Ganpat University
॥ विद्यया समाजोत्कर्षः ॥

University Journal of Research: Computer Science and Applications

An Annual, Peer-reviewed Research Journal of FCA
UJRCSA
University Journal of Research : Computer Science and Applications

Speech Emotion Recognition: A Machine Learning Framework Utilizing MFCC Features and the RAVDESS Dataset

Srushti Hareshbhai Lathiya[a], Maitri Piyushkumar Thakkar[b], Dr. Dhara Ashish Darji[c],

[a]Student, FCA, Ganpat University, Mehsana-384012, India

[b]Student, FCA, Ganpat University, Mehsana-384012, India

[c]Assistant Professor, FCA, Ganpat University, Mehsana-384012, India

Corresponding Author Email: dnd01@ganpatuniversity.ac.in

Abstract

The aim of this paper is to design and evaluate a machine learning framework for speech emotion recognition using the RAVDESS dataset. This framework uses Mel Frequency Cepstral Coefficients (MFCC) as audio features to identify and categorize emotions such as happiness, sadness, anger, and surprise. It uses comparative analyses on a host of machine learning algorithms: SVM, Logistic Regression, Random Forest, and Decision Tree to discern their ability towards classifying the emotion. This indicates that, based on SVM models, this classification shows its peak performance in having a high accuracy at about 85.4%.

Keywords: SER, SVM, MFCC, RAVDESS, Librosa

1. Introduction

Speech is not only a communication tool. It is an excellent medium through which people can demonstrate their emotions. The understanding and interpretation of these from speech influence various directions, most definitely in health care, human-computer interaction, customer service, and entertainment. SER focuses on developing the ability to classify vocal signals into happiness, sadness, anger, and surprise, among others. This can make human interactive systems more responsive and sensitive, making them more effective.

Advances in machine learning and signal processing have allowed scientists to develop sophisticated models that can decode emotions in speech signals. A good number of features applied in SER indicate that Mel Frequency Cepstral Coefficients, or MFCC for short, constitute some fundamental representation of audio signals. MFCCs pick the nature of human speech since they model how the ear perceives sound, and most are also very effective for emotion classification tasks. The feature of converting an audio signal into a compact feature space allows MFCCs to pick up very subtle emotional cues hidden in speech.

The RAVDESS dataset is a more robust resource for training and testing an SER model. RAVDESS presents a collection of professionally recorded emotional speech episodes by professional actors. This promises a high degree of realism and variety in data. This dataset could help in significant experimentation with the development and testing of frameworks using machine learning approaches that could be developed to be successfully developed for the recognition of emotions from speech.

This paper introduces a learning framework in which MFCC features extracted from the RAVDESS dataset are applied to classify emotions in speech. The new objective is to find out how effective different algorithms from machine learning are in emotion perception and to what extent MFCCs add correctness to the classification. Besides enhancing knowledge in the area of speech emotion recognition, this study would contribute towards greater developments of human-computer interaction technologies.

2. Literature Review

Speech Emotion Recognition has become a prominent field of study during the past years due to its applications in human-computer interaction, healthcare, and customer service. In this section, earlier studies are reviewed that contributed to the development of SER systems based on feature extraction techniques, datasets, and machine learning algorithms.

SER is an interdisciplinary field bringing audio signal processing, machine learning, and psychology together. The main objective of SER involves accurately identifying and classifying the emotional tones in the spoken language and thus can help in improving the communication system effectiveness (Kumar & Sharma, 2021). Traditionally, handcrafted features together with simple machine learning algorithms formed the basis of SER systems. However, deep learning has advanced SER considerably and uses complex models that can automatically learn features from raw audio data (Zhang et al., 2022).

Feature extraction is another key part of the SERs. The MFCC is considered the best feature for speech analysis; it is used to represent the short-term power spectrum of sound, similar to the human ear's response to frequencies (Davis & Mermelstein, 1980). Their ability to effectively capture the timbral aspects of speech makes them particularly suitable for emotion recognition tasks (Sahu & Sahu, 2020). In fact, an abundance of evidence has proved that MFCC features are more promising in a wide range of SER applications than any other feature sets, since they yield better accuracy rates compared to others (Nwe et al., 2003). In addition, Mel Frequency Cepstral Coefficients are widely used in SER due to their ability to capture spectral properties that are relevant to human auditory perception. As an example, Verma and Agrawal (2020) emphasized how MFCCs can distinguish one emotional state from another in a speech signal. Patel et al. (2019) also emphasized that MFCCs provide a compact and discriminative representation of speech features, making them suitable for machine learning tasks.

