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.

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 Review

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 picked the Voice Gender Dataset and Voice Emotion Dataset, which consist of 3168 data entries consisting of two, each having the values concerning the attributes of the below-mentioned type

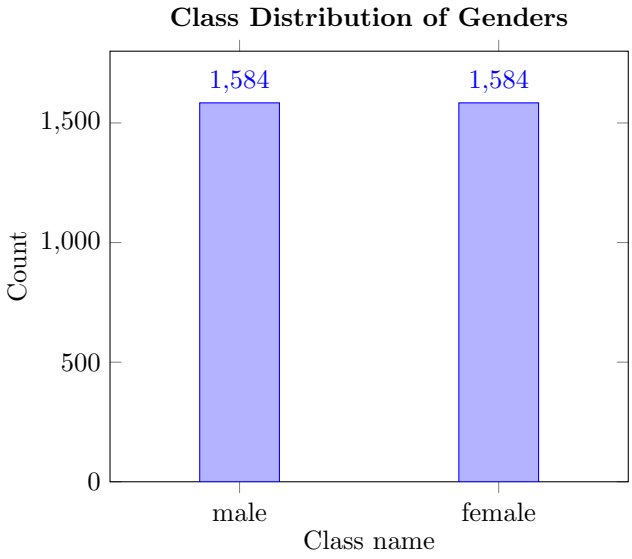
Table 1: Feature Descriptions

MODEL	DATATYPE
MEANFREQ	float
SD	float
MEDIAN	float
Q25	float
Q75	float
IQR	float
SKEW	float
KURT	float
SP.ENT	float
SFM	float
MODE	float
CENTROID	float
MEANFUN	float
MINFUN	float
MAXFUN	float
MEANDOM	float
MINDOM	float
MAXDOM	float
DFRANGE	float
MODINDEX	float

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 1440 supervised audio samples from 24 actors (12 male + 12 female), capturing a diverse range of emotions. It encompasses 8 distinct emotions neutral, calm, happy, sad, angry, fearful, disgust, and surprised.

