International Journal of Electronics Communication and Computer Engineering
Volume 3, Issue 1, ISSN 2249 –071X

Feature Extraction Techniques for Voice Operated PC Application

[1.] Ms. Gayatri V. Kapse, [1] Ms. Vishakha M. Kadam
[2.] Vijay G. Neve, [3.]Atul D. Raut

[1.] U.G. Student of Info. Tech. Jawaharlal Darda Institute of Engg. & Tech.Yavatmal (M.S.)
[2.] Assoc. Prof. & Head of Electrical Engg Dept., Jagadambha Coll. Of Engg. & Tech. Yavatmal (M.S.)
Member of Editorial Board-IJECCE
[3.] Lecturer, of Info. Tech Dept. Jawaharlal Darda Institute of Engg. & Tech. Yavatmal (M.S.)
[1.] kapsegayatri8@gmail.com , [1.]vishukadam28@gmail.com
[2.] vijay_neve@rediffmail.com, [3.]atuldraut@gmail.com

***Abstract* — Feature extraction is process of obtaining different features such as power, pitch, and vocal tract configuration from the speech signal. Feature extraction of speech is one of the most important problems in the field of speech recognition and representative of the speech. This phase is proceeded right after the input speech is pre-emphasized and windowed. A novel method for speech recognition is presented, utilizing nonlinear signal processing techniques to extract time-domain based, reconstructed phase space derived features. These nonlinear methodologies differ strongly from the traditional linear signal processing techniques typically employed for speech recognition. In this paper briefly discuss the signal modeling approach for speech recognition. It is followed by overview of basic operations involved in signal modeling to discover the discriminatory strength of these reconstructed phase space derived features, isolated phoneme are executed and are compared to a baseline classifier that uses Mel frequency cepstral coefficient features. Statistical methods are implemented to model these features. The results demonstrate that reconstructed phase space derived features contain substantial discriminatory power, even though the Mel frequency cepstral coefficient features outperformed them on direct comparisons. When the two feature sets are combined, improvement is made over the baseline, suggesting that the features extracted using the nonlinear techniques contain different discriminatory information than the features extracted from linear approaches. These nonlinear methods are particularly interesting, because they attack the speech recognition problem in a radically different manner and are an attractive research opportunity for improved speech recognition accuracy. Further commonly used spectral and temporal analysis techniques of feature extraction are also discussed.**

***Key Words* — A-MFCC, Cepstral, Feature extraction, MFCC, Speech signal, Spectral**

I. INTRODUCTION

Feature extraction of speech is important problems in the field of speech recognition. Feature extraction is process of obtaining different features such as power, pitch, and vocal tract configuration from the speech signal.

The speech feature extraction in a categorization problem is about reducing the dimensionality of the input vector whereas maintaining the discriminating power of the signal. As from fundamental formation of speaker identification and verification system, that the number of training and test vector needed for the classification problem grows with the dimension of the given input so need feature extraction of speech signal.

The purpose of feature extraction is to convert the speech waveform to some type of parametric representation for further analysis and processing, which is referred as the signal-processing front end. In speaker independent speech recognition difficulty found on extracting features that are somewhat invariant to changes in the speaker. Some examples of these variations include punctuation differences, male-female vocal tract difference etc. The speech signal is a slowly time varying signal. When examined over a sufficiently short period of time (between 5 and 100 ms), its characteristics are fairly stationary. However, over long periods of time (on the order of 0.2s or more) the signal characteristics change to reflect the different speech sounds being spoken. Therefore, short-time spectral analysis is the most common way to characterize the speech signal. So feature extraction involves analysis of speech signal. Broadly the feature extraction techniques are classified as spectral analysis and temporal analysis technique. In spectral analysis spectral representation of speech signal is used for analysis. In temporal analysis the speech waveform itself is used for analysis.

II. LITERATURE REVIEW

The first speech recognition system was built at Bell Labs in the early 1950's [1]. The task of the system was to recognize digits separated by pauses spoken from a single speaker. The system was built using analog electronics and it performed recognition by detecting the resonant frequency peaks of the uttered digit. Despite its crudeness, the system was able to achieve 98% accuracy, and it proved the concept that a machine could recognize human speech [1].

Research continued into the 1960's and 1970's fueled by the advent of digital computing technology with focus both on the signal processing and the pattern recognition aspects of developing a speech recognizer. The most significant contributions to speech analysis included to the development of the cepstral analysis, and linear predictive coding (LPC), which replaced the antiquated method of

1990's, research focused on expanding the capabilities of ASR systems to more complex tasks including speaker independent data sets, larger vocabularies, and noise robust recognition. Progress was made by incorporating speech tailored signal processing methods for feature extraction such as Mel frequency cepstral coefficients. Also the research community became more organized for facilitation of data and software sharing so that comparisons among researchers could be made.

