AIVP ADVANCES IN IMAGE AND VIDEO PROCESSING

SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM

VOLUME 5, ISSUE 1
ISSN 2054-7412

Automatic Non-native Dialect and Accent Voice Detection of South Indian English

[1]Ibrahim Patel, [2]Raghavendra Kulkarni, and [3]Y Srinivasa Rao
[1]Dept. of ECE B.V. Raju Inst. of Tech., Narsapur Medak,(T. S) India;
[2]Dept. of ECE K.M.C.E&T, Devarkonda under JNT University, Hyderabad, T. S, India
[3]Department of Instrument Technology, Andhra University, Visakhapatnam, A. P, India
ptlibrahim@gmail.com; srinniwasarau@gmail.com; raghavendrakulkarni444@gmail.com

ABSTRACT

Speech recognition has achieved enormous improvements presently. However, robustness is still one of the big tribulations, e.g. performance of recognition fluctuates penetratingly depending on the speaker, particularly as the speaker has robust accent that is not coated in the training corpus. The speaker variability, such like gender, accent, age, speaking rate, and phone realizations, are vital problems in speech recognition. The mainly South Indian accent identification is a recent challenging problem closely related to other relatively recent fields of the multilinguality area like non native speech identification and language identification. This paper explains an automatic recognition system for English accents from 5 different South Indian State. The approach is based on a corresponding set of random nets with situation independent HMM units. The random topology was in addition substituted by pronunciation transcription constraints so as to integrate accent specific automatic word recognizers.

Keywords: - MFCC, HMM, Accent modeling, Speaker Identification, South Indian Accent, ASR

1 Introduction

Speech recognition is a difficult task and it is still an active research area. Automatic speech recognition works based on the premise that a person's speech exhibits characteristics that are unique to the speech. However this task has been challenged by the highly variant of input speech signals. The principle source of variance is the speech itself. Speech signals in training and testing sessions can be greatly different due to many facts such as non-native accents with time, health conditions (e.g. the speech has a cold), speaking rates, etc. There are also other factors, beyond speech variability, that present a challenge to speech recognition technology. Examples of these are acoustical noise and variations in recording environments (e.g. speech uses different telephone handsets).

Non-native accents can be also found in travelling and tourist centers or automatic international phone call services. Since English is probably the mostly used second language in the world, many speakers use it when addressing such services abroad, expecting some immediately feedback. If the listener is able to identify the accent, he may be able to find a suitable attendant knowing the first language of the specific client. For an automatic system, this means selecting an automatic speech recognizer for that first language or, if there is no such recognizer, another one adapted for the given English accent.

Non-native accent identification is a new challenging problem closely related to other new fields from the multilinguality area. One of the most important ones is Speech Identification (SI). In this field, there

DOI: 10.14738/aivp.51.2749
Publication Date: 08th March, 2017
URL: http://dx.doi.org/10.14738/aivp.51.2749

is a vast amount of knowledge for each specific language (phoneme inventories, for instance) which allows us to discriminate between languages. A related field is automatic Vernacular Identification (VI), a relatively recent area of research which is very close to non-native accent identification. Dialectal differences are often proudly marked and native speakers do not generally attempt to conform to a standard variant. Non-native speakers, on the other hand, show different degrees of reading competence and pronunciation competence [7], that is, their knowledge of the grapheme-to phoneme conventions of the foreign language may vary a lot, as well as their ability to pronounce sounds which are not part of their native sound inventory.

Whenever someone makes an utterance, it not only communicates message made up of words and sentences with a specific meaning, but also the same speech signal carries information about the person and his mental or physical state of affairs. This information can sometimes be utilized by the listeners albeit the applied technology to describe and classify different speakers and accordingly, the age, gender, accent, language, emotion, physical conditions etc., of speakers may be categorized to serve a purpose.

