A CRITICAL REVIEW ON FEATURE EXTRACTION TECHNIQUES IN SPEECH PROCESSING

R.SWATHIKA
Ph.D. Research Scholar
Department of Computer Science
Bharathiar University, Coimbatore.
swathi19cs@gmail.com

Dr.K.GEETHA
Assistant Professor
Department of Computer Science
Bharatiar University, Coimbatore.
geethakab@gmail.com

***Abstract*— Today's trend in Speech recognition applications include automatic answering machines, dictation systems, command control applications, speaker identification system Speaker verification system etc. Feature extraction is performed by converting the speech waveform to a parametric representation for speech processing and analysis at a low data rate. As a result, exceptional and quality characteristics of speech dare are to be utilised for better results. Early feature extraction approaches and their processes are detailed in this study. Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC), Line Spectral Frequencies (LSF), Discrete Wavelet Transform (DWT), Perceptual Linear Prediction (PLP), Mel Frequency Cepstral Coefficients(MFCC), constant-Q cepstral coefficients (CQCC). Finally this paper concludes the best strategies to be adopted for applications.**

Keywords — *Linear Prediction Coefficients, Linear Prediction Cepstral Coefficients, Line Spectral Frequencies, Discrete Wavelet Transform, Perceptual Linear Prediction, Mel Frequency Cepstral Coefficients, constant-Q cepstral coefficients*

I. INTRODUCTION

Speech processing is the study of audio signals and how they are processed in various applications. Since signals are usually processed in digital representation, audio processing can be seen as a special case of digital signal processing.

Speech recognition research has been going on for a long time and it converts spoken words into text form. There are two types of speech recognition algorithms: speaker dependent and speaker independent. Speaker dependent systems focus on developing a system that can detect an individual's unique voiceprint, whereas Speaker independent systems focus solely on identifying the word pronounced by the speaker [1].

Speech recognition systems can be classified into isolated, connected, continuous, and spontaneous speech recognition systems based on the style of the speech data, and speech applications can be classified into small vocabulary, medium vocabulary and large vocabulary speech recognition systems based on the size of the vocabulary.

Method of deriving Linear Prediction coefficients (LPC's) is found by Atal and Hanauer in 1971 to analyse speech signals and linear prediction cepstral coefficients (LPCCs) derived by Atal and Sambur in 1974. In 1975, Itakura proposed the LSF representation with the help of theoretical physicist Alex Grossmann. Jean Morlet introduced the wavelet transform which allows high-frequency events to be identified with improved temporal precision. H. Hermansky introduced Perceptual Linear Prediction (PLP) in 1988 which can also be used in speech processing. Mel Frequency Cepstral Coefficients (MFCCs) are a common element in artificial speech and speaker recognition systems. Davis and Mermelstein created the MFCC in the 1980s, and used in speech applications which have been a state of the art ever since. Youngberg and Boll [22] introduced constant-Q cepstral coefficients (CQCC) in 1978 and it was refined later in 1991 by Brown[2] .

II. Speech Recognition Process

Speech Recognition system involves following steps:

A. Speech Acquisition

B. Pre-Processing

C. Feature Extraction

A. Speech Acquisition

Voice captured through any device is to be recorded as a database and further it cane used for speech applications. A combination of hardware such as headsets, microphones with the of software tools can be used to create the digitized voice data.

B. Pre processing

The speech signal is decomposed into a sequence of overlapping frames. The frames with the size of 10ms to 25ms with or without shift can be used for the analysis of speech signal. The input speech data are pre emphasized with co-efficient of 0.97 using a first order digital filter. The samples are weighted by a Hamming window for avoiding spectral distortions[3].

C. Feature Extraction

a. Linear Prediction Coefficients (LPC)

LPC is a popular formant estimation method and a powerful speech processing tool [4]. Linear prediction

coefficients (LPC) are a voice characteristic that mimics the human vocal tract [6]. It estimates the concentration and frequency of the left-over residue by estimating the formants, removing their effects from the speech signal and evaluating the speech signal. Each sample of the signal is stated to be a direct assimilation of previous samples in the result.