Standard speech data was compiled and distributed such as the TIMIT corpus and the Re-source Management corpus. Also, speech software toolkits with open source code were made available; the Hidden Markov Modeling Toolkit (HTK) being one example [5].

III. Analysis of Problems

The problem addressed in this concern the investigation of novel acoustic modeling techniques that exploit the theoretical results of nonlinear dynamics, and applies them to the problem of speech recognition. The judgment of ASR systems is driven by results and this concept reduces the problem to the following question: "Does the use of nonlinear signal processing techniques improve the accuracy of current state of computer speech-to-text systems?" Although this is the central question addressed in this work, this research also seeks to further the understanding of the nonlinear methods employed, as well as to determine the limitations of the current techniques conventionally used in ASR [1].

IV. Spectral Analysis Techniques

The speech signal is a slowly time-varying signal. Over long periods of time the signal characteristic change to reflect the different speech sounds being spoken. Therefore, the short-time spectral analysis is the most common way to characterize the speech signal. In spectral analysis spectral representation of speech signal is used for analysis. Following are the Spectral analysis techniques [2].

A. Cepstral Analysis

The aim of cepstral analysis methods is to extract the vocal tract characteristics from the excitation source, because the vocal tract characteristics are what contain the information about the phoneme utterance [5]. One typical model used to represent the entire speech production mechanism is given in Fig. 1,

Voice

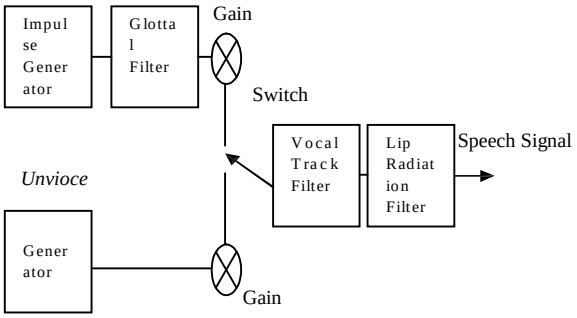

Fig 1. Block diagram of speech production model [2]

Although this model is accurate, the analysis can be made simpler by replacing the glottal, vocal tract, and lip radiation filters, by a single vocal tract filter given in the Fig. 2, this model is obtained by collapsing all these separate filters into a vocal tract filter by the convolution operation. In the linear acoustic model of speech production, the composite speech spectrum, consist of excitation signal filtered by a time-varying linear filter representing the vocal tract filter as shown in Fig 2.

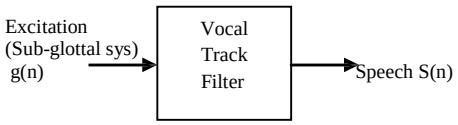

Fig 2. Linear acoustic model of speech production [2]

The speech signal is given as

$$s(n)=g(n)*v(n) \quad (1)$$

Where,
v (n): vocal tract impulse response
g (n): excitation signal
The frequency domain representation

$$s(f)=g(f)*v(f) \quad (2)$$

Taking log on both sides

$$Log_2S(f)=Log_2G(f)+Log_2V(f) \quad (3)$$

Hence in log domain the excitation and the vocal tract shape are superimposed, and can be separated. Cepstrum is computed by taking inverse discrete Fourier transform (IDFT) of logarithm of magnitude of discrete Fourier transform to obtain finite length input signal. In speech recognition cepstral analysis is used for formant tracking and pitch detection. If speech signals exhibits sharp periodic pulses it is viewed as voiced. If no such structure is visible in speech signal, the speech is considered unvoiced [2].

B. Mel frequency cepstral coefficients (MFCC)

MFCC is the best known and most popular, and has been used. MFCC is based on the known variation of the human ear's critical bandwidths with frequencies, with

filters spaced linearly at low frequencies and logarithmically at high frequencies. This is expressed in the Mel-frequency scale. MFCC provides a substantial data reduction, because a few coefficients are sufficient to represent the cepstrum of the audio signal.

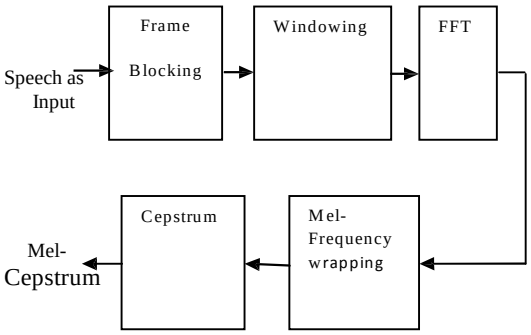

Fig 3. Block Diagram of MFCC Processor

Computation of MFCC has been explained in Fig.3, [6] currently, the most popular speech feature is the Mel filter bank cepstral coefficients (MFCC). MFCC feature vector is usually a 39 dimensional vector, consisting of 13 basic features, and their first and second derivatives. The procedure of feature extraction is summarized as shown in Fig. 4.

