Journal of Information Systems and Communication
ISSN: 0976-8742 & E-ISSN: 0976-8750, Volume 3, Issue 1, 2012, pp.-358-364.
Available online at http://www.bioinfo.in/contents.php?id=45

A NOVEL ALGORITHM USING MFCC AND ERB GAMMATONE FILTERS IN SPEECH RECOGNITION

LAVNEET SINGH[1], GIRIJA CHETTY[2] AND SAVLEEN SINGH[3]

ISE, University of Canberra, Australia
*Corresponding Author: Email- Lavneet_agra@yahoo.co.in[1], Girija.Chetty@canberra.edu.au[2] and savleenagra@gmail.com[3]

Received: January 12, 2012; Accepted: February 15, 2012

Abstract- Automatic Speech Recognition has been an active topic of research for the past four decades. The main objective of the automatic speech recognition task is to convert a speech segment into an interpretable text message without the need of human intervention. Many different algorithms and schemes based on different mathematical paradigms have been proposed in an attempt to improve recognition rates. Cepstral coefficients play an important part in speech theory and in automatic speech recognition in particular due to their ability to compactly represent relevant information that is contained in a short time sample of a continuous speech signal. The goal of this paper is to discuss comparison of speech parameterization methods: Mel-Frequency Cepstrum Coefficients (MFCC) and improved Mel-Frequency Cepstrum Coefficients (MFCC) using ERB GAMMATONE filters. First, we remove signal correlation through normalization, then we use ERB GAMMATONE filter to filtering the cepstral coefficients. Finally, we reduce dimension of the cepstral coefficients by the variances of cepstral coefficients in different dimension and obtain our features. By using various classifiers, we try to simulate the speech feature extraction at much optimal and least error rate providing robust method for Automatic Speech Recognition (ASRs) .
Keywords- Automatic Speech Recognition, Mel frequency Cepstrum Coefficients (MFCC's), ERB Gammatone Filtering, Hidden Markov Model

Citation: Lavneet Singh, Girija Chetty and Savleen Singh (2012) A Novel Algorithm Using MFCC and ERB Gammatone Filters in Speech Recognition. Journal of Information Systems and Communication, ISSN: 0976-8742 & E-ISSN: 0976-8750, Volume 3, Issue 1, pp.-358-364.

Introduction

Automatic Speech Recognition has been an active topic of research for the past four decades. The main objective of the automatic speech recognition task is to convert a speech segment into an interpretable text message without the need of human intervention. Many different algorithms and schemes based on different mathematical paradigms have been proposed in an attempt to improve recognition rates. Since the problem of speech recognition is complex, under certain circumstances, recognition rates are far from optimal. In addition other constraints such as computational complexity and real-time constraints come into play in the design and implementation of a working product. Computer hardware and software have significantly improved in terms of speed, memory, cost and availability, which have enabled the use of more sophisticated and computationally demanding algorithms to be implemented even on low-power low-cost handheld electronic devices. However, we prefer algorithms with low computational and memory requirements since they can be implemented easily and at lower cost. Due to improvements both in algorithms and in hardware, automatic speech recognition has become more affordable and available. Automatic speech recognition is still an open topic of research, where improvement and changes are constantly made in a hope for better recognition rates(J.C. Junqua and J.P. Haton)[1].

Automatic speech recognition (ASR) attempts to map from a speech signal to the corresponding sequence of Words it represents. To perform this, a series of acoustic features are extracted from the speech signal, and then pattern recognition algorithms are used. Thus, the choice of acoustic features is critical for the system performance. If the feature vectors do not represent the

underlying content of the speech, the system will perform poorly regardless of the algorithms applied. This task is not easy and has been the subject of much research over the past few decades. The task is complex due to the inherent variability of the speech signal. The speech signal varies for a given word both between speakers and for multiple utterances by the same speaker. Accent will differ between speakers. Changes in the physiology of the organs of speech production will produce variability in the speech waveform. For instance, a difference in height or gender will have an impact upon the shape of the spectral envelope produced. The speech signal will also vary considerably according to emphasis or stress on words. Environmental or recording differences also change the signal. Although humans listeners can cope well with these variations, the performance of state of the art ASR systems is still below that achieved by humans (H.G. Hirsh and D. Pearce) [2].

