Gender Recognition by Voice

Priyanka Makwana (Team-4)
California State University, Sacramento
priyankamakwana@csus.edu

Abstract

The task of identifying a human's gender by voice seems an easy task when a human identifies it. It becomes difficult when a computer has to identify it whether the voice is of male or female. A human has natural capability of identifying the difference but when it comes to computer we need to teach it by providing inputs, methodology or different training data and make it learn. In this project, the focus is on training computer to identify the gender based on input of acoustic attributes using various Machine Learning algorithms and get the best results.

Keywords: Gender Recognition; acoustic attributes, voice information

1. Background

The most common means of communication is speech signal. The recorded speech can serve as our input to the system. The system processes this speech signal to get acoustic attributes. Analyze the input and compare it with the trained data, perform computations based on the algorithm used and provides the closest matching output. There are various applications where gender recognition can be helpful. Some of them are:

- for further identification of human sounds like male laughing, female singing and more.
- categorizing audios/videos by adding tags and simplifying and reducing search space
- automatic salutations
- muting/saving sounds for a gender
- can help personal assistants like Siri, Google Assistant to answer the question with female generic or male generic results.

2. Method

There is a direct relation between data and model. A model is only as good as the data. The dataset used for recognizing gender based on audio events is retrieved from Kaggle (Gender Recognition by Voice). It is a large-scale collection of all frequency attributes of voice of both genders. The dataset is a comma separated values file which is derived from converting sound waves for both genders. It has 21 attributes and more than 3K entries of sound clip information equally distributed of both genders. Also, dataset has equal entries for male and female. This problem falls under supervised learning as the system is trained by providing large number of labelled inputs for both genders. Also, it is a classification problem as the output is based on classifying into male or female. The goal is to compare outputs of various models and suggest the best model that can be used for gender recognition by voice.

2.1 Approach

Considering a simple case, one can say there is difference of frequency of voice of male and female and based on frequency one can differentiate genders. However, frequency can vary widely within a spoken word, let alone an entire sentence. Frequency rises and falls with intonation, often to communicate certain emotion within words and speech. This can make it difficult to pinpoint an exact frequency. Figure 1 shows the work flow of this project. The dataset is made up of acoustic attributes of different voices. It is passed through the specan function of warbler package to get all attributes. Considering all the attributes, we found correlation between them and reduced the number of attributes by preprocessing it using PCA (Principle Component Analysis). The outcome is applied with following algorithms: Random Forest, CART model, SVM, XGBoost and Stacked Ensemble model.

