GENDER RECOGINITION USING VOICE

Arjun Mehra
arjun20173@iiitd.ac.in

Aryman Srivastava
aryman20184@iiitd.ac.in

Dheeraj
Dheeraj20194@iiitd.ac.in

Khushdev Pandit
Khushdev20211@iiitd.ac.in

Abstract

Speech is the most popular means of human communication. It is used to express their emotions, cognitive states, and intentions to each other. In general, a speech and voice recognition system can be used for gender identification. The human ear is a natural voice recognition system. We can teach a machine to do the same task using modern machine-learning algorithms and visualization tools. By incorporating the algorithms with the correct set of features, we can prepare a machine to recognize the voice for us. We will use features such as mean frequency(kHz), Standard Deviation, First Quantile, Third Quantile, etc., to make a model learn voice features and recognize the gender of the speaker.

1. Introduction

The Voice contains many effective communication methods consisting of 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 voice 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 in improving the classifier's efficiency. After getting the extracted features and labels, ML techniques are used to build a high-quality classifier for gender recognition. The most efficient classifiers and feature extractors for gender recognition by voice include Deep Neural Networks and Convolution Neural Networks since both techniques show robustness to the changes and have low noise sensitivity.

The progress in technological fields 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 Decision Tree, for classifying the speaker's gender to 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 class by gender categorization, automatic salutations, muting sounds for gender, and audio/video categorization with tagging.

2. Literature Review

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

2.1 Gender Recognition by Voice using an Improved Self-Labeled Algorithm **[1]** by Ioannis Liveris, Emmanuel G Pintelas, and P.E. Pintelas uses a hybrid of Ensemble Learning and Semi-Supervised Learning (SSL) algorithms called iCST-Voting, for Gender Recognition by Voice. They demonstrate the classification efficiency of the proposed algorithm in terms of accuracy for stable and robust predictive models.

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 **[2]** by Mohammad Amaz

Uddin, Md Sayem Hossain, Refat Khan Pathan, and Munmun Biswas narrates about extraction of the features from the audio speech to recognize gender as male or female and use those features to recognize the gender of the speaker.

At first, the authors performed preprocessing to get noise-free data. They used a multi-layer architecture model to extract features such as fundamental frequency, spectral entropy, and flatness. They mapped the data into a suitable range and used Mel Frequency Cepstral Coefficient (MFCC) to extract the features from the mapped data.

Three different datasets were extracted, and accuracy was plotted using two classifiers: - Support Vector Machine and K-Nearest Neighbors. The best accuracy was about 96.8% using K-Nearest Neighbors. We will be implementing SVM in the further progress of the project.

3. Dataset Features

We have picked the Voice Gender Dataset, which consists of 3168 data entries consisting of two, each having the values concerning the attributes of the below-mentioned type: -

FEATURES	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

Table 1: Raw Features

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.

Example Voice Data Value:

{‘meanfreq’: 0.059781, ‘sd’: 0.064241,
‘median’: 0.032027, ‘Q25’: 0.015071,
‘Q75’: 0.90193, ‘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}

The class distribution of genders is shown below:

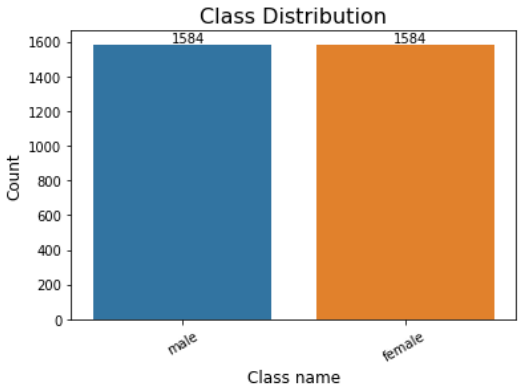

Figure 1: Class Distribution of Genders.

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 heatmaps, which tell 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.’

