Multivariate Gaussian Mixture Model based Automatic Phoneme Recognizer for Kannada

Prashanth Kannadaguli[1], Vidya Bhat[2]
Department of Electronics and Communication Engineering,
Manipal Institute of Technology, Manipal, India

***Abstract*—We build an automatic phoneme recognition system based on Gaussian Mixture Modeling (GMM) which is a static modeling scheme. Models were built by using Stochastic pattern recognition and Acoustic phonetic schemes to recognize phonemes. Since our native language is Kannada, a rich South Indian Language, we have used 15 Kannada phonemes to train and test these models. Mel – Frequency Cepstral Coefficients (MFCC) are well known Acoustic features of speech[3][4]. Hence we have used the same in speech feature extraction. Finally performance analysis of models in terms of Phoneme Error Rate (PER) justifies the fact that though static modeling yields good results, improvisation is necessary in order to use it in developing Automatic Speech Recognition systems.**

***Keywords*—Phoneme Modeling; GMM; Pattern Recognition; MFCC; PER; Kannada**

I. INTRODUCTION

The Automatic Speech Recognition (ASR) System of any language must be able to recognize spoken sentences, words, syllables and phonemes of that particular language [3]. Here sentences 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 sentences 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 [1][2][5][6][9], which evidently proves that we get the PER in the range 5% to 30%. The most successful results are for HMM which have used MFCC as speech features[5]. Since speech is a pseudo-random signal having quasi-periodic nature, we can also use stochastic analysis for its features' pattern recognition. Hence we have used GaussianMixtureModeling which uses 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 includes the mean and covariance of their MFCC to generate a probability density function using multivariate modeling. 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. WORKING OF GMM

The basic idea here is to develop a model that aims at the production of the most probable phoneme Q* when we give an acoustic 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 parameters that are used to estimate the probability distribution, then the Bayesian or MAP decision rule can be given by[7]

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

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

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

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 the samples were recorded for the same native Kannada speakerboth for training and testing.Details of the database are shown in Table1.

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

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 [8] for noise removal. Our database has different folders arranged by phoneme Unicode 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 computation of Discrete Cosine Transform(DCT) as explained in [4][9]. The output of DCT is of 12 dimensions. For pictorial representation of phonemes, we have used first two 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 2D vector space.

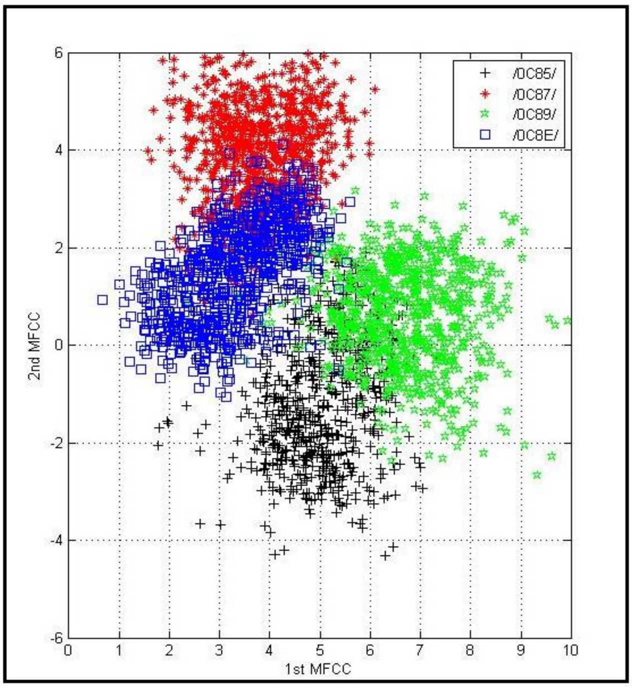

Fig.1: 2D scatter plot for four phonemes

D. *Phoneme recognition using GMM*

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

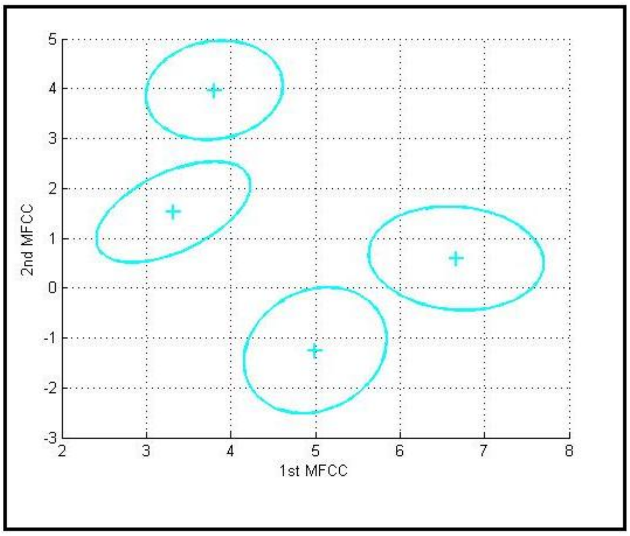

Fig.2: Mean and Standard deviation ellipse for the multivariate process in Fig.1.

