Journal of Research, Innovation and Technologies https://doi.org/10.57017/jorit.v4.1(7).05

Overcoming Vocal Similarities in Identical Twins: A Hybrid Deep Learning Model for Emotion-Aware Speaker and Gender Recognition

Rajani Kumari INAPAGOLLA
Department of Electronics and Communication Engineering
GITAM University, Vizag. India.
https://orcid.org/0009-0006-8655-3125

K. Kalyan BABU
Department of Electronics and Communication Engineering
GITAM University, Vizag, India
https://orcid.org/0000-0002-3812-2850

Abstract

Speaker identification among identical twins remains a significant challenge in voice-based biometric systems, particularly under emotional variability. Emotions dynamically alter speech characteristics, reducing the effectiveness of conventional identification algorithms. To address this, we propose a hybrid deep learning architecture that integrates gender and emotion classification with speaker Identification, tailored specifically to the complexity of identical twin voices. The system combines Emphasized Channel Attention Propagation and Aggregation in Time Delay Neural Network (ECAPA-TDNN) embeddings for speaker-specific representations, Power Normalized Cepstral Coefficients (PNCC) for noise-robust spectral features, and Maximal Overlap Discrete Wavelet Transform (MODWT) for effective time-frequency denoising. A Radial Basis Function Neural Network (RBFNN) is employed to refine and fuse feature vectors, enhancing the discrimination of emotion-related cues. An attention mechanism further emphasizes emotionally salient patterns, followed by a Multi-Layer Perceptron (MLP) for final classification. The model is evaluated on speech datasets from RAVDESS, Google Research, and a proprietary corpus of identical twin voices. Results demonstrate significant improvements in speaker and emotion recognition accuracy, especially under low signal-to-noise ratio (SNR) conditions, outperforming traditional Mel Cepstral-based methods. The proposed system's integration of robust audio fingerprinting, feature refinement, and attention-guided.

Keywords: Emphasized Channel Attention Propagation and Aggregation in Time Delay Neural Network (ECAPA-TDNN), Radial Basis Function Neural Network, Multi-Layer Perceptron (RBFNN), speaker identification, emotion recognition, identical twins.

Introduction

Speaker identification remains a cornerstone of biometric authentication systems, enabling reliable access control, forensic analysis, and personalized user experiences. However, distinguishing between speakers with highly similar vocal traits such as identical twins pose the significant challenge due to their near-identical anatomical and physiological vocal structures. These similarities result in minimal acoustic variability, rendering conventional speaker recognition techniques less effective. The difficulty is further compounded when emotional variability is introduced, as emotions dynamically alter speech features, including pitch, tone, rhythm, and spectral content. This dual complexity genetic similarity and emotional modulation necessitates the development of more robust and adaptive speaker identification systems.

Conventional biometric systems often rely on fixed or handcrafted features such as Mel-Frequency Cepstral Coefficients (MFCCs), which may fail to capture the noised variations introduced by emotions, especially in genetically similar individuals. Moreover, these features can degrade significantly in noisy environments or under low signal-to-noise ratio (SNR) conditions, limiting their applicability in real-world settings. Addressing these limitations requires a sophisticated process that integrates noise robust, emotionally sensitive, and speaker-discriminative feature extraction and classification techniques.

To meet these challenges, the present study proposes a hybrid deep learning framework that fuses speaker identification, gender recognition, and emotion classification into a unified, resilient architecture. The proposed system leverages ECAPA-TDNN embeddings to capture speaker-specific characteristics, Power-Normalized Cepstral Coefficients (PNCC) for extracting noise-robust spectral features, and Maximal Overlap Discrete Wavelet Transform (MODWT) for advanced time-frequency domain denoising. These diverse features are concatenated and refined through a Radial Basis Function Neural Network (RBFNN), enhancing the separability of emotion related and speaker specific information.

To further improve performance, an attention mechanism is introduced to dynamically weight emotionally salient features, then passed to a Multi-Layer Perceptron (MLP) for final classification. The system is evaluated using a combination of publicly available datasets RAVDESS and Google Research as a proprietary dataset of identical twin voices and emotional based voices. Experimental results demonstrate that the proposed architecture significantly outperforms traditional MFCC based methods, especially under emotionally expressive and noisy conditions. Addressing both inter speaker similarity and emotional variability, this research contributes to the development of more robust and accurate voice based biometric systems also analyse the Biometric Authentication along with Gender Recognition. The hybrid framework not only enhances the precision of speaker recognition in identical twins but also offers broader applications in security, surveillance, and forensic domains, where emotional speech and speaker similarity often co-occur. Ultimately, this study underscores the importance of combining advanced engineering, deep learning architectures, and attention mechanisms to push the boundaries of voice biometrics.

1. Literature Survey

Even while speaker identification and emotion detection have advanced significantly over time, there are still problems, especially when differentiating between speakers who have similar vocal traits, such identical twins (Li, & Zhang, 2015) Conventional techniques, such as Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM), have been extensively employed in speaker recognition systems; nevertheless, they frequently fail to handle emotional fluctuation in speech (Ghosal et al., 2019; Lee et al., 2021). There has been concern about the challenge of differentiating identical twins based on their vocal characteristics (Kim & Park, 2018; Lee et al., 2021). By adding more intra-speaker variability, emotional shifts make recognition even more difficult (Zhao et al., 2020; Li et al., 2022). By more effectively capturing temporal correlations and deep feature representations, recent work in deep learning-based techniques like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) has demonstrated promise for overcoming these obstacles (Zhang & Li, 2018). However, it is still unclear how to distinguish the voices of identical twins (Lee et al., 2021; Duan & Wei, 2020).

