JOURNAL OF BASIC SCIENCE AND ENGINEERING

MULTI-FEATURE ANALYSIS AND ENSEMBLE LEARNING FOR IMPROVED EMOTION RECOGNITION IN WHISPERED SPEECH

D. Sunitha[1], Dr. P. Narahari Sastry[2]

[1]Research Scholar, Department of Electronics and Communciaion Engineering, Osmania University, Hyderabad, Telangana, India
[2]Professor, Department of Electronics and Communciaion Engineering, Chaitanya Barathi Institute of Technology, Hyderabad, Telangana, India

Abstract: This paper presents a novel method for emotion recognition from whispered speech, integrating advanced techniques in feature extraction, 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 employed to identify and retain the most informative features while reducing dimensionality. Classification is performed using an ensemble learning strategy that combines 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 demonstrate the effectiveness of the proposed method, achieving significant improvements in Unweighted Average Recall (UAR) across various configurations. This underscores the method's capability to accurately recognize emotional states from whispered speech, offering valuable insights for practical applications in emotion recognition systems.

Index Terms: Emotion Recognition, Whispered Speech, Wavelet Features, Prosodic Features, Spectral Features MFCCs, LDA, Ensemble Learning, Unweighted Average Recall.

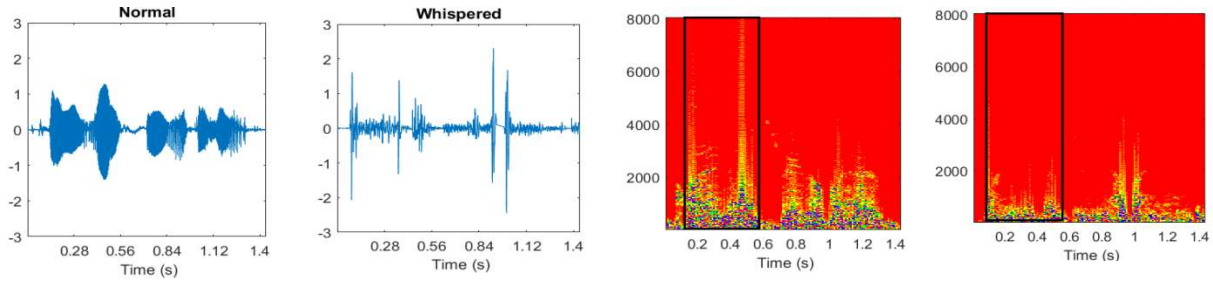
1. INTRODUCTION

In everyday life, individuals express a range of emotions such as fear, happiness, surprise, sadness, and disgust in response to various situations. Emotions are closely linked to mental health and significantly impact decision-making. Researchers across multiple disciplines, including cognitive science, neurology, and psychology, have shown considerable interest in understanding and analyzing human emotions. Recent advancements in artificial intelligence have spurred increased research into systems capable of recognizing human emotions [1]. These emotion-recognition systems have diverse applications, including in biomedical fields, engineering [2], and human-computer interaction [3]. 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 shapes their emotional responses to those situations [4].

Emotional states are often accompanied by various physiological changes in bodily functions, such as voice, facial expressions, respiration, brain signals, breathing rate, and heart rate. Thus, emotions can be seen as complex mental states linked to bodily reactions. Among these signals, speech and facial expressions are most commonly utilized by researchers for emotional state identification. Compared to images, speech signals are less complex and contain more

compact information regarding emotions. Additionally, recording speech is easier and more convenient than capturing other signals that require specialized equipment [6].

Speech Emotion Recognition (SER) has garnered significant interest due to its wide range of applications, including call centers, smart TVs, computer games, robot interactions, criminal investigations, and psychological medical diagnosis [7-10]. SER mainly uses machine learning methods to automatically predict correct 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 high breathiness and no periodic excitation, leading to significant alterations in its structure, reduced perceptibility, and decreased intelligibility due to the absence of periodic vocal fold vibrations.