As the performance of ASR systems has advanced, the domains to which they have been applied have expanded. The first speech recognition systems were based on isolated word or letter recognition on very limited vocabularies of up to ten symbols and were typically speaker dependent. The next step was to develop medium vocabulary systems for continuous speech, such as the Resource Management (RM) task, with a vocabulary of approximately a thousand words. Next, large vocabulary systems on read or broadcast speech with an unlimited scope were considered. Recognition systems on these tasks would use large vocabularies of up to 65,000 words, although it is not possible to guarantee that all observed words will be in the vocabulary. An example of a full vocabulary task would be the Wall Street Journal task (WSJ) where passages were read from the Wall Street Journal. Current state of the art systems have been applied recognizing conversational or spontaneous speech in noisy and limited bandwidth domains. An example of such a task would be the Switchboard corpus. The most common approach to the problem of classifying speech signals is the use of hidden markov model. Before delving into the worlds of phonology, we present an overview of automatic speech recognition and give insight to some commonly used techniques that attempt to solve this formidable task.

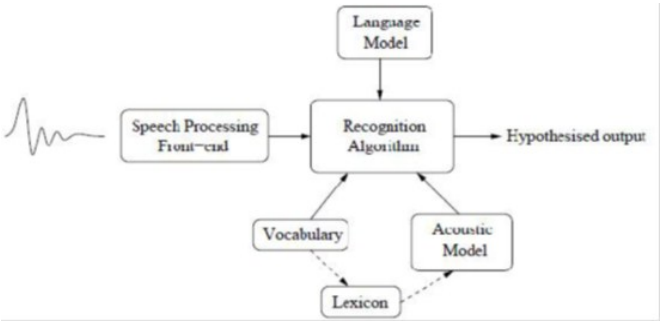

Fig. 1- General Speech Recognition Systems

A speaker recognition system mainly consists of two main modules, speaker specific feature extractor as a front end followed by a speaker modeling technique for generalized representation of extracted features as defined by (S. Saha and D. Bobbert and M. Wolska) [3, 4]. Since long time MFCC is considered as a reliable front end for a speaker recognition application because it has coefficients that represents audio based on perception mentioned by (K. Fujita et al and D. OShaughnessy) [5, 6]. In MFCC the frequency bands are positioned logarithmically (on the Melscale) which approximates the human auditory systems response more closely than the linear spaced frequency bands of FFT or DCT. This allows for better processing of data. Fig.1 shows the speaker recognition system used in this investigation. Accuracy of automatic speaker recognition is known to degrade severely when there is acoustic mismatch between the training and testing material which is clearly defined by (Renals S. et. al and B.H. Juang and L.R. Rabiner) [7,8].

Mel Frequency Cepstral Coefficients

Cepstral coefficients play an important part in speech theory and in automatic speech recognition in particular due to their ability to compactly represent relevant information that is contained in a short time sample of a continuous speech signal (N. Morgan and Bourlard) [9]. The definition for real Cepstral coefficients is given by the following equation-

Cepstrum (x) = IDFT (log (DFT(x)) (1.1)

We also note that

Cepstrum (x*y) = Cepstrum (x) + Cepstrum (y) (1.2)

Equation 1.2 can be easily derived from 1.1 and is useful in case we model the speech signal as a result of an excitation convolved with an impulse response of the vocal tract filter. DFT is the Discrete Fourier Transform often implemented by the Fast Fourier Transform algorithm. The Mel Frequency Cepstral Coefficients (MFCCs) are obtained by converting the result of the log- absolute value frequency spectrum to a Mel perceptually-based spectrum and taking an inverse discrete cosine transform of the result. Using Cepstral terminology we regard the Mel mapping to be a rectangular low frequency filter followed by a discrete cosine transform. The result is a smoothed cepstrum which can be further sampled to a specific number of coefficients. Qfrequency is a cepstrum value ('cepstrum frequency value') while a lifter is a weighted cepstrum or in other words a filter for the cepstrum coefficients.