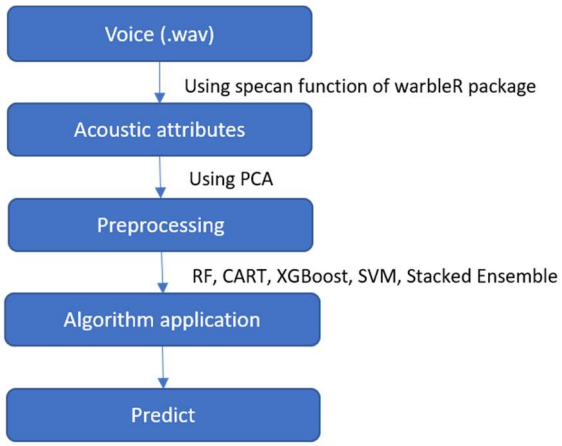


Fig.: 1 Work flow to recognize gender of voice

2.2 Technology and Resources

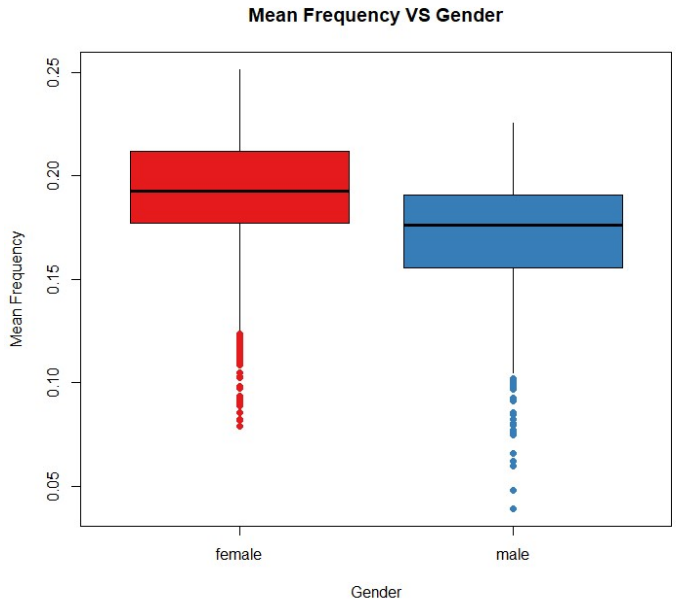
This project is implemented using:

- Language: R
- IDE: R Studio

The packages used are caTools, corrplot, caret randomForest, party, Matrix, ggplot2, dplry, RColorBrewer, ElemStatLearn, e1071, rpart, and xgboost.

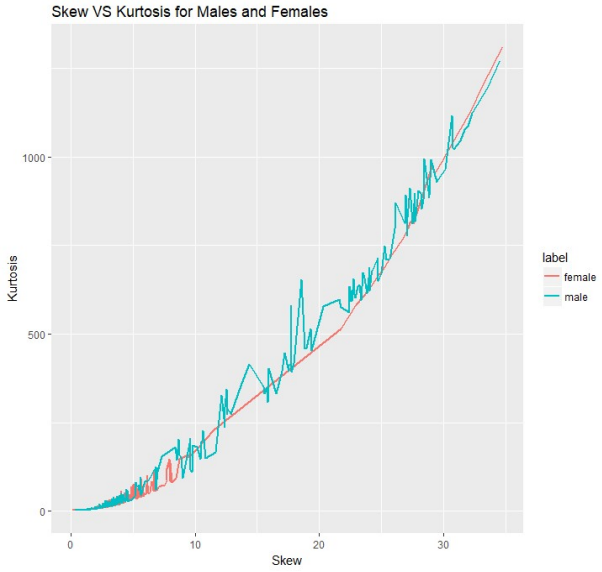
3. Implementation

In the implementation process, initial thing done is understanding of how the dataset is derived. The sound samples are collected from [5], [6], [7] and [8]. These samples are stored as a .WAV file, which are then pre-processed for acoustic analysis using warbleR package in R. The result is the 22 acoustic attributes of the provided sound. Our dataset is madeup of these attributes ignoring one attribute i.e. duration, as duration of all samples is 20 seconds. Therefore, the dataset has 21 attributes. The attributes consist of mainly statistical computations of the frequency (Appendix A). The frequency of female voice is higher than a male voice. After understanding the dataset formulation, data exploration was done. Visualizing the parameters and its dependency gives a better understanding.

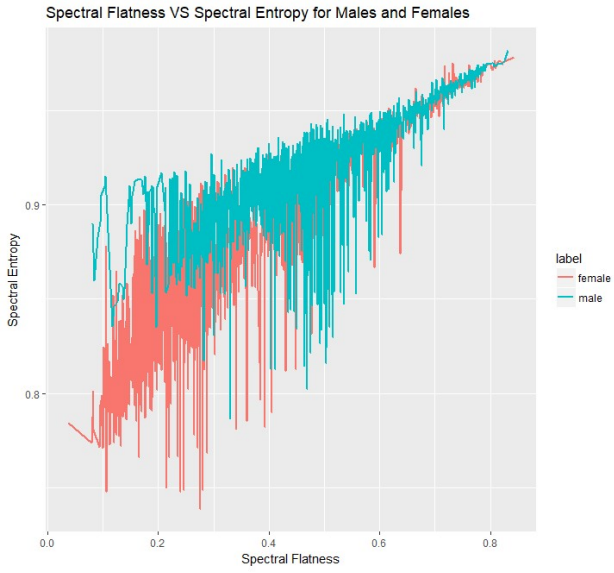


The image shows difference in mean frequency of male and female voices which

clearly shows male and female voices cannot be determined only based on mean frequency.

