A Comparison of Gaussian Mixture Modeling (GMM) and Hidden Markov Modeling (HMM) based approaches for Automatic Phoneme Recognition in Kannada

Prashanth Kannadaguli[1], Vidya Bhat[2]
Department of Electronics and Communication Engineering,
Manipal Institute of Technology, Manipal, India
[1]prashscd@gmail.com
[2]bhatk_vidya@yahoo.in

***Abstract*-- We build and compare phoneme recognition systems based on Gaussian Mixture Modeling (GMM) which is a static modeling scheme and Hidden Markov Modeling (HMM) which is a Dynamic modeling scheme. Both models were built by using Stochastic pattern recognition and Acoustic phonetic schemes to recognise phonemes. Since our native language is Kannada, a rich South Indian Language, we have used 15 Kannada phonemes to train and test these models. Since Mel – Frequency Cepstral Coefficients (MFCC) are well known Acoustic features of speech, we have used the same in speech feature extraction. Finally performance analysis of both models in terms of Phoneme Error Rate (PER) justifies the fact that Dynamic modeling yields better results over Static modeling and can be used in developing Automatic Speech Recognition systems.**

Keywords*-- Phoneme Modeling; GMM; HMM; Pattern Recognition; MFCC; PER; Kannada

I. INTRODUCTION

The Automatic Speech Recognition (ASR) system of any language must be able to recognize spoken sentenses, words, syllables and phonemes of that particular language [1]. Here sentenses consist of many utterances of different words, words are made up of many syllables and each syllable is a meaningful utterance of phonemes. Hence it is very clear that phoneme is the smallest part of speech and it is absolutely necessary to build a phoneme recognition system which can be later used for syllable or word recognition which in term can be used for recognizing sentenses leading to a language model basically which works in controlled environments. Keeping this in mind, in order to build a language model for Kannada, this work is our first approach to build a phoneme recognition system.

For phoneme recognition there are several signal processing techniques that have been proposed [3][4][5], which evidently proves that we get the PER in the range 30% to 60%. The most successful results are for HMM which have used MFCC as speech features[3]. Since speech is a pseudo-random signal having quasi-periodic nature, we can also use stochastic analysis for its features' pattern recognition. Hence we are comparing two different modeling techniques namely Gaussian Mixture Modeling and Hidden Markov Modeling for phoneme recognition. In these techniques, we use Bayesian decision rule, also known as Maximum a Posteriori (MAP).

To demonstrate these concepts, we have built a database of 15 Kannada phonemes. Each phoneme is recorded 500 times for training and 200 times for testing with a sampling rate of 8kHz. While recording the phonemes, we have recorded the same phoneme under different background noise but using the same microphone and software tool. Hence we have 7500 phonemes in the training database and 3000 phonemes in testing database. The training phase of phonemes include the mean and covariance of their MFCC to generate a probability density function for each model. Given this model in testing phase, we can estimate the likelihood of any testing sample belonging to all 15 classes and that class which gives higher likelihood is the recognized phoneme.

II. ACOUSTIC PHONETIC MODELING

The basic idea here is to develop a model that aims at the production of the most probable phoneme Q* when we give an accoustic observation sequence S as an input. If Q_i is the i-th possible phoneme sequence and the conditional probability is evaluated over all the possible phonemes and ψ represents the parmeters that are used to estimate the probability distribution, then the Bayesian or MAP decision rule can be given by[6]

$$Q^* = {}^{argmax}_{Q_i} P(Q_i / S, \Psi) \qquad (1)$$

Since each phoneme Q* has to be realized in infinite number of possible acoustic ways, it can be represented by its model M_i which yields

$$M^* = {}^{argmax}_{M_i} P(M_i / S, \Psi) \qquad (2)$$

Here M* is the model of the sequence of phoneme data which represents the linguistic message in the speech input S, $M_i$is the possible phoneme data sequence Q_i, $P(M_i / S, \Psi)$ is the posterior probability model of phoneme data sequence given the acoustic input S and the maximum is evaluated over

978-1-4799-6761-2/15/$31.00 ©2015 IEEE

all the possible models. Now we can apply Baye's ruleas follows

$$P(M_i / S, \Psi) = \frac{P(S / M_i, \Psi) P(M_i / \Psi)}{P(S / \Psi)} \quad (3)$$

Here $P(S / M_i, \Psi)$ is called as acoustic model which accounts for the likelihood that a specific model M_i has produced the acoustic observation S, $P(M_i / \Psi)$ represents language model holding the priory probability of the corresponding phonemes and $P(S / \Psi)$ stands for the apriory probability of the acoustic sequence.

