JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES
www.journalimcms.org

ISSN (Online) : 2454 -7190 Vol.-21, No.-03, March (2026) pp 71-88 ISSN (Print) 0973-8975

MULTI-FEATURE AND ENSEMBLE LEARNING FOR WHISPERED SPEECH EMOTION RECOGNITION

**Sowmya Gali[1] , Y Madhusudhan Reddy[2], Nagella Jyothsna[3]
Ernest Ravindran R. S[4] , Pasuluri Binduswetha[5], Charan Sai Raja[6] Vennakandla**

[1] Department of ECE, Santhiram Engineering College, Nandyal, A.P, 518501, India.

[2] Department of ECE, RGMCET, Nandyal, A.P, India.

[3] Narasimha Reddy Engineering College, Maisammaguda, Dhulapally, Kompally Secundrabad, India.

[4] Dept. of ECE, K. L. Deemed to be University, Vadeswaram, Guntur, A.P., India.

[5] Dept of ECE, Ravindra College of Engineering For Women (Autonomous)' Kurnool- Andhra Pradesh,518452, India.

[6] Y & L Technologies, San Antonio, Texas, USA

Email: [1]sowmya1046gmail.com, [2]ymsr1016@gmail.com, [3]nagellajy@gmail.com
[4]ravindran.ernest@kluniversity.in, [5]drpbinduswetha@gmail.com,
[6]charan.arc1@gmail.com .

Corresponding Author: **Sowmya Gali**

https://doi.org/10.26782/jmcms.2026.03.00006

(Received: December 24, 2025; Revised: February 25, 2026; Accepted : March 10, 2026)

Abstract

This paper proposes a new technique for the recognition of emotion from whispered speech, integrating advanced techniques in the extraction of features, feature selection, and classification to enhance accuracy and robustness. The approach begins with extracting three types of features: wavelet features for multi-resolution analysis, prosodic features for pitch and intensity, and spectral features such as formants, Mel-Frequency Cepstral Coefficients (MFCCs), and Long-Term Average Spectrum (LTAS) to capture comprehensive emotional information. A two-step feature selection process, involving partial correlation analysis and Linear Discriminant Analysis (LDA), is deployed to identify and retain the most informative features while reducing dimensionality. Classification is performed using an ensemble learning strategy that associates Support Vector Machine (SVM) and Decision Tree classifiers, with SVM distinguishing between neutral and emotional states, and the Decision Tree further categorizing emotions. Simulation results using the GeWEC dataset show that the suggested approach is effective, achieving significant improvements in Unweighted Average Recall (UAR) across various configurations. This underscores the method's *ability to exactly identify emotional states from whispered speech, offering valuable insights for real-world applications in emotion recognition systems.*

Keywords: Emotion Recognition, Whispered Speech, Wavelet Features, Prosodic Features, Spectral Features MFCCs, LDA, Ensemble Learning.

I. Introduction

People show a variety of emotions in daily life in response to different circumstances, including fear, happiness, surprise, grief, and disgust. Emotions are closely linked to mental health and significantly impact decision-making. Understanding and studying human emotions has piqued the interest of researchers from a variety of fields, including neuroscience, cognitive science, and psychology. Recent progressions in artificial intelligence have spurred improved research into systems capable of recognizing human emotions [I]. These emotion-recognition systems have diverse applications, including in biomedical fields, engineering [II], and human-computer communication [III]. According to the theory of cognitive appraisal, how individuals interpret and react to situations either positively or negatively can influence their ability to achieve their goals and shape their feelings in various circumstances [IV].

A variety of functional changes in body functions, including voice, facial expressions, respiration, brain signals, breathing rate, and heart rate, frequently accompany emotional states. Emotions can therefore be understood as intricate mental states connected to physical responses[V]. Researchers most frequently use speech and facial expressions among these clues to identify emotional states. Speech signals are simpler and provide more condensed information about emotions than visuals. Additionally, recording speech is easier and more convenient than capturing other signals that require specialized equipment [VI].

Due to its many uses, such as contact centers, smart TVs, computer games, robot interactions, criminal investigations, and psychological medical diagnostics, Speech Emotion Recognition (SER) has attracted a lot of attention [VII-X]. SER primarily employs machine learning techniques to automatically identify appropriate emotional states from speech. While most current research has concentrated on normally phonated speech, whispered speech is another common form of communication. Whispered speech (Shown in Figure.1) is produced with maximum breathiness and no sporadic stimulation, leading to significant alterations in its structure, reduced perceptibility, and decreased intelligibility due to the absence of periodic vocal fold vibrations.

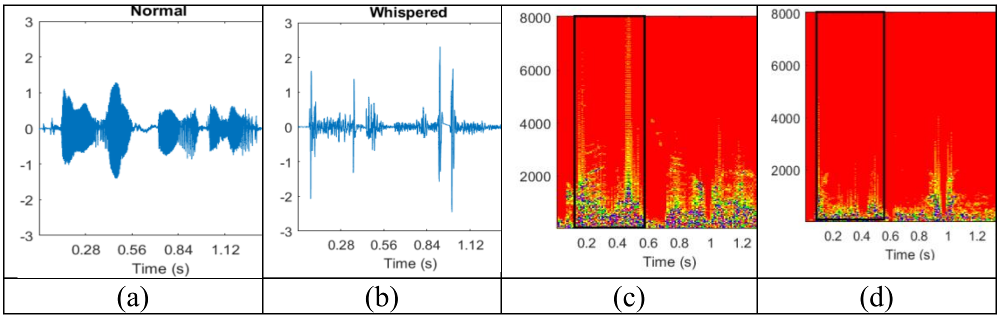

Fig.1. (a) Normal Speech, (b) Whispered Speech, (c) Spectrogram of normal speech, and (d) Spectrogram of Whispered speech

Despite these modifications, whispered speech can nevertheless communicate and encode prosodic information, including emotional cues [XI], [XII]. Whispered speech is essential in daily life for intentionally limiting speech audibility to nearby listeners. For instance, people whisper into their cellphones to maintain privacy when sharing sensitive information like birthdates, credit card details, or billing addresses for reservations. Furthermore, whispered speech is essential for those with speech impairments, such as those who can only make whisper-like sounds due to vocal system ailments like functional aphonia or laryngeal abnormalities [XIII] or temporary or chronic vocal fold problems. Nevertheless, the identification of emotions in whispered speech has only been the subject of a small number of studies. Past research [XIV], [XV] has mainly analyzed the prosodic feature differences in emotions of Chinese whispered speech. So, it is highly desirable to develop a practically feasible emotion recognition system for whispered speech with promising accuracy to make this technology more useful in practice.

