International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 3, Issue 1 (August 2013) PP: 26-30

Speaker Recognition Using Vocal Tract Features

Prasanth P. S.
Sree Chitra Thirunal College of Engineering, Pappanamcode, Thiruvanthapuram

Abstract: *Speaker recognition refers to the process of automatically determining or verifying the identity of a person based on his/her voice. A method of robust feature estimation of vocal tract features from the speech utterances leads to speaker classification. This paper deals with estimation of pitch and Mel Frequency Cepstral Coefficient (MFCC) and how those features can be used for speaker recognition. The Support Vector Machine (SVM) is well suited for binary speaker classication*

I. Introduction

The objective of automatic speaker recognition is to recognize a person from a spoken utterance. A speaker recognition system can be operated in either identification mode or verification mode. In speaker identification, the goal is to identify the speaker of an utterance from a given population, where as speaker verification involves validating the identity claim of a person. Speaker recognition systems can be classified into text-dependent systems and text-independent systems. Text-dependent systems require the recitation of a predetermined text, whereas text-independent systems accept speech utterances of unrestricted text. This letter deals with text-independent speaker classification.

A speech signal can be decomposed into two parts: the source part [1] and the system part. The system part consists of the smooth envelope of the power spectrum and is represented in the form of cepstrum coefficients, which can be computed by using either the linear prediction analysis or the mel filter-bank analysis. Most of the automatic speaker recognition systems reported in the literature utilize the system information in the form of cepstral coefficients. These systems perform reasonably well. The source information has been rarely used in the past for speaker recognition systems. The source contains information about pitch and voicing. This information is very important for humans to identify a person from his/her voice. A few studies have been reported where pitch information is used as a feature for speaker recognition. However results are not very encouraging. The main reason for this is that pitch estimation is not very reliable, and this affects the performance of the speaker recognition system.

Pitch detection [3] relies on the periodic qualities of the sound waveform, therefore any attempt to determine pitch is only valid on voiced segments of an utterance. The beauty of human speech lies in the complexity of the different sounds that can be produced by a few tubes and muscles, which also makes it non stationary over periods of 10 ms.. This intricacy, however, makes speech processing a challenging task. One defining characteristic of speech is its pitch, or fundamental frequency. Pitch detectors are used in vocoders, speaker identification and verification systems and also as aids to the handicapped. Because of its importance many solutions to estimate pitch has been proposed.

The speaker-specific vocal tract information is mainly represented by spectral features like Mel-Frequency Cepstral Coefficients (MFCCs). MFCC's are based on the known variation of the human ear's critical bandwidths with frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. Speaker recognition has a different objective, i.e., differentiating one speaker from the others. However, as a matter of fact, most existing speaker recognition systems use vocal tract features like MFCC [2] . This indicates that the MFCC features do contain important speaker-specific information, in addition to the intended phonetic information. Ideally, if a large amount of phonetically balanced speech data is available for speaker modeling, the phonetic variability tends to be smoothed out so that speaker-specific aspects can be captured.

Support Vector Machines (SVM's) are new learning method used for binary speaker classification. The set of vectors are optimally separated by a hyper plane if it is separated without error and the distance between the closest vector to the hyper plane is maximal. Support Vector Machine generates a hyper plane exactly in the middle of feature vectors for binary speaker classification.

II. Vocal source and vocal tract features

The output speech signal is convolution of the vocal tract features with vocal source features or excitation signal. If we can extract those features, we can recognize the speaker.

2.1 Vocal source features: pitch estimation

One defining characteristic of speech is its pitch, but it is often times difficult to obtain this value precisely because speech segments often do not contain the pitch or fundamental component F0, but only its harmonics. F0 is the rate of vibration of vocal cords located in larynx. Pitch period is the frequency interval between harmonics present in the signal and is estimated using FFT method is shown in the Figure 1.

