International Journal of Research in Advent Technology, Vol.7, No.4S, April 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

GENDER CLASSIFICATION BASED ON SPEECH RECOGNITION USING SVM

B Rajasekhar
Asst.Professor
Anurag Group of Institutions

Abstract: The term gender identification deals with finding out the gender of a person from his or her voice. Gender identification has been implemented in several Automatic Speaker Recognition (ASR) systems and has proved to be of great significance. The use of gender identification in today's technology makes it easier for user authentication and identification in high security systems. In this paper, we have discussed about the gender identification process for speech signals using three different features namely Pitch using autocorrelation, Signal energy and Mel Frequency Cepstral Coefficients (MFCCs). A linear Support Vector Machine (SVM) classifier was used for classification of features extracted from the speech signal using signal processing methods. Two sets of experiments were performed - in the first experiment, one speech file was tested against one training file as a one-on-one experiment. In the second experiment, one speech file was tested against three training files. The average accuracy of the second experiment was slightly higher than the first experiment. Performance evaluation results are encouraging. The approach can be used in wide range of applications.
Keywords: gender identification, MFCC, pitch, fundamental frequency, autocorrelation, signal energy, SVM classifier.

1. INTRODUCTION:

Gender identification through a speakers voice deals with finding out whether the speech is spoken by a male or a female. There are several advantages to automatic detection or prediction of the speakers' gender [1]. In the context of Automatic Speaker Recognition (ASR), gender identification and detection is very crucial since gender-dependent systems prove to be more accurate than gender-independent systems. The common applications of these gender-dependent systems are speaker recognition and identification, multimedia annotation, and speaker indexing, annotation in multimedia, speaker recognition [2]. Speaker diarization and speech synthesis are greatly enhanced with the application of gender identification. Other common use of gender-dependent systems include gender-wise sorting of telephone calls for surveys related to gender-sensitization, and detecting telephone call speeches from unsatisfied or angry callers [3], [4]. This paper discusses an approach for gender identification through speech with the help of three different features and a classifier, namely, Pitch using Autocorrelation, Signal Energy and Mel Frequency Cepstral Coefficients (MFCC), and Support Vector Machine (SVM) classifier. Comparison of classification accuracy results was done for each feature. Among the three features, MFCC was found to exhibit the highest accuracy automatically operated fire and to control the spread of fire.

In this paper, sample speeches were modeled as an autoregressive (AR) process and represented in the state-space domain by the Spectral Subtraction. The original speech signal and the reconstructed speech signal obtained from the output of the filter were compared. The idea of this comparison is to pursue an output speech signal, which is similar to the original one. It was concluded that Spectral Subtraction is a good constructing method for speech.

2. LITERATURE SURVEY:

A supervised dictionary learning (SDL) approach based on the Hilbert-Schmidt independence criterion (HSIC) has been proposed that learns the dictionary and the corresponding sparse coefficients in a space where the dependency between the data and the corresponding labels is maximized [1]

Mel Frequency Ceptral Coefficient is a very common and efficient technique for signal processing. This paper presents a new purpose of working with MFCC by using it for Hand gesture recognition. The objective of using MFCC for hand gesture recognition is to explore the utility of the MFCC for image processing. Till now it has been used in speech recognition, for speaker identification. The present system is based on converting the hand gesture into one dimensional (1-D) signal and then extracting first 13 MFCCs from the converted 1-D signal. Classification is performed by using Support Vector Machine [2]

Emotion recognition helps to recognize the internal expressions of the individuals from the speech database. In this paper, Dynamic time warping (DTW) technique is utilized to recognize speaker independent Emotion recognition based on 39 MFCC features. A large audio of around 960 samples of isolated words of five different emotions are collected and recorded at 20 to 300 KHz sampling frequency. Training and test templates are generated using 39 MFCC features. In the proposed work, we have extracted the MFCC coefficients from the speech database and DTW is used to store a

prototypical version of each word in the vocabulary and compute incoming emotion with each word [3]

3. PROPOSED METHOD:

A. Features

The features MFCC, Pitch using Autocorrelation, and Signal energy are discussed in details as follows: 1) MFCC: MFCC is a representation of the real cepstral of a windowed short-time signal derived from the Fast Fourier Transform (FFT) of that signal. It represents the short time spectral features of a signal. MFCC is a cepstral method which converts speech into parameters according to the Mel Scale. Usually, about 20 coefficients are used in Automatic Speech Recognition (ASR), but 10-12 coefficients are sufficient enough for coding speech. The steps involved in MFCC are shown in Fig. 2. In MFCC, the speech signal is divided into several random samples of time, usually 20 msec or 30 msec. The samples may or may not be overlapped, although overlapping is more commonly used since it smoothened the transmission from one frame to the next. A Hamming window is then used to remove discontinuities and smooth out the edges of the samples. A Hamming window with length n is computed as

$$W(n) = 0.54 - 0.46cos(\frac{2\pi n}{N-1}), 0 \leq n \leq N-1 \quad (1)$$

where n is the current sample, N is the total number of samples, and W (n) is the coefficient. After framing and windowing FFT is calculated so as to extract features frame wise. This is followed by computation of magnitude spectrum and triangular filterbank which results in filterbank energies. Discrete Cosine Transformation (DCT) from the filterbank energies is then calculated. The coefficients ranged according to significance. The 0th coefficient is often disregarded. Only the first 13 coefficients were taken for this study since they are the most relevant and can likely give the characteristics of the speech signals.