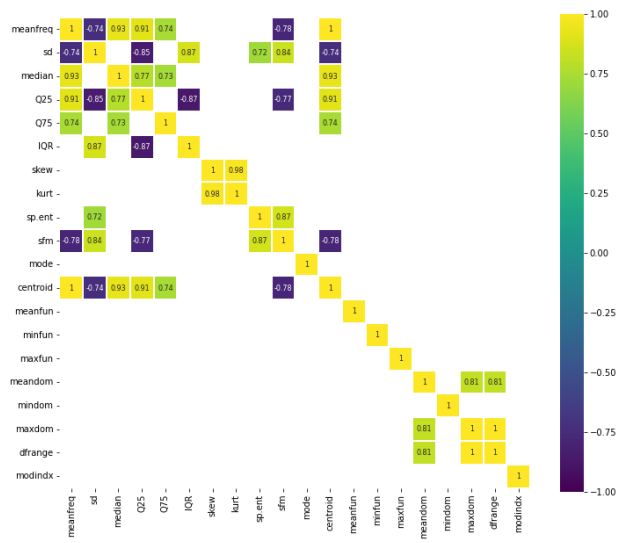

Figure 2: Correlation plot for attributes correlating above 0.7

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.

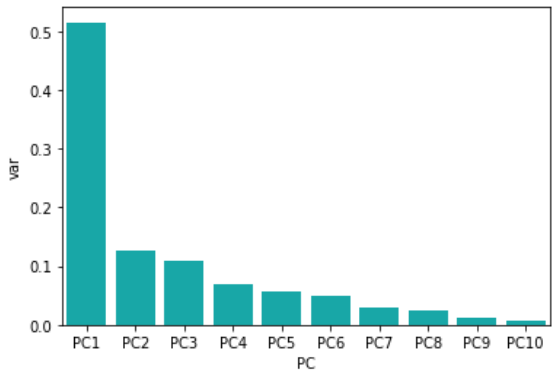

Figure 3: PCA Explained Variance

As we can 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 increase. 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.

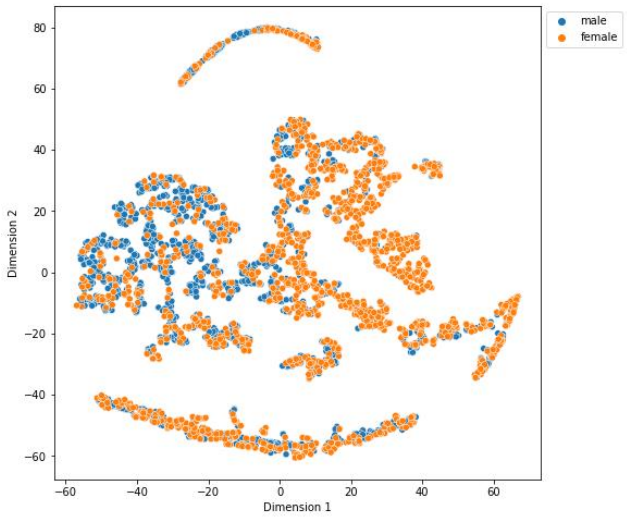

Figure 3: t-SNE of Logistic Regression of dimension 2

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:

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

x is the mean of the feature values; s is the standard deviation of the feature values.

4. Methodologies

Our objective is to classify the gender of the speaker using their voice parameters. For 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, and Decision Tree (Depth = 4, 8, 10, 15, 20) for classification purposes. To further optimize various parameters in the previously mentioned models, we performed 10 – fold Cross Validation.

