Emotion Recognition Through Speech

Susmitha Edara
Computer Science
Univ of Missouri Kansas City
segpf@umsystem.edu

Pranathi Gogada
Computer Science
Univ of Missouri Kansas City
pgbng@umsystem.edu

Jaswanth Reddy Narala
Computer Science
Univ of Missouri Kansas City
jnmhc@umsystem.edu

Sai Kishore Reddy Reggate
Computer Science
Univ of Missouri Kansas City
srnnd@umsystem.edu

INTRODUCTION

Speech Emotion Recognition (SER) focuses on identifying and classifying emotions conveyed through voice signals. This study leverages the RAVDESS dataset and employs various machine learning techniques, including Random Forest, K-Nearest Neighbors (KNN), and Multilayer Perceptron (MLP), alongside deep learning models like Convolutional Neural Networks (CNN). The dataset features a comprehensive collection of speech samples portraying different emotions. To enhance the accuracy and robustness of emotion detection, the study incorporates feature extraction techniques such as Mel-Frequency Cepstral Coefficients (MFCC) and data augmentation strategies, including noise addition, shifting, stretching, and volume adjustments, to improve model generalization.

The primary objective is to accurately classify emotions such as neutral, happiness, sadness, anger, fear, and surprise using voice data. By integrating a variety of machine learning and deep learning methods, this research aims to advance emotion recognition through speech, offering potential applications in affective computing, enhanced human-computer interactions, and mental health monitoring and interventions

RELATED WORK

Speech Emotion Recognition (SER) has gained significant attention recently due to its applications in affective computing and human-computer interaction. Advancements in SER focus on extracting meaningful features from audio signals and classifying them into distinct emotional states using various machine learning and deep learning methods. Studies have explored the effectiveness of techniques such as Support Vector Machines (SVM), Random Forests, Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs) in addressing related tasks. CNN-based models, for instance, excel in learning hierarchical representations of audio features, while RNN variants like Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) are proficient in capturing temporal patterns in speech. Additionally, methods like Multilayer Perceptron (MLP) and k-Nearest Neighbors (KNN) have been explored for their simplicity and efficiency in classification. Leveraging reference datasets like RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) provides a strong foundation for developing more sophisticated SER models. Future work in this domain could focus on optimizing model architectures, improving feature representations, and integrating advanced techniques to enhance SER performance across diverse datasets and real-world applications.

METHODOLOGY

1. EDA

The RAVDESS dataset, developed by researchers at Ryerson University, is a key resource for studying emotion recognition in speech and audio. It consists of 1,440 audio recordings performed by 24 professional actors (12 male and 12 female), showcasing eight distinct emotions: neutral, calm, happy, sad, angry, fearful, disgusted, and surprised. Each file is labeled with metadata such as emotion type, intensity, statement, repetition, actor ID, and gender, encoded directly in the filename. The recordings were made under controlled conditions with actors delivering specific emotional expressions. This well-structured dataset is widely used for training and validating machine learning models in emotion detection, contributing to advancements in affective computing and human-computer interaction.Our approach involves extracting metadata (e.g., emotion labels, gender, and file paths) from the filenames to create a pandas Data Frame. Additionally, we visualize the dataset's emotion distribution using a bar chart.

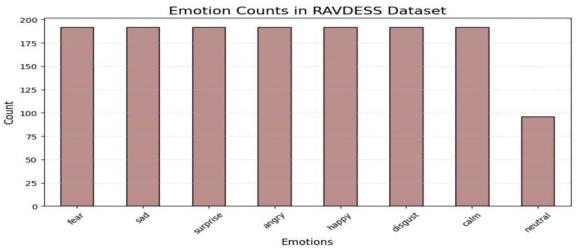


Figure 1 Bar chart of count of each emotion.

