Voice Based Gender Detection

Anchita Goel
Indraprastha Institute of
Information Technology
Delhi
2012019

Sarthak Ahuja
Indraprastha Institute of
Information Technology
Delhi
2012088

I. Problem Statement

Detect gender of the speaker based on various features like MFCC, pitch, short-time energy, energy entropy, zero-crossing rate and spectral centroid.

II. Work Done

A. Past Work

Gender detection has been done using MFCC, pitch and other general features. Here we incorporate other energy based features like short-time energy, energy entropy and others like zero-crossing rate and spectral centroid.

B. Dataset Description

Source: Azreda (http://www.azreda.org/audio.html). Our training dataset has 1140 seconds of female voice and 1866 seconds of male voice. Testing dataset consists of 1284 seconds of female recording and 1790 seconds of male recording. The sampling rate of the dataset is 8000 Hz. The classifier can be tested on any live recording or any pre-recorded voice file stored in .wav format. Response is given by

$$y_i = \begin{cases} 1 & \text{if female} \\ 0 & \text{if male} \end{cases}$$

C. Data Preprocessing

Median filtering is applied to the entire dataset. Median filters can do an excellent job of rejecting certain types of noise, in particular, impulse noise in which some individual frames have extreme values. After median filtering, the entire dataset is split into samples each of duration two seconds.

D. Features extracted

We have extracted the following features from the dataset:

1) Mel Frequency Cepstral Coefficient (MFCC): MFCC is based on the human peripheral auditor system. The human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency t measured in Hz, a subjective pitch is measured on a scale called the Mel Scale .The mel frequency scale is a linear frequency spacing below 1000 Hz and logarithmic spacing above 1kHz.

Fig. 1. Steps involved in obtaining MFCC

2) Energy Entropy: While energy and entropy are common features used for speech detection, they both have limitations in noisy environments. With the energy E(i) and entropy H(i) of each frame i computed in parallel, both parameters are adjusted by shifting their respective baselines. This is achieved by subtracting the average amount of the first 10 frames accordingly. We denote the average energy and entropy of the first 10 frames as C_E and C_H, respectively. The energy-entropy is formally calculated as:

$$EE(i) = \sqrt{1 + |(E(i) - C_E)(H(i) - C_H)|}$$

Males have low and distributed energy-entropy while females have high and stays for short period of time.

3) Short-time Energy: This measurement can distinguish between voiced and unvoiced speech segments, since unvoiced speech has significantly smaller short- time energy.

$E_m = \sum_{n=-\infty}^{+\infty} (x[n]w[m-n])^2$
where w[m] is the window function used to extract a frame from the speech waveform. Males have low Short-time energy as compared to females.

4) Zero-Crossing Rate: The zero-crossing rate is the rate of sign-changes along a signal, i.e., the rate at which the signal changes from positive to negative or back.

$$Z = \frac{1}{T-1} \sum_{t=1}^{T-1} I(s_t s_{t-1} < 0)$$

where s is a signal of length T and the indicator function I(A) is 1 if its argument A is true and 0 otherwise. Females have higher zero-crossing rate than males.

5) Spectral Centroid: The spectral centroid is a measure used in digital signal processing to characterise 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.

$$\frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)}$$

E. Pattern Recognition via SVMs

The main idea of Support Vector Machines is to define a boundary between two classes by maximal separation of the closest observations. SVMs are powerful algorithm on binary classification tasks.

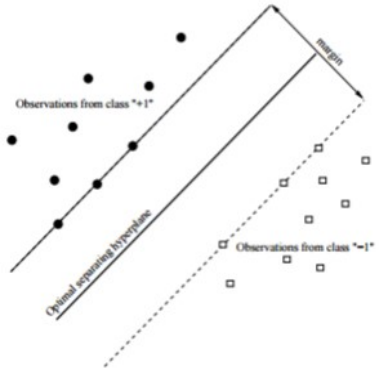

Fig. 2. Goal is to maximise the margin

Consider a dataset D = $\{(x_i, y_i), x_i \in R^p, y_i \in \{-1, 1\}\}$ that is linearly separable. The margin is defined as the shortest perpendicular distance between the hyperplane and the observations, referring to the width of the blank region separating two data clouds. The goal for SVMs is to maximize the margin. Any hyperplane can be written as

$$w.x - b = 0$$

where w is a coefficient vector and b is a constant.When the data is linearly separable, data can be completely separated by two hyperplanes such that no points fall in between the hyperplanes. Such hyperplanes are defined as

$$w.x - b = -1$$

and

$$w.x - b = 1$$

The region between the two hyperplanes is called margin. The goal is to maximise this margin which is equivalent to minimising $|w|^2$. The data can be classified as

$$class(x_i) = \begin{cases} 1 & \text{w.x}_i - b > 0 \\ \text{-1} & \text{w.x}_i - b \leq 0 \end{cases}$$