$$\text{MFCC}= \sum_{k=1}^{13} X_k \cos\left[\frac{\left(k-\frac{1}{2}\right)\pi}{13}\right]$$

i=1,2.....M

M is the number of Cepstral Coefficients and Σ n represents the log energy output of the k^{th} Mel filter. The triangular lifters are linearly spaced up to 1000 Hz and logarithmically spaced afterwards up to 4000 Hz. The hidden assumption is that more important speech information is encapsulated in the low frequency band of 0 - 1000 Hz while the higher 1000-4000 Hz band contains less information per Hz. The triangular lifters can be regarded as a possibility function which serves as an upper bound to a symmetrical distribution where only the mean and variance are known. The possibility function entails all the possible distributions that might occur and is the coarsest upper bound we can obtain knowing only the mean and variance of a stochastic process. The human ear filters sound linearly for lower frequencies and logarithmically for higher frequencies. Partitioning the frequency range into two different spacing schemes that also resemble the Bark scale yields an efficient representation of the spectrum.

MFCC's are based on the known variation of the human ears criti-

cal bandwidths with frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. The characteristics are expressed on the mel-frequency scale, which is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. In addition, rather than the speech waveforms themselves, MFCC's are shown to be less susceptible to the above mentioned variation of the speakers voice and surrounding environment. The basic concept of a mel-frequency cepstral coefficient processor is described below.

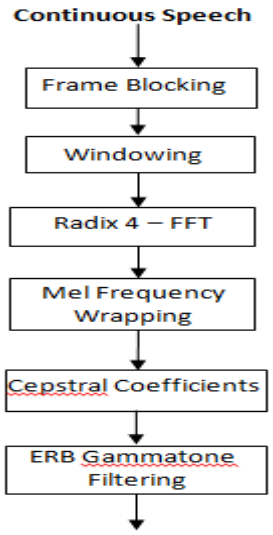

Fig. 2- Block diagram of MFCC processor with ERB Gammatone Filtering

ERB Gammatone Filters

After the development of third-octave filter banks, psycho acousticians performed further studies to obtain more accurate estimates of the auditory filter bandwidths. Most recently, they arrived at a formula they use to refer to Equivalent Rectangular Bandwidth (ERB). While a formula to convert frequency values into ERB-based frequencies is the bandwidth of an ERB filter centered at a given frequency fc is

$BW_{ERB} = 24.7\ (0.00437 f_c + 1)$

It is important to note that the formula above converts a frequency (in Hz) to a bandwidth (also in Hz). To convert a frequency in Hz to a frequency in units of ERB-bands, the formula should be used, namely

$ERB_{rate} = 21.4 \log\ (0.00437 f_c + 1)$

The bandwidths of the filters are set by a critical band function and so filter bandwidth increases with center frequency. If the energy at the output of each filter is calculated at a given point in time, and the values are plotted as a function of filter center frequency, the result is essentially the excitation pattern described by Moore and Glasberg (1983) [10]. The gammatone auditory filter can be described by its impulse response:

$$\gamma tone(t) = at^{n-1}\, e^{-2n\nu t} \cos(2\pi f_c t + \emptyset)\,(t > 0) \quad (1)$$

This function was introduced by Aertsen and Johannesma (1980) [11] and used by de Boer and de Jongh (1978) [12] to characterize "recover" data from cats. The primary parameters of the filter are b and n. b largely determines the duration of the impulse response; n is the order of the filter and it largely determines the slope of the skirts of the filter. When the order of the filter is in the range 3-5, the shape of the magnitude characteristic of the gammatone filter is very similar to that of the roex(p) filter commonly used to represent the magnitude characteristic of the human auditory filter (Patterson and Moore, 1986) [13].

Glasberg and Moore (1990) [14] have summarized human data on the equivalent rectangular bandwidth (ERB) of the auditory filter with the function:

ERB=24.7+0.108*fc (2)

Together, equations (1) and (2) define a gammatone auditory filterbank if one includes the common assumption that the filter center frequencies are distributed across frequency in proportion to their bandwidth. When the order of the filter is 4, b is 1.018 ERB. The 3-dB bandwidth of the gammatone filter is 0.887 times the ERB.

ERB GAMMATONE Filtering of MFCC's

This function computes the filter coefficients for a bank of Gammatone filters. These filters were defined by Patterson and Holdworth for simulating the cochlea. The result is returned as an array of filter coefficients. Each row of the filter arrays contains the coefficients for four second order filters. The transfer function for these four filters shares the same denominator (poles) but have different numerators (zeros). All of these coefficients are assembled into one vector that the ERB FilterBank function can take apart to implement the filter.