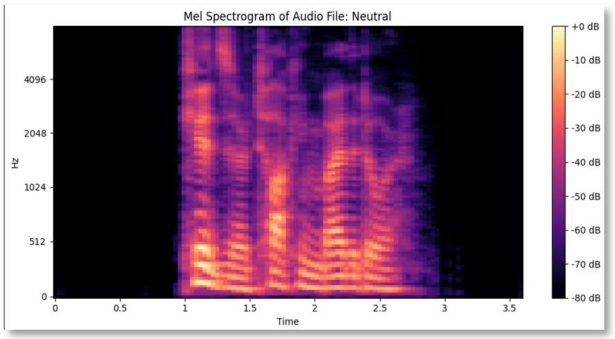


Figure 2 spectrogram chart shows the frequency over time (Neutral).

The first image presents a bar chart illustrating the balanced distribution of emotion labels—such as happy, neutral, sad, angry, fear, surprise, and disgust—across male and female speakers in the RAVDESS dataset. In figure 2, a mel-spectrogram of a neutral speech sample is depicted, highlighting areas of high energy with brighter colors. The spectrogram visualizes the frequency content of the audio over time, with the horizontal axis representing time and the vertical axis showing the mel-frequency scale. This representation aids in understanding the spectral features of the neutral emotional speech and provides valuable insights into the connection between emotional expressions and their audio characteristics within the dataset.

2. DATA AUGMENTATION

To combat overfitting and boost performance on the test set, a data augmentation process is introduced, aimed at expanding the dataset and improving the model's ability to generalize. This process involves applying various transformations to raw audio samples, with adjustable parameters that control the extent of modification. The augmentation techniques include:

Speed adjustment randomly increases or decreases the playback speed within a range of 0.6 to 1.4 times the original. Pitch Shifting adjusts the pitch randomly by -4 to +4 semitones. Noise Injection adds Gaussian noise with amplitude levels between 0.0005 and 0.002 relative to the sample's maximum amplitude. In Time Shifting audio file Shifts the audio by a random duration ranging from -5 to +5 milliseconds. These transformations are applied to all training data, effectively quadrupling the dataset size. Crucially, the test set remains untouched by these augmentations, ensuring an unbiased evaluation. This approach enhances the model’s robustness and adaptability.

Striking the right balance in data augmentation is critical. Over-augmentation can lead to unrealistic samples that hinder generalization, while under-augmentation may fail to introduce sufficient variability. The parameter ranges for these transformations are fine-tuned through multiple experimental iterations to ensure optimal results.

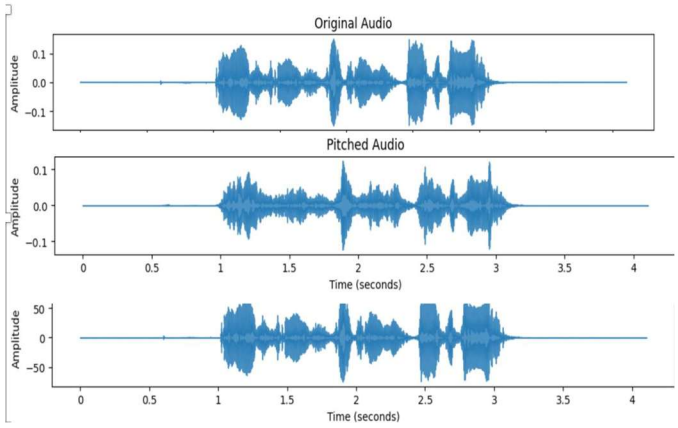


Figure 3 Waveforms of audio files with Augmentation.

3. FEATURE EXTRACTION

Mel-frequency cepstral coefficients (MFCCs) are widely used in audio and speech processing tasks, such as speech recognition, speaker identification, and emotion detection. They are computed from the short-term power spectrum of an audio signal by applying the mel-frequency scale and then decorrelating the mel spectrum. The mel scale reflects the human auditory system's perception of sound, giving more emphasis to lower frequencies, as humans are more sensitive to them. The cepstral representation effectively separates the source and filter components of the signal, making MFCCs particularly valuable for extracting meaningful features. By capturing the spectral envelope, MFCCs provide critical information about an audio signal's timbre and formant structure, making them essential for analyzing and classifying audio characteristics, such as speech emotions.