i. DC offset removal and pre-emphasis

This module removes the DC offset of the Speech signal and pre-emphasis the signal spectrum by approximately 20 dB per decade to compress the spectrum of the speech signal. The pre-emphasis filter is used to cancel the negative spectral slope of voiced speech signal to improve the efficiency of the spectral analysis.

ii. Framing

Human speech signal is slowly time varying and can be treated as a stationary process when considered under a short time frame. Therefore, the speech signal is usually separated into small duration blocks, called frames, and the spectral analysis is performed on these frames. The speech signal is divided into frames of 256 samples each, and a pre-emphasis filter is applied on each frame. The neighboring blocks are overlapped by 1/2 to 2/3 length of the frame and the frame shift is the frame length minus the frame overlap. The commonly used frame length and frame shift are 20-30 ms and 10 ms respectively for speech recognition task because the positions of the articulators do not change much in the period of frame length.

iii. Windowing

After being partitioned into frames, each frame is multiplied by a window function prior to the spectral analysis to reduce the effect of discontinuity introduced by the framing process by attenuating the values of the samples at the beginning and end of each frame. Commonly used windows include Hamming and Hanning windows. If no window is used, the case can be treated as the rectangular window. Each window has its own pros and cons. Compared to rectangular window, the Hamming and Hanning windows decrease the frequency resolution of the spectral analysis.

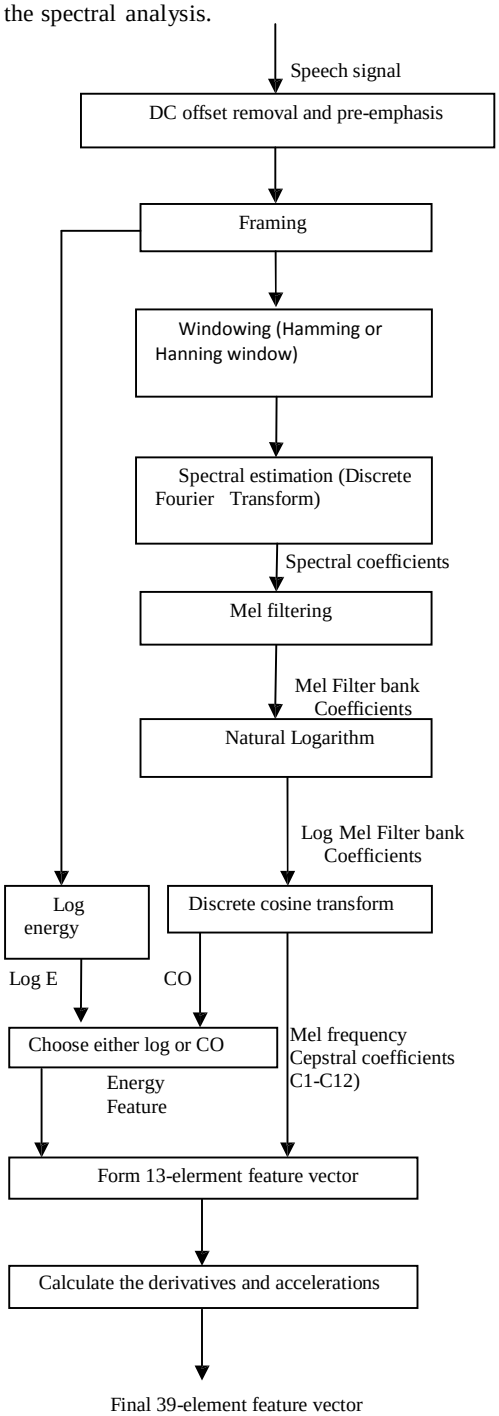

Fig 4. The common procedure of feature extraction [MFCC]

iv. Spectral estimation

The spectral coefficients of the speech frames are estimated using the fast Fourier transform (FFT) algorithm for MFCC. These coefficients are complex numbers containing both magnitude and phase information. For speech recognition tasks, the phase information is usually discarded and only the magnitudes of the spectral coefficients are extracted. It is also common to use the power of the spectral coefficients. Besides FFT, there is another spectral estimation technique called linear predictive coding (LPC) analysis which is used to extract the LPC cepstral coefficients. One difference between the LPC spectral analysis and FFT spectral analysis is that the LPC spectrum is a parametric estimate of the smoothed spectral envelope, whereas the FFT spectrum tends to provide more details of the spectrum of the speech frame as shown in Fig.5.