III. METHODOLOGY

There are two phases in our work, Training and Testing.

A. *Construction of Database*

Though the ultimate goal is to develop a speaker independent system, to start with, we have decided to build a speaker dependent system. So all samples were recorded for the same native Kannada speaker both for training and testing. Details of the database is shown in Table1.

B. *Pre-processing*

Since the recordings of speech samples were made in normal conditions with different background noise, it becomes absolutely necessary to isolate speech from noise including end point detection of speech. We have used the method proposed in [7] for noise removal. Our database has different folders arranged by phoneme unicodes inside which, all corresponding phonemes are saved after pre-processing in .wav format.

C. *Feature Extraction*

Mel-Frequency Cepstral Coefficients were used as the acoustic phonetic features. The MFCC extraction includes Pre-emphasis, Framing, Windowing, computation of Fast Fourier Transform (FFT), Mel Frequency warping, its logarithm and finally finding Discrete Cosine Transform (DCT) as explained in [1][8][9]. The output of DCT is of 12 dimensions. For pictorial representation of phonemes, we have used first three dimensions of MFCC data. Such a plot for four phonemes is as shown in Fig.1 and it can be observed that phonemes have serious overlap in 3D vector space.

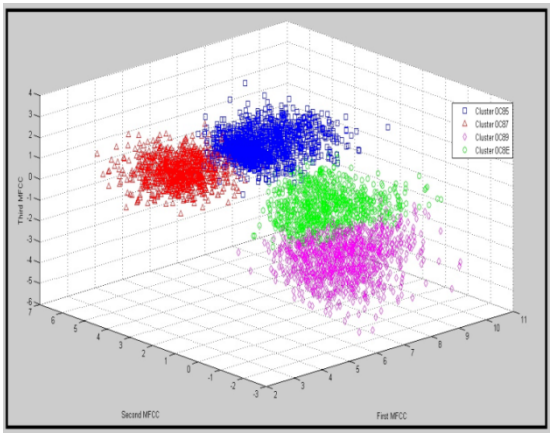

Fig.1: 3D scatter plot for first four phonemes

TABLE1: DETAILS OF PHONEME DATABASE

Unicode	Kannada Character	Number of Training samples	Number of Testing Samples
0C85	ಅ	500	200
0C87	ಇ	500	200
0C89	ಉ	500	200
0C8E	ಎ	500	200
0C92	ಒ	500	200
0C950CBD	ಕ	500	200
0C950CBF	ಕಿ	500	200
0C950CC1	ಕು	500	200
0C950CC6	ಕೆ	500	200
0C950CCA	ಕೊ	500	200
0C970CBD	ಗ	500	200
0C970CBF	ಗಿ	500	200
0C970CC1	ಗು	500	200
0C970CC6	ಗೆ	500	200
0C970CCA	ಗೊ	500	200

D. *Phoneme recognition using GMM*

To recognize an unknown phoneme from our testing database given its MFCC, we perform multivariate model for each class by calculating the mean and covariance matrices of corresponding phoneme sequences[10]. The mean and standard deviation ellipse of the multivariate processes shown in Fig.1 is plotted in Fig.2.

