Effective Study Of Feature Extraction Methods For Speaker Identification

Mahesh Adhav, Dr. Jagdish D. Kene

***Abstract*— Speech processing is emerged as one of the important application in the area of digital signal processing. Various fields for research in speech processing are speech recognition, speaker recognition, speech synthesis, speech coding etc. The objective of automatic speaker recognition is to extract, characterize and recognize the information about speaker identity. Feature extraction is the first step for speaker recognition. Many algorithms are suggested by the researchers for feature extraction such as pitch extraction, formant extraction, Liner Predictive Coding (LPC) and Mel Frequency Cepstral Coefficients (MFCCs) etc. The characteristics that are commonly considered include fundamental frequency f0, duration, intensity, spectral variation and wavelet based feature. This paper helps in choosing the techniques along with their relative merits & demerits. The conclusion is based on comparison of different extraction techniques with reference to above mentioned parameters.**

***Index Terms*—Feature extraction, Pitch, Formant, LPC, MFC, and PLP.**

I. INTRODUCTION

In daily acoustic environments, the sound arriving at our ears often comes from multiple sources but human ear has ability to focus on the speech signal from a conversation partner while ignoring the acoustic signal from the other talkers in the environment, this ability is known as the "cocktail party problem". It is a psychoacoustic phenomenon that refers to the remarkable human ability to selectively attend to and recognize one source of auditory input in a noisy environment. The cocktail party problem (CPP) first proposed by Colin Cherry [1]. He has described the concept of human ear as a speech recognizer. To model an equivalent human hearing system, it is important to understand the working of human auditory system. At the linguistic level of communication, first the idea is formed in the mind of the speaker. The idea is then transformed to words, phrases and sentences according to the grammatical rules of the language [2]. At the physiological level of communication the brain creates electric signal that moves along the motor nerves. These electric signals activate muscles in the vocal tract and vocal cords. This vocal tract and vocal cord movements results in pressure changes within the vocal tract and in particular at the lips, initiates a sound wave that propagates in space. Speech signals are composed of a sequence of sounds and the sequence of sounds are produced as a result of acoustical excitation of the vocal tract when air is expelled from the lungs. The sound is connected with human emotion, health and environment, etc. In the environment varieties of speakers available, the above concept can be used to recognize an individual.

This paper is organized as follows; Section I gives brief introduction about human auditory system. An overview of feature extraction with various techniques is described in Section II, the conclusion and future work is defined in Section III.

***Mahesh Adhav**, Department of Electronics and Telecommunication System, Government College of Engineering, Amravati, Amravati, India, 8421646534*

***Jagdish D. Kene**, Department of Electronics and Telecommunication System, Government College of Engineering, Amravati, 7038727032,*

II. FEATURE EXTRACTION

The feature extraction process aims to extract a compact, efficient set of parameters that represent the acoustic properties observed from input speech signal, for subsequent utilization by acoustic modeling. The feature extraction is a lossy (non-invertible) transformation. It is not possible to reconstruct the original speech from its features. At the core of the feature extraction lies the short-term spectral analysis like discrete Fourier transform, accompanied with several signal processing operations. The basic principle here is to extract a sequence of features for each short-time frame of the input signal, with an assumption that such a small segment of speech is sufficiently stationary to allow meaningful modelling [3]. The efficiency or response of this phase is important for the next phase since it affects the behavior of modelling process. The reason for computing the short-term spectrum is that the cochlea of the human ear performs a quasi-frequency analysis. The analysis in the cochlea takes place on a nonlinear frequency scale (known as the Bark scale or the mel scale). This scale is approximately linear up to about 1000 Hz and is approximately logarithmic thereafter. So, in the feature extraction, it is very common to perform a frequency warping of the frequency axis after the spectral computation [4]-6].

A number of speech feature extraction methods have been studied, such as Pitch extraction, Formant extraction, Linear Predictive Cepstral Coefficients (LPCCs), Mel-Frequency Cepstral Coefficients (MFCCs) and Perceptual Linear Predictive Coefficients (PLPs) which are discuss in the following sub-sections.

