Conference Paper Title*

*Note: Sub-titles are not captured in Xplore and should not be used

Manjiri Bane
Department of Information Technology
DJSCE
Mumbai, India
manjiribane0224@gmail.com

Minori Wakade
Department of Information Technology
DJSCE
Mumbai, India
minoriwakade@gmail.com

Manav Pathak
Department of Information Technology
DJSCE
Mumbai, India
manav.pathak04@gmail.com

Krish Mehta
Department of Information Technology
DJSCE
Mumbai, India
krishmehta986@gmail.com

Monika Mangla
Department of Information Technology
DJSCE
Mumbai India
email address or ORCID

Sharvari Patil
Department of Information Technology
DJSCE
Mumbai India
email address or ORCID

***Abstract*—This paper proposes a Speech emotion recognition (SER) system focused on identifying human emotions from speech signals. In this study, the RAVDESS dataset is used, with emotions grouped into two categories: Positive (happiness, calm, surprise, neutral) and Negative (sadness, anger, fear, disgust). To capture both spectral and prosodic aspects of speech, five acoustic features are extracted: Mel Frequency Cepstral Coefficients (MFCC), Chroma, Mel Spectrogram, Spectral Contrast, and Tonnetz. The extracted features are reduced using an autoencoder and then classified using Random Forest and Extreme Gradient Boosting (XGBoost) Models. The results show that Random Forest reaches 71.88%, while XGBoost achieves an accuracy of 78%. These results highlight the effectiveness of combining acoustic features with tree-based models for emotion recognition.**

***Index Terms*—Speech Emotion Recognition, Acoustic Features, Classification Models, Machine Learning.**

I. Introduction

Speech is one of the most direct and natural forms of human communication. In addition to transmitting words, it conveys emotional states that reveal the speaker's confidence, engagement, and attitude. These emotional cues play a critical role in contexts such as public speaking, interviews, and classroom learning. Consequently, speech emotion recognition (SER) has become an important area of research in recent years.

Automatic recognition of emotions from the voice has promising applications in the context of public speaking. Speakers can receive feedback on their emotional tone, helping them improve their speech delivery.

Emotions influence not only physiological responses such as pulse, brain waves, or posture, but also translate in speech through acoustic features. Since acoustic signals can be captured directly, they remain a practical and widely used modality for SER.

In this paper,we present a speech emotion recognition system suited for public speaking and educational purposes. Using the RAVDESS dataset, we extract acoustic features including MFCC, Chroma, Mel spectrogram, Spectral Contrast, and Tonnetz. To evaluate performance, we implement and compare Random Forest and XGBoost classifiers. Our findings highlight the potential of ensemble methods for classifying emotions relevant to public speaking.

II. Related Work

A. Literature

Speech Emotion Recognition (SER) methodologies generally fall into two categories: (1) basic feature extraction followed by traditional classifiers, and (2) deep learning systems learning directly from raw audio or spectrograms. The first is highly relevant in settings with limited data or where clarity and computational efficiency are vital.

1) Literature based on Machine Learning: An IEEE review titled *"A Comprehensive Review of Speech Emotion Recognition Systems,"* [2] emphasizes that the combination of spectral and prosodic features with conventional classifiers continues to be widely adopted and effective in computing research. Building on this foundation, recent work at Stanford applied machine learning methods on the RAVDESS dataset [7], demonstrating that standardized datasets are essential for benchmarking SER models and that traditional algorithms can capture emotional patterns effectively. Furthermore, Vu et al. in *"Amplifying Emotional Signals: Data-Efficient Deep Learning for Robust Speech Emotion Recognition,"* [1] proposed a deep learning framework for SER to address the challenges of limited datasets. By employing a ResNet34 model with transfer learning and data augmentation, their system achieved an accuracy of 66.7% and an F1-score of 0.631 on a combined RAVDESS and SAVEE dataset. This study highlights how data-efficient deep learning methods can bridge the gap between classical machine learning approaches and modern end-to-end architectures.

