SPEECH RECOGNITION USING HMM WITH MFCC- AN ANALYSIS USING FREQUENCY SPECRAL DECOMPOSION TECHNIQUE

Ibrahim Patel[1] Dr. Y. Srinivas Rao[2]

[1]Assoc. Prof., Department of BME, Padmasri.Dr.B.V.Raju Institute of Technology.Narsapur s

Ptlibrahim@gmail.com

[2] Assoc. Prof., Department of Instrument Technology, Andhra University, Vizag, A.P.
srinniwasarau@gmail.com

ABSTRACT

This paper presents an approach to the recognition of speech signal using frequency spectral information with Mel frequency for the improvement of speech feature representation in a HMM based recognition approach. A frequency spectral information is incorporated to the conventional Mel spectrum base speech recognition approach. The Mel frequency approach exploits the frequency observation for speech signal in a given resolution which results in resolution feature overlapping resulting in recognition limit. Resolution decomposition with separating frequency is mapping approach for a HMM based speech recognition system. The Simulation results show an improvement in the quality metrics of speech recognition with respect to computational time, learning accuracy for a speech recognition system.

KEYWORDS

Speech-recognition, Mel-frequencies, DCT, frequency decomposition, Mapping Approach, HMM, MFCC.

1. INTRODUCTION

Speech recognition is a process used to recognize speech uttered by a speaker and has been in the field of research for more than five decades since 1950s [1]. Voice communication is the most effective mode of communication used by humans. Speech recognition is an important and emerging technology with great potential. The significance of speech recognition lies in its simplicity. This simplicity together with the ease of operating a device using speech has lots of advantages. It can be used in many applications like, security devices, household appliances, cellular phones, ATM machines and computers.

With the advancement of automated system the complexity for integration & recognition problem is increasing. The problem is found more complex when processing on randomly varying analog signals such as speech signals. Although various methods are proposed for efficient extraction of speech parameter for recognition, the MFCC method with advanced recognition method such as HMM is more dominant used. This system found to be more accurate under low varying environment but fails to recognition speech under highly varying environment. This need to the development of an efficient recognition system which can provide is efficient varying system.

Research and development on speaker recognition method and technique has been undertaken for well over four decade and it continues to be an active area. Approaches have spanned from human auditory [2] and spectrogram comparisons [2], to simple template matching, to dynamic time-warping approaches, to more modern statistical pattern recognition [3], such as neural networks and Hidden Markov Model (HMM's) [4].

DOI : 10.5121/sipij.2010.1209

It is observed that, to extract and recognize different information from a speech signal at variable environment, many algorithms for efficient speech recognition is proposed in past. Masakiyo Fujimoto and Yasuo Ariki in their paper "Robust Speech Recognition in Additive and channel noise environments using GMM and EM Algorithm" [5] evaluate the speech recognition in real driving car environments by using a GMM based speech estimation method [6] and an EM algorithm based channel noise estimation method.

A Gaussian mixture model (GMM) based speech estimation method proposed in J.C.Segura et al [6] estimates the expectation of the mismatch factor between clean speech and noisy speech at each frame by using GMM of clean speech and mean vector of noise. This approach shows a significant improvement in recognition accuracy. However, the Segura's method considered only the additive noise environments and it did not consider about the channel noise problem such as an acoustic transfer function, a microphone characteristic etc

A Parallel model combination (PMC) method [7] has been proposed by M.J.F Gales, and S.J.Young adapts the speech recognition system to any kinds of noises. However, PMC has a problem, of taking huge quantity of computation to recognize the speech signal. Another method for speech recognition called "spectral subtraction" (SS) is also proposed as a conventional noise reduction method [3]. However, using spectral subtraction method degrades the recognition rate due to spectral distortion caused by over or under subtraction. Additionally, spectral subtraction method does not consider the time varying property of noise spectra, because it estimates the noise spectra as mean spectra within the time section assumed to be noise.

