AN INTRODUCTION TO THE FEATURES EXTRACTED FROM THE AUDIO SIGNAL

Navid Samimi Behbahan *

Shabnam Azari*

Hedayat Bahadori*

Abstract: *Voice recognition is the holder of two recognition types. Speech recognition and speaker identification. By analyzing the features of a sound wave can be estimated that this attribute means for extremity speech speaker identification and verification to provide a bioassay performed. In contrast, speech recognition systems have been trying to understand the concept of the sound wave. Most existing research on speech recognition technology developed for the speaker-independent systems which can convert all speakers have spoken. In this paper, how to extract the audio and video features and two popular methods in this field (LPC and MFCC) will be investigated.*

***Keywords**: speech recognition, feature extraction, LPCC, MFCC.*

*Sama Technical and Vocational Training College, Islamic Azad University, Omidiyeh Branch, Omidiyeh, Iran

1- INTRODUCTION

In order to decrease volume and time of assessment, useful information of utterance signal being useful at separating various categories of utterance is extracted and useless redundancies are omitted. Kinds of methods of feature extraction have been created which contain of two total sets of coefficients based on current prediction and coefficients based on filter bank. First category of features using more for identifying teller's identity, models polarity of utterance signal and second category using to identify utterance and noisy models are more rebuts against noise. In this chapter, several methods of performing the duty describe.

2 Steps of extracting feature:

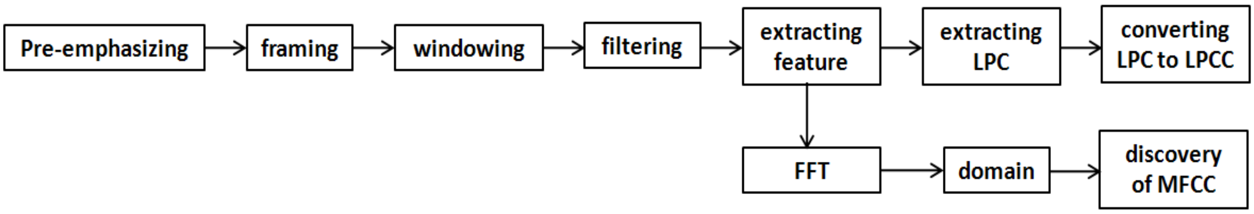

Figure 1: feature recognition

2-1- Pre-emphasizing

In order to omit the effect of sudden changes of continuous time signal, the signal should be passed from a first-order filter called pre emphasizing filter. α is the coefficient of pre emphasizing and the sum of it is selected 0.9 ≤ α≤ 1. Output of the filter i.e. $S_1(n)$ can be easily computed following equation:

$$s_1(n) = s(n) - \alpha\, s(n-1) \qquad (1)$$

2-2- Framing

In order signal is stationary and characteristics of talk signal is almost constant, utterance signal is divided into frames 20 to 30 ms and features are extracted from each frame. Usually, frames are selected how correlate with each other. The scale of correlation is usually selected between 10 and 15 ms. [1]

2-3 windowing

In this step, each frame is separately multiplied in a window to reduce the effect of un-continuity of signal at beginning and end of each frame.

If the window is showed with W(n), the applying of the window will be as following:

$$\overline{X_k} = X_k(n)\, W(n) \qquad 0 \le n \le N-1 \qquad (2)$$

Which N is the number of samples in a frame and k is the number of frame.

Hamming and Hanning windows are those windows that usually use in this application. The mathematical relationship of them is as following:

$$\overline{X_k} = X_k(n)\,W(n) \qquad 0 \leq n \ \leq N-1 \tag{3}$$

$$w(n) = \frac{1}{2}[1 - \cos(\frac{2n\pi}{N-1})] \qquad 0 \leq n \ \leq N-1 \tag{4}$$

$$w(n) = 0.54 - 0.46\cos\left(\frac{2n\pi}{N-1}\right) \qquad 0 \leq n \ \ N-1 \tag{5}$$

2-4- Methods of feature extraction

This operation can be performed in two realms of time and frequency.