Identify applicable funding agency here. If none, delete this.

2) Literature based on Deep Learning: Khalil et al. [4], in their survey paper, provided a comprehensive overview of deep learning methods applied to SER. The review discusses CNNs, RNNs, DNNs, and hybrid models, as well as the transition from hand-crafted acoustic features (e.g., MFCC, pitch, energy) to end-to-end deep learning using spectrograms and raw waveforms. It also highlights key challenges such as limited dataset size, class imbalance, cross-corpus generalization, and the high computational demands of deep architectures. Building on this direction, Gao, Shi, Chu, and Kawahara [3] proposed a model that integrates acoustic and semantic information with a strong emphasis on dynamic cross-modal interaction. Their approach leverages HuBERT-large as an acoustic feature extractor and a BERT-based semantic extractor, integrated via a cross-modal gated interaction (CmGI) module with a temporal-aware gated fusion mechanism. This design improved unweighted accuracy by 3.32% on IEMOCAP [8] under speaker-independent 5-fold cross-validation, demonstrating state-of-the-art performance. However, unlike the broader but less computationally demanding approaches discussed by Khalil et al. [4], Gao et al.'s system is more resource-intensive and relies on ASR outputs, which may introduce error sources.

Together, these studies underscore the growing importance of acoustic–semantic fusion in SER while highlighting the trade-offs between robustness, efficiency, and computational complexity. They also underline that both feature-based and deep learning–based approaches are valuable, and motivate our work on acoustic feature extraction and machine learning for public speaking analysis, as they are efficient with limited datasets and computationally lightweight compared to deep-learning systems.

B. Extending Existing Approaches

Along with emotion recognition, our work extends to incorporate speech pacing analysis. With the help of an automatic speech-to-text transcription step, we estimated words per minute (WPM), a widely recognized parameter for evaluating speech pace in communication research [5], [6]. This enables the system to categorize speech as "Slow" , "Normal" and "Fast" and provide quick feedback alongside emotion based insights. The integration of pacing complements emotion cues, making the system valuable specifically for public speaking, where both emotional parameter and delivery style play a crucial role.

C. Future Directions within Research Work

As an extension to the current research, considering directions highlighted by prior studies are helpful. These directions not only strengthen the robustness but also help to expand its application to real-world scenarios.

- Multimodal fusion (e.g. combining audio with visual cues) improves robustness.
- Hybrid systems (e.g. mixing extracted features with learned embeddings) , offers adaptability.
- Training on larger emotion datasets (eg. IEMOCAP [8]) boosts model generalization and performance.

III. Methods

A. Feature Extraction

In this work, five complementary acoustic features were extracted: Mel Frequency Cepstral Coefficients (MFCC), Chroma, Mel Spectrogram, Spectral Contrast, and Tonnetz. Together, these features capture spectral, prosodic, and tonal information essential for emotion recognition and public speaking analysis, including aspects such as clarity and intensity control. These features were selected for extraction following prior studies [9]

- **MFCC:** MFCCs approximate human auditory perception by mapping linear frequency f (Hz) to the Mel scale. Applying a discrete cosine transform (DCT) to log-scaled Mel spectra yields compact representations of vocal tract characteristics, capturing timbre and formant patterns.
- **Chroma:** Chroma features map spectral energy to 12 pitch classes. The chroma coefficient for pitch class p at time t is computed from the spectrogram energy $X(f, t)$. These features reflect pitch and intonation patterns relevant to expressiveness.
- **Mel Spectrogram:** Derived from Short-Time Fourier Transform (STFT) magnitudes projected onto the Mel scale, the Mel spectrogram preserves localized spectral energy patterns, which are informative of stress, rhythm, and intensity.
- **Spectral Contrast:** Spectral contrast measures the difference between peaks and valleys in sub-bands. This feature captures vocal timbre, helping distinguish resonant emotions (e.g., anger) from flatter speech (e.g., sadness).
- **Tonnetz:** Tonal centroid features project audio onto harmonic relations such as fifths and intervals. In speech, Tonnetz captures tonal balance shifts associated with emotional tone and stress.