When the data is not linearly separable, kernel functions play an important role by linearizing the data. Data can be transformed by a kernel function K(x, y) into the inner product space in which it is feasible to separate them linearly. Common kernel functions are the radial basis function kernel (RBF kernel):

$$\text{K(x,y)} = \exp \frac{|x-y|^2}{2\sigma^2}$$

and the linear kernel:

$$\text{K(x,y)} = \text{x}^T y + c$$

and the polynomial degree kernel:

$$\text{K(x,y)} = (\alpha x^T y + c)^d$$

F. Result and Discussion

We used the SVM classifier using various combinations of kernel functions and features for classification. The polynomial degree kernel is used with default degree=3. Fetaures extracted from the data have been divided into two groups:
Group 1 of features is described as the features extracted by performing MFCC and
Group 2 of features include the Energy Entropy, Short-time energy, Zero-Crossing Rate and Spectral centroid. The accuracy of classification are reported as follows:

Kernel Function	Accuracy
Linear	64.4%
Polynomial	45.9%
RBF	58.2%

TABLE I. ACCURACY FOR DIFFERENT KERNEL FUNCTIONS WHEN GROUP 1 FEATURES ARE USED.

1) Using only Group 1 features: Table 1 shows the accuracy for different kernel functions when Group 1 features are used. Fig 3 is the ROC plot obtained when the above stated kernel functions and Group 1 of features are used:

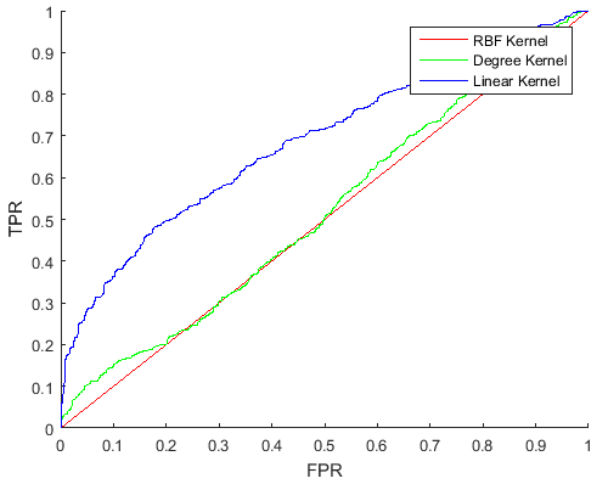

Fig. 3. ROC curve when using Group 1 features for different kernel functions

Kernel Function	Accuracy
Linear	41.7%
Polynomial	58.2%
RBF	41.7%

TABLE II. ACCURACY FOR DIFFERENT KERNEL FUNCTIONS WHEN GROUP 2 FEATURES ARE USED.

2) Using only Group 2 features: Table II shows the accuracy for different kernel functions when Group 2 features are used. Fig 4 is the ROC plot obtained when the above stated kernel functions and Group 2 of features are used:

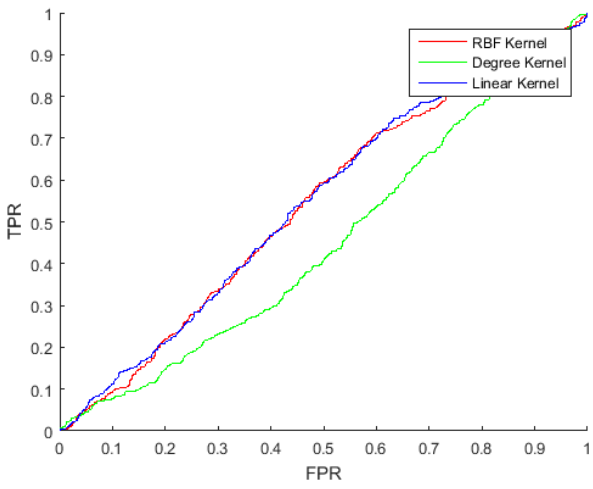

Fig. 4. ROC curve when using Group 2 features for different kernel functions

Kernel Function	Accuracy
Linear	71.6%
Polynomial	56.9%
RBF	58.2%

TABLE III. ACCURACY FOR DIFFERENT KERNEL FUNCTIONS WHEN GROUP 1 AND GROUP 2 FEATURES ARE USED.

3) Using both Group 1 and Group 2 features: Table III shows the accuracy for different kernel functions when both Group 1 and Group 2 features are used. Fig 5 is the ROC plot obtained when the above stated kernel functions and both Group 1 and Group 2 of features are used:

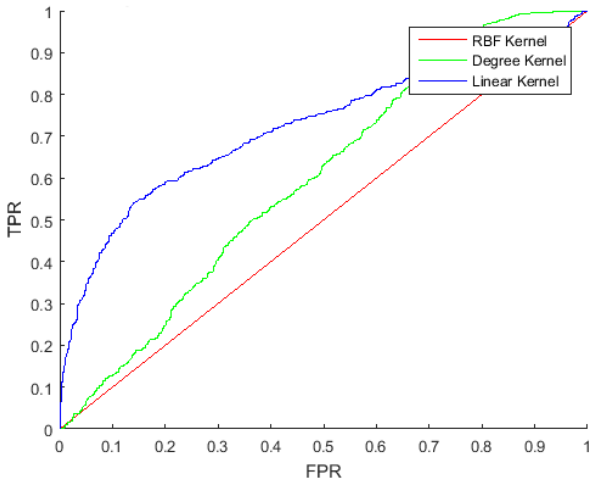

Fig. 5. ROC curve when using Group 1 and Group 2 features for different kernel functions

Our classifier can also be used for classifying a live recording of voice: The green color indicates that the voice recognised is of a male and the red color indicates that the voice recognised if of a female.

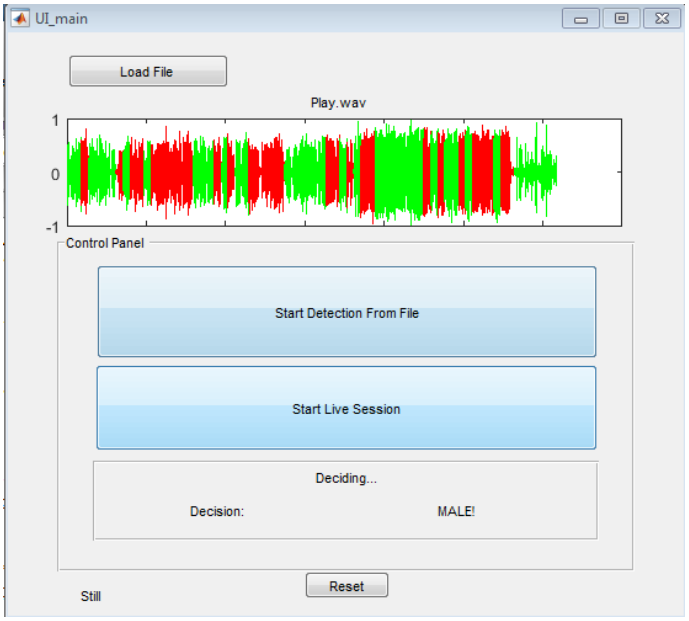

Fig. 6. A capture of the classifier recognising gender from a file uploaaded in .wav format in Matlab

III. CONCLUSION

We conclude that the SVM classifier using a linear kernel function gives the best accuracy of all the classifiers. Group 1 of features just comprises of MFCC components and Group 2 consists of energy-based as well as some advanced features. Using both the groups gives the most accurate classifier than using each group individually. Our code and dataset can be found by clicking here.

REFERENCES

[1] Ma,Z. and Fokou,E. *Speaker Gender Recognition via MFCCs and SVMs.*

[2] Chen,S.-H. and Luo,Y.-R. *Speaker Verification Using MFCC and Support Vector Machine.*