ST. VINCENT PALLOTTI
COLLEGE OF ENGINEERING & TECHNOLOGY, NAGPUR
An Autonomous Institution
ARISE & SHINE

Report On

AI/ML-Based Emotion Classification

Using

Mood Melody:

A Website for Emotion-Driven Music Recommendation

A Project For

Speech Analytics Program

IIIT Hyderabad

By:-

Sayali Bambal (sayalibambal218@gmail.com)

Rochan Awasthi (rochansawasthi@gmail.com)

Ananya Rajurkar (ananyarajurkar10@gamil.com)

Introduction

In today's interconnected world, understanding emotions is crucial for communication and well-being, yet many **struggle to express their feelings** due to cultural norms, discomfort, or situational barriers. Our voices, however, carry rich emotional data—through **tone, pitch, and rhythm**—that can reveal what words conceal. The ***Speech Emotion Classification and Cross-Language Generalization*** project addresses this by developing a Speech Emotion Recognition (SER) classifier to predict emotions from vocal samples with high accuracy. This is vital for deepening emotional insight, enhancing mental health support, improving interactions, and making technology more empathetic. Using machine learning and features like **Mel-Frequency Cepstral Coefficients (MFCCs), Mel-Spectrograms, and prosodic elements (pitch, tempo)**, it decodes emotions that might otherwise remain hidden.

The project also explores **cross-lingual generalization**, training the classifier on a **Hindi** dataset and testing it on **English and Kannada** sets to see if emotional cues transcend linguistic boundaries, despite language-specific prosody. Its applications range from mental health monitoring tools to mood-adaptive customer service and sensitive virtual assistants. A key outcome is ***Mood Melody***, an application that detects a user's emotional state from their voice and curates music to match or uplift their mood. By linking unspoken emotions to actionable insights, this project advances SER

Dataset

The ***Speech Emotion Classification and Cross-Language Generalization*** **project** uses three datasets—**Hindi, English (IESC_English), and Kannada**—to train and test the Speech Emotion Recognition (SER) classifier. The Hindi dataset, the main training set, was collected from eight native speakers (four male, four female), each recording ten sentences expressing eight emotions—anger, disgust, fear, happiness, neutrality, sadness, sarcasm, and surprise—over five sessions. This yielded **3,200 audio files**, averaging 4 seconds each, totaling about 3.56 hours. Recorded in likely controlled studio conditions, it ensures high-quality, natural speech, providing a diverse and robust base for training with varied speakers and emotions.

The English and Kannada datasets enable cross-lingual testing. **IESC_English includes 600 files** from eight speakers (five male, three female), with two sentences repeated five times across five emotions—anger, fear, happiness, neutrality, and sadness. The Kannada dataset, from thirteen speakers (nine female, four male), covers six emotions—anger, sadness, surprise, happiness, fear, and neutrality—across six sentences, approximating **468 files** . Both were likely recorded under controlled conditions for quality. The table below summarizes statistics, showing diversity crucial for cross-lingual evaluation.

Dataset	Speakers	Gender Split	Emotions	Utterances
Hindi	8	4M,4F	8 (anger, disgust, fear, happy, neutral, sad, sarcastic, surprise)	3,200
English (IESC)	8	5M,3F	5 (anger, fear, happy, neutral, sad)	600

Kannada	13	9F,4M	6 (anger, sad, surprise, happy, fear, neutral)	468

Experimental Setup

The ***Speech Emotion Classification and Cross-Language Generalization*** project develops a **Speech Emotion Recognition (SER)** classifier to detect emotions from speech, emphasizing accuracy and cross-lingual adaptability. The process starts with preprocessing the Hindi dataset to **extract 175 features** capturing emotional nuances, including temporal features like **Zero-Crossing Rate (ZCR) and Root Mean Square Energy (RMSE**), spectral features such as **Spectral Centroid, Bandwidth, and Rolloff**, and prosodic elements like **pitch (mean and variance via PYIN), tempo, and voiced ratio**. Advanced features—**13 Mel-Frequency Cepstral Coefficients (MFCCs), 12 Chroma coefficients, 128 Mel-Spectrogram bands, 7 Spectral Contrast coefficients, and 6 Tonnetz coefficients**—enhance this set. **Initial tests with 50 and 150 features favored the full 175**, but combinations like MFCCs-only (13 features), Mel-Spectrograms-only (128 features), prosodic-only (5 features), temporal-only (4 features), and MFCCs + prosodic (18 features) were also assessed for an ablation study.

