Speech Emotion Recognition in Male Speech Using 58-Dimensional MFCC Features and Random Forest Classification

Aris Rakhmadi[1]✉, Dewi Soyusiawaty[2], Irma Yuliana[3] and Dimara Kusuma Hakim[4]
[1,3]Universitas Muhammadiyah Surakarta, Indonesia
[2]Universitas Ahmad Dahlan, Indonesia
[4]Universitas Muhammadiyah Purwokerto, Indonesia

✉Corresponding Author: aris.rakhmadi@ums.ac.id

ABSTRACT

Speech emotion recognition (SER) has developed into a significant research topic in affective computing and human–computer interaction because emotional cues embedded in speech signals can enhance communication between humans and intelligent systems. However, accurately identifying emotional states from speech remains challenging due to dissimilarities in acoustic patterns, speaker features, and recording situations. This study investigates the effectiveness of Mel-Frequency Cepstral Coefficient (MFCC) acoustic features for emotion recognition in male speech using a Random Forest classification model. The dataset used in this research consists of 35,910 male speech samples, each represented by a 58-dimensional MFCC feature vector extracted from emotional speech recordings. The speech samples are categorized into eight emotional classes: angry, fear, calm, disgust, neutral, happy, sad, and surprise. To develop and evaluate the model's performance, the MFCC data were divided into 80% for training and 20% for testing. The Random Forest model was trained to learn emotional patterns embedded in MFCC features. The experimental findings reveal that the proposed approach achieved an overall classification accuracy of 84.33% with a macro-average F1-score of 0.856, indicating relatively stable performance across multiple emotional categories. Feature importance analysis further reveals that lower-order MFCC coefficients play a dominant role in emotion classification. These findings demonstrate that MFCC features combined with Random Forest classification provide an effective baseline approach for SER and offer valuable insights for future research involving more advanced machine learning models.

Keywords: speech emotion recognition, MFCC, random forest, acoustic features, machine learning.

A. Introduction

Speech produced by humans contains both linguistic meaning and emotional indicators that reflect a speaker's psychological and affective state. The ability to automatically recognize emotions from speech signals has therefore become an imperative research topic in the fields of affective computing, human–computer interaction, and intelligent systems. SER intends to recognize emotional statuses such as anger, happiness, sadness, or fear based on acoustic characteristics contained in speech signals. Accurate emotion recognition systems can support a wide range of applications, including intelligent virtual assistants, healthcare monitoring systems, customer service analytics, and adaptive learning environments [1].

Despite significant advancements in speech processing technologies, accurately identifying emotions from speech remains a challenging task. Multiple factors, including speaker characteristics, vocal variability, recording conditions, and language differences, influence emotional expression in speech [2], [3]. These factors often introduce variability in acoustic patterns, making it difficult for classification models to reliably

distinguish between emotional states. Consequently, the development of robust feature extraction techniques and classification models remains a central focus in SER research.

One of the major widely used feature extraction methods in speech processing is the MFCC. MFCC features are designed to capture the perceptual characteristics of human hearing by representing the spectral envelope of speech signals in the Mel frequency scale [4], [5], [6]. Due to their ability to effectively represent the acoustic properties of speech, MFCC features have been widely used in speech-related tasks, consisting of speech recognition, speaker identification, and emotion detection. Previous studies have demonstrated that MFCC-based representations can capture relevant spectral information that reflects emotional variations in speech production.

In addition to feature extraction, the model selection of classification algorithms plays a crucial role in defining the performance of speech emotion recognition systems. Various machine learning systems have been analysed in previous studies, including k-Nearest Neighbors (k-NN), Decision Trees, Support Vector Machines (SVMs), and Random Forest classifiers [7], [8]. Among these approaches, Random Forest has attracted considerable attention because it delivers reliable performance on high-dimensional data while exhibiting a lower tendency toward overfitting. Random Forest constructs various decision trees and employs their predictions, enabling the model to learn complex relationships between acoustic features and emotional states.