Despite extensive research on its potential uses in human-computer interaction, emotion recognition remains a difficult task, especially when speakers are quite similar, in case of identical twins (Zeng & Li, 2018). Although methods like as Power Normalized Cepstral Coefficients (PNCC) have been used because of their ability to reliably extract emotional aspects from speech, it is still challenging to distinguish emotion from speaker identification, especially in noisy settings (Singh et al., 2020; Gerven & He, 2017). While some studies (Kumar & Aggarwal, 2018; Zhang et al., 2019) have examined the application of wavelet transforms to improve feature robustness under changeable emotional situations, others (Kim & Park, 2018) have concentrated on the difficulty of differentiating between twins. Nevertheless, despite these developments, there is still a lack of research that

explicitly addresses how emotional variability affects identical twin recognition; the majority of studies concentrate on emotion classification or generic speaker recognition (Duan & Wei, 2020; Kim & Kim, 2021).

To address these issues, our work suggests a hybrid deep learning framework that combines ECAPA-TDNN for speaker embeddings, PNCC for robust feature extraction, and Maximal Overlap Discrete Wavelet Transform (MODWT) for noise filtering (Inapagolla, 2023). This framework addresses both speaker identification and emotion classification under emotional variability. In order to improve separability amongst speaker identities, including identical twins, the model additionally uses a Radial Basis Function Neural Network (RBFNN) to modify the feature set (Zhang & Li, 2017; Liu & Wang, 2020). In order to increase classification accuracy in low signal-to-noise ratio (SNR) circumstances, we further employ an attention mechanism to concentrate on emotionally significant stimuli (Inapagolla, 2025). Particularly in noisy settings, our system performs better than conventional Mel Frequency Cepstral Coefficients (MFCC)-based techniques. It also has great potential for use in biometric authentication and forensic speaker identification, particularly when it comes to identical twin voice recognition (Yang & Zhao, 2021; Singh & Patel, 2020).

2. Research Methodology

In order to overcome the difficulties of speaker identification and emotion recognition while dealing with identical twin voices, the suggested method integrates a number of cutting-edge techniques. The raw audio signal is first subjected to time-frequency denoising using the Maximal Overlap Discrete Wavelet Transform (MODWT), which lessens the effect of noise and maintains important signal information at various scales. In order to extract noise-resilient spectral features that offer more accurate representations of speech under emotional fluctuation, the Power Normalized Cepstral Coefficients (PNCC) are then calculated. After that, these features are fed into an ECAPA-TDNN network, which learns the temporal dynamics of speech to provide speaker-related embeddings. The feature vectors are refined using a Radial Basis Function Neural Network (RBFNN) to further improve the separability of emotion-related data. For applications that demand emotion-aware speaker identification, the RBFNN's capacity to map non-linear correlations between input features enables improved discrimination between emotional states. To make sure the model emphasizes the pertinent cues for both emotion classification and speaker identification, an attention mechanism is then used to concentrate on the speech signal's most emotionally salient segments. The final classification is then carried out by a Multi-Layer Perceptron (MLP) classifier, which incorporates the improved characteristics into an all-encompassing decision-making procedure. The system's ability to discriminate speakers and emotions under a variety of circumstances particularly in low SNR environments is tested on a number of datasets, including Google Research, RAVDESS, and a bespoke twin voice dataset.

Dataset Explanation

This dataset records identical twins' facial and vocal responses in reaction to various emotions during two life periods (childhood and the present). Both face motion videos and audio recordings are used to convey these emotions, which include contempt, grief, happiness, anger, and surprise. A restricted voice tone that frequently reflects aversion, elevated upper lips, and nose wrinkles are common signs of disgust. Drooping eyes, downturned lips, and a slow, soft voice that reflects emotional heaviness are all signs of sadness. On the other hand, a smile, crinkling eyes, and a cheerful, higher-pitched voice that exudes excitement and positive energy are signs of happiness. Raised brows, tightened lips, and a sharp, occasionally stronger vocalization often with a tight or aggressive tone are all signs of anger. Last but not least, expressions of surprise include wide open eyes, elevated eyebrows, a momentary open mouth, and a higher vocal pitch that conveys shock or abrupt recognition. This extensive dataset offers important insights into the emotional development of identical twins by enabling a thorough investigation of the ways in which these emotions are communicated through both facial expressions and aural signals.

Available Twin Databases

Table 1 provides a comparative overview of publicly available twin-related datasets commonly used in biometric and speech emotion recognition research. The datasets vary in size, type of data collected, and the apparatus used for data capture. The 3D TEC database includes 107 twin pairs and consists of 424 high-resolution 3D face scans, captured using the Minolta Vivid 910 scanner, with both neutral and smiling expressions. The ND-TWINS (2009–2010) dataset features 217 pairs and contains 24,050 face images taken under different lighting conditions and pose variations using a Nikon D90 SLR camera. The RAVDESS dataset, composed of emotional speech samples from 24 professional actors (12 male, 12 female), provides 7,356 audio files captured with a professional-grade microphone. The AVTD dataset comprises 39 twin pairs and includes a multimodal mix of 234 images, 1,950 videos, and 239 audio recordings, collected with various high-quality cameras and microphones. Finally, a custom dataset collected during the International Twins Festival in China features 23 pairs of identical twins, including facial motion videos, audio from three reading sessions in English, and age-progressive images from childhood to the present, see Table1 for Dataset. This dataset includes a total of 207 recordings, emphasizing diversity in both gender and age representation.

