# FEATURE EXTRACTION FOR SPEECH RECOGNITON

**Manish P. Kesarkar (Roll No: 03307003)**
**Supervisor: Prof. Preeti Rao**

## Abstract

Automatic speech recognition (ASR) has made great strides with the development of digital signal processing hardware and software. But despite of all these advances, machines can not match the performance of their human counterparts in terms of accuracy and speed, specially in case of speaker independent speech recognition. So today significant portion of speech recognition research is focussed on speaker independent speech recognition problem. The reasons are its wide range of applications, and limitations of available techniques of speech recognition. In this report we briefly discuss the signal modeling approach for speech recognition. It is followed by overview of basic operations involved in signal modeling. Further commonly used temporal and spectral analysis techniques of feature extraction are discussed in detail.

## 1. Introduction

Speech recognition system performs two fundamental operations: signal modeling and pattern matching [1]. Signal modeling represents process of converting speech signal into a set of parameters. Pattern matching is the task of finding parameter set from memory which closely matches the parameter set obtained from the input speech signal.

### 1.1 Motivation for signal modeling [1]

1. To obtain the perceptually meaningful parameters i.e. parameters which are analogous to those used by human auditory system.
2. To obtain the invariant parameters i.e. parameters which are robust to variations in channel, speaker and transducer.
3. To obtain parameters that capture spectral dynamics, or changes of spectrum with time.

The signal modeling involves four basic operations: spectral shaping, feature extraction, parametric transformation, and statistical modeling [1]. Spectral shaping is the process of converting the speech signal from sound pressure wave to a digital signal; and emphasizing important frequency components in the signal. Feature extraction is process of obtaining different features such as power, pitch, and vocal tract configuration from the speech signal. Parameter transformation is the process of converting these features into signal parameters through process of differentiation and concatenation. Statistical modeling involves conversion of parameters in signal observation vectors. In this report we focus on analysis techniques used for feature extraction. Section 2 briefly discusses basic operations involved in spectral shaping. Section 3.1 discusses spectral analysis techniques of feature extraction in detail. The temporal analysis techniques for feature extraction are discussed in section 3.2. Section 4 summarizes the report and conclusions are drawn.

## 2 Spectral Shaping

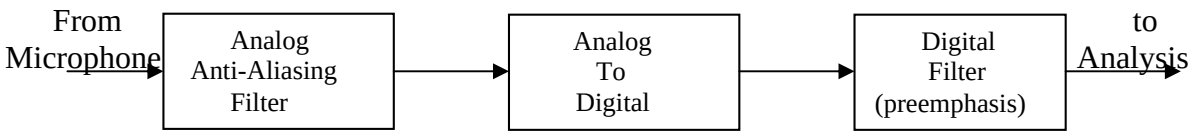


Fig. 1.Basic operations in spectral shaping "from[1]".

Spectral shaping [1] involves two basic operations: digitisation i.e.conversion of analog speech siganl from sound pressure wave to digital signal; and digital filtering i.e.emphasizing important frequency components in the signal. This process is shown in Fig.1.

The main purpose of digitisation process is to produce a sampled data representation of speech signal with as high signal-to-noise ratio (SNR) as possible. Once signal conversion is complete, the last step of digital post filtering is most often executed using a Finite Impulse Response (FIR) filter given as

$$H_{pre}(z) = \sum_{k=0}^{N_{pre}} a_{pre}(k) z^{-k} \quad (1)$$

Normally, a one coefficient digital filter known as pre-emphasis filter, is used

$$H_{pre}(z) = 1 + a_{pre}\, z^{-1} \quad (2)$$

A typical range of values for $a_{pre}$ is [-1.0,-0.4]. The preemphasis filter boosts the signal spectrum approximately 20 dB per decade.

Advantages of preemphasis filter

1. The voiced sections of speech signal naturally have a negative spectral slope (attenuation of approximately 20 dB per decade due to physiology of speech production system [3]). The preemphasis filter serves to offset this natural slope before spectral analysis, thereby improving the efficiency of the analysis [2].
2. The hearing is more sensative above the 1-kHz region of the spectrum. The preemphasis filter amplifies this area of the spectrum. This assists the spectral analysis algorithm in modelling the perceptually important aspects of speech spectrum [1].

# 3 Feature Extraction

In speaker independent speech recogniton, a premium is placed on extracting features that are somewhat invariant to changes in the speaker. So feture extraction involves analysis of speech siganl. Broadly the feature extraction techniques are classified as temporal analysis and spectral analysis technique. In temporal analysis the speech waveform itself is used for analysis. In spectral analysis spectral representation of speech signal is used for analysis.