Another important aspect in SER research is the characteristics of the speech dataset used for training and evaluation. Many previous studies rely on datasets that contain recordings from multiple speakers and both genders [9]. While such datasets provide diverse speech samples, the mixture of male and female speech signals may introduce additional variability in acoustic patterns [10], [11]. Gender differences in vocal tract configuration, pitch range, and speech characteristics may influence the distribution of acoustic features, potentially affecting the performance of emotion classification models.

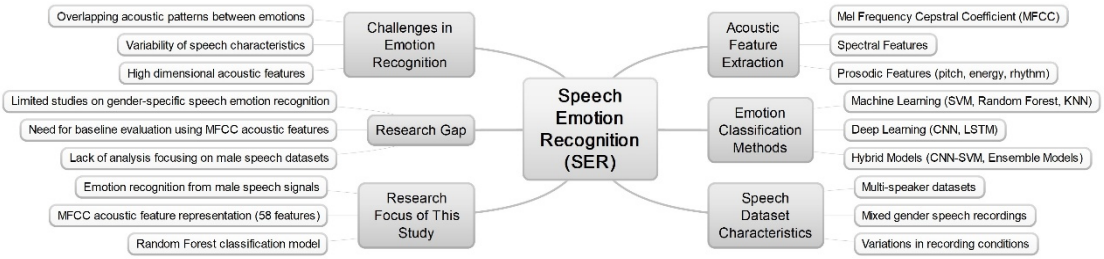

Figure 1. Research Landscape and Gap in SER Using Acoustic Features.

Figure 1 presents a conceptual mind-map illustrating the research landscape of SER. The mind-map highlights several important components of SER research, including acoustic feature extraction methods, classification approaches, dataset characteristics, and common challenges in emotion recognition tasks. The figure also identifies an important research gap in the limited exploration of gender-specific emotional speech analysis [12], [13]. While many studies investigate emotion recognition using mixed-gender datasets, relatively few focus on analyzing emotional patterns in single-gender speech datasets. This gap suggests the need for further investigation to understand better how acoustic features capture emotional variations in gender-specific speech data.

Motivated by this research gap, this study evaluates SER on a male speech dataset using MFCC acoustic features. By focusing on a single-gender dataset, variability due to gender differences in speech characteristics can be reduced, enabling clearer analysis of

emotional patterns in speech signals. In this research, each speech sample is represented by a 58-dimensional MFCC feature vector that captures the spectral characteristics of speech associated with emotional expression.

Therefore, the present study focuses on assessing the effectiveness of MFCC acoustic features for SER using a Random Forest classification model. The research aims to evaluate how well machine learning algorithms can identify emotional states in male speech signals using MFCC feature representations. The results of this study are expected to provide a reliable baseline for SER research and contribute to the development of more advanced emotion recognition models in future studies.

B. Methodology

This research adopts a machine learning–based approach to perform SER using acoustic feature representations extracted from emotional speech recordings. The research process consists of several stages, including dataset preparation, feature representation, data preprocessing, model development, and performance evaluation. The overall workflow of the research process is illustrated in Figure 2.

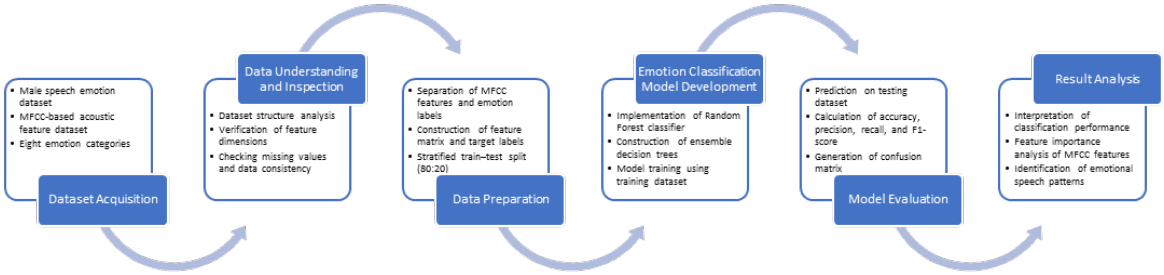

Figure 2. Research Workflow for SERusing MFCC Features and Random Forest Classification

