Estimation of HMM-GMM Parameter for Tamil Words

Geetha K [#1], Dr.Vadivel R [*2]

[#] Ph.D Scholar, D.J. Academy for Managerial Excellence,
Coimbatore, India

[1] geethakab@gmail.com

[*] Department of Information Technology,
Bharathiar University

[2] vlr_vadivel@yahoo.co.in

Abstract—**An *isolated word based Hidden Markov Model(HMM) is built for Tamil spoken words. Tamil utterances of monosyllable, disyllable and trisyllable words of Tamil speech utterances are considered. The Hidden Markov Tool Kit(HTK) is used to build and estimate the HMM parameters of the acoustic units. The Mel- Frequency Cepstral Coefficients (MFCC) are extracted from the speech utterances and Multivariate Gaussian Mixture Model with different number of components is used to estimate the state emission probabilities of the HMM. Finally, Viterbi Decoder employed to recognize the test speech utterances. The performance of the models is measured in terms of percentage of correctness, percentage of accuracy and Word Error Rate (WER) and found that five state model with four Gaussian components per state produced the best result of monosyllable words comparing with disyllable and trisyllable words of Tamil Language.***

Keyword-Hidden Markov Model, Gaussian Mixture Model, Speech Recognition

I. Introduction

Speech recognition System (SRS) is a technique which aims to convert a speaker's spoken utterance into a text string[1]. Even though many researchers involved in speech recognition systems/application, it is still far from the solved problem.Continuous HMMs are the best choice for the construction of the acoustic model for speech recognition system[2]. In this statistical model, if a sufficient amount of training data is available, it yields a large number of parameters which leads to the better accuracy in recognition systems. Evaluation of more number of elementary Gaussians during recognition leads to the degradation of the speed of the system. Further, if the training data are less or not sufficient, the system performance will be degraded[3].

Subspace Gaussian Mixture Model (SGMM) in which the state dependent GMM parameters are derived from global shared model subspace and low-dimensional state-dependent vectors. This type of statistical model is suitable for the low resource speech recognition systems[4][5].

Indian languages like Tamil, is syllabic in nature. It also has a close relation between what is spoken to what is written. A word that consists of a single syllable is called a monosyllabic word[6]. A word containing two and three syllables is called as disyllabic word and trisyllabic word respectively. Polysyllable refers either to a word contains more than three syllables or to any word of more than one syllable.

II. Related Work

Even though developing speech recognition system has enormous issues and challenge, there are many research works have been grown and some of the related works are presented here. Riedhammeret. al.[7]. compared classic and multiple-codebook semi continuous models using diagonal and full covariance matrices with continuous HMMs and subspace Gaussian mixture models. They experimented on the RM and WSJ corpora and found that a classical semi continuous system does not perform as well as a continuous one, multiple-codebook semi-continuous systems can perform better, particularly when using full-covariance Gaussians.

Poveyet. al.[4] described an acoustic model in which all phonetic states share a common Gaussian Mixture Model structure. The means and mixture weights vary in a subspace of the total parameter space and called as a Subspace Gaussian Mixture Model (SGMM) in which globally shared parameters define the subspace. This type of acoustic model suits when the training data is less since it allows for a compact representation of the signal and gives better results.

Many researchers are involved in the research for Tamil speech recognition/synthesis system. The Phoneme Recognition System[8] is improved by using language models at the recognition phase of the system. Speech signals were segmented using language models and recognition was done using similarity measure, based on the acoustic characteristics of the phoneme signal. The errors in the recognized phoneme sequence were detected and corrected using the integrated model of variable length phoneme model and inter-word hybrid language model.

Karpagavalli and Sabitha[9] implemented a small vocabulary, speaker independent isolated word recognition system trained with ten Tamil words uttered five times by 20 speakers and achieved 93.5 percentages of accuracy

III. Hidden Markov Model

Hidden Markov Model is a statistical framework and it can be defined as a finite set of states, each of which is associated with a probability distribution. Transitions among the states are governed by a set of probabilities called transition probabilities. In a particular state an outcome or observation can be generated, according to the associated probability distribution[10][11].