## 3.1 Spectral Analysis techniques

### 3.1.1 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 [1]

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 [8].
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 [3], [8].

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 bandpass 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]

$$Bark = 13\text{atan}(0.76f/1000) + 3.5\text{atan}(f^2/(7500)^2) \quad (3)$$

$$mel \text{ frequency} = 2595\log_{10}(1 + f/700) \quad (4)$$

An expression for critical bandwidth is [1]

$$BW_{critical} = 25 + 75[1 + 1.4(f/1000)^2]^{0.69} \quad (5)$$

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 [1]

$$S_i(n) = \sum_{j=(N_i-1)/2}^{(N_i-1)/2} \alpha_i(j)\, s(n+j) \quad (6)$$

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 [1].

Table 1. Two Critical Filter banks, source [1].

| Index | Bark Scale | | Mel Scale | |
|---|---|---|---|---|
| | Center Freq. (Hz) | BW (Hz) | Center Freq. (Hz) | BW (Hz) |
| 1 | 50 | 100 | 100 | 100 |
| 2 | 150 | 100 | 200 | 100 |
| 3 | 250 | 100 | 300 | 100 |
| 4 | 350 | 100 | 400 | 100 |
| 5 | 450 | 110 | 500 | 100 |
| 6 | 570 | 120 | 600 | 100 |
| 7 | 700 | 140 | 700 | 100 |
| 8 | 840 | 150 | 800 | 100 |
| 9 | 1000 | 160 | 900 | 100 |
| 10 | 1170 | 190 | 1000 | 124 |
| 11 | 1370 | 210 | 1149 | 160 |
| 12 | 1600 | 240 | 1320 | 184 |
| 13 | 1850 | 280 | 1516 | 211 |
| 14 | 2150 | 320 | 1741 | 242 |
| 15 | 2500 | 380 | 2000 | 278 |
| 16 | 2900 | 450 | 2297 | 320 |
| 17 | 3400 | 550 | 2639 | 367 |
| 18 | 4000 | 700 | 3031 | 422 |
| 19 | 4800 | 900 | 3482 | 484 |
| 20 | 5800 | 1100 | 4000 | 556 |
| 21 | 7000 | 1300 | 4595 | 639 |
| 22 | 8500 | 1800 | 5287 | 734 |
| 23 | 10500 | 2500 | 6063 | 843 |
| 24 | 13500 | 3500 | 6964 | 969 |

### 3.1.2 Cepstral Analysis

This analysis technique is very useful as it provides methodology for separating the excitation from the vocal tract shape [2]. 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 shape as shown in fig.2

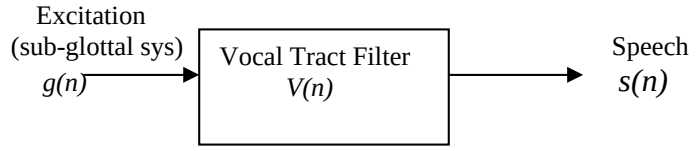


Fig.2. Linear acoustic model of speech production "from [1]".

The speech signal is given as

$$s(n) = g(n) * v(n) \tag{7}$$

where $v(n)$: vocal tract impulse response

$g(n)$: excitation signal

The frequency domain representation

$$S(f) = G(f) \cdot V(f) \tag{8}$$

Taking log on both sides

$$\log(S(f)) = \log(G(f)) + \log(V(f)) \tag{9}$$

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.3.


Fig. 3. System for obtaining cepstrum "adapted from[2]".

$$S(k) = \sum_{n=0}^{N-1} s(n)\exp(-j2\pi/N)nk \tag{10}$$

$$\hat{S}(k) = \log (S(K)) \tag{11}$$

$$\hat{s}(n) = (1/N)\sum_{k=0}^{N-1} \hat{S}(k)\, \exp(j2\pi/N)nk \tag{12}$$

$\hat{s}(n)$ is defined as cepstrum. In speech recognition cepstral analysis is used for formant tracking and pitch ($f_0$) detection. The samples of $\hat{s}(n)$ in its first 3ms describe $v(n)$ and can be separated from the excitation. The later is viewed as voiced if $\hat{s}(n)$ exhibits sharp periodic pulses. Then the interval between these pulses is considered as pitch period. If no such structure is visible in $\hat{s}(n)$, the speech is considered unvoiced.

### 3.1.3 Mel Cepstrum Analysis

This analysis technique uses cepstrum with a nonlinear frequency axis following *mel* scale [3]. 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.