The dataset applied in this research is composed of MFCC-based acoustic feature representations derived from emotional speech recordings collected from several widely used benchmark datasets. The original audio recordings were obtained from four well-known emotional speech datasets: the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [14], the Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D) [15], the Toronto Emotional-Speech Set (TESS) [16], and the Surrey Audio-Visual Expressed Emotion (SAVEE) [17]. These datasets contain speech recordings expressing multiple emotional states, produced by different speakers and recorded under varying conditions. The MFCC features used in this study were extracted from these audio recordings and organized into structured feature datasets.

Each speech sample is represented by a 58-dimensional Mel-Frequency Cepstral Coefficient (MFCC) feature vector, which captures the spectral characteristics of speech signals using the Mel frequency scale, which approximates human auditory perception [18], [19]. MFCC features are widely used in speech processing because they effectively represent the spectral envelope of speech signals and preserve important acoustic information related to emotional expression. In this study, the analysis focuses on the male speech dataset, which contains 35,910 speech samples distributed across eight emotional classes: angry, fear, happy, calm, disgust, sad, neutral, and surprise. Each instance in the dataset represents a speech sample with MFCC feature values and a corresponding emotion label.

Before model training, the dataset undergoes a data inspection stage to ensure its consistency and appropriateness for machine learning analysis. This stage includes examining the dataset structure, verifying the MFCC feature dimensionality, and checking for missing values [20], [21]. After this inspection process, the MFCC feature vectors are separated from the emotion labels to form the input feature matrix and the target classification variable.

To develop and evaluate the classification model's performance, eighty percent of the data were used for model training, while the remaining twenty percent was reserved for testing. Stratified sampling is applied during the splitting process to preserve the distribution of emotion classes in both subsets. The training dataset is used to develop the classification model, while the testing dataset is used to evaluate the model's ability to recognize emotional patterns from unseen speech samples.

The classification model employed in this study is the Random Forest algorithm, a learning model that integrates multiple decision trees. Random Forest constructs a collection of decision trees using randomly selected subsets of training data and feature variables. The final prediction is determined by majority voting among the individual decision trees. In this research, the Random Forest classifier is implemented using 100 decision trees to capture complex relationships between MFCC acoustic features and emotional states.

The effectiveness of the proposed model was examined using typical classification indicators: accuracy, precision, recall, and F1-score. These evaluation metrics provide a comprehensive assessment of the model's ability to classify emotional speech signals across multiple categories. In addition, a confusion matrix is used to visualize classification results and identify potential misclassification patterns among emotional categories. Feature importance analysis is also conducted to determine the contribution of individual MFCC features to the classification process and to understand better which acoustic characteristics play significant roles in emotion recognition.

C. Results and Discussion

The performance of the SER system developed in this study was evaluated using the Random Forest classification algorithm with MFCC acoustic features as input variables. The evaluation results are summarized in Table 1, which presents the classification report for each emotional category, including recall, precision, F1-score, and support values. Overall, the model achieved a classification accuracy of 84.33%, indicating that the combination of MFCC acoustic features and the Random Forest algorithm provides a reliable approach for recognizing emotional states in speech signals.

The macro-average F1-score from the evaluation is 0.856, indicating that the classification model performs moderately consistently across emotional categories. This metric is important because it reflects the average performance across all classes without being influenced by class imbalance. The result indicates that the model can learn acoustic patterns associated with emotional expression in speech signals across multiple categories.

A closer examination of the classification performance reveals that several emotional categories achieve relatively high recognition accuracy. Among the evaluated emotions, the surprise class achieved the highest F1-score of 0.922, followed by calm at 0.897 and angry at 0.886. These results indicate that these emotional states produce distinctive acoustic characteristics that can be effectively captured through MFCC feature representation. Emotional speech expressing surprise and anger often involves noticeable variations in pitch dynamics, vocal intensity, and spectral energy distribution, which contribute to their distinguishable acoustic signatures.