The filter bank contains numChannels channels that extend from half the sampling rate (fs) to lowFreq. The ERB filter function computes four separate second order filters. This avoids a problem with round off errors in cases with very small characteristic frequencies (<100Hz) and large sample rates (>44kHz). The problem is caused by roundoff error when a number of poles are combined, all very close to the unit circle. Small errors in the eighth-order coefficient are magnified when the eighth root is taken to give the pole location. These small errors lead to poles outside the unit circle and instability.

The robustness of PLP after ERB GAMMATONE filtering shows improved performance. Because ERB GAMMATONE is based on human auditory perception, this technology was taken to MFCC in recent years. In fact that human perception tends to tract the relative value of input rather than to its absolute values is very obvious in vision. Similarly, we can take knowledge of this fact in human auditory perception. Some circumstantial evidence indicates that there is a preference for sounds with a certain rate of change too. The ERB GAMMATONE filtering technique suppresses the spectral components that change more slowly or quickly than typical range of change of speech, and enhance the dynamic parts of noisy speech. The low cut-off frequency of the filter determines the fastest spectral change of the log spectrum, which is ignored in the output, whereas the high cut-off frequency determines the fastest spectral change that is preserved in the output parameters. The high-pass portion of the equivalent bandpass filter is expected to alleviate the effect of convolution noise introduced in the channel. The lowpass filtering helps to smooth some of the fast frame-to frame spectral changes present in the short term spectral estimate due to analysis artifacts. Because Mel frequency domain

also is nonlinear frequency domain, we can filter the cepstral domain with ERB GAMMATONE filtering technique, in other word, we can append a filtering processing after DCT. And the DCT essentially is a linear transformation, it is not distinct between before the DCT and after, that means it is equivalent to filtering in the cepstral domain.

We found that recognition accuracy using MFCC features was disappointing however, the insertion errors were very high. In order to decrease the insertion errors, and filtering features signals in the time domain, we use the ERB GAMMATONE filtering technique proposed after MFCC.

Experimental Results

The feature extraction and classification algorithm was implemented using Matlab with TIMIT databases. In Matlab, we make an audio function folder for root directory for accessing the matlab files and functions. The various functions created and scripted in matlab were used to extract the features from the wav files of TIMIT database and then classify using various classifiers. MFCC function is used to create Cepstral coefficients. Finally, the Cepstral Coefficients are classified according to various classification algorithms. In this study, we have use Hidden Markov Model as classifiers using HTK tools. The following results show the whole implementation.

This phase includes converting the speech waveform into a parametric representation with a considerably low information rate for further analysis and processing. This phase is often referred to as the signal processing front end. The speech signal can be described as a slowly timed varying signal, or quasistationary. A sample of speech from the well known speech database TIMIT, in this case from a version of TIMIT with noise added and a sample rate of 8000 Hz, can be seen below.

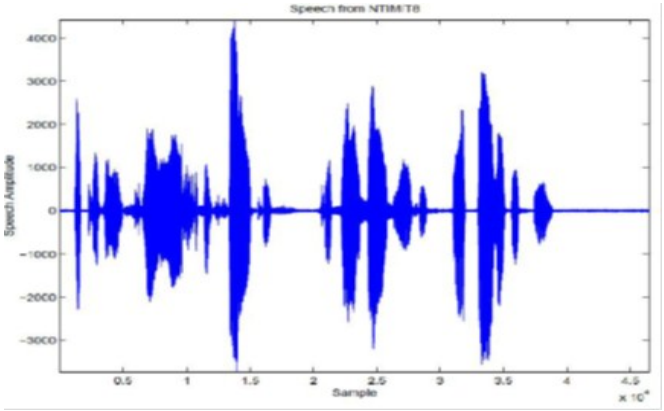

Fig. 3- Speech data from TIMIT, Fs = 16000Hz, 16-bits, telephone noise added

Frame Blocking

The first step of the feature extraction is to frame the speech into frames of approximately 30 msec (30 msec at Fs = 16000Hz gives 312 samples). To be able to extract as much features as possible from a speech sample, the technique of overlapping frames is used. The speech is blocked into frames of N samples (N = 312 in our case). With a overlapping of 50% one will get M number of frames out of a speech sample consisting of S samples:

Windowing

The next step in the processing is to window each individual frame so as to minimize the signal discontinuity at the beginning and end of each frame. The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame. A typical window utilized for speaker verification is the Hamming window.