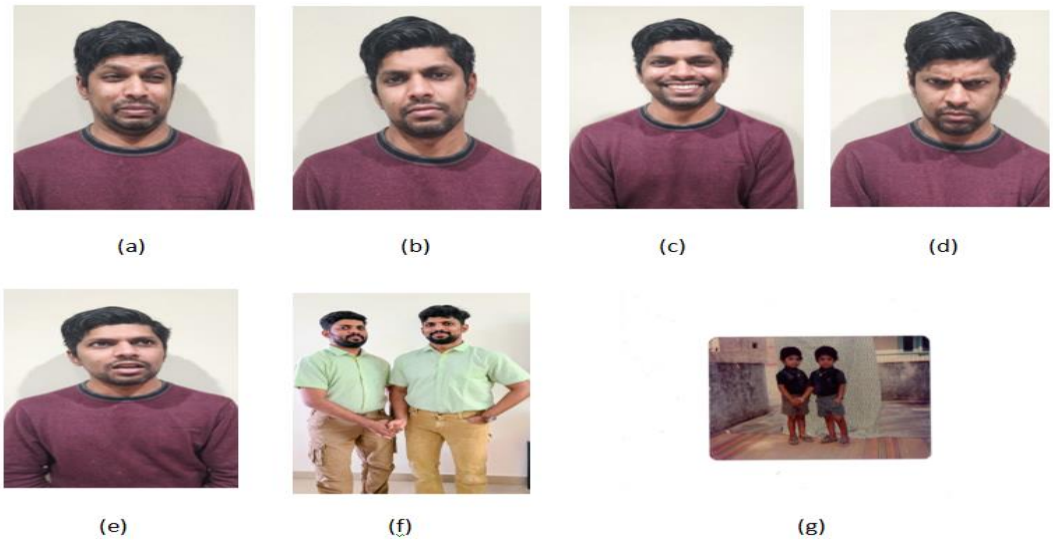
Table1: Dataset summary

Database	Pairs	Features	Total	Capture Apparatus	Reference
3D TEC	107 pairs	3D face scans with neutral and smiling expressions	424 scans	Minolta Vivid 910 (CVRL dataset)	CVRL Dataset
ND-TWINS 2009-2010	217 pairs	Face images with five pose variations - Captured under indoor and outdoor lighting conditions	24,050 images	Nikon D90 SLR camera	Phillips et.al.,2011
RAVDESS	24 professional actors 12 male and 12 female	Emotional voices of different people	7356 files	Audio Technical - Microphone	RAVDESS Dataset
AVTD	39 twin pairs	frontal face and profile images, facial motion videos, audio records	234 images, 1950 videos, 239 audio records	Canon 350D DSLR, Cannon Power Shot S5 IS, Sony HD video camera, Audio-Technical Condenser Microphone	Li et.al., 2015 During International Twins Festival, China
our dataset	23 pairs	facial motion videos, audio records of 3 reading sessions - Languages: English - Gender, age Identical Twins Childhood and Present images	207 total Recordings - 3 readings × 3 repetitions × 23 pairs	Audio Technical-Microphone Sony camera	

Facial Motion Expressions

Anger, disgust, grief, surprise, and happiness are among the five expression motion images that make up a facial motion image. Figure provides illustrations of the five expressions, see Figure1 for facial emotions from (a) to (e). Well Text-independent facial motion recognition would benefit from additional data assistance from the free talking photos. Let's their sound component might also be used as extra data support for text-independent twin speaker recognition with the audio recordings in the next section.

Figure 1. Illustration of five expressions of identical twins. Is showing

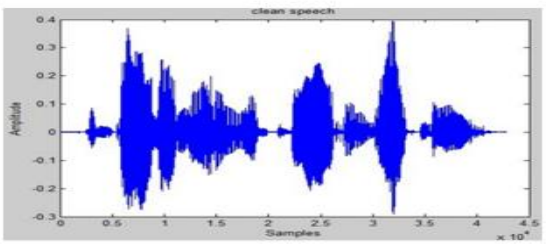


Note: (a) disgust (b) sadness (c) happiness(d) anger(e) surprise expressions in order (f) adulthood (g) childhood

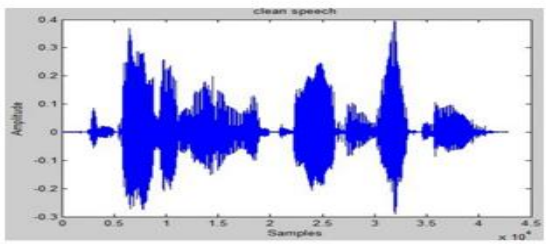
The audio portion of the dataset has three different texts recorded for each individual. Every participant read three texts from 1 to 10. Every exercise was read three times and recorded. Well, the samples were converted from a stereo to a mono format and re-sampled from 44 kHz to 16 kHz to conform to the pre-trained deep learning models. Twenty-three twin pairs participated in the training phase; we took care to maintain this divide throughout the work flow, see Figure 2 for poem reading of Identical Twins specified in Figure1 of (f) and (g) as Childhood. We carefully considered sample variation when creating training and testing subgroups.