Table 1. Classification Performance of the Random Forest Model for Each Emotion Class

Emotion	Recall	Precision	F1-Score	Support
Angry	0.917	0.857	0.886	1158
Fear	0.779	0.877	0.825	1158
Calm	0.910	0.884	0.897	134
Disgust	0.788	0.859	0.822	1158
Happy	0.813	0.824	0.819	1158
Sad	0.883	0.805	0.842	1158
Neutral	0.863	0.821	0.841	1040
Surprise	0.899	0.947	0.922	218
Accuracy			0.843	
Macro Average	0.857	0.859	0.857	7182
Weighted Average	0.843	0.845	0.843	7182

However, several emotional categories demonstrate slightly lower classification performance. The fear, disgust, and happy classes achieved F1-scores around 0.82, indicating that these emotional states are relatively more difficult for the model to distinguish. This phenomenon is commonly observed in SER research because certain emotions share similar acoustic characteristics. For instance, fear and sadness may exhibit comparable speech patterns, such as lower speech energy and similar pitch contours, which can lead to classification ambiguity.

A profounder understanding of the model's classification behavior can be obtained by analyzing the confusion matrix shown in Figure 3. The confusion matrix provides a detailed representation of how predicted emotional categories correspond to the actual labels in the dataset. The strong diagonal pattern observed in the matrix indicates that most speech samples are correctly classified. Several emotional classes demonstrate a high number of correct predictions, confirming that the model successfully captures relevant acoustic features associated with emotional expression.

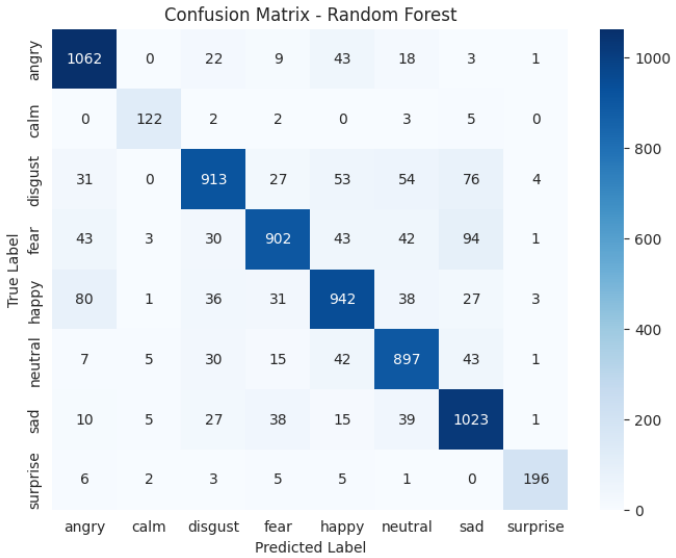

Figure 3. Confusion Matrix for Speech Emotion Classification using MFCC Acoustic Features and the Random Forest Model

Nevertheless, the confusion matrix also reveals several patterns of misclassification between emotionally similar categories. Several fear samples are classified as sad, while several disgust samples are also predicted as sad. This pattern suggests that these emotional states may share overlapping acoustic characteristics such as similar pitch variation and vocal energy distribution. In addition, some happy samples are misclassified as angry, possibly due to similarities in expressive speech patterns, such as higher speech intensity and dynamic pitch changes.

In addition to evaluating classification performance, it is important to analyze how individual acoustic features contribute to the classification model's decision-making method. To investigate this aspect, feature importance was evaluated using the Random Forest algorithm. The most influential MFCC features identified by the model are summarized in Table 2, which lists the ten features with the highest importance scores.

