I J C T A, 10(9), 2017, pp. 25-31 ISSN: 0974-5572

Analysis of Feature Extraction Behavior for Speaker Identification and Verification in Adverse Acoustic Condition

Shrikant Upadhyay*, Sudhir Kumar Sharma*, Sandip Vijay*, Aditi Upadhyay* and Pawan Kumar*

ABSTRACT

The behavior of speech signal is almost different and varies person to person in this universe and it completely depends upon the existing condition. The real time application voice recogniton faces a serious major problem when it trained and tested using any devices. Feature extraction technique is one the useful method which is basically used to figure out the behavior and characteristics of different voices which are very essential to deploy according the existing condition. The person staying at remote location using their voice as password or authentication verification (key) is possible only when it is properly identified and verified efficiently with the existing condition in term of acoustic condition or changes in external surrounding. So, behavior analysis of feature extraction is very essential to identify and verify the original user from destination to sources or vice-versa is crucial. This paper tries to analyze and identify the suitable feature extraction behavior that would suitable and helpful to judge the authenticate person of different group.

Keywords: Different feature extraction technique, Speaker Identification Speaker Verification, Acoustic condition, Behavior prediction.

1. INTRODUCTION

Language is the engine of civilization, and speech is its most powerful and natural form. Textual language has become extremely important in modern life, but speech has dimensions of richness that text cannot approximate [1].

The increase in application demand has resulted in increased need in speech recognition research area over the past decades, a large variety of speech processing method have propose and speech recognition has been in main centre for the whole world today.

2. SPEAKER IDENTIFICATION AND VERIFICATION

Speaker recognition is the general term used to include different ways of discriminating people based on their voices. The main categories are: speaker identification system and

2.1. Speaker Identification

The objective of speaker identification is to classify an unlabeled utterance belonging to one of the N reference speakers [2]. It can be closed set identification or open set identification.

* Cambridge Institute of Technology, Department of Electronics & Communication Engineering, Tatisilwai, Ranchi 835103, Jharkhand, India.

* Jaipur National University, Department of Electronics & Communication Engineering, Jagatpura, Jaipur 302017, Rajasthan, India.

* ICFAI University, Deparment of Electronics & Communication Engineering, Dehradun, Uttarakhand, India, *Emails: shri.kant.yay@gmail.com, sudhir.732000@gmail.com, vijaysandip@gmail.com, sweetcaditi@gmail.com, pawan_aloysius1@yahoo.com*

2.2. Speaker Verification

The objective of speaker verification is to accept or reject the identity claim of speaker [3]. If the match between test and reference is above threshold level, the claim is accepted.

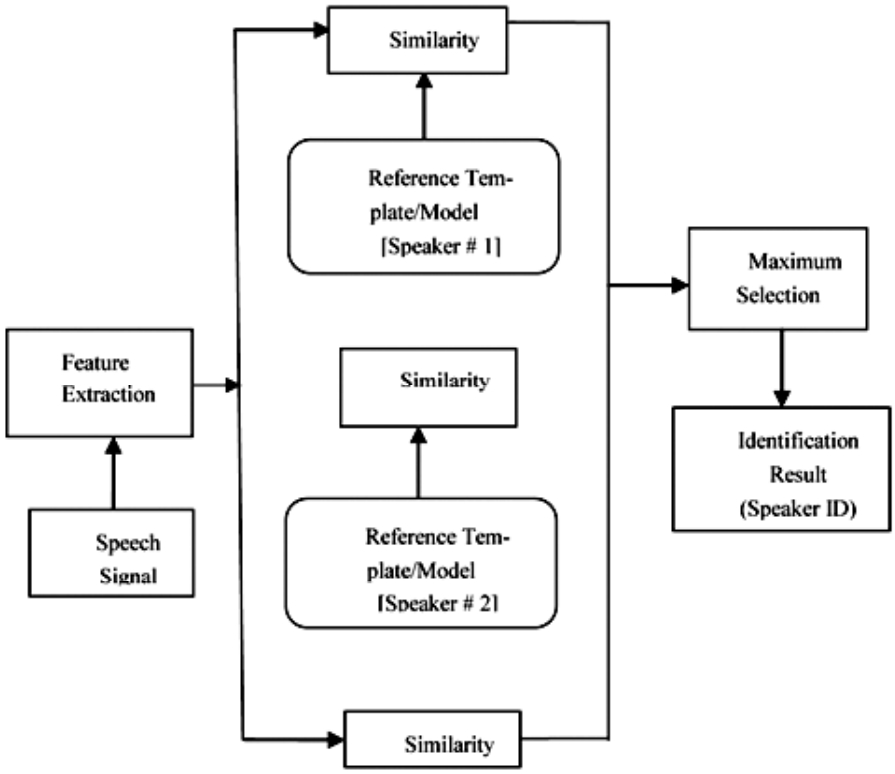

Figure 1: Structure of speaker identification system

3. FEATURE EXTRACTION TECHNIQUE

Feature extraction involved in signal modeling that performs temporal and spectral analysis. The need of feature extraction arises because the raw speech signal contains information to convey message to the observer or receiver and has a high dimensionality. Feature extraction algorithm derives a characteristics feature vector with lower physical or spatial properties.