The next step in determining the *mel* cepstrum is to compute the energy in this weighted sequence. If $V_l(k)$ is the frequency response of $l^{th}$ *mel* scale filter. The resulting energies are given for each speech frame at a time n and for the $l^{th}$ *mel* scale filter are

$$E_{mel}(n,l) = (1/A_l) \sum_{k=L_l}^{U_l} | V_l(k)\, S(k) |^2 \quad (13)$$

Where $U_l$ and $L_l$ are upper and lower frequency indices over which each filter is nonzero and $A_l$ is the energy of filter which normalizes the filter according to their varying bandwidths so as to give equal energy for flat spectrum.

The real cepstrum associated with $E_{mel}(n,l)$ is referred as the *mel*-cepstrum and is computed for the speech frame at time *n* as

$$C_{mel}(n,m) = (1/N) \sum_{l=0}^{N-1} \log\{E_{mel}(n,l)\} \cos[2\pi(l+1/2)/N] \quad (14)$$

Such *mel* cepstral coefficients $C_{mel}$ provide alternative representation for speech spectra which exploits auditory principles as well as decorrelating property of cepstrum.

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

The basic idea behind the linear predictive coding (LPC) analysis [2] 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 difference equation describing relation between speech samples *s(n)* and excitation *u(n)* for AR model is as follows

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + G\, u(n) \quad (15)$$

The system function is of the form

$$H(Z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \quad (16)$$

A linear predictor of order *p* with prediction coefficients $\alpha_k$ is defined as a system whose output is [2]

$$\hat{s}(n) = \sum_{k=1}^{p} \alpha_k\, s(n-k) \quad (17)$$

The system function is $p^{th}$ order polynomial

$$P(z) = \sum_{k=1}^{p} \alpha_k\, z^{-k} \quad (18)$$

The prediction error *e(n)* is defined as

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k\, s(n-k) \quad (19)$$

The prediction error sequence is the output of the system whose transfer function is

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \quad (20)$$

it can be seen by comparing eq.15 and eq.19 if $\alpha_k = a_k$ ,, then prediction error filter, *A(z),*Will be an inverse filter for the system, *H(z)* of eq.16

$$H(Z) = \frac{G}{A(z)} \quad (21)$$

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, *H(z)*, in the model for speech production. The short-time average prediction error is defined as[2]

$$E_n = \sum_m (e_n(m))^2 \quad (22)$$

$$= \sum_m \{s_n(m) - \sum_{k=1}^{p} \alpha_k s_n(m-k)\} \quad (23)$$

Where $s_n(m)$is segment of speech in vicinity of sample *n*, i.e.

$$s_n(m) = s(n+ m) \quad (24)$$

The values of $\alpha_k$ that minimize $E_n$ are obtained by setting
$\partial E_n/\partial \alpha_i = 0$ , $i = 1,2,\ldots\ldots p$, thereby obtaining the equations

$$\sum_m s_n(m-i) s_n(m) = \sum_{k=1}^{p} \alpha_k \sum_m s_n(m-i)\, s_n(m-k) \quad (25)$$

$$if \ \ \phi_n(i,k) = \sum_m s_n(m-i)\, s_n(m-k) \quad (26)$$

Then Eq. 25 is written as

$$\sum_{k=1}^{p} \alpha_k \phi_n(i,k) = \phi_n(i,0) \qquad i=1,2,3....p \quad (27)$$

There are three basic ways to solve above set of equations

1. Lattice method
2. Covariance method
3. Autocorrelation method

In speech recognition the autocorrelation method is almost exclusively used because of its computational efficiency and inherent stability. The autocorrelation method always produces a prediction filter whose zero lies inside the circle in z-plane [2].

In autocorrelation method the speech segment is windowed as follows

$$s_n(m) = s(m+n)w(m) \quad (28)$$

where *w(m)* is finite length window i.e .zero outside interval $0 \le m \le N-1$