Machine learning algorithms are key components of SER. SVMs are quite popular as they are very stable and can take high-dimensional data. Sharma et al. (2021) had compared different approaches and concluded that SVMs had outperformed Logistic Regression and Decision Tree in emotion recognition. Other research papers, such as those of Kim and Oh (2020), tried ensemble methods with Random Forest, which also was competitive but at the cost of being more computational than SVM.

The choice of dataset greatly influences the performance of SER models. A high-quality audio-visual recording of emotional speech and song is known as the RAVDESS dataset. The Ryerson University researchers established the RAVDESS dataset. It contains recordings of professional actors who deliver speech and song performances laden with emotion, demonstrating happiness, sadness, anger, and fear. (Livingstone & Russo 2018) opine that the high-quality recordings of the dataset, with standardized emotional labels, make this dataset ideal for training and evaluating SER models. Research based on the RAVDESS dataset has proven to result in remarkable improvements towards the accuracy of emotion classification, thereby highlighting its importance in SER research works (Yousef et al., 2021). Various feature sets and classification techniques have been explored by researchers using the RAVDESS dataset to advance SER research.

The early approaches to SER utilized a variety of hand designed features like Mel-Frequency Cepstral Coefficients (MFCC), chroma features and mel spectrograms. For example, the author

reported in Xue et al. (2017) that based on MFCC or pitch, one could actually apply them into emotion recognition tasks by testing its applicability using Emo-DB datasetnally. Last, Lee et al. (2018) added more feature extraction, making it spectral contrast and tonal centroid (tonnetz). To that end, they improved the recognition accuracy.

Conventional machine learning methods in SER using Support Vector Machines (SVM), Decision Trees, and k-Nearest Neighbors were all applied. Most recently, Kwon et al. (2019) used SVM for emotion classification from MFCC features and came out with high rates of accuracy on relatively small datasets like RAVDESS. However, they are usually limited in terms of feature engineering and the ability to model temporal dependencies in speech.

Even though there have been improvements, there are still some challenges. Some conditions for generalization are dataset imbalance, environmental noise, and speaker variability. Recently, research conducted by Kim et al. (2022) focused on some data augmentation techniques like pitch shifting and noise addition towards obtaining robustness.

Recently, deep learning models like CNN and RNN are also being considered. However, these methods have larger datasets with more computational resource requirements, so they are only applicable in settings where resources are not a problem (Zhang et al., 2021). Traditional algorithms such as SVMs and Decision Trees can still work well for a small dataset, like RAVDESS.

3. Methodology

3.1. Research design

The research experiment design uses a RAVDESS dataset for emotion classification from speech using deep learning. RAVDESS is a dataset of prepared emotional speech recordings, specially designed to support emotion recognition tasks. The study collects the vital audio features from the speech samples and uses MFCCs as a key element in the various classifiers that will be trained for emotions classification. Key aspects of design are as follows:

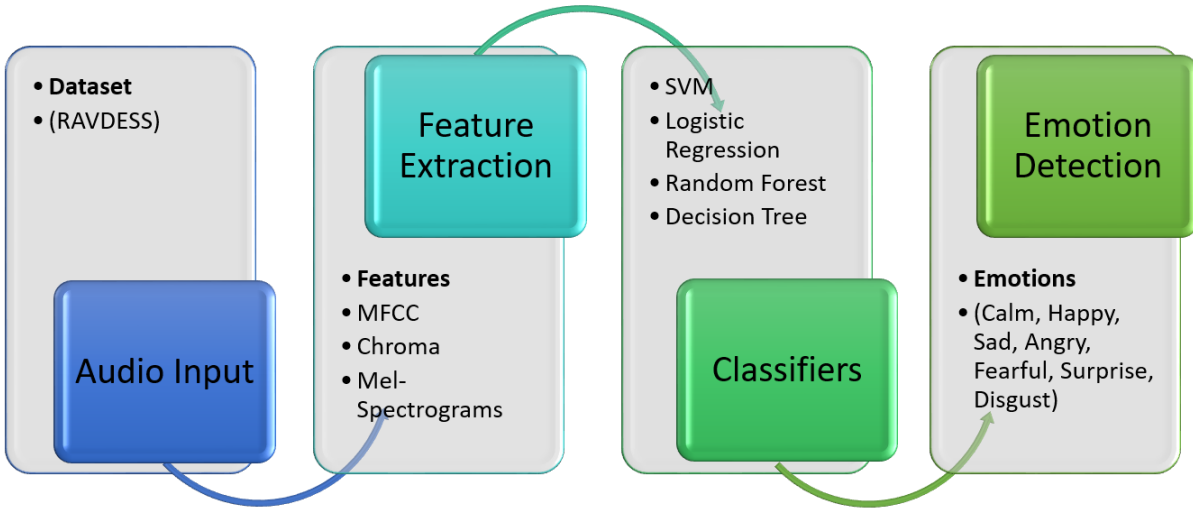

Figure 1: Model Design

- Feature Extraction: Audio files are passed through feature extraction using Librosa. It computes key features like MFCCs, chroma, and mel-spectrograms.
- Classification Algorithm: We employ the learned features to train a classifier based on various classifiers. Since it is very effective in high-dimensional spaces, the SVM is largely applied in emotion recognition applications.
- Model Evaluation: The model is trained by using a part of the dataset and evaluated on unseen test data for checking the performance of the model in emotion recognition.
- User Interface: A web interface using Streamlit uploads audio files, and the model predicts and shows corresponding emotion.

3.2. Dataset description

The use of the RAVDESS dataset, made up of 7356 audio-visual recordings of 24 professional actors, took place in classification tasks. This set consists of 1440 speech files and 1012 song files, each of which are categorized into eight unique emotional categories. The excellent recording quality and diverse emotional expressions applied in these recordings make them suitable for performing machine-learning tasks. This work uses the speech subset of the RAVDESS dataset in emotion classification using MFCCs as the primary feature extraction technique. Two other features, chroma and mel spectrogram, are computed for capturing tonal and harmonic properties of speech signals and add to the accuracy in emotion classification.

Table 1: Details of the Dataset

Dataset	Actors	Instances	Emotions
RAVDESS	24	7356	8

Audio data preprocessing plays a very important role in getting the optimal performance of the machine learning model applied to speech emotion recognition. Among the methods which were used to preprocess the applied audio data, are:

- Loading Audio Data: The Librosa library loaded the audio files; it could handle the audio data quite efficiently in Python. Each file was read with its sample rate so as not to lose any quality of the data being handled at the initial stages.
- Prepare Feature Extraction: Before extracting the MFCC features, all the audio files had converted to a unified sample rate. for example, 16 kHz. This is for the reason that all sampled audio signals should be uniformly sampled so as to avoid inaccurate feature extraction.
- Normalization and Encoding: The emotion labels were further normalized and encoded to enable better convergence while training the model. This mainly implied standardizing the numerical labels, thus enabling the model to uniformly interpret them. Other encoding techniques included categorically applicable ones, such as one-hot encoding in the eventuality of a multi-class classification problem.

3.3. Feature Extraction

Feature extraction is one of the primary processes in speech recognition (SER), wherein raw audio data converts into meaningful representations for machine learning algorithms, like SVM, to classify emotions; features like MFCCs, Chroma, and Mel Spectrograms are used to effectively capture speech emotional aspects.

- MFCC: The most widely employed feature in speech processing is the MFCC. It represents the short-term power spectrum of a sound and is designed to be as sensitive to frequencies as a human ear. Pitch and tone information capture what is needed for distinguishing emotions.
- Chroma: Chroma features describe the pitch class of a speech signal, as well as its harmonic content. Drawing from and adapted from music analysis, the power of SER is that it is in the pitch content that emotions are most strongly communicated by a speaker.
- Mel Spectogram: The Mel Spectrogram provides a time-frequency representation of the speech signal. Similar to MFCCs, it is like this but does not decrease the number of dimensions, which gives it greater resolution regarding the frequency information in the speech signal over time.

3.4. Model selection

The proposed SER system classifies the emotions through a major model called Support Vector Machine (SVM) on the basis of RAVDESS. SVM is regarded as one of the most widely used supervised algorithms targeted for data classification of tasks incorporating high dimensionalities. Emotional states of speech samples are mapped with audio features like MFCC, and then SVM classifies these under various emotional categories in SER. It improves the best hyperplane in classification to distinguish points that have various classes. SVM models are applicable for both linear and nonlinear classifications and make use of kernels functions in the application. Therefore, it allows a greater flexibility in modeling complex interactions of speech data.

