Large Vocabulary in Continuous Speech Recognition Using HMM and Normal Fit

Hemakumar G[#1], Punithavalli M [*2], Thippeswamy K [#3]

[1#] *Research Scholar, Bharathiar University, Coimbatore, Tamil Nadu, India and Government College for Women (Autonomous), Mandya, Karnataka.*
[2*] *Department of Computer Application, Bharathair University, Coimbatore, Tamil Nadu, India.*
[3#] *Department of Computer Science, Visvesvarya Technological University, Mysore Regional Centre, Mysuru, Karnataka, India.*

Abstract— this paper addresses the problem of large vocabulary speaker independent continuous speech recognition using the phonemes, Hidden Markov Model (HMM) and Normal fit method. Here we first detect for the voiced part in speech signal through computing dynamic threshold in each frame. Real Cepstrum coefficients are extracted as features from the voiced frames. The Baum–Welch algorithm is applied for training those features. Then normal fit technique is applied, the outputted values are labelled using correspondent phoneme or syllable. The model is tested for 5 languages namely English, Kannada, Hindi, Tamil and Telugu. The automatic segmentation of speech signals average accuracy rate is 95.42% and miss rate of about 4.58%. In the large vocabulary, average Word Recognition Rate (WRR) is 85.16% and average Word Error Rate (WER) is 14.84%. All computations are done using mat lab.

Keywords — *Automatic Speech Recognition (ASR), Speech Enhancement, Speech Perception, HMM and Normal fit method.*

I. INTRODUCTION

Automatic Speech Recognition is a computerized process where machine shall receive as its input a speech recording and it produces as its output a transcription. The main aim of an ASR system is to accurately and efficiently convert a speech signal into a text message transcription of the spoken words, independent of the device used to record the speech (i.e., the transducer or microphone), the speaker's accent or the acoustic environment in which the speaker is located (e.g., office, noisy room, outdoors). The major problem that complicates ASR implementation is speaker's variability. Because ASR systems are supposed to be general use systems they have to support multiple speakers and be able to adapt to all the variations that introduces. There are variations in speech styles, pitch and anatomy that make each speaker unique. Also things like background noise, utterances, and dialects can negatively affect the interpretation of speech. Even words that sound alike can create problems for ASR systems [1]. In the ASR, problems occur in performance when moving from speaker-dependent (SD) to speaker-independent (SI) conditions for connectionist HMM or Artificial Neural Network (ANN) systems in the context of large vocabulary in continuous speech recognition (LVCSR) [2].

The performance of ASR system may also reduce due to degrade in speech signal by noises. Those noises are like Gaussian white noise, pink, red and gray noises which occur during recording time. The noise occurs by multi speakers during recording is most challenging task to handle. Those noises should be reduced before signal segmentation and feature extraction. The enhancement of speech signal is required