Fast Fourier Transform (FFT)

The next step is to apply a Fourier Transform on the windowed speech frame. A Radix-4 Fast Fourier Transform is utilized, converting each frame from the time domain into the frequency domain. The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT). To get a better display of the Fourier Transform, the process of zero padding is applied. It is important to note that zero padding does not provide any additional information about the spectrum Y(w) of the sequence {x(n)}.

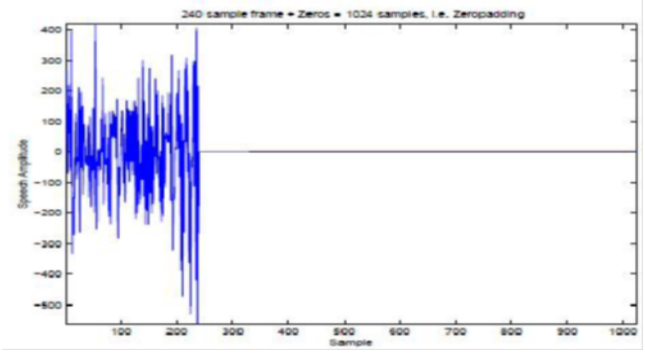

Fig 4- Windowing of audio sample

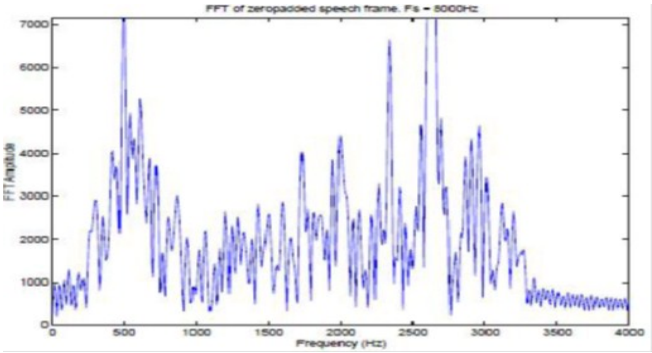

Fig. 5- Converting the signal into frequency amplitude domain using Fast Fourier Discrete Transform

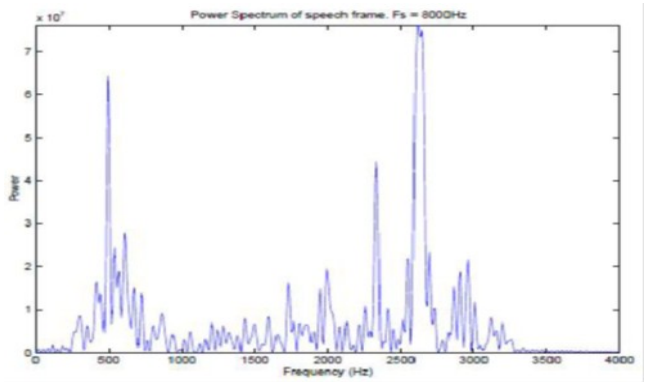

Fig. 6- Power Spectrum of SpeechFrame

Mel-Frequency Wrapping

As mentioned above, studies have been conducted that show that the human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the Mel scale. The Mel-frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 mels. Our approach to simulate the easier way of extracting the power from the speech is to apply a filterbank to the Power Spectrum. This filterbank is uniformly spaced on the mel scale, has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval. The number of Mel spectrum coefficients, K, is typically chosen as 13, but will vary a little depending on the sampling frequency. To be observed is that we are applying these filters in the frequency domain; therefore we simply multiply those triangle-shape windows in figure 7 on the Power Spectrum.

