GENDER RECOGNITION BY VOICE

Arjun Mehra
arjun20178@iiitd.ac.in

Aryman Srivastava
aryman20184@iiitd.ac.in

Dheeraj
dheeraj20194@iiitd.ac.in

Khushdev Pandit
khushdev20211@iiitd.ac.in

Abstract

Speech is one of the significant attributes of expressing human communication. Humans use voices to express their emotions, thoughts, and intentions to other people. Many attributes, such as facial structure, hair length, body structure, and voice, can be used for gender recognition. In our project, we used voice as a focal point to classify the gender of the speaker. Modern machine learning techniques and visualization tools can help us find important features and accurate models to classify gender. By incorporating the algorithms with the correct set of features, we can prepare a machine to recognize the voice for us. The feature set we will use for this are the mean frequency (kHz), Standard Deviation, First Quantile, Third Quantile, etc., to make the models learn voice features, find the best set of attributes as a classifier and ultimately recognize the gender of the speaker.

Project GitHub link: [GitHub]

1. Introduction

The Voice has a variety of efficient communication techniques that include linguistic and paralinguistic criteria, including gender, age, language, and others. Voice and sound analysts who use a variety of applications, such as analyzing criminal voices in crime situations, emotion detection, and strengthening human-computer interface, have found it challenging to determine human gender based on the voice. The dataset's characteristics determine the resilience and efficiency of classification models; hence, extracting information from the raw data is essential to increase the classifier's performance. A high-quality classifier for gender recognition is created using ML approaches once the extracted features and labels have been acquired. 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 ways of doing work have also improved due to technological advancement. We gathered the datasets for this study from internet sources, concentrating primarily on the qualities common to

most of the datasets. To further evaluate the efficacy of these models, we have suggested employing binary and multiclass classification methods, such as Logistic Regression, Naive Bayes, and Decision Tree, for categorizing the speaker's gender. Additionally, boosting approaches and graphical-based procedures have been used to enhance the learning of the models by selecting features and eliminating noise from the collected raw data. Our research has applications in voice emotion identification, human-to-machine interaction, gender-based telephone class sorting, automatic salutations, gender-specific sound muting, and audio/video classification 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 pre-processing to get noise-free data. They used a multi-layer architecture model to extract fundamental frequency, spectral entropy, and flatness features. 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 Neighbours. The best accuracy was about 96.8% using K-Nearest Neighbours. We will be implementing SVM in the further progress of the project.

3. Dataset

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}

3.1. Pre-processing

The Voice Gender Dataset consists of extended attributes. Hence, we used EDA plots, such as heatmaps, to get the correlation between the features and removed the features with a correlation of more than 0.7. We also removed extra features by looking at the data points of the features. Features with a repetitive amount of repeat data points were removed as well. Using the above methods, we performed feature extraction and data cleaning for better performance of the model.

3.1.1. Feature Selection

We plot the correlation heatmap of the features to make the following observations:

1. Meanfreq has a high positive correlation with the centroid (1), median (0.93), Q25 (0.91), Q75(0.74), and a high negative correlation with sfm (-0.78), and sd (-0.74).
2. Sd has a high positive correlation with IQR (0.87), sp.ent (0.72), sfm (0.84), and a high negative correlation with Q25 (-0.85), centroid (-0.74).

We then used pair plots to verify correlations observed from the heatmaps. We observed that meanfreq and centroid follow the same increasing path and approximately increasing similarly with median and Q25. Hence after attending the same for other feature sets, features such as skew, maxfreq, centroid, and dfrange were removed from the final collection of features. These features were dropped because of a high positive and negative correlation with features like centroid.

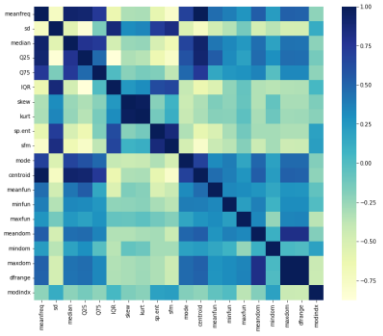

Figure 1: Correlation plot for various attributes in Dataset

3.1.2. 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. After data standardization, the remaining datasets had zero mean and unit standard deviation.

