International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
Volume No.2, Issue No.6, pp : 479-484 1 June 2013

# Review of Feature Extraction Techniques in Automatic Speech Recognition

**Shanthi Therese S.[1],Chelpa Lingam[2]**

[1]Thadomal Shahnai Engineering College, Mumbai, Affiliated to Mumbai University, INDIA

[2]Pillai's HOC College of Engineering & Technology, Rasayani, Affiliated to Mumbai University, INDIA

[1]stherese@rediffmail.com, [2]chelpa.lingam@gmail.com

## Abstract

***Speech has evolved as a primary form of communication between humans. The advent of digital technology, gave us highly versatile digital processors with high speed, low cost and high power which enable researchers to transform the analog speech signals in to digital speech signals that can be scientifically studied. Achieving higher recognition accuracy, low word error rate and addressing the issues of sources of variability are the major considerations for developing an efficient Automatic Speech Recognition system. In speech recognition, feature extraction requires much attention because recognition performance depends heavily on this phase. In this paper, an effort has been made to highlight the progress made so far in the feature extraction phase of speech recognition system and an overview of technological perspective of an Automatic Speech Recognition system are discussed.***

## Keywords

***Feature Extraction, Mel Cepstrum, MFCC, Fusion MFCC, Speech Recognition, Linear predictive coding***

## 1 Introduction

Speech is the most natural form of human communication. Automatic Speech Recognition (ASR) is the process of converting a speech signal to a sequence of words, by means of an algorithm. ASR system involves two phases. Training phase and Recognition phase. In training phase, known speech is recorded and parametric representation of the speech is extracted and stored in the speech database. In the recognition phase, for the given input speech signal the features are extracted and the ASR system compares it with the reference templates to recognize the utterance. In a speech recognition system, many parameters affect the accuracy of recognition such as vocabulary size, speaker dependency, speaker independence, time for recognition, type of speech (continuous, isolated) and recognition environment condition. The extraction and selection of the best parametric representation of acoustic signals is an important task in the design of any speech recognition system. Speech recognition algorithm consists of several stages in which feature extraction and classification are mainly important. Let us understand the classification of speech and speech recognition system.

### 1.1 Human auditory system.

To model a 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. At the physiological level of communication the brain creates electric signals that move 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. Finally at the linguistic level of the listener, the brain performs speech recognition and understanding. [1][2][3]

### 1.2 Speech sounds and its categorization

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. There are various ways to categorize speech sounds. Speech sounds based on different sources to the vocal tract. Speech sounds generated with a periodic glottal source are termed Voiced. Voiced speech is produced when the vocal cords play an active role (i.e. vibrates) in the production of the sound. Examples /a/, /e/, /i/. Likewise, sounds not generated are called unvoiced. Unvoiced sounds are produced when vocal cords are inactive. Examples /s/,/f/. Other classes are Nasal Sounds and Plosives. Nasal sounds are the one in which

sound gets radiated from nostrils and lips. Examples include /m/, /n/, /ing/. Plosive sounds are those characterized by complete closure /constriction towards front of the vocal tract. Examples inlude /p/,/t/. [2][3]

### 1.3 Elements of a Language

A fundamental distinctive unit of a language is a phoneme. Different languages contain different phoneme sets. Syllables contain one or more phonemes, while words are formed with one or more syllables, concatenated to form phrases and sentences. One broad phoneme classification for English is in terms of vowels, consonants, diphthongs, affricates and semi vowels. [2][3]

## 2 Types of Speech Recognition

Speech recognition systems can be separated in several different classes by describing what types of utterances they have the ability to recognize. These classes are classified as following.

### 2.1 Isolated Words

Isolated word recognizers usually require each utterance to have quiet (lack of an audio signal) on both sides of the sample window. It accepts single words or single utterance at a time. These systems have "Listen/Not-Listen" states, where they require the speaker to wait between utterances (usually doing processing during the pauses). Isolated Utterance might be a better name for this class.

### 2.2 Connected Words

Connected word systems (or more correctly 'connected utterances') are similar to isolated words, but allows separate utterances to be run-together with a minimal pause between them.