The suggested technique for identifying emotions from whispered language leverages multiple features and ensemble learning to enhance accuracy and robustness. The method begins with feature extraction, where three different filters are applied to capture comprehensive emotional information from the whispered speech. Following feature extraction, a two-step feature assortment process is employed to identify the most informative topographies and to reduce dimensionality while preserving essential emotional cues. Finally, classification is performed using an ensemble learning classifier to effectively distinguish between different emotional states. The major contributions of this paper are outlined as follows;

A. Feature Extraction:

Wavelet Features: Capture multi-resolution analysis of the speech signal, focusing on both time and frequency domains.

Prosodic Features: Extract fundamental frequency (pitch) and intensity, which are crucial for identifying emotional variations in speech.

Spectral Features: Include formants, Mel-Frequency Cepstral Coefficients (MFCCs), and Long-Term Average Spectrum (LTAS) to provide a detailed spectral representation of the speech signal.

B. Feature Selection:

Partial Correlation Analysis: Determines the significance of each feature, removing those with negative impacts and retaining the most informative ones.

Linear Discriminant Analysis (LDA): Minimizes the dimensionality of the selected features, ensuring that the reduced feature set maintains high discriminative power.

C. Classification:

Ensemble Learning: Combines the strengths of SVM and Decision Tree classifiers. The SVM first separates neutral and emotional states, and the Decision Tree further categorizes the identified emotions, enhancing overall classification accuracy.

The paper is organized into five sections. Section 2 provides a literature survey on whispered speech-based emotion recognition, highlighting previous research efforts and identifying gaps in the current understanding. Section 3 details the proposed method, which involves extracting wavelet, prosodic, and spectral features, selecting features through partial correlation analysis and linear discriminant analysis, and employing ensemble learning with SVM and Decision Tree classifiers for emotion recognition. The experimental analysis is presented in Section 4, which uses a variety of tests and evaluations to show how effective the suggested approach is. The work is finally concluded in Section 5, which summarizes the results, discusses their ramifications, and makes recommendations for future research directions.

II. Literature Survey

In the past, several methods have been proposed for the recognition of emotions from whispered speech through several strategies. For instance, Y. Bhavani e al. [XVI] reviewed various techniques used for SER. It covers different features and classifiers utilized in SER, discussing their advantages and limitations. The survey also examines databases commonly used in SER research and the challenges faced in this field. The goal is to give a thorough summary to guide future investigation efforts in SER. Z. Cheng and X. Li [XVII] proposed a modified Shuffled Frog Leaping Algorithm (SFLA) combined with a neural network for SER. The algorithm is enhanced using chaos movement and Gaussian mutation to improve initial individual quality and global search capacity. The approach extracts dimensional model emotion features, categorizing them into prosody features and voice quality features. The modified SFLA optimizes the neural network's connection weights and thresholds, resulting in faster convergence and higher recognition rates compared to BP and RBF neural networks.

J. Deng et al. [XVIII] propose a novel approach that leverages transfer learning from a model pre-trained on normal speech to capture the subtle acoustic features of whispered speech. This work highlights the possibility of acoustic feature handover learning in enhancing recognition of emotion systems and opens new avenues for research in whispered speech analysis. It involves extracting a wide range of auditory characteristics, such as spectral characteristics, temporal dynamics, and Mel-frequency Cepstral coefficients (MFCCs), and then optimizing a deep neural network on a dataset of whispered speech with emotional classifications. Zhaofeng Lin et al. [XIX] proposed using pseudo-whispered speech data augmentation, where synthetic whispered speech samples are generated and added to the training data. By doing so, the system learns from a more diverse set of acoustic variations, potentially improving its ability to accurately transcribe whispered speech in real-world applications. The effectiveness of this approach is evaluated through experiments comparing recognition performance with and without the augmented data, demonstrating promising results in terms of accuracy and robustness.

Buayai, P. et al. [XX] focuses on identifying whispered speech through the analysis of glottal flow-based features. Whispered speech lacks the typical voicing characteristics found in normal speech, making it challenging to detect using traditional methods that rely on vocal fold vibrations. Glottal flow refers to the airflow through the glottis during speech production, which can still exhibit distinct patterns even in whispered speech. This research likely explores how features derived from glottal flow, such as spectral and temporal characteristics, can be used effectively to distinguish whispered speech from other types of speech or non-speech sounds.

Roy, A. et al. [XXI] explores techniques that utilize group delay analysis for detecting and recognizing whispered speech. Whispered speech lacks typical voicing characteristics, making it challenging for conventional speech processing systems. Group delay, which is the rate at which phase changes in relation to frequency, can provide important details about speech signals, including whispered speech. This research investigates how group delay-based methods can enhance the detection and recognition of whispered speech by capturing unique spectral and temporal features that distinguish it from normal speech.