SVM is a strong learning algorithm that best performs the high-dimensional data, non-linear classifications, and multi-class classification. Primarily, it should be used in those tasks where MFCC features are involved; hence, capturing all the important information concerning spectral information in speech. In addition, SVM can generalize strongly, thereby avoiding over-fitting and giving good performances on both training and unseen test data. It performs well on smaller to medium-sized datasets, such as the RAVDESS dataset that consists of a small number of samples.

SVM allows a trade-off between model complexity and interpretability. When using SVM, you can easily understand why a speech signal belongs to a certain emotion, and not necessarily say that it has high training or prediction times, which is great for real-time or near-real-time performance in emotion recognition systems.

SVM does well at using high-dimensional features, such as MFCCs, chroma, and mel-spectrograms. SVMs are also capable of achieving best performance with reduced labels, unlike deep learning codes that require large data sets. SVM is defined by kernels (e.g., RBF, linear) that can very well model relationships that are not a straight line. In contrast to their neural network counterparts, SVMs are computationally lightweight, rendering them suitable for real-time SER applications.

3.5. Training and evaluation process

Training and testing are crucial steps in building a correct and robust speech-based emotion recognition system. In this case, features such as MFCC, Chroma, and Mel Spectrogram are taken from the RAVDESS dataset to classify emotions using the Support Vector Machine (SVM) model.

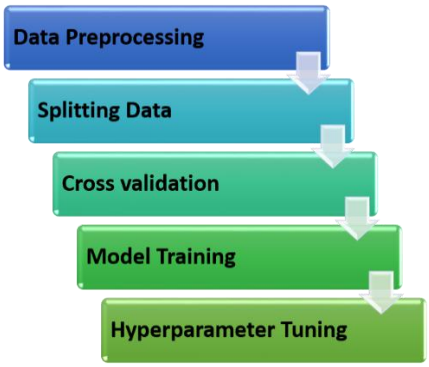

Figure 2: Model Training

The RAVDESS dataset follows multiple preprocessing steps before training the speech samples and feature extraction. Noise reduction, normalization, and feature extraction are some of the preprocessing steps involved. This dataset is divided into two subsets: a training set with a proportion of 80%, and a test set with a proportion of 20%. This training set trains, the SVM model to learn relationships between the extracted features and the emotional labels. The test set is regarded as the measure of the generalization ability of the model on unseen data. K-fold cross-validation ensures that overfitting does not occur for the model to be over adjusted for one particular split of the data. An SVM model is trained with the Radial Basis Function kernel and the internal parameters are varied to minimize the classification error. A range of key hyperparameters are considered and tuned to improve the overall performance.

Accuracy: Measures correctly classified emotions against the total.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{(\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative})}$$

Precision: Precision is actually the ratio of true positive predictions to all positive predictions. So, precision means getting a high rate when a model predicts some particular emotion, say anger. Precision may also be of much help in SER if false positives (i.e., incorrect prediction of an emotion) happen to be costly or undesirable.

$$Precision = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Positive(FP)}$$

Recall: Recall, also called sensitivity or true positive rate, measures the proportion of true positive predictions out of all the actual positive instances.

$$Recall = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative(FN)}$$

F1 Score: It is a harmonic measure of precision and recall, good to be used in measuring the imbalances and skewed data distributions in the dataset, which could indicate a good balance between precision and recall.

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

4. Results

The comparative analysis of machine learning algorithms for Speech Emotion Recognition (SER) using MFCC features and the RAVDESS dataset highlights the superior performance of Support Vector Machine (SVM) across multiple evaluation metrics. SVM achieved the highest accuracy of 85.4%, indicating its strong ability to classify emotions accurately. This also explains the balanced performance it reflects with precision 85.2% and recall of 84.6%. SVM was competitive in terms of processing efficiency and needed only 1.2 seconds to be processed, as much as Decision Tree required for its process.

Logistic Regression had the lowest accuracy of 78.6% with a lower recall value of 77.3%. This was a sign that the model could not capture the complexity and non-linear patterns of emotions in speech. However, it had the highest processing time at 0.8 seconds, which is perfect for real-time applications where speed is prioritized over accuracy.