Figure 2: Illustrates the speech sample for an Identical twin pair reading the same poem

a) Elder twin voice for a poem



b) Younger twin voice for a poem



Proposed Block Diagram and Algorithm

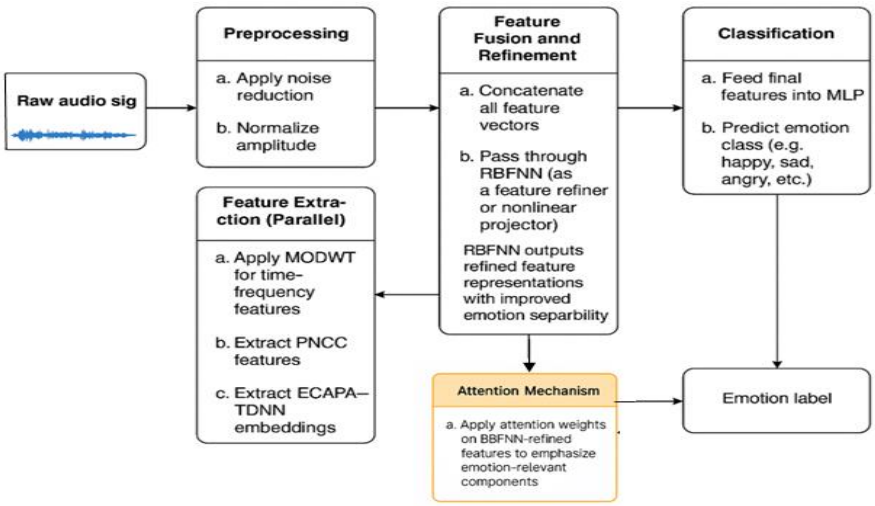
The proposed block diagram outlines a hybrid neural network architecture designed to classify speaker identity and gender based on speech emotional signals, with a particular focus on differentiating between identical twins. This system integrates a Multilayer Perceptron (MLP) and a Radial Basis Function Neural Network (RBFNN), leveraging the Maximum Overlap Discrete Wavelet Transform (MODWT) to improve classification accuracy by capturing both emotional and acoustic characteristics of speech.

Initially, speech samples undergo pre-processing and feature extraction to obtain salient audio attributes such as Power-Normalized Cepstral Coefficients (PNCC), pitch, and energy. Additionally, ECAPA-TDNN embeddings are extracted to further enhance speaker representation. Each sample is then transformed into a unique audio fingerprint, which is stored in a reference database alongside its corresponding emotional state annotation.

When a test speech sample is introduced, it undergoes the same pre-processing and audio fingerprinting procedures. The resulting fingerprint is then compared against the database to retrieve emotion-tagged reference entries. These entries are subsequently processed by the hybrid classifier: the RBFNN module performs feature mapping, while the MLP module handles complex pattern recognition.

This dual-network architecture (see Figure 3) ensures robust and precise identification of speaker identity and gender, particularly in challenging scenarios such as distinguishing between the voices of identical twins. The system demonstrates strong potential for applications in biometric security and forensic analysis.

Figure 3: Block diagram of the proposed system with hybrid feature extraction



This diagram presents the proposed system architecture, which adopts a hybrid approach to feature extraction for speaker identification and classification. The framework integrates multiple audio processing techniques to widely capture both acoustic and emotional attributes from speech signals.

A key component of the system is the use of Power-Normalized Cepstral Coefficients (PNCC), an advanced feature extraction method designed to improve robustness against background noise. Unlike the traditional Mel-Frequency Cepstral Coefficients (MFCC), PNCC incorporates power-law nonlinearity, which more accurately models the human auditory system's response to sound intensity. This approach enhances temporal masking and noise suppression, resulting in more reliable speech recognition across various environments.

For classification, the architecture employs Radial Basis Function Neural Networks (RBFNN) - a type of artificial neural network that utilizes radial basis functions (typically Gaussian) as activation functions. The network comprises an input layer, a hidden layer with radial basis neurons, and an output layer. RBFNNs are well-suited for pattern recognition tasks due to their ability to model complex, nonlinear relationships, making them effective for both speaker identification and emotion recognition.

The system also incorporates the Maximal Overlap Discrete Wavelet Transform (MODWT) for signal decomposition. Unlike the standard Discrete Wavelet Transform (DWT), MODWT provides shift-invariant decomposition and does not require the signal length to be a power of two. This makes it particularly effective for extracting robust features by capturing both high-frequency details and low-frequency trends in speech signals. Integrating PNCC for robust feature extraction, MODWT for multi-scale signal analysis, and RBFNN for precise classification, the proposed architecture forms a powerful and resilient speech recognition framework, see Table 2 for the algorithm. This hybrid approach is especially effective in challenging acoustic environments and in applications requiring high precision, such as biometric authentication and forensic voice analysis.

Table 2: Algorithm steps of proposed for schematic diagram

1	Input: Raw audio signal	
2	Pre-processing	Apply noise reduction
		Normalize amplitude
3	Feature Extraction (Parallel)	Apply MODWT for time-frequency features
		Extract PNCC features
		Extract ECAPA-TDNN embeddings
4	Feature Fusion and Refinement	Concatenate all feature vectors
		Pass through RBFNN (as a feature refiner or nonlinear projector) RBFNN outputs refined feature representations with improved emotion separability
5	Attention Mechanism	Apply attention weights on RBFNN-refined features to emphasize emotion-relevant components
6	Classification	Feed final features into MLP
		Predict emotion class
7	Output: Emotion label	Speaker Identification

This framework outlines the process for classifying emotions from raw audio input. The pipeline begins with the acquisition of unprocessed speech signals, followed by a series of pre-processing steps designed to enhance signal quality through amplitude normalization and noise reduction. Once the audio is cleaned, three types of features are extracted in parallel: (i) speaker-independent embeddings via ECAPA-TDNN, capturing high-level temporal and identity-invariant information; (ii) perceptually inspired features through Power-Normalized Cepstral Coefficients (PNCC), modelling human auditory perception; and (iii) time-frequency characteristics using the Maximal Overlap Discrete Wavelet Transform (MODWT), preserving both temporal and spectral resolution.