Hidden Markov Model (HMM) [4] is a natural and highly robust statistical methodology for automatic speech recognition. It was tested and proved considerably in a wide range of applications. The model parameters of the HMM are essence in describing the behavior of the utterance of the speech segments. Many successful heuristic algorithms are developed to optimize the model parameters in order to best describe the trained observation sequences. The objective of this paper is to develop an efficient speech recognition algorithm with the existing system following HMM algorithm. The paper integrates the frequency isolation concept celled as sub band decomposition to the existing MFCC approach for extraction of speech feature. The additional feature concept of provides the information of varying speech coefficient at multiple band level this feature could enhancement the recognition approach then the existing one.

2. Hidden Markov Modeling

An HMM is a stochastic finite state automation defined by the parameter $\lambda = (A, p, B)$, where A is a state transition probability, p is the initial state probability and B is the emission probability density function of each state, defined by a finite multivariate Gaussian mixture as shown in figure below. Each model can be used to compute the probability of observing a discrete input sequence O = O1, . . . , OT, P(O|λ) to find the corresponding state sequence that maximizes the probability of the input sequence, P(Q|O, λ), and to induce the model that maximizes the probability of a given sequence P(O|λ) > P(O|λ).

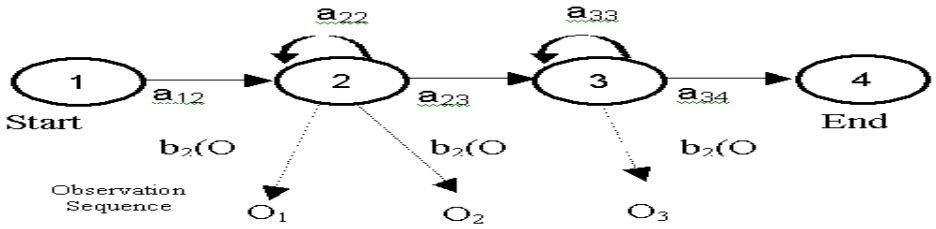


Figure.1 A typical left-right HMM (aij is the station transition probability from state i to state j; OT is the observation vector at time T and bi(OT) is the probability that OT is generated by state i).Given the form of the HMM, there are three basic problems to be solved.

Problem 1: calculation of probability P (O|λ) for a given observation sequence.

For a given observation sequence O =O1, O2, . . . , OT and a model λ, the calculation for probability of occurrence for a given feature is one of the difficult and time consuming task. For a given HMM for speech recognition the probability is given by,

$$P_{ij}(\mathbf{O}_i|\lambda_j) = \log \sum_{l=1}^{N} \alpha_T(l),$$

Here αT is the forward component of the Forward-Backward procedure.
Problem 2: For a given observation sequence O = O1O2, . . . , OT and a model λ, the selection to corresponding state sequence Q = q1q2, . . . , qT which best "explains" the observation, is one more problem faced in speech recognition for HMM.
This problem is solved by using Viterbi algorithm.
Problem 3: Parameters Optimization.
For a given model λ, the HMM parameters {A, b, ð} should be such chosen so as to maximize the probability P(Oλ) called a learning procedure. Considering all the temporal feature values as continuous using the continuous density HMM can optimize the learning process.

2.1 Clustering

Regarding dissimilarity measure selection the distance function d (i, j) is given by

$$d[(i,j)|(i-1,j-1)] = \sum_{i=1}^{F} |O_1(i) - O_2(i)|$$

with the assumption that the similar speech forming one cluster corresponds to one HMM model. In this project work a complete link hierarchical clustering technique is been used. The first step is to compute a measure matrix 'W' for HMM dissimilarity measurement. Once the measure matrix 'W' is computed the clustering technique is applied to obtain the K clusters. The algorithm produces a sequence of clustering of decreasing number of clusters at each step. The clustering produced at each step results from the previous one by merging two clusters into one. Finally, isolating the sequences into K groups we fit each HMM to each cluster using all the observation sequences in cluster for HMM training.

3. Mel Spectrum Approach

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. By applying the procedure described above, for each speech frame of around 30msec with overlap, a set of mel-frequency cepstrum coefficients is computed. These are result of a cosine transform of the logarithm of the short-term power spectrum expressed on a mel-frequency scale. This set of coefficients is called an acoustic vector. Therefore each input utterance is transformed into a sequence of acoustic vectors. As these vectors are evaluated using a distinct filter spectrum the feature information obtained is limited to certain frequency resolution information only and needs to be improved. In the following section a frequency decomposition method incorporated with existing architecture is suggested for HMM training and recognition. This mel spectrum is used recognition information in conventional speech recognition system. The spectrum doesn't exploit the variations in fundamental resolution & hence is lower in accuracy to improve the accuracy of operation a spectral decomposition approach is respected as shown in figure (2).