2) Signal Energy: Signal Energy of a continuous-time signal x(t) is computed as

$$E_x = \int_{-\infty}^{\infty} |x(t)|^2 dt \quad (2)$$

3) Pitch using Autocorrelation: The correlation between two waveforms is a measure of their similarities. At different time intervals, the waveforms are compared to find their sameness at each interval. Hence the autocorrelation function is a correlation of a waveform with itself. The autocorrelation function is calculated by

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

TABLE I
PITCH FREQUENCY VALUES OF 5 FEMALE AND 5 MALE SPEAKERS FOR THE SPEECH, "SHE HAD YOUR DARK SUIT IN GREASY WASH WATER ALL YEAR": (A) SPEAKER NUMBER, (B) PITCH FREQUENCY VALUES FOR A CORRESPONDING SPEAKER.

(a) Speaker	*(b) Pitch Frequency (%)*
S1 (F)	193.33
S2 (F)	183.69
S3 (F)	188.35
S4 (F)	189.84
S5 (F)	207.18
S11 (M)	138.67
S12 (M)	120.41
S13 (M)	117.23
S14 (M)	137.16
S15 (M)	141.43

B. Classifiers

The classification for speech signals was accomplished using Support Vector Machines (SVM). The objective in SVM is to adjudicate the decision boundary or the hyperplane in a multidimensional space that separates different class labels based on statistical learning theory. Using this hyperplane, the SVM executes a binary classification, which are decisions having a true or false value [4]. The separation of two classes using SVM is shown in Fig. 3. The data points which are closest to the hyperplane or decision surface are called the support vectors. These support vectors are the most difficult to classify as compared to other data points, as they lie so close to the hyperplane. The line perpendicular to the hyperplane is called Margin. Margin is also the distance between the support vectors of Class-1 and
Class-2
The separating function in SVM can be connected with the support vectors, which is expressed in the form of linearly combined kernels.

TABLE II
CLASSIFICATION ACCURACY OF THE FEATURES FOR TWO EXPERIMENTS: (A) FEATURES EXTRACTED, (B) TRAINING FILES, (C) TESTING FILES, (D) ACCURACY (%) OF EXP. 1, (USING 1 TRAINING FILE AND 1 TESTING FILE), (E) ACCURACY (%) OF EXP. 2, (USING 2 TRAINING FILES AND 1 TESTING FILE).

(a) Feature	*(b) Training*	*(c) Testing*	*(d) Exp.1 Accuracy(%)*	*(e) Exp.2 Accuracy(%)*
Pitch	Female	Female	56.52	50.00
	Female	Male	26.09	33.33
	Male	Male	57.14	57.14
	Male	Female	33.33	27.70
Energy	Female	Female	50.87	55.81
	Female	Male	29.63	29.62
	Male	Male	55.17	54.55
	Male	Female	30.13	33.33
MFCC	Female	Female	53.84	69.23
	Female	Male	23.08	30.76
	Male	Male	61.53	61.54
	Male	Female	30.76	30.76

4. RESULTS

Two experiments were performed. In the first experiment i.e., Exp.1, for both male and female gender, one speech file has been taken as training file and tested against another speech file as a one-on-one testing. Whereas in the Exp.2, three speech files were trained and tested against one speaker file, and checked for classification accuracy. In each experiment, the training file and testing file are different and non-repetitive. A total of 6 utterances each from 10 male and 10 female speakers were used for experimentation. The contours for fundamental frequency of a male (speaker 11) and a female speaker (speaker 1) are shown in FIG. It can clearly be seen from the figures that, the F0 contour of male speaker is lower than that of a female speaker. The Pitch frequency in both male and female speakers lies within the desired range, which are 100 Hz to 180 Hz for male speakers and 165 Hz to 300 Hz for female speakers. The Pitch frequencies for 5 male speakers and 5 female speakers

