Gender Recognition using Voice

Vineet Ahirkar, Naveen Bansal
MS in CS, UMBC
1000 Hilltop Cir, Baltimore,
MD 21250

Abstract

Voice classification goes beyond just the frequency of the voice and thus requires additional feature detection and learning based on these features. The paper aims at learning different models to classify voice samples as male or female based on the gender of the person. We used nine different models to get the initial accuracies and then used some ensemble methods such as bagging and feature extraction techniques such as PCA to improve the accuracies.

Introduction

Human ear has an excellent mechanism of perceiving the voice. It distinguishes the voice based on factors such as the loudness, frequency, the pitch and the resonating frequency. In our paper, we will be distinguishing the human voice based on the genders - male or female. We will also be analyzing the be model that distinguishes the voices based on the training models.

The paper is organized as follows: Problem section describes the problem we are trying to solve, Proposed method section describes different approaches we took to solve the problem, Experiment section describes the overall result and accuracies we got from all the different approaches and overall conclusions are mentioned in conclusion section.

Problem

A human ear can distinguish between a male and female voice easily. If we want to teach a machine to do the same then what features of a voice would the machine require to classify ?

The voice of an adult male can range between 85 to 180 Hz and that of adult female ranges between 165 to 255 Hz. Thus we can see that although the frequency ranges differ quite a lot, there is a mid range where they seem to overlay. This is why differentiating voices only on the basis of frequency is inadequate.

Men and women have different vocal folds due to variance in the sizes of larynx. Breath can be transfered at varying pressures over the vocal folds thus having different results in different pitches in voice of men and women. Genetics also causes a huge variance in the voice.

Proposed method

Dataset

We used the dataset provided by kaggle to train our models. The dataset consists of 3168 voice samples.

Features

Features in the dataset were obtained from voice samples using warble R package which are as follows:

1. **Length of signal (duration)**
   Length of the signal measured in miliseconds.

2. **Mean frequency (meanfreq)**
   Mean normalized frequency of the spectrum of the audio signal, measured in kHz.

3. **Standard Deviation of Frequency (sd)**
   Standard deviation measures the amount of variation or dispersion of data values. A low standard deviation indicates that the values are more closer to the mean, whereas a high standard deviation indicates that the values are more spread out.

4. **Median Frequency (median)**
   Median frequency is the middle value of a dataset and is measured in kHz.

5. **First Quantile (Q25)**
   Quantiles are the points dividing range of probability distribution into contiguous intervals with equal probabilities. It is the data value when the standard distribution goes beyond the first threshold. It is measured in kHz.

6. **Third Quantile (Q75)**
   Similar to the first quantile, the third quantile is a data point when the standard deviation reaches the one third of the highest in the range. It is measured in kHz.

7. **Interquantile Range (IQR)**
   Interquantile Range is the difference between the first (lower) and the third (upper) quartile and is measured in KHz. It is a measure of statistical dispersion.

8. **Skewness (skew)**
   Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative, or even undefined. Skew can be thought to refer to the direction opposite to that where curve appears to be leaning.

9. **Kurtosis (kurt)**
   Kurtosis is a measure of the "tailedness" of the probability distribution of a real-valued random variable. Similar to skewness, kurtosis is a descriptor of the shape of a probability distribution and just as for skewness, there are different ways of quantifying it for a theoretical distribution and corresponding ways of estimating it from a sample from a population.

10. **Spectral Entropy (sp.ent)**
    In general, entropy is nothing more than the measure of amount of disorders in a system. Spectrum Entropy tells us how different the distribution of energy is in the system. High value of spectral entropy indicates the existence of a constant similarity of energy (small variations). Low value of spectral entropy indicates high variances and irregularity.

11. **Spectral Flatness (sfm)**
    Spectral Flatness is a measure used in digital signal processing to characterize an audio spectrum. It is also known as Wiener entropy. It is typically measured in decibels and provides a way to quantify how noise-like a sound is, as opposed to being tone-like.

12. **Mode frequency (mode)**
    Mode frequency is the one occurring most in the entire dataset.

13. **Frequency centroid (centroid)**
    The frequency centroid is a measure used in digital signal processing to characterize a spectrum. It indicates where the "center of mass" of the spectrum is. Perceptually, it has a robust connection with the impression of "brightness" of a sound.

14. **Peak frequency (peakf)**
    Peak frequency is the frequency with highest energy.

15. **Mean Frequency (meanfun)**
    Average of fundamental frequency measured across acoustic signal.

16. **Minimum frequency (minfun)**
    Minimum fundamental frequency measured across acoustic signal.

17. **Maximum frequency (maxfun)**
    Maximum fundamental frequency measured across acoustic signal.

18. **Average dominant frequency (meandom)**
    Average of dominant frequency measured across acoustic signal.

19. **Minimum dominant frequency (mindom)**
    Minimum of dominant frequency measured across acoustic signal.

20. **Maximum dominant frequency (maxdom)**
    Maximum of dominant frequency measured across acoustic signal.

21. **Range of dominant frequency (dfrange)**
    Range of dominant frequency measured across acoustic signal.