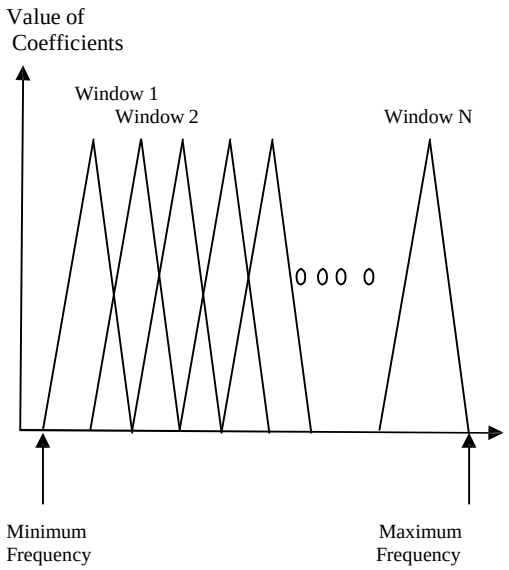

Fig 5. Example of MEL window [7]

v. Mel filtering

The spectrum of speech signal is then filtered by a group of triangle band pass filters that simulate the characteristics of human's ear. These windows are called the Mel windows and the filtering process is called Mel filtering. The Mel filtering is to model the human auditory system that perceives sound in a nonlinear frequency binning. For example, the musical pitch described in octaves and semitones is basically proportional to the logarithm of frequency. The ears analyze the spectrum of the sound in groups according to a series of overlapped critical bands. The critical bands are distributed in a way that the frequency resolution is high in low frequency region and low in high frequency region. There are several ways to distribute the critical bands and Mel frequency scale is one of them. The bandwidth of the window is narrow in low frequency and gradually increases for higher frequency. The edge of the window is arranged so that it coincides with the center of the neighbor window. To decide the location of the Mel frequency of the center of the windows, the Mel frequencies for minimum and Maximum linear frequency are first calculated using

$$fmel = 2595 \times \log(1 + f / 700) \qquad (4)$$

Where f_{Mel} is the Mel frequency for the linear frequency. The windows are evenly distributed in the Mel frequency, and the center frequencies of the windows, when converted back to linear frequency, is not linear.

vi. Natural logarithm

Whereas the Mel filtering approximates the nonlinear characteristics of human auditory system in frequency, the natural logarithm deals with the loudness nonlinearity. It approximates the relationship between the human's perceptions of the loudness and the sound intensity. Besides this, it converts the multiplication relationship between parameters into addition relationship. The convolution distortions, such as the filtering effect of microphone and channel, and the multiplication in frequency domain, such as the amplification of soft sound, become simple addition after the logarithm. Hence they can be easily removed by subtracting the mean of the coefficients. This technique is called cepstral mean subtraction/normalization.

vii. Discrete cosine transform

The DCT is applied on the log Mel filter bank coefficients to generate the cepstral coefficients and this process is a modified Homomorphic processing The Homomorphic processing is very useful in speech recognition, as it can separate the vocal tract shape function from the excitation signal of the speech production model. The lower order cepstral coefficients represent the smooth spectral shape or vocal tract shape, whereas the higher order coefficients represent the periodicity in the waveform, or the excitation information. Only the lower order coefficients are used in speech recognition systems, hence a dimension reduction is achieved. Another benefit of DCT is that the generated cepstral coefficients are less correlated than the log Mel filter bank coefficients. Therefore, it is possible to use diagonal matrix for the covariance matrix of the Gaussian in the HMM acoustical model and this significantly reduces the number of parameters in the acoustical model.

viii. Log energy calculation

In addition to the normal MFCC features, the energy of the speech frame is also used as a feature. The log energy, called log E, is calculated directly from the time-domain signal of a frame. Sometimes, it is replaced by C0.

ix. *Derivatives and accelerations calculation*

The trend of the speech signals in time is lost in the frame-by-frame analysis. To recover the trend information, the time derivatives (the first delta) and accelerations (second delta) are used. For speaker independent speech recognition system, the derivatives and accelerations are especially important. Although the location of the formant of the speech varies from person to person, the time trend of the formant are quite constant among different speakers. These derived features are simply concatenated to the original cepstral features to form the final feature.

x. Normalization

The normalization process ensures that all the features contribute equally. Without normalization, the feature with large dynamic range, such as the energy Feature C0, may dominate the Euclidean distance.

C. Robust feature extraction [A-MFCC]

Mel frequency cepstral coefficients (MFCCs) are perhaps the most widely used front ends in the state of the art speaker identification systems. One of the major problems with MFCCs is that they are very sensitive to additive noise. To overcome this bottleneck, a temporal filtering procedure on the autocorrelation sequence is proposed to minimize the effect of additive noise. The proposed feature is called Relative Autocorrelation Mel Frequency Cepstral Coefficients (A-MFCC) which is derived based on filtering the temporal trajectories of short time one sided autocorrelation sequence. This filtering process minimizes the effect of additive noise. It proposes a new approach, utilizing peaks obtained from the autocorrelation spectrum of the speech signal. This approach preserves the autocorrelation spectral peaks [4]. Computation of AMFCCs has been explained in Fig. 6.