Table 2. Top 10 Most Important MFCC Features Identified by the Random Forest Model

Rank	MFCC Feature	Importance Score
1	MFCC 0	0.0488
2	MFCC 2	0.0345
3	MFCC 5	0.0291
4	MFCC 1	0.0271
5	MFCC 3	0.0256
6	MFCC 7	0.0217
7	MFCC 4	0.0216
8	MFCC 8	0.0211
9	MFCC 12	0.0195
10	MFCC 11	0.0193

The results show that MFCC 0 has the highest importance score among all features, indicating that the spectral energy component of speech signals plays a crucial role in emotion recognition. Emotional speech often modifies vocal intensity and overall energy distribution, and lower-order MFCC coefficients capture these changes. Other features such as MFCC 2, MFCC 5, and MFCC 1 also contribute significantly to the classification process, suggesting that multiple spectral components are involved in representing emotional information in speech signals.

The distribution of feature importance values is more clearly shown in the feature importance visualization in Figure 4. The plot illustrates that lower-order MFCC coefficients dominate the classification process, indicating that the overall spectral envelope of speech signals carries substantial emotional information. These coefficients capture broad spectral characteristics related to vocal tract configuration and energy distribution during speech production.

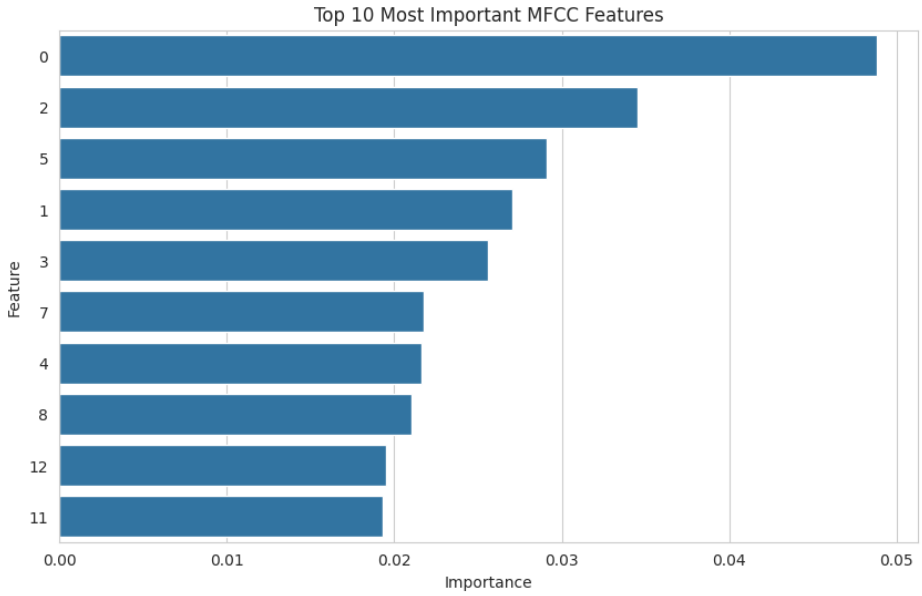

Figure 4. Visualization of the Top Ten Most Influential MFCC Features Identified by the Random Forest Classifier

The dominance of lower-order MFCC coefficients is consistent with established findings in speech signal processing research. Emotional speech typically alters vocal tension, pitch variation, and spectral energy distribution, which are reflected in these coefficients. The presence of several MFCC coefficients among the most influential features indicates that emotional information is distributed across multiple spectral components rather than being represented by a single acoustic parameter.

The experimental results demonstrate that MFCC acoustic features combined with the Random Forest classifier provide an effective baseline approach for speech emotion recognition. The model achieves relatively strong classification performance while also offering interpretable insights into feature contributions and classification behavior. These findings confirm the effectiveness of MFCC features in capturing emotional characteristics in speech signals and highlight the continued relevance of traditional machine learning approaches in SER research.

D. Conclusion

This study investigated the effectiveness of MFCC acoustic features combined with the Random Forest algorithm for SER. The experimental results demonstrate that the proposed approach effectively identifies emotional states from speech signals. The model achieved an overall classification accuracy of 84.33%, with a macro-average F1-score of 0.856, indicating relatively stable performance across the eight evaluated emotional categories. These findings confirm that MFCC features provide meaningful spectral representations that enable machine learning models to capture emotional characteristics embedded in speech signals.