22. **Modulation index (modindx)**
    Calculated as the accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range. It describes by how much the modulated variable of the carrier signal varies around its unmodulated level.

Models

We propose to train the conventional models to learn the task of classifying male and female voice samples. The models we chose are as follows.

1. Decision tree
2. K Nearest Neighbors
3. Logistic Regression
4. Naive Bayes
5. Neural Network
6. Perceptron
7. Random Forest
8. SVM
9. Gradient Boosting

Techniques

We also propose some techniques to reduce variance by reducing the dimensionality using PCA algorithm and pruned decision tree. The detailed description of our approaches are as follows

- **Full Feature Set**
  We trained all the models mentioned above on the dataset containing all the 22 features.

- **PCA**
  We reduced the dimension of our data set to 5 dimensions and then trained all the models to observe the change in accuracy.

- **Handpicking Top features**
  We built a decision tree and then picked the top five features from the decision tree i.e the features which provides maximum information gain.

- **Voting**
  We conducted voting on all the models and used the mode of all results as our prediction.

Intuition

High accuracy in all the models indicated that the problem might be simple and we may not need all the attributes to classify voice samples. Reducing the dimensions will prevent the model from over fitting and might provide more accurate results.

Experiment Description

Below are the details of the models we used in the experiments.

Decision tree

The decision tree with the gini criterion is used in the experiments. There is no max_depth specified, and no kind of pruning is done. This rules out any under-fitting.

K Nearest Neighbor

The value of K is set to 10 ,i.e., the class label of 10 neighbors is checked to predict the class of the data point. All data points are given equal weights. The distance calculation is done by the default 'Minkowski' metric with p as 2, thus actually resulting in a standard Euclidean metric.

Logistic Regression

Logistic Regression with primal formulation is used having the L1 regularizer and alpha set to 0.0001 which ensures lower weights. 'liblinear' solver is used for the optimization with max iterations set to 100.

Naive Bayes

Gaussian Naive Bayes is used because it supports continuous valued features and models each as conforming to a Gaussian (normal) distribution.

Neural Network

A neural network with 2 hidden layers each containing 5 neurons. A 'relu' activation function is used with the learning rate initialized to 0.001 and then increasing by a step at each epoch.

Perceptron

A perceptron with L1 regularizer and alpha set to 0.0001 ensures that over-fitting is tackled.

Random Forest

Random Forest with the gini criterion is used. Bootstrapping is set to true which means that samples are drawn with replacement. Max of 10 estimators are used.

SVM

A non-linear Support Vector Machine with the Radial Basis Function kernel (RBF kernel) with a degree of 3. Penalty parameter C is set to default value 1.

Gradient Boosting

Gradient Boosting with the 'friedman_mse' criterion and the 'deviance' loss function is used. The number of estimators used are 100 which means that 100 different regression trees are created having a max depth of 3 to get an estimate of the class.

Most Informative Features

Information gain is used to decide which feature to split on at each step in building the tree. Information gain is calculated on the basis of inverse of the impurity at that particular feature split. We used a gini impurity which is a measured as the probability of a randomly chosen element from the set being incorrectly labeled if distribution of labels is random in the subset.

PCA

Principle Component Analysis (PCA) is a technique in which correlations are found in features and the top performing features are extracted by applying orthogonal transformation on linearly uncorrelated variables called 'Principal Components'. PCA actually selects a linear combination of features and uses them as the new features. These new features can then be used to train the models. PCA has proved to be effective in removing features having no correlation with the classes and thus increasing accuracy. Each Principle

Component has one dimension with mid-point having value 0. The sign of the variable tells indicates the direction of the variable in that PC on a single dimension vector. A positive sign indicates a direct relationship between the vector and that variable, whereas a negative sign indicates a inverse relationship between the vector and variable. The magnitude of the variable signifies strength of the relationship.

Voting

Voting is a ensemble learning technique in which multiple models are trained and then they vote for a particular class. A majority or a average vote is then taken to decide the final class. In our experiment, we used a majority voting technique on the 9 models discussed above, i.e., we considered the class with the maximum votes for the final class.
A single model has the tendency to have some presumptions and generalize the function being learnt. Such an error is called as bias. Voting reduces the error due to bias which may have led to over-fitting.

Experiment Observations

Full Feature Set

Neural network achieved best accuracy when trained on all the 22 features which is around 93%. Neural network has a flexible decision boundary and thus is able to classify the data more accurately than the other models. The perceptron classifier performs poorly compared to other models and achieves accuracy of around 81%. The perceptron algorithm may not be able to perform as the data might not be linearly separable. All the models gave more than 81% accurate results which indicates that the data might not be that hard to learn and there might be a subset of features which alone could predict the label of the data.