(a) (b) (c) (d)

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

Despite these changes, whispered speech can still encode prosodic information and convey emotional cues [11], [12]. 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. Additionally, whispered speech is crucial for individuals with speech disabilities, such as those affected by temporary or long-term vocal fold issues or vocal system diseases like functional aphonia or laryngeal disorders [13], who can only produce whisper-like sounds. However, only a few studies have focused on recognizing emotions in whispered speech. Past research [14], [15] has mainly analyzed the prosodic feature differences in emotions of Chinese whispered speech. Therefore, 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 proposed method for emotion recognition from whispered speech leverages multiple features and ensemble learning to enhance accuracy and robustness. The approach 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 selection process is employed to identify the most informative features and to reduce dimensionality while preserving essential emotional cues. Finally, classification is performed using an ensemble learning classifier is used to effectively distinguish between different emotional states. The major contributions of this paper are outlined as follows;

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

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): Reduces 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. Section 4 presents the experimental analysis, demonstrating the effectiveness of the proposed method through various tests and evaluations. Finally, Section 5 concludes the paper by summarizing the findings, discussing their implications, and suggesting directions for future research.

2. LITERATUE SURVEY

In the past, several method shave been proposed for the recognition of emotions form whispered speech through several strategies. For instance, Y. Bhavani e al. [16] 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 provide a comprehensive overview to guide future research efforts in SER. Z. Cheng and X. Li [17] 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. [18] 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 potential of acoustic feature transfer learning in enhancing emotion recognition systems and opens new avenues for research in whispered speech analysis. It involves extracting a comprehensive set of acoustic features, including Mel-frequency Cepstral coefficients (MFCCs), spectral features, and temporal dynamics, and then fine-tuning a deep neural network on a

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

whispered speech dataset annotated with emotional labels. Zhaofeng Lin et al. [19] 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. [20] 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. [21] 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 refers to the rate of change of phase with respect to frequency and can provide valuable information about speech signals, including whispered speech. This research investigate 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. [22] 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 enhance the accuracy and robustness of whispered speech recognition systems.

Sharma, V., et al. [23] explored the techniques for converting whispered speech into normal speech using the inversion of Mel Frequency Cepstral Coefficients (MFCCs). Whispered speech differs significantly from normal speech due to the absence of voicing and altered acoustic properties. MFCCs are commonly used to represent the spectral envelope of speech signals. In this context, the inversion of MFCC features involves reconstructing or transforming whispered speech signals to sound more like normal speech. The study investigate 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. [24] 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. [25] 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. [26] investigates 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 [27] 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. MFCC, which are commonly used to represent the spectral envelope of speech signals, and IMFCC, which capture complementary spectral information, can be fused to enhance the robustness and accuracy of whispered speech recognition. Zhang, Li., et al. [28] proposed a method for improving whispered speech recognition by integrating Deep Denoising Autoencoder (DDAE) and Inverse Filtering techniques. 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.

3. PROPOSED APPROACH

3.1 Overview

As illustrated in Figure 2, the proposed method consists of three main phases: pre-processing, feature extraction, and classification.

Pre-Processing: The input speech signal undergoes segmentation to divide it into short-time segments. 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. The features with reduced dimensions are then concatenated to form a final feature vector.

Classification: The final feature vector is processed through an ensemble classifier. This ensemble consists of two classifiers: a Support Vector Machine (SVM) and a Decision Tree. The SVM

differentiates between emotional and neutral states, while the Decision Tree further categorizes each emotion into distinct branches, ultimately identifying the specific emotion.

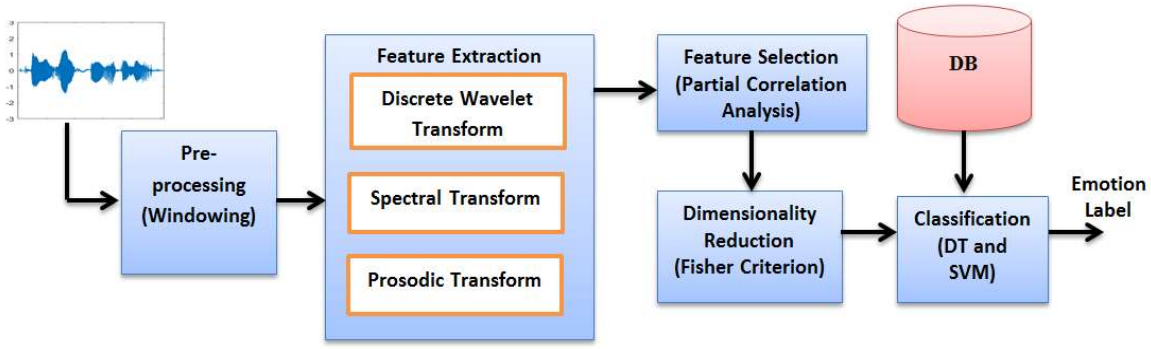


Figure.2 Overall block diagram of proposed method

3.2 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 used a coefficient of 0.97 and it is done as

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

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.