3.1. Mel-frequency cepstrum co-efficient (MFCC)

MFCC technique is basically used to generate the fingerprints of the audio files.

Let us consider each frame consist of 'N' samples and let its adjacent frames be separated by 'M' samples where M is less than N. Hamming window is used in which each frame is multiplied. Mathematically, Hamming window equation is given by:

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \tag{1}$$

Now, Fourier Transform (FT) is used to convert the signal from time domain to its frequency domain. Mathematically, it is given by:

$$X_k = \sum_{i=0}^{N-1} x_i \frac{2\pi ik}{e^{N-1}} \tag{2}$$

$$M = 2595\log_{10}\left(1+\frac{f}{700}\right) \tag{3}$$

In the next step log Mel scale spectrum is converted to time domain using Discrete Cosine Transform (DCT). Mathematically, DCT is defined as follow:

$$X_k = \alpha \sum_{i=0}^{N-1} x_i \left(2i + 1/2N\right) \tag{4}$$

The result of the conversion is known as MFCC and the set of co-efficient is called acoustic vectors.

3.2. Linear Predictive Coding Analysis (LPC)

It is a frame based analysis of the speech signal performed to provide observation vectors [4]. The relation between speech sample S (n) and excitation X(n) for auto regressive model (system assume all pole mode) is explained mathematically as:

$$S(n) = \sum_{k=1}^{p} a_k s(n-k) + G.X(n) \tag{5}$$

The system function is defined as:

$$H(z) = \frac{S(Z)}{X(Z)} \tag{6}$$

A linear predictor of order 'p' with prediction co-efficient (α_k) is defined as a system whose output is defined as:

$$\hat{s}(n) = \sum_{k=1}^{p} \alpha_k S(n-k) \tag{7}$$

The system function is p^{th} order polynomial and it follows:

$$P(z) = \alpha_k \; z^{-k} \tag{8}$$

The prediction error e(n) is defined as:

$$e(n) = s(n) - \hat{s}(n)$$

$$= s(n) - \sum_{k=1}^{p} \alpha_k S(n-k) \tag{9}$$

The transfer function of prediction error sequence is:

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \tag{10}$$

Now, by comparing equation (5) and (10), if $\alpha_k = \alpha_k$ then A(z) will be inverse filter for the system H(z) of equation (6).

$$H(z) = G/A(z) \tag{11}$$

The purpose is to find out set of predictor coefficients that will minimize the mean squared error over a short segment of speech waveform. So, short-time average prediction error is defined as [5].

$$E(n) = \sum_m \left(e_n(m)\right)^2$$

$$= \sum_m \left\{ s_n(m) - \sum_{k-1}^{p} \alpha_k s_n(m-k) \right\} \tag{12}$$

where, $s_n(m)$ is segment of speech in surrounding of n samples i.e. = $s_n(m) = s\,(n + m)$

Now, the value of α_k minimize E_n are obtained by taking $\partial E_n / \partial a_i = 0$ & $i = 0, 1, 2....p$ thus getting the equation:

$$\sum_m s_n(m-i) s_n(m) = \sum_{k-1}^{p} \alpha_k \sum s_n(m-i) s_n(m-k) \tag{13}$$

If $\phi_n(i,k) = \sum_m s_n(m-i)\, s_n(m-k)$ (14)

Thus, equation (13) rewritten as:

$$\sum_{k=1}^{p} \alpha_k \varphi_k(i,k) = \varphi_k(i,0), \text{ for } i = 1, 2, 3 \ldots p \quad (15)$$

The three ways available to solve above equation i.e. autocorrelation method, lattice method and covariance method. In speech recognition the autocorrelation is widely used because of its computational efficiency and inherent stability [5].