### 2.3 Continuous Speech

Continuous speech recognizers allow users to speak almost naturally, while the computer determines the content. (Basically, it's computer dictation). Recognizers with continuous speech capabilities are some of the most difficult to create because they utilize special methods to determine utterance boundaries.

### 2.4 Spontaneous Speech

At a basic level, it can be thought of as speech that is natural sounding and not rehearsed. An ASR system with spontaneous speech ability should be able to handle a variety of natural speech features.[4][5]

## 3 Automatic Speech Recognition system classifications

The following tree structure emphasizes the speech processing applications. Depending on the chosen criterion, Automatic Speech Recognition systems can be classified as shown in Fig.1.

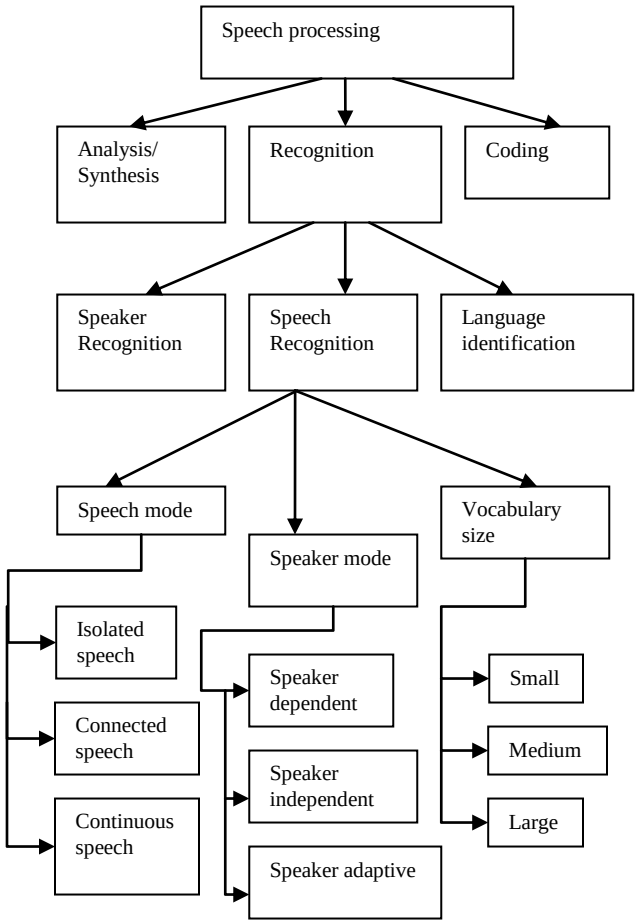



Fig 1 Speech Processing Classification

## 4. Modules of ASR

Modules that are identified to develop a speech recognition system are

1) Speech Signal acquisition
2) Feature Extraction

3) Acoustic Modelling
4) Language & Lexical Modelling
5) Recognition

**4.1 Speech signal Acquisition**

Much of the success of a speech recording depends on the recording environment and microphone placement. Ideally, speech recordings should take place in soundproof studios or labs. If those are not available, one should try to find a relatively quiet room with as little low-frequency noise as possible. Most typical sources of low-frequency noise include 60-Hz hum from electrical equipment, heating and air-conditioning ducts, elevators, doors, water pipes, computer fans, and other mechanical systems in the building. If possible, those devices should be switched off during recording. [6] MICROPHONES, PRAAT, AUDACITY, SPHINX, JULIUS are the various tools which are being used by researchers for recording speech database. [7]

**4.2 Feature Extraction in speech recognition.**

In speech recognition, feature extraction requires much attention because recognition performance depends heavily on this phase. The main goal of the feature extraction step is to compute a parsimonious sequence of feature vectors providing a compact representation of the given input signal. The feature extraction is usually performed in three stages. The first stage is called the speech analysis or the acoustic front end. It performs spectro temporal analysis of the signal and generates raw features describing the envelope of the power spectrum of short speech intervals. The second stage compiles an extended feature vector composed of static and dynamic features. Finally, the last stage transforms these extended feature vectors into more compact and robust vectors that are then supplied to the recognizer. [8][20]