3.2.1 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 measured using the autocorrelation method, with a pitch range of 75–300 Hz for males and 100–500 Hz for females. Intensity, or energy, refers to the loudness of speech. Both pitch and intensity are strongly correlated with a speaker's emotional state, making them valuable for speech emotion recognition [29]. Additionally, statistical features such as mean, standard deviation, maximum, minimum, and range are computed to reflect variations in pitch and intensity.

3.2.2 Spectral Features

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

Formants represent the resonance frequencies of the vocal tract where peaks of high energy occur. They vary with emotion, making them useful for speech emotion recognition [30].

MFCCs are derived from a nonlinear Mel-scale and highlight the significance of low-frequency components relative to high-frequency ones. They are commonly used in speaker and speech recognition systems due to their ability to mimic the human auditory system, being sensitive to sound variations at lower frequencies.

LTAS reflects the logarithmic signal power density of the voiced parts of the speech signal and adjusts for pitch effects. It is less computationally complex compared to MFCCs [31], [32].

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, maximum, and minimum.

3.2.3 Wavelet Features

Wavelet transform is a multi-resolution analysis method used for analyzing acoustic signals [33], [34]. 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. 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 [35].

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.

3.3 Feature Selection

Feature selection is performed to determine the importance of each feature. 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 final set with reduced dimensions.

3.3.1. 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

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

relationship. Consider the group of independent variables as $X = \{x_1, x_2, \dots, x_n\}$, the partial correlation is computed as

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

For the above Matrix the inverse is calculated as

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

Finally the partial correlation between two variables is calculated as

$$\Upsilon_{ij} = \frac{-\lambda_{ij}}{\sqrt{\lambda_{ii}}\sqrt{\lambda_{jj}}} \quad (4)$$

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.

3.3.2 Fisher criterion

In pattern recognition applications, the dimensionality of the feature set can pose several challenges. Methods that operate in lower-dimensional spaces generally have lower computational complexity and can achieve optimal performance. After feature selection, which often yields a large feature set, dimensionality reduction transforms these features into a lower-dimensional space with minimal information loss. One of the primary concerns in dimensionality reduction is preserving information. To obtain an optimal feature set in a reduced-dimensional space, we use the Fisher Criterion, which focuses on linear relationships for dimensionality reduction. While Principal Component Analysis (PCA) is another popular method for dimensionality reduction, it may not effectively capture discriminative information from high-dimensional emotional features. In this work, both PCA and Fisher Criterion are applied, with results demonstrating the superior performance of the Fisher Criterion. Mathematically, the Fisher Criterion [36-38] is calculated as follows:

$$\lambda_F = \frac{\sigma_B}{\sigma_W} \quad (5)$$

Where λ_F is Fisher rate of features, σ_B is the variance between different classes and σ_W is the variance within the class. σ_B is calculated as

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

Where $\bar{E}$ is the mean of the entire data set and is defined as

$$\bar{E} = \frac{1}{M}\sum_{i=1}^{M} x_i \quad (7)$$

And E_c is the sample mean for i^{th} Emotion class E_i, defined by

$$E_c = \frac{1}{N_P}\sum_{x\in E_c} x_i \quad (8)$$

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

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

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

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

3.4 Classification

For classification, we employed two machine learning algorithms they are Support Vector Machine (SVM) and Decision Tree. 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. Following the SVM classification, the Decision Tree algorithm is used for further classification. During the testing phase, if a speech signal is labeled as emotional, it is processed using a binary tree to determine the specific emotion. A simple illustration of this classification process using SVM and Decision Tree is shown in Figure 3.