Then

$$\phi_n(i,k) = \sum_{m=0}^{N+p-1} s_n(m-i)\, s_n(m-k) \qquad \begin{matrix} 1 \le i \le p \\ 0 \le k \le p \end{matrix} \quad (29)$$

$$\phi_n(i,k) = R_n(i-k)$$

where

$$R_n(k) = \sum_{m=0}^{N-1-k} s_n(m)\, s_n(m+k) \quad (30)$$

$R_n(k)$ is autocorrelation function eq. 27 is simplified as [2]

$$\sum_{k=1}^{p} \alpha_k R_n(|i-k|) = R_n(i) \qquad 1 \le i \le p \tag{31}$$

This results in p×p matrix of autocorrelation values, it is topelitz matrix; i.e. symmetric and all elements along diagonal are equal. The resulting equations are solved using Durbins recursive procedure as [2]

$$E^{(0)} = R(0) \tag{32}$$

$$k_i = \{ R(i) - \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R(i-j) \} / E^{(i-1)} \tag{33}$$

$$\alpha_i^{(i)} = k_i \tag{34}$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{(j-1)}^{(i-1)} \tag{35}$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)} \tag{36}$$

Eq. 31 to 34 are solved recursively for i =1,2…p and final solution is given as

$\alpha_j$ = LPC coefficients = $\alpha_j^{(P)}$

$k_i$ = PACOR coefficients

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.

A very important LPC parameter set which is derived directly from LPC coefficients is LPC cepstral coefficients $c_m$. The recursion used for this is [6]

$$c_0 = \ln G \tag{37}$$

$$c_m = \alpha_m + \sum_{k=1}^{m-1} (k/m) c_k a_{m-k} \qquad 1 \le m \le p \tag{38}$$

$$c_m = \sum_{k=1}^{m-1} (k/m) c_k a_{m-k} \qquad m > p \tag{39}$$

Where $G$ is the gain term in LPC model. This method is efficient, as it does not require explicit cepstral computation. Hence combines decorrelating property of cepstrum with computational efficiency of LPC analysis.

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

PLP analysis [5] models perceptually motivated auditory spectrum by a low order all pole function, using the autocorrelation LP technique. Basic concept of PLP method is shown in block diagram of Fig. 4.

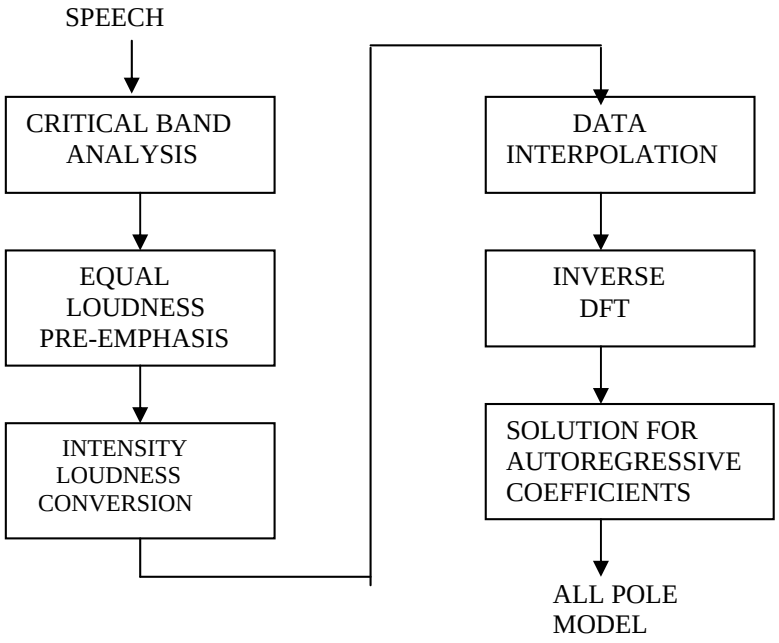


Fig 4 Block diagram of PLP speech analysis method "from [5]".

It involves two major steps: obtaining auditory spectrum, 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. Eighteen critical band filter outputs with their center frequencies equally spaced in bark domain, are defined as

$$\Omega(w) = 6 \ln \left( (w/1200\pi) + \left( (w/1200\pi) + 1 \right)^{0.5} \right) \tag{40}$$

Center frequency of $k$th critical band $\Omega_k = 0.994$k

They cover the frequency range $0 \le f \le 5$kHz ($0 \le \Omega \le 16.9$ Bark). Each filter is simulated by spectral weighing $c_k(w)$.The filter output is multiplied by equal loudness function $E(w_k)$ and converted into loudness domain using Stevens power law with exponent $r = 3$ as in[5]

$$Q_k = \left[ E(w_k) \int_0^{\pi} c_k(w).P(w)\, dw \right]^{1/r} \tag{41}$$

Critical band weighing function is [5]

$$\begin{aligned} c_k(w) &= 10^{(\Omega - \Omega k + 0.5)} && \Omega \le \Omega_k - 0.5 \\ &= 1 && \Omega_k - 0.5 \le \Omega \le \Omega_k + 0.5 \\ &= 10^{2.5(\Omega - \Omega k + 0.5)} && \Omega_k + 0.5 \le \Omega \end{aligned} \tag{42}$$

The output thus obtained is linearly interpolated to give interpolated auditory spectrum. The interpolated auditory spectrum is approximated by fifth order all pole model spectrum. The IDFT of interpolated auditory spectrum provides first six terms of autocorrelation function. These are used in solution of Yule Walker equations [6] to obtain five autoregressive coefficients of all-pole filter. 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. Also it provides better performance to cross speaker ASR [7].