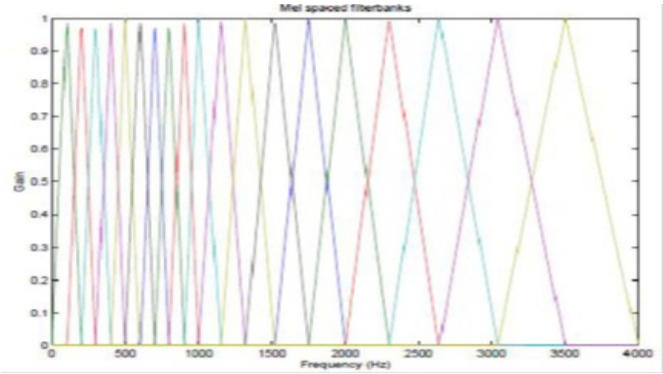

Fig. 7- Mel Frequency Filterbanks

Cepstral Coefficients

The next step is to convert the log Mel spectrum back to time. The result is called the Mel frequency cepstral coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the Mel spectrum coefficients are real numbers, we can convert them to the time domain using the Discrete Cosine Transform (DCT)-

$$\text{MFCC}= \sum_{k=1}^{13} X_k \cos\left[\frac{\left(k-\frac{1}{2}\right)\pi}{13}\right]$$

i=1,2……M

The filter bank is constructed using 13 linearly-spaced filters (133.33Hz between center frequencies,) followed by 27 log-spaced filters (separated by a factor of 1.0711703 in frequency.) Each filter is constructed by combining the amplitude of FFT bin as shown in the figure 8.

The outputs from this routine implemented in Matlab are the MFCC coefficients and several optional intermediate results and inverse results. reqresp the detailed fft magnitude used in MFCC calculation, 256 rows. fb the mel-scale filter bank output, 40 rows. Here is the result of calculating the cepstral coefficients of the 'A huge tapestry hung in her hallway' utterance from the TIMIT database (TRAIN/DR5/FCDR1/SX106/ SX106.ADC) spoken by 7 speakers. The utterance is 50189 samples long at 16kHz, and all pictures are sampled at 100Hz and there are 312 frames. Note, the top row of the mfcc-cepstrum, ceps is known as C_0 and is a function of the power in the signal. Since the waveform in our work is normalized to be between -1 and 1, the C_0 coefficients are all negative. The other coefficients, C_1-C_{12}, are generally zero-mean.

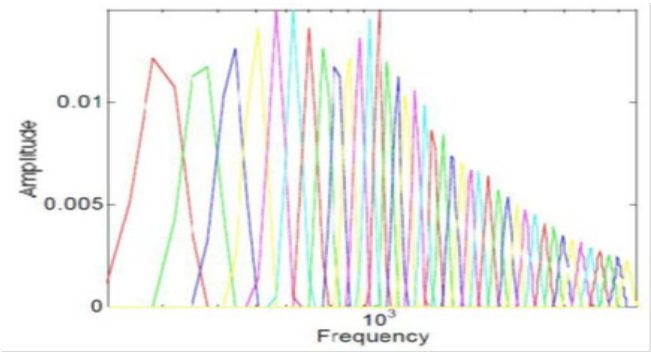

Fig. 8- Mel Frequency Cepstrum Coefficients Representation

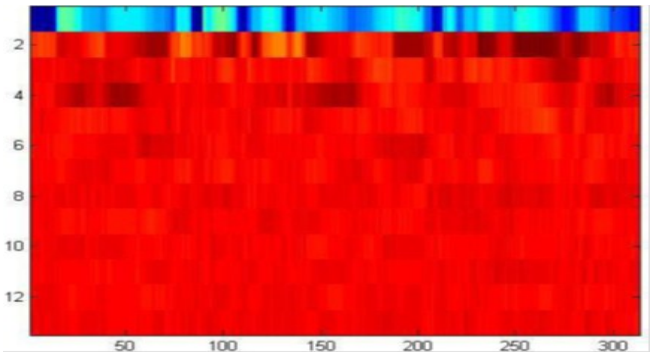

Fig. 9- Spectrogram of sample audio signal with power spectrum

After combining several FFT channels into a single Mel-scale channel, the result is the filter bank output. This is shown below (the fb output of the mfcc command includes the log10 calculation.).