Speech segment is windowed in autocorrelation method as discuss below:

$$S_n = S(m+n) + w(m) \text{ for } 0 \le m \le N-1 \quad (16)$$

Where, $w(m)$ is finite window length

Then, we have

$$\varphi_n(i,\ k) = \sum_{m=0}^{N+p-1} s_n(m-i)\, s_n(m-k) \text{ for } 1 \le i \le p,\ 0 \le k \le p \quad (17)$$

$$\varphi_n(i,k) = R_n(i-k) \quad (18)$$

Where, $R_n(k) = \sum_{m=0}^{N-1-k} s_n(m)\, s_n(m+k)$

$R_k(k)$ is autocorrelation function then equation (15) is simplified as [6]

$$\sum_{k-1}^{p} \alpha_k\ R_n(|i-k|) = R_i(i) \text{ for } 1 \le i \le p \quad (19)$$

Thus using Durbin's recursive procedure the resulting equation is solved as:

$$E^{(i)} = \left(1 - k_i^2\right) E^{(i-1)} \quad (20)$$

Then from equation (19) to (22) are solved recursively for i = 1, 2 ...p and this give final equation as:

α_j = LPC coefficient = $\alpha_j^{(p)}$

k_i = PACOR coefficients

A very essential LPC parameter set which is derived directly from LPC coefficients is LPC cepestral coefficients C_m. The recursion used for this discussed as [7]:

$$C_0 = \ln G \quad (21)$$

$$C_m = \alpha_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) C_k a_{m-k} \text{ for } 1 \le m \le p \quad (22)$$

$$C_m = \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) C_k a_{m-k} \text{ for } m > p \quad (23)$$

3.3. Linear Prediction Cepstral Coefficients (LPCC)

The basic parameters for estimating a speech signal, LPCC play a very dominant role. This method is that where one speech sample at the current time can be predicated as a linear combination of past speech sequence or sample. LPCC algorithm in term of block diagram is shown in figure (2) below:

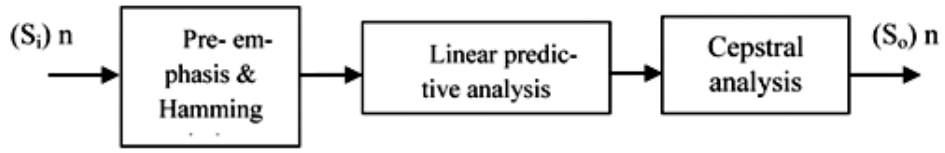

Figure 2: Steps involved in LPCC algorithm processing.

A digital all-pole filter is used to model the vocal tract and has a transfer function represented in z-domain as:

$$V(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k \, z^k} \tag{24}$$

where, $V(z)$ is the vocal tract transfer function, G is the gain of the filter, is the set of auto regression coefficients known as linear prediction coefficients (LPC) and p is the order of all-pole filter. One of the efficient method for estimating the LPC coefficients and filter gain is autocorrelation [7]. The inverse FFT transform of the logarithm of the speech magnitude spectrum and it is defined as:

$$\hat{s}[n] = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \ln\left[s(w) \right] e^{jwn} \, dw \tag{25}$$

4. RELATED WORK

This paper completely based on analyzing the best possible solution in signal processing domain by detecting and identifying the behavior of above three feature extractions while training and testing the real time voice sample and judges the efficient one which suites to be the best and suitable for real time applications. Acoustic condition is one the major issues that must be taken into consideration.

Since the random behaviour of speech signal is difficult to judge and come to conclusion. So, it is very important to extract the feature of signal and analyze the exact behaviour of feature extraction for speaker identification and its verification considering the three important feature extraction methods i.e. MFCC, LPC and LPCCC. Above figure 3 shows the random behaviour of signal for 5 KHz signal for 5 msec.

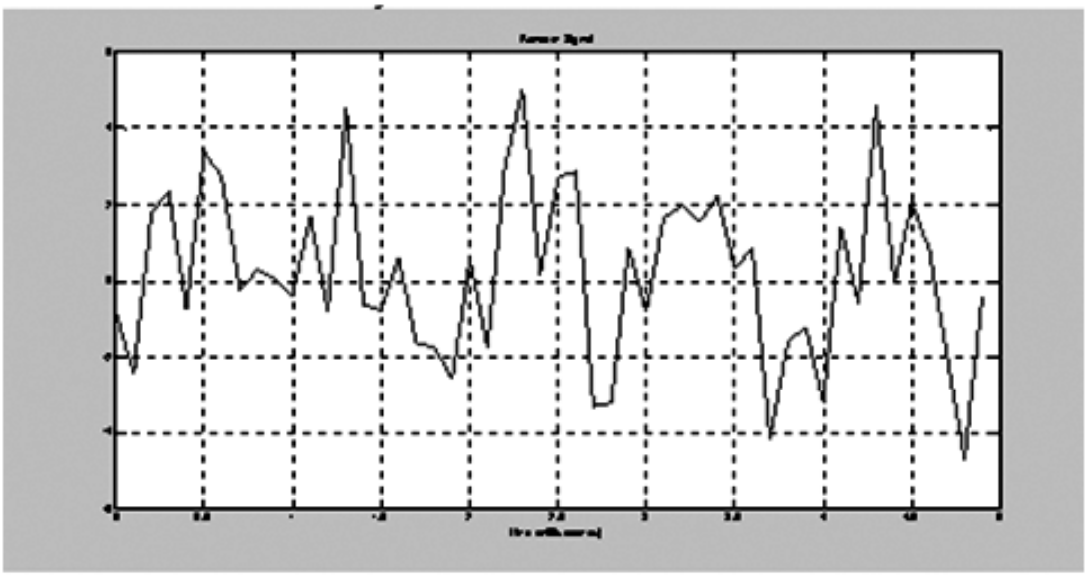

Figure 3: Behaviour of random signal for 5 KHz for time 5 millisecond.