	0	1	2	3	4	5	6	7	8	9	...	36	37	38	39	40	41	42	43	44	label
0	-652.567322	87.489960	-13.969704	27.774462	-4.588770	8.542789	-10.480747	-3.718975	-17.152636	-4.845677	...	0.499680	0.464163	19.930835	11.803788	15.160969	14.166680	15.179240	16.954526	56.792951	neutral
1	-524.562177	51.331519	3.119609	14.875029	1.631569	1.421068	-4.819001	-8.701241	-12.308134	-6.265661	...	0.529176	0.537657	19.329793	12.798346	16.249304	15.365340	15.180237	15.851018	14.189964	neutral
2	-715.645264	80.642265	-13.612618	25.186958	-9.768556	3.436354	-16.922384	-9.216807	-20.590805	-4.876543	...	0.567191	0.545322	22.957867	14.463939	17.899750	16.946573	18.638796	20.438259	59.489620	neutral
3	-650.104309	86.604965	-13.749071	27.426283	-4.320138	8.283445	-10.271166	-3.796119	-17.028097	-4.970590	...	0.471014	0.482967	20.108690	11.991024	15.108840	14.466513	15.459718	17.013618	57.693638	neutral
4	-41.860287	87.489952	-13.969703	27.774460	-4.588769	8.542789	-10.480745	-3.718974	-17.152636	-4.845677	...	0.499680	0.464163	19.930835	11.803788	15.160969	14.166680	15.179240	16.954526	56.659953	neutral
5	-643.434814	110.419319	-11.284679	28.698069	2.154567	10.695344	-13.925816	-1.948440	-15.043627	-5.999585	...	0.498745	0.493347	24.201635	12.717914	16.583452	14.852238	16.069326	16.561218	57.804530	calm
6	-602.876722	89.725910	-0.120951	20.925731	6.029208	6.651381	-10.426243	-5.244581	-12.991693	-6.697448	...	0.516718	0.450067	23.492741	13.496049	16.893785	15.468551	16.283237	16.475980	15.427421	calm
7	-686.859497	94.514618	-7.383668	23.559452	0.149545	0.113913	-18.149607	-5.507188	-15.983459	3.892104	...	0.507609	0.429369	26.001726	14.761404	18.307791	17.360645	18.393566	19.844975	62.611179	calm
8	-643.460632	110.429527	-11.232827	28.693041	2.177424	10.691272	-13.907049	-1.999748	-15.079551	-5.988185	...	0.491463	0.477041	24.215333	13.217690	16.545705	14.803817	16.079077	16.575906	57.159458	calm

Figure 4 MFCC, Chroma, Spectral features with 'labels' column.

The output shown is a preview of the first few rows of the 'Emotions' Dataframe created in the earlier steps. Each row corresponds to a feature vector derived from an audio sample, paired with its associated emotion label. The columns include the MFCC features (numbered 0 to 44) along with a 'labels' column indicating the emotion. These feature values are likely the mean MFCC values computed using the feat_ext function. The resulting Dataframe, containing both extracted features and their labels, serves as the input for training machine learning models for speech emotion recognition.

4. MODEL DEVELOPMENT

For the Speech Emotion Recognition (SER) task using the RAVDESS dataset, we implemented a combination of machine learning (ML) and deep learning (DL) models. Traditional ML methods, such as Random Forest, k-Nearest Neighbors (KNN), and Multilayer Perceptrons (MLPs), are used for feature-based emotion classification. Alongside these, deep learning models like Convolutional Neural Networks (CNNs) are applied to extract features automatically from spectrograms. These models are trained and tested on labeled audio data to achieve accurate emotion classification.