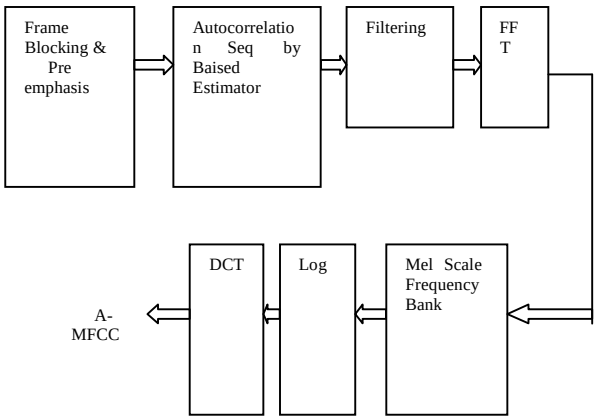

Fig 6: Block Diagram of A-MFCC Processor [6]

Firstly, calculate the autocorrelation of the noisy signal. As the temporal autocorrelation of noise is a DC or slowly varying signal, its effect is suppressed by a high-pass filter. The autocorrelation sequence of the frame signal is obtained using a biased estimator. A temporal filtering is then applied to the autocorrelation sequence to obtain the relative autocorrelation sequence in order to suppress the additive noise. A set of robust Mel-frequency cepstral coefficients are derived from the magnitude of the relative autocorrelation power spectrum by applying it to a conventional mel frequency filter-bank and finally passing its logarithm to the DCT block.

D. Critical Band Filter Bank Analysis

It is one of the most fundamental concepts in speech processing. It can be regarded as crude model of the initial stages of transduction in human auditory system. Motivation for filter bank representation [8]

1. According to "place theory" the position of maximum displacement along the basilar membrane for stimuli such as pure tones is proportional to the logarithm of the frequency of the tone [3].
2. The experiments in human perception have shown that frequencies of a complex sound within a certain bandwidth of some nominal frequency cannot be individually identified unless one of the components of this sound falls outside the bandwidth. This bandwidth is known as critical bandwidth [4], [3]. Combination of these two theories gave rise to the critical band filter bank analysis technique. Critical bank filter bank is simply bank of linear phase FIR band pass filters that are arranged linearly along the *Bark* (or *Mel*) scale. The bandwidths are chosen to be equal to a critical bandwidth for corresponding center frequency. *Bark* i.e. critical band rate scale and *Mel* scale are perceptual frequency scale defined as[3]

$$Bark = 13\arctan(0.76f/1000) + 3.5\arctan\left(f^2/(7500)^2\right) \quad (5)$$

$$Melfrequency = 2595\log_{10}(1 + f/700) \quad (6)$$

An expression for critical bandwidth is [1]

$$BW_{critical}\ 25 + 75\left[1 + 1.4(f/1000)_2\right]^{0.69} \quad (7)$$

Table 1 shows the critical filter banks based on *Bark* scale and *mel* scale. Each filter in digital filter bank is usually implemented as a linear phase filter so that the group delay for all filters is equal and the output signal from the filters are synchronized in time. The filter equations for linear phase filter implementation can be summarized as follows [8]

$$S_i(n) = \sum_{j=-(N_i-1)/2}^{(N_i-1)/2} \alpha_i(j)s(n+j) \quad (8)$$

Where $\alpha_i(j)$ denotes j^{th} coefficient for i^{th} critical band filter.

The output of this analysis is a vector of power values for each frame of data. These are usually combined with other parameters, such as total power, to form a signal

measurement vector. The Filter bank attempts to decompose the signal into discrete set of spectral samples that contain information similar to what is presented to higher levels of processing in auditory system. Because the analysis technique is largely based on linear processing, it is generally robust to ambient noise [8].

E. Principal Component Analysis (PCA)

Principal component analysis (also known as principal components analysis) (PCA) is a technique from statistics for simplifying a data set. It was developed by Pearson (1901) and Hotelling (1933), whilst the best modern reference is Jolliffe (2002). The aim of the method is to reduce the dimensionality of multivariate data whilst preserving as much of the relevant information as possible. It is a form of unsupervised learning in that it relies entirely on the input data itself without reference to the corresponding target data (the criterion to be maximized is the variance). PCA is a linear transformation that transforms the data to a new coordinate system such that the new set of variables, the principal components, are linear functions of the original variables, are uncorrelated, and the greatest variance by any projection of the data comes to lie on the first coordinate, the second greatest variance on the second coordinate, and so on. In practice, this is achieved by computing the covariance matrix for the full data set. Next, the eigenvectors and eigenvalues of the covariance matrix are computed, and sorted according to decreasing eigenvalue. Note that PCA's bias is not always appropriate; features with low variance might actually have high predictive relevance, it depends on the application.