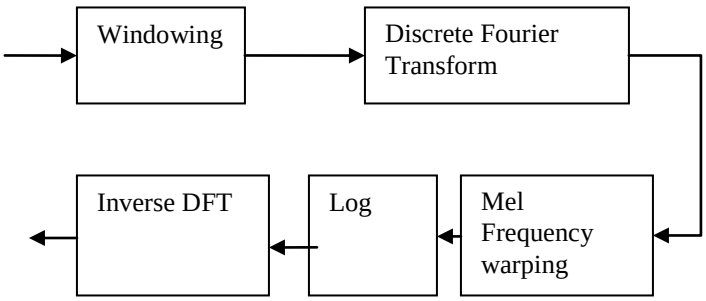



Fig 2 Feature Extraction Techniques

**4.2.1 Cepstral Analysis**

This analysis technique is very useful as it provides methodology for separating the excitation from the vocal tract shape [9] [20]. In the linear acoustic model of speech production, the composite speech spectrum consists of excitation signal filtered by a time-varying linear filter representing the vocal

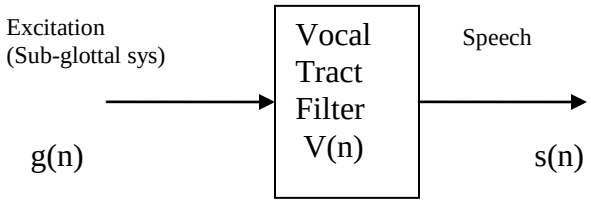



Fig 3 Linear Model of Speech Production

tract shape as shown in Fig. 3.

The speech signal is given as

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

Where $v(n)$: vocal tract impulse response
$g(n)$: excitation signal

Following is the frequency domain representation

$$s(f) = G(f) \cdot V(f) \quad \ldots \quad (2)$$

Taking log on both sides

$$\log(S(f)) = \log(G(f)) + \log(V(f)) \quad \ldots \quad (3)$$

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 finite length input signal as shown in Fig.2.

**4.2.2 Mel Cepstrum Analysis**

This analysis technique uses cepstrum with a nonlinear frequency axis following *mel* scale [10]. For obtaining *mel* cepstrum the speech waveform $s(n)$ is first windowed with analysis window $w(n)$ and then its DFT $S(k)$ is computed. The magnitude of $S(k)$ is then weighted by a series of *mel* filter frequency responses whose center frequencies and bandwidth roughly match those of auditory critical band filters.

**4.2.3 Mel-Frequency Cepstrum Coefficients (MFCC)**

A compact representation would be provided by a set of Mel-Frequency Cepstrum Coefficients (MFCC), which is the result of a cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale. [11] The performance of the Mel-Frequency Cepstrum Coefficients (MFCC) may be affected by the number of filters, the shape of filters, the way that filters are spaced and the way that the power spectrum is warped. The traditional MFCC calculation excludes the 0th coefficient. Fang Zheng, Guoliang Zhang and Zhanjiang Song have proposed that it can be regarded as the generalized Frequency Band Energy (FBE) and is hence useful, which results in the FBE-MFCC. [12][13]

**4.2.4 Linear Discriminant Analysis (LDA)**

Linear Discriminant Analysis (LDA) is commonly used technique for data classification and dimensionality reduction. Oh-Wook Kwon, Kwokleung Chan and Te-Won Lee have proposed a new kind of transformations to LDA technique using mixtures of Variational Bayesian Principal Component Analyzers (VBPCA) to analyze mel-frequency band energies and obtain proper transformations. LDA algorithm provides better classification compared to Principal Components Analysis. [14] It easily handles the case where the within-class frequencies are unequal and their performances have been examined on randomly generated test data. This method maximizes the ratio of between-class variance to the within-class variance in any particular data set thereby guaranteeing maximal reparability. The use of Linear Discriminant Analysis for data classification is applied to classification problem of speech recognition.