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. 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.

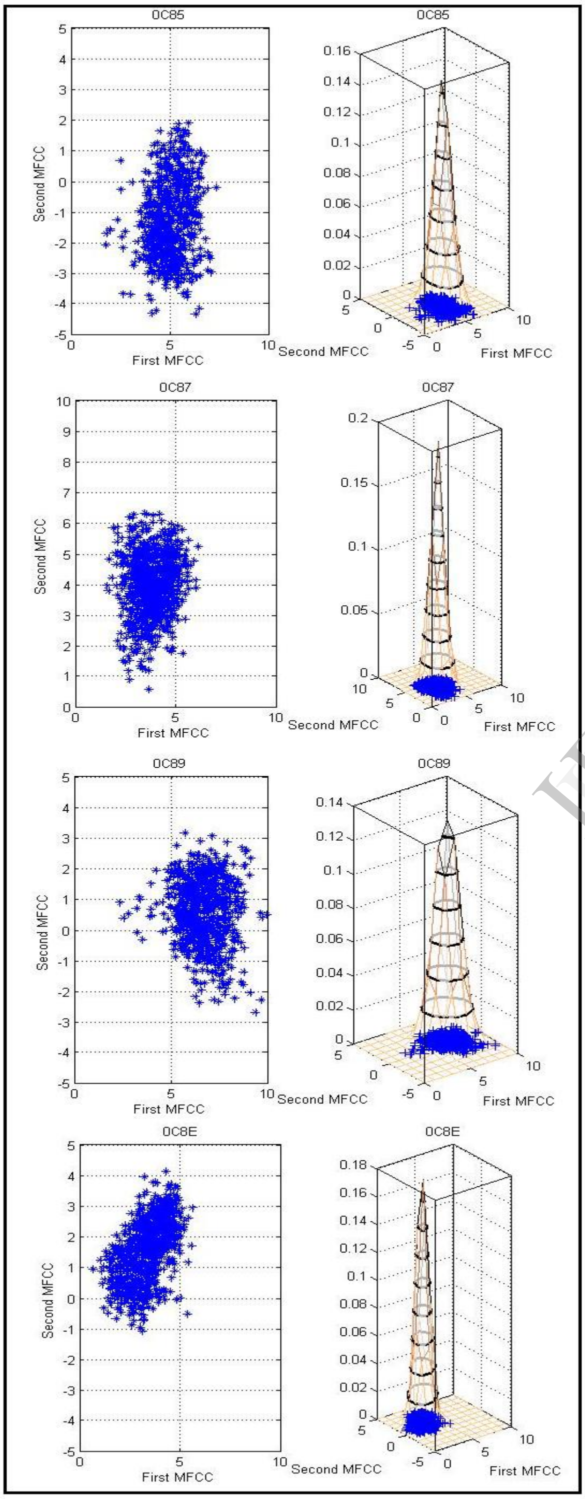

Fig.3: 2D scatterplots and 3D Gaussian PDF plots for phonemes 0C85, 0C87, 0C89 and 0C8E

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.

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 calculation is as shown in Table2.

TABLE 2: PER CALCULATIONS

Unicode	Kannada Character	PER (%)
0C85	ಅ	0
0C87	ಇ	0
0C89	ಉ	0
0C8E	ಎ	0
0C92	ಒ	0
0C950CBD	ಕ	1.6
0C950CBF	ಕಿ	0
0C950CC1	ಕು	0.4
0C950CC6	ಕೆ	2.4
0C950CCA	ಕೊ	5.6
0C970CBD	ಗ	0
0C970CBF	ಗಿ	2.2
0C970CC1	ಗು	6.2
0C970CC6	ಗೆ	0
0C970CCA	ಗೊ	0

Histograms of the GMM based phoneme recognizer is as shown in Fig.4. It is clear that the phonemes having similar PDF or similar pronunciations lead to misclassification. Above results are consistent with the results obtained from other traditional methods [5] and the results are better than Bayesian phoneme recognizer [10].