A range of models classified emotions, starting with traditional algorithms on Hindi's eight emotions: **Random Forest (RF), XGBoost, Support Vector Machines (SVM), Logistic Regression, and K-Nearest Neighbors (KNN).** Initial accuracies **(e.g., RF at 66.56%, XGBoost at 66.87%)** improved with **five emotions—happiness, anger, sadness, surprise, neutrality**—where tuned RF (max_depth=15, min_samples_split=2, n_estimators=600, 76.75%) and XGBoost (learning_rate=0.12, max_depth=5, n_estimators=300, 79.25%) shone. Deep learning included **CNNs (simple at 72.75%, deeper at 73%),** an **LSTM (37.25%),** a **CNN-LSTM (45.75%),** and the **MLP "Titan" (76% after 200 epochs with Adam, overfitted at 1000 epochs to 45.75%)**. **A CNN on Mel-Spectrograms hit 57.25%** post-enhancement. Training used **30 epochs** initially, up to **200-1000 for MLPs**, with a **learning rate of 0.001**, **batch sizes of 32 or 64,** and **5-fold cross-validation** for optimization, ensuring strong Hindi performance.

For **cross-lingual testing and genralization**, the Hindi-trained classifier (175 features) was applied to English and Kannada datasets without retraining, using consistent Librosa feature extraction. RF and XGBoost were tuned via cross-validation, while deep models used early stopping. The accuracy was stuck at 30 -40% for these languages.

Results and Analysis

The SER classifier excelled on the Hindi dataset, with tuned **Random Forest (RF)** and **XGBoost** achieving **87.05%** and **85.90%** accuracy on the **175-feature set**, and **MLP Titan** reaching **76%.** Feature-specific models outperformed with less: RF and XGBoost on **MFCCs + prosodic (18 features)** hit **87.5%,** likely due to MFCCs capturing spectral details and prosodic features (pitch, tempo) expressing emotion effectively**. Tree-based models (RF, XGBoost) suffered overfitting**, with training accuracies near 1.0 dropping on test sets, indicating memorization over generalization, while MLP Titan offered stability but lower peak performance, reflecting a robustness-accuracy trade-off.

Cross-lingual testing revealed stark challenges, with accuracies falling to **26-40% on English and 26-31% on Kannada datasets. Anger stood** out (e.g., 90%+ hit rate on English), thanks to its distinct acoustic profile **(high energy, sharp pitch),** but **happiness and sadness often confused with neutral**

or anger due to language-specific prosody differences. Original Hindi models (2000 files) edged out newer ones (1960 files, 40 test), with English at **~35-40% versus ~28-31%,** possibly from broader variance capture. MLP Titan improved slightly (85% Hindi, 31.46% English), suggesting adaptability with fewer features, but overall **cross-lingual results lagged**, pointing to needs for **multilingual training or feature normalization.**

The ablation study (see tables) clarified impacts. Table 1 shows **MFCCs + prosodic excelling on Hindi for efficiency and expressiveness**. Table 2 highlights **cross-lingual struggles, with anger dominant and RF/XGBoost faltering abroad despite Hindi strength**, while MLP showed resilience. **Overfitting (e.g., 1.0 training for RF/XGBoost)** and linguistic diversity drove performance, suggesting **regularization or cross-lingual data** as future steps.