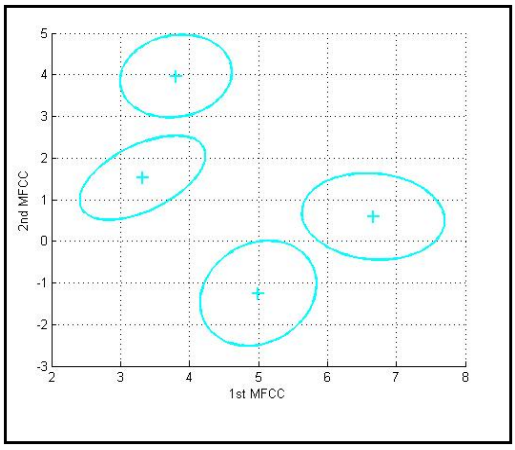

Fig.2: Mean and Standard deviation ellipse for the multivariate process in Fig.1

Later we estimate the likelihood of the given test feature vector using the multivariate model for each class. We have used the standard Gaussian Probability Density Function[12][13]. This implicitly assumes that the MFCC vectors in each class have a uni-modal normal distribution, which returns the estimates of mean and covariance matrix of Gaussian multivariate data samples. 2D plot of data samples of four phonemes using first two MFCC values and the

equivalent 3D plots using Gaussian Mixture Modeling are shown in Fig.3.

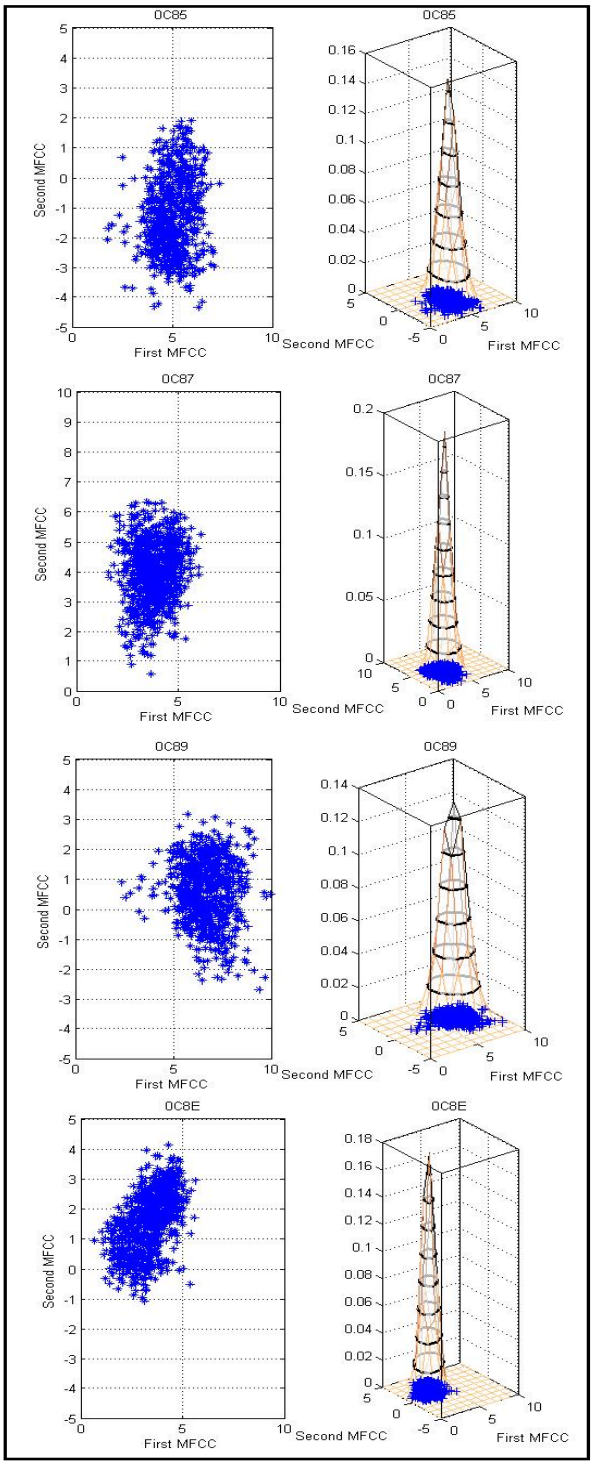

Fig.3: 2D scatterplots and 3D Gaussian PDF plots for phonemes 0C85,0C87,0C89 and 0C8E

The Expectation Maximization (EM) algorithm is used as the main training function in GMM. The EM tries to maximize the likelihood of the data, for the given GMM parameters like mean, covariance. The estimation step is purely soft classification wherein for each feature vector it calculates probability of a class, given that feature vector. In maximization step, mean and covariance of each class is updated using all features and a weight. The algorithm iterates on both steps until the total likelihood increases for the training data. During testing we use the features of unknown signal to estimate the likelihood of the sequences in the feature vector and obtain the posterior probabilities.