F. Independent component analysis (ICA)

Recently, the concept of sparse coding and ICA has been successfully applied to natural signal representation. [12]. ICA was also used to elucidate the basic functions of natural images [13] and sound signals [14,15], assuming that the underlying causes have sparse or in general super-Gaussian distributions. Along this line of research, assume that speech signals are generated by a generative model where speech signals are represented as a linear combination of basic functions weighted by independent source coefficients. A frame of N observed speech samples is represented by a linear combination of N source signals as

$x = As;$

Where x is an N ×1 column vector of the speech samples, A is an N ×N mixing matrix whose column vectors constitute a set of basic functions and s is an N ×1 column vector of the source signals. In this work, assume that the source signals follow a sparse distribution. This sparseness assumption is reasonable when trying to obtain basis functions that produce an coefficient coding scheme. On the other hand, one can also adapt the source distribution using a parameterized model of the source density, such as the generalized Gaussian or exponential power density [5]. For representing speech signals however, this parameterized approach leads to source densities that have Laplacian or even sparser density models [16,15]. Both directions, namely a parameterized density model with an independence cost function and the Laplacian prior model yield similar basis functions and properties for speech signal representation. For simplicity, assume the Laplacian source model to learn the basis functions. With the assumption of the Laplacian source density. [17]

V. DATA SET AND EXPERIMENTAL RESULTS

A digital database of 200 English words spoken by 30 speakers has been used for the experiment of speaker identification system. The spoken samples are recorded by 15 male, 10 female and 5 child speakers in the studio environment using the microphone and a tape recorder. Each speaker pronounced 5 repetitions of words. The resulting database was partitioned for the use of training and testing.

1. Language: English
2. Vocabulary Size: A set of 200 most frequently use English words
3. Speakers: 30 speakers
4. Utterances: (15 male, 10 female and 5 children) 5 repetitions each
5. Audio Recording: Recording on a cassette tape in studio SNR>40dB
6. Digitization: 16KHz. Sampling, 16 bit quantization.

i. Testing on a clean Speech

The purpose of this experiment is to evaluate the performance of MFCC, A-MFCC when training data and the testing data are in a clean environment, i.e. assuming 40 dB signals to noise ratio (SNR).

I. Speaker identification rate(%)for clean speech

Feature type	Recognition rate (%)
MFCC	*98.24*
A-MFCC	*99.27*

It was observed that recognition rates are approximately identical for MFCC and A-MFCC. With the use of MFCC front end, the speaker identification rate was 98.24% and with A-MFCC it was 99.27% as given in Table (1).

ii. Testing on noisy speech

The polluted testing utterances are generated by adding the artificial noises at five SNR levels. The white noise is generated by a random number generator program, and other colored noises such as factory noise, F16 noise are extracted from the NATO RSG-10 corpus [18]. The noises are added to the clean speech signal at 20, 15, 10, 5 and 0 dB of SNR. Both MFCC and A-MFCC are evaluated and the speaker identification rates are compared with the

traditional MFCC front-end. Table II (a)-(c) show the results obtained by using MFCC, A-MFCC front-ends respectively. From the result it is obvious that A-MFCC are quite robust to the additive noise.

Table II (a) Speaker Identification rate (%) for testing speech corrupted by white noise.

Feature type	Noise levels(dB)					
	40	20	15	10	5	0
MFCC	98.2	83.8	55.8	29.3	10.5	3.7
A-MFCC	99.2	85.8	58.9	34.8	15.0	7.4

Table II (b) Speaker Identification rate (%) for testing speech corrupted by factory noise.

Feature type	Noise levels(dB)					
	40	20	15	10	5	0
MFCC	99.24	84.10	56.17	30.18	11.50	4.20
A-MFCC	99.31	86.11	58.90	35.16	16.15	8.20

Table II (c) Speaker Identification rate (%) for testing speech corrupted by f 16 noise.

Feature type	Noise levels(dB)					
	40	20	15	10	5	0
MFCC	98.0	83.2	57.1	31.4	10.8	3.7
A-MFCC	98.9	85.9	59.1	35.7	14.9	7.2

VI. TEMPORAL ANALYSIS

It involves processing of the waveform of speech signal directly. It involves less computation compared to spectral analysis but limited to simple speech parameters, e.g. power and periodicity.