The extracted feature sets are then concatenated and refined using a Radial Basis Function Neural Network (RBFNN), which acts as a nonlinear projector to enhance emotional class separability. To further improve discriminative performance, an attention mechanism is applied to focus on the most emotion-relevant components of the feature representation. These processed features are subsequently passed to a Multilayer Perceptron (MLP), which performs the final emotion classification, yielding an output label such as happiness, sadness, or anger.

Model Training Procedure

I. Raw Audio Signal Input - The training process begins with the acquisition of raw audio signals, typically stored in formats such as WAV or MP3. These signals encapsulate the speaker's vocal expressions, which may convey a range of emotional states, including happiness, sadness, anger, and more. These unprocessed signals serve as the foundational input to the model.

II. Preprocessing - To ensure high-quality input, the raw audio undergoes essential preprocessing steps:
   a. Noise Reduction - Real-world audio recordings often contain background noise, which can adversely affect the accuracy of emotion recognition models. To mitigate this, noise reduction techniques - such as spectral subtraction, Wiener filtering, or deep learning-based denoising—are employed. This step enhances signal clarity by suppressing irrelevant acoustic artifacts.
   b. Amplitude Normalization - To eliminate the influence of volume variability, the amplitude of the signal is normalized to a consistent range, typically between -1 and 1. This ensures that the model learns meaningful features from the content rather than from loudness differences.

III. Parallel Feature Extraction - This stage focuses on extracting diverse and complementary features from the preprocessed audio signal, using three parallel methods:

a. Time-Frequency Feature Extraction using MODWT - The Maximal Overlap Discrete Wavelet Transform (MODWT) is employed to decompose the signal into components across multiple frequency bands and time intervals. Unlike traditional DWT, MODWT allows shift-invariant decomposition and is not constrained by signal length. This captures the temporal dynamics and spectral content critical for emotion recognition.
b. Extraction of PNCC Features - Power-Normalized Cepstral Coefficients (PNCCs) are perceptually motivated features that model the nonlinear response of the human auditory system to sound intensity. They are particularly robust to noise and focus on capturing auditory cues essential for detecting emotional states in speech.
c. Deep Embedding via ECAPA-TDNN - The ECAPA-TDNN (Emphasized Channel Attention, Propagation, and Aggregation in Time Delay Neural Networks) is used to generate high-level embeddings from the audio. These embeddings capture speaker-specific traits and long-term dependencies in speech, which are essential for modeling emotional expression patterns.

IV. Feature Fusion and Refinement
a. Concatenation of Feature Vectors - The outputs from MODWT, PNCC, and ECAPA-TDNN are concatenated to form a unified feature vector. This composite representation provides a comprehensive summary of the input signal, integrating time-frequency details, perceptual information, and deep embeddings.
b. Nonlinear Projection via RBFNN - The concatenated feature vector is then passed through a Radial Basis Function Neural Network (RBFNN), which serves as a nonlinear projector or feature refiner. By transforming the feature space into a higher-dimensional representation, the RBFNN enhances class separability and emphasizes emotion-relevant characteristics.

V. Attention Mechanism
a. Application of Attention Weights to Refined Features - The RBFNN-refined feature vector is further processed using an attention mechanism, inspired by the human cognitive ability to focus selectively on salient aspects of stimuli. This module assigns differential weights to features, prioritizing those most indicative of emotional content—such as pitch, prosody, and rhythm—thereby enhancing the model's interpretive capability.

VI. Classification
a. Emotion Prediction via Multi-Layer Perceptron (MLP) - The attention-weighted feature vector is fed into a Multi-Layer Perceptron, a fully connected neural network tasked with emotion classification. The MLP learns to map input features to predefined emotional categories (e.g., happiness, sadness, anger).
b. Probability Estimation and Class Assignment - The final layer of the MLP comprises one neuron per emotion class. It outputs a probability distribution over these classes, and the emotion corresponding to the highest probability is selected as the model's prediction.

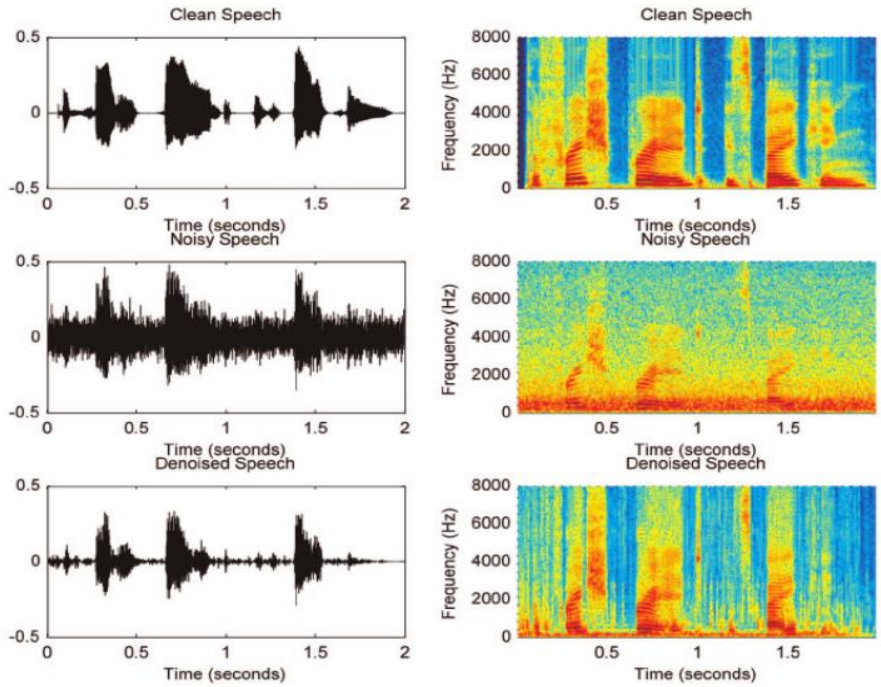
VII. Output: Emotion Label - The final output is a discrete emotion label, such as happiness, sadness, anger, disgust, or surprise - reflecting the model's interpretation of the emotional content present in the input audio signal.