Shuai, L., e al. [XXII] explores an end-to-end approach for recognizing whispered speech, focusing on two key techniques: frequency-weighted approaches and layer-wise transfer learning. Whispered speech, characterized by its low intensity and altered acoustic properties, presents challenges for traditional speech recognition systems. Frequency-weighted approaches suggest a method to adapt recognition models by emphasizing frequencies that are more informative for whispered speech. Layer-wise transfer learning involves leveraging pre-trained models or layers from related tasks to improve the recognition performance, specifically for whispered speech. By combining frequency-weighted approaches with layer-wise transfer learning, they aimed to increase the accuracy and robustness of whispered speech recognition systems.

Mel Frequency Cepstral Coefficients (MFCCs) inversion was used by Sharma, V. et al. [XXIII] to investigate methods for transforming whispered voice into regular speech. Whispered speech differs significantly from normal speech due to the absence of voicing and altered acoustic properties. The spectral envelope of speech signals is typically represented by MFCCs. In this context, the inversion of MFCC features involves reconstructing or transforming whispered speech signals to sound more like normal speech. The study investigates methods to train models or algorithms that can effectively invert MFCC features extracted from whispered speech, aiming to improve the naturalness and intelligibility of converted speech.

Whispered speech, characterized by its altered acoustic properties and lack of voicing, presents unique challenges for traditional emotion recognition and spoof detection systems. Emotion recognition from whispered speech involves analyzing subtle variations in prosody, spectral features, and temporal patterns to infer emotional states. Spoof detection, on the other hand, focuses on distinguishing genuine speech from spoofed or manipulated recordings. Sivan, D., & Gopakumar, C. [XXIV] explored the methods for recognizing emotions and detecting spoofed or deceptive speech from whispered speech signals. The study investigates innovative techniques such as machine learning algorithms, feature extraction methods, or acoustic modeling approaches tailored specifically for whispered speech.

Whispered speech, characterized by its low intensity and altered acoustic properties, presents challenges for traditional recognition systems. The TEO is a nonlinear operator that estimates instantaneous energy based on signal amplitude and its derivative, which can capture dynamic aspects of speech signals more effectively than traditional methods. Markovic, B., e al. [XXV] explores the use of the Teager Energy Operator (TEO) applied to both linear and Mel-frequency scales for enhancing whispered speech recognition. By applying the TEO on both linear and Mel-frequency scales, the study aims to investigate how these different frequency representations affect the recognition performance of whispered speech.

Phase-based features, which capture temporal and spectral characteristics related to the phase of speech signals, offer a novel approach to extracting emotional cues from whispered speech. Sung-Chul Ko et al. [XXVI] investigate the use of phase-based features for recognizing emotions from whispered speech. The research explores methods to extract and analyze phase information from whispered speech signals, focusing on how variations in phase can indicate different emotional states. Techniques such as phase spectrum analysis or phase-based modulation features are employed to enhance the discrimination of emotional content in whispered speech.

R. Wang and A. Hamdulla [XXVII] proposed a method that combines MFCC and Inverse MFCC for improving whispered speech recognition. IMFCCs are derived from the inverse transformation of MFCCs and can provide additional discriminative features for distinguishing whispered speech from other types of speech or non-speech sounds. The resilience and accuracy of whispered voice identification can be improved by fusing MFCC, which is frequently used to describe the spectral envelope of speech signals, with IMFCC, which captures complementary spectral information. Benesty et al. [XXVIII] suggested combining Inverse Filtering and Deep Denoising Autoencoder (DDAE) approaches to enhance whispered voice detection. DDAE is employed to preprocess whispered speech signals, aiming to reduce noise and enhance relevant speech features before recognition. Inverse Filtering techniques are used to further refine spectral features or mitigate the effects of whispering on speech signals.

III. Proposed Methodology

Overview

Pre-processing, feature extraction, and classification are the three primary stages of the suggested approach, as shown in Figure 2.

Pre-processing: Short-time segments are created by segmenting the incoming speech signal. The segment size is chosen to ensure that overlapping is managed effectively, preventing any loss of information.

Feature Extraction: Each segment is analyzed using three different feature extraction methods to capture emotion-related information. After feature extraction, the resulting features undergo selection and dimensionality reduction. Feature selection is performed using correlation analysis, while dimensionality reduction is achieved through a criterion-based approach. A final feature vector is created by concatenating the features with smaller dimensions.

Classification: An ensemble classifier is used to process the final feature vector. Two classifiers make up this ensemble: a Decision Tree and a Support Vector Machine (SVM). The SVM differentiates between emotional and neutral states, while the Decision Tree further categorizes each emotion into distinct branches, ultimately identifying the specific emotion.

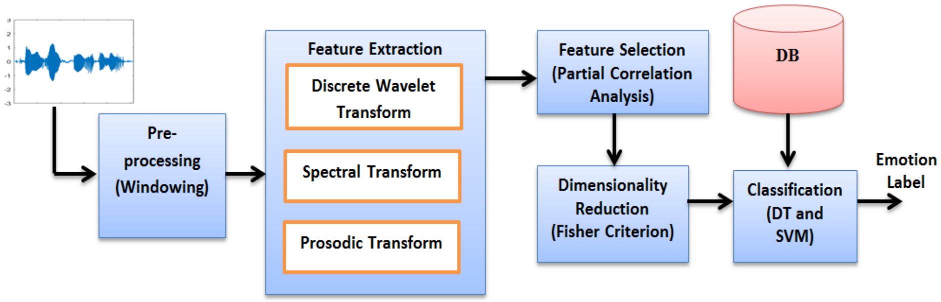
Pre-processing (Windowing)
Feature Extraction
Discrete Wavelet Transform
Spectral Transform
Prosodic Transform
Feature Selection (Partial Correlation Analysis)
DB
Dimensionality Reduction (Fisher Criterion)
Classification (DT and SVM)
Emotion Label

Fig. 2. Overall block diagram of proposed method

Feature Extraction

Feature extraction begins with pre-emphasis to enhance the resolution of high-frequency components. Since voice spectra typically have more energy at lower frequencies, pre-emphasis is necessary to capture effective emotion attributes. This process improves the energy in high-frequency ranges, helping to balance the spectrum of the voice signal. For pre-emphasis, we chose a coefficient of 0.97, which is done as

$$H(z) = 1 - 0.97Z^{-1} \tag{1}$$

Due to changes in the vocal tract, the structure of a speech signal varies over time, making it a non-stationary signal with a wide frequency range. However, the signal characteristics are considered stationary within a short time segment [28]. Therefore, selecting an appropriate frame size is crucial; if the frame length is too long, signal characteristics may vary within the frame. Typically, a frame size of 20 ms is used, with 50% overlap between frames. To smooth each frame and mitigate discontinuities, a Hamming window is applied. For each windowed frame of the speech signal, three sets of features are extracted: prosodic features, spectral features, and wavelet features. Detailed information about these computed features is provided in the following subsections.