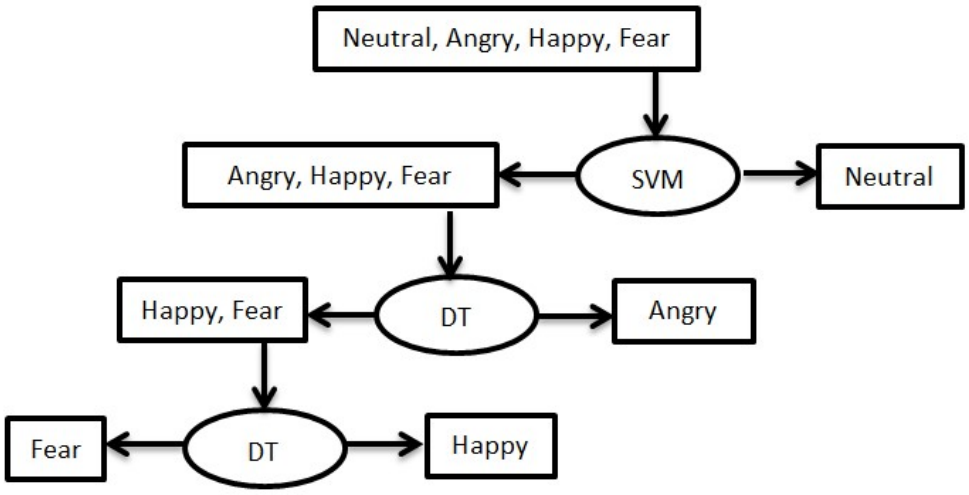


Figure.3 Ensemble Learning Assisted Classification of emotions

4. 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 effectiveness of the proposed approach is measured using A metric called Unweighted Average Recall (UAR). These metrics offer a comprehensive assessment of the model's performance in correctly identifying and classifying emotions. Additionally, the results are compared with state-of-the-art methods to benchmark the proposed approach against existing techniques, highlighting its advantages and areas for improvement.

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

4.1 Dataset and Setup

Geneva Whispered Emotion Corpus (GeWEC) [39] is used here to evaluate the effectiveness of the proposed system. The corpus includes paired utterances of normal phonated and whispered speech. Two male and two female professional French-speaking actors from Geneva were recruited to speak eight predefined French pseudo-words ("*belam", "molen", "namil", "nodag", "lagod", "minad", and "nolan*") in both normal and whispered modes, expressing one of four emotional states: *Angry, Fear, Happiness, and Neutral*. Each actor was asked to express each word in all four emotional states five times, resulting in labeled utterances corresponding to the intended emotion. Consequently, GeWEC comprises a total of 1,280 instances. To provide an in-depth evaluation of the proposed method, we also generated binary valence/arousal labels from the emotion categories. In the valence space, angry and fear are categorized as negative valence, while happiness and neutral are categorized as positive valence. In the arousal space, neutral is categorized as low arousal, whereas angry, happiness, and fear are categorized as high arousal.

4.2 Results

We use Unweighted Average Recall (UAR) as a performance metric, which has also been the competition measure of the first challenge on emotion recognition from speech [40] and follow-up ones. It equals the sum of the recalls per class divided by the number of the classes, and appears more meaningful than overall accuracy in the given case of presence of class imbalance.

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

		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 UAR for Binary Valence in leave-one-speaker-out testing for different train/test combinations.

		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

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

Table.3 UAR for Binary Arousal in leave-one-speaker-out testing for different train/test combinations.

		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 show that when the system is trained and tested on normal phonated speech, it achieves the highest 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.

From the result, it can be seen that the proposed recognition system using multiple features and ensemble learning features performs best when both the training and test data are exclusively drawn from normal phonated speech, achieving the highest UAR of 74.23% for the four-class emotion classification problem. Conversely, whispered speech in a matched condition yields a significantly lower UAR of 46.32%. Training with whispered speech seems to particularly affect the recognition of valence. It appears that using a training set composed of whispered speech would be more effective for whispered speech emotion recognition (i.e., matched condition learning). However, Table 2 indicates that there is no significant reduction in performance when using a system trained with normal phonated speech. Surprisingly, for binary valence and binary arousal, the system trained with normal phonated speech sometimes achieves slightly higher UAR than when trained with whispered speech. Furthermore, multi-condition training is only truly beneficial for whispered speech.