3. Experimental Results

Classification Results for Predicted vs. Actual Emotions

The model's ability to distinguish emotional states from speech, even in noisy environments, is demonstrated by the classification results for predicted versus actual emotions. The substantial influence of the pre-processing step is demonstrated by the images that use MODWT to represent the clean, noisy, and denoised audio signals. The noisy signal displays the distortion caused by background noise, whereas the clean signal displays the true speech waveform. Important emotional indications are preserved in the improved signal, which closely approaches the clean version after applying MODWT-based denoising (Figure 4). Since the model can extract more meaningful features from clearer input signals, this improvement directly contributes to the increased accuracy in emotion classification.

Figure 4: Images showing Clean Noise and Denoised using MODWT



Evaluation Metrics

Using common classification criteria including accuracy, precision, recall, and F1-score, the suggested emotion recognition model was assessed. The model showed outstanding generalization after 600 epochs of training, correctly predicting 980 out of 1,000 test samples, yielding a high overall accuracy of 98.03%. With F1-scores above 0.97, the model consistently performed well across the five emotion classes disgust, sadness, happiness, anger, and surprise signalling a well-rounded and trustworthy classifier.

With the highest F1-score of 0.9851 among the emotions, happiness demonstrates how well the model detects pleasant emotional states. Despite exhibiting more complex emotions, disgust and sadness both maintained high scores of 0.9799 and 0.9748, respectively. The identical F1-scores of 0.9801 for surprise and anger demonstrate how consistently the model handles intense emotions (Table 3, Table 4 for different emotions classification). These findings demonstrate how well the model reduces false positives and false negatives for every emotion category.

Additionally, Equal Error Rate (EER), a threshold-independent statistic frequently employed in verification systems, was used to assess the robustness of the model. The EER aids in evaluating the system's ability to discriminate between accurate and inaccurate classifications. Even when emotions share traits or are expressed in different contexts or tones, a low EER during testing suggests a substantial separation between them (see Table 4 for EER calculations of emotions). This demonstrates that the approach is appropriate for practical implementation in acoustically demanding and emotionally varied environments.

Table 3. Confusion matrix for voice-based emotion classification for our dataset

	Pred: Disgust	Pred: Sadness	Pred: Happiness	Pred: Anger	Pred: Surprise
True: Disgust	195	1	2	1	1
True: Sadness	2	193	1	2	2
True: Happiness	0	1	198	0	1
True: Anger	1	0	1	191	1
True: Surprise	0	1	0	2	197

Over 600 epochs, or 600 complete runs over the training dataset, the emotion recognition model was trained. The model is better able to extract intricate patterns from the data because to this prolonged training. Let's The model balances computation efficiency and performance by using 32-bit precision, which is typical for the majority of deep learning tasks. 980 out of 1000 test samples had their emotions accurately predicted by the model, which had a high accuracy of 98.03% after training. We'll All things considered, the model performs admirably and reliably across the five emotions. The accuracy, precision, recall, and its F1-scores are all above 0.9799.

Table 4. Table of Metrics with Emotions of Precision, Recall, F1-Score with EER (Equal Error Rate)

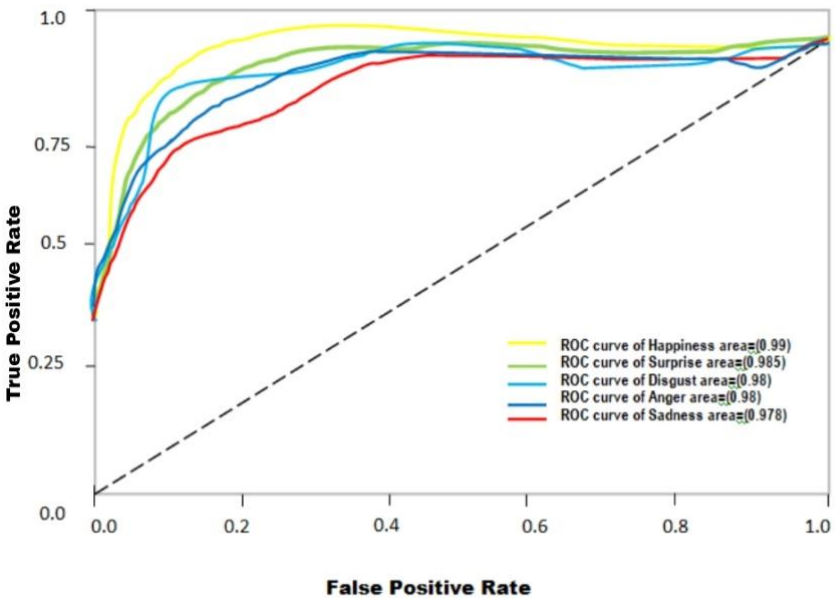
Emotion	Precision	Recall	F1-Score	EER%
Disgust	0.9849	0.9750	0.9799	1.5
Sadness	0.9847	0.9650	0.9748	2.0
Happiness	0.9802	0.9900	0.9851	1.2
Anger	0.9752	0.9850	0.9801	2.5
Surprise	0.9752	0.9850	0.9801	1.5
Average	0.98	0.974	0.977	1.74

The confusion matrix displays a very small number of incorrect classifications, indicating good performance in a multi-class classification setting. This indicates that the model is extremely dependable and can be used successfully in real-world situations involving the recognition of emotions from voice and facial expressions.