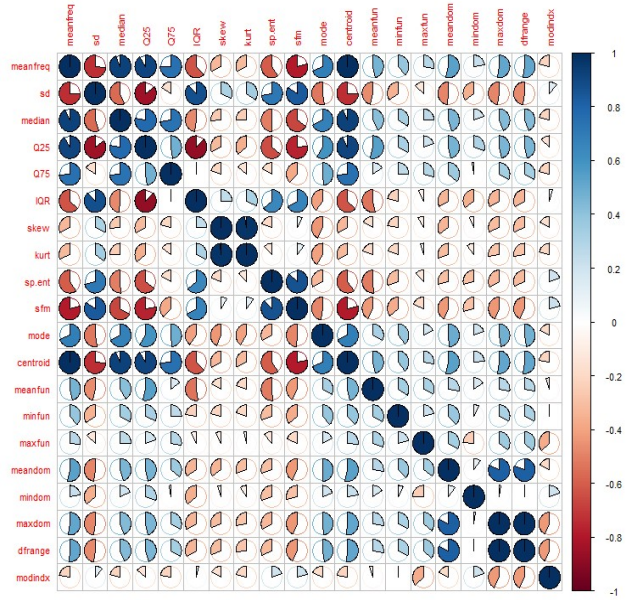


The skewness and kurtosis of male and female acoustic data.

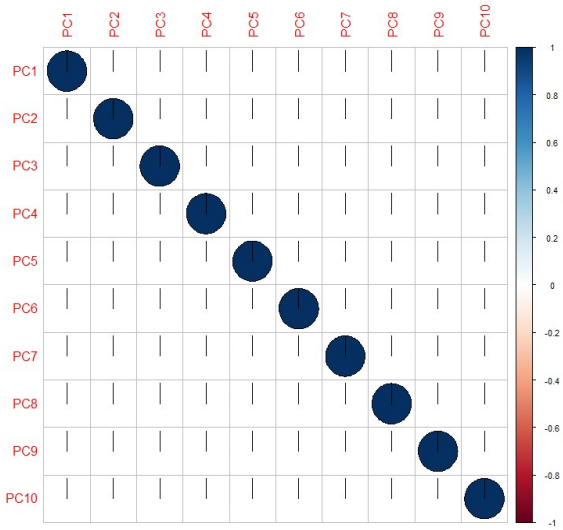


The spectral flatness and spectral entropy of male and female acoustic data in the dataset.

This analysis showed there is lot of dependencies among all attributes. As, all the data is numerical we can get correlation between them and analyze further. Following graph shows correlation among all attributes:



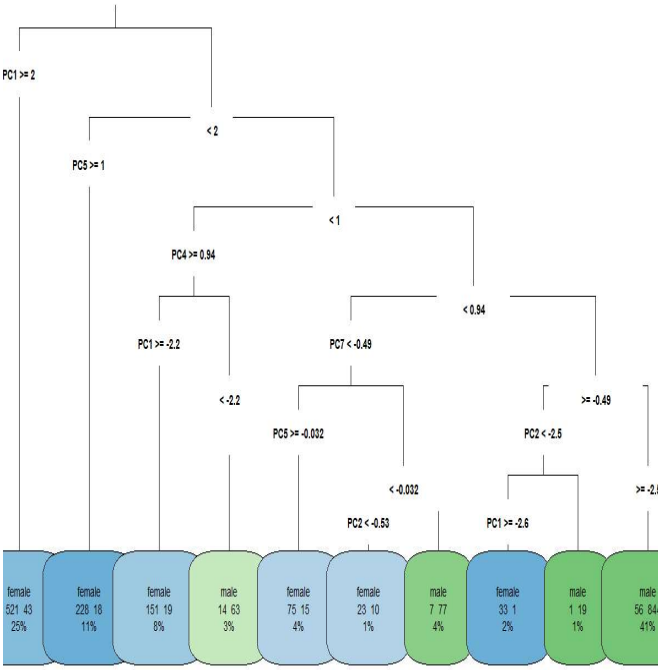
By applying PCA (Principle Component Analysis), one can see the correlation being removed among the attributes.