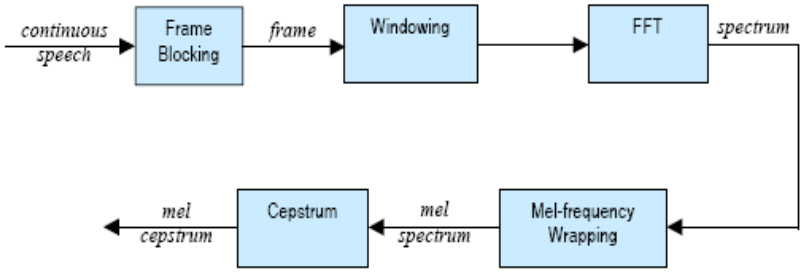


Figure (2): Speech process models

Therefore we can use the following approximate formula to compute the mels for a given frequency 'f' in Hz;

$$\textbf{mel (f) = 2595*log}_{10}\textbf{ (1 + f / 700)}$$

One approach to simulating the subjective spectrum is to use a filter bank, spaced uniformly on the mel scale where the filter bank has a triangular band pass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval. The modified spectrum of S(ω) thus consists of the output power of these filters when S(ω) is the input. The number of Mel spectrum coefficients, K, is typically chosen as 20.

Note that this filter bank is applied in the frequency domain; therefore it simply amounts to taking those triangle-shape windows in the Figure.3 on the spectrum. A useful way of thinking about this Mel-wrapping filter bank is to view each filter as a histogram bin (where bins have overlap) in the frequency domain.

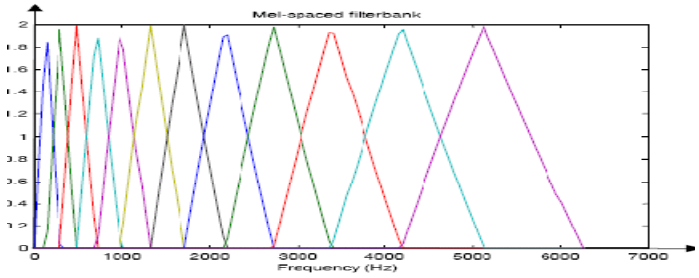


Figure. 3. An example of mel-spaced filterbank

The log Mel spectrum is converted back to time. The result is called the Mel frequency cepstrum 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 (and so their logarithm) are real numbers, we can convert them to the time domain using the Discrete Cosine Transform (DCT). Therefore if we denote those Mel power spectrum coefficients that are the result of the last step are we

$\tilde{S}_k, k = 1,2,...,K$, can calculate the MFCC's, $\tilde{c}_n$ as the first component,

$$\tilde{c}_n = \sum_{k=1}^{K} (\log \tilde{S}_k) \cos\left[n\left(k - \frac{1}{2} \right) \frac{\pi}{K} \right], n = 1,2,3...K$$

$\tilde{c}_0$ from the DCT since it represents the mean value of the input signal, which carried little speaker specific information. As shown in the figure. 3

The normalized cross-correlation based method is used for pitch estimation. The algorithm assumes a monophonic signal. The method follows the assumption that the signal has a periodicity corresponding to the fundamental frequency or pitch. Starting with a signal that is assumed to be periodic, or more precisely, quasi-periodic, it follows that $s(t) \approx \alpha s(t + T)$ where T is the quasi-period of the signal and the scalar α accounts for amplitude variations. For training

HMM, twelve frequency spectral coefficients (Mel) using 10 millisecond Hamming windowed frames were extracted. The Mel scaled is defined by;

$$Mel(f) = 2595\log_{10}(1 + \frac{f}{700})$$

The employed filters were triangular and equally spaced along the entire Mel-scale. The pitch frequency was also computed in the same window. The zero order energy and Frequency zero were included with the above-mentioned features. In order to take the advantage of pitch and spectral dynamics, the velocity (delta) and acceleration parameters were also added to the feature space as shown in figure (4).