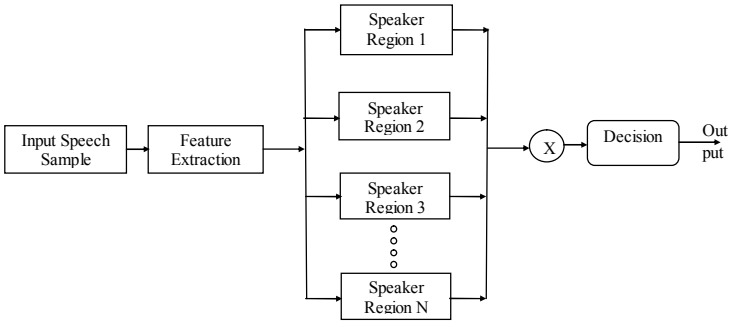

Figure. 1: Speaker Identification System

The reason behind going for algorithms like MFCC HMM is the fact that they are very popular and also produce accurate results. The importance of the field of biometrics is going stronger day by day and even a lot of the popular operating systems are incorporating these features and not only Operating systems, we are seeing several popular social websites are using these features to classify different persons. A typical speaker recognition system is made up of two components; feature extraction and classification. Speaker recognition (SR) can be divided into speaker identification and speaker verification. Speaker identification system determines who amongst a closed set of known speakers is providing the given utterance as depicted by the block diagram in Fig. 1. Speaker specific features are extracted from the speech data, and compared with speaker models created from voice templates previously enrolled. The model with which the feature matches the most is selected as the legitimate speaker. In most cases, the model generates a likelihood score and the model that generates the maximum likelihood score is selected.

Five major accents of SIE are identified in the literature: (persons whose native Indian language is (a) Kannada, (b) Tamil, (c) Tegulu, (d) Malayalam and (e) Marathi Speaking English language). The term 'South Indian English' is used to refer collectively to all the accents of English spoken in south Indian states. The aim is to develop speaker accent independent speech recognition system for South-Indian languages (Devanagari script) based on accent and interpret it to Standard English Accent.

Some of the factors that make the speech recognition problem easier or harder are the following:

- The quality of the speech and the channel over which it is received.

- The number of possible languages from which the system must choose.
- The length of the utterance on which the decision must be made. The amount of training data that is available for making the models for each language. Both total duration of the training speech and the number of different training speakers are important factors.
- The availability of transcripts of the training speech and phonetic dictionaries for the language to assist in the creation of the models for the language.

2 State-of-the-Arts

Several studies have considered acoustic modeling for different accents of the same language. Approaches include the pooling of data across accents, leading to a single accent-independent acoustic model; the isolation of data for each accent, leading to individual accent-specific acoustic models; and adaptation techniques in which models trained on one accent are adapted using data from another. Recently, selective data sharing across accents through the use of appropriate decision-tree state clustering algorithms have also received some attention. These studies extend the multilingual acoustic modeling approach first proposed by Schultz and Waibel to apply to multiple accents of the same language.

Most of the above studies consider the scenario in which the accent of the incoming speech is known and each utterance is presented only to the matching set of acoustic models. This approach is appropriate when the aim is to evaluate different acoustic modeling strategies without allowing performance to be influenced by the effects of accent misclassification. However, in many practical situations, the accent of the incoming speech would not be known. In such cases a single system should be able to process multiple accents.

With the advancement of automated systems the difficulty for integration & recognition problem is increasing. The problem is found more complex when processing on randomly varying analog signals such as speech signals. The acoustic-phonetic approach is the straightforward way of decoding the speech signal in a sequential manner based on the observed acoustic features of the signal and the known relations between acoustic features and phonetic symbols. The acoustic-phonetic approaches have not achieved the same success in practical systems as have alternative methods. An overview of several proposed approaches to automatic speech recognition by machine and the basic strengths and weaknesses of each approach is provided in this section.

There are three approaches to Speech Recognition, namely

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

Robust speech recognition is presented by Yunxin Zhao in his paper [2]. In his proposal an EM algorithm is formulated in the DFT domain for joint estimation of parameters for distortion channel and additive noise from online degraded speech, and the posterior estimates of short-time speech power spectra are obtained at the convergence of the EM algorithm. Any speech features derivable from power spectra can then be approximately estimated by minimum mean-squared error estimation. For their testing the speech data were taken from the TIMIT database and were degraded by a distortion channel and colored noise at various SNR levels.

Yunxin Zhao with Shaojun Wangf and Kuan-Chieh Yen analyses an EM type recursive estimation of time varying environment for speech recognition [3]. Their experimental results showed significant improvement in recognition word accuracy due to the proposed recursive estimation as compared

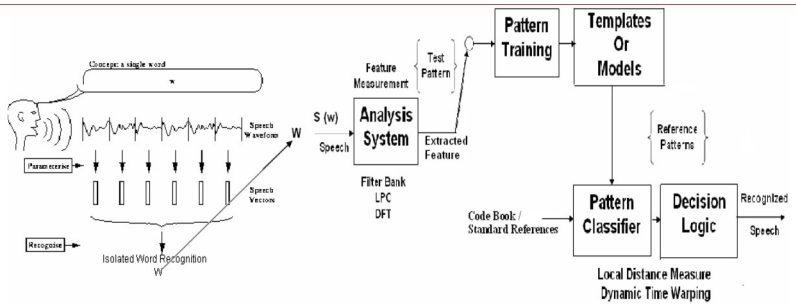

Figure.2: The System Models of the isolated word identification.

with the results from direct recognition using a baseline system and from performing speech feature estimation using a batch EM algorithm. In their proposal the recursive estimation is proposed for tracking both channel and noise parameters in time-varying degradation environments to improve robustness of speech recognition. This method reduces error significantly at high SNR but is ineffective at low SNR.

