International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-0072

Voice based Gender Recognition

Remna R. Nair[1], Bhagya Vijayan[2]

[1]Master of Computer Applications, College of Engineering, Trivandrum, Kerala, India
[2]Asst.Professor, Master of Computer Applications, College of Engineering, Trivandrum, Kerala, India

---

Abstract - *An Is a human's voice "UNIQUE"? This is actually a good question with which most of us are really confused. Yes!! How much a fingerprint is unique that much a human's voice is also unique. Because of this uniqueness of the voice, the human voice can be used for many recognition processes. One of the most heard recognition processes is "gender". It is easier for an individual to recognize or identify a human gender by hearing the voice. So this paper is developed with a mindset to make the machine learn and identify the gender of the given voice(real-world input).*

Key Words: Decision Trees, Gradient Tree Boosting, Gender Recognition, Random forests, Support Vector Machine(SVM).

1. INTRODUCTION

One of the most common means of communication in the world is through voice. In the real world, it is possible for a person to verify the gender of a person through voice. Voice is filled with lots of linguistic features. These voice features are considered as the voice prints to recognize the gender of a speaker. The recorded voice is considered as the input to the system, which then the system process to get voice features. Examine the input and compare it with the trained model, carry out calculations based on the algorithm used and gives the latest matching output. Gender recognition can be used along with various other applications. Some are:

- For detecting feeling like male sad, female anger, etc.
- Differentiating audios and videos using tags.
- Spontaneous salutations.
- Helping personal assistants to answer questions with gender-specific results etc.

2. METHODOLOGY

The dataset used for detecting gender from the audio files is retrieved from VoxForge, which is a free speech corpus and acoustic model repository for open source speech engines. It is a large-scale collection of voices of both genders. From the collected audio files the powerful discriminating features are extracted with which a CSV file is created. With this CSV file various models are trained using Support Vector Machine, Decision Trees, Gradient Tree Boosting, Random forests, and accuracy is calculated. The goal is to compare outputs of various models and suggest the best model that can be used for gender recognition by voice in real-world inputs.

In short, the following are the steps followed -

- First, the voice(.wav files) needs to be converted to a form that the system can understand.
- Preprocessing needs to be done on the file to avoid external noises.
- After the noises have been removed the feature extraction process can be carried out. It is much necessary to find out the acoustic features that have a high discriminative power to classify the genders.
- Next step, is to train the machine with collected features from the dataset to make the machine capable to classify the genders of the voice. Here the machines are trained with four algorithms Support Vector Machine, Decision tree, Gradient Tree Boosting, and Random Forest.
- Calculate the accuracy of each algorithm with the dataset.
- After finding out the best algorithm over the dataset (based on the accuracy), a particular algorithm will be used to find out the gender of the real world input given while testing.

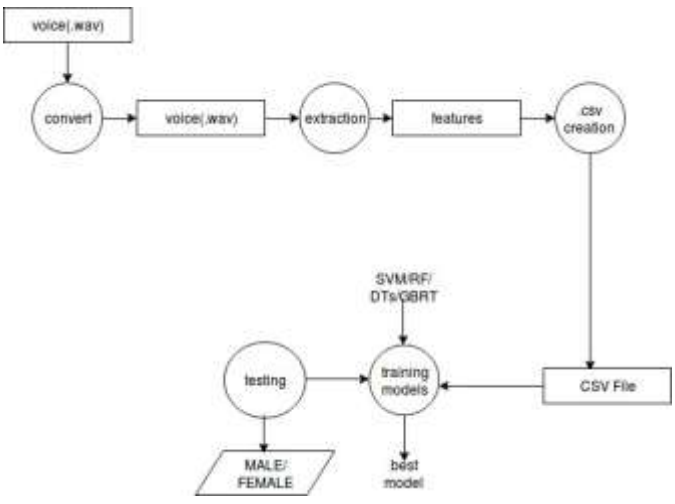

Fig -1: Overall process of the system

3. IMPLEMENTATION

3.1 Data Collection and Preprocessing

For Voice Based Gender Recognition system, the dataset collected consist of 62,440 audio samples compressed (.tgz) in sets of 10 files that can be automatically downloaded from the url http://www.repository.voxforge1.org/downloads/SpeechCorpus/Trunk/Audio/Main/16kHz_16bit. Next step is to uncompress the compressed files. After uncompressing the .wav files are obtained. These .wav files are not understood by the machines, so it needs to be converted into the machine-understandable format. First, the contents of every wave file are read, then the features are extracted and saved into a pandas data frame. In addition to the wave file, the README files(available with the downloaded dataset) are also parsed to extract metadata: gender, age range, and pronunciation. Notably, the python package Scipy wavefile is used to get the audio data, Scipy stats to extract the main features and Numpy and its fast fourier transform fft and fftfreq to deduce the .wav data to frequencies.

