Gender and Emotion Recognition Using Voice

Anikait Agrawal
anikait22072@iiitd.ac.in

Ansh Varshney
ansh22083@iiitd.ac.in

Anant Kaushal
anant22067@iiitd.ac.in

Abdullah Shujat
abdullah22013@iiitd.ac.in

Abstract

Speech is the most prevalent means of human communication, used to convey emotions, cognitive states, and intentions. Beyond gender identification, speech and voice recognition systems can also classify emotions expressed by speakers. While the human ear naturally recognizes voice characteristics and emotional cues, we can train machines to perform these tasks using modern machine-learning algorithms and visualization tools. By integrating algorithms with the correct set of features—such as mean frequency (kHz), standard deviation, first and third quantiles, and others—we can prepare a machine to recognize voice attributes and determine both the gender and emotional state of the speaker. In our project, we will collect data not only for gender recognition but also for emotion classification by recording speech samples that reflect various emotional states. This comprehensive dataset will enable our model to learn and identify the subtle features associated with different emotions, enhancing its ability to recognize and interpret human speech. {**GitHub Link**}

1 Introduction

The Voice contains many effective communication methods, including linguistic and paralinguistic parameters such as gender, age, language, etc. Identifying human gender based on the voice has been challenging for voice and sound analysts who deploy numerous applications, including investigating criminal voices in crime scenarios, emotion recognition, and enhancing human-computer interaction. The robustness and effectiveness of classification models depend on the dataset's features; hence, extracting information from the raw data is vital to improving the efficiency.

After getting the extracted features and labels, ML techniques are used to build a high-quality classifier for gender recognition. Progress in technological fields has also improved the methods of doing one task. In this project, we collected datasets from online resources and mainly focused on the features present in most of the datasets.

Further, we have proposed using binary and multiclass classification algorithms, such as Logistic Regression, Naive Bayes, and Support Vector Machines, to classify the speaker's gender and check the effectiveness of these models. Graphical-based methodologies have also been implemented, along with boosting techniques, to improve the learning of the models by feature selection and noise removal from the gathered raw data. Our work can find its usage in fields like speech-emotion recognition, human-to-machine interaction, sorting telephone classes by gender categorization, automatic salutations, muting sounds for gender, and audio/video categorization with tagging. We have been based our start of our next project for emotion recognition by first collecting the dataset.

2 Literature Survey

Gender Recognition by Voice and Emotion is a broad problem, with various ways to overcome it.

2.1 Gender Recognition by Voice Using an Improved Self-Labeled Algorithm

Gender Recognition by Voice using an Improved Self-Labeled Algorithm [2] uses a hybrid of Ensemble Learning and Semi-Supervised Learning (SSL) algorithms called iCST-Votinga for Gender Recognition by Voice. One of the main problems faced by the authors is highly time-varying and has very high randomness. This problem is mainly due to less data availability for efficient training of the classifiers. Finding more data is expensive and time-consuming, while finding unlabeled information is more effortless. The authors have suggested two methods to tackle this problem: semi-supervised learning (SSL) algorithms and Ensemble Learning (EL). The authors proposed a hybrid plan combining the SSL (using the Self-labeled algorithm of SSLs) and EL, called iCST-Voting.

2.2 Gender Recognition from Human Voice using Multi-Layer Architecture

Gender Recognition from Human Voice using Multi-Layer Architecture [3] by Mohammad Amaz Uddin, Md Sayem Hossain, Refat Khan Pathan, and Munmun Biswas describes about extraction of the features from the audio speech to recognize gender as male or female and employs those features to recognize the gender of the speaker. The authors first applied preprocessing to get noise-free data. They were using a multi-layer architecture model with the feature extraction capabilities like fundamental frequency, spectral entropy, and flatness. They mapped the data into a suitable range and extracted their features from the mapped data by use of Mel Frequency Cepstral Coefficient (MFCC).

2.3 Emotion recognition from Human Voice using pitch, tone, and rate

Emotion recognition from Human Voice using pitch, tone, and rate [1] Apart from that, El Ayadi, M., Kamel, M. S., Karray, F. (2011), ”Survey on speech emotion recognition: Features, classification schemes, and databases,” Pattern Recognition, presents a comprehensive review of all the work done in that area. It then talks of extracting these features for the purpose of emotion detection and discusses several classification schemes, multiple of which include hidden Markov models and neural networks, amongst others, and public databases for training and testing of emotion recognition systems.