3 Automatic Accent Identification

Speaker variability, such as gender, accent, age, speaking rate, and phones realizations, is one of the main difficulties in speech recognition task. It is shown in [14], that gender and accent are the two most important factors in speaker variability. Usually, gender-dependent model is used to deal with the gender variability problem.

In south India, almost every state has its own dialect. When speaking accent, the speaker's dialect greatly affects his/her accent. Some typical accents, such as Kannada, Marathi, Telugu and Tamil and Malayalam, are quite different from each other in acoustic characteristics. Similar to gender variability, a simple method to deal with accent problem is to build multiple models of smaller accent variances, and then use a model selector for the adaptation. A cross-accents experiment in performance of accent-independent system is generally 30% worse than that of accent-dependent one. Thus it is meaningful to develop an accent identification method with acceptable error rate.

Although foreign accent identification is extensively explored, little has been done to domestic one, to the best of our knowledge. Actually, domestic accent identification is more challenging: 1) Some linguistic knowledge, such as syllable structure used in [15], is of little use since people seldom make such mistakes in their mother language; 2) Difference among domestic speakers is relatively smaller than that among foreign speakers. In our work, we want to identify different accent types spoken by people with the same mother language.

Most of current accent identification systems, as mentioned above, are built based on the HMM framework, while some investigated accent specific features to improve the performance. Although HMM is effective in classifying accents, its training procedure is time-consuming. Also, using HMM to model every phoneme or phoneme-class is not economic. We just want to know which accent type the given utterances belong to. Furthermore, HMM training is a supervised one: it needs phone transcriptions. The transcriptions are either manually labeled, or obtained from a speaker independent model, in which the word error rate will certainly degrade the identification performance.

In this section, we train two HMMs for each accent: one for male, the other for female, since gender is the greatest speaker variability. Given test utterances, the speaker's gender and accent can be

identified at the same time, compared with the two-stage method in [16]. The commonly used feature in speech recognition systems, MFCC, is adopted to train the HMMs. The relationship between HMM parameter and recognition accuracy is examined. We also investigate how many utterances per speaker are sufficient to reliably recognize his/her accent. We randomly select N utterances from each test speaker and averaged their log-likelihood in each HMM. It is hoped that the more the averaged utterances, the more robust the identification results. Experiments show that with 4 test utterances per speaker, about 11.7% and 15.5% error rate in accent classification is achieved for female and male, respectively. Finally, we show the correlations among accents, and give some explanations is shown in the fig. 2.

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.

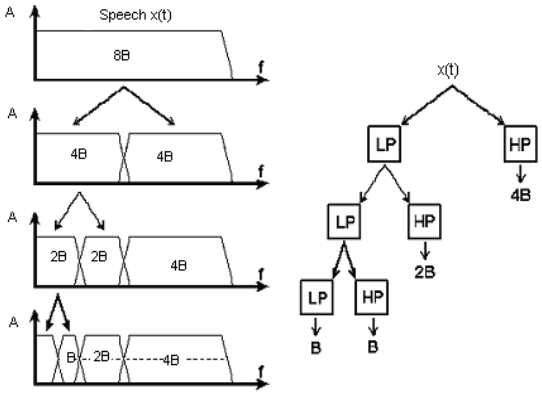

Figure 3: Splitting the signal spectrum with an iterate filter bank.

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 be analyzed is covered in the places of interest. The disadvantage of this scheme 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 minimum details of 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 numbers 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 3. The spectral decomposition obtained coefficient could be observed as shown in figures 4 & 5.