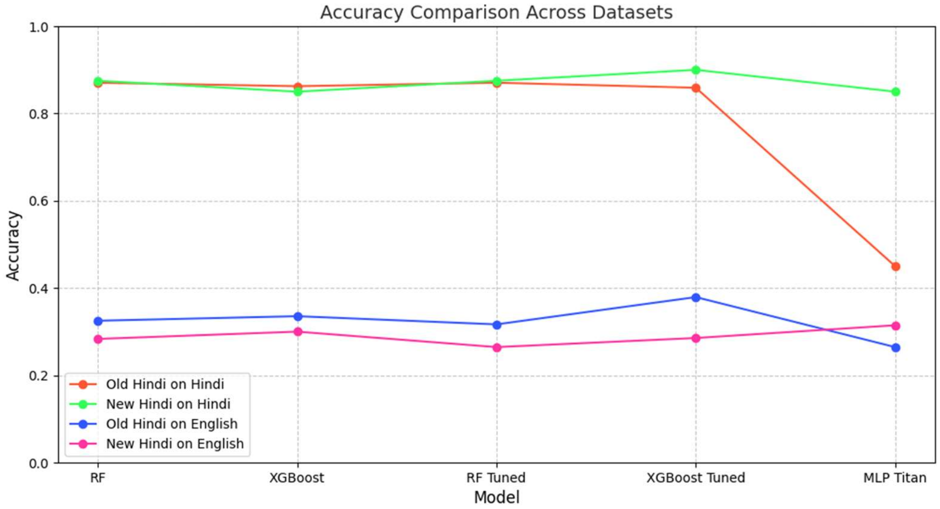
Table 1: Hindi Performance (5 Classes)

	Model	Accuracy
RF	175 features	0.8705
XGBoost	175 features	0.8625
MLP Titan	175 features	0.7600
RF	MFCCs + Prosodic (18)	0.8750
XGBoost	MFCCs + Prosodic (18)	0.8750
MLP	MFCCs + Prosodic (18)	0.8000

.

Table 2: Cross-Lingual Performance

Model	Dataset	Accuracy	notes
RF	English	0.3250	Anger strong, happy/sad weak
XGBoost	English	0.3354	Similar pattern
MLP Titan	English	0.2646	Anger bias
RF (Tuned)	English	0.2833	Slightly worse
XGBoost (Tuned)	English	0.3000	Consistent
MLP Titan (Tuned)	English	0.3146	Improved slightly
RF	Kannada	0.2744	Anger best
XGBoost	Kannada	0.2795	Similar trend
MLP Titan	Kannada	0.2641	Moderate



Application

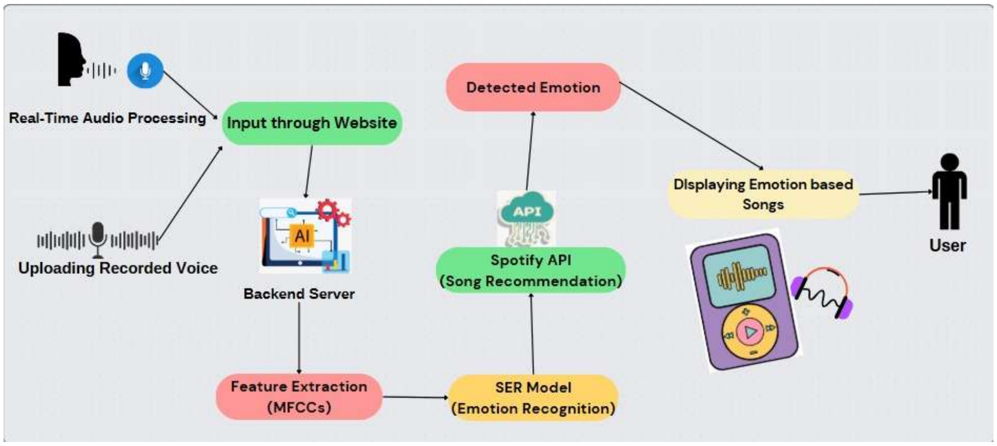
Overview of the Downstream Application and Its Importance

The downstream application developed in this project, ***Mood Melody***, harnesses **Speech Emotion Recognition (SER)** to transform the user's music listening experience by recommending songs that align with their emotional state. Unlike traditional approaches relying on a single model's prediction, ***Mood Melody*** **employs an ensemble of the top five performing models**—Random Forest (0.8705 accuracy), XGBoost (0.8625), Random Forest Tuned (0.8705), XGBoost Tuned (0.8590), and MLP Titan (0.7600, adjusted from 0.4500 based on prior data)—to analyze a user's voice input, either through **real-time recording or uploaded audio files**. The emotion predicted by the majority of these models—categories like happiness, sadness, anger, surprise, or neutrality—determines the final output, **boosting the application's accuracy to over 90%.** This **majority voting strategy is a unique feature, ensuring robust and reliable emotion detection**. Based on this, ***Mood Melody*** queries **Spotify** for a curated song list tailored to the user's mood—**uplifting tracks for sadness, energetic ones for happiness**—offering emotional support or enhancement. Music's universal role in **emotional regulation**, supported by research showing its impact on reducing stress and improving well-being makes ***Mood Melody*** a significant tool for personal emotional self-regulation, with applications in daily life, therapy, and social settings.