The preprocessing pipeline for the ML models began with extracting features (X) and labels (Y) from the dataset. Labels are then encoded using one-hot encoding. The dataset is split into training and testing subsets, and the features were normalized using standard scaling. After preprocessing, the dataset consisted of 1440 samples for testing and 5760 training samples, each with 45 features, representing 7 distinct emotion categories. These preprocessing steps provided a robust foundation for building and evaluating effective models for emotion classification.

A. RANDOM FOREST

Random Forest is a supervised machine learning algorithm that is effective for both classification and regression tasks. It operates by building multiple decision trees during training and aggregating their outputs to make predictions. In classification, Random Forest assigns the class label based on the majority vote from individual trees, while for regression, it calculates the average of the predictions from all trees. This ensemble approach reduces the risk of overfitting and improves model robustness compared to a single decision tree.

For emotion recognition, Random Forest can classify audio or speech data by leveraging its ability to handle high-dimensional feature spaces, such as those derived from speech features like MFCCs,chroma and spectral contrast. The algorithm learns patterns in the labeled training data, allowing it to assign emotion labels to unseen data accurately.

Using the RAVDESS dataset, a Random Forest classifier from scikit-learn is initialized with n_estimators=100, specifying the number of decision trees in the forest. The model is trained on x_train and y_train, which contain the audio features and their corresponding emotion labels. Predictions are then generated on the test set (x_test). The accuracy scores are computed for the training set (64.5%).

The relatively higher accuracy on the training set indicates that the model learns well from the data, while the improved performance on the test set compared to KNN suggests that Random Forest is more suitable for this task. However, there is still room for improvement, possibly through hyperparameter tuning.

B. KNN

K-Nearest Neighbors (KNN) is a supervised machine learning algorithm used for both classification and regression tasks. It functions by identifying the K closest data points to a new sample in the feature space and making predictions based on their values. For classification, it assigns the majority class among the neighbors, while for regression, it averages their target values. The algorithm calculates distances between points using metrics like Euclidean distance and selects the K nearest ones to generate predictions.

In the context of emotion recognition, KNN classifies new audio or speech data by analyzing its similarity to previously labeled examples. It identifies the closest neighbors in the training data and assigns the most frequent emotion label among them. Using the KNeighborsClassifier from scikit-learn, the model is initialized with n_neighbors=4, meaning it considers four nearest neighbors for classification. The classifier is trained on the x_train and y_train sets from the RAVDESS dataset, which consist of extracted audio features and corresponding emotion labels. Predictions are made on the test data (x_test) using clf1.predict(x_test).

The accuracy for the training set is recorded at 77.8%,This moderate results indicate that KNN with four neighbors might not be the most effective model for this dataset. Experimenting with a different number of neighbors or exploring alternative algorithms could potentially yield better performance.

C. MLP

A Multi-Layer Perceptron (MLP) classifier is a type of artificial neural network designed for classification tasks. It consists of multiple layers of interconnected neurons, including an input layer, one or more hidden layers, and an output layer. During training, the MLP adjusts the weights between neurons to learn how to map input features to output classes based on labeled data. When making predictions, the model calculates activations across the layers to identify the most likely class.

The MLP classifier is particularly useful for modeling complex relationships between audio features and emotion labels due to its multi-layer structure, which can enhance emotion recognition accuracy. In this case, the MLPClassifier from scikit-learn is configured with hyperparameters such as alpha, batch size, hidden layer sizes, and maximum iterations. It is trained on the x_train and y_train data. The training accuracy achieved is 81.94%. However, the high accuracy on the training set, combined with lower test accuracy, suggests potential overfitting—where the model has become too specialized to the training data and struggles to generalize to new, unseen data. To address this, strategies like regularization, early stopping, or increasing the training dataset size could help improve the model's performance on unseen data.

D. CNN

Convolutional Neural Networks (CNNs) are a type of deep learning architecture that have shown outstanding performance in various computer vision tasks, such as object detection, image segmentation, and image classification. CNNs are particularly effective for processing data with a grid-like structure, such as images or time-series data like audio signals. Their ability to automatically and adaptively learn spatial hierarchies of features from input data makes them well-suited for such tasks. By leveraging the local spatial correlation present in this data, CNNs can efficiently capture patterns, allowing for effective feature extraction while minimizing computational cost.