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 4 it is seen that it is left with, after the repeated spectrum splitting, a series of band-pass bands with doubling bandwidth and one low-pass band. The first

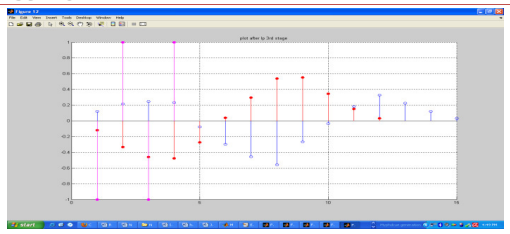

Figure. 4: Output after 3rd stage decomposition for a given speech signal

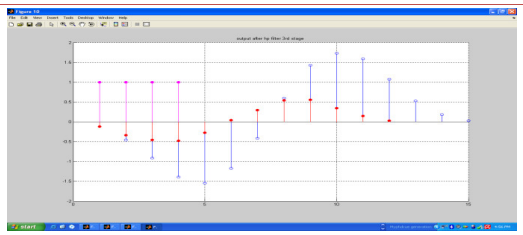

Figure. 5: Plot after 4th stage decomposition

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

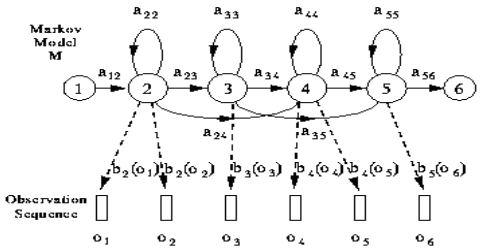

Figure. 6:The HMM Generation Model

5 Markov Modeling

Stationary finite state Markov chains are a simple class of processes from which more complex process models can be derived. Hidden Markov Models belong to this class of processes, since they are composed of probabilistic functions describing the outputs that are associated to the states of a Markov Chain. For non-stationary process, such as the sequence of the speech observation vectors Hidden Markov Models turns out to be a good model. The definition of Markov Chains and of Hidden Markov models is introduced and their relation with the random processes to be modeled is outlined with some examples in this chapter. The specific case of the application of the Hidden Markov Models to speech recognition, in which the output probabilistic functions associated to the Markov Chain are mixtures of normal distributions, is considered. Methods for the initialization of Hidden Markov Models parameters are addressed. These methods supply starting values that are used by training iterative methods to find the final estimate of the parameters. Let each spoken word be represented by a sequence of speech vectors or observations ***O***, defined as

$$O = o_1, o_2, o_3, \ldots, o_i \tag{1}$$

Where Oi is the speech vector observed at time *t*. The isolated word recognition problem can then be regarded as that of computing

$$\arg\max\{P(u/O)\} \tag{2}$$

Where ω_i is the ***i***'th vocabulary word? This probability is not computable directly but using Bayes' Rule gives

$$\{P(u/O)\} = \frac{P(O/u_i)P(u_i)}{P(O)} \tag{3}$$

Thus, for a given set of prior probabilities $P(\omega_i)$, the most probable spoken word depends only on the likelihood $P(O|\omega)$. Given the dimensionality of the observation sequence, O, the direct estimation of the joint conditional probability $P(o_1,o_2,\ldots..,|\omega_i)$ row examples of spoken words is not practicable. However, if a parametric model of word production such as a Markov model is assumed, then estimation from data is possible since the problem of estimating the class conditional observation densities $P(O|\omega_i)$. is replaced by the much simpler problem of estimating the Markov model parameters. Figure.6 A typical left-right HMM (aij is the stationary transition probability from state i to state j; O_t is the observation vector at time t and $b_j(O_t)$. is the probability that O_t is generated by state i).

In HMM based speech recognition, it is assumed that the sequence of observed speech vectors corresponding to each word is generated by a Markov model as shown in Fig.6 A Markov model is a finite state machine which changes state once every time unit and each time t that a state j is entered; a speech vector O_t s generated from the probability density $b_j(O_t)$. Furthermore, the transition from state i to state j is also probabilistic and is governed by the discrete probability a_{ij}. Fig. 6 shows an example of this process where the six state model moves through the state sequence X=1,2,2,3,4,4,5,5,6 in order to generate the sequence o_1 to o_6. Notice that in HTK, the entry and exit states of a HMM are non-emitting. This is to facilitate the construction of composite models as explained in more detail later. The joint probability that O is generated by the model M moving through the state sequence X is calculated simply as the product of the transition probabilities and the output probabilities. So for the state sequences is X in the Fig.6. However, in practice, only the observation sequence O is known and the underlying state sequence X is hidden. This is why it is called a Hidden Markov Model.