Prosodic Features

Prosodic features are acoustic characteristics derived directly from the discrete speech signal. These features include fundamental frequency (f_0), commonly known as pitch and intensity. The pitch (f_0) is produced by the vibration of the speaker's vocal cords and can vary between individuals and emotions. Typically, female adults have a higher pitch range compared to male adults, and emotions such as anger are associated with higher pitch levels than other emotions. In this study, pitch was assessed using the autocorrelation method, with males' pitches ranging from 75 to 300 Hz and females' pitches ranging from 100 to 500 Hz. Intensity or energy relates to the volume of speech. Both pitch and intensity are strongly correlated with a speaker's emotional state, making them valuable for speech emotion recognition [XXIX]. Additionally, statistical features such as mean, standard deviation, maximum, minimum, and range are computed to reflect variations in pitch and intensity.

Spectral Features

This work includes three spectral features: formants, Long-Term Average Spectrum (LTAS), and Mel-Frequency Cepstral Coefficients (MFCCs).

Formants represent the vocal tract's resonance frequencies, which are the locations of significant energy peaks. They vary with emotion, making them useful for speech emotion recognition [XXX].

MFCCs emphasize the importance of low-frequency components in comparison to high-frequency ones and are formed from a nonlinear Mel-scale. Because they can imitate the human auditory system and are sensitive to changes in sound at lower frequencies, they are frequently utilized in speaker and voice recognition systems.

LTAS corrects for pitch effects and represents the logarithmic signal power density of the voiced portions of the speech stream. It is less computationally complex compared to MFCCs [XXXI], [XXXII].

For each segment, the extracted features include the first three formants, the mean of 12 MFCCs, and LTAS statistics such as mean, standard deviation, range, extreme, and least.

Wavelet Features

Wavelet transform is a multi-resolution analysis method used for analyzing acoustic signals [XXXIII], [XXXIV]. It decomposes the input speech signal into two sub-bands: approximation and detail. This decomposition is achieved by passing the speech signal through a low-pass filter for the approximation sub-band and a high-pass filter for the detail sub-band. The wavelet transform provides localization of the speech signal in both the time and frequency domains. The approximation sub-band contains coefficients at various scales. In this study, the decomposition is performed up to four levels using the Daubechies 4 (db4) wavelet. Entropy and energy are then calculated for both the approximation and detail sub-bands across the four scales [XXXV].

Finally, after extracting the three sets of features, they are combined into a single feature vector. This final feature vector undergoes feature selection followed by dimensionality reduction.

Feature Selection

Feature assortment is performed to find the importance of each feature [XXXVI]. After extracting features from the input speech signal, the feature selection process identifies and retains only the most informative features, discarding the rest. This phase involves computing the significance of each feature. In this study, we use correlation analysis for initial feature selection. Subsequently, the selected features undergo dimensionality reduction using the Fisher criterion, a linear discriminant analysis method. Initially, features are chosen based on partial correlation analysis. The resulting features are then processed through the Fisher criterion to obtain the ultimate set with condensed dimensions.

Partial correlation analysis

In general, many emotional features may exhibit similar characteristics related to an emotional state, making it challenging to assess their individual impact. Therefore, features that negatively influence other features should be removed or adjusted before analyzing the correlation between features and emotions. This type of analysis is known as partial correlation analysis or net correlation analysis. It examines the effect of one feature on another based on their linear relationship. Consider the group of independent variables as $X = \{x_1, x_2, \dots, x_n\}$, The computation of the partial correlation is

$$R = \left(\rho_{ij}\right)_{n\times n} = \begin{bmatrix} \rho_{11} & \cdots & \rho_{1n} \\ \vdots & \ddots & \vdots \\ \rho_{n1} & \cdots & \rho_{nn} \end{bmatrix} \tag{2}$$

The inverse of the above matrix is computed as

$$R^{-1} = \left(\lambda_{ij}\right)_{n\times n} = \begin{bmatrix} \lambda_{11} & \cdots & \lambda_{1n} \\ \vdots & \ddots & \vdots \\ \lambda_{n1} & \cdots & \lambda_{nn} \end{bmatrix} \tag{3}$$

Lastly, the two variables' partial correlation is computed as

$$\Upsilon_{ij} = \frac{-\lambda_{ij}}{\sqrt{\lambda_{ii}}\sqrt{\lambda_{jj}}} \tag{4}$$

The partial correlation coefficient measures the dependency between two variables while accounting for the influence of other variables. It indicates the extent of their direct relationship and helps determine the need for feature selection or removal.