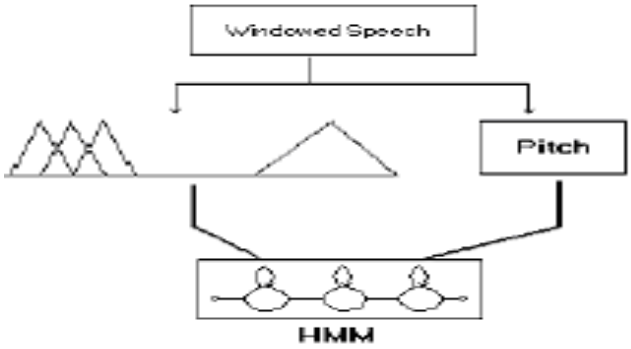


Figure. 4 The pitch frequency was also computed in the same window

4. Spectral Decomposition Approach

Filter bank can be regarded as wavelet transform in multi resolution band. Wavelet transform of a signal is passing the signal through this filter bank. The outputs of the different filter stages are the wavelet and scaling function transform coefficients. Analyzing a signal by passing it through a filter bank is not a new idea and has been around for many years under the name sub band coding. It is used for instance in computer vision applications.

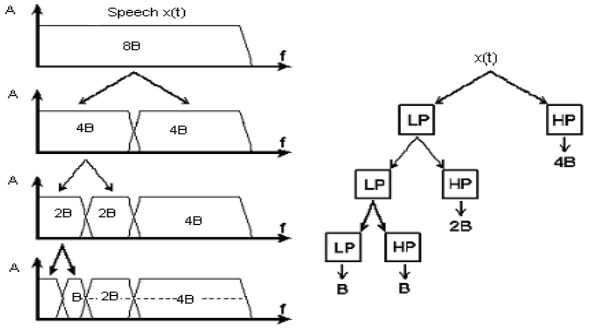


Figure: 5. splitting the signal spectrum with an iterated filter bank.

The filter bank needed in sub band coding can be built in several ways. One way is to build many band pass filters to split the spectrum into frequency bands. The advantage is that the width of every band can be chosen freely, in such a way that the spectrum of the signal to analyze is covered in the places of interest the disadvantage is that it is necessary to design every filter separately and this can be a time consuming process. Another way is to split the signal spectrum in two equal parts, a low pass and a high-pass part. The high-pass part contains the smallest details importance that is to be considered

here. The low-pass part still contains some details and therefore it can be split again. And again, until desired number of bands are created. In this way an iterated filter bank is created.

Usually the number of bands is limited by for instance the amount of data or computation power available. The process of splitting the spectrum is graphically displayed in figure: 6. The spectral decomposition obtained coefficient could be observed as,