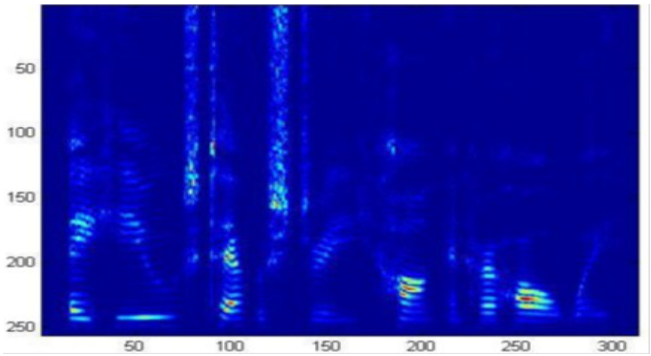

Fig. 10- MFCC's histrogram

The ERB GAMMATONEOUT function implements the ERB GAMMATONE (Relative Spectra) algorithm. The ERB GAMMATONE algorithm is a common piece of a speech-recognition system's front-end processing. It originally was designed to model adapta-

tion processes in the auditory system, and to correct for environmental effects. Broadly speaking, it filters out the very low-frequency temporal components (below 1Hz) which are often due to a changing auditory environment or microphone. High frequency temporal components, above 13 Hz, are also removed since they represent changes that are faster than the speech articulators can move.

The first input to this routine in Matlab is an array of spectral data, as produced by the MFCC routines. Each row contains one "channel" of data; each column is one time slice. The f*s* parameter specifies the sampling rate, 100Hz in many speech recognition systems. The original ERB GAMMATONE filter is defined only for a frame rate of 100Hz. This code is equal to the original at 100Hz, but scales to other frame rates. Here the ERB GAMMATONE filter is approximated by a simple fourth order Butterworth bandpass filter.

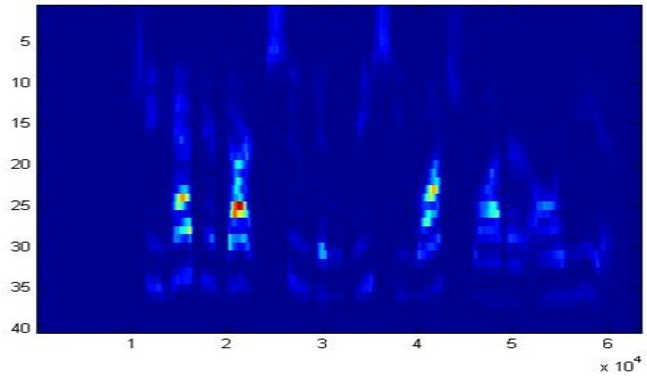

Fig. 11- MFCC's after ERB Gammatone Filtering with more sharpen spectrum images

Feature Matching

In our experiments, we take voice sets of TIMIT databases. The various functions of applying feature extraction, filtering and classification is done in Matlab. For each voice in the test sets, we added Gaussian white noise to them, SNR level is from- 5db to 20db, interval 5db, to get the clean voices. The size of each frame is 30ms, and frame shift is 15ms. We pre emphasize each frame after enframing the speech, pre-emphasis formula is:

$H(z) = 1 - \mu z^{-1}$

Where pre-emphasis factor $\mu = 0.9372$, because the frame length is 30ms and the sampling frequency is 16 kHz, we can use the 512-point FFT to obtain speech power spectrum. After Mel subband filtering, we obtain the cepstral coefficients of each frame and use the ERB GAMMATONE filtering technique to process the cepstral coefficients, then we get the features coefficients which we need. We select the first 13 coefficients of feature vector which has been sorted in a descending order according to the variance, and we discard the other coefficients. What's more, the logarithmic energy of each frame is very important to reflect characteristic of voice, we append it as the supplement of feature vectors. At the same time, in order to obtain the dynamic characteristics of voice, we calculate the first and second order differential as a supplementary factor in the end of feature vector. Finally, the feature vector of each frame consists of 39 dimensional feature parameters. The recognition model is built with the non-jump from left to right continuous Hidden Markov Model (HMM). Each HMM has five states, the probability density function of the values observed under each state is the mixed Gaussian probability density function, and the transfer matrix is diagonal. Model is trained and tested by HTK.