TABLE I
Feature Types and Corresponding Dimensionality

Feature Type	**Dimension**
MFCC	13
Chroma	12
Mel Spectrogram	128
Spectral Contrast	7
Tonnetz	6
Total	**166**

By concatenating these features, each utterance is represented as a multidimensional vector, encoding spectral detail, prosody, and tonal information. This representation enhances emotion recognition accuracy and provides cues relevant for evaluating public speaking performance.

B. Speech Pacing Analysis

To assess the delivery speed of speakers, a module was implemented to calculate Words Per Minute (WPM). The audio uploaded is first processed using the Google Speech Recognition API, which transcribes the speech into text. The WPM is then computed using the formula:

$$\text{WPM} = \frac{\text{Word Count from Transcript}}{\text{Speech Duration (Minutes)}} \tag{1}$$

The speech duration is obtained through Librosa's time-domain processing, while the transcription provides the total word count. Based on communication research and best practices referenced in the ”Extending Existing Approaches” part of the ”Related Work” Section, we classified 130 – 165 WPM as the *Normal* speech range. Speeches below 130 WPM were labeled *Slow* and those above 165 WPM were labeled *Fast*.

This pacing classification is particularly significant in the context of public speaking. Slow speech may cause disengagement, while fast delivery may reduce clarity and affect audience comprehension. By providing feedback on speech pacing within our system, users can enhance their delivery style to remain within the defined range, thereby improving both clarity and audience engagement.

C. Preprocessing

Post feature extraction, we preprocessed the dataset to prepare it for classification. The target variable, sentiment, was categorical and it was encoded into numeric labels using the LabelEncoder. The dataset was then divided into training and testing sets in 80:20 ratio, ensuring class categorization to preserve the distribution of available classes across both subsets.

Based on the extracted acoustic features (MFCC, Chroma, Mel Spectrogram, Spectral Contrast, Tonnetz), we applied standardization using StandardScaler. The scaler was fitted on the training data and further applied to both training and testing sets to prevent leakage of data. Standardization assured zero-mean, unit-variance features, thus improving model convergence and classification performance.

D. Classification Models

In this system, we employed two ensemble-based machine learning algorithms for sentiment and speech emotion classification: Random Forest (RF) and Extreme Gradient Boosting (XGBoost).

1) ***Random Forest (RF):*** RF is an ensemble method that builds multiple decision trees using bootstrap samples of the data and averages their predictions. This reduces variance and helps avoid overfitting. In our work, the RF model was trained with 300 decision trees and a maximum depth of 20. RF was selected due to stability, it also works well with high-dimensional features, and provides reliable classification performance.

2) ***Extreme Gradient Boosting (XGBoost):*** XGBoost is a boosting-based ensemble method that builds decision trees sequentially, where each new tree focuses on correcting the errors of the previous ones. It is known for efficient results, regularization techniques, and strong performance in classification tasks. In this work, the XGBoost model was trained with 500 trees, a maximum depth of 8, a learning rate of 0.05, and both subsampling (0.8) and column sampling (0.8) strategies to achieve generalization.

Ensemble methods such as RF and XGBoost are particularly effective for speech-related tasks.

- Ability to capture nonlinear relationships in acoustic features.
- Less sensitive to noise in speech signals.
- Notably better and stable predictions than single models.
- Balanced predictive accuracy with clarity.

By applying both Random Forest (bagging) and XGBoost (boosting), we compared two complementary ensemble learning approaches for this application.

E. System Workflow

The system integrates well-defined modules that support efficient emotion recognition and speech pacing evaluation.