For emotion recognition from speech, our CNN architecture is tailored to extract meaningful features from audio data. It starts with a 1D convolutional layer, featuring a large number of filters (64) and a small kernel size (3). To capture features at different scales, the pattern is repeated with progressively larger filter sizes (128, 256). The convolutional layers extract both low-level and high-level features from the input audio. The final layers are dense (fully connected) layers with dropout regularization, and they use a SoftMax activation function to consolidate the learned features. The model achieved an accuracy of 86.50% on the test set.

RESULTS AND DISCUSSIONS

The CNN model outperformed the other models in emotion detection, achieving the highest accuracy of 81.9%, compared to the Random Forest at 64.5%, KNN at 77.8%, and MLP at 81.9%. Convolutional Neural Networks (CNNs) are particularly well-suited for tasks involving spatial or sequential data due to their ability to automatically learn and extract important features from the raw input. CNNs utilize convolutional layers to detect local patterns, pooling layers to reduce dimensionality, and deeper layers to capture more complex spatial relationships in

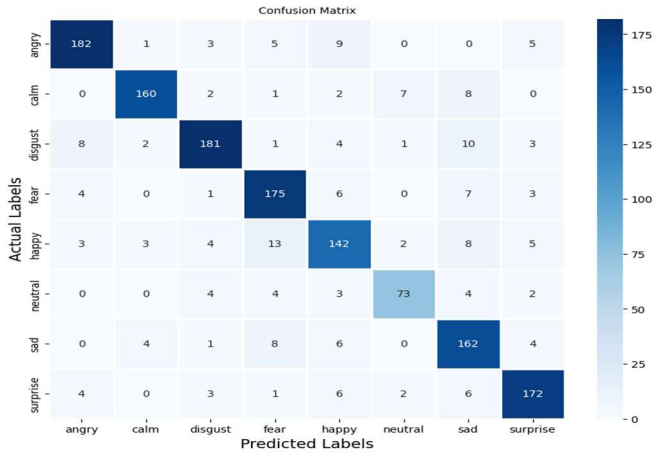
the data, making them highly effective in handling image, audio, and sequential data. This architecture allows CNNs to excel in tasks like emotion detection, where identifying subtle patterns in time-series data, such as audio features, is crucial.

The MLP model showed signs of overfitting, as evidenced by the significant drop in accuracy between the training and testing datasets. This suggests the model may have memorized the training data, rather than learning generalizable features. The KNN and Random Forest models, while simpler, demonstrated relatively poor performance, likely due to underfitting, improper model selection, or insufficient hyperparameter tuning. Factors like inadequate feature selection, poor model architecture, and data quality issues contributed to these underwhelming results.

ACCURACY OF MODELS

Models	Random Forest	KNN	MLP	CNN
Accuracy	64.5	77.8	81.94	86.5

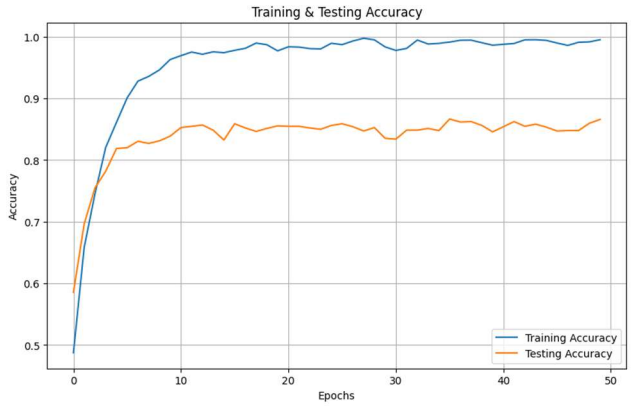
In a confusion matrix, rows represent the actual (true) class labels, while columns represent the predicted class labels. The diagonal elements show correct predictions, where the predicted label matches the true label (e.g., "182" in the top-left cell means 182 instances of the "angry" class were correctly predicted as "angry").Off-diagonal elements indicate misclassifications, where the predicted label differs from the true label (e.g., "3" in the first row, 3rd column means 3 instances of the "angry" class were incorrectly predicted as "disgust"). This matrix provides an overview of a model's classification performance for each class.