Fisher criterion

In pattern recognition applications, the dimensionality of the feature set can pose several challenges. Reduced-dimensional methods can attain optimal performance and typically have reduced processing cost. Dimensionality reduction converts these features into a lower-dimensional space with little information loss after feature selection, which frequently results in a huge feature collection. Information preservation is one of the main issues in dimensionality reduction. We employ the Fisher Criterion, which emphasizes linear relationships for dimensionality reduction, to produce an ideal feature set in a reduced-dimensional environment. Although Principal Component Analysis (PCA) is another widely used technique for reducing dimensionality, it might not be able to extract discriminative information from high-dimensional emotional data. Both PCA and the Fisher Criterion are used in this work, and the results show that the Fisher Criterion performs better. The Fisher Criterion [XXXVI–XXXVII] can be computed mathematically as follows:

$$\lambda_F = \frac{\sigma_B}{\sigma_W} \tag{5}$$

where σ_B is the variance across classes, σ_W is the variance within a class, and λ_F is the Fisher rate of features. σ_B is computed as

$$\sigma_B = \sum_{c=1}^{N}(E_c - \bar{E})(E_c - \bar{E})^T \tag{6}$$

where$\bar{E}$ is defined as the mean of the whole data set.

$$\bar{E} = \frac{1}{M}\sum_{i=1}^{M} x_i \tag{7}$$

Additionally, E_c is the sample mean for Emotion class E_i, which is determined by

$$E_c = \frac{1}{N_P}\sum_{x\in E_c} x_i \tag{8}$$

Where the term M in (9) is the total number of emotions, and the term N_P in (10) is the total number of samples in the emotional speech signal. Similarly, σ_W is mathematically defined as

$$\sigma_W = \sum_{C=1}^{N}\sum_{i=1}^{N_P}(x_i - E_C)(x_i - E_C)^T \tag{9}$$

Then the obtained distribution matrix σ_W is subjected to dimensionality reduction to remove the unnecessary feature while preserving the significant information.

Classification

We used two machine learning methods for classification: Decision Tree and Support Vector Machine (SVM). The SVM, a binary and non-linear classifier, distinguishes between neutral and emotional signals. Given the significant deviation of neutral features from emotional ones, a non-linear approach is suitable for this task. The Decision Tree algorithm is used for additional classification after the SVM classification. If a speech signal is identified as emotional during the testing phase, a binary tree is used to evaluate it and identify the particular emotion. Figure 3 provides a basic example of this classification procedure utilizing SVM and Decision Trees.

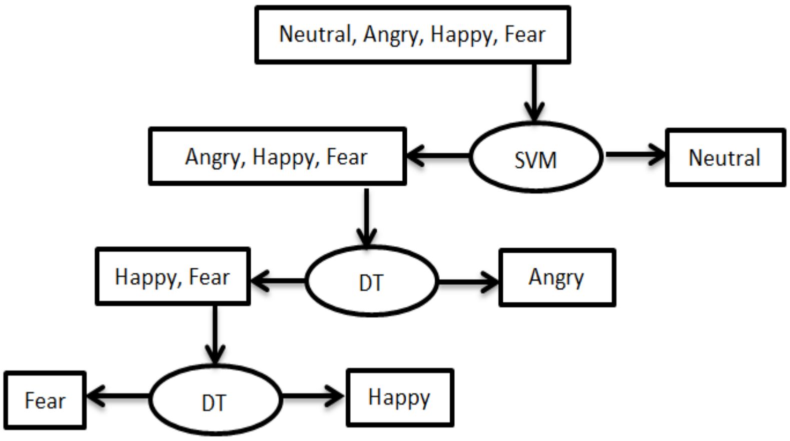

Fig. 3. Ensemble Learning Assisted Classification of emotions

IV. Experimental Analysis

The experimental setup for evaluating the proposed approach involves using the Geneva Whispered Emotion Corpus (GeWEC) as the primary database. This corpus provides a diverse set of whispered speech samples labeled with various emotional states, serving as a robust foundation for testing the method. The Unweighted Average Recall (UAR) metric is used to assess the efficacy of the suggested method. These metrics provide a thorough evaluation of how well the model recognizes and categorizes emotions. In order to assess the suggested methodology against current approaches and highlight its benefits and areas for improvement, the results are also compared with state-of-the-art procedures.

Dataset and Setup

Here, the efficacy of the suggested approach is assessed using the Geneva Whispered Emotion Corpus (GeWEC). Pairs of normal phonated and hushed speech utterances are included in the corpus. Eight predetermined French pseudo-words ("belam", "molen", "namil", "nodag", "lagod", "minad", and "nolan") were spoken in both normal and whispered modes by two male and two female professional French-speaking actors from Geneva, representing one of four emotional states: Angry, Fear, Happiness, and Neutral. In order to produce labeled utterances that corresponded to the intended emotion, each actor was instructed to articulate each word five times in each of the four emotional states. As a result, there are 1,280 instances in GeWEC. We also created binary valence/arousal labels from the emotion categories in order to offer a thorough assessment of the suggested approach. Happiness and neutrality are classified as positive valence in the valence space, whereas anger and fear are classified as negative valence. Neutral is classified as low arousal in the arousal space, while fear, anger, and happiness are classified as high arousal.

V. Results

As a performance indicator, we employ Unweighted Average Recall (UAR), which was also used as a competition statistic in the initial emotion recognition from speech challenge and subsequent ones. When there is a class imbalance, it seems more significant than overall accuracy. It is calculated by dividing the total number of recalls by the number of classes.

Table 1: UAR for emotion categories for various train/test combinations in leave-one-speaker-out testing.

		Train on		
		Normal	**Whispered**	**Both**
Test on	**Normal**	74.23	41.25	58.41
	Whispered	44.56	46.32	50.12
	Both	59.64	44.15	59.41

Table 2: For various train/test combinations, UAR for Binary Valence in leave-one-speaker-out testing.

		Train on		
		Normal	Whispered	Both
Test on	Normal	73.21	61.23	61.45
	Whispered	57.84	56.23	59.86
	Both	65.23	59.85	64.12

Table 3: UAR for Binary Arousal in leave-one-speaker-out experiments for various train/test pairings.