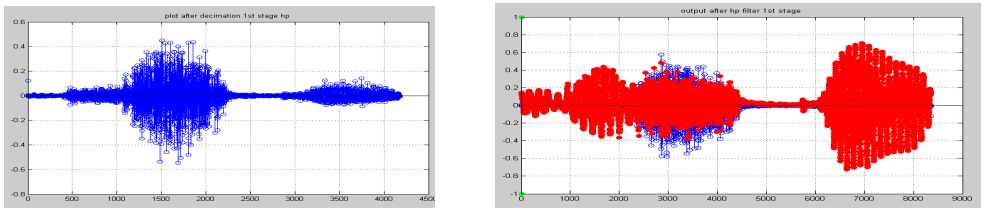


Figure: 6. Output after 1st stage decomposition for a given speech signal

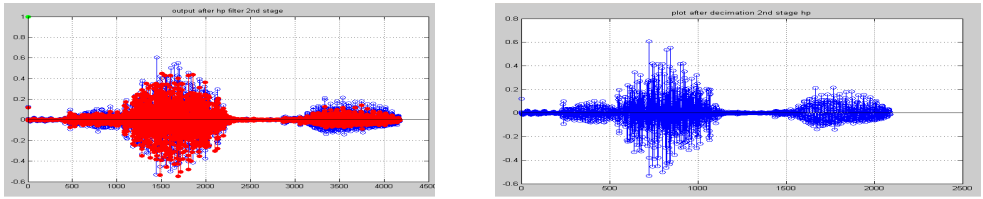


Figure: 7. Plot after 2nd stage decomposition

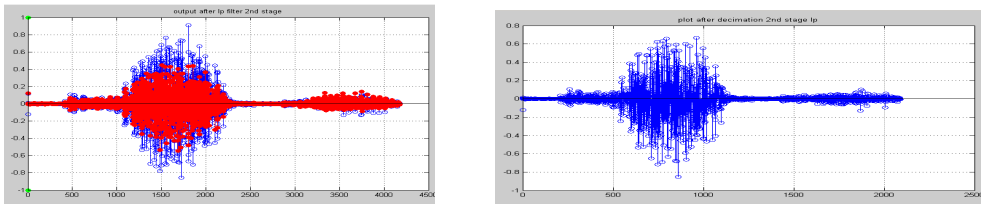


Figure:8. Output after 3rd Stage decomposition

The advantage of this scheme is that it is necessary to design only two filters; the disadvantage is that the signal spectrum coverage is fixed.

Looking at figure 5 it is observed that it is left with lower spectrum, after the repeated spectrum splitting is a series of band-pass bands with doubling bandwidth and one low-pass band. The first split gave a high-pass band and a low-pass band; in reality the high-pass band is a band-pass band due to the limited bandwidth of the signal. In other words, the same sub band analysis can be performed by feeding the signal into a bank of band-pass filters of which each filter has a band width twice as wide as its left neighbor and a low-pass filter. The wavelets give us the band-pass bands with doubling bandwidth and the scaling function provides with the low-pass band. From this it can be concluded that a wavelet transform is the same thing as a sub band coding scheme using a constant-Q filter bank. It can be summarized, as in implementation of the wavelet transform as an iterated filter bank, it is not necessary to specify the wavelets explicitly The actual lengths of the detail and approximation coefficient vectors are slightly more than half the length of the original signal. This has to do with the filtering process, which is implemented by convolving the signal with a filter. The spectral decomposition reveals the accuracy of individual resolution which was not explored in mel spectrum. This approach is developed with a mapping concept for speech recognition as outlined between evaluation of the suggested system for a

simulation of the purposed system is carried out on mat lab tool & the resolution obtained are as outlined below.

5. Mapping Approach

After the enrolment session, the acoustic vectors extracted from input speech of a speaker provide a set of training vectors. A set of 'L' training vectors from the frequency information is derived using well-known LBG algorithm [4]. The algorithm is formally implemented by the following recursive procedure. The LBG algorithm designs an M-vector codebook in stages. It starts first by designing a 1-vector codebook, then uses a splitting technique on the codewords to initialize the search for a 2-vector codebook, and continues the splitting process until the desired M-vector codebook is obtained. Figure 7 shows, in a flow diagram, the detailed steps of the LBG algorithm. "Cluster vectors" is the nearest-neighbor search procedure, which assigns each training vector to a cluster associated with the closest codeword. "Find centroids" is the centroid update procedure. "Compute D (distortion)" sums the distances of all training vectors in the nearest-neighbor search so as to determine whether the procedure has converged. A general operational flow diagram of the suggested LBG mapping approach is shown in fig.7