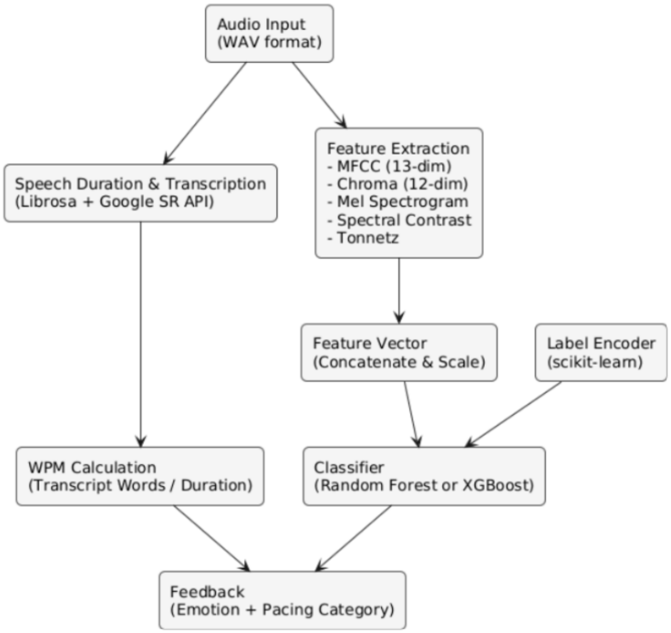

Fig. 1. Balanced workflow of system

The system emphasizes on robust emotion classification. By leveraging selected acoustic features that effectively capture emotional cues, it minimizes overhead computations without compromising accuracy. Ensemble classifiers provide model stability and transparency, thus facilitating straightforward analysis and retraining. This design supports practical deployment with clear and reliable feedback.

IV. Experiments and Results

A. Dataset Description

For this work, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [7] was used, focusing on the speech-only portion. The dataset contains 1,440 audio samples recorded by 24 professional actors (12 female, 12 male), each delivering two lexically-matched statements in a North American accent. Emotional expressions include neutral, calm, happy, sad, angry, fearful, disgust, and surprised, with each emotion produced at normal and strong intensity levels (except neutral).

To simplify the classification task for our public speaking application, the original eight emotions were mapped into two sentiment categories:

- **Positive:** Neutral, Calm, Happy, Surprised
- **Negative:** Sad, Angry, Fearful, Disgust

TABLE II
RAVDESS FILE NAMING CONVENTION

Part	Description
1	Modality
2	Vocal channel
3	Emotion
4	Intensity
5	Statement Number
6	Repetition
7	Actor ID

Each file is uniquely identified using a seven-part numeric code, encoding modality, vocal channel, emotion, intensity, statement, repetition, and actor, as shown in Table II.

An example filename from the RAVDESS dataset is:
`03-01-05-02-02-01-12.wav`
where:

- **03** → Modality (audio-only)
- **01** → Vocal channel (speech)
- **05** → Emotion (happy)
- **02** → Intensity (strong)
- **02** → Statement number (2nd statement)
- **01** → Repetition (first repetition)
- **12** → Actor ID (12th actor)

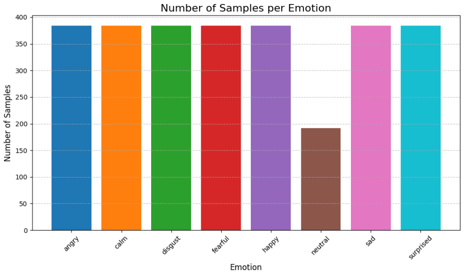

Fig. 2. Distribution of audio samples for each emotion in the RAVDESS dataset.

This structure ensures that each audio sample can be uniquely identified, which is crucial for dataset management and model training. The distribution of audio samples across the eight emotions is shown in Fig. 2, provides a clear overview of dataset balance and the variety of emotional expressions captured. This visualization ensures that the dataset is suitable for training and evaluating speech emotion recognition models for applications providing real-time feedback to speakers or learners.