3.1.3 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.

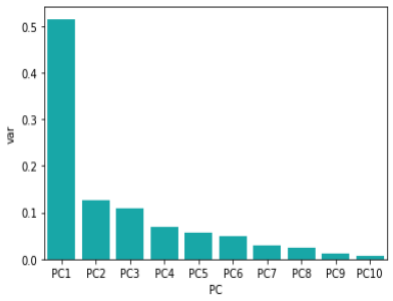

Figure 2: PCA Explained Variance

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 increases. This implies that as components increase, the variation between datasets decreases. We used PCA to reduce the dimensions of the data points into multiple dimensions and plot the variation in the dataset for each component against the number of components.

3.1.4 t-SNE:

T-Distributed Stochastic Neighbour 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.

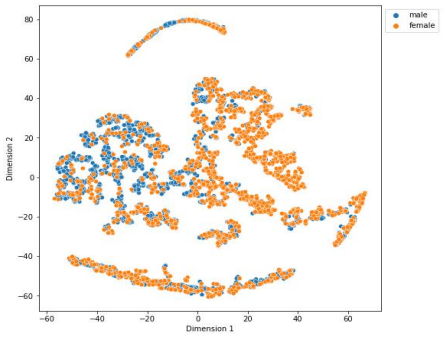

Figure 3: t-SNE of Logistic Regression of dimension 2

4. Methodologies

We want to categorize a speech based on its auditory characteristics, such as mean frequencies, quantiles, spectral entropy, etc., and binary classifications by identifying whether it is male or female. We employed Support Vector Machine, Naive Bayes, Logistic Regression, Decision Tree, Random Forest, and Artificial Neural Network. Accuracy, precision, recall, and F1 are overall grading criteria. Curves such as ROC-AUC and Loss curves are also used for analysis.

4.1. Model and Details

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. To further optimize various parameters in the previously mentioned models, we used 10 – fold Cross Validation.

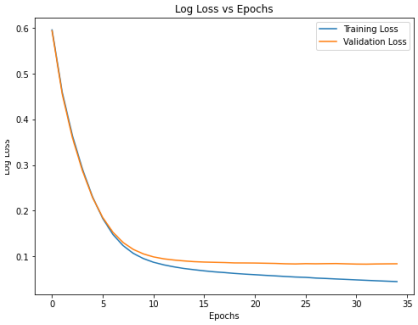

Figure 4: Loss curve for ANN (MLP)

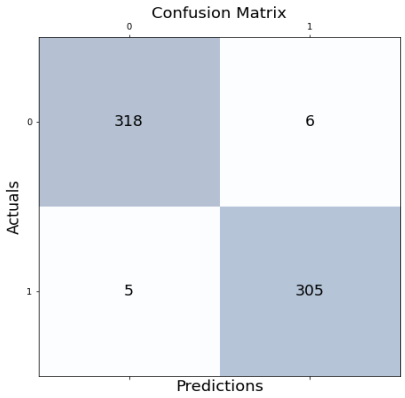

Figure 5: Confusion Matrix for Random Forest with max depth=10

1. **Logistic Regression:** A logistic function is employed in a linear binary classification model. We applied L2 regularisation to it.
2. **Decision Tree:** A classification model categorizes data using decision boundaries. Using the features, it creates a tree-based model and constantly divides the data according to the parameters to achieve a final result.
3. **Random Forest:** It is a decision tree ensemble model that integrates the results from several decision trees. It functions more effectively than standard decision trees because of ensemble approaches.
4. **Naïve Bayes:** It consists of various probabilistic models that assume feature independence to assign probabilities to output classes. For our purpose, we used Naïve Bayes models, Gaussian and Bernoulli.
5. **MLP:** MLP or Artificial Neural Network is a class of Neural Networks that mainly comprises of completely interconnected neurons.
6. **SVM:** It is a classification model that generates optimal hyperplanes for separating classes. We can employ strategies like soft-margin and kernel trick if the data cannot be linearly separated (to increase the dimensions of the dataset). Cover's theorem is used when increasing the dimensions of the dataset for classification.