1) Feature extraction in realm of time:

Features utilizing time characteristic of talk signal have created various methods in processing digital signal [2]. These features include:

- Average
- Zero crossing rate
- Energy
- Autocorrelation
- Average Magnitude Difference

2) Feature extraction in realm of frequency

There are many methods in the field of feature extraction of utterance frequency, it is attempted to study the most important method including cepestrum analysis of MFCC and LPC.

2-5- Filter Bank

In structure of filter banks, signal S(n) passes between Q filter that totally covers the width of filter band. Separating filter correlate with each other in frequency realm.

Various kinds of filter banks are used in utterance recognition field. In monotonous filter bank, central frequency f_i for passing filter I is equal to:

$$f_i = \frac{F_s\, i}{N} \qquad 1 < i < Q \qquad (6)$$

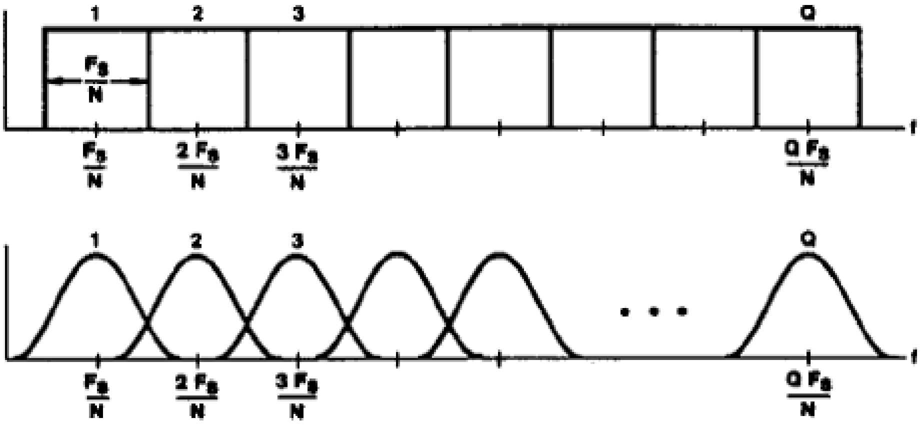

Figure 2: Exhibition of filter bank

F_s is sampling frequency and N is the number of essential filters for covering frequency realm. The number of filter must be equal to the relation of $Q \leq N/2$ and the width of *i* filter (b_i) must follow the relation of $b_i \leq \frac{F_s}{N}$

Non- monotonous filter bank is designed based on characteristic of auditory system.

2-6- Linear predictive analysis

One of the most foremost techniques of analysis and synthesis of utterance is linear predictive which is used in most of coding applications of utterance due to the capability in exhibition of utterance waveform concerning variable parameters with time [3].

In linear predictive analysis, one sample of talk signal in determined time can be assessed with linear combination from number of its previous samples. On the other hand, if S(n) is a sample in n time, it can be assessed with previous sample P as following:

$$s(n) \cong \alpha_1 S(n-1) + \alpha_2 S(n-2) + \cdots + \alpha_p S(n-p) \qquad (7)$$

If the coefficients of α_1 to α_p are assumed constancy during talk signal, below equation can be written as following:

$$s(n) = \sum_{i=1} \alpha_i z^{-1}\, S(z) + Gu(z) \qquad (8)$$

$$v(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{i-1}^{p} u_i z^{-1}} = \frac{G}{1 - p(z)} \qquad (9)$$

V(*z*) is all polar filter that was mentioned as *H*(*z*) in mathematical model of utterance produce, relation (3). The degree of the filter is p and is approximately appropriate model for channel utterance and audio platform and oral cavity. The goal of linear predictive analysis is to obtain the coefficients of the filter i.e. α_1 to α_p. Whatever the degree of the model (p) is more, more exact form of talk signal will be obtained as well as more characteristic and accurate are shown by coefficient α_1 and α_p.