A. Elements of HMM

HMM for discrete symbol observations is characterized by $(N, M, A, B, \pi\,)$, where:

- N, the number of states in the model, the individual states are labeled as $\{1,2,3\ \ldots N\}$ and the state at time t is denoted by q_t
- M, the number of distinct observation symbols per state. The individual symbol is denoted by
$$V = \{v_1, v_2, v_3, \ldots \ldots., v_M\}$$
- State transition probability distribution $A = a_{ij}$ where
$$a_{ij} = P\{q_{t+1} = j | q_t = i\}, \quad 1 \le i, j \le N$$
- Observation symbol probability distribution in state j, $B = \{b_j(k)\}$, in which
$$b_j(k) = P(o_t = v_k \mid q_t = j)$$
$$1 \le k \le M, j = 1,2,,\ldots N$$
- Initial state distribution
$$\pi = P[q_1 = i] \qquad 1 \le i \le N$$

For convenience, a HMM is defined in compact notation (N, M, λ) where $\lambda = (A, B, \pi\,)$.

B. Three Problems of HMM

Problem 1: Given the observation sequence $O = (\,o_1, o_2, o_3, \ldots \ldots \ldots, o_T)$and a model $\lambda = (A, B, \pi\,)$, how to efficiently compute $P(O|\,\lambda)$, the probability of the observation sequence?

Problem 2: Given the observation sequence $O = (o_1, o_2, o_3, \ldots \ldots \ldots, o_T)$and a model $\lambda = (A, B, \pi\,)$, how to choose a corresponding state sequence $q = (\,q_1, q_2, q_3, \ldots \ldots \ldots, q_T)$ that is optimal?

Problem 3: How to adjust the model parameters $\lambda = (A, B, \pi\,)$, to maximize $P(O|\,\lambda)$?

Problem 1 is known as evaluation problem. Problem 2 is used to test and find the optimal sequence fit in the probabilistic model. Problem 3 adjusts the parameter of the HMM which is also known as training HMM.

C. Continuous Density HMM(HMM-GMM)

The observations generated by all distinct states of this type are infinite and continuous, i.e. $V = \{v | v \in R^L\}$ The observation probability distribution in state j, $B = \{b_j(v)\}$, is a continuous probability density function and is often a mixture of L dimensional Multivariate Gaussian distributions given by Eq. 1.

$$b_j(v) = \sum_{k=1}^{M} c_{jk} (\frac{1}{(2\pi)^{L/2} |\Sigma_{jk}|^{1/2}} \exp\left(-\frac{1}{2}(v - \mu_{jk})^T \Sigma_{jk}{}^{-1} (v - \mu_{jk})\right)) \qquad \ldots(1)$$

where Mis the number of Gaussians assigned to state j, and the$\boldsymbol{\mu_{jk}}$ and $\Sigma_{jk}$are the means and covariance matrices of the mixtures. $c_{jk}$is the mixture coefficient for the k^th^ mixture in state j.

IV. Experimental Results

HMM GMM parameters for Tamil Syllables are estimated using Hidden Markov Model Toolkit. The Windows 7 operating system has been used for the experiment. Speech recognition based on the Hidden Markov Model-Gaussian Mixture Model (HMM-GMM) framework generally involves training a completely separate GMM in each HMM state. The outline of the process of training and testing followed in HTK is presented in the Fig 1.

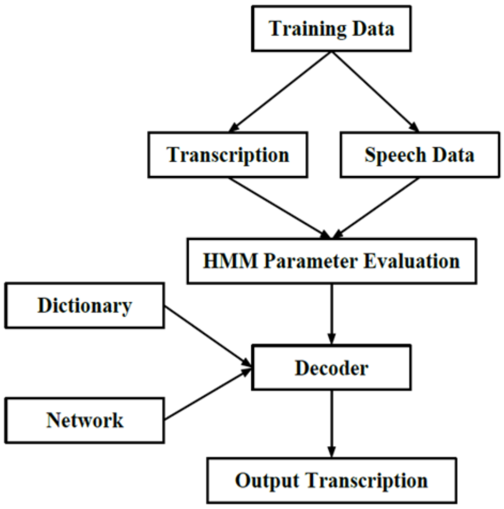

Figure 1. HTK Processing Stages

A. Data

Data are recorded with the help of a high quality microphone in a closed room using a recording tool audacity. The recorded audiofiles are saved as HTK transcription. The sampling rate used for recording is 16 kHz. 30 isolated Tamil words in each category (Monosyllable, Disyllable and Trisyllable) spoken by five native speakers (3 Male and 2 Female) were recorded. Each word is uttered five times by each speaker. 80 % of the utterances is used for training and 20 % of the utterances used for testing. Sample words uttered by the speaker is presented in the table I.