**4.2.5 Fusion MFCC**

Santosh Gaikwad et al. [15] performed feature extraction by combining MFCC and LDA techniques. Fusion MFCC technique was applied to the database, of vocabulary size of 625 continuous sentences. In this work, preprocessing and training phase are concentrated. In preprocessing, voiced period of speech signal was identified. By randomly selecting 500 sentences, the speech recognition system was trained using all combinations of feature extraction techniques. i.e. MFCC, LDA and Fusion MFCC. From the results obtained it is found that this feature extraction technique yields better recognition in continuous speech recognition system.

**4.2.6 Linear Predictive Coding (LPC) Analysis**

The basic idea behind the Linear Predictive Coding (LPC) analysis is that a speech sample can be approximated as linear combination of past speech samples. By minimizing the sum of the squared differences over a finite interval between the actual speech samples and the linearly predicted ones, a unique set of predictor coefficients is determined. Speech is modeled as the output of linear, time-varying system excited by either quasi-periodic pulses (during voiced speech), or random noise (during unvoiced speech).

The linear prediction method provides a robust, reliable, and accurate method for estimating the parameters that characterize the linear time-varying system representing vocal tract. Most recognition systems assume all pole model known as Auto Regressive (AR) model for speech production. The basic approach is to find set of predictor coefficients that will minimize the mean squared error over a short segment of speech waveform. The resulting parameters are then assumed to be the parameters of the system function in the model for speech production. [16]

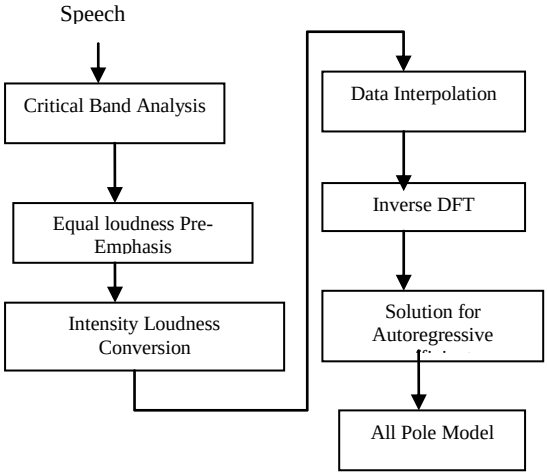
For voiced regions of speech all pole model of LPC provides a good approximation to the vocal tract spectral envelope. During unvoiced and nasalized regions of speech the LPC model is less effective than voiced region. The computation involved in LPC processing is considerably less than cepstrum analysis. Thus the importance of method lies in ability to provide accurate estimates of speech parameters, and in its relative speed.

The technique proposed by Anup Kumar Paul1, Dipankar Das, Md. Mustafa Kamal, [16] in which speech samples first get pre processed and cepstral coefficients are extracted using LPC. Using vector quantization, code book was constructed. ANN was used in the recognition phase. Multilayer perceptron approach using different number of hidden layer were tested for speech recognition in Bangla language and found that MLP using 5 hidden layers is more generic than using MLP with 3 layers. The results were found to give better recognition accuracy in an isolated word recognition and digit recognition.

**4.2.7 Perceptually Based Linear Predictive Analysis (PLP)**

H.Hermansky, B. A. Hanson, H. Wakita proposed a new PLP analysis [17], which models perceptually motivated auditory spectrum by a low order all pole function, using the autocorrelation LP technique. 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. Basic concept of PLP method is shown in block diagram of Fig. 4. It involves two major steps: Obtaining auditory spectrum and approximating the auditory spectrum by an all pole model. Auditory spectrum is derived from the speech waveform by critical-band filtering, equal loudness curve pre-emphasis, and intensity loudness root compression. The PLP analysis provides similar results as with LPC analysis but the order of PLP model is half of LP model. This allows computational and storage saving for ASR. [18] [19]




## 5 Approaches to ASR by Machine

One of the distinguishing characteristics of speech is that it is dynamic. Even within a small segment such as a phone, [21] the speech sound changes gradually. The beginning of a phone is affected by the previous phones, the middle portion of the phone is generally stable, and the end is affected by the following phones. The temporal information of speech feature vectors plays an important role in recognition process.