5. PARAMETERS AND ITS CONSIDERED VALUES

Table 1
Parameters considered for analysis

Feature Extraction	*Frequency*	*Group*	*Condition*
MFCC	5KHz	Child	Acoustic (Noisy)
LPC	10KHz	Adult	Acoustic (Crowd)
LPCC	15KHz	Old	Acoustic (Environmental Effect)

Table 2
Variations in extraction techniques.

Feature Extraction	*Variations in %*	*Group*	*Condition*
MFCC	19.66	Child	Acoustic (Noisy)
LPC	14.23	Adult	Acoustic (Crowd)
LPCC	18.51	Old	Acoustic (Environmental Effect)

6. ANALYSIS RESULT

The analysis result concluded from above data for the variations in sample at different frequencies and the behavior of feature extraction can be identified clearly.

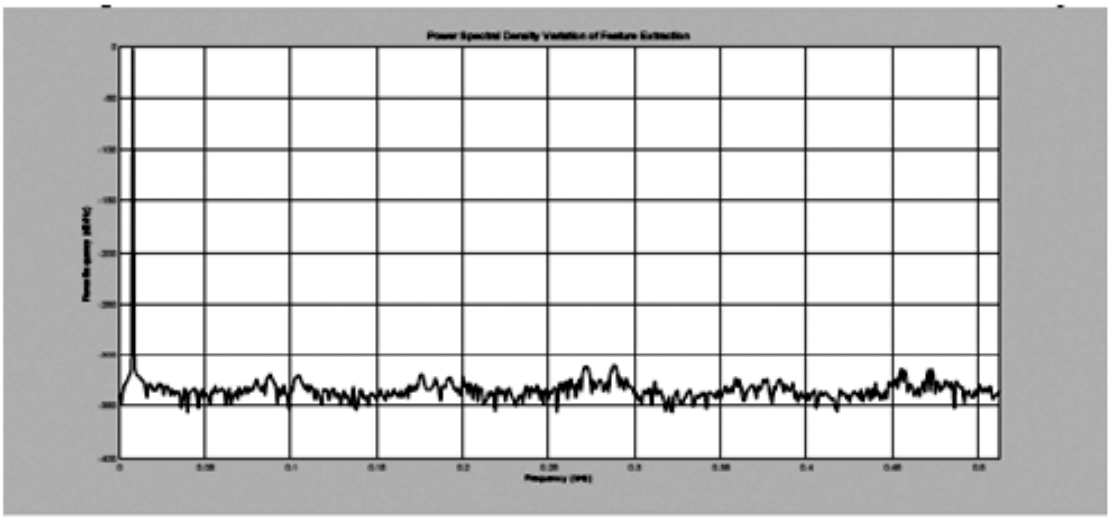

Figure 4: Power spectral analysis of extraction technique for different group.

7. CONCLUSION

The analysis result for the above mention feature technique shows that the MFCC shows the greater variation for child group having a smaller frequency where as LPC and LPCC proves to give smaller deviation. So, we can go for LPCC for smaller frequency deviation. Next, the PSD (power spectrum density) for the considered feature extraction in which LPCC proves to be efficient and shows higher value (dB/Hz) compared to MFCC and LPC. So, this may help to identify the exact feature extraction technique for different samples. This gives proper idea of behavior feature extraction technique for different speech signal. This result is useful for acoustic condition also but the more challenging would be the nature of the condition or noise.

REFERENCES

[1] M. Sigmund, “Speaker recognition identifying people by their voices,” Habilitation thesis, Brno University of Technology, Institute of radio electronics (2000).

[2] H. Gish and M. Schmidt, "Text-dependent speaker identification," IEEE Signal Processing Magazine, pp. 18-32, October (1994).

[3] S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Trans. Acoust., Speech, Signal Processing, vol. 29, pp. 254-272, April (1984).

[4] Abdelnaiem, "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, pp. 55-66, June (2013).

[5] L. R. Rabiner and R.W. Schafer, Digital processing of speech signals. Englewood cliffs, New Jersey: Prentice-Hall (1978).

[6] L. R. Rabiner and B. H. Juang, Fundamental of speech recognition, Englewood cliffs, New Jersey: Prentice-Hall (1978).

[7] Han Y, Wang G and Y Yang, "Speech emotion recognition based on mfcc", Journal of Chong Qing. University of Posts and Telecommunication (Natural Science Ediion) vol.69, pp. 34-39 (2008).