The formants are defined by the coefficients of the difference equation, hence LPC must estimate these coefficients [5]. The formant frequencies are the frequencies at which the resonant crests occur. By computing the linear predictive coefficients above a sliding window and locating the crests in the spectrum of the following linear prediction filter, the positions of the formants in a speech signal can be predicted. LPC is beneficial for encoding high-quality speech at a low bit rate.

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \qquad (1)$$

where $\hat{s}$ is the predicted sample, s is the speech sample, p is the predictor coefficients.

$$e(n) = s(n) - \hat{s}(n) \qquad (2)$$

From that, each frame of the windowed signal is auto correlated, and the linear prediction analysis order is determined by the highest autocorrelation value[8]. The LPC analysis follows, with each frame of the autocorrelations being turned into an LPC parameters set, which contains the LPC coefficients. The block diagram for the LPC feature extraction procedure is shown in Figure 1.

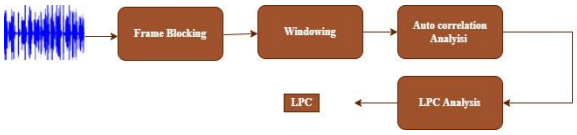

Figure.1. Block diagram for the feature extraction process of LPC

$$a_m = \log\left[\frac{1-k_m}{1+k_m}\right] \qquad (3)$$

The linear prediction coefficient is a_m while the reflection coefficient is k_m.

The vocal tract information from a particular utterance is efficiently selected using linear predictive analysis. It is well-known for its precision and speed of computation.

b. *LPCC*

LPCC has proven to be one of the most popular ways to analyze speech signals. The key idea underlying this technique is that one current speech sample may be predicted using a linear combination of previous speech samples. The LPCC algorithm is shown in Figure.2.

Pre-processing is the initial step to derive coefficients in which speech signal is separated into frames. This is the same as using a windowing function to multiply the full voice sequence. When the windowing procedure is finished, it follows LPA. LPA is based on the idea that the type of the sound being delivered is determined by the form of the vocal tract. To depict the vocal tract, a mechanized all-post channel is used, with a move work spoken to in the z area, as shown in equation 4.

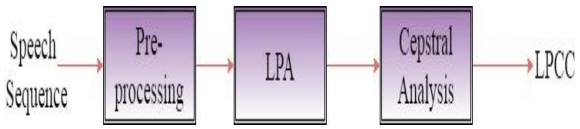

Figure.2. Block diagram for the feature extraction process of LPCC

$$V(z) = \frac{G}{1-\sum_{K=1}^{P} a_K Z^{-K}} \qquad (4)$$

where V(z) represents the work done on the vocal tract transfer. G is the filter's gain, a_k is the set of auto regression coefficients known as Linear Prediction Coefficients (LPC), and p is the all-pole filter's request. The Autocorrelation process [4] is one of the most effective methods for determining the LPC coefficients and filter gain. Cepstral analysis is the final phase of this technique, and it refers to the process of determining the cepstrum of a voice sequence. FFT cepstrum and LPC cepstrum are the two basic forms of cepstral methods. The real cepstrum in the previous scenario is defined as the backwards FFT transform of the logarithm of the speech size range, as defined by equation 2.

$$\hat{s}[n] = \frac{1}{2\pi}\int_{-\pi}^{+\pi} \ln[s(\omega)]\, e^{j\omega n} d\omega \qquad (5)$$

However, another method for evaluating these cepstral coefficients is to use a series of recursive techniques to extract them from the LPC, and the coefficients obtained in this way are known as straight Linear Prediction Cepstral Coefficients (LPCC).

c. *Linear Spectral Frequencies (LSF)*

Line spectral frequencies refer to the individual lines of the Line Spectral Pairs (LSP) (LSF). In the interconnected tube model of the human vocal tract, LSF describes the two resonance scenarios that occur. The nasal cavity and mouth shape are taken into account in the model, which provides the foundation for the linear prediction illustration's physiological significance. The vocal tract can be entirely open or completely closed at the glottis depending on the two resonance circumstances [9]. The two conditions produce two groups of resonant frequencies, with the number of resonances in each group determined by the number of connected tubes. The odd and even line spectra,

respectively, represent the resonances of each circumstance, and they are interlaced into an uniquely ascending group of LSF[10,11].