The data from the .wav files are recorded as amplitude in the time domain, but the potentially valuable features (with a higher discriminative power male/female) are frequencies. To convert the data to frequencies we use DFT (Discrete Fourier Transforms), particularly the FFT implementation (Fast Fourier Transform) of the algorithm. The Fourier transform takes a signal in the time domain (set of measurements over time) and turns it into a spectrum - a set of frequencies with corresponding (complex) values. The spectrum does not contain any information about time!

So, to find both the frequencies and the time at which they were recorded, the audio signal is splitted into small, overlapping slices, and the Fourier transform is applied to each (short time Fourier transform).np.fft.fft returns a complex spectrum. np.fft.fftfreq returns the sample frequencies. A sample folder consists of 10 audio recordings from one particular user. Every .wav file in the folder is processed in a way that the dominating frequencies in sliding windows of 200ms (1/5th of a second) are extracted. If a .wav file is 4 seconds long, a list containing 20 frequencies will be extracted. The list of frequencies is then filtered to contain only values in the human voice range (20Hz < frequency < 280Hz) In addition, values around 50Hz are most likely noise and also filtered (https://en.wikipedia.org/wiki/Mains_hum).

3.2 Feature Extraction

The most important feature extracted is the frequencies with which more powerful detection can be done. The rest of the feature depends on these frequencies. Here, the dominating frequencies are extracted in sliding windows of 200ms.After the dominating frequencies are extracted other properties such as skew, kurtosis, etc are calculated. With the following features a CSV file is created to train the models.

- Mean - Mean specifies the average of all the dominating frequencies.
- Median - Median is the middle value in the list of frequencies.
- Mode - Here, mode refers to the frequency that occurs most often.
- Standard Deviation - By calculating the standard deviation we could get how much the values in the dominating frequency list varies from the mean value of the list.
- Skewness - Skewness shows the asymmetry of the voice frequency spectrum around the sample means. If skewness is positive, the spectrum spreads out more to the right of the mean than to the left side, and vice versa.
- Kurtosis - Kurtosis shows how much outlier-prone a distribution is. The frequency spectrum will exhibit a normal shape if kurtosis value is 3.Otherwise if the value is less than 3, the frequency spectrum will have fewer items at the center and the tails than the normal curve but with more items in the shoulders. If the frequency spectrum is having more items near the center and at the tails, and few items in the shoulders compared to a normal distribution with the same mean and variance, then the kurtosis value is greater than 3.
- Low and High - Low points out the minimum frequency in the dominating frequency list and high shows the maximum frequency in the dominating frequency list.
- First Quartile(Q25) - Q25 is the median of the lower half of the dominating frequency list. It means that about 25% of the frequency values in the dominating frequency list lie below Q25 and 75% of dominating frequency values larger than Q25.
- Third Quartile(Q75) - Q75 is the median of the upper half of the dominating frequency list. It suggests that about 75% of the frequency values in the dominating frequency list lie below Q75 and only 25% of dominating frequency values lie above Q75.
- Interquartile Range(IQR) - IQR means the frequency ranging between Q25 and Q75 and the value is the difference between Q75 and Q25 of an audio.

3.3 Modeling Using Different Machine Learning Algorithms

The different ML algorithms used are:

Support Vector Machine(SVM) - SVM is a discriminative classifier that classifies the dataset into two parts of a hyperplane based on the extracted features. When a new data

occurs, SVM compares with the existing groups, and with which greater similarity is found it will be binded with that particular group.

Decision Trees (DTs) - DTs non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the gender by learning simple decision rules inferred from the data features.

Gradient Tree Boosting or Gradient Boosted RegressionTrees(GBRT) - Gradient Boosting is a generalization of boosting to arbitrary differentiable loss functions.

Random Forest - Random Forest is a supervised classification algorithm. As of the name suggests, this algorithm creates the forest with a number of trees. In Random Forest classifier as the number of the tree increases, the forest gives the high accuracy results.

After modeling with these datasets the best algorithm for voice-based gender recognition is identified. With the best model, the real world inputs gender can be identified.

3.4 Testing The Models

For testing the models an interface needs to be created for recording human voice. After the human voice has been recorded the noises are cleared out. Then this recorded input is given to the best model(model with high accuracy compared to others) and the gender is detected.

4. RESULT

The result is the calculated accuracies from all models. The accuracies with the different models are compared and the best model is the one which has the highest accuracy with the data. The aim of using multiple models is to identify the best gender recognition model, which can be helpful in developing future applications. The table below shows the obtained accuracies of different models.

Table -1: Accuracy Table

Algorithm Used	*Training*	*Testing*
Decision Tree	1.000	0.869
Gradient Boosting	0.958	0.937
Random Forests	0.987	0.893
Support Vector Machine	0.940	0.905

The result points out that the algorithm with the highest accuracy is Random Forest. So the best algorithm for Voice Based Gender Classification is Random Forest.