4.2. Performance Metric

The main parameter we used to assess overall performance was accuracy. The effectiveness of the models was further evaluated using other measures, including accuracy, recall, and F1. In general, high accuracy is preferred. Accuracy is crucial because it indicates how many accurate classifications the model can produce given a new data point to categorize. The ROC curve is plotted, and Area under the Curve (AUC) is observed. Higher AUC is desired and indicates better model prediction ability and good performance.

5. Result and Analysis

To determine the ideal parameters, we train our models using 10-fold cross-validation. Using the best model discovered by the best parameter search, we calculated accuracy, precision, recall, and F1-Score (using accuracy as the scoring metric since we want more correct classifications). Below is an overview of the models used with the test data:

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
ANN (ReLU activation function)	0.9823	0.9824	0.9823	0.9823
Random Forest	**0.9825**	**0.9838**	**0.9807**	**0.9822**
SVM	0.9754	0.9755	0.9754	0.9754

Table 2: ML model's Performance Metrics

Random Forest performed best among all the models. This is not surprising as the random forest is an ensemble model of decision trees; it combines the output of various decision trees.

As evident from the t-SNE plot, the data is not linearly separable. The ANN also performed exceptionally in this case, given that the network chosen was not very dense (2-layer).

After running multiple SMV, the linear SVM gave the best results. It gave better results than LR, NB, and DT models. This is because it tries to find the optimal hyperplane which classifies the unseen data better.

Logistic Regression and Decision Trees have comparable results even though the data was not linearly separable, as evident from the t-SNE plot.

The Naïve Bayes models gave the worst result as these models work on the assumption that features are independent, which is not valid.

We also performed feature importance in Random Forest.

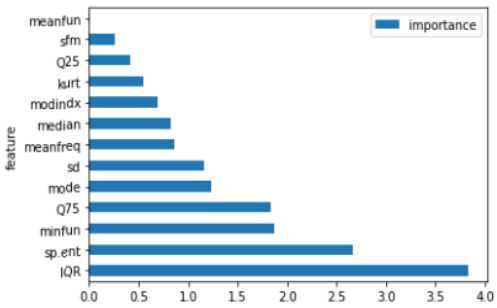

Figure 7: Feature Importance plot

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. Outcome

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 Random Forest has provided the best score for each of the metrics chosen to judge each model. The Support Vector Machine has similar results as compared to Random Forest. The other models also gave comparable results, Naïve Bayes being the worst.

6.2. Learning

We gained knowledge about data visualization and analysis techniques via the project. EDA approaches, and data pre-processing procedures were covered. We used models including logistic regression, decision trees, random forests, MLP, SVM, and Naive Bayes. We also investigated several dataset types and learned about the significance of data in machine learning applications. We learned how to evaluate the performance of our models' using measures like recall, precision, and accuracy and to illustrate this performance using curves like the ROC-AUC and Loss Curve.

6.3. Future Work

Better results can be obtained by making the MLP deeper. Grid Search can be used to get a better insight into the optimum parameter rather than the trial-and-error method. The further development in the project will be to increase the number of genders we can classify using different models.

6.4. Member Contribution

1. Arjun Mehra: Data Collection and Analysis, Naïve Bayes Model, Random Forest, SVM.

2. Aryman Srivastava: Exploratory Data Analysis, Feature Selection, MLP, SVM, Report.

3. Dheeraj: Data collection and Pre-processing, Decision Tree, Random Forest, Report, PPT.

4. Khushdev Pandit: Pre-processing and Data Visualization, EDA, Logistic Regression, MLP, SVM, PPT.

7. References

[1] https://www.researchgate.net/publication/331536653_Gender_Recognition_by_Voice_using_an_Improved_SelfLabeled_Algorithm/fulltext/5c7f272f299bf1268d3cdc17/GenderRecognition-by-Voice-using-an-Improved-SelfLabeledAlgorithm.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