A. Pitch extraction

The pitch has aroused the periodicity through vocal cords vibration when madding voiced sound, pitch frequency is a

ISSN: 2278 – 7798

very important parameter using to describe the characteristic of voice excitation source. The variational range of pitch frequency for any human being is generally from 50 Hz to 500 Hz, the cycle of the male voice is 50 Hz - 300 Hz, and the female is 100 Hz - 500 Hz [5].The range of different peoples pitch frequency has a small gap and intersection. So it is difficult to recognize speaker only with it's a frequency value. The male voice's pitch frequency is generally lower than the female, therefore, pitch can be used to distribute male and female speakers.

B. Formant

The vocal tract (the portion between throat and mouth) forms the tube, which is characterized by its resonance frequency, which gives rise to formants. Formants are defined as the spectral peaks of sound spectrum, of the voice of a person. They are often measured as amplitude peaks in the frequency spectrum of the sound wave. Formant information include in spectral envelope. Formant frequency will change when the same person speak different word. Therefore formant parameter cannot be the effective in speaker recognition. Methods of fetching formant contain cepstrum method and linear forecasting method [5].

C. Linear predictive coefficient (LPC)

Linear predictive coding (LPC) method was introduced 60 years ago by Bishnu S. Atal and being used for speech vocal tracing because it represents vocal tract parameters and the data size are very suitable for speech compression [7]. LPC analyses the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal is called the residue. In LPC system, each sample of the signal is expressed as a linear combination of the previous samples. This equation is called a linear predictor and hence it is called as linear predictive coding. But in LPC frequencies are weighted equally on a linear scale while the frequency sensitivity of the human ear is close to the logarithmic

D. Mel Frequency Cepstral Coefficient (MFCC)

Mel Frequency Cepstral Coefficients (MFCC) is commonly used as feature extraction technique in speech recognition system such as the system which can be automatically recognize which is the task of recognition people from their voice [6][9]. Basic concept of MFCC method is shown in Figure 1.

a. Pre-Emphasis

Noise has a greater effect on the higher modulating frequencies than the lower ones. Hence, higher frequencies are artificially boosted to increase the signal-to-noise ratio by using pre-emphasis.

Let sound signal is x(n), then

$$X_2(n) = X(n) - a*X(n-1) \qquad 0.9 \leq a \leq 1$$

Where $X_2(n)$ is the output of filter and a is the normalization factor.

b. Framing

Speech is a non-stationary signal. If the frame is too long, it will affect time resolution and if the frame is too short, it will affect frequency resolution. There is a trade-off between time resolution and frequency resolution. Therefore pre-emphasized signal is segmented into frames of 30-40 ms with optional overlap of 1/3-1/2 of the frame size. Framing is very important part for good results because variation of amplitude is more in larger signals as compared to smaller signals.

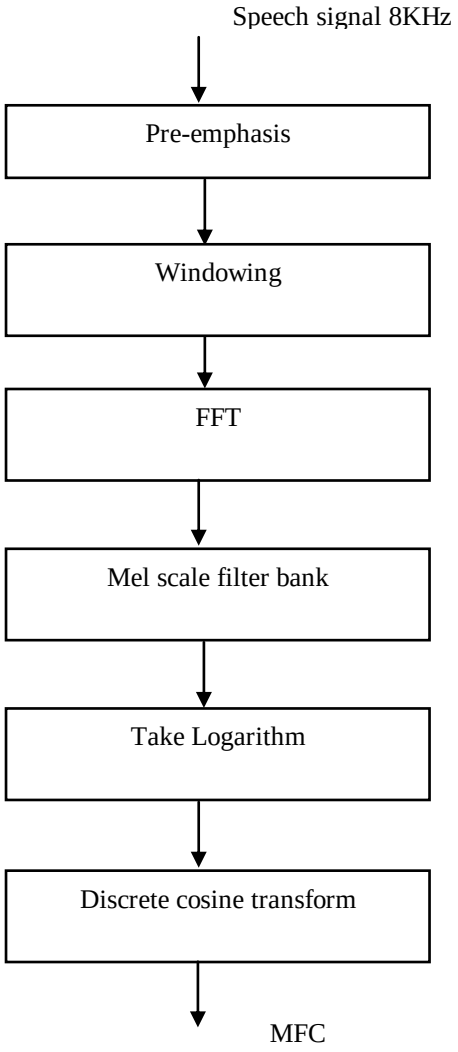

Figure 1- Block diagram of MFCC

c. Hanning Window

Hanning window is used to remove discontinuities which arise at the beginning and end of the frame, these discontinuity introduce undesirable effects in the frequency response. All frames will be multiplied with a hanning window in order to keep the continuity of the first and last points in the frame as well as achieving better efficiency compared to other windowing techniques. The input sound signal is denoted by X(n), where n = 0,1,2.....N-1, then after multiplying signal with hanning window, the output of this block is