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \quad (6)$$

$\hat{s}$ (n) is the predictor signal, and a_k is the LPC coefficients, where k is the time index and p is the order of the linear prediction. The a_k coefficients are calculated using the autocorrelation or covariance methods to lower the prediction error.

$$H(Z) = \frac{1}{A(Z)} = \frac{1}{1-\sum_{i=1}^{p} aiz^{-1}} \quad (7)$$

where $A(z)$ is the LPC analysis filter and $H(z)$ is the all-pole filter. An inverse polynomial $P(z)$ and $Q(z)$ are used to calculate the LSF coefficients.

$$P(z) = A(Z) + Z^{-(p+1)} A(Z^{-1}) \quad (8)$$

$$Q(z) = A(Z) - Z^{-(p+1)} A(Z^{-1}) \quad (9)$$

where $P(z)$ is the vocal tract with the glottis closed, $Q(z)$ is the LPC analysis filter of order P.

In order to convert LSF back to LPC, the equation below is used

$$A(z) = 0.5[P(Z) + Q(Z)] \quad (10)$$

The LSF processor is shown in Figure.3 as a block diagram. LSF's most well-known use is in the field of voice compression, with extensions into speaker and speech recognition.

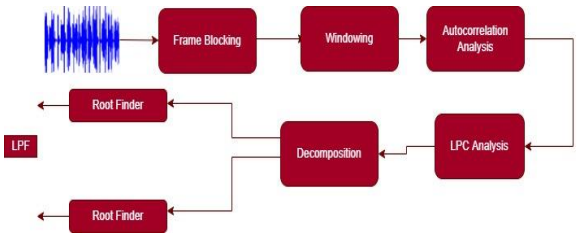

Figure.3. Block diagram of the LSF

d. Discrete wavelet transform (dwt)

A wavelet is a waveform with an effective duration of zero and an average value of zero. Many wavelets are orthogonal, which is an excellent quality for compact signal representation. Wavelet is a signal processing approach that can effectively describe real-world non-stationary signals [12,13]. It can extract information from transient signals in both the time and frequency domains at the same time. A continuous-time function is broken into wavelets using the continuous wavelet transform (CWT). Moreover, there is information redundancy, and significant computer efforts are necessary to calculate all possible CWT scales and translations, limiting its applicability. The discrete wavelet transform (DWT) is an extension of the WT that improves the decomposition process flexibility.

A wavelet is a waveform having duration of zero and an average value of zero. Many wavelets are orthogonal, making them ideal for concise signal representation. WT is a signal processing approach that can accurately describe non-stationary signals in real life. It can simultaneously extract information from temporal and frequency domains from transitory signals. The continuous wavelet transform divides a continuous-time function into wavelets (CWT). Furthermore, there is information duplication, and calculating all possible CWT scales and translations requires significant computer work, restricting its applicability.

$$W_2(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \left(\frac{t-b}{a}\right) dt \quad (11)$$

where $\varphi(t)$ is the mother wavelet, a and b are uninterrupted parameter.

A particular shift denotes how well the original signal correlates to the translated and dilated mother wavelet, and the wavelet coefficient is an expansion. As a result, the wavelet representation of the original signal in reference to the mother wavelet is the coefficient group of CWT (a, b) associated with a particular signal.

$$\emptyset(t) = \sum_{n=0}^{N-1} h(n)\sqrt{2}\emptyset(2t-1) \quad (12)$$

$$\varphi(t) = \sum_{n=0}^{N-1} g(n)\sqrt{2}\emptyset(2t-n) \quad (13)$$

where ϕ(t)Þ is the scaling function, ψ(t)Þ is the wavelet function, h[n] is the an impulse v- pass response of a low-pass filter, and g[n] is an impulse response of a high-pass filter.

$$(DWT)(m,p) = \int_{-\infty}^{+\infty} x(t)\varphi_{m,p}\,dt \quad (14)$$

where ψm, p is the wavelet function bases, m is the dilation parameter and p is the translation parameter. Thus $\varphi_{m,p}$ is defined as:

$$\varphi_{m,p} = \frac{1}{\sqrt{a_0^m}} \left(\frac{t - pb_0a_0^m}{a_0^m}\right) \quad (15)$$

$$(DWT)(m,k) = \frac{1}{\sqrt{a_0^m}} \sum_n x(n) \cdot g\left(\frac{n - nb_0 a_0^m}{a_0^m}\right) \quad (16)$$

where g(*) is the mother wavelet and x[n] is the discredited signal.