Assessment of the variable importance - Methods that use ensembles of decision trees (Tree, Forest, and Gradient Boosting) can compute the relative importance of each attribute. These importance values can be used to inform a feature selection process.

The feature importance depends on how well a feature discriminates between the classes. The std, dev seems to be in the top 3 most important features in the three tree-based models.

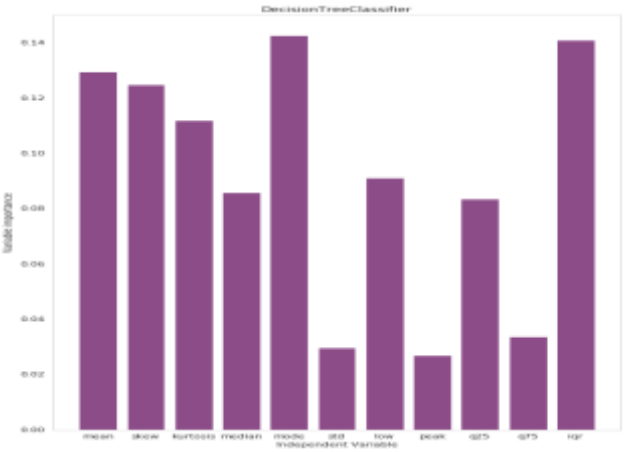

Chart -1: Decision tree classifier

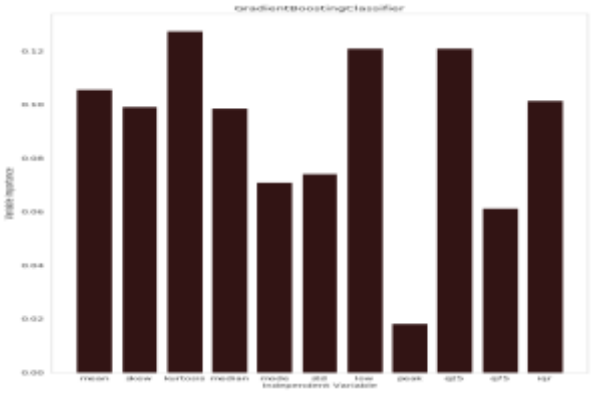

Chart -2: Gradient Boosting Classifier

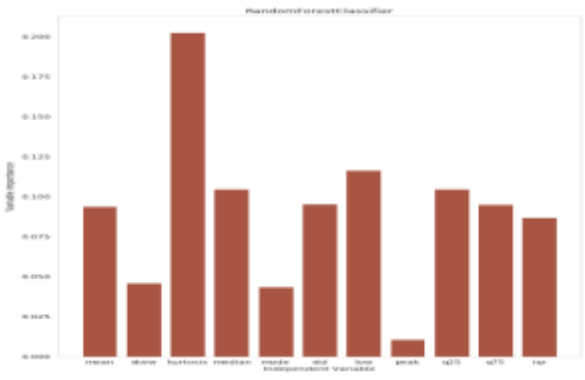

Chart -3: Random Forest Classifier

5. CONCLUSIONS

The aim of this paper is to build a gender recognition system based on voice. As stated above std, dev, kurtosis, and skewness seem to have the most discriminant power. Some unexpected behavior is in the peaks at very low frequencies (< 50 Hz) that can be viewed in mode and median. This can be due to the presence of noise in the audio recordings. Even if it was filtered, not all sources of noise were accounted for and further analysis in this way would be interesting but falls beyond the scope of this project.

Further exploration of potentially interesting features could also be pictured out. Other features that were not used here but could also have good differentiating power are: spectral flatness spectral entropy, fundamental frequency

At the same time, the set of selected features seem to be enough for the tested models to achieve good accuracy. The Random Forest gave the best accuracy followed by the GradientBoosting, SVM and Decision Tree.

As for the applicability of the tested models, the analysis of this dataset was supervised and one might not always have access to the labels of the different categories. But they could be of great utility in a variety of situations, like discerning music/speech or a correctly functioning machine/instrument from a faulty one.

REFERENCES

[1] H. Harb and L. Chen, "Voice-based gender identification in multimedia applications," Journal of Intelligent Information Systems, vol. 24, no. 2-3, pp. 179–198, 2005. [Online]. Available: http://dx.doi.org/10.1007/s10844-005-0322-8.

[2] Kunyu Chen, Gender Identification By Voice, Stanford University kunyu@stanford.edu.

[3] T. Vogt and E. André, "Improving automatic emotion recognition from speech via gender differentiation," in Proc. Language Resources and Evaluation Conference (LREC 2006), Genoa. Citeseer, 2006.