3 Dataset Features

We have managed to select datasets for both Gender Recognition and Emotion Recognition, which collectively make up 3168 entries. Every entry gets characterized by attributes related to voice signals be it mean frequency, spectral entropy and modulation index. So, they help in analyzing and separating/distinguishing characteristics in voice to shape it up with gender and the state of emotion it may be in. Such models can be used to be trained upon to be able to distinguish and differentiate voices based on their acoustic properties.

Example Voice Data Value:

```
{
  ’meanfreq’: 0.059781, ’sd’: 0.064241,
  ’median’: 0.032027, ’Q25’: 0.015071,
  ’Q75’: 0.090193, ’IQR’: 0.075122,
  ’skew’: 12.863462, ’kurt’: 274.402906,
  ’sp.ent’: 0.893369, ’sfm’: 0.491918,
  ’mode’: 0.059780, ’centroid’: 0.059781,
  ’meanfun’: 0.084279, ’minfun’: 0.015702,
  ’maxfun’: 0.275862, ’meandom’: 0.007812,
  ’mindom’: 0.007812, ’maxdom’: 0.007812,
  ’dfrange’: 0.000000, ’modindx’: 0.000000
}
```

We performed EDA on the dataset to extract the valuable attributes in this project and discard the noises, which will be discussed in the later subsections.

For our emotion-based data collection, the data in general consists of 2452 audio files, with 12 male speakers and 12 Female speakers, the lexical features (vocabulary) of the utterances are kept constant by speaking only 2 statements of equal lengths in 4 different emotions by all speakers. This dataset was chosen because it consists of speech and song files classified by 247 untrained Americans to 4 different emotions at two intensity levels: Happy, Sad, Angry and Neutral along with a baseline of Neutral for each actor

3.1 Preprocessing

The dataset underwent several preprocessing operations to ensure it is clean and suitable for model training. We take benefit of two packages which makes our task easier. - LibROSA - for processing and extracting features from the audio file. - soundfile - to read and write audio files in the storage. Only audio files representing the emotions angry, sad, neutral, and happy were retained. Relevant audio features, such as MFCC, Chroma, Mel spectrogram, Spectral Contrast, and Tonnetz, were extracted and averaged to form fixed-length vectors. The data was then split into training and testing sets to ensure unbiased model evaluation. These steps are essential for training an effective emotion recognition model.

3.1.1 Feature Selection

Features selection is used for recognizing the relevant features with respect to emotional recognition, wherein features like extracting MFCCs, Chroma, Mel-spectrogram, Spectral contrast, and tonnetz are evaluated for feature importance with tree-based models like the Random Forest. That is, after selecting the most important features, redundancy can be reduced, and it can also improve the model's performance. An alternative is the use of correlation analysis to remove highly correlated features, while also employing various dimensionality reduction techniques like PCA for further optimization of the feature set.

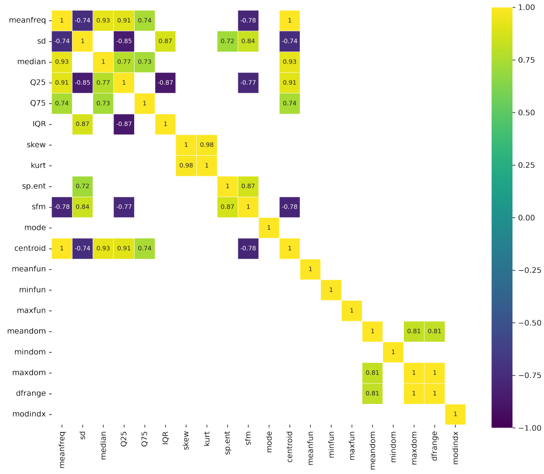

Figure 1: Correlation plot for attributes correlating above 0.7 .

3.1.2 Dimensionality Reduction using Principal Component Analysis (PCA)

PCA is a statistical procedure that uses an orthogonal transformation that converts a set of correlated variables to a group of uncorrelated variables. It is mainly used to reduce the dimensionality of the dataset, retaining most of the information. The various steps include the construction of a covariance matrix, computing eigenvectors, and using the attributes with greater values of eigenvectors. We have used PCA to detect the importance of the components of the dataset.

As we can see from Figure 3 the variance is high when there is only 1 component; however, the variance has significantly dropped to two components and drops even further as the number of components increases. This implies that as components increase, the variation between datasets decreases.

3.1.3 Visualizing High Dimension Data using t-SNE

t-Distributed Stochastic Neighbor Embedding (t-SNE) is an unsupervised, non-linear technique primarily used to explore and visualize high-dimensional data. t-SNE uses Gaussian Probabilistic Distribution to define the relationships between points in high-dimension space.