Further analysis of the classification results revealed that certain emotional categories, such as surprise, calm, and anger, were recognized with higher accuracy than others. This suggests that these emotional expressions produce distinctive acoustic patterns that are easier for the model to identify. Meanwhile, emotions such as fear, disgust, and happiness showed slightly lower classification performance due to their similar acoustic characteristics. The

confusion patterns observed in the classification results highlight the inherent challenges of distinguishing emotionally similar speech signals using spectral features alone.

In addition, the feature-importance analysis indicates that lower-order MFCC coefficients, particularly MFCC 0, play a significant role in classification. These coefficients capture the overall spectral envelope of speech signals, which reflects the energy distribution and vocal-tract configuration during speech production. The experimental findings suggest that combining MFCC features with traditional machine learning methods, such as Random Forest, can provide an effective baseline for SER tasks. Future research may explore integrating additional acoustic features or deep learning architectures to improve classification performance further and address the challenges of distinguishing emotionally similar speech patterns.

Acknowledgment: The authors would like to express their sincere gratitude to Prof. Harun Joko Prayitno, Rector of Universitas Muhammadiyah Surakarta, for his continuous support in encouraging academic research and scholarly publication. The authors also extend their appreciation to Prof. Abdul Fadlil from the Informatics Doctoral Program at Universitas Ahmad Dahlan for delivering valuable lectures in the *Intelligent Pattern Recognition* course, which inspired the development of this research and the preparation of this article.

Daftar Pustaka

[1] B. Raviteja, N. Adithi, P. S. Jashwanth, and L. Sunitha, "Affective Speech Processing Using MFCC," in *2025 International Conference on New Trends in Computing Sciences, ICTCS 2025*, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 239–244. doi: 10.1109/ICTCS65341.2025.10989400.

[2] K. Zahra Nurbana and E. Sudarmilah, "Non-fungible token modeling: the enthusiasm of music fans for the digital-collectible revolution," *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 4, pp. 2880–2888, Aug. 2025, doi: 10.11591/eei.v14i4.9269.

[3] Y. I. Kurniawan, F. Razi, N. Nofiyati, B. Wijayanto, and M. L. Hidayat, "Naive Bayes modification for intrusion detection system classification with zero probability," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2751–2758, Oct. 2021, doi: 10.11591/eei.v10i5.2833.

[4] G. P. Chaitra and M. Farida Begam, "Real Time Dialectal Speech Recognition Using MFCCs on Mobile Applications," in *Proceedings of 2025 International Conference on Emerging Technologies in Computing and Communication, ETCC 2025*, Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/ETCC65847.2025.11108382.

[5] A. Al-Shidi, F. H. B. Nordin, I. I. Mohamed, M. A. Younis, E. E. Abusham, and I. E. Elmutasim, "Study of MFCC, Spectral Peaks, and Hashing for Real-Time Audio Fingerprinting in Quran Classification Systems," in *2024 5th International Conference on Innovative Computing, ICIC 2024 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICIC63915.2024.11116163.

[6] Y. Bai, "A Study on Speech Emotion Recognition Based on MFCC and KNN Models," in *2024 IEEE 2nd International Conference on Image Processing and Computer Applications, ICIPCA 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 729–732. doi: 10.1109/ICIPCA61593.2024.10709039.

[7] A. G. Putrada, N. Alamsyah, S. F. Pane, M. N. Fauzan, and D. Perdana, "Virtual Sensors Method and Architecture for a Smart Home Environment with Random

Forest," in *2023 10th International Conference on ICT for Smart Society (ICISS)*, IEEE, Sep. 2023, pp. 1–6. doi: 10.1109/ICISS59129.2023.10292065.

[8] A. G. Putrada, N. Alamsyah, M. N. Fauzan, and D. Perdana, "PCA-SVM for a Lightweight ASL Hand Gesture Image Recognition," in *Proceedings of the International Conference on Electrical Engineering and Informatics*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICEEI59426.2023.10346744.