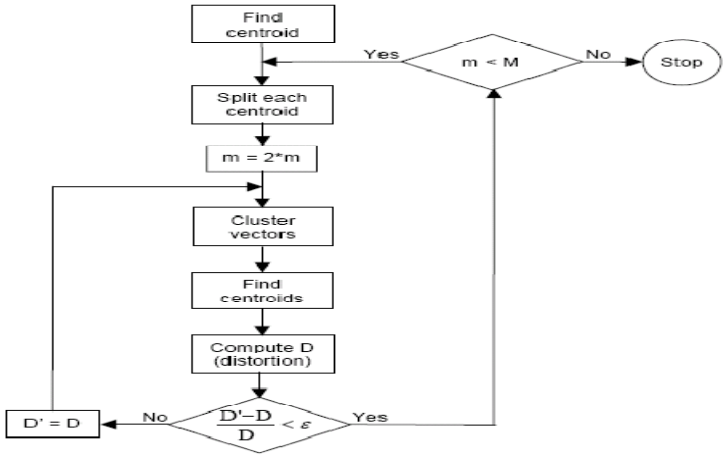
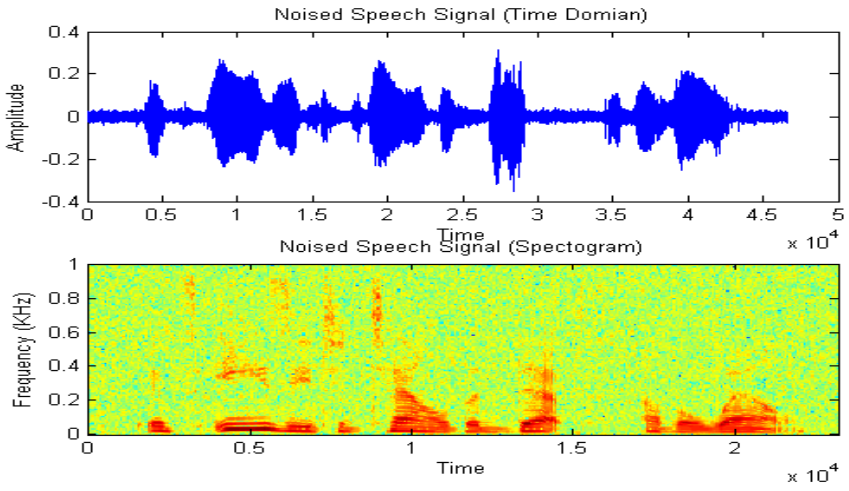


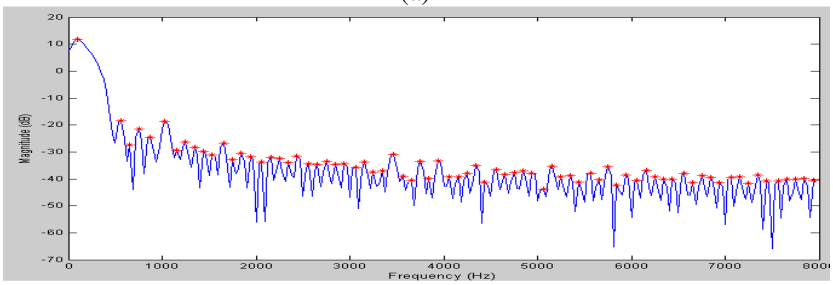
Figure: 9. Flow diagram of the LBG algorithm

6. Simulation Observation

For the training of HMM network for the recognition of speech a vocabulary consist of collection words are maintained. The vocabulary consists of words given as, "DISCRETE", "FOURIER", "TRANSFORM", "WISY", "EASY", "TELL", "FELL", "THE", "DEPTH", "WELL", "CELL", "FIVE", each word in the vocabulary is stored in correspondence to a feature define as a knowledge to each speech word during training of HMM network. The features are extracted on only voice sample for the corresponding word.
The test speech utterance: used for testing given as "its easy to tell the depth of a well", at 16KHz sampling.


Noised Speech Signal (Time Domian)
Amplitude
Time
x 10^4
Noised Speech Signal (Spectogram)
Frequency (KHz)
Time
x 10^4

(a)


Magnitude (dB)
Frequency (Hz)

(b)

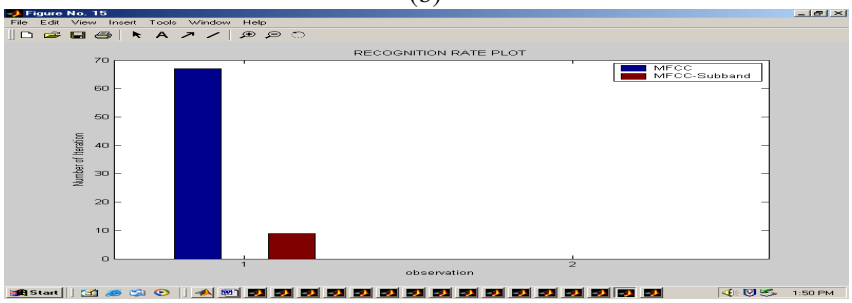


Figure No. 15
File Edit View Insert Tools Window Help
RECOGNITION RATE PLOT
MFCC
MFCC-Subband
Number of Iteration
observation
Start
1:50 PM