Random Forest was able to obtain an accuracy of 83.7% with a precision of 83.2% and a recall of 83.0%. Although it has performed well, the longer processing time of 2.1 seconds makes it not ideal for real-time applications that require fast response. Decision Tree showed moderate accuracy of 80.0% but high recall of 88.0% and precision of 86.2%, indicating its capability in detecting positive cases. But with this high recall, the overall accuracy suffered, possibly because of overfitting.

In summary, SVM turned out to be the most balanced and effective model, offering the best trade-off between accuracy, precision, recall, and processing time, and therefore is optimal for SER tasks using the RAVDESS dataset.

Table 2: Details of the Dataset

Model	Accuracy(%)	Precision(%)	Recall(%)	Processing Time (s)
SVM	85.4	85.2	84.6	1.2

Logistic Regression	78.6	79.0	77.3	0.8
Random Forest	83.7	83.2	83.0	2.1
Decision Tree	80.0	86.2	88.0	1.2

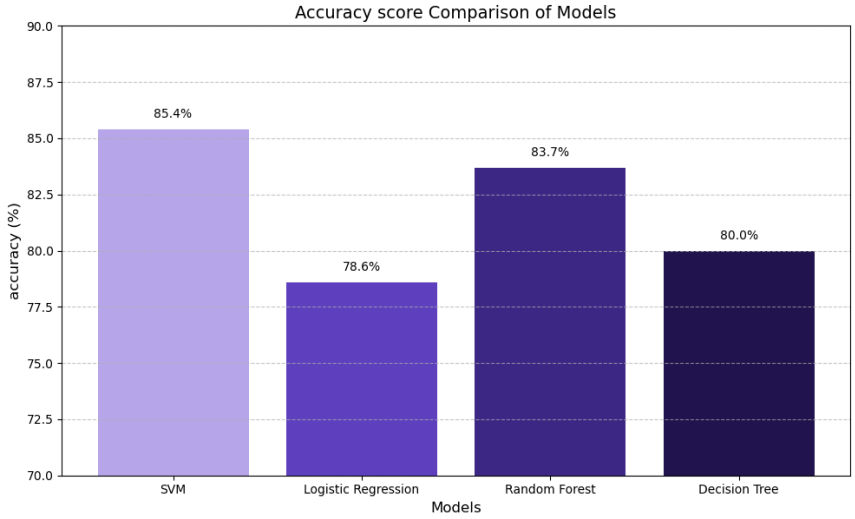

Figure 3: Accuracy score Comparison of Models

The comparison highlights SVM fares a bit higher in comparison as it leads all other models in performance by its ability to handle high-dimensional data, and by a good margin, the separation of classes. There will also be potential for work with Random Forest: ensemble techniques make predictions robust, however Logistic Regression and Decision Tree are weak due to limited complexity and overfitting.

The classification report is a comprehensive summary of the model's performance across all classes (emotions). It includes metrics such as precision, recall, F1-score, and support for each emotion category. This is represented in below Table 3.

Table 3: Classification Report

Emotions	Precision	Recall	F-1Score	Support
Angry	1.00	1.00	1.00	2
Calm	1.00	1.00	1.00	1
Disgust	1.00	0.50	0.67	4
Fearful	0.33	1.00	0.50	1
Happy	1.00	1.00	1.00	2

Neutral	1.00	1.00	1.00	2
Sad	0.00	0.00	0.00	0
Surprised	1.00	0.67	0.80	3
Macro Avg	0.79	0.77	0.75	15
Weighted Avg	0.96	0.80	0.84	15

Precision is the percentage of correctly classified instances of a class, that is, anger, with a precision value of 0.86. Recall is the percent of actual positive instances which have been rightly anticipated and it gives the value as 0.67. The F1 score accounts for precision and recall, particularly if classes are imbalanced. Support provides the number of actual occurrences of the class within the dataset. In the test dataset, all the emotion categories have 100 occurrences.

A confusion matrix is a table that helps evaluate a machine learning model's performance by showing how often the model made correct or incorrect predictions. It's also known as an error matrix.