A. Power Estimation

The use of power measures in speech recognition is fairly standard today. [1]In most speech recognition system Hamming window is almost exclusively used. In practice power in signal after windowing is approximately equal to the power of signal before windowing. The purpose of window is to weight, samples towards the center of the window this characteristic coupled with overlapping analysis performs important function of obtaining smoothly varying parametric estimates. The Window duration controls amount of averaging or smoothing in power calculation. Excessive smoothing can obscure true variation in the signal. Rather than using power directly in speech recognition systems use the logarithm of power multiplied by 10, defined as the power in decibels, in an effort to emulate logarithmic response of human auditory system [4]. It is calculated as

$$PowerindB = 10\log_{10}(P(n)) \qquad (9)$$

The major significance of P (n) is that it provides basis for distinguishing voiced speech segments from unvoiced speech segments. The values of P (n) for the unvoiced segments are significantly smaller than that for voiced segments. The power can be used to locate approximately the time at which voiced speech becomes unvoiced and vice versa.

B. Fundamental Frequency Estimation

Fundamental Frequency is defined as the frequency at which the vocal cords vibrate during a voiced sound. Fundamental frequency has long been difficult parameter to reliably estimate from the speech signal. Previously it was neglected for number of reasons, including large computational burden required for accurate estimation, the concern that unreliable estimation would be a barrier to achieving high performance [4]. It is useful in speech recognition of tonal languages (e.g. Chinese) and languages that have some tonal components (e.g. Japanese). Fundamental frequency is often processed on logarithmic scale, rather than a linear scale to match the resolution of human auditory system.

C. Gold and Rabiner Algorithm

It is one of earliest and simplest algorithm for *f0* estimation. In this algorithm [1] the speech signal is processed so as to create a number of impulse trains which retain the periodicity of the original signal and discard features which are irrelevant to the pitch detection process. This enables use of very simple pitch detectors to estimate the period of each impulse train. The estimates of several of these pitch detectors are logically combined to infer the period of the speech waveform. The algorithm can be efficiently implemented either in special purpose hardware or on general-purpose computer.

D. Cepstrum based pitch determination

In the cepstrum [5], we observe that for the voiced speech there is a peak in the cepstrum at the fundamental period of the input speech segment. No such a peak appears in the cepsrtum for unvioced speech segment. If the cepstrum peak is above the preset threshold, the input speech is likely to be voiced, and position of peak is good estimate of pitch period. Its inverse provides f_0. If the peak does not exceed the threshold, it is likely that the input speech segment is unvoiced.

VII. RESULTS AND CONCLUSION

Analysis techniques for feature extraction have been studied in detail and following conclusions are drawn

1. Temporal analysis techniques involve less computation, ease of implementation. But they are limited to determination simple speech parameters like power, energy and periodicity of speech. For finding vocal tract parameters we require spectral analysis techniques.

2. Critical band filter bank decomposes the speech signal into discrete set of spectral samples containing information, which is similar to information, presented to higher levels processing in auditory system.
3. Cepstral analysis separates the speech signal into component representing excitation source and a component representing vocal tract impulse response. So it provides information about pitch and vocal tract configuration. But it is computationally more intensive.
4. Mel cepstral analysis has decorrelating property of cepstral analysis and also includes some aspects of audition.
5. MFCC is the best known and most popular method, A-MFCC contributes to better performance in terms of speaker identification. Experimental results show that the proposed approach is more effective in overcoming additive noises which are stationary in nature at low SNR's. Furthermore, this proposed method works well for different types of noises including white, F16 and factory noise.
6. ICA coefficient by using an additional transform was an effective method to reduce the dependencies among the source coefficients.
7. This paper, discussed the various technique developed in each stage of speech recognition system,
8. Following Table *II* represent the list of technique with their properties for Feature extraction.

III. Various techniques of feature extraction

Sr. No	Method	Property	Procedure for Implementation
1.	Principal Component analysis (PCA)	Non liner feature extraction method, Linear map, fast, Eigenvector-based	Traditional, eigenvector Base method, also known as karhuneu-Loeve expantion; good for Gaussian data
2.	Cepstral Analysis	Static feature extraction method.power Spectrum	Used to represent Spectrum envelope
3.	Filter Bank Analysis	Filters tuned required frequencies	
4.	Mel-Frequency cepstrum coefficients (MFCC)	Power spectrum is computed by performing Fourier Analysis	This method is used for find out the features.
5.	Independent component analysis(ICA)	Non liner feature extraction method, Linear map, iterative non-Gaussian	Blind course separation , used for de-mixing non-Gaussian distributed sources(features)