TABLE I
Sample words considered for the experiment

Type of the word	*Examples*
Monosyllable	தா,பூ,கை, பை, வா,கோ
Disyllable	*டெங்கு,கடி, கொக்கி, கொசு,*
Trisyllablle	*கமுதி, குரங்கு, பெரிது, பிடிப்பு*

B. Feature Extraction

The steps involved in the feature extraction are pre-emphasis, frame blocking, windowing, filter bank analysis, logarithmic compression, and discrete cosine transform[12][13]. The overall process of the MFCC feature computation is illustrated in fig. 2. The input speech data are pre emphasized with a coefficient of 0.97 using a first order digital filter. It is then segmented into twenty millisecond frames with an overlap of fifty percent between adjacent frames and windowed using Hamming window. The filter bank analysis is performed to convert each time domain frame of samples into the frequency domain. The log Energy and twelve MFCC coefficients are thereafter extracted. Delta and acceleration coefficients were also derived. Models are constructed with twelve MFCC coefficients with log energy (MFCC), twelve MFCC coefficients with log energy and its velocity (MFCC+D) and twelve MFCC coefficients with log energy, velocity and acceleration coefficients (MFCC+D+A).

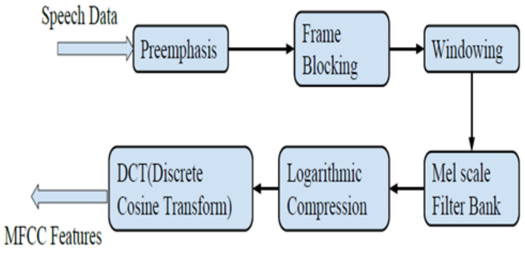

Fig. 2. MFCC Feature Extraction

C. HMM-GMM

MFCC feature vectors extracted from the frames of the speech signal corresponding to each word are given as input to estimate the parameters of HMM. With variant HMM parameters, the number of states (N) and the number of Gaussian mixture per state were used to test the performance of the each case of the implemented system. Left-right model with 4 states, 5 states and 6 states were implemented in which, the state 1 and state N are treated as non emitting states. Fig 3 shows the Bakis model with 6 states in which state 1 and state 6 act as non emitting states.

Initially, a flat start mechanism is used for initialization of HMM-GMM model with one Gaussian mixture per each emitting state of 4 state model, then the parameters for each word is generated. Once an initial set of models is created, re estimation of the entire training set is done. Each of the models in each case is then re estimated by increasing the mixture component until it reaches six. The same procedure is followed to evaluate 5 state and 6 state model.

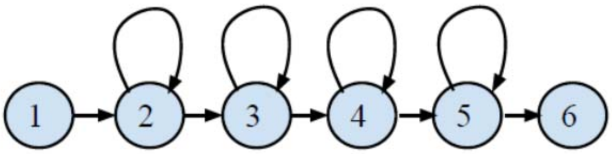

Fig.3. Six state HMM with state 1 and state 6 as non emitting states

D. Experimental Results

The performance of the system is tested with trained speaker. The formula for evaluating performance of monosyllable, disyllable and trisyllable words of Tamil utterances is given in eqn. 2, where N is the number of words in the test set, D is the number of deletions, S is the number of substitutions and I is the number of insertions. Percentage of word accuracy is obtained by eqn. 3.

$$\text{Percentage of Correctness (PC)} = (N-D-S)/N \times 100 \qquad (2)$$
$$\text{Percentage of Accuracy (PA)} = (N-D-S-I)/N \times 100 \qquad (3)$$

Word Error Rate (WER) is used as the criterion to evaluate the performance of the system and it is calculated by eqn. 4.

$$\text{WER} = 100 - \text{PA} \qquad (4)$$

Percentage of accuracy of the isolated syllabic words of Tamil words with MFCC with log energy and the number of Gaussian mixture is four is presented in the table II. The result shows that five state model with double data acceleration coefficient better suits to design a syllable based Tamil recognition systems.

TABLE III
Recognition Accuracy Percentge

Type of the Word	**No. of words used for training**	**No. of words used for testing**	**No. of recognized words**	**Insertion/Substitution /Deletion**	**%Accuracy**	**WER**
Monosyllable	600	150	134	I=6, S=10, D=0	89.3%	10.7%
Disyllable	600	150	128	I=5, S=17, D=0	85.3%	14.7%
Trisyllable	600	150	120	I=7,S=23, D=0	80%	20%

V. Conclusion

Automatic Speech Recognition is a challenging area of research in developing the interaction between man and machine. Since Tamil language is syllabic in nature, this work is carried out to fix the HMM parameters for developing syllable based Tamil Speech Recognition System using the Gaussian Mixer Model. In this paper, experiments are carried out to estimate the HMM model parameters for the construction of the syllabic based speech recognition system for the Tamil language. So, HMM for monosyllable, disyllable and trisyllable Tamil words are constructed and tested with variant MFCC feature sets, states and Gaussian mixtures. The performance is evaluated and the result shows that the five state model with four mixture component is enough to implement syllable based Tamil isolated word recognition systems.