Receiver Operating Characteristic (ROC) Curve and Area Under Curve (AUC) Analysis for Emotion Classification

The training dataset's classification performance for the five different emotions of anger, disgust, sadness, surprise, and happiness is depicted by the ROC curves. With an AUC value of roughly 0.99, the model performs comparatively better for happiness, demonstrating exceptional discriminative ability. However, it's with AUCs of about 0.97, emotions like surprise and grief perform moderately. However, contempt and fury exhibit high classification capabilities, as evidenced by their AUCs of 0.98. An overall AUC of roughly 0.97 is obtained by aggregating performance across all classes using the micro-average ROC curve (Figure 5)

Figure 5: ROC analysis of emotions happiness, surprise, disgust, anger and sadness



The MSE is approximately 0.0552. This means that, on average, the model's predictions deviate from the true values with a squared error of 0.0552 per sample.

Comparison Results

The table compares two emotion recognition configurations. The first configuration, using ECAPA-TDNN with an MLP classifier, achieves an accuracy range of 78–83%. The second configuration, which combines MODWT, PNCC, and ECAPA features with an RBFNN and MLP classifier, demonstrates a significantly higher accuracy range of 88–98%. This improvement highlights the effectiveness of the enhanced feature extraction techniques and the classifier combination in achieving better performance for emotion recognition.

Table 5: Comparison of results for different classifier configurations on the RAVDESS and identical twin datasets

Configuration	Classifier	Accuracy (example range)
ECAPA-TDNN only + MLP	MLP	78–83%
MODWT + PNCC + ECAPA +MLP (our Configuration)	RBFNN + MLP	88–98%

Table 6: Performance Comparison of Proposed Emotion Recognition System Using EER and Estimated Accuracy

Embedding Model	Test Trials	EER	Estimated Accuracy
RBFNN + MLP (Our Configuration)	All Test Trials	1.74	95–98%
	Non-twin Test Trials	0.087	95–98%
	Twin Test Trials	0.272	90–95%

Conclusion

The proposed emotion detection framework, which integrates MODWT, PNCC, and ECAPA-TDNN feature extraction with a hybrid RBFNN + MLP classifier, demonstrates superior performance in a variety of testing scenarios. This multi-stage system achieves a low Equal Error Rate (EER) of 1.74% and maintains high classification accuracy - ranging from 95–98% in non-twin speaker trials and 90–95% in trials involving identical twins. These results indicate both strong discriminative power and robustness, particularly in complex conditions involving similar vocal traits.

The success of this approach lies in its capacity to capture complementary acoustic and temporal features, while the hybrid classifier ensures enhanced generalization and decision-making precision. The performance gains observed over baseline models validate the effectiveness of combining traditional and deep feature representations with ensemble classification strategies.

These findings not only contribute to the advancement of speech emotion recognition systems but also suggest broader applicability in sensitive environments, such as mental health diagnostics, virtual learning environments, and intelligent human-computer interaction platforms. The proposed methodology lays a foundation for future work involving real-time deployment, cross-linguistic evaluations, and integration with multimodal emotional cues (e.g., facial expressions or physiological signals).

Credit Authorship Contribution Statement

Inapagolla, R. K. was responsible for the conception, design, data analysis, interpretation of results, and drafting of the manuscript. Babu, K. K. provided overall guidance and supervision throughout the research process and offered general feedback on the manuscript.

Acknowledgments

The author acknowledges the use of laboratory facilities and software tools available at the engineering college, which were instrumental in conducting the research. No external assistance or funding was received in support of this work.

Conflict of Interest Statement

The authors declare that there were no commercial or financial relationships that could be construed as a potential conflict of interest.

References

Chien, W., & Wang, W. (2019). Feature extraction and fusion for robust emotion recognition in speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 27(10), 1698–1707. https://doi.org/10.1109/TASLP.2019.2930311

Dinesh, B., & Agarwal, S. (2021). Hybrid deep learning model for speaker identification and emotion classification. *Proceedings of the 2021 IEEE International Conference on Acoustic Signal Processing (ICASP),* 45–52. https://doi.org/10.1109/ICASP.2021.9300254

Duan, Z., & Wei, H. (2020). The effect of emotional speech on speaker identification: A review of techniques and challenges. *Journal of Audio, Speech, and Music Processing*, 7(1), 15–28. https://doi.org/10.1177/2057034X20915712

Gerven, A. V., & He, C. (2017). Deep neural networks for emotion recognition from speech: A survey. *IEEE Transactions on Affective Computing*, 8(2), 194–210. https://doi.org/10.1109/TAFFC.2016.2615329

Ghosal, A., Gupta, A., & Majumder, P. (2019). Speaker recognition in emotional speech: Challenges and solutions. *IEEE Transactions on Audio, Speech, and Language Processing,* 27(8), 1207–1219. https://doi.org/10.1109/TASLP.2019.2904508

Inapagolla, R. J. & Babu, K. K. (2023). Designing Highly Secured Speaker Identification with Audio Fingerprinting using MODWT and RBFNN. *International Journal of Intelligent Systems and Applications in Engineering*, 25–30. https://www.ijisae.org/index.php/IJISAE/article/view/4779

Inapagolla, R.K & Babu, K.K. (2025). Audio Fingerprinting to Achieve Greater Accuracy and Maximum Speed with Multi Model CNN-RNN-LSTM in Speaker Identification. *International Journal of Computational and Experimental Science and Engineering*,1108–1116.https://doi.org/10.22399/ijcesen.1138