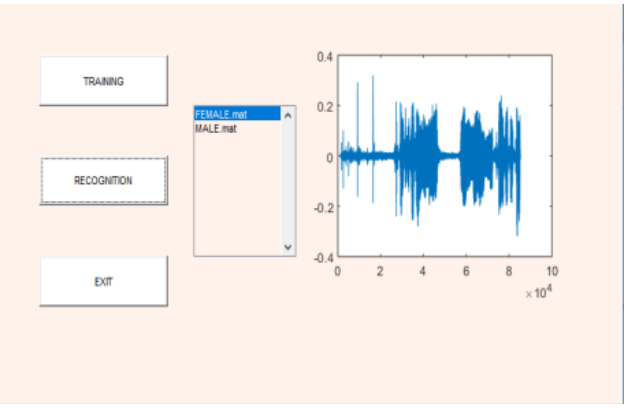

Figure 1: Graphical User Interface

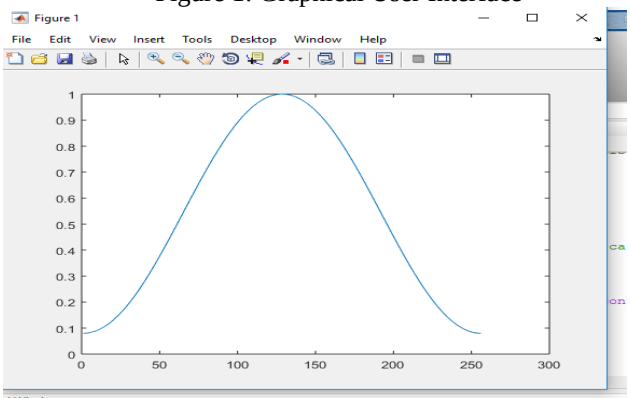

Figure 2: Hamming window output of the per processing block.

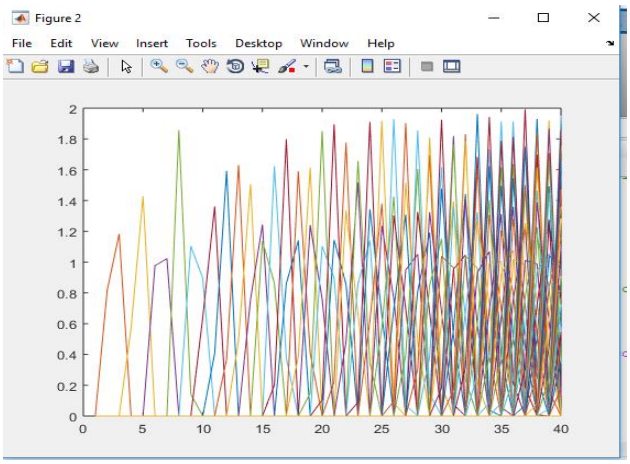

Figure 3**:** MFCC (Mel Frequency Cepstral Coefficient) feature extraction of the data signal

5. CONCLUSION:

This paper discussed about the existing features and classifiers used in the identification of gender. We have used TIMIT database for all the training and testing files. A total of 50 speech files were used which consisted of 6 utterances from 25 male speakers and 25 female speakers. Two experiments were performed, namely Exp.1 and Exp.2. The former takes one training file and test it against one testing file, whereas the latter tests takes three training files and test it against one testing file. It was found that MFCC gave the best overall classification accuracy for both Exp. 1 and Exp. 2, when compared to Signal energy and Pitch. The accuracy of the latter two is below 60% in both Exp. 1 and Exp. 2. The overall best accuracy achieved was 69.23% by MFCC, whereas the other two features were below 60% in both the experiments. It is seen that Pitch frequency accuracy was lower due to noise in between the signals rather than at the start and end of the signal since pre-processing was done to remove the noise at the beginning and end of the signals. This can be further processed using better algorithms to detect voiced and unvoiced region. The performance of signal energy was also satisfactory. It was seen that the Signal energy of female speech is higher than that of a male speech, which was likely due to the higher frequency in a female speech as compared to a male speech. More features and data can be used and experimented to get better accuracy. This paper can further be extended by adding the gender features in emotion detection. Even under normal situations, the emotions of a male and female usually differs to a certain level. Therefore, gender identification will assist in finding the difference in male and female emotions depending on different situations. One other way it can be further improved by using and combining more expert systems in order to get better accuracy rate. For the future extension of this paper, classifiers like Hidden Markov Model and Gaussian Mixture Model along with SVM classifier can be used to improve the classification accuracy.

REFERENCES

[1] Mehrdad J. Gangeh, AliGhodsi,Mohamed S. Kamel,"Multiview Supervised Dictionary Learning in Speech Emotion Recognition," IEEE Transaction on audio, speech, and language processing.

[2] Shikha Gupta1, Jafreezal Jaafar2, Wan Fatimah wan Ahmad3 and Arpit Bansal4 J. Clerk Maxwell, "Feature extraction using mfcc" Signal & Image Processing : An International Journal (SIPIJ) Vol.4, No.4, August 2013

[3] N.Murali Krishna1,P.V. Lakshmi2,Y. Srinivas3 J.Sirisha Devi4," Emotion Recognition using Dynamic Time Warping Technique for Isolated Words," IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 1, September 2011

[4] Aastha Joshi," Speech Emotion Recognition Using Combined Features of HMM & SVM Algorithm International Journal of Advanced Research in Computer Science and Software Engineering Research Paper Volume 3, Issue 8, August 2013.

[5] Eslam Mansour mohammed1, Mohammed Sharaf Sayed2," LPC and MFCC Performance Evaluation with Artificial Neural Network for Spoken Language Identification ," International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 6, No. 3, June,2013