B. Results and Analysis

*1) **Random Forest (RF)**:* : This model showcased reliable performance in classifying speech sentiment, correctly identifying approximately 71.88% of the test samples. The ensemble nature of RF, which aggregates predictions from multiple decision trees, contributed to its stability and reduced overfitting. Its precision and recall values indicate slightly better detection of Negative speech, suggesting that the model is more sensitive to cues of emotional intensity, such as sharper spectral variations or prosodic shifts. However, performance on Positive speech, while reasonably accurate, lagged slightly, reflecting challenges in capturing subtle acoustic markers of positivity. Despite this, the balanced overall F1-score highlights RF's ability to generalize well across both sentiment categories, making it a strong baseline for comparison against more advanced models.

TABLE III
RANDOM FOREST PERFORMANCE METRICS

Class	Precision	Recall	F1-score	Support
Negative	0.73	0.75	0.74	154
Positive	0.70	0.69	0.69	134
Overall	0.72	0.72	0.72	288

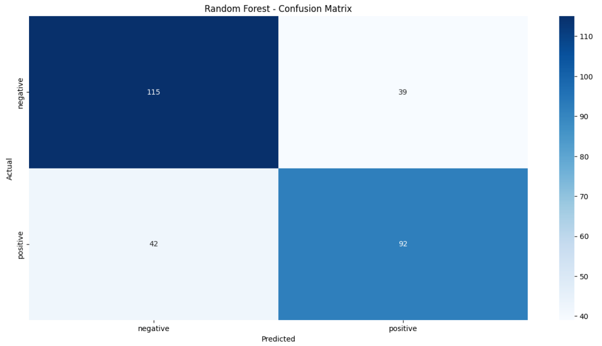

Fig. 3. Confusion Matrix for Random Matrix

*2) **Extreme Gradient Boosting (XGBoost)**:* This model achieved stronger classification performance, correctly predicting approximately 78% of the samples. Unlike Random Forest, which builds trees independently, XGBoost leverages gradient boosting to sequentially construct decision trees, with each tree correcting the errors of its predecessors. This iterative refinement enables the model to capture complex, nonlinear relationships among the acoustic features, such as subtle prosodic variations and spectral patterns that distinguish Positive from Negative speech. The higher precision and recall scores for both classes underscore its robustness in handling imbalanced feature cues and noisy data. Furthermore, the relatively balanced F1-scores highlight its effectiveness across sentiment categories, minimizing bias toward either Positive or Negative labels. The improvement over RF demonstrates the advantage of boosting, not only in raw accuracy but also in enhancing generalization, making XGBoost a more powerful approach for speech sentiment recognition in constrained data environments.

TABLE IV
XGBoost Performance Metrics

Class	**Precision**	**Recall**	**F1-score**	**Support**
Negative	0.81	0.79	0.80	154
Positive	0.76	0.78	0.77	134
Overall	0.78	0.78	0.78	288

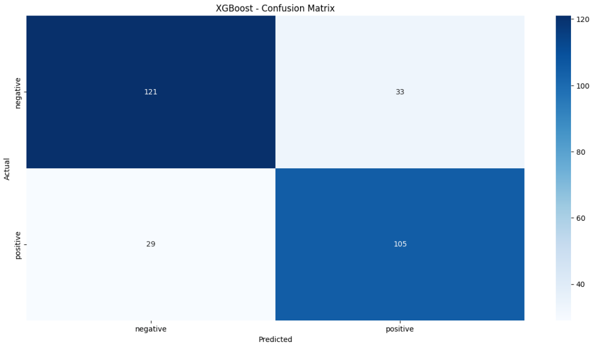

Fig. 4. Confusion Matrix for XGBoost

C. Comparative Discussion

XGBoost outperformed Random Forest by approximately 6.6% in accuracy.It is more capable to capture complex, nonlinear relationships among acoustic features while iteratively correcting errors. Random Forest, though providing stable predictions, is inferior to the results of XGBoost. Both models successfully leverage the acoustic feature set, highlighting the effectiveness of ensemble learning for SER.

Further, the integration of Words Per Minute (WPM) classification complements emotion detection. While emotion analysis educates users about the tone of their speech, WPM categorization offers insight on delivery pace—ensuring clarity, engagement, and overall presentation effectiveness.