[9] M. Tummala, L. Harish, M. E. Malkhed, S. S. Kumar, N. Neelima, and V. Venugopal, "Optimizing Gender Identification with MFCC Feature Engineering and Enhanced Gradient Boosting," in *2024 Asian Conference on Intelligent Technologies, ACOIT 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ACOIT62457.2024.10939536.

[10] S.-H. Chen, W.-T. Huang, C.-H. Lai, Y.-L. Lin, and M.-H. Su, "Analysis and Discussion of Feature Extraction Technology for Musical Genre Classification," in *2024 27th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, IEEE, Oct. 2024, pp. 1–4. doi: 10.1109/O-COCOSDA64382.2024.10800024.

[11] G. Ru, X. Zhang, J. Wang, N. Cheng, and J. Xiao, "Improving Music Genre Classification from multi-modal Properties of Music and Genre Correlations Perspective," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Jun. 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10097241.

[12] C. Ahmadi, S. H. Wang, S. P. Chiu, and J. L. Chen, "Dual Acoustic Feature Fusion for Enhanced Audio Deepfake Detection Using VGG-16 Architecture: Mitigating Speech Tampering with MFCC and ELTP," in *Proceedings - 2024 RIVF International Conference on Computing and Communication Technologies, RIVF 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 216–220. doi: 10.1109/RIVF64335.2024.11009070.

[13] T. Lu, "A Study on the Effectiveness of AI-Assisted Teaching Tools for English Speaking Under the Assistance of AI Speech Evaluation Technology," in *Learning and Analytics in Intelligent Systems*, 2025, pp. 216–227. doi: 10.1007/978-3-031-95252-4_20.

[14] D. P. Major and M. Chatterjee, "Acoustic analyses of the RAVDESS corpus of emotional stimuli," *JASA Express Lett.*, vol. 6, no. 2, Feb. 2026, doi: 10.1121/10.0042364.

[15] Z. S. Kahhoul, N. Terki, M. L. Tiar, I. Benaissa, and S. Boutiba, "Ensemble Learning for Improved Speech Emotion Recognition: A Control Dimension Analysis of Log-Mel Spectrograms from the CREMA-D Dataset," *Signal Image Video Process.*, vol. 19, no. 13, p. 1135, Dec. 2025, doi: 10.1007/s11760-025-04763-8.

[16] A. Benzirar, M. Hamidi, and M. Filali Bouami, "Building a speech emotion recognition system using RNN, GRU and LSTM," *Int. J. Speech Technol.*, vol. 28, no. 3, pp. 745–759, Sep. 2025, doi: 10.1007/s10772-025-10214-z.

[17] S. Nath, A. K. Shahi, T. Martin, N. Choudhury, and R. Mandal, "Speech Emotion Recognition Using Machine Learning: A Comparative Analysis," *SN Comput. Sci.*, vol. 5, no. 4, p. 390, Apr. 2024, doi: 10.1007/s42979-024-02656-0.

[18] R. Nutenki, A. Thatipudi, A. K. Perikala, and H. Medida, "Analysis and Classification of Arcing Signals by Using MFCC," in *2024 4th International Conference on*

Advances in Electrical, Computing, Communication and Sustainable Technologies, ICAECT 2024, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICAECT60202.2024.10468692.

[19] V. Rraci, L. Anderson, and S. Chakrabarty, "Classification of Right Hemisphere Damage Using MFCC Paralinguistic Voice Features," in *2025 Intermountain Engineering, Technology and Computing, IETC 2025*, Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/IETC64455.2025.11039385.

[20] M. Sivaramakrishnan, A. Rajput, and M. Saravanan, "Classification of Deep Fake Audio Using MFCC Technique," in *2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems, ICITEICS 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICITEICS61368.2024.10625305.

[21] W. Liu and Y. Duan, "Speaker recognition based on MFCC and GFCC feature parameter extraction," in *2025 5th International Symposium on Computer Technology and Information Science, ISCTIS 2025*, Institute of Electrical and Electronics Engineers Inc., 2025, pp. 33–37. doi: 10.1109/ISCTIS65944.2025.11065212.