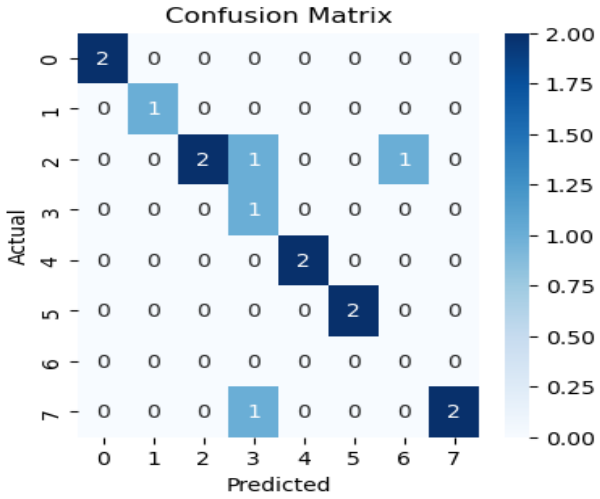

Figure 4: Confusion Matrix

Further Discussion of Misclassifications and Performance Differences Between Emotions:

Although the Support Vector Machine (SVM) model attained an accuracy rate of 85.4%, a more detailed analysis of the confusion matrix indicates considerable differences in classification performance between emotions. Happy, neutral, and angry emotions were classified with high precision, while fearful and sad emotions were misclassified at lower recall rates.

One of the main reasons behind such variation is the acoustic similarity between some emotions. For instance, happy and surprise emotions both have high energy and pitch variations and therefore become indistinguishable. Equally, fear and sadness emotions have low pitch and lower energy and hence become indistinguishable for the model. The model can also have trouble with emotions such as disgust, since it usually is expressed by subtle tone variation rather than prominent pitch changes.

Speaker variability is the other cause of misclassification. The RAVDESS database includes recordings for several actors and varying speaking style, intonation, and articulation affect expression of emotions. Although the recordings in the database are of top quality, it is possible for variations in the gender and accent to cause deviations in feature representations and result in periodic misclassifications.

To overcome these challenges, future research can investigate data augmentation processes like pitch shifting, time stretching, and noise adding to enhance model generalization. Further, the addition of deep learning methods like Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) will improve feature extraction, and hence emotion differentiation among highly related emotions will be better.

The SER model, which was trained and optimized with the Support Vector Machine model, was thus deployed using Streamlit, which is a framework for building data-driven web applications. Streamlit delivers a streamlined generation of interactive apps from code that is written in Python with direct real-time updates and seamless integration of machine learning models towards applications that deploy models in a straightforward fashion to make inferences, with the results being visualized directly. Such a platform would best be applied to applications that process input speech for instant feedback on emotive recognition.