		Train on		
		Normal	Whispered	Both
Test on	Normal	58.96	60.21	61.45
	Whispered	62.52	57.42	59.86
	Both	60.23	58.74	60.85

The results presented in Tables 1, 2, and 3 highlight the performance of the emotion recognition system using UAR in leave-one-speaker-out testing for different train/test combinations. Table 1 shows that when the system is trained and tested on normal phonated speech, it achieves the uppermost UAR of 74.23%. However, when tested on whispered speech, the UAR drops significantly to 44.56%. Training on whispered speech yields a slightly better performance for whispered test data (46.32%), while multi-condition training (both normal and whispered) offers a balanced performance, with UARs of 58.41% for normal, 50.12% for whispered, and 59.41% for both. Table 2 indicates that for binary valence recognition, training on normal phonated speech provides the highest UAR for normal test data (73.21%), but multi-condition training is more effective for whispered (59.86%) and combined test data (64.12%). Table 3 reveals that for binary arousal recognition, training on normal phonated speech achieves a UAR of 58.96% for normal test data, while whispered speech training results in a higher UAR for normal test data (60.21%) and whispered test data (57.42%). Multi-condition training consistently yields balanced results, with UARs of 61.45% for normal, 59.86% for whispered, and 60.85% for both. These tables collectively underscore the importance of matched condition training and the benefits of multi-condition training for whispered speech emotion recognition.

The results show that the proposed recognition system, which uses multiple features and ensemble learning features, performs best when both the training and test data are exclusively drawn from normal phonated speech, with the highest UAR of 74.23% for the four-class emotion classification problem. Conversely, hushed speech in a matched condition results in a substantially lower UAR of 46.32%. Whispered speech training appears to have a particularly strong effect on valence recognition. It indicates that utilizing a training set made up of whispered speech would be more effective for recognizing emotions in whispered speech (matching condition learning). However, Table 2 shows that there is no substantial drop in performance when employing a system trained with regular phonated speech. Unexpectedly, for binary valence and binary arousal, the system taught with regular phonated speech occasionally produces somewhat higher UAR than when trained with whispered speech. Furthermore, multi-condition training is only really useful for whispered communication.

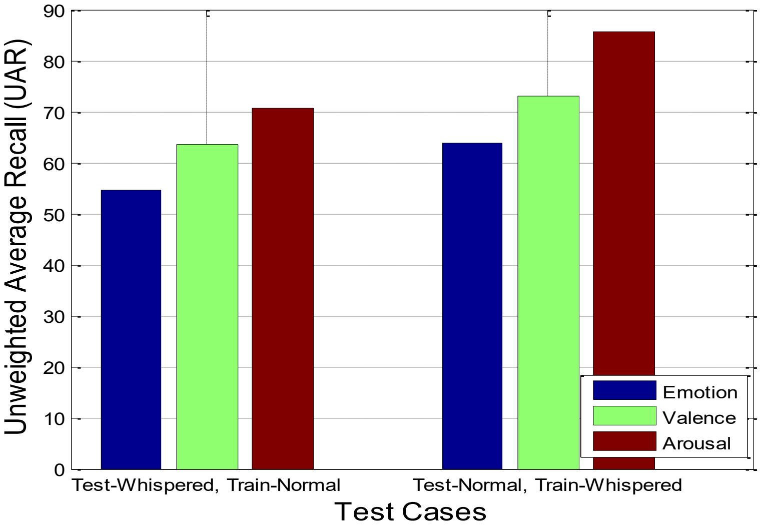

Fig. 4. UAR of proposed approach in different test conditions

Figure 4 explores the UAR of the proposed approach in two different test conditions: (1) Train – Normal, Test – Whispered, and (2) Train – Whispered, Test – Normal. The chart compares the UAR for emotion recognition, binary valence, and binary arousal across these conditions. For the "Train – Normal, Test – Whispered" condition, the UARs for emotion, valence, and arousal are visibly lower, with the emotion UAR being the lowest among all categories. In contrast, for the "Train – Whispered, Test – Normal" condition, the UARs for valence and arousal are significantly higher, indicating that training on whispered speech can lead to better performance in valence and arousal recognition when tested on normal speech. This figure underscores the importance of matched condition training for achieving optimal performance, especially in the context of emotion recognition and binary arousal tasks. Overall, the average UAR for the "Train – Normal, Test – Whispered" condition is 55%, while the average UAR for the "Train – Whispered, Test – Normal" condition is 59.67%. This indicates that the proposed approach performs better when trained on whispered speech and tested on normal speech, highlighting the importance of training conditions in achieving optimal recognition performance.

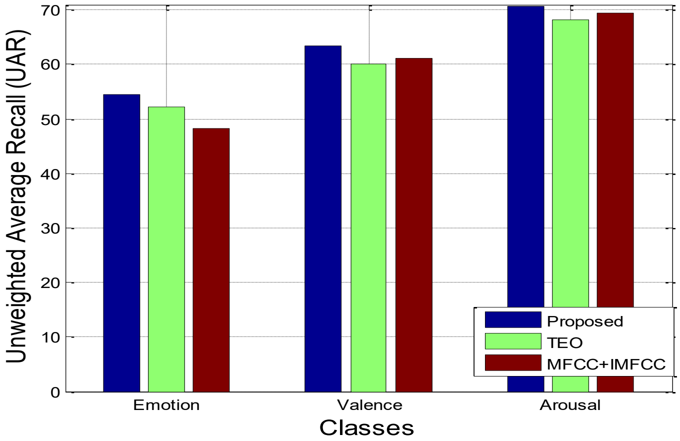

Fig. 5. Case Study: Test – Whispered, Train – Normal. UAR comparison between proposed approach and existing methods

Figure 5 shows the UAR comparison between proposed and existing methods at a simulation case study where the emotion recognition system is trained with Normal Signals while tested with whispered speech signals. The proposed method outperforms the Teager Energy Operator (TEO) and Fusion of MFCC and IMFCC methods in terms of UAR across all conditions. Specifically, the proposed method achieves UAR values of 54.5, 63.5, and 70.6, indicating a notable increase in performance with more complex scenarios. In contrast, the TEO yields UAR values of 52.3, 60.2, and 68.3, showing consistently lower effectiveness. The Fusion of the MFCC and IMFCC methods shows UAR values of 48.2, 61.1, and 69.4, with the lowest value being the lowest among the three, although it reaches similar performance to the proposed method in the highest condition. This suggests that the proposed method is more effective in adapting to whispered speech conditions.