After feature extraction, to model the distribution of the feature vectors x (1:N) any of the following modeling technique can be used .Basically there exist three approaches of speech recognition.
They are

- Acoustic Phonetic Approach
- Pattern Recognition Approach
- Artificial Intelligence Approach

In the Acoustic Phonetic approach the speech recognition were based on finding speech sounds and providing appropriate labels to these sounds. This is the basis of the acoustic phonetic approach which postulates that there exist finite, distinctive phonetic units (phonemes) in spoken language and that these units are broadly characterized by a set of acoustics properties that are manifested in the speech signal over time.

The pattern-matching approach involves two essential steps namely, pattern training and pattern comparison. The essential feature of this approach is that it uses a well formulated mathematical framework and establishes consistent speech pattern representations, for reliable pattern comparison, from a set of labeled training samples via a formal training algorithm.

The Artificial Intelligence approach is a hybrid of the acoustic phonetic approach and pattern recognition approach. In this, it exploits the ideas and concepts of Acoustic Phonetic and Pattern Recognition methods. Knowledge based approach uses the information regarding linguistic, phonetic and spectrogram. Existing modeling approaches for speech recognition have been represented diagrammatically in the following Fig 5.

Fig 4. PLP Speech Analysis Method

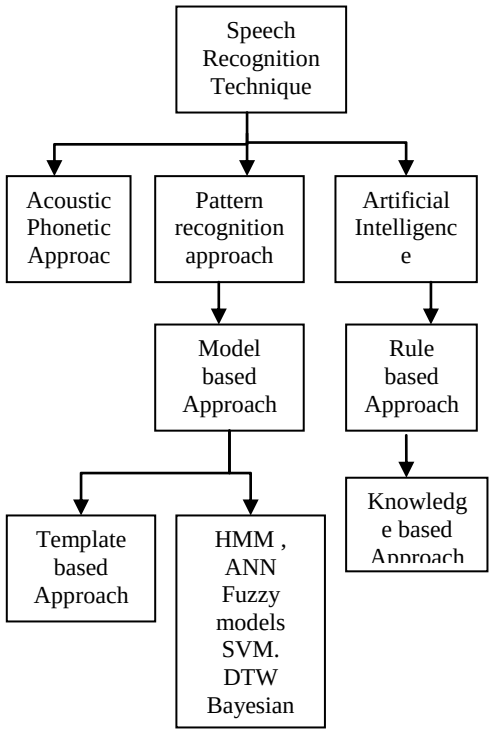



Fig 5 Speech Recognition Techniques

### 6. Performance Evaluation of Speech Recognition Systems

The performance of speech recognition systems is usually specified in terms of accuracy and speed. Accuracy may be measured in terms of performance accuracy which is usually rated with Word Error Rate (WER), whereas speed is measured with the real time factor. Other measures of accuracy include Single Word Error Rate (SWER) and Command Success Rate (CSR).

Word Error Rate (WER): Word error rate is a common metric of the performance of a speech recognition or machine translation system. The general difficulty of measuring performance lies in the fact that the recognized word sequence can have a different length from the reference word sequence. The WER is derived from the Levenshtein distance, working at the word level instead of the phoneme level. This problem is solved by first aligning the recognized word sequence with the reference (spoken) word sequence using dynamic string alignment.

Word error rate can then be computed as

$$WER = (S+D+I)/N \quad \ldots \quad (4)$$

*S* is the number of substitutions,
*D* is the number of the deletions,
*I* is the number of the insertions,
*N is* the number of words in the reference.

When reporting the performance of a speech recognition system, sometimes Word Recognition Rate (WRR**)** is used instead:

$$WRR = 1 - WER = 1 - (S+D+I) / N \quad \ldots \quad (5)$$

$$= (H - I) / N \quad \ldots \quad (6)$$

Where H = (N-S-D) is the correctly recognized words.

## 7. Conclusion and Future Work

In this paper we have reviewed, the classification and development of speech recognition system. We have also discussed the commonly used feature extraction techniques which contributes maximum recognition accuracy in any speech recognition application.