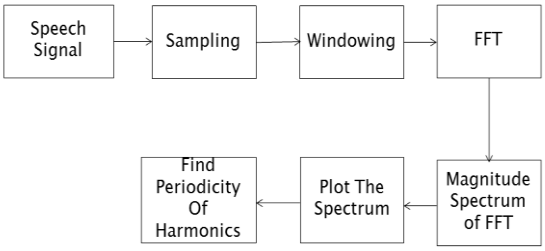

Figure 1: Block diagram of pitch estimation using FFT method

2.2 Vocal tract features: Mel Frequency Cepstral Coefficients(MFCC)

Human ear does not follow a linear frequency spacing.. Mel Frequency Cepstral Coefficients (MFCC) are based on the known variation of the human ear's critical bandwidths with frequency, and use filters spaced linearly at low frequencies and logarithmically at high to capture the phonetically important characteristics of speech. The mel-frequency scale, is linear at frequency spacing below 1000 Hz and logarithmic above 1000 Hz. The speech input is typically recorded at a sampling rate above 10000 Hz. This sampling frequency was chosen to minimize the effects of aliasing in the analog-to-digital conversion. These sampled signals can capture all frequencies up to 5 kHz, which cover most energy of sounds that are generated by humans. The main purpose of the MFCC processor is to mimic the behavior of the human ears. In addition, rather than the speech waveforms themselves, MFCC's are shown to be less susceptible to mentioned variations. Block diagram of Mel Frequency Cepstral Coefficient generation is shown in Figure 2.

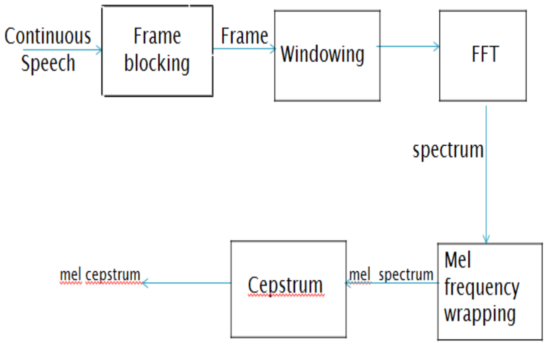

Figure 2: Block diagram of Mel Frequency Cepstral coefficient generation

In this paper, we use the standard procedures of extracting MFCC [2] on a short-time frame basis as described as follows:

1) Short-time Fourier transform is applied every 10 ms with a 30-ms Hamming window.
2) The magnitude spectrum is warped with a set of nonlinearly spaced triangular filters that are centered on equally spaced frequencies in Mel-scale is shown in Figure 3.
3) The log-energy of each filter output is computed.
4) Discrete cosine transform (DCT) [2] is applied to the filter bank output to produce the cepstral coefficients.

The 20 Mel Cepstral Coefficients are generated for each frame.

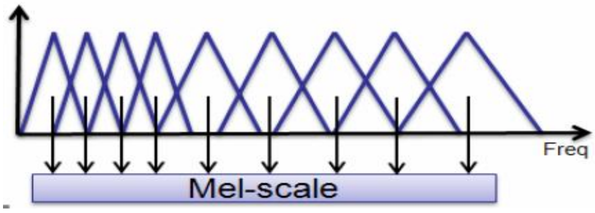

Figure 3: The triangular filter bank based on Mel frequency scale

III. Support Vector Machine (SVM)

Support Vector Machine has been used as a classifier in speaker recognition.. Currently, SVM is widely used in object detection & recognition, content-based image retrieval, text recognition, biometrics, speech recognition also used for regression. SVM's introduce the notion of a “kernel induced feature space” which casts the data into a higher dimensional space where the data is separable. We are given *l* training examples $\{x_i; y_i\}$ i = 1; _ _ _ ; *l*, where each example has d inputs ($x_i \epsilon R^d$), and a class label with one of two values ($y_i \epsilon \{-1,1\}$) . Now, all hyper planes in R^d are parameterized by a vector (w), and a constant (b), expressed in the equation, g(x) is a linear function is shown in the Figure 4.

$$g(x)=W^T X + b \qquad (1)$$

If the support vectors are not linearly separable more complex functions are used to describe the margin or a method of introducing an additional cost function associated with misclassification is appropriate.

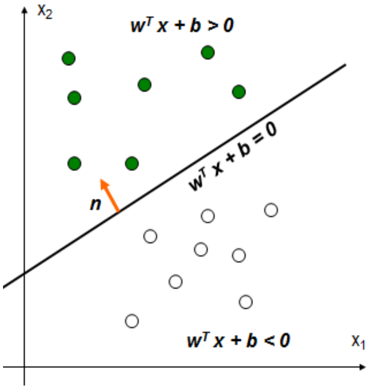

Figure 4: Generation of hyper plane in the feature space

The Support Vector Machine (SVM) which works as large margin speaker classier is well suited for binary speaker classification. The linear discriminant SVM classifier provides maximum margin, compared to other SVMs. Margin is defined as the width that the boundary could be increased by before hitting a data point is shown in Figure 5.