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

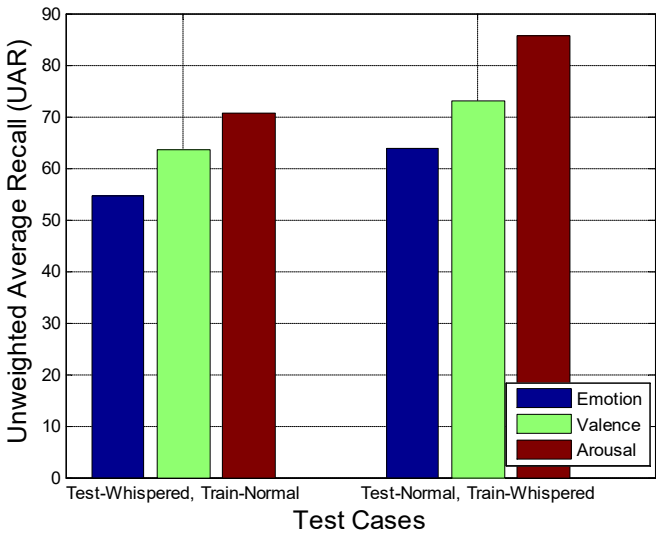
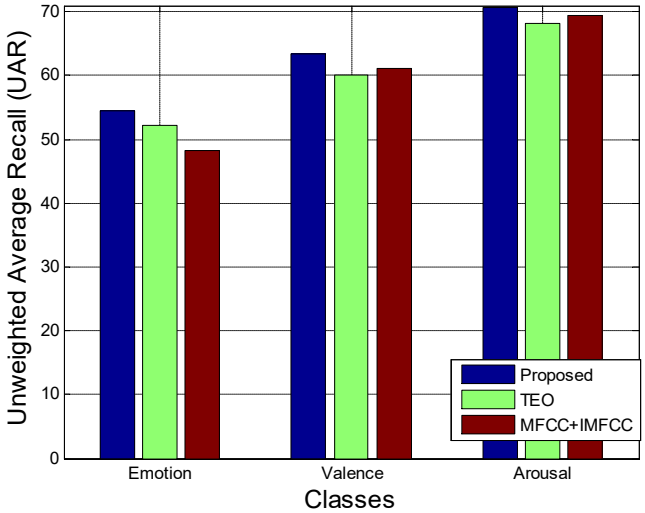


Figure.4 UAR of proposed approach in different test conditions

Figure.4 explores the UAR of proposed approach in two different test conditions; they are (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.



Vol. 21, No. 1, (2024)
ISSN: 1005-0930

Figure.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 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 MFCC and IMFCC method 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.

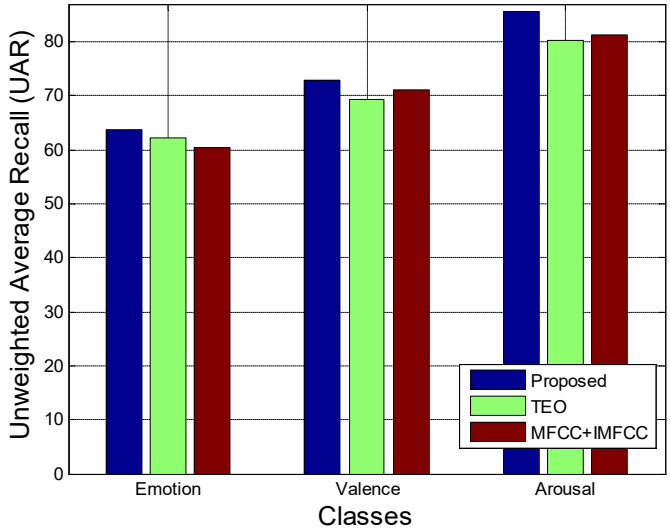


Figure.6 Case Study: Test – Normal, Train – Whispered. UAR comparison between proposed approach and existing methods

Figure.6 shows the UAR comparison between proposed and existing methods at a simulation case study where 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.

5. 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.

References

[1] Swain, M., Routray, A., Kabisatpathy, P., “Databases, features and classifiers for speech emotion recognition: a review”, *Int. J. Speech Technol.* 21, 93–120, 2018.
[2] M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo, “Recognizing emotions induced by affective sounds through heart rate variability”, *IEEE Trans. Affect. Comput.*, vol. 6, no. 4, pp. 385-394, Oct. 2015.
[3] D. Poªap, ``Model of identity veri_cation support system based on voice and image samples," *J. Univers. Comput. Sci.*, vol. 24, pp. 460-474, Jan. 2018.
[4] Thagard, P. , 2019. Mind Society: From Brains to Social Sciences and Professions. Oxford University Press (March 1, 2019) .
[5] Mohammadi, Z., Frounchi, J., Amiri, M., 2017. Wavelet-based emotion recognition system using EEG signal. Neural Comput. Appl. 28, 1985–1990.
[6] Tawari, A., Trivedi, M.M., 2010. Speech emotion analysis: exploring the role of context. *IEEE Trans. Multimed*. 12, 502–509.
[7] 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.
[8] Meddeb, M., Karray, H. , Alimi, A.M., 2017. Content-based Arabic speech similarity search and emotion detection. In: Hassanien, A.E., Shaalan, K., Gaber, T., Azar, A.T., Tolba, M.F. (Eds.), Proceedings of the International Conference On Advanced Intelligent Systems and Informatics, 2016. Springer International Publishing, Cham, pp. 530–539.
[9] Sinith, M.S., Aswathi, E., Deepa, T.M., Shameema, C.P., Rajan, S., 2016. Emotion recognition from audio signals using Support Vector Machine. In: Proceedings of the IEEE Recent Advances in Intelligent Computational Systems, RAICS 2015. IEEE, pp. 139–144.
[10] Likitha, M.S., Gupta, S.R.R., Hasitha, K., Raju, A.U., 2018. Speech based human emotion recognition using MFCC. In: Proceedings of the International Conference on Wireless Communications, Signal Processing and Networking 2017, pp. 2257–2260.
[11] F. H. Knower, ``Analysis of some experimental variations of simulated vocal expressions of the emotions," J. Social Psychol., vol. 14, no. 2, pp. 369-372, 1941.