E. Phoneme recognition using HMM

Here we follow the same steps until we get the Gaussian Multivariate PDF as in Fig.3. The Expectation Maximization (EM) algorithm is used as the main training function in HMM. The EM tries to maximize the likelihood of the data, for the given HMM parameters like gaussian mean, covariance and state transition probabilities. We have used Baum-Welch (Forward-Backward) algorithm for training. Since phonemes can be properly modelled using single emission state [1], for each phoneme we hace used the HMM model as shown in Fig.4.

The estimation step is purely soft classification wherein for each feature vector it calculates probability of a class, given that feature vector. In maximization step, mean and covariance of each class is updated using all features and a weight. The algorithm iterates on both steps until the total likelihood increases for the training data. After the initialization of HMM parameters the training of the model is carried out using forward-backward algorithm [1]. The training results in updation of all elements of transition and emission probability matrices. The newly obtained matrices for state transition and emission represent the trained HMM and need to be saved so as to use them in testing stage. During testing we use the features of unknown signal to estimate the likelihood of the sequences in the feature vector and obtain the posterior probabilities.

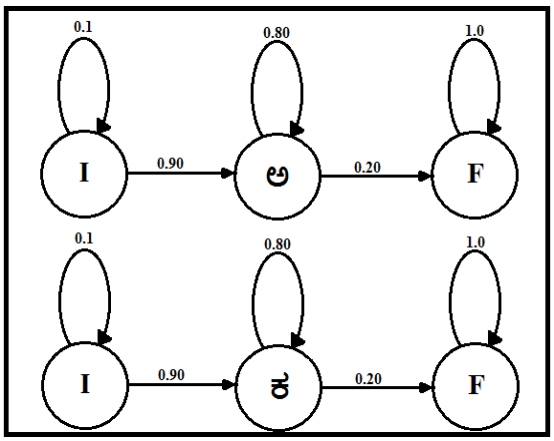

Fig.4: HMM models for phonemes 0C85 and 0C950CBD

IV. RESULTS AND DISCUSSIONS

The result analysis was done by using Phoneme Error Rate (PER), which can be defined as the ratio of the number of phonemes misclassified to the total number of phonemes used for testing. The PER comparison is as shown in Table2.

TABLE 2: PER COMPARISON

Unicode	PER for GMM based Recognizer (%)	PER for HMM based Recognizer (%)
0C85	0	0
0C87	0	0
0C89	0	0
0C8E	0	0
0C92	0	0
0C950CBD	1.6	0
0C950CBF	0	0
0C950CC1	0.4	0
0C950CC6	2.4	0
0C950CCA	5.6	0
0C970CBD	0	0
0C970CBF	2.2	0
0C970CC1	6.2	0
0C970CC6	0	0
0C970CCA	0	0

In the PER calculation for GMM based phoneme recognizer, though the phonemes having similar PDF or similar pronunciations lead to misclassification[14], the results are better than Bayesian phoneme recognizer[11]. From the PER for HMM based phoneme recognizer it is clear that the phonemes modeled using HMM can be accurately recognized. The HMM results are consistant and better than GMM based phoneme recognizer because of the fact that the GMM recognizer considers the successive feature vectors are statically independant and they do not take time into account but the HMM recognizer does.