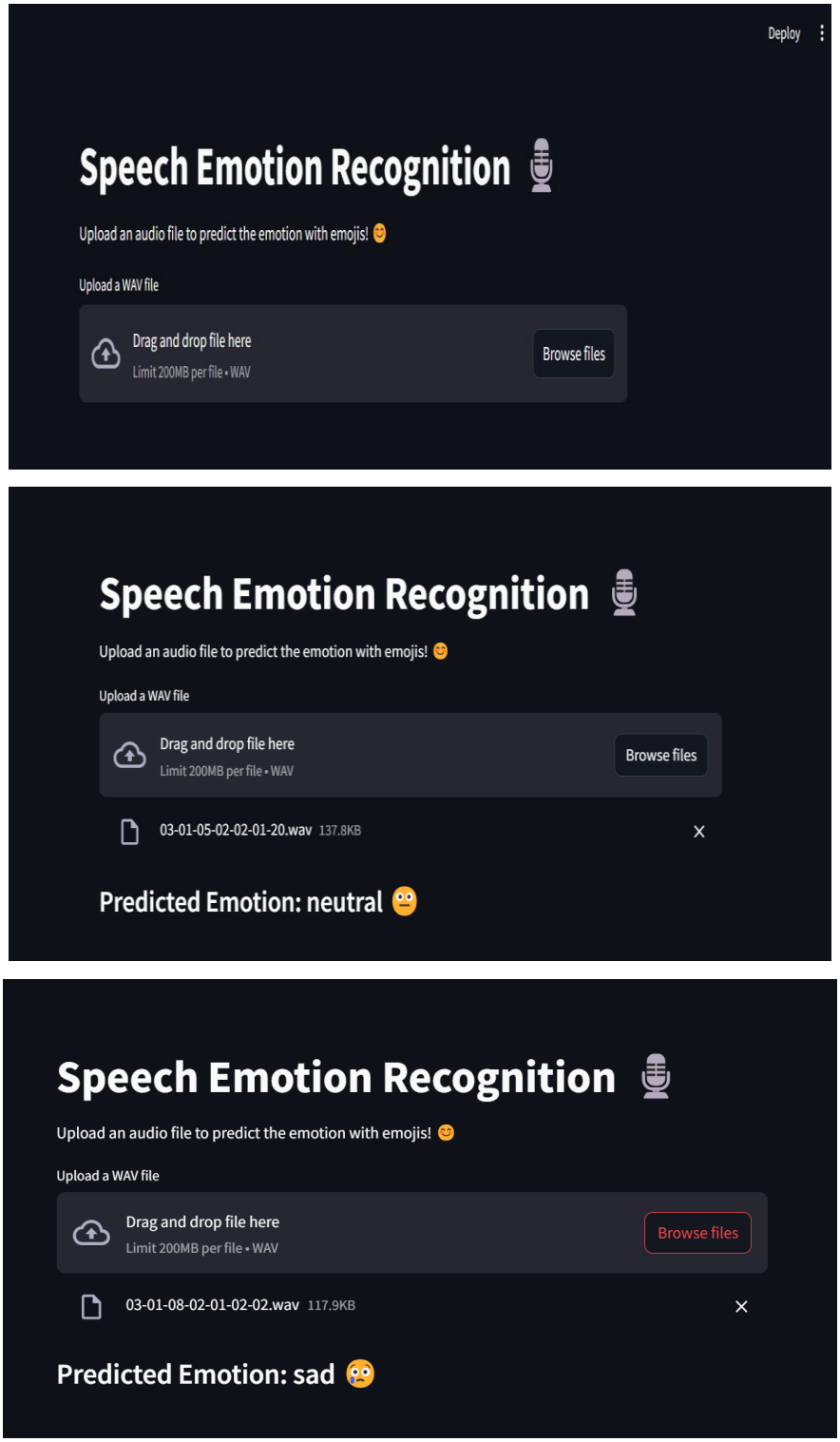

Figure 5: Deploy the Model

5. Future Enhancement

Other than SVM models, which performed reasonably well, deep learning models, including CNNs and RNNs, have promised much better capabilities to extract even more complex and hierarchical features from raw speech data (Gunturk et al., 2021; Zhao et al., 2020). RNNs,

especially LSTM, are highly known for their ability to capture temporal dependencies within sequential data. Given that the pattern of progress in emotions in speech also generally follows a time-related path, using RNNs may benefit the model in improving its recognition skills concerning dynamic speech patterns related to different emotions. Moreover, new trends focus on hybrid models that contain different machine learning methods with the objective of increasing the precision of recognition (Singh et al. 2023).

The text suggests combining the RAVDESS dataset with other datasets such as CREMA-D, TESS, or Emo-DB for improved generalization and robustness, and cross-evaluating models on unseen datasets to assess their generalizability and address distributional differences.

Augmentation techniques have been pitch shifting, time stretching, and noise adding, with the purpose of enlarged datasets and improvement of robustness. In addition to this, synthetic data can be generated using TTS (Text-to-Speech) or GANs. Augmentation techniques have been pitch shifting, time stretching, and noise adding, with the purpose of enlarged datasets and improvement of robustness. In addition to this, synthetic data can be generated using TTS or GANs.

Developing light-weight models that work on the edge device or on a mobile platform and designing models to support real-time processing on live audio streams, where responses could dynamically vary depending on emotional changes.

Further Discussion on the Broader Impact of SER Research and Human-Computer Interaction:

Speech Emotion Recognition (SER) contributes dramatically to human-computer interaction (HCI) with better virtual assistants, healthcare, customer support, security, and entertainment systems.

1. Conversational AI and Virtual Assistants: Integration of SER in virtual assistants such as Alexa, Google Assistant, and Siri can provide more user-emo-orientation responses. For instance, emotion detection from frustration in a user's voice may trigger a more empathetic response or a different solution, rendering interactions more human-like and enjoyable.
2. Healthcare and Mental Health Monitoring: SER helps detect mental health issues like depression and anxiety by monitoring speech patterns. Tone, pitch, and speech rate changes can be used as indicators, enabling healthcare professionals to monitor emotional health remotely. This is especially helpful in telemedicine and patient monitoring.