2-7- Campestral Analysis

One of characteristics is extracted from utterance signal and using in most of applications is campestral analysis. Not only these coefficients have channel utterance, but also, contain of information of excitation signal. There are two campestral analyses: Analysis of campestral FFT or (DFT) and linear predictive campestral analysis. Totally, analysis of FFT campestral is divided into real cepstrum and complex cepestrum. The difference between real and complex cepestrum is that the analysis of complex cepestrum involves information of talk signal phase. Figure 3 shows how calculation of the coefficients of FFT campestral [4].

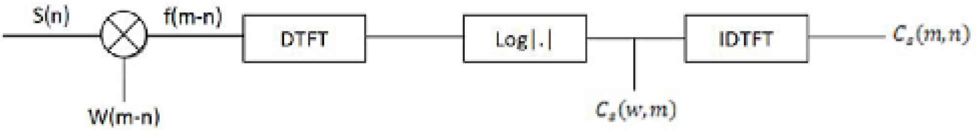

Figure 3: Calculation of real cepestrum by using DTFT

In analysis of linear predictive campestral, the coefficients of LPC campestral called LPCC are directly calculated from the coefficients linear predictive. If the coefficients of linear predictive are coefficients of filter of equation (9) and the rank P, the coefficients of LPCC are calculated as following:

$$\alpha[n] = \begin{cases} \alpha[n] + \displaystyle\sum_{k=1}^{n-1} \frac{k}{n}\, c[k]\, \alpha[n-k] & 1 \le n \le p \\ \displaystyle\sum_{k=n-p}^{n-1} \frac{k}{n}\, c[k]\, \alpha[n-k] & 1 \le n \le p \end{cases} \qquad (10)$$

In abovementioned equation, the number of coefficients of PLCC is q-appropriate. If Q in more than the number of coefficients of linear predictive i.e. p, the coefficients of LPPC are calculated to rank of p from first law and higher coefficients law of p are calculated from second law of equation (10). Then if Q is equal to or smaller than p, the coefficients of PLCC are just calculated from first law. Usually, is selected. If the equation of (9) is stated as following:

$$v(z) = \frac{G}{1 + \sum_{i=1}^{p} \alpha_i z^{-1}} \qquad (11)$$

The right side of equation of (11) must be multiplied in a negative. The coefficients of PLCC have extensive applications at utterance recognition and teller.

2-8- The Coefficients of MFCC

The main idea in using the coefficients of MFCC is to inspire with the properties of auditory of human's ear in conceiving and understanding speech. A unit of Mel of measuring of step is understood and actively is not dependent on step frequency because the operation of human's ear is how this frequency does not understand as well as physical one. These two frequencies are related to each other as following:

$$F_{mel} = 2595 \log(1 + \frac{F_{HZ}}{700}) \qquad (12)$$

Based on this fact, following steps are performed in feature extraction of MFCC. At first, immediate spectrum of frame is calculated by using method of FFT. Then on obtained spectrum, the filter must be placed algorithmically and based on mentioned equation. Figure 4 shows the distribution of kind of filter [5].

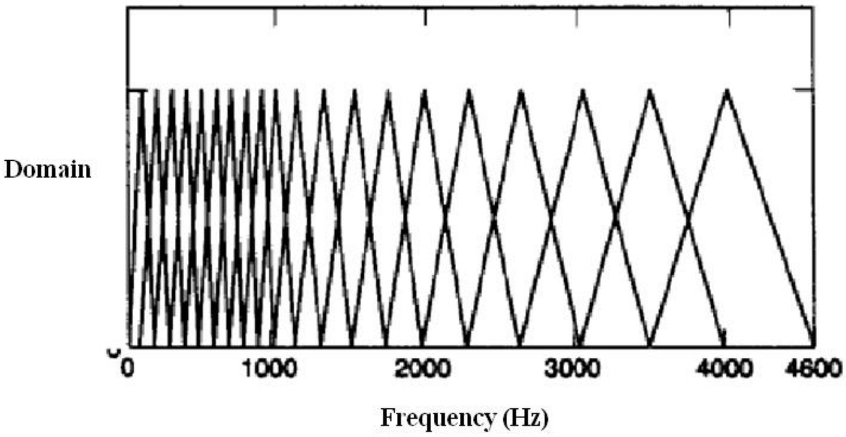

Figure 4: Triangular filters distributed based on 'Mel' criteria

After placing filter on signal spectrum, output of filters is calculated. Using these sums, the coefficients of MFCC are obtained by using below equation:

$$c(i) = \sum_{j=1}^{F} \log(x_j) \cos\left(\frac{x_i(j-0.5)}{F}\right) \qquad , \qquad 1 \le i \le F \tag{13}$$

Which in it, the number of filters and x_j obtained output are achieved from *J* filter and the coefficients of MFCC. N specified the number of coefficients. These sums are considered 13.

2-9- Derivations of Campestral Coefficients

The coefficients of campestral states appropriately the features of speech signal and have impact in increasing accuracy of speech identification. But for increasing the accuracy of system, it can be used derivations of these coefficients over time. Derivations of campestral coefficients contain of dynamic and transition information among various states of dialect and therefore combination of campestral coefficients and derivations can state better features of speech signal. If t is number of frame and I is the number of coefficient, the derivation of campestral coefficients or other hand the coefficients of campestral delta are obtained as following:

$$d_t(i) = \sum_{\tau=1}^{N} \left[\frac{\tau(c_{t+\tau}(i) - c_{t-\tau}(i))}{2\sum_{t=1}^{N} \tau^2}\right] \qquad , \qquad N \le t \le T - N \tag{14}$$

Which in this equation C_t is campestral coefficient or campestral delta coefficient over time. T is the total number of frames and N is determinant of the length of window on which is applied derivation. In first and end frames, following equations are used for calculation of derivations of coefficients [6].

$$d_t(i) = c_{t+1}(i) - c_t(i) \qquad , \qquad t < N \tag{15}$$

$$d_t(i) = c_t(i) - c_{t-1}(i) \qquad , \qquad t \ge T - N \tag{16}$$

The number of N can be optionally selected. But, usually the numbers of 2 or 3 are the most appropriate number for it. Sometimes, in denominator of equation of (14), the constancy number is used. For calculating second derivation of campestral coefficients or campestral delta coefficients, the number of campestral delta coefficients is placed in equations of (14) to (16). For calculating derivation of higher rank is performed the same methods, as well.

3- CONCLUSION

Synthesis and speech recognition technologies have created tremendous value. Talk to the fastest and most efficient way to contact people. Speaking identify potential replacement for writing, typing, keyboard input and key-button electronic control that has applied for admission and only requires that the small business market work better. In addition to talking the talk can be detected using the same computer for all physically disabled people with speech or hearing abilities that are appropriate to a flash Netter be replaced by the environments friendly output visual symptoms (of lights) and auditory (tones of warning ...) All messages express the words used to address the system needs to optimize these messages.

4- REFERENCES

[1] G. E. Hinton, S. Osindero, and Y. Teh, “A fast learning algorithm for deep belief nets”, Neural Comput., vol. 18, pp. 1527–1554, 2006.

[2] N. Morgan, “Deep and wide: Multiple layers in automatic speech recognition”, IEEE Trans. Audio Speech Lang. Processing, vol. 20, no. 1, Jan. 2012.

[3] A. Mohamed, G. Dahl, and G. Hinton, “Deep belief networks for phone recognition”, in Proc. NIPS Workshop Deep Learning for Speech Recognition and Related Applications, 2009

[4] Y. Hifny and S. Renals, “Speech recognition using augmented conditional random fields”, IEEE Trans. Audio Speech Lang. Processing, vol. 17, no. 2, pp. 354–365, 2009.

[5] S. Young, “Large vocabulary continuous speech recognition: A review,” IEEE Signal Processing Mag., vol. 13, no. 5, pp. 45–57, 1996.

[6] J. Martens, “Deep learning via Hessian-free optimization”, in Proc. 27th Int. Conf. Machine learning, 2010.