[12] I. R. Murray and J. L. Arnott, ``Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," J. Acoust. Soc. Amer., vol. 93, no. 2, pp. 1097-1108, 1993.
[13] P. Mitev and S. Hadjitodorov, ``Fundamental frequency estimation of voice of patients with laryngeal disorders," Inf. Sci., vol. 156, nos. 1-2, pp. 3-19, 2003.
[14] Y. Jin, Y. Zhao, C. Huang, and L. Zhao, ``Study on the emotion recognition of whispered speech," in Proc. GCIS, Xiamen, China, 2009, vol. 3, pp. 242-246.
[15] G. Chenghui, Z. Heming, Z. Wei, W. Yanlei, and W. Min, ``A preliminary study on emotions of Chinese whispered speech," in Proc. IFCSTA, Chongqing, China, 2009, vol. 2, pp. 429-433.
[16] 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.
[17] 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.
[18] 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,
[19] 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.
[20] Buayai, P., Uthansakul, M., & Uthansakul, P. (2022). Whispered Speech Detection Using Glottal Flow-Based Features. *Symmetry*, 14(4), 777
[21] 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.
[22] 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*
[23] 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.
[24] Sivan, D., & Gopakumar, C. (2017). Emotion recognition and spoof detection from whispered speech. *2017 International Conference on Computing Methodologies and Communication (ICCMC)*.
[25] 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.
[26] Sung-Chul Ko , Young Sik, & Kyu-Young Kim (2016). Exploitation of phase-based features for whispered speech emotion recognition. *IEEE Access, 4*, 6074-6082.
[27] 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
[28] 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.

[28] Rabiner LR, Schafer RW. “Digital processing of speech signals”. *Pearson Educ (Singapore) Pte. Ltd*., (Indian reprint) 2004.
[29] Meftah, A. , Alotaibi, Y. , Selouani, S.-.A. , 2014. “Designing, building, and analyzing an Arabic speech emotional corpus”. *Work. Free. Arab. Corpora Corpora Process. Tools Work*. Program. 22
[30] 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.
[31] Bahmanbiglu, S.A., Mojiri, F., Abnavi, F., 2017. “The Impact of Language on Voice: an LTAS Study”. *J. Voice* 31 (249).
[32] Yüksel, M., Gündüz, B., 2018. “Long term average speech spectra of Turkish”. *Logop. Phoniatr. Vocology* 43, 101–105
[33] 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.
[34] 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.
[35] Coifman, R.R., Wickerhauser, M.V., 1992. “Entropy-based algorithms for best basis selection”. *IEEE Trans. Inf. Theory* 38, 713–718.
[36] W. Malina , “On an extended fisher criterion for feature selection”, *IEEE Trans. Pattern Anal. Mach. Intell*. 5 (1981) 611–614.
[37] J. Yang , J. Yang , “Why can LDA be performed in PCA transformed space?”, *Pattern Recogn*. 36 (2) (2003) 563–566.
[38] S.Q. Zhang , B.C. Lei , A.H. Chen , “Spoken emotion recognition using local fisher discriminant analysis”, in: *Proceedings of the Tenth IEEE International Conference on Signal Processing Proceedings*, 2010, pp. 538–50.
[39] Bänziger T, Mortillaro M, Scherer KR. Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. Emotion. 2012 Oct;12(5):1161-79. doi: 10.1037/a0025827. Epub 2011 Nov 14.
[40] B. W. Schuller, S. Steidl, and A. Batliner, ``The INTERSPEECH 2009 emotion challenge," in Proc. INTERSPEECH, Brighton, U.K., 2009, pp. 312315.