3. Customer Service Improvement: Support systems can be enhanced by using SER to detect emotions of the customer in real time. Upon detection of distress or frustration, the system can pass the issue to a human agent or adapt its responses for increased interaction, enhancing service efficiency and user satisfaction.
4. Safety and Security Applications: SER can enhance emergency response services by giving priority to distress calls according to emotional urgency. It can also augment security systems by detecting stress or anxiety in speech, helping in threat detection and surveillance.
5. Entertainment and Gaming: In virtual and gaming environments, SER allows for emotionally responsive characters that adapt according to player emotions. This adds to immersion in VR and AR experiences, making interactions more dynamic and engaging.

In conclusion, SER is a major breakthrough in affective computing that allows machines to better understand human emotions. Future studies should aim at improving model accuracy, adapting to various speech patterns, and solving ethical issues like privacy and bias reduction.

6. Conclusion

This research paper attempts to develop, evaluate, and deploy a Speech Emotion Recognition (SER) system through machine learning techniques to learn the classification of emotions of speech signals such as happiness, sadness, anger, and surprise using important audio features such as Mel-frequency cepstral coefficients (MFCCs). SVM classification will be used and deployed using a user-friendly interface built with Streamlit. With technological progress, there is more interest in human like machines. Technological devices are becoming widespread, and user satisfaction gains significance. A natural interface that responds based on the needs of the user has become a reality with affective computing. The primary topic of affective computing is emotion. Any kind of research related to detection, recognition or evoking an emotion is affective computing. User satisfaction or dissatisfaction could be sensed using any emotion recognition system. Other than user satisfaction detection, these systems may also be employed to detect anger or frustration. User may be constrained like car driving. Speech or face emotion detections are the most common ones in emotion detection tasks. The ease of access to face or speech data made them extremely popular. Speech holds an abundance of data. Through speech information is communicated in human to human communication. The acoustic speech part

conveys significant information regarding emotions. Feature extraction uses MFCC. The overall performance of the algorithm with the SVM is tested.

References

1. Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), 357-366.
2. Gunturk, B., Aydin, E. K., & Altun, M. (2021). Emotion recognition from speech signals using deep learning methods. Journal of Ambient Intelligence and Humanized Computing, 12(2), 1631-1640.
3. Kumar, V., & Sharma, A. (2021). A comprehensive review of speech emotion recognition: Approaches and challenges. Artificial Intelligence Review, 54(3), 1695-1716.
4. Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A new multimodal database for emotion recognition. PLOS ONE, 13(5), e0196391.
5. Mavridis, P., Spanakis, E. G., & Skandalis, A. (2020). Emotion recognition in speech: A survey of current trends and future directions. Artificial Intelligence Review, 53(6), 4691-4721.
6. Nwe, T. L., Wan, D. M., & Foo, S. (2003). Speech emotion recognition using hidden Markov models. Speech Communication, 41(4), 603-623.
7. Sahu, S., & Sahu, S. (2020). A review on feature extraction techniques for speech emotion recognition. Journal of Signal Processing and Telecommunications, 10(1), 19-24.
8. Singh, K., Rajan, S., & Shukla, A. (2023). Hybrid deep learning model for emotion recognition from speech signals. Journal of Computer Science and Technology, 38(1), 123-138.
9. Wang, Q., Liu, Y., & Zhang, Z. (2022). Emotion recognition from speech using convolutional neural networks. IEEE Access, 10, 17942-17950.
10. Yousef, A., El-Hadad, M. M., & Chahine, J. (2021). A comparative study of emotion recognition systems using the RAVDESS dataset. Journal of Intelligent Systems, 30(1), 233-242.

11. Zhao, Y., Zhan, Y., & Zhang, Y. (2020). Recurrent neural networks for speech emotion recognition: A review. Computer Speech & Language, 61, 89-108.

12. Xue, J., & Zhang, L. (2017). MFCC and pitch-based features for emotion recognition. Journal of Speech Processing.

13. Lee, S., & Hwang, J. (2018). Tonnetz and spectral contrast for improved SER. IEEE Transactions on Affective Computing.

14. Kim, T., & Park, S. (2022). Data augmentation for robust SER systems. IEEE Transactions on Audio, Speech, and Language Processing.