V. CONCLUSION

In this work, we compared speaker dependent phoneme recognition in Kannada language using GMM and HMM based approaches. From the acoustic phonetic approach, a phoneme based modeling is adopted. Furthermore, the template classifier and pattern recognizer are built using two methods, which are of stochastic approaches. Phonemes are used as the smallest unit for recognition. End point detection, automatic silence removal of the training samples are performed with an algorithm based approach. Detected phonemes are used to train the models which use 12 dimensional MFCC feature matrices as the observation vectors. During the recognition phase, the speech features of an unknown utterance are searched against the different phoneme models trained. The concatenation of the scores of those parts makes up the whole result. In this thesis, the recognition strategy was described in detail and the experiment results were presented. From these results it can be inferred that, each of the valid combinations of Kannada phonemes, that results in Kannada vowels or consonants can be successfully modeled using HMMs, with one emitting state per model. Results reveal that, this method is suitable for building automatic phoneme recognition systems. In future it is necessary to investigate the performance of different acoustic models such as phoneme level models, syllable level models and word level models, so that the best one among them may be used for building Continuous time speech recognizing system in Kannada language. This work can be further extended by including various acoustic phonetic features and by investigating the influence of the number of states used in HMM on the performance of the system.

ACKNOWLEDGMENT

We thank everyone who supported us with valuable suggestions during this work.

REFERENCES

[1] Lawrence R. Rabiner, B. H. Juang, “Fundamentals of speech recognition”, 2nd Indian Reprint, Pearson Education, pp. 103-455, Delhi, 1993.

[2] Y. Lee and K.W. Hwang, “Selecting good speech features for recognition,” ETRI, vol. 18, Apr. 1996.

[3] S. Young, “The general use of tying in phoneme based HMM speech recognition”, proceedings of ICASSP, 1992, pp. 569-572.

[4] C. H. Lee, J. L. Gauvain, R. Pieraccini and L. R Rabiner, “Large vocabulary speech recognition using subword units”, Speech Communication, vol. 13, pp. 263–279, 1993.

[5] Atal. B, Rabiner, L., “A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition” Acoustics, Speech, and Signal Processing, IEEE Transactions , Volume: 24, Issue: 3, pp. 201 – 212, Jun 1976.

[6] T.Dutoit, F. Marques,”Applied signal processing”, Springer 2008.

[7] G. Saha, Sandipan, “A new silence removal and endpoint detection algorithm for speech and speaker recognition applications”, Department of Electronics and Electrical Communication Engineering Indian Institute of Technology, Khragpur, Kharagpur, India.

[8] M.A.Anusuya and S.K.Katti, "Speaker independent kannada speech recognition using vector quantization", Proceedings published by International Journal of Computer Applications® (IJCA)ISSN: 0975 – 8887, MPGINMC-2012 ,7-8 April, 2012.

[9] M.A.Anusuya and S.K.Katti, " Wavelet packet based kannada speech recognition ", Proceedings published by International Journal of Computer Applications® (IJCA)ISSN: 0975 – 8887, MPGINMC-2011, 7-8 April, 2012.

[10] J. O. Berger, “Statistical decision theory and bayesian analysis”, Springer, 1993.

[11] Prashanth Kannadaguli, Vidya Bhat, “Bayesian classifier based automatic phoneme recognizer for Kannada”, International Journal Of Electronics And Communication Technology (IJECT), ISSN: 2230-7109 (Online), 2230-9543 (Print), Volume. 5, Issue. 4, October – December 2014, pp 64-66.

[12] J. MacQueen, “Some methods for classification and Analysis of multivariate observations”, Proc. Of Fifth Berkely symposium on Mathematical Statistics and Probability, June 21-July 18, 1965 and December 27, pp-281-297, 7th January- 1966.

[13] R. K. Aggarwal and M. Dave, “Using gaussian mixture for hindi speech recognition system,” International Journal of Speech processing, image Processing and Pattern Recognition, vol. 4, no. 4, December 2011.

[14] Prashanth Kannadaguli, Vidya Bhat, “Multivariate gaussian mixture model based automatic phoneme recognizer for kannada”, International Journal of Engineering Research and Technology (IJERT), ISSN: 2278-0181, Vol.3 Issue 10 October 2014, pp-725-728.