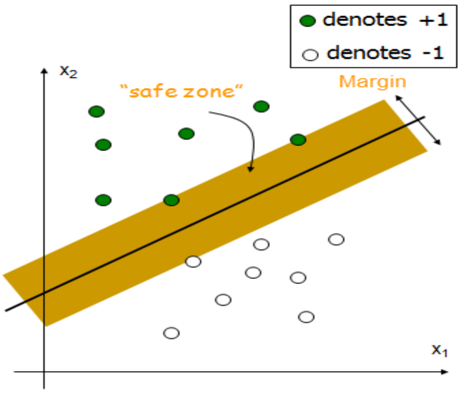

Figure 5: Large margin linear classifier

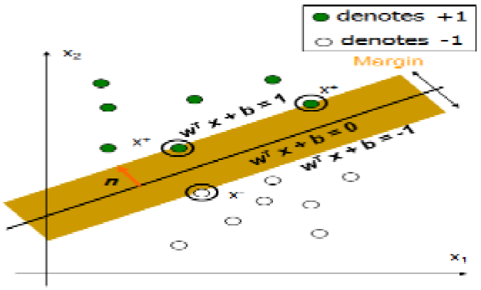

Figure 6: Maximum margin linear classifier

Given set of data points $\{(X_i, y_i)\}$, i=1,2,…….n, where

For y_i=+1 $\qquad W^T X_i + b > 0 \qquad (2)$

For y_i=-1 $\qquad W^T X_i + b < 0 \qquad (3)$

With a scale transformation on both w and b ,the above is equivalent to

For y_i=+1 $\qquad W^T X_i + b >= 1 \qquad (4)$

For y_i=-1 $\qquad W^T X_i + b <= -1 \qquad (5)$

We know that

$$W^T x^+ + b = 1 \qquad (6)$$

$$W^T x^- + b = -1 \qquad (7)$$

Where x^+ , x^- are the positive and negative support vectors on the both sides of the hyper plane

The margin width is $(x^+ - x^-).n$ (8)

Where n is the normal vector such that $n = \frac{w}{||w||}$

$$(x^+ - x^-).\frac{w}{||w||} = \frac{2}{||w||} \qquad (9)$$

So we have to maximize the margin width, maximize the $\frac{2}{||w||}$ is shown in Figure 6.

IV. Experimental results and Discussion

4.1 Pitch estimation

A sample utterance of a person who has to be classified is recorded which is given as input. Speech signal is sampled at 8 kHz. Speech signal is non-stationary, so the speech signal is windowed with 10 ms Box – car window and FFT is calculated. From the short time spectrum find the periodicity of harmonics, which define the pitch period F0.Pitch period is the non cepstral coefficient which can be used for speaker recognition. Pitch estimation by FFT method in MAT LAB is shown in Figure 7 .Pitch estimation can also find out by the auto correlation method or by Average Magnitude Difference Function (AMDF) [3].In auto correlation method compute autocorrelation of the speech signal. Normalize the autocorrelation function by divides the value at n=0.Divides the higher portion of the autocorrelation function into N equal parts. Find the maximum value of normalized autocorrelation for each of the N divisions. These N maximum autocorrelation (MACV) [4] correspond to N MACV features which can be utilized for speaker recognition.