Kim, S., & Park, J. (2018). Emotional variance in twin speech recognition and its applications. *International Journal of Speech Technology*, 21(4), 589–600. https://doi.org/10.1007/s10772-018-9431-5

Kim, K., & Kim, J. (2021). Attention-based deep learning models for emotion-aware speaker identification. *IEEE Access*, 9, 172134–172145. https://doi.org/10.1109/ACCESS.2021.3068586

Kumar, S., & Aggarwal, N. (2018). Wavelet-based feature extraction for emotion and speaker recognition. *Proceedings of the International Conference on Signal Processing and Communication (SPCOM*), 1–6. https://doi.org/10.1109/SPCOM.2018.8766801

Lee, J., Park, H., & Jeong, H. (2021). Wavelet transform-based features for noise-robust speaker recognition. *Speech Communication*, 134, 12–25. https://doi.org/10.1016/j.specom.2021.03.002

Lee, J., & Cho, H. (2020). Speaker identification for identical twins using deep neural networks. *IEEE Transactions on Audio, Speech, and Language Processing,* 28, 1235–1245. https://doi.org/10.1109/TASLP.2020.2975365

Li, J., Wang, L., & Xu, C. (2022). Wavelet transform for robust emotion recognition in speech signals. *Journal of Electrical Engineering & Technology,* 17(4), 1853–1862. https://doi.org/10.1007/s42835-021-00952-6

Li, J., Zhang, L., Guo, D., Zhuo, S., & Sim, T. (2015). Audio-visual twin's database. *International Conference on Biometrics*, 493–500. https://doi.org/10.1109/ICB.2015.7139115

Liu, J., & Wang, X. (2020). Towards better speaker emotion recognition via fine-grained feature fusion. *IEEE Access*, 8, 178491–178501. https://doi.org/10.1109/ACCESS.2020.3018085

Mavroforakis, M., & Koutroumpis, E. (2020). Robust emotion classification from speech using joint spectral features and temporal attention networks. *Speech Communication,* 124, 1–10. https://doi.org/10.1016/j.specom.2020.02.003

Moore, S., & Li, P. (2019). Emotion-robust speaker recognition systems: Current approaches and challenges. *Journal of Voice,* 33(5), 776–789. https://doi.org/10.1016/j.jvoice.2018.09.004

Park, Y., & Kim, T. (2019). Discriminating identical twins in speaker identification under emotional variability. *Speech Communication*, 108, 23–32. https://doi.org/10.1016/j.specom.2019.02.005

Peng, H., & Zhang, Y. (2019). Cross-corpus speech emotion recognition using deep learning techniques. *IEEE Transactions on Multimedia*, 21(7), 1804–1815. https://doi.org/10.1109/TMM.2019.2898553

Poria, S., & Cambria, E. (2020). Deep learning for emotion recognition: A review. *IEEE Transactions on Affective Computing,* 11(2), 101–110. https://doi.org/10.1109/TAFFC.2020.2975365

Singh, M., & Patel, P. (2020). Feature extraction techniques for speech emotion recognition: A comparative review. *Journal of Signal Processing*, 44(3), 247–261. https://doi.org/10.1109/JSP.2020.2960332

Singh, R., Gupta, A., & Rana, A. (2020). PNCC-based emotion recognition from speech: A comparative study. *Journal of Signal Processing Systems,* 92(2), 179–191. https://doi.org/10.1007/s11265-020-01431-5

Snyder, D., Garcia, M., & Karbasi, A. (2021). ECAPA-TDNN: A deep learning architecture for speaker recognition under emotional variability. *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2021, 6063–6067. https://doi.org/10.1109/ICASSP39728.2021.9413595

Yang, B., & Zhao, Z. (2021). Emotional speech recognition using deep convolutional networks: A comparative study. *Proceedings of the International Conference on Speech Processing,* 112–118. https://doi.org/10.1109/ICSP50755.2021.9473446

Zhang, T., & Li, X. (2018). Speaker recognition under stress and emotion: *Challenges and techniques. Speech Communication*, 102, 38–50. https://doi.org/10.1016/j.specom.2018.04.001

Zhao, X., Li, F., & Zhang, Y. (2020). Deep learning techniques for speech emotion recognition: A survey. *IEEE Access*, 8, 106728–106740. https://doi.org/10.1109/ACCESS.2020.2998756

Zeng, Z., & Li, Z. (2018). Multimodal emotion recognition: A review and new directions. *IEEE Transactions on Affective Computing*, 9(2), 226–240. https://doi.org/10.1109/TAFFC.2017.2709823

Zhang, Z., & Li, J. (2017). Improving speaker identification under emotional variability. *Journal of Voice,* 31(6), 779–788. https://doi.org/10.1016/j.jvoice.2017.02.006

Cite this article:

Inapagolla, R. K., & Babu, K. K. (2025). Overcoming vocal similarities in identical twins: A hybrid deep learning model for emotion-aware speaker and gender recognition. *Journal of Research, Innovation and Technologies*, Volume IV, 1(7), 69-81. https://doi.org/10.57017/jorit.v4.1(7).05

Article's history:

Received 17th of February, 2025; Revised 19th of March, 2025;
Accepted for publication 29th of March, 2025; Available online: 30th of March, 2025
Published as article in Volume IV, Issue 1(7), 2025

© The Author(s) 2025. Published by RITHA Publishing. This article is distributed under the terms of the license CC-BY 4.0., which permits any further distribution in any medium, provided the original work is properly cited maintaining attribution to the author(s) and the title of the work, journal citation and URL DOI.