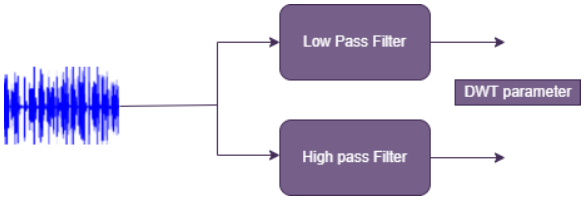

Figure.4. Block diagram of the DWT

e. Perceptual linear prediction (PLP)

Pperceptual linear prediction (PLP) technique combines crucial bands, intensity-to-loudness compression, and equal loudness pre-emphasis to extract useful information from speech [15]. Its origin can be traced back to the nonlinear bark scale, and it was created with the goal of removing speaker-dependent variables from speech recognition tasks. PLP is comparable to the MFCC in that it provides a representation that conforms to a smoothed short-term spectrum that has been equalised and compressed in a manner similar to human hearing. The PLP technique replicates numerous key properties of hearing, and an autoregressive all–pole model approximates the resulting auditory like spectrum of speech. The PLP's block diagram is shown in Fig.5.

$$bark(f) = \frac{26.81\, f}{1960 + f} - 0.53 \quad (17)$$

where bark(f) is the frequency (bark) and f is the frequency (Hz).

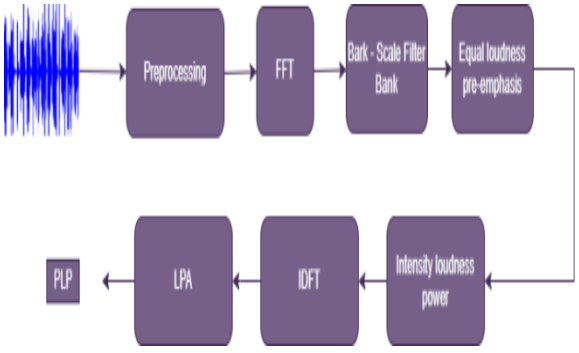

Figure.5. Block diagram of the PLP

f. Mel Frequency Cepstral Coefficient

Mel-Spectrum is computed by passing the Fourier transformed signal through a set of band-pass filters known as mel-filter bank. The Mel filter bank is built from triangular filters. Figure.6. shows block diagram of extraction of MFCC feature from a signal[17].

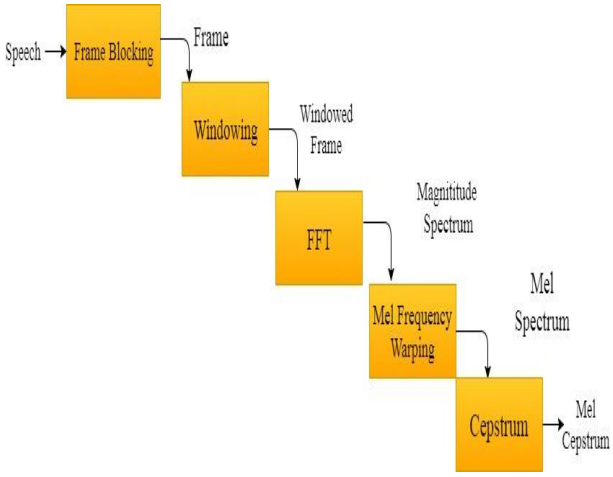

Figure.6. Block diagram for the feature extraction process of MFCC

The filters are overlapped in such a way that the lower boundary of one filter is situated at the center frequency of the next filter. 1000 Hz was defined as 1000mels. An approximate formula to compute the Mels for a given frequency in Hz is using equation 18 [18].