## 3.2Temporal 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.

### 3.2.1 Power Estimation

The use of some sort of power measures in speech recognition is fairly standard today. Power is rather simple to compute. It is computed on frame by frame basis as [1]

$$P(n) = (1/N_s) \sum_{m=0}^{N_s-1} \left( w(m)\, s(n - N_s/2 + m) \right) \quad (43)$$

Where $N_s$ is the number of samples used to compute the power, $s(n)$ denotes the signal, $w(m)$ denotes the window function, and the ,and $n$ denotes the sample index of center of the window. In most speech recognition system Hamming window is almost exclusively used. The Hamming window is specific case of Hanning window, which is given as

$$w(n) = \frac{\alpha_w - (1 - \alpha_w)\cos(2n\pi/(N_s-1))}{\beta_w} \qquad \text{For } 0 \le n \le N_s - 1$$

$$w(n) = 0 \qquad \text{elsewhere.} \quad (44)$$

$\alpha_w$ is window constant in the range[0,1],and $Ns$ is the window duration in samples. For Hamming window $\alpha_w$=0.54, $\beta_w$ is normalization constant.

In practice normalization is done so 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. Large amount of over lap results in reduction of amount of noise introduced in measurements by artifacts such as window placement and nonstationary channel noise [2]. However 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 [3]. It is calculated as

$$\text{Power in dB} = 10 \log_{10}(P(n)) \quad (45)$$

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.

### 3.2.2 Fundamental Frequency Estimation

Fundamental Frequency ($f_0$) or pitch 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, and difficulty in characterizing complex interactions between $f_0$ and suprasegmental phenomenon [3]. 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.

There are various algorithms to estimate $f_0$ we will consider two widely used algorithms: Gold and Rabiner algorithm [4], cepstrum based pitch determination algorithm [2].

#### 3.2.2.1 Gold and Rabiner algorithm

It is one of earliest and simplest algorithm for $f_0$ estimation. In this algorithm [4] 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.

#### 3.2.2.2 Cepstrum based pitch determination

In the cepstrum [2], 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.

## 4 Conclusions

The basic operations in speech recognition system have been discussed briefly. Different temporal and spectral 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. LPC analysis provides compact representation of vocal tract configuration by relatively simple computation compared to cepstral analysis. To minimize analysis complexity it assumes all pole model for speech production system. But speech has zeros due to nasals so in these cases the result are not as good as in case of vowels but still reasonably acceptable if order of model is sufficiently high.
6. LP derived cepstral coefficients have decorrelating property of cepstrum and computational ease of LPC analysis.
7. PLP analysis uses includes certain aspects of audition provides similar spectral estimate of speech as LPC analysis but with lower order model. Also it provides better performance to cross speaker ASR [9].

**Acknowlegdement**

I wish to express my sincere gratitude to Prof. Preeti Rao for her constant guidance throughout the course of the work and many useful discussions which enabled me to know the subtleties of the subject in proper way.

## References

[1] J. W. Picone, "Signal modelling technique in speech recognition," *Proc. Of the IEEE,* vol. 81, no.9, pp. 1215-1247, Sep. 1993.

[2] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, New Jersey: Prentice-Hall, 1978.

[3] D.O. Shaughnessy, *Speech Communication: Human and Machine.* India:University Press ,2001.

[4] B. Gold and L. R. Rabiner,"Parallel processing techniques for estimating pitch periods of speech in the time domain," *J. Acoust. Soc. America,* vol.46, pt. 2, no. 2, pp 442-448, Aug. 1969.

[5] H. Hermansky, B. A. Hanson, and H. Wakita, "Perceptually based linear predictive analysis of speech," *Proc. IEEE Int. Conf. on Acoustic, speech, and Signal Processing,"* pp. 509-512, Aug.1985.

[6] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition,* Englewood Cliffs, New Jersey: Prentice- Hall, 1978.

[7] 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.

[8] L. Roderer, *The Physics and Psychophysics of Music: An Introduction,* New York, Springer Verlag, 1995.