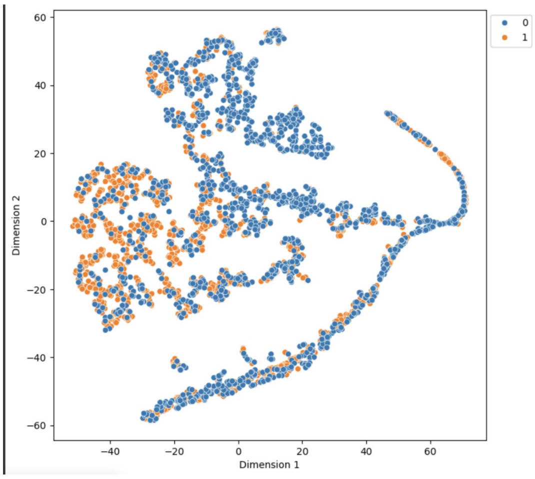

Figure 2: t-SNE of Logistic Regression of dimension 2

3.1.4 Feature Scaling / Data Standardization

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This makes the dataset center around zero mean, and the resultant distribution has a unit standard deviation. The formula for standardization:

$$z_i = \frac{x_i - \bar{x}}{s}$$

$\bar{x}$ is the mean of the feature values; s is the standard deviation of the feature values.

Our objective is to classify the speaker's gender using their voice parameters. For the Classification Problem, we used four different methodologies. We tried classification using Machine Learning models to classify the voice using acoustic parameters such as mean frequencies, Quantiles, Spectral Entropy, etc.

4 Methodologies

Our objective is to classify the gender and emotion of the speaker using their voice parameters.We want to classify a speech based on its acoustic features, such as mean frequencies, quantiles, spectral entropy, etc., and binary classification by determining whether it is male or female. We used Support Vector Machine, Naive Bayes, Logistic Regression, Decision Tree and Random Forest, .Accuracy, precision,recall, and F1 are overall grading criteria. Curves like ROC-AUC and Loss curves are also used for analysis.

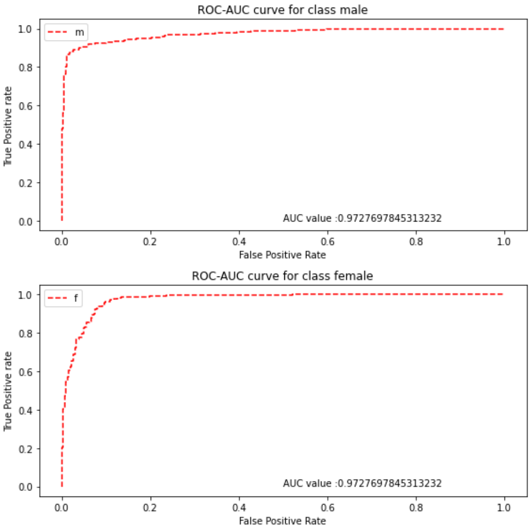

Figure 3: ROC-AUC curve for Gender Recognition

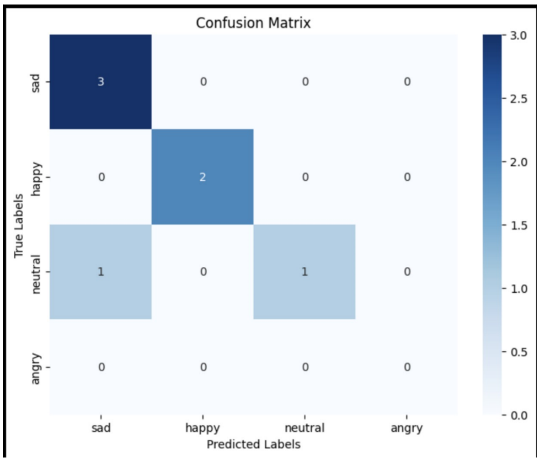

Figure 4: Confusion Matrix For Random Forest(Emotion Recognition)

- 1. **Logistic Regression**: logistic model with L2 regularization, suitable for binary classification.
- 2. **Decision Trees**: The data gets split into categories with tree-like decision boundaries.
- 3. **Random Forests**: An ensemble of decision trees, where collective results improve performance.
- 4. **Naïve Bayes**: Probabilistic model assuming feature independence; Gaussian and Bernoulli variations used.
- 5. **SVM**: This makes the optimal hyperplanes for classification by using kernel tricks and Cover's theorem for non-linear separations.