References

[1] R. Rabiner, and B. H. Juang, *Fundamentals of Speech Recognition*, (Prentice-Hall International, 1993).

[2] L.R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proceedings of the IEEE*, vol. 77, Issue 2, pp. 257–286,1989.

[3] Lukas Burget., Petr Schwarz., MohitAgarwal., Pinar Akyazi., Kai Feng., ArnabGhoshal., OndrejGlembek., NagendraGoel., Martin Karafiat., Daniel Povey., AriyaRastrow., Richard C., Rose., and Samuel Thomas., Multilingual Acoustic Modeling for Speech Recognition Based on Subspace Gaussian Mixture Models, *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, (Page: 4334 - 4337, 2010).

[4] Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O.Glembek, N. Goel, M. Karafi´at, A. Rastrow, R. Rose, P. Schwarz, and S. Thomas, The subspace Gaussian mixture model - A structured model for speech recognition, (2011) *Computer Speech and Language*, vol. 25, pp. 404–439,.

[5] Jiyong Zhang, Fang Zheng, MingxingXu, Ditang Fang, Semi-Continuous Segmental Probability Modeling For Continuous Speech Recognition, *Proceedings of theInternational Conference of Spoken Language Processing*, (Oct. 2000, Beijing).

[6] R.Thangarajan,A.M. Natarajan and M. Selvam, Word and Triphone Based Approach in Continuous Speech Recognition for Tamil Language, (2008) *WSEAS Transaction on Signal Processing*, 4(3), pp 76-85, ISSN:1790-5022.

[7] K. Riedhammer, T. Bocklet, A. Ghoshal, D. Povey, Revisiting Semi-Continuous Hidden Markov Models, *Proceeding of IEEE-ICASSP*, (pp.4721-4724, 2012).

[8] S.Saraswathi and T.V. Geetha, Improvement in Performance of Tamil Phoneme Recognition using Variable Length and Hybrid Language Models, *Proceeding of IEEE-ICSCN*, (Page:11-15, 2007).

[9] S.Karpagavalli, P.V.Sabitha, Isolated Tamil Words Speech Recognition Using Linear Predictive Coding And Neural Networks, (2012) *International Journal of Computer Science and Management Research,* ISSN 2278-733X.

[10] Hidden Markov Model Toolkit - available at http://htk.eng.cam.ac.uk,2012.

[11] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying (Andrew) Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, ValtchoValtchev, Phil Woodland, *The HTK Book*, Cambridge University Engineering Department, December 2006.

[12] Manish P., Kesarka., Feature Extraction for Speech Recognition, *Seminar Report*, *Electronic Systems Group, EE. Dept, IIT Bombay*, pp. 1-11, 2003.

[13] Anusuya, M. A., &Katti, S. K.. *Front end analysis of speech recognition: A review, (2011*) International Journal of Speech Technology, Springer, Vol.14, pp. 99–145.

AUTHOR PROFILE

Mrs.Geetha. K completed her B.Sc. and M.Sc. in Bharathidhasan University, Thiruchirappalli. She perceived her M.Phil. in the area of Database Management System from Mother Theresa Women's University. She is currently working as Assistant Professor in the Department of Computer Science, Bharathiar University, Coimbatore. She has attended many seminars and workshops conducted by various educational Institutions and presented her research papers. She is guiding M.Phil. Scholars. Her research interest lies in the area of Speech Recognition Systems, Data Management System and Neural Networks. She is a life member of CSI and IAENG.

Dr. R. Vadivel Completed his B.E. in Periyar University, Salem and M.E. in Annamali University, Chidambaram. He obtained his PhD degree from ManonmaniamSundaranar University, Thirunelvely, Tamil Nadu in 2013. At present he is working as an Assistant Professor in the Department of Information Technology, Bharathiar University, Coimbatore. He has published more than 40 research papers in National, International Journals and Conferences. He is guiding M.Phil and Ph.D. research scholars. His research interest lies in the area of Computer Networks & Security, Mobile Computing, Mobile Ad-Hoc Networks, Wireless Sensor Networks, Data mining and Digital Signal Processing. He is a life member of CSI, ISTE, ACS, ISCA, AMIE, IACSIT and IAENG.