V. Conclusion

In this paper, we presented a Speech Emotion Recognition (SER) system leveraging the RAVDESS dataset, where eight emotions were assigned to Positive and Negative sentiment categories. The following five acoustic features namely MFCC, Chroma, Mel Spectrogram, Spectral Contrast, and Tonnetz were extracted to capture both spectral and prosodic characteristics of speech. Two ensemble-based models, Random Forest (RF) and Extreme Gradient Boosting (XGBoost), were evaluated for classification. While the RF model achieved an accuracy of 71.88%, the XGBoost model outperformed it with 78% accuracy, highlighting the effectiveness of boosting in handling complex and nonlinear relationships in acoustic features.

A distinct contribution of this work is the integration of Words Per Minute (WPM) analysis, which enables classification of speech pacing as slow, normal, or fast. This extension along with predicted sentiment offers valuable feedback for public speaking training.

VI. Future Work

Based on the obtained results, we have identified directions for our future work to enhance the performance, robustness, and applications of the proposed SER system.

A. Dataset Expansion and Diversity

Immediate step to diversify the model is to train on larger and varied emotional speech datasets, such as IEMOCAP and MELD. These datasets provide better emotional content, multimodal samples, and more generic speech, which could improve the system's generalization and performance.

B. Deep Learning Architectures

While ensemble learning methods provide partial effectiveness, future work must involve deep neural networks, including Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and transformer-based architectures. These models have the potential to automatically learn high-level acoustics directly from raw audio or even through spectrograms, thus performing better than feature-engineered pipelines.

C. Multimodal Emotion Recognition

Extension of the system to incorporate visual cues, such as facial expressions, gaze, gestures, and posture, is a important step towards making the system adapt to real-world usage. Multimodal fusion of speech and body language parameters provide holistic assessment of emotional state and communication style, making the system more useful for training and real-world coaching.

By pursuing these directions, the SER system can evolve into a comprehensive multimodal communication coaching tool, capable of delivering personalized feedback to enhance overall delivery quality.

References

[1] T. Vu, "Amplifying Emotional Signals: Data-Efficient Deep Learning for Robust Speech Emotion Recognition," *arXiv preprint arXiv:2509.00077*, 2025.

[2] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A Comprehensive Review of Speech Emotion Recognition Systems," *IEEE Access*, vol. 9, pp. 47795–47820, 2021, doi: 10.1109/ACCESS.2021.3068045.

[3] Y. Gao, H. Shi, C. Chu, and T. Kawahara, "Speech Emotion Recognition with Multi-level Acoustic and Semantic Information Extraction and Interaction," in *Proc. Interspeech*, Kos, Greece, Sep. 2024, Paper ID: 2385, doi: 10.21437/Interspeech.2024-2385.

[4] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019, doi: 10.1109/ACCESS.2019.2936124.

[5] M. Rao, “Finding how many words per minute is ideal for an effective speech at work,” *Prezent.ai Blog*, Sep. 17, 2024. [Online]. Available: https://www.prezent.ai/blog/words-per-minute-speech

[6] M. Pujadas-Farreras, “The effect of speech rate on easy language audios in Catalan: comprehension and acceptability of different speech rates,” *System*, 2025, doi: 10.1016/j.system.2025.100921.

[7] S. R. Livingstone and F. A. Russo, “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English,” *PLoS ONE*, vol. 13, no. 5, p. e0196391, 2018, doi: 10.1371/journal.pone.0196391.

[8] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, “IEMOCAP: Interactive Emotional Dyadic Motion Capture Database,” *Language Resources & Evaluation*, vol. 42, no. 4, pp. 335–359, Dec. 2008, doi: 10.1007/s10579-008-9076-6.

[9] S. Samyuktha and S. Sarwath Unnisa, “Emotional Speech Recognition Using CNN Model,” *Int. J. Inf. Technol., Res. Appl. (IJITRA)*, vol. 4, no. 1, pp. 30–38, Mar. 2025, doi: 10.59461/ijitra.v4i1.164.