$$\mathrm{F(Mel)} = [2595 * \log_{10}\left[1 + \tfrac{\mathrm{f}}{700}\right]] \quad (18)$$

The most commonly used filter shaper is triangular, and in some cases the Hanning filter can be used. The mel spectrum of the magnitude spectrum X.k/ is computed by multiplying the magnitude spectrum by each of the triangular mel weighting filters.

$$s(m) = \sum_{k=0}^{N-1} [|X(k)|^2 H_m(k)] \quad (19)$$

Discrete Cosine Transform (DCT) takes after logarithm for these 26 energy values. DCT is ascertained utilizing condition appeared in equation 4.

$$C_n = \sum_{k=1}^{k} (\log S_k)\left[n(k) - {}^1/_2\right]\frac{\pi}{K} \quad (20)$$

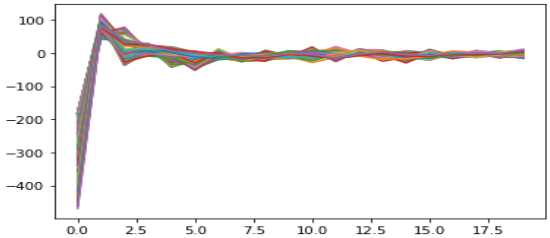

Figure.7. MFCC features using Librosa.

Figure.7. shows the diagram of Mel Frequency Scale using Librosa**.** The overlapping windows in the frequency domain can be directly used. The energy within each triangular window is obtained and followed by the DCT to achieve better compaction within a small number of coefficients and results are known as MFCC. The data will be stored in the database and take to compare with the voice input at the testing phase with same steps of process. First and derivative coefficients are calculated for each frame as given in equations 23 and 24[19]. Figure.8. Shows the various speakers using MFCC Extraction.

$$\Delta f_k[\mathrm{n}] = \mathrm{f_{k+M}}[\mathrm{n}] - \mathrm{f_{k-M}}[\mathrm{n}] \quad (21)$$

$$\Delta^2 \mathrm{f_k}[\mathrm{n}] = \Delta \mathrm{f_{k+M}}[\mathrm{n}] - \Delta \mathrm{f_{k-M}}[\mathrm{n}] \quad (22)$$