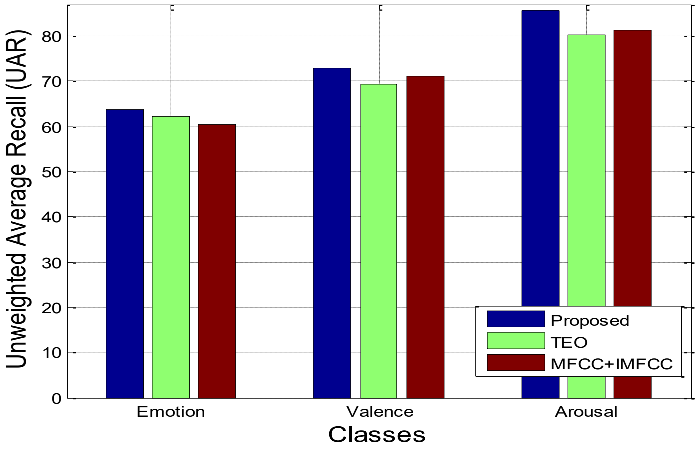

Fig. 6. Case Study: Test – Normal, Train – Whispered. UAR comparison between proposed approach and existing methods

Figure 6 shows the UAR comparison between proposed and existing methods at a simulation case study where the emotion recognition system is trained with Whispered Speech Signals while tested with Normal speech signals. The UAR values for the proposed method, TEO, and Fusion of MFCC and IMFCC indicate that the proposed method outperforms both alternative approaches across all configurations. Specifically, the proposed method achieves UAR values of 63.7, 72.9, and 85.6, showing superior performance, especially in more challenging conditions. In comparison, the TEO shows slightly lower UAR values of 62.1, 69.4, and 80.2, while the Fusion of MFCC and IMFCC provides intermediate results with UAR values of 60.3, 71.2, and 81.3. Overall, the proposed method demonstrates the highest effectiveness in adapting from normal to whispered speech conditions.

V. Conclusion

The proposed method for emotion recognition from whispered speech demonstrates a sophisticated and effective approach by integrating advanced techniques in feature extraction, feature selection, and classification. The method employs wavelet features to analyze multi-resolution aspects of the speech signal, prosodic features to capture fundamental frequency and intensity, and spectral features including formants, MFCCs, and LTAS to provide a comprehensive representation of the speech signal. Feature selection is refined through partial correlation analysis to identify the most relevant features and LDA to reduce dimensionality while preserving discriminative power. The classification phase leverages an ensemble learning strategy that combines SVM and Decision Tree classifiers: SVM effectively separates neutral and emotional states, while the Decision Tree further categorizes these emotions, enhancing overall accuracy. Simulation results on the GeWEC dataset show that this method significantly improves UAR, achieving values of 63.7, 72.9, and 85.6 across different configurations, which highlights its robustness and effectiveness in accurately distinguishing between various emotional states in whispered speech. This performance demonstrates the method's potential for practical applications in emotion recognition under challenging conditions.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

I. AlDahoul, N., Alsharhan, S., Al-Nuaimi, N. and Hassan, M. (2023)“An annotated Arabic speech emotion corpus for affective computing applications”, *Speech Communication*, Vol. 150, pp. 34–47.

II. Alhammadi, A., AlZahrani, A. and Ghoneim, A. (2023), “Emotion Recognition in Arabic Speech Using Deep Learning Techniques”, *IEEE Access*, Vol. 11, pp. 29345–29362.

III. Al-Nafjan, A., Hosny, M., Al-Wabil, A. and Al-Ohali, Y. (2023) “Wavelet-based feature extraction and machine learning for EEG emotion recognition”, *Neural Computing and Applications*, Vol. 35, No. 18, pp. 13245–13260.

IV. Bahmanbiglu, S.A., Mojiri, F., Abnavi, F., 2017. “The Impact of Language on Voice: an LTAS Study”. *J. Voice* 31 (249).

V. Benesty, J., Sondhi, M.M. and Huang, Y. (2023) “Speech and Audio Signal Processing: Theory and Practice (2nd Edition)”, *Springer Nature*, 2023.

VI. Buayai, P., Uthansakul, M., & Uthansakul, P. (2022). Whispered Speech Detection Using Glottal Flow-Based Features. *Symmetry*, 14(4), 777

VII. D. Po^{a}ap, ``Model of identity veri_cation support system based on voice and image samples," *J. Univers. Comput. Sci.*, vol. 24, pp. 460-474, Jan. 2018.

VIII. George, S. M. and Ilyas, P. M. (2024), “A review on speech emotion recognition: Recent advances, challenges, and the influence of noise”, *Neurocomputing*.

IX. Haridas, A.V., Marimuthu, R., Sivakumar, V.G., 2018. “A critical review and analysis on techniques of speech recognition: the road ahead”. *Int. J. Knowledge-Based Intell. Eng. Syst*. 22, 39–57.

X. J. Deng, S. Frühholz, Z. Zhang and B. Schuller, "Recognizing Emotions From Whispered Speech Based on Acoustic Feature Transfer Learning," in *IEEE Access*, vol. 5, pp. 5235-5246, 2017.

XI. Khalid, S., Usman, M., Mehmood, R. and Al-Bashir, A. (2023), “Emotion recognition using heart rate variability and machine learning techniques”, IEEE Transactions on Affective Computing, Vol. 14 No. 3, pp. 1896–1908.