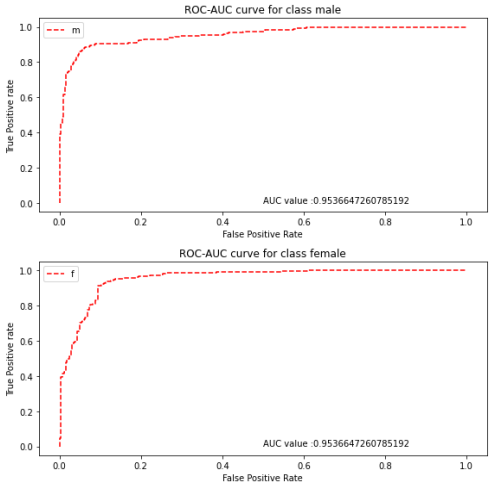

Figure 4: ROC-AUC curve for Gaussian Naïve-Bayes for Male and Female Classes

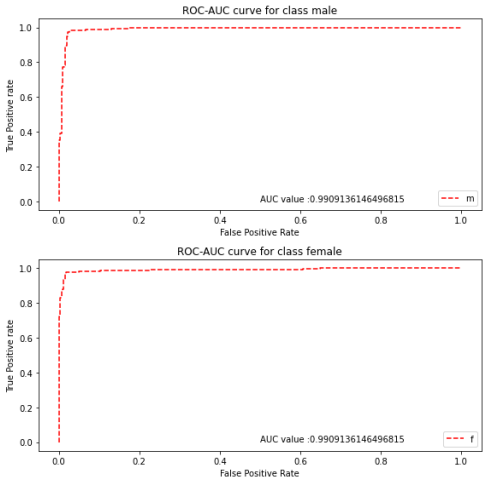

Figure 5: ROC-AUC curve for Logistic Regression for Male and Female Classes

5. Results 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 Logistic Regression gave the best performance of the two models, with an Average Accuracy score of 97.25%, across all the classes. The Decision Tree gave a similar result with depth=10, using "Gini" as the criterion, and a Recall score of 97.51%, going further down with Gaussian Naïve-Bayes having a Precision accuracy of 92.87 %, and Bernoulli having a Precision score of 87.98%. Below mentioned are all the metrics of the models that were used:

MODEL	ACCURACY	PRECISION	RECALL	F1
LR	0.9725	0.9727	0.9725	0.9726
GNB	0.9287	0.9307	0.9287	0.9286
BNB	0.8747	0.8798	0.8747	0.8744
DT (WITH GINI AND DEPTH=10)	0.9716	0.9690	0.9751	0.9720

Table 2: Binary Classification Metrics

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

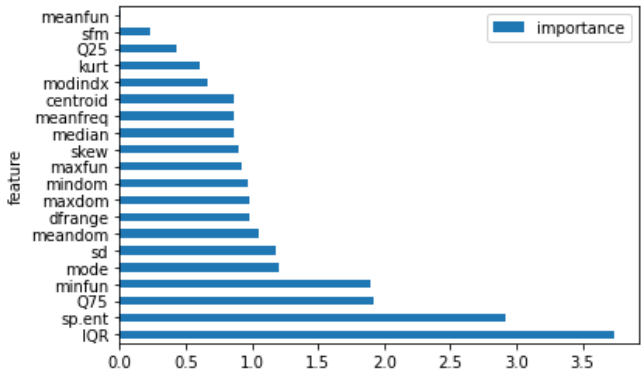

Figure 6: Feature Importance before feature selection

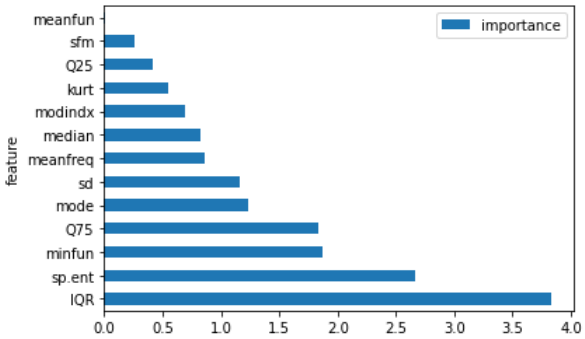

Figure 7: Feature Importance after feature selection

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 “minfun,” “sp.ent”, and “Q75” are also necessary for voice analysis.

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 Logistic Regression has provided the best score for each of the metrics chosen to judge each model. The decision Tree having a depth of ten and Gini as a criterion gave satisfactory results compared to the Logistic Regression model.

6.2. Work Left

The project progress has been on and forwards to the schedule that had been proposed. The work that remains ahead of us is to work on the models such as SVM, Random Forest, Analysis, and Performance of models, as well as checking for overfitting and underfitting of the model.

6.3. Member Contribution

Arjun Mehra: - Data Collection and Analysis, Naïve Bayes Model, PPT

Aryman Srivastava: - Exploratory Data Analysis, Feature Selection, Report

Dheeraj: - Data Collection and Pre-processing, Decision Tree Model, Report

Khushdev Pandit: - Pre-processing and Data Visualization, EDA, Logistic Regression, PPT

7. References

[1] https://www.researchgate.net/publication/331536653_Gender_Recognition_by_Voice_using_an_Improved_SelfLabeled_Algorithm/fulltext/5c7f272f299bf1268d3cdc17/Gender-Recognition-by-Voice-using-an-Improved-SelfLabeled-Algorithm.pdf

[2] https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9194654

[3] https://www.hindawi.com/journals/sp/2019/7213717/

[4] https://www.mdpi.com/2504-4990/1/1/30/pdf