(c)

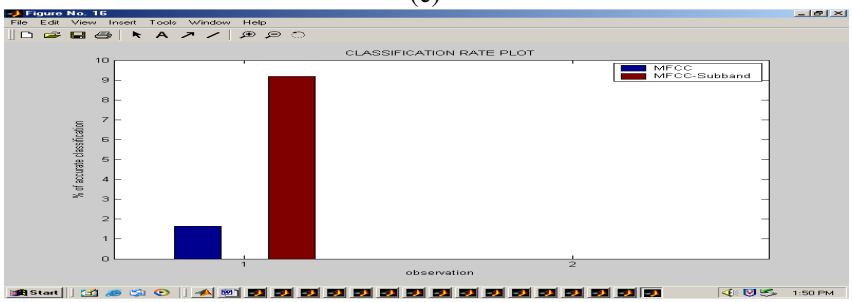
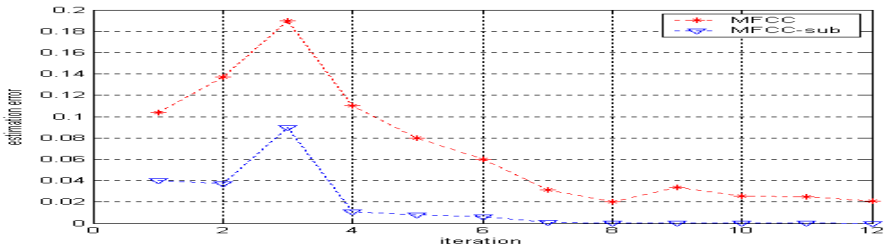
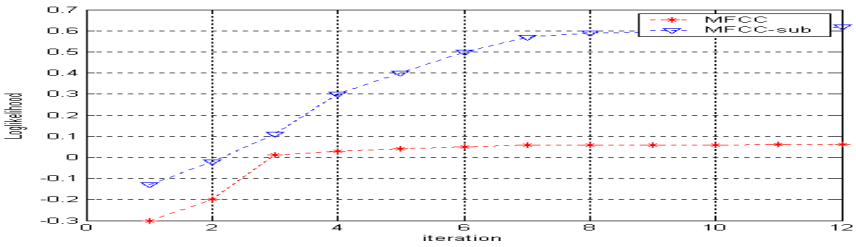


Figure No. 16
File Edit View Insert Tools Window Help
CLASSIFICATION RATE PLOT
MFCC
MFCC-Subband
% of accurate classification
observation
Start
1:50 PM

(d)



(e)



(f)

Figure : (10) (a) original speech signal and it's noise effect speech signal, (b) the energy peak points picked for training, (c) the recognition computation time for the MFCC based and the modified MFCC system, (d) the observed correct classified symbols for the two method, (e) the estimation error for the two methods with respect to. Iteration, (f) the likelihood variation with respect to iteration for the two methods.

7. Conclusion

A speech recognition system for robust to noise effect is developed. The MFCC conventional approach & extracting the feature of speech signal at lower frequency & is modified in this paper. An efficient speech recognition system with the integration of MFCC feature with frequency sub band decomposition using subband coding is proposed. The two features passed to the HMM network result in better recognition compared to existing MFCC method. From the observation made for the implemented system it is observed to have better efficiency for accurate classification & recognition compared to the existing system.

8. References

[1] Varga A.P, and Moore R.K.: Hidden Markov Model decomposition of speech and noise, Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, pp. 845-48, (1990)

[2] Allen, J.B.: How do humans process and recognize speech IEEE Trans. on Speech and Audio Processing, vol. 2, no. 4, pp.567--577 (1994.)