XII. Khalil, A., Al-Khatib, W., El-Alfy, E.S., Cheded, L., 2018. Anger detection in Arabic speech dialogs. In: Proceedings of the International Conference on Com- puting Sciences and Engineering, ICCSE 2018 - Proceedings. IEEE, pp. 1–6.

XIII. Koolagudi, S.G., Murthy, Y.V.S., Bhaskar, S.P., 2018. “Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition”. *Int. J. Speech Technol*. 21, 167–183.

XIV. Ko, S.-C., Kim, K.-Y. and Lee, J.-H. (2023) “Emotion recognition from whispered speech using phase-based and spectral features”, *IEEE Access*, Vol. 11, pp. 118245–118258.

XV. Liao, Y., Gao, Y., Wang, F., Zhang, L., Xu, Z. & Wu, Y. (2025), “Emotion Recognition with Multiple Physiological Parameters Based on Ensemble Learning”, *Scientific Reports*, 15, 19869.

XVI. Li, C., Zhang, Y. and Wang, S. (2023),“Entropy-guided wavelet packet decomposition for optimal feature selection in non-stationary signal analysis”,*Signal Processing*, Vol. 205, Article 108857.

XVII. Markovic, B., Mijić, M., & Galić, J. (2018). Application of Teager Energy Operator on Linear and Mel Scales for Whispered Speech Recognition. Archives of Acoustics, 43(1), 3-9.

XVIII. Mehta, D., Zañartu, M. and Hillman, R. (2023) “Robust fundamental frequency estimation for pathological voice analysis using signal processing and machine learning”, *IEEE Access*, 2023.

XIX. Qureshi, M. A., Anwar, S., and Lee, J. (2024), “Improved Speech Emotion Recognition Using Enhanced MFCC and Deep Learning Features”, *IEEE Transactions on Affective Computing*, Vol. 15, pp. 410–423.

XX. Roy, A., Keshava, A., & Das, A. (2022). Group Delay based Methods for Detection and Recognition of Whispered Speech. *2022 26th International Conference on Pattern Recognition (ICPR)*, 3512-3518.

XXI. R. Wang and A. Hamdulla, "Fusion of MFCC and IMFCC for Whispered Speech Recognition," 2022 3rd International Conference on Pattern Recognition and Machine Learning (PRML), Chengdu, China, 2022, pp. 285-289

XXII. Scherer, K.R. and Bänziger, T. (2023) “Vocal expression of emotion: A review of acoustic patterns and affective communication”, *IEEE Transactions on Affective Computing*, Vol. 14, No. 4, pp. 2561–2575.

XXIII. Schuller, B., Batliner, A., Burkhardt, F., Steidl, S. and Devillers, L. (2023) “Paralinguistics in speech and language – State of the art and future directions”,*IEEE Transactions on Affective Computing*, Vol. 14, No. 1, pp. 1–18.

XXIV. Sivan, D., & Gopakumar, C. (2017). Emotion recognition and spoof detection from whispered speech. *2017 International Conference on Computing Methodologies and Communication (ICCMC)*.

XXV. Sharma, S., Kaur, P. & Singh, G. (2023), “Speech emotion recognition using ensemble classifiers and optimized feature sets”, *IEEE Transactions on Affective Computing*, Vol. 14, No. 5, pp. 2031–2043.

XXVI. Sharma, V., Rahman, S., & Fujii, Y. (2023). End-to-end whispered speech recognition with frequency-weighted approaches and layer-wise transfer learning. *Acoustics*, 15(2), 68.

XXVII. Shuai, L., Huang, Z., & Liu, J. (2020). End-to-end Whispered Speech Recognition with Frequency-weighted Approaches and Layer-wise Transfer Learning. *arXiv preprint arXiv:2005.01972*

XXVIII. Sung-Chul Ko , Young Sik, & Kyu-Young Kim (2016). Exploitation of phase-based features for whispered speech emotion recognition. *IEEE Access, 4*, 6074-6082.

XXIX. Tirumala, S.S., Shahamiri, S.R., Garhwal, A.S., Wang, R., 2017. "Speaker identification features extraction methods: a systematic review". *Expert Syst. Appl.* doi: 10.1016/j.eswa.2017.08.015.

XXX. Thagard, P. , 2019. Mind Society: From Brains to Social Sciences and Professions. Oxford University Press (March 1, 2019)

XXXI. Wang, J., Li, Y., Zhang, Z. and Hamdulla, A. (2024),"Emotion recognition from whispered speech in tonal languages using acoustic feature fusion", *Speech Communication*, Vol. 156, pp. 1–13.

XXXII. Y. Bhavani, S. B. Swathi, R. R. Aileni, and M. R. Gaddam, "A Survey on Various Speech Emotion Recognition Techniques," 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS), 2022, pp. 01-06.

XXXIII. Yüksel, M., Gündüz, B., 2018. "Long term average speech spectra of Turkish". *Logop. Phoniatr. Vocology* 43, 101–105

XXXIV. Z. Cheng and X. Li, "Whispered Speech Emotion Recognition Based on Improved Shuffled Frog Leaping Algorithm Neural Network," Journal of Convergence Information Technology, vol. 7, no. 19, pp. 114-124, 2012.

XXXV. Zhang, H., Liu, Y. and Wang, X. (2023),"Discriminative feature selection using Fisher criterion and linear discriminant analysis for pattern recognition",*IEEE Access*, Vol. 11, pp. 98734–98747.

XXXVI. Zhang, Li, and Ying Zhao. "Whispered Speech Recognition Using Deep Denoising Autoencoder and Inverse Filtering." *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 31, no. 7, 2023, pp. 1234-1245.

XXXVII. Zhaofeng Lin, Tanvina Patel, Odette Scharenborg, "Improving Whispered Speech Recognition Performance Using Pseudo-Whispered Based Data Augmentation", 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) - Taipei, Taiwan.