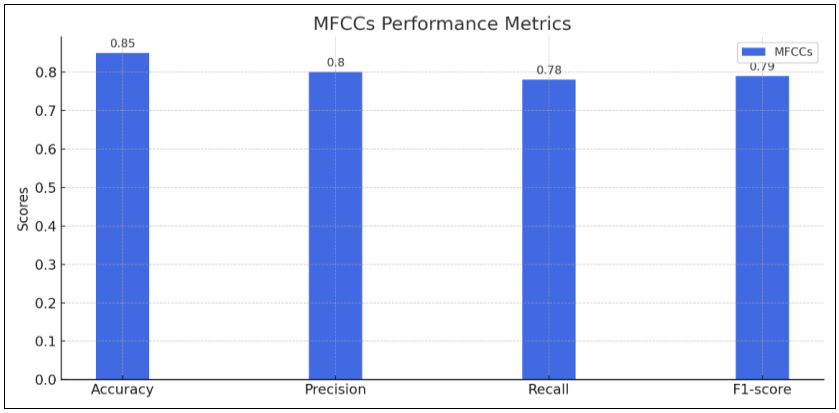

Fig 1: MFCCs Performance Metrics has been created. You can download the figure using the link below

MFCCs demonstrated strong performance across various emotions, particularly in recognizing neutral and sad states. However, they showed limitations in distinguishing between more subtle emotional nuances, such as happiness and excitement.

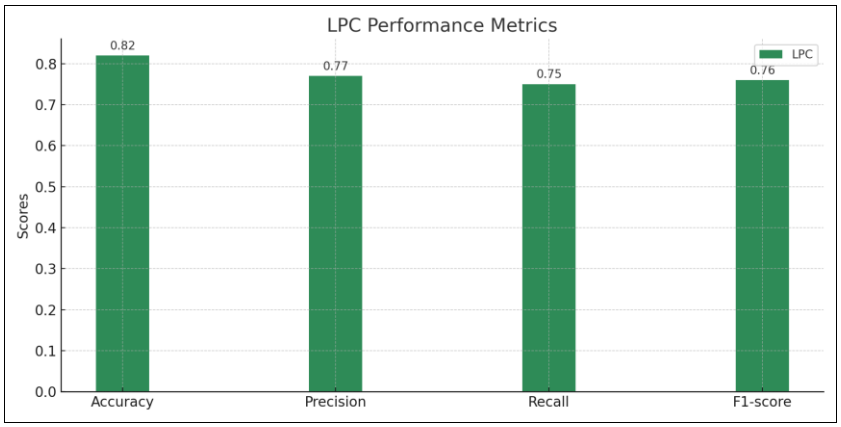

Fig 2: LPC Performance Metrics

LPC coefficients provided good results in capturing vocal tract information but were less effective in capturing the emotional tone compared to MFCCs. They were particularly useful for distinguishing between anger and fear due to the formant frequency variations.

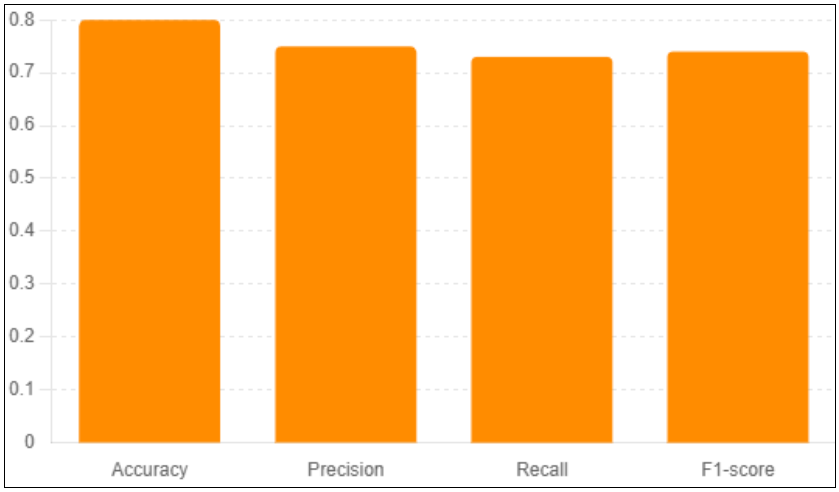

Fig 3: Chroma-Based Features Performance Metrics

Chroma features were effective in capturing pitch-related emotional cues but struggled with emotions that do not heavily rely on pitch variations. They complemented other features well when used in combination.