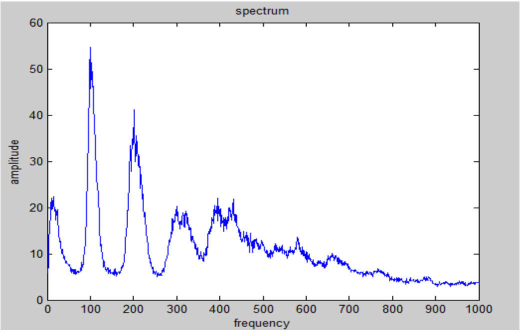

Figure 7: Pitch estimation by FFT method

4.2 MFCC coefficient generation

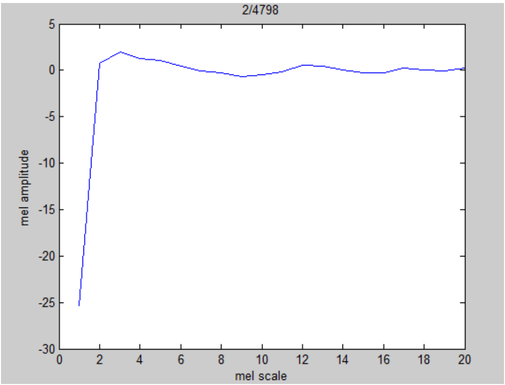

Figure 8: MFCC coefficient generations

MFCC coefficient generation [1] in MAT LAB is shown in Figure 8. 20 Mel cepstral coefficients are generated for each speech segment.

4.3 Support Vector Machine classification

The Support Vector Machine is used as the classifier for speaker classification. For the binary speaker classification takes the four sample utterance of each speaker. For the speaker classification MFCC coefficients are given as the input to the SVM classifier. The support vector machine trains the MFCC coefficients of the utterance of the speaker. . Support Vector Machine (SVM) generates a hyper plane which divides the MFCC coefficients of the each speaker in the middle. Thus the training phase is completed. During the testing phase an utterance is spoken by the speaker. The support vector machine calculates the MFCC coefficients of the each spoken utterance and which determine the spoken utterance is falls on which side of the hyper plane .Thus the speaker is verified. In this work speech signal is mixed with white Gaussian noise signal. The SVM trains the MFCC coefficients of the noisy speech signal. The SVM classify the MFCC coefficients by the information from SVM classifier structure and then find the recognition accuracy of the SVM classifier. The maximum accuracy of the SVM classifier when MFCC coeficients are given to the input of the SVM classifier is 0.9442 and the minimum accuracy is 0.4701.When plot the SNR verses accuracy in the MAT LAB the accuracy of the classification increas when SNR increases. This result is shown by SNR verses accuracy plot in the Figure 9.

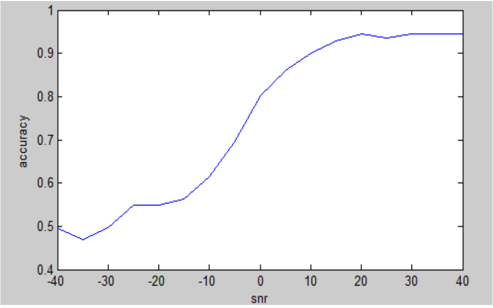

Figure 9: SNR verses accuracy of the speaker classification by SVM

V. Conclusion

This paper mainly deals with how the vocal source features and vocal tract features can be estimated. Pitch is a vocal source feature which is estimated by FFT method. MFCC is a vocal tract feature which is also estimated. The pitch and MFCC can be together used for speaker recognition. The support vector machine is used as the classifier for the speaker recognition. SVM generates the hyper plane for the speaker classification. SVM classifier trains the MFCC of the spoken utterances and determines the utterances of the speaker falls on which side of the hyper plane. Thus the speaker is verified. Here speech signal is mixed with noise signal and then find out the accuracy of the classification when the MFCC coefficients are input by plotting the SNR verses accuracy in the MAT LAB

References

[1] Combining Evidence From Residual Phase and MFCC Features for Speaker Recognition, K. Sri Rama Murty and B. Yegnanarayana, *Senior Member,* IEEE, *IEEE SIGNAL PROCESSING LETTERS, VOL. 13, NO. 1, JANUARY 2006.*

[2**]** **"**Robust Speaker Recognition Using Denoised Vocal Source and Vocal Tract Features" , C. Ching, Senior Member, IEEE, Nengheng Zheng, Member, IEEE, and Tan Lee, Member, IEEE. *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 19, NO. 1, JANUARY 2011*

[3] Pitch Estimation using Autocorrelation Method and AMDF-International Conference on Advances in Computing and Management – 2012, Savitha S Upadhya1, Nilashree Wankhede.

[4] Use of voicing and pitch information for speaker recognition Brett R. Wilder moth and Kuldip k Paliwal School of Microelectronic Engineering, Griffth University,Brisbane,Australia.

[5] Discrete –time speech signal processing priniples and practice-Thomas F.Quatieri