### REFERENCES

[1] *L.Rabiner and Biing-Hwang Juang, "Fundamentals of Speech approach", Prentice Hall PTR, c1993.*

*[2] Thomas F. Quatieri, Discrete-Time Speech Signal Processing: Principles and Practice, Pearson Education, Inc. 2002.*

*[3] Lawrence R. Rabiner, et. Al. Speech Recognition by Machine. 2000 CRC Press LLC.*

*[4] Meysam Mohamad pour, Fardad Farokhi, "An Advanced Method for Speech Recognition", World Academy of Science, Engineering and Technology 25, 2009.*

*[5] Simon Kinga and Joe Frankel, Recognition, "Speech production knowledge in automatic speech recognition", Journal of Acoustic Society of America, Oct 2006.*

*[6] "Best Practices in the Acquisition, Processing, and Analysis of Acoustic Speech Signals" Bartek Plichta, Michigan State University. Historicalvoices,org/flint/extras/Audio-technology.pdf*

*[7] Mathur, R., Babita, Kansal, A., "Domain specific speaker independent continuous speech recognizer using Julius", Proceedings of ASCNT – 2010, CDAC, Noida, India, pp. 55 – 60.*

*[8] Jain, R. And Saxena, S. K., "Advanced Feature Extraction & Its Implementation In Speech Recognition System", IJSTM, Vol. 2 Issue 3, July 2011*

*[9] H.Hermansky, "Perceptual Linear Predictive Analysis of Speech," The Journal of the Acoustical Society of America, vol. 87, no. 4, pp. 1738–1752, 1990*

*[10] D.O. Shaughnessy, Speech Communication: Human and Machine. Second Edition India: University Press (India) Private Limited, 2001*

*[11] Pols, L.C.W., "Spectral analysis and identification of Dutch vowels in monosyllabic words," Doctoral Dissertation, Free University, Amsterdam, The Netherlands,1966*

*[12] Fang Zheng , Guoliang Zhang , and Zhanjiang Song "Comparison of Different Implementations of MFCC", The journal of Computer Science & Technology, pp. 582-589, Sept. 2001*

*[13] Hossan, M.A. "A Novel Approach for MFCC Feature Extraction", 4th International Conference on Signal Processing and Communication Systems (ICSPCS), pp. 1-5, Dec 2010*

*[14] Oh-Wook Known, Kwokleung Chan, Te-Won Lee, "Speech Feature Analysis Using Variational Bayesian PCA", in IEEE Signal Processing letters, Vol. 10, pp.137 – 140, May 2003*

*[15] Santosh Gaikwad, Bharti Gawali, Pravin Yannawar, Suresh Mehrotra, "Feature Extraction Using Fusion MFCC For Continuous Marathi Speech Recognition", in IEEE conference (INDICON), pp 1-5, Dec.2011*

*[16] Anup Kumar Paul, Dipankar Das and Md. Mustafa Kamal, "Bangla Speech Recognition System using LPC and ANN," Seventh International Conference on Advances in Pattern Recognition, pp. 171 – 174, Feb.2009*

*[17] H.Hermansky, B. A. Hanson, H. Wakita, "Perceptually based Linear Predictive Analysis of Speech," Proc. IEEE Int. Conf. on Acoustic, speech, and Signal processing, pp. 509-512, Aug.1985*

*[18] H. Hermansky, B. A. Hanson, and H. Wakita, "Perceptually based Processing in Automatic Speech Recognition," Proc. IEEE Int. Conf. on Acoustic, speech, and Signal processing, pp. 1971-1974, Apr.1986*

*[19] B. S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," Journal of the Acoustical Society of America, vol. 55, no. 6, pp. 1304–1312, June 1974*

*[20] Audio Signal Classification M.Tech. Credit Seminar Report, Electronic Systems Group, EE. Dept, IIT Bombay, November 2004*

*[21] Dongsuk Yook , "Introduction to Automatic Speech Recognition" Department of computer science, Korea University, 2003*