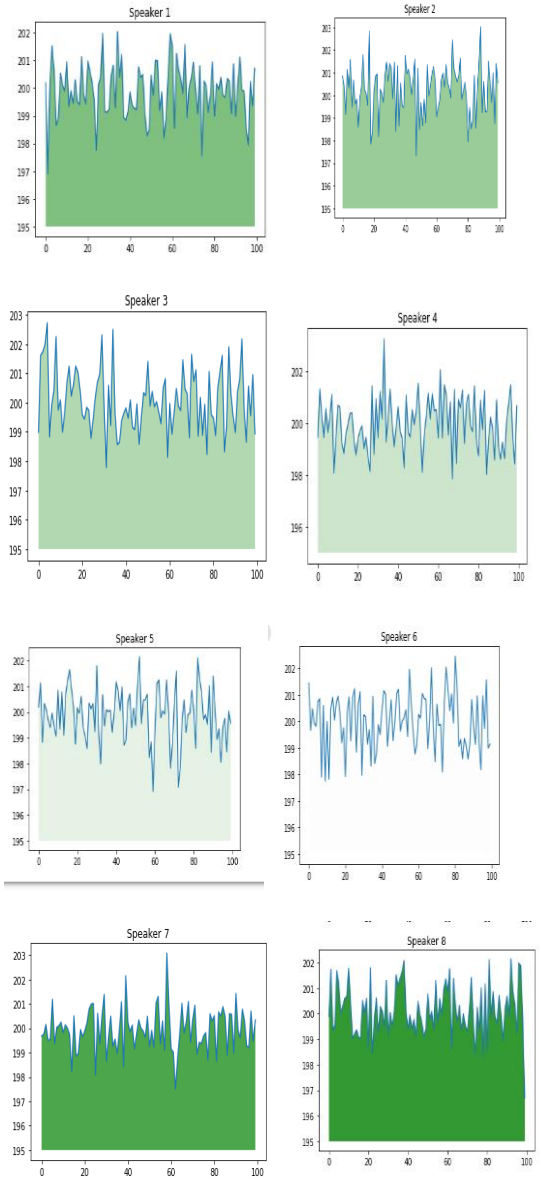

Figure.8. MFCC Features of 8 Speakers

g. *Constant Q Cepstral Coefficients (CQCC)*

CQCCs (Constant Q Cepstral Coefficients) were recently introduced in the area of ASV spoofing detection[16]. The constant Q transform and cepstral analysis are combined in the CQCC extraction. CQCCs are a more enticing option than regular MFCCs because they provide a time-frequency resolution that is closer to that of human perception.

The CQT of a discrete signal x(n) is defined by

$$X^{CQ}(k,n) = \sum_{j=[N_K/2]}^{n+[\frac{N_k}{2}]} x(j) a_k^*(j - n + \frac{N_k}{2}) \quad (23)$$

where k = 1, 2, ..., K is the frequency bin index, $a_k(n)$ are the basis functions, $*$ is the complex conjugate and N_k is a variable window length – full details are presented in equation [17]. The center frequencies f_k are defined according to $f_k = 2^{(k-1)/(B)} f_1$, where f_k is the center frequency of bin k, f1 is the center frequency of the lowest frequency bin and B is the number of bins per octave. In practice, B determines the time-frequency resolution trade-off. The Q-factor is a measure of the filter selectivity and reflects the ratio between the center frequency and the bandwidth:

$$Q = \frac{f_k}{f_{k+1} - f_k} = (2^{\frac{1}{B}} - 1)^2 \quad (24)$$

$$CC(r) = \sum_{k=0}^{k-1} \log |X^{DFT}(k)|^2 \cos\left[\frac{r(k-\frac{1}{2})\pi}{K}\right] \quad (25)$$

$$CQCC(p) = \sum_{l=0}^{L-1} \log |X^{CQ}(l)^2 \cos\left[\frac{p(l-\frac{1}{2})\pi}{L}\right] \quad (26)$$

where p = 0...L − 1 and where l is the linear scale index. The full CQCC extraction algorithm is described in [23]. Fig.9 shows the block diagram o the CQCC.

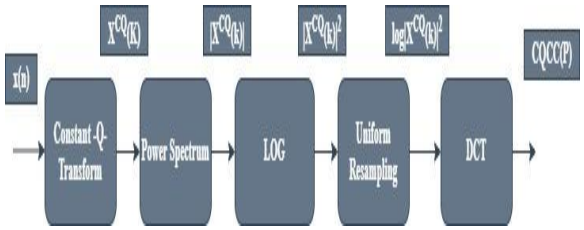

Figure.9. Block diagram of the CQCC

Table 1 shows a comparison between the seven featuere extraction techniques

Feature Extraction Techniques	Type of filter	Shape of Filter	Speed of Computation	Type of Coefficient	Noise Resistance	Frequency Captured
Linear Prediction Coefficient(LPC)	Linear Prediction	Linear	High	Autocorrelation Coefficient	High	Low and Medium
Linear prediction Cepstral Coefficient(LPCC)	Linear Prediction	Linear	Medium	Cepstral	High	Low and Medium
Line Spectral Frequencies(LSF)	Linear Prediction	Linear	Medium	Spectral	High	Low and Medium
Discrete Wavelet Transform (DWT)	LOW and High Pass Filter	---	High	Wavelet	Medium	Low and High
Perceptual Linear Prediction(PLP)	Bark	Tra[izodal	Medium	Cepstral	Medium	Low and Medium
Constant Q Cepstral Coefficients (CQCC)	Linear Prediction	Linear	High	Cepstral	Medium	Low
Mel Frequency Cepstral Coefficient(MFCC)	Mel	Triangular	High	Cepstral	Medium	Low

III. CONCLUSION

This paper discussed about the types features used for speech processing and how they can be extracted. Features like LPC, LPCC, LSF, PLP DWT, CQCC and MFCC were taken for consideration. Features like MFCC, LPCC and PLP are popularly used in the recent days. Finally a comparative study is presented based on seven features.