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
}
```

3.1 Preprocessing

The dataset obtained from voice genders was vast; hence we used the EDA plots, removed the attributes with the same values for each datapoint, etc., to remove useless columns, i.e., we performed feature selection.

3.1.1 Feature Selection

We performed feature selection by looking at the heat maps, which show the correlation between attributes. After completing EDA plots, we checked which characteristics have the most repetitive value for each datapoint and removed that attribute, such as 'mindom.'

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.

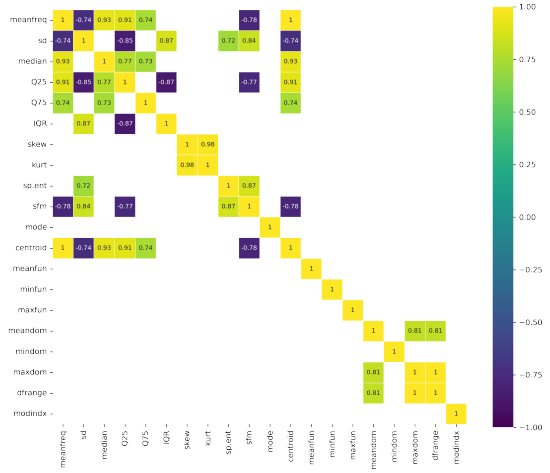


Figure 1: Correlation plot for attributes correlating above 0.7 .

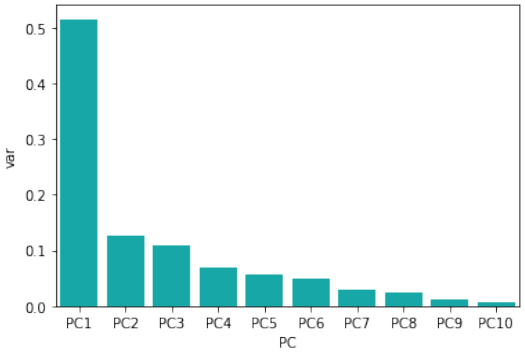


Figure 2: **PCA explained Variance.**

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.

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:

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

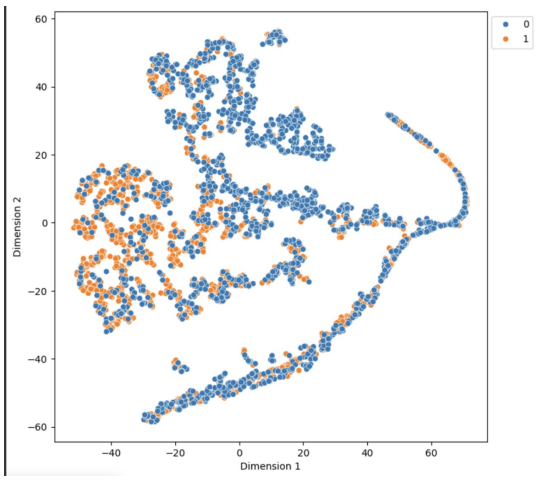


Figure 3: t-SNE of Logistic Regression of dimension 2

$\bar{x}$ 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 Methodoligies

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

4.1 Classification

We applied binary and multi-class classification models to the data points available in the dataset. We performed an 80:20 train validation split and standardized the data in the data frame.
We used the following classification models: Logistic Regression, Gaussian Naïve Bayes, Bernoulli Naïve Bayes for classification purposes. To further optimize various parameters in the previously mentioned models, we performed 10 – fold Cross Validation

5 Result and Analysis

5.1 Classification

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 percent, across all the classes. The Logistic Regression model gave a similar result with the mean accuracy being 97.1591

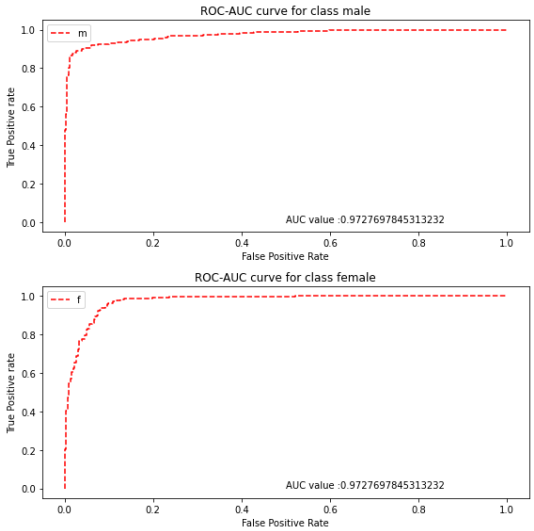


Figure 4: ROC-AUC curve for Logistic Regression

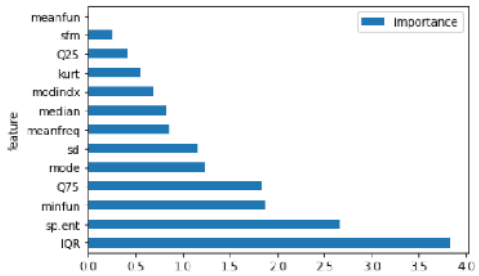


Figure 5: Feature Importance before feature selection

percent. 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 percent and BNB having an Average Accuracy of 87.404 percent.

Below mentioned are all the metrics of the models that were used:

Table 2: 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

We have also computed feature importance of the features before and after feature selection, as shown below:

As we can see from the plots, "IQR" has the maximum importance for song classification. This implies that the interquartile range of a voice is an essential attribute while classifying the voice as male or female. Other features such as "minifun", "sp.ent", and "Q75" are also necessary for voice analysis.

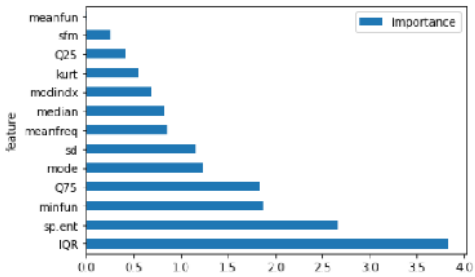


Figure 6: Feature Importance after feature selection

6 Conclusion

6.1 Outcomes

So far, the work has proposed using acoustic features (meanfun, sfm, Q25, kurt, modindex, median, meanfreq, sd, mode, Q75, minfun, sp.ent, IQR) to predict the Gender of Human Beings using their voice as input. The work has examined using a dataset of speech parameters with different machine-learning models. The results tell that Linear SVM has provided the best score for each of the metrics chosen to judge each model. The Logistic Regression model gave a satisfactory results compared to other models like Gaussian and Bernoulli Naive Bayes.

6.2 Work Left

The project progress has been on and forward to the schedule that had been proposed. The work that remains ahead of us is to work on the models such as Random Forest, Decision Tree, Analysis, and Performance of models, as well as checking for overfitting and underfitting of the model for gender-based and emotion-based datasets.

6.3 Member Contibution

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

Abdullah Shujat: - Pre-processing and Data Visualization, Logistic Regression, Naive Bayes

Anant Kaushal: - Data Collection and Visualization, SVM

Anikait Agrawal: - Data Collection, Logistic Regression, Naive Bayes

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

[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. 1

[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. 1