$$W(n,\alpha) = X(n) * W(n)$$

ISSN: 2278 – 7798

$W(n,\alpha)= (1-\alpha)- \alpha \cos((2*pi*n)/(N-1))$,for $0<=n<=N-1$
Values of α will be different for different windows, normally it is 0.46 [6].

d. Fast Fourier Transform (FFT)

Fast Fourier Transform converts time domain signal in frequency domain signal [11]. While doing this transform it can be assume that signal is periodic within frame. If signal is not periodic within the frame, then compute this transform but discontinuity comes at start and end points of the frame. For dealing with this situation there are two options.

1. For increasing signal continuity at first and last points multiply all frames with hanning window.
2. Take a frame of variable size.

e. Mel-scale filter bank

It has been proved that human ears are more sensitive and have higher resolution to low frequency compared to high frequency. Hence, the filter bank is designed to emphasize the low frequency over the high frequency. Hence output of fast Fourier transform block multiplied by a set of 32 triangular bandpass filters for getting log energies of each filter. All these filters are equally spaced along mel frequency. Basic formula for converting frequency to mel-scale is as follows.

$Mel(f) = 1125 * \ln(1 + f/700)$

f. Discrete Cosine Transform (DCT)

Output of Mel-scale filter bank then passed through logarithm block. This block is used for normalization purpose. After that normalized signal is then passed through DCT block that de-correlates the log energies of filters. Finally the output of DCT block provides the MFCC coefficient values.

E. Perceptually Based Linear Predictive Analysis (PLP)

The PLP analysis has been proposed by H. Hermansky, B. A. Hanson, H. Wakita in early days, which models perceptually motivated auditory spectrum by a low order all pole function, using the autocorrelation LP technique [8]. This technique was mainly focused in cross-speaker isolated word recognition. PLP analysis results also demonstrated that speech representation is more consistent than the standard LP method. The order of PLP model is half of LPC model. This allows computational and storage saving for Automatic Speech Recognition (ASR). But the drawback of this method is that the PLP function provides limited capability of dealing with distortion. Basic concept of PLP method is shown in Figure 2.

The PLP coefficients are computed as follows:

1. The discrete time domain input signal x(n) is subject to the N - point DFT
2. The critical-band power spectrum is computed through discrete convolution of the power spectrum with the piece-wise approximation of the critical-band curve, where B is the Bark warped frequency obtained through the Hertz-to-Bark conversion.
3. Pre-emphasize the spectrum to approximate the unequal sensitivity of human hearing versus frequency that is to approximate equal loudness curves.
4. Compress the spectral amplitudes with a logarithmic compressor (approximates power-law relation between intensity and loudness).
5. Perform an inverse DFT to give cepstral coefficients.
6. Finally, the PLP coefficients are computed after autoregressive modelling and conversion of the autoregressive coefficients to cepstral coefficients.