References

[1]. Hariharan M, Vijean V, Fook CY, Yaacob S. Speech stuttering assessment using sample entropy and Least Square Support vector machine. In: 8th International Colloquium on Signal Processing and its Applications (CSPA). 2012. pp. 240-245.

[2]. Manjula GN, Kumar MS. Stuttered speech recognition for robotic control. International Journal of Engineering and Innovative Technology (IJEIT). 2014;3(12):174-177.

[3]. Duffy JR. Motor speech disorders: Clues to neurologic diagnosis. In: Parkinson's Disease and Movement Disorders. Totowa, NJ: Humana Press; 2000. pp. 35-53.

[4]. Kurzekar PK, Deshmukh RR, Waghmare VB, Shrishrimal PP. A comparative study of feature extraction techniques for speech recognition system. International Journal of Innovative Research in Science, Engineering and Technology. 2014;3(12):18006-18016.

[5]. Ahmad AM, Ismail S, Samaon DF Recurrent neural network with back propagation through time for speech recognition. In: IEEE International Symposium on Communications and Information Technology (ISCIT 2004). Vol. 1. Sapporo, Japan: IEEE; 2004. pp. 98- 102.

[6]. Shaneh M, Taheri A. Voice command recognition system based on MFCC and VQ algorithms. World academy of science. Engineering and Technology. 2009;57:534-538

[7]. Mosa GS, Ali AA. Arabic phoneme recognition using hierarchical neural fuzzy petri net and LPC feature extraction. Signal Processing: An International Journal (SPIJ). 2009;3(5): 161.

[8]. Yousefian N, Analoui M. Using radial basis probabilistic neural network for speech recognition. In: Proceeding of 3rd International Conference on Information and Knowledge (IKT07), Mashhad, Iran. 2007.

[9]. Cornaz C, Hunkeler U, Velisavljevic V. An Automatic Speaker Recognition System. Switzerland:Lausanne; 2003. Retrieved from:http://read.pudn.com/downloads60/sourcecode/multimedia/audio/209082/asr_project.pdf

[10]. Shah SAA, ul Asar A, Shaukat SF. Neural network solution for secure interactive voice response.World Applied Sciences Journal. 2009;6(9):1264-1269.

[11]. Ravikumar KM, Rajagopal R, Nagaraj HC. An approach for objective assessment of uttered speech using MFCC features. ICGST International Journal on Digital Signal Processing, DSP. 2009;9(1):19-24

[12]. Kumar PP, Vardhan KSN, Krishna KSR. Performance evaluation of MLP for speech recognition in noisy environments using MFCC & wavelets. International Journal of Computer Science & Communication (IJCSC). 2010;1(2):41-45

[13]. Kumar R, Ranjan R, Singh SK, Kala R, Shukla A, Tiwari R. Multilingual speaker recognition using neural network. In: Proceedings of the Frontiers of Research on Speech and Music, FRSM. 2009. pp. 1-8

[14]. Narang S, Gupta MD. Speech feature extraction techniques: A review. International Journal of Computer Science and Mobile Computing. 2015;4(3):107-114

[15]. Al-Alaoui MA, Al-Kanj L, Azar J, Yaacoub E. Speech recognition using artificial neural networks and hidden Markov models. IEEE Multidisciplinary Engineering Education Magazine. 2008;3(3):77-86

[16]. Al-Sarayreh KT, Al-Qutaish RE, Al-Kasasbeh BM. Using the sound recognition techniques to reduce the electricity consumption in highways. Journal of American Science. 2009;5(2):1-12

[17]. Gill AS. A review on feature extraction techniques for speech processing. International Journal Of Engineering and Computer Science. 2016;5(10):18551-18556

[18]. Othman AM, Riadh MH. Speech recognition using scaly neural networks.World academy of science. Engineering and Technology. 2008;38:253-258

[19]. Chakroborty S, Roy A, Saha G. Fusion of a complementary feature set with MFCC for improved closed set text-independent speaker identification. In: IEEE International Conference on Industrial Technology, 2006. ICIT 2006. pp. 387-390