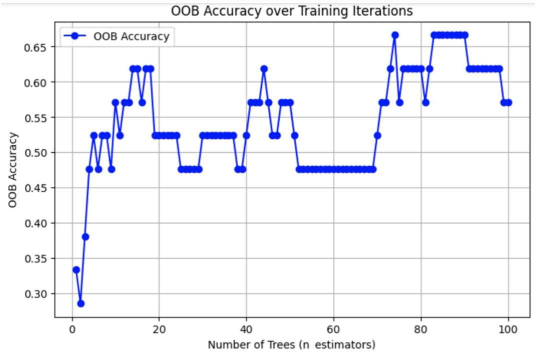

Figure 5: Plot for accuracy for the Random Forest Model (Emotion Recognition))

4.1 Model and Details

We used binary and multi-class models with an 80:20 train-test split, standardized data, and optimized parameters using 10-fold cross-validation. Audio features like MFCCs and Mel spectrograms were extracted, and SVM, Random Forest, and Decision Tree models were trained for emotion classification.

4.2 Performance Metrics

We evaluated the performance mainly based on accuracy, which measures how well the model classifies new data. Other metrics such as recall and F1 score were also considered. High accuracy is important because it is an indicator of the ability of the model to generate correct predictions. The ROC curve was analyzed, and the Area Under the Curve (AUC) was observed. Higher AUC values indicate stronger predictive performance and overall model effectiveness.

5 Result and Analysis

5.1 Classification for Gender Recognition

We performed Binary Classification to classify the speaker's voice as Male or Female. The dataset of randomly sampled to produce better results.
The Linear SVM model gave the best performance of the 4 models used with a mean Accuracy score of 97.538 %, across all the classes. The Logistic Regression model gave a similar result with the mean accuracy being 97.1591 %. Though, the other 2 models namely , Gaussian Naive Bayes and Bernoulli Naive Bayes had good accuracy , it was less than the Linear SVM and Logistic Regression Model , with GNB having an Average Accuracy of 92.8030 % and BNB having an Average Accuracy of 87.404 %.

Table 1: Binary Classification Metrics

Model	Accuracy	Precision	Recall	F1
LR	0.97	0.97	0.97	0.97
GNB	0.92	0.92	0.92	0.92
BNB	0.87	0.87	0.87	0.87
Linear SVM	0.97	0.97	0.97	0.97

5.2 Classification for Emotion Recognition

The Random Forest model gave the best performance of the 3 models used with a mean Accuracy score of 76.595 %, across all the classes. The other models, like Decision Tree and SVM, did give satisfactory results, with the lowest being from the Decision Tree's side with an accuracy of 66 %. The SVM model had an accuracy of well over 75 %.

Table 2: Binary Classification Metrics

Model	Accuracy	Precision	Recall	F1
DT	0.66	0.67	0.66	0.66
RF	0.76	0.70	0.72	0.71
SVM	0.74	0.74	0.75	0.74

6 Conclusion

6.1 Outcomes

In our study, we managed to go about two different datasets to handle two aspects of voice recognition: Gender Recognition and Voice Recognition. As mentioned, we used a dataset with parameters of the type like meanfun, sfm, Q25, etc., with RBF Kernel for SVM showing the best performance across models used for this dataset. As opposed to that, for emotion detection we use a different type of dataset derived from .wav files for which preprocessing is a bit unique in its own format. For emotion recognition, the Random Forest model outperforms the other models quite effectively being able to capture the robustness in all the features as represented in it all. These methods when combined can easily provide a robust nature in model-processing and an understanding for future technologies.

6.2 Member Contibution

Ansh Varshney: - Pre-processing and Data Visualization, Feature Extraction, SVM, Random Forests, PPT

Abdullah Shujat: - Pre-processing and Data Visualization, Logistic Regression, Naive Bayes, Random Forests, Report Writing

Anant Kaushal: - Data Collection and Visualization, SVM, Decision Trees, PPT

Anikait Agrawal: - Data Collection, Logistic Regression, Naive Bayes, Decision Trees, Report Writing

References

[1] Sung-Woo Byun and Seok-Pil Lee. Emotion recognition using tone and tempo based on voice for iot. *The transactions of The Korean Institute of Electrical Engineers*, 65(1):116–121, 2016.

[2] Ioannis E Livieris, Emmanuel Pintelas, and Panagiotis Pintelas. Gender recognition by voice using an improved self-labeled algorithm. *Machine Learning and Knowledge Extraction*, 1(1):492–503, 2019.

[3] Mohammad Amaz Uddin, Md Sayem Hossain, Refat Khan Pathan, and Munmun Biswas. Gender recognition from human voice using multi-layer architecture. In *2020 International conference on innovations in intelligent systems and applications (INISTA)*, pages 1–7. IEEE, 2020.