$$P\left[O_j, X/M\right] = o_{12}/(o_1)o_{22}/(o_2)o_{23}/(o_3) \tag{4}$$

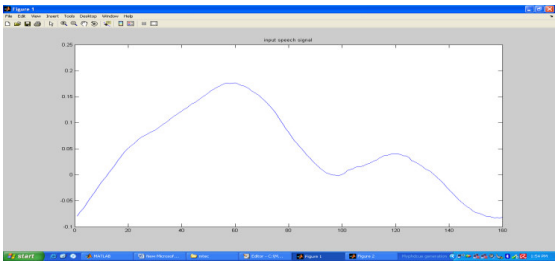

Figure7: (a) Input speech signal and its noise effect on speech signal

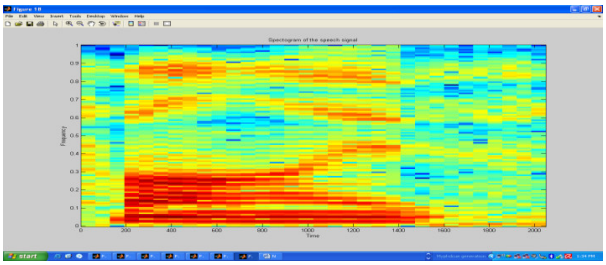

Figure7: (b) the spectral plot for the noised speech

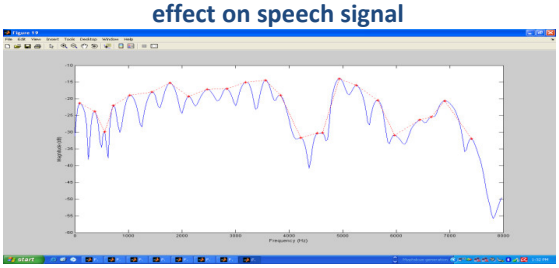

Figure7: (c) the energy peak points picked for training

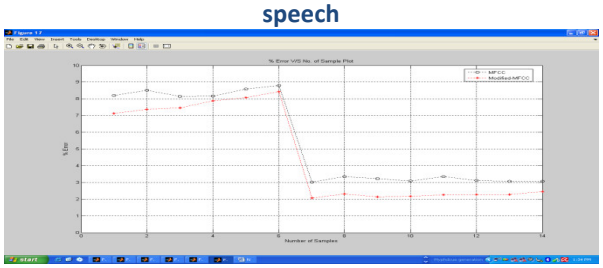

Figure7: (d) the recognition computation time for the MFCC based and the modified MFCC system.

$$Retrieval\ Accuracy\ (\%) = \left(\frac{No.of\ truely\ recognized\ Words}{Total\ No.of\ Words}\right)\times 100$$

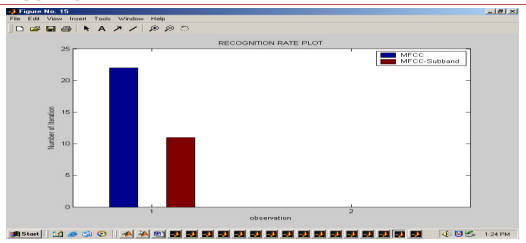

Figure.7 (e) Performance plot showing accurate iteration of classification

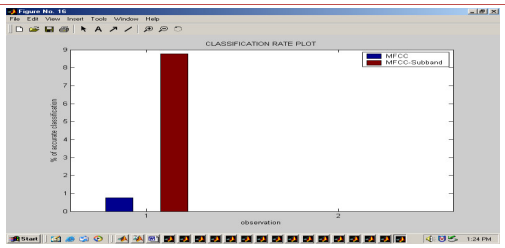

Figure.7 (f) Performance plot showing percentage of accurate classification

Table1. Accuracy performance for the developed methods

Samples		Recognition		True Match count		Accuracy (%)	
	Utterance	Sub-MFCC	MFCC	Sub-MFCC	MFCC	Sub-MFCC	MFCC
S1	"Its easy to tell the depth of well"	"it easy to tell the depth of tell"	"it easy to well effect well"	7/8	6/8	87	50
S2	" Discrete Fourier transform"	" Discrete Fourier transform"	" Discrete For transform"	3/3	2/3	100	66

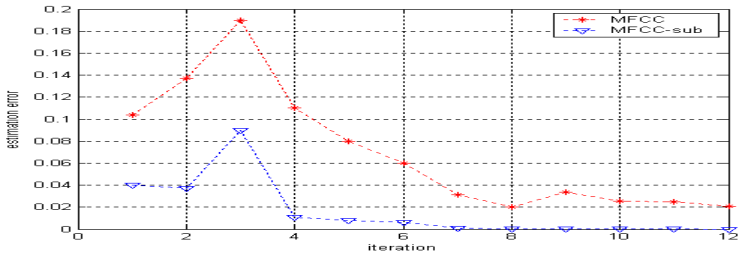

Figure 7(g): Estimation error plot for developed methods with variation in computing iteration.