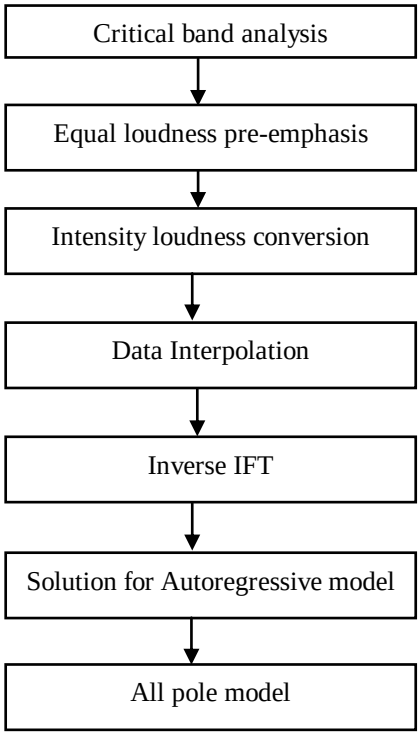

Figure 2- PLP Speech Analysis Method

III. CONCLUSION AND FUTURE WORK

In this review paper basics of feature extraction have discussed, we also discussed some features extraction techniques and their pros and cons. Through this review it is found that MFCC is used widely for feature extraction of speech. Some new methods are developed using combination of more techniques. There is another scope to develop new hybrid methods that will give better performance in robust speech recognition area.

REFERENCES

[1] E. C. Cherry, "Some experiments on the recognition of speech with one and with two ears", Journal Acoustic Society America, volume 25, 1953, pp. 975–979.
[2] L. R. Rabiner and B. H. Juang, "Fundamentals of Speech Recognition" Prentice Hall, Englewood Cliffs, N.J., 2013, pp. 54-60.
[3] L. Muda, M. Begam and I. Elamvazuthi, "Voice recognition algorithm using MFCC & DTW techniques", Journal of Computing, Volume 2, Issue 3, March 2010, pp. 138-143.
[4] N. Do Minh, "An Automatic Speaker Recognition System", Audio Visual Communications Laboratory, Swiss Federal Institute of Technology, Lausanne, 2001.

ISSN: 2278 – 7798

[5] Qiyue Liu, Mingqiu Yao, Han Xu, Fang Wang, "Research on Different Feature Parameters in Speaker Recognition", Journal of Signal and Information Processing. Volume 4, 2013, pp. 106-110.
[6] V. Tiwari, "MFCC and its applications in speaker recognition", International Journal on Emerging Technologies, Volume 1, Issue 1, 2010, pp. 19-22.
[7] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave", Journal of the acoustical society of America, Volume 50, 1971, pp. 637-655.
[8] H. Hermansky, "Perceptual linear predictive (PLP) analysis for speech", Journal of Acoustic Society of America, Volume 87, 1990 pp. 1738-1752.
[9] S. C. Joshi, and A. N. Cheeran, "MATLAB Based Feature Extraction Using Mel Frequency Cepstrum Coefficients for Automatic Speech Recognition", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 3, Issue 6, June 2014, pp. 1820-1823.
[10] S. K. Saksamudre, P. P. Shrishrimal, R. R. Deshmukh, "A Review on Different Approaches for Speech Recognition System", International Journal of Computer Applications, Volume 115, Issue 22, April 2015, pp. 23-28.
[11] J. D. Kene, K. D. Kulat, "Performance Evaluation of Physical Layer of Mobile WiMax System by Implementing Hybrid Channel Coding Scheme", Published in International Journal of Computer Applications (IJCA), vo. 94, no. 1, pp. 30-34.

Mahesh B. Adhav received the B.E. degree from the department of Electronics and Telecommunication Engineering, SSGMCE, Shegaon, India, in 2013 and pursuing M.Tech in GCOE, Amravati. His research of interest include Speech Processing and Signal Processing

Dr. Jagdish D. Kene did his bachelor degree in Electronics Engineering in 2001, from Manoharbhai Patel Institute of Engineering and Technology (MPIET), Nagpur University, Nagpur and Master degree in Electronics Engineering in 2005, from Yashavantrao Chowan College of Engineering (YCCE), Nagpur University, Nagpur, M.S. India. He did Ph-D in the field of wireless communication from Visvesvaraya National Institute of Technology (VNIT), Nagpur, M.S., India. He is currently associated with Government College of Engineering, Amravati (GCoEA), Maharashtra state, India as Associate Professor in Electronics Engineering Department having total experience of 13 years. His research work is related to Performance evaluation and optimization solution of physical layer by implementing various error correction coding techniques in mobile WiMax environment. He has published 4 Journal Papers, 6 papers in International Conferences in his research area. He also published more than 8 papers in National Conferences in his academic carrier. He is member of Professional societies like ISTE. He believes that Trust and Honesty is the secrets of success.

ISSN: 2278 – 7798