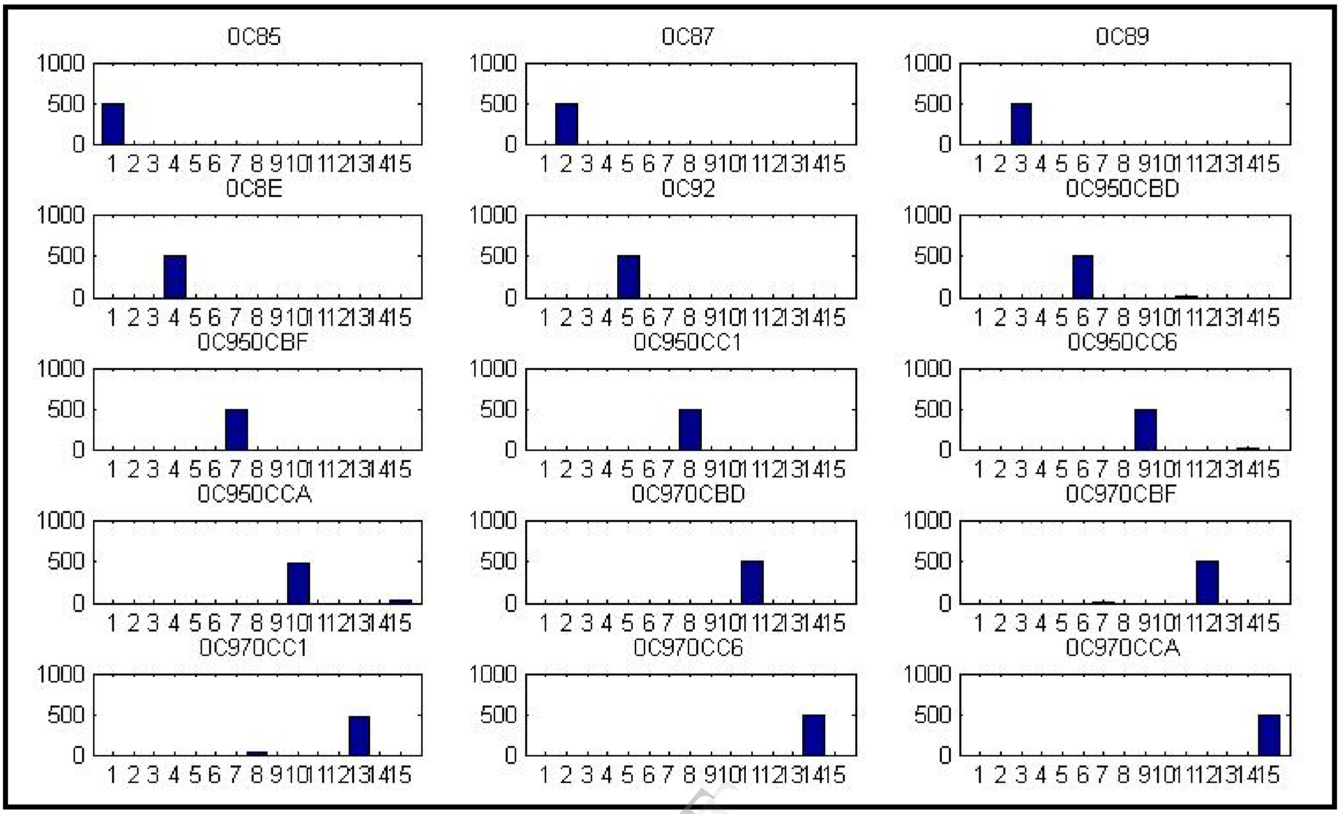

Fig. 4: Histogram of the outputs of the GMM based phoneme recognizer, for samples of each of the fifteen possible input phonemes. The integer values along x-axis refer to the index of the phoneme.

V. CONCLUSION

In this work, we presented Gaussian Mixture Modeling for phoneme recognition. This is a novel approach, different from traditional methods. Results reveal that, this method is suitable for building automatic phoneme recognition systems. This work can be further extended by including various acoustic phonetic features and by using Hidden Markov Modeling as a different approach in automatic phoneme recognition for Kannada language.

ACKNOWLEDGMENTS

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

REFERENCES

[1] 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, December2011.

[2] 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.

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

[4] Y. Lee and K.W. Hwang, “Selecting Good Speech Features for Recognition,” ETRI, vol. 18, Apr. 1996.

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

[6] S. A. Zahorian, P. Silsbee, and X. Wang, “Phone classification with segmental features and a binarypair partitioned neural network classifier,” proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-97), 1997, pp. 1011 -1014.

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

[8] 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.

[9] M. A. Anusuya and S. K. Katti, "Mel Frequency Discrete Wavelet Coefficients for Kannada Speech Recognition using PCA" in Proceedings of International Conference on Advances in Computer Science, 2010.

[10] Prashanth Kannadaguli, Vidya Bhat, “Phoneme modeling for speech recognition in kannada using bayesian multivariate modeling”, Unpublished.