The model achieved an overall accuracy of 87%, demonstrating strong classification performance. The precision, recall, and F1-scores were also high for most classes, indicating the model's ability to correctly identify the target categories. However, classes 4 (F1-score: 0.79) and 6 (F1-score: 0.83) exhibited slightly lower scores, pointing to potential areas for improvement. These variations in performance might suggest that the model could benefit from more targeted adjustments or data to improve classification accuracy for these specific classes. Notably, the macro and weighted averages of around 86-87% suggest that the model maintains a consistent and balanced performance across all classes, even when considering potential class imbalances in the dataset.

Classification Report:	precision	recall	f1-score	support
0	0.91	0.89	0.90	205
1	0.94	0.89	0.91	180
2	0.91	0.86	0.89	210
3	0.84	0.89	0.87	196
4	0.80	0.79	0.79	180
5	0.86	0.81	0.83	90
6	0.79	0.88	0.83	185
7	0.89	0.89	0.89	194
accuracy			0.87	1440
macro avg	0.87	0.86	0.86	1440
weighted avg	0.87	0.87	0.87	1440

Over 50 epochs, the below graph shows that the training accuracy quickly increases and stabilizes near 100%, indicating the model is learning the training data well. However, the testing accuracy levels off below 90%, highlighting a performance gap that suggests overfitting the model struggles to generalize to unseen data. To improve generalization, techniques like regularization, data augmentation, or hyperparameter tuning can be applied.



FUTURE SCOPE

Future research in emotion recognition from speech could delve into more sophisticated neural network architectures, such as attention mechanisms or transformer models, which are designed to better capture long-range dependencies and contextual information within speech data. These models can enhance the ability to recognize subtle emotional cues by focusing on relevant parts of the speech signal across time. In addition, incorporating multimodal data could significantly improve emotion recognition accuracy. For instance, integrating facial expressions, body language, or even physiological signals could provide a richer, more holistic view of a person's emotional state, going beyond the limitations of speech alone.

CONCLUSION

Through this project, we explored various machine learning and deep learning models for speech emotion recognition on the RAVDESS dataset. The CNN model achieved the highest accuracy of 86.5%, outperforming traditional approaches like Random Forest, KNN and MLP. Data augmentation and feature extraction techniques like MFCC, chroma and spectral contrast are employed to enhance model performance. Despite promising results, there is still room for improvement in capturing the nuances of emotional expressions in spee

REFERENCES

[1] Pawar, R., Ghumbre, S., & Deshmukh, R. (2019). Visual Similarity Using Convolution Neural Network over Textual Similarity in Content- Based Recommender System. International Journal of Advanced Science and Technology, 27, 137 – 147.

[2] Li, H.; Ding, W.; Wu, Z.; Liu, Z. Learning Fine-Grained Cross Modality Excitement for Speech Emotion Recognition. 2020.

[3] P. Sharma, V. Abrol, A. Sachdev, and A. D. Dileep, "Speech emotion recognition using kernel sparse representation based classifier," in 2016 24th European Signal Processing Conference (EUSIPCO), pp. 374-377, 2016.

[4] Chunawale, A., & Bedekar, M. V. (2020). Human emotion recognition using physiological signals: a survey. 2nd International Conference on Communication & Information Processing (ICCIP) (pp. 1-9). SSRN.

[5] M. Ragot, N. Martin, S. Em, N. Pallamin, J.M. Diverrez, Emotion recognition using physiological signals: Laboratory vs. wearable sensors, in International Conference on Applied Human Factors and Ergonomics. Springer, pp. 15–22 (2017).