This pre-processed data is now ready to for algorithm application.

The algorithms selected are based on research work and papers that showed which algorithms works best.

Following image shows the classification tree generated using CART model.



4. Result

The result is the accuracies from all models. These accuracies are compared between and model with the highest accuracy is considered best model for this kind of data. The purpose of using multiple models is to identify the best sound classification model, which can be helpful in developing future applications. The table shows the achieved accuracies of different models compared to the prior accuracies.

Results achieved after applying algorithms

Algorithm	Prior Accuracy	Current Accuracy
Random Forest	87%	96.74%
CART	81%	89.89%
SVM	85%	98%
XGBoost	-	97.26%
Stacked Ensemble	89%	98%

The results suggest, the algorithms with highest accuracy are SVM and Stacked Ensemble. Stacked Ensemble is derived using results of all algorithms. As SVM is derived directly, it can be considered as best for this

5. Conclusion

For this dataset SVM can be considered best as it provided best results. There are many factors that can affect and prove different dataset better for acoustic results. Came up with this conclusion, because I understood the process of how the dataset was built so that I can apply the same process and provide an input voice to check its gender. The update of the library by the author produced different result than how the dataset was built. This made an issue to match the attributes with dataset.

6. References

[1] Gender identification from speech signal by examining the speech production characteristics (http://ieeexplore.ieee.org/document/7980584/)

[2] Speech Recognition with Deep Learning (https://medium.com/@ageitgey/machine-learning-is-fun-part-6-how-to-do-speech-recognition-with-deep-learning-28293c162f7a)

[3] Gender Clustering and Classification Algorithms in Speech Processing: A Comprehensive Performance Analysis (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.258.9728&rep=rep1&type=pdf)

[4] Dataset: Voice gender dataset (https://www.kaggle.com/primaryobjects/voicegender/data)

[5] Dataset derived from: The Harvard-Haskins Database of Regularly-Timed Speech

[6] Dataset derived from: Telecommunications & Signal Processing Laboratory (TSP) Speech Database at McGill University

[7] Dataset derived from: VoxForge Speech Corpus

[8] Dataset derived from: Festvox CMU_ARCTIC Speech Database at Carnegie Mellon University

[9] Tutorial for R is very helpful and easy to understand here - https://www.tutorialspoint.com/r/index.htm

[10] Working demo http://www.primaryobjects.com/2016/06/22/identifying-the-gender-of-a-voice-using-machine-learning/

[11] Referring in-class examples for various models

APPENDIX – A

Following is the list of attributes in dataset:

- **meanfreq**: mean frequency (in kHz)
- **sd**: standard deviation of frequency
- **median**: median frequency (in kHz)
- **Q25**: first quantile (in kHz)
- **Q75**: third quantile (in kHz)
- **IQR**: interquantile range (in kHz)
- **skew**: skewness (see note in specprop description)
- **kurt**: kurtosis (see note in specprop description)
- **sp.ent**: spectral entropy
- **sfm**: spectral flatness
- **mode**: mode frequency
- **centroid**: frequency centroid (see specprop)
- **peakf**: peak frequency (frequency with highest energy)
- **meanfun**: average of fundamental frequency measured across acoustic signal
- **minfun**: minimum fundamental frequency measured across acoustic signal
- **maxfun**: maximum fundamental frequency measured across acoustic signal
- **meandom**: average of dominant frequency measured across acoustic signal
- **mindom**: minimum of dominant frequency measured across acoustic signal
- **maxdom**: maximum of dominant frequency measured across acoustic signal
- **dfrange**: range of dominant frequency measured across acoustic signal
- **modindx**: modulation index. Calculated as the accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range
- **label**: male or female