In Table 1, we compare the word recognition correct, recognition accuracy and the sentence recognition correct under different SNR level, they can be indicated with Corr, Acc and Correct respectively. ERB means ERB GAMMATONE, "+" means combing two methods. The most of the experiment results show us that the whole robustness the ERB GAMMATONE+MFCC method offers much higher than classical PLP and classical MFCC method. Although compared to MFCC method, its average value of Corr slightly increased while its average value of Acc and Correct, it is also much higher than that of the MFCC and PLP method. The improved performance is obvious as compared to some classical feature extraction, and especially under slight high level SNR (>10db), we can get more robust feature for ASR. [*See Table 1*]

Conclusions

There are several motivations for using spectral-peak or formant features. Formants are considered to be representative of the underlying phonetic content of speech. They are also believed to be relatively robust to the presence of noise, and useful in low-bandwidth applications. Additionally, it has been hypothesized that formants or spectral peak positions can be easily adapted to different speakers. However, the extraction of robust and reliable formant estimates is a nontrivial task. Recently, there has been increased interest in other methods for estimating spectral peaks, for example, using the HMM or gravity centroid features.

We use a combination ERB GAMMATONE filtering technique for MFCC feature extraction in this study. One way is to replace Mel filters with ERB Gammatone filters, and another is to append a ERB GAMMATONE filtering in time domain after transformation. By two methods we obtain more robust feature, and we also refer that the speech enhancement can help improving robustness. Finally, because ERB Gammatone filtering used in this model is based on a linear assumption, and the voice is only similar to a linear model, in fact it is still nonlinear, we believe that the nonlinear filtering is potential, and the nonlinear filtering will become our direction in future.

References

[1] Junqua J.C. and Haton J.P. (1996) *Robustness in utomatic Speech Recognition., Norwell, Massachusetts, USA*.

[2] Hirsh H.G. and Pearce D. (2000) *The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions., ISCA ITRW ASR, Paris, France*.

[3] Saha S. (1995) *Biomedical Engineering Conference, Proceedings of the Fourteenth Southern*. 134-137.

[4] Bobbert D., Wolska M. (2007) *11th Workshop on the Semantics and Pragmatics of Dialogue*, 159-160.Trento, Italy.

[5] Fujita K. et al. (2003) *IEEE Trans. on Consumer Electronics*, 49, (3), 765-769.

[6] Shaughnessy D.O. (1987) *Speech Communication*.

[7] Renals S. et. al. (1994) *IEEE Tran. on Speech and Audio Processing*, 2(1), Part 11, 161-174.

[8] Juang B.H., Rabiner L.R. (1992) *Neural Networks for Signal*

Processing II., *IEEE-SP Workshop*, 214-222.
[9] Morgan N., Bourlard H.A. (1995) *IEEE*. 83(5), 742-772.
[10] Brian C., Moore J. and Brian R. Glasberg (1983) *British Journal of Audiology*, 17(1), 31-48.
[11] Aertsen A.M.H.J., Johannesma P.I.M. (1980) *Biological Cybernetics, Springer Berlin / Heidelberg*, 38(4), 235-248.
[12] Boer and de Jongh (1978) *J. Acoust. Soc. Am*. 63(1), 115-135.
[13] Patterson and Moore (1986) *The Journal of Organic Chemistry,* 51 (26), 5300-5306.
[14] Glasberg and Moore (1990) *Hearing Research*, 47(1-2), 103-138.

Table 1- Classification Results

Features	SNR	Clean	20dB	15dB	10dB	5Db	0dB	-5dB	Average
PLP	Corr	85.43	85.43	81.64	37.86	27.12	21.03	9.92	49.77
	Accurate	-3.6	-3.6	-0.5	26.82	26.33	21.03	9.92	10.91
	Correct	0	0	0	0	0	0	0	0.36
MFCC	Corr	98.96	82.12	75.35	59.49	22.15	8.55	8.55	50.73
	Accurate	65.28	-7.14	-7.4	2.8	13.42	8.55	8.55	12..00
	Correct	32.5	0	0	0	0	0	0	4.64
MFCC + ERB	Corr	100	98.26	97.56	84.12	`52.78	21.45	10.32	66.35
	Accurate	83.22	92.51	58.46	1.23	-25.62	5.23	10.32	32.19
	Correct	59.69	70.56	41	5	0	0	0	25.17