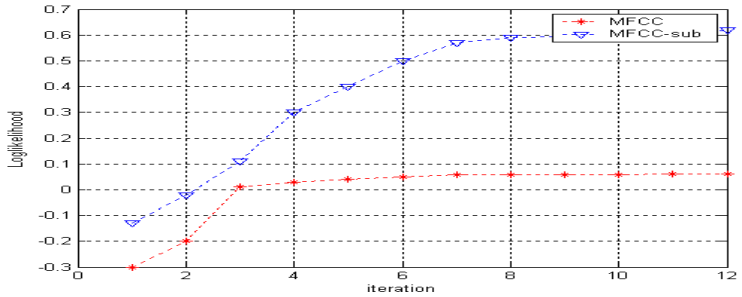

Figure 7 (h): the observed correct classified symbols for the two methods

Performance fig. 7 (a to d) shows the simulation plot for number of iteration taken for the recognition of given speech signal from trained data base from the plot. It is observed that for the recognition speech signal MFCC with HMM takes above loglikelihood iteration to which the error limit where as the same process is done with iteration using MFCC –subband recognition method about 45% of iteration are saved using MFCC-subband. From the result obtained as shows fig. 7 (e) ,(f), (g) and (h) it is seen that the accuracy of classify the given speech signal is found to be 90% more in case of MFCC –subband method then the MFCC method.

6 Conclusion

A speech recognition system for speech recognition at noise corruption is developed. The MFCC algorithm, which cannot extract 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 sub band 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 implement system it is observed to have better efficiency for accurate classification & recognition compared to the existing method.

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] Yunxin Zhao, "Maximum likelihood joint estimation of channel and noise for robust speech recognition" Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference Volume 2, 5-9 June

[3] Yunxin Zhao; Shaojun Wang; Kuan-Chieh Yen "Recursive estimation of time- varying environments for robust speech recognition" Acoustics, Speech, and Signal Processing, 2001.

[4] Rabiner and B.H. Juang, "Fundamentals of Speech Recognition", Prentice Hall, Englewood Cliffs, NJ, 1993

[5] sadaoki- Furui " Digital Speech Processing , synthesis and Recognition "

[6] Donglai Zhu; Paliwal, K.K. "Product of power spectrum and group delay function for speech recognition" Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference Volume 1, 17-21 May

[7] H.A. Murthy and V. Gadde, "The Modified Group Delay Function and Its Application to Phoneme Recognition", Proc. ICASSP, vol. 1, pp. 68-71, 2003

[8] B. Yegnanarayana and H.A. Murthy, "Significance of Group Delay Functions in Spectrum Estimation", IEEE Trans. Signal Processing, vol. 40, pp. 2281-2289, 1992

[9] Jen-Tzung Chien; Chih-Hsien Huang, "Bayesian duration modeling and learning for speech recognition" Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference Volume 1, 17-21 May 2004

[10] Ramirez, J.; Segura, J.C.; Benirez, C.; de la Torre, A.; Rubio, A. "A new voice activity detector using sub band order-statistics filters for robust speech recognition "Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference Volume 1, 17-21 May 2004 Page(s):I - 849-52 vol.1

[11] K. Woo, T. Yang, K. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," Electronics Letters, vol.36, no. 2, pp. 180–181, 2000.

[12] Smith, N.D.; Gales, M.J.F. "Using SVMs and discriminative models for speech recognition" Acoustics, Speech, and Signal Processing, 2002. Proceedings.

[13] N.Smith and M. Gales ," Speech Recognition using SVMs", in Advances in Neural Information Processing Systems, T.G. Dietterich , S.Becker , and Z. Ghahramani, Eds.,vol. 14. MIT Press,2002.

[14] C. Huang, T. Chen, S. Li, E. Chang and J.L. Zhou, "Analysis of Speaker Variability," in Proc. Eurospeech'2001, vol.2, pp.1377-1380, 2001.

[15] K. Berkling, M. Zissman, J. Vonwiller and C. Cleirigh, "Improving Accent Identification Through Knowledge of English Syllable Structure," in *Proc. ICSLP'98*, vol.2, pp. 89-92, 1998.

[16] C. Teixeira, I. Trancoso and A. Serralheiro, "Accent Identification," in *Proc. ICSLP'96*, vol.3, pp. 1784-1787, 1996.