REFERENCES

[1] B. Gold and N. Morgan, Speech and Audio Signal Processing. New York: John Wiley & Sons Inc., 2000. www.pearsonhighered.com/assets/hip/us/hip_us.../0130226165.pdf
[2] Speech Recognition Using Features Extracted from Phase Space Reconstructions by Andrew Carl Lindgren, B.S. at Marquette University Milwaukee, Wisconsin May 2003 http://speechlab.eece.mu.edu/papers/Lindgren_thesis.pdf
[3] L. Roderer , the Physics and Psychophysics of Music: An Introduction, New York, Springer Verlag , 1995. http://www.gi.alaska.edu/~Roederer/pdf/religion_science.pdf
[4] D.O. Shaughnessy, Speech Communication: Human and Machine. India: University Press, 2001. http://www.ee.iitb.ac.in/~esgroup/es_mtech03_sem/sem03_paper_03307003.pdf
[5] L. R. Rainer and R. W. Schafer, Digital Processing of Speech Signals. Englewood Cliffs, New Jersey: Prentice Hall, 1978. http://mi.eng.cam.ac.uk/~ajr/SpeechAnalysis/node93.html
[6] A Novel Feature Extraction Technique for Speaker Identification International Journal of Computer Applications (0975 – 8887) Volume 16– No.6, February 2011 http://www.ijcaonline.org/volume16/number6/pxc3872720.pdf
[7] Speech Enhancement with Applications in Speech Recognition http://www.ntu.edu.sg/home/ASESChng/SpeechTechWeb/members/xiaoxiong/xiao_FYR_v2.pdf
[8] J. W. Picone, "Signal modelling technique in speech recognition," *Proc. Of the IEEE*, vol. 81, no.9, pp. 1215-1247, Sep. 1993.
[9] HOTELLING, Harold, 1933. Analysis of a Complex of Statistical Variables into Principal Components. Journal of Educational Psychology, 24(6 & 7),417–441 & 498–520.
[10] JOLLIFFE, I. T., 2002. Principal Component Analysis. Seconed. Springer Series in Statistics. New York: Springer-Verlag New York.
[11] PEARSON, Karl, 1901. On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine, Series 6 2(11), 559–572.
[12] B.A. Olshausen, D.J. Field, Emergence of simple-cell receptive Geld properties by learning a sparse code for natural images, Nature 381 (1996) 607–609.
[13] A.J. Bell, T.J. Sejnowski, The 'independent components' of natural scenes are edge Glters, Vision Res. 37 (23) (1997) 3327–3338.
[14] A.J. Bell, T.J. Sejnowski, Learning the higher-order structure of a natural sound, Network Comput. Neural Syst. 7 (1996) 261–266
[15] M.S. Lewicki, E5cient coding of natural sounds, Nat. Neurosci. 5 (4) (2002) 356–363.
[16] G.-J. Jang, T.-W. Lee, A probabilistic approach to single channel blind signal separation, in: Advances in Neural Information Processing Systems, 15, MIT Press, Cambridge, MA, 2003.
[17] T.-W. Lee, Independent Component Analysis: Theory and Applications, Kluwer Academic Publishers, Dordrecht, 1998.
[18] A. VARGA AND H. J. M. STEENEKEN, Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, Vol. 12, (1993), pp. 247–251

Author's Profile

Vijay Gajanan Neve was born at Khamgaon, Maharashtra, India on 31st August 1972. He received the B.E. degree in Electrical Engineering and the M.E degree in Electrical Power System Engineering from the Amravarti university Amravarti, INDIA, in 1994 and 2000, respectively.

Currently, he is pursuing his Ph.D. degree in power quality from the same university. He is currently working as a Associate Professor and Head of Electrical Department at Jagdhambha College of engineering, Yavatmal (M.S.), INDIA, , since 2010 to till date. He has 14 years teaching and 2 years Industrial experience. He has published several papers in International conferences and International journals. He has published in the Journal of the ***Institution of***

Engineer's (India) in Electrical Engineering, *Vol 85, Sept 2004 on Page No. 83 on* Transient Stability Analysis by Transient Energy Function Method : Closest and Controlling Unstable Equilibrium Point Approach" and also he was awarded as a Merit Certificate for the same publication. He has membership in **ISTE Life Membership.** Also he is member in editorial board of IJECCE.

His area of research interest includes power system, power quality, etc.

Atul D. Raut was born at Amravati, Maharashtra, India on 25th May 1975. He received the B.E. degree in Computer science and Engineering and the M.E degree in Computer science and Engineering from the Amravarti University, Amravarti.

He has 11 years teaching experience since 2002. He currently working as a Associate Professor at Jawaharlal Darda Institute of Engineering & Technology, Yavatmal (M.S.), INDIA, since2010 to till date. He has published several papers in International conferences and attended many seminars.

His area of research interest includes Indexing XML, etc.

Gayatri V. Kapse was born at Amravati, Maharashtra, India on 11th October 1990. She is pursuing her B.E degree in Information Technology, from Jawaharlal Darda Institute of Engineering & Technology Yavatmal (M.S.). Her area of interest includes hardware networking and communication system.

Vishakha Mukund Kadam was born at Ghatanji, Maharashtra, India on 28thJuly 1989. She is pursuing her B.E degree in Information Technology, from Jawaharlal Darda Institute of Engineering & Technology Yavatmal (M.S.). Her area of interest includes web designing.