[3] Kim W. Kang, S. and Ko, H. : Spectral subtraction based on phonetic dependency and masking effects, IEEE Proc.- Vision, Image and Signal Processing, 147(5), pp 423--27 (2000)

[4] R.J. Elliott, L. Aggoun and J.B. : Moore Hidden Markov Models: Estimation and Control", Springer Verlag, (1995)

[5] Fujimoto, M.; Riki, Y.A.: Robust speech recognition in additive and channel noise environments using GMM and EM algorithm. Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference , , Vol1 17--21 May (2004)

[6] J .C.Segura, A.de la Torre, M.C.Benitez and A.M.Peinado: Model Based Compensation of the Additive Noise for Continuous Speech Recognition. Experiments Using AURORA II Database and Tasks, EuroSpeech'01,Vol.I, Page(s):I – 941--944
[7] M.J.F.Gales and S.J.Young : Robust Continuous Speech Recognition Using Parallel Model Combination, IEEE Trans. Speech and Audio Processing, Vol.4, No.5, pp.352--359(1996).
[8] Renals, S., Morgan, N., Bourlard, H., Cohen, M, and Franco, H. : Connectionist Probability Estimators in HMM Speech Recognition, IEEE Trans. on Speech and Audio Processing, vol. 2, no. 1, pp. 161--174, (1994.)
[9] J. Neto, C. Martins and L. Almeida: *Speaker-Adaptation in a Hybrid HMM-MLP Recognizer"*, in Proceedings ICASSP '96, Atlanta, Vol. 6, pp.3383--3386 (1996.)
[10] Sadaoki Furui : Digital speech processing , synthesis and recognition second edition

[11] Gajic, B.; Paliwal, Kuldip .K. "Robust speech recognition using features based on zero crossings with peak amplitudes" Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference Volume 1, 6-10
[15] Lockwood P. and Boudy J (1992), "Experiments with a non-linear Spectral Subtractor (NSS), HMM and the projection, for robust speech recognition in cars", *Speech Communication,* Vol. 11, Nos. 2-3, pp.215-228.
[16] Das **S.,** Bakis R., Nadas A., Nahamoo D., and Picheny M., (1 993), "Influence of background noise and microphone on the performance of the IBM TANGORA speech recognition system", *Proc. IEEE ICASSP* , *Australia,* April 1994, VI, pp. 21-23.
[17] Gong Y., (1995), "Speech recognition in noisy environments: A survey". *SpeechCommunication,* 16, pp.261-291
[18] Junqua J-C., Haton J-P., (1996), "Robustness in ASR: Fundamentals and Applications", *Kluwer Academic Publishers.*
[19 Tamura **S.,** (1 987), "An analysis of a noise reduction multi-layer neural network",
[20] Tamura **S.,** (1 990),"Improvements to the noise reduction neural network", *IEEE*
[21] Xie F. and Campernolle D., (1994),"A family of 'MLP' based non-linear spectral estimators for noise reduction", *IEEE ICASSP '94,* pp. 53-56.

Authors

1) **Ibrahim Patel** Working as Associate Professor BME department Vishnupur. Narsapur, Medak, (Dist), Andhra Pradesh India. He has received B.Tech in (E&CE), M.Tech Degree in Biomedical Instrumentation and currently pursuing Ph.D at Andhra University. He is having 16+ years of teaching and research experience and published 12 research papers in the International conference & journals and His main research interest includes Voice to sign language.

2) **.Y.Srinivasa Rao** received his Ph.D in Electrical Communication Engineering from Indian Institute of Science Bangalore in 1998. At present, he is an Associate Professor in Instrument Technology Department, AU College of Engineering, Andhra University, Visakhapatnam, AP, India. He is having 15+ years of teaching and research experience and published 38 research papers in the International journals and presented few of them in the International Conferences. He had received **"Emerald Literati Network 2008 Outstanding Paper Award (UK)", "Best Paper Award"** and **"Best Student paper Award"** in **ICSCI –**International Conference in **2007, "Vignan Pratibha Award – 2006"** from Andhra University and selected as **"Science Researcher"** for Asia – Pacific region in **2005** by UNESCO and Australian Expert group in Industry Studies (AEGIS) at the University of Western Sydney (UWS) and received **"Young Scientist Award"** from Department of Science and Technology, Govt. of India in **2002**. His present research focuses on nanotechnology and VLSI. He is a life member of ISTE and member of IMAPS (India).