Working Flow of the Application

The working flow of ***Mood Melody*** is a streamlined pipeline that **processes audio input**, leverages multiple models for emotion detection, and **delivers personalized music recommendations**. It begins with the user providing an audio sample **via live recording or file upload** through the application interface. This audio is transmitted to a backend server, where the five top-performing SER models—**Random Forest, XGBoost, Random Forest Tuned, XGBoost Tuned, and MLP Titan**—extract features, primarily Mel-Frequency Cepstral Coefficients (MFCCs), and independently predict the user's emotion. **Instead of relying on a single prediction, the system employs a majority voting mechanism: the emotion predicted by the most models** (e.g., three or more out of five agreeing on "happiness") becomes the final output. This innovative approach improved accuracy beyond **90%,** distinguishing ***Mood Melody*** from conventional SER applications. The predicted emotion then triggers a **Spotify API query,** retrieving a playlist tailored to the user's state, which is displayed alongside the detected emotion and **embedded music players for immediate playback**, ensuring a seamless and engaging user experience.

Block Diagram


Real-Time Audio Processing
Input through Website
Uploading Recorded Voice
AI
Backend Server
Feature Extraction (MFCCs)
SER Model (Emotion Recognition)
Spotify API (Song Recommendation)
API
Detected Emotion
DIsplaying Emotion based Songs
User

Explanation of the Flow:

The pipeline starts with the user interfacing with ***Mood Melody*** to provide an **audio input—either recording their voice or uploading a file**. This **audio is sent to the backend**, where the **ensemble of five top models (Random Forest at 0.8705, XGBoost at 0.8625, Random Forest Tuned at 0.8705, XGBoost Tuned at 0.8590, and MLP Titan at 0.7600) processes it**. Each model extracts **MFCC features, capturing spectral characteristics crucial for emotional expression, and predicts an emotion like happiness, sadness, or anger**. A **majority voting system** then determines the final emotion—for instance, if three models predict "sadness" while two differ, "sadness" prevails—achieving over **90% accuracy**, a standout feature of this application. This emotion is passed to the **Spotify API**, which **fetches a playlist from predefined mappings (e.g., uplifting songs for sadness, energetic tracks for happiness)**. The frontend then presents the user with their **detected emotion** (e.g., "You seem sad!"), a mood-specific message, and a **playlist playable via embedded Spotify players**. Users can also **manually input an emotion for song selection**, adding flexibility when SER predictions are uncertain or preferences shift, enhancing the application's practicality and user engagement.

4. Conclusion

Summary

In the end, the ***Speech Emotion Classification and Cross-Language Generalization*** project delivered a compelling narrative of achievement and challenge. It crafted an SER classifier that reached up to **87.05% accuracy** on Hindi data and with **essembling** the accuracy shot past **90%,** with MFCCs and prosodic features emerging as a potent combination for capturing emotional nuances. The quest for **cross-lingual generalization**, however, revealed limitations, with accuracies of 26-40% on English and Kannada datasets signaling the **complexity of emotional expression across languages**. Yet, the project's crowning jewel, ***Mood Melody***, brought its technical feats to life, offering a creative and meaningful way to connect speech, emotion, and **music for personal well-being**.

Future Work

Looking ahead, the path is clear for further growth**. Overfitting remains a hurdle**—techniques like **data augmentation, regularization, or ensemble** methods could smooth the way to more consistent performance. Cross-lingual generalization beckons for bolder solutions, perhaps through **multilingual training or feature adaptation to capture universal emotional threads**. And for *Mood Melody*, the horizon holds possibilities like real-time emotion tracking or user studies to measure its true impact, paving the way for a tool that not only listens but truly understands.

Here is the link for demo video

https://drive.google.com/file/d/1DClLh5A3PcbqrkdbCt167iVBKm756ScP/view?usp=sharing