Most Informative Features

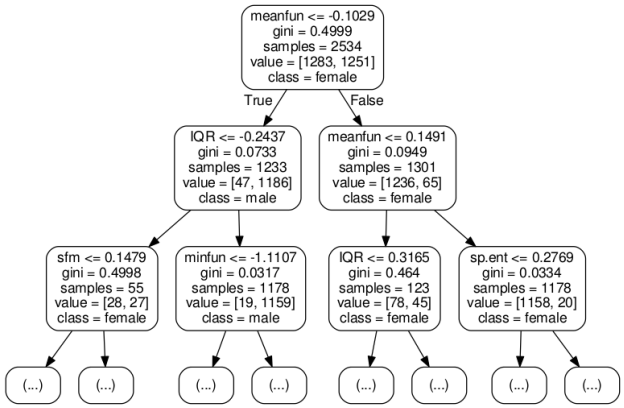

Figure 1: Decision Tree with depth 2

We picked the 5 top performing features from the decision tree based on high information gain. They are as follows:

- Mean Frequency
- Inter-Quantile Range
- Spectral Flatness
- Minimum Frequency
- Spectral Entropy

As per intuition, the accuracy of all the models got increased after reducing the dimensions. The accuracy of the perceptron model increased to 94% which might be because of the reason that data become linearly separable after the dimensionality reduction. Random forest got the highest accuracy when trained on the top five features and gives around 97% accuracy. Thus we were able to increase the overall accuracy by 3% using our proposed dimensionality reduction technique.

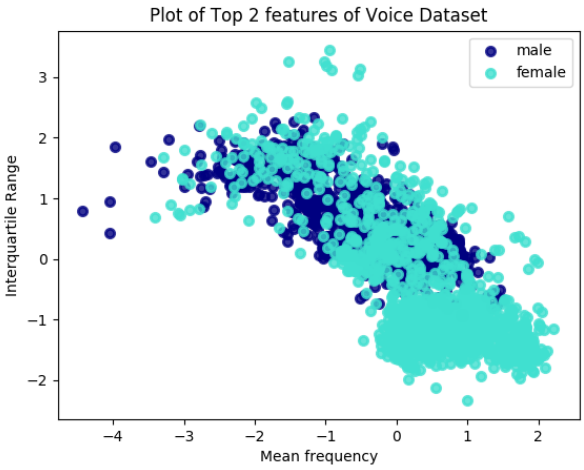

Figure 2: Plot of top 2 attributes from decision tree

PCA

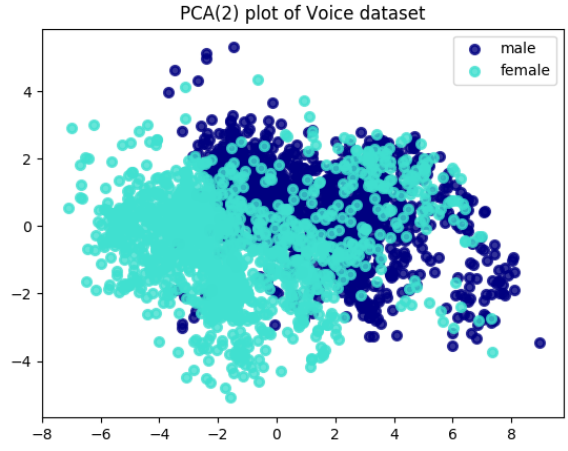

Figure 3: Plot of PCA with 2 components

Similar pattern were observed with the dimensionality reduction using PCA. The accuracy of all the models were

higher compared to the accuracies obtained without dimensionality reduction. Again, Random forest outperformed the other models and achieved an accuracy of around 97%.

Voting

We observed that the accuracy in each of the above approach got increased when we conducted voting on all the models and used mode of their predictions as the predicted label. A model might be biased towards a particular class and voting helps us overcome that problem.

Accuracy	**Feature Set**		
	Full(22)	**PCA(5)**	**Top 5**
Decision tree	0.8659	0.9479	0.9589
K nearest neighbour	0.9021	0.9416	0.9621
Logistic regression	0.8927	0.9684	0.9637
Naive bayes	0.8849	0.8690	0.9494
Neural network	0.9306	0.9558	0.9653
Perceptron	0.8186	0.9558	0.9447
Random forest	0.9006	0.9701	0.9716
SVM	0.9259	0.9685	0.9637
Gradient boosting	0.9148	0.9684	0.9669
Voting	0.9384	0.9826	0.9873

Table 1: Results

Conclusion

Nine different models were trained to classify a voice sample as male or female. Two different techniques were proposed to reduce the dimensionality of data and voting among all the models was proposed to counter variance. The experimentation on the dataset shows an increase in accuracy for all the proposed methods. Random forest performed best in both the approaches involving dimensionality reduction. Voting among all the models seems to always increase the accuracy which is expected due to reduction in variance.

References

1. Hassam Ullah Sheikh, Who is speaking ? Male or female ? Ph.D dissertation, dept. of engineering and sciences, University of Manchester.

2. Musaed Alhussein, Zulfiqar Ali, Muhammad Imran, and Wadood Abdul, Automatic Gender Detection Based on Characteristics of Vocal Folds for Mobile Healthcare System, Dept. of Computer Engineering, King Saud University.

3. Kory Becker, Gender Recognition by Voice and Speech Analysis, http://www.kaggle.com.

4. Phonetically balanced, US English single speaker databases designed for unit selection speech synthesis research, Language Technologies Institute, Carnegie Mellon University.