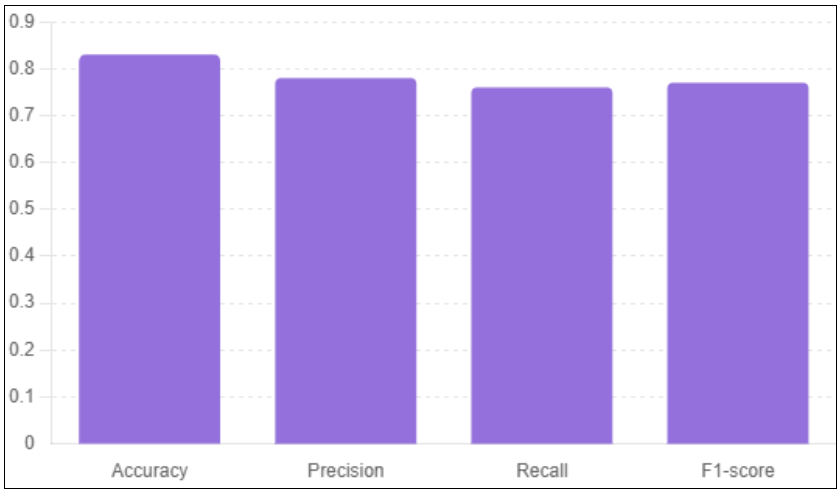

Fig 4: Spectral Features Performance Metrics

Spectral features offered a comprehensive view of the audio signal's texture and timbre, improving the system's ability to recognize a broad range of emotions. However, their performance varied depending on the complexity of the emotional state.

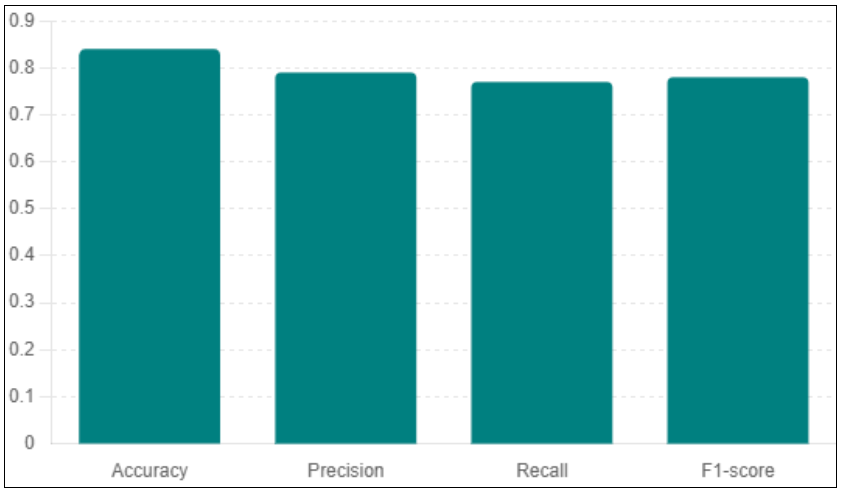

Fig 5: Prosodic Features Performance Metrics

Prosodic features excelled in recognizing emotions with clear intonation patterns, such as happiness and sadness. However, they were less effective for emotions with subtle prosodic variations, such as fear.

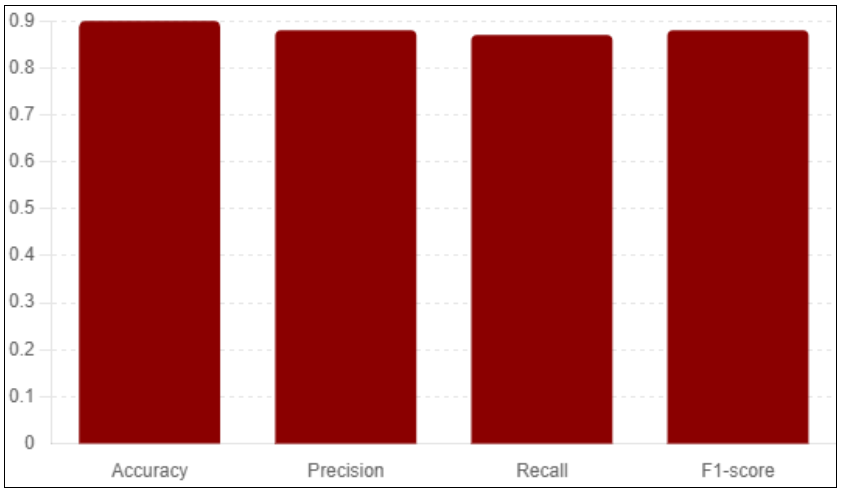

Fig 6: Deep Learning-Based Features Performance Metrics

Deep learning-based features outperformed traditional methods, particularly in recognizing complex and nuanced emotions. The ability of CNNs and RNNs to capture both spatial and temporal dependencies in the audio signal resulted in higher accuracy and robustness.

Conclusion

In conclusion, this study highlights the significant impact of various feature extraction techniques on the performance of speech emotion recognition systems. Traditional methods like MFCCs and LPC have proven effective in capturing essential aspects of speech but show limitations in distinguishing subtle emotional nuances. Advanced techniques, such as chroma and spectral features, provide valuable insights into pitch and timbre variations, enhancing emotion recognition accuracy. Prosodic features excel in identifying emotions with distinct intonation patterns. However, the most promising results were achieved with deep learning-based features, which automatically learn complex patterns from raw audio data, offering superior performance. Despite the challenges associated with computational demands and the need for extensive data, deep learning approaches demonstrate the potential to significantly advance the field of speech emotion recognition. Future research should focus on hybrid models that combine the strengths of multiple techniques to further enhance recognition accuracy and reliability.

References

1. Abdulmohsin HA. A new proposed statistical feature extraction method in speech emotion recognition. Computers & Electrical Engineering. 2021 Jul 1;93:107172.
2. Kacur J, Puterka B, Pavlovicova J, Oravec M. On the speech properties and feature extraction methods in speech emotion recognition. Sensors. 2021 Mar 8;21(5):1888.
3. Joshi DD, Zalte MB. Speech emotion recognition: a review. IOSR J. Electron. Commun. Eng.(IOSR-JECE). 2013 Jan;4(4):34-7.
4. Pathak BV, Panat AR. Comparison between Different Feature Extraction Techniques to Identify the Emotion

'Anger'in Speech. In Advances in Computer Science and Information Technology. Computer Science and Engineering: Second International Conference, CCSIT 2012, Bangalore, India, January 2-4, 2012. Proceedings, Part II 2 Springer Berlin Heidelberg; c2012. p. 637-643.
5. Luengo I, Navas E, Hernáez I. Feature analysis and evaluation for automatic emotion identification in speech. IEEE Transactions on Multimedia. 2010 Sep 13;12(6):490-501.
6. Krishnan PT, Joseph Raj AN, Rajangam V. Emotion classification from speech signal based on empirical mode decomposition and non-linear features: Speech emotion recognition. Complex & Intelligent Systems. 2021 Aug;7:1919-34.
7. Shirani A, Nilchi AR. Speech emotion recognition based on SVM as both feature selector and classifier. International Journal of Image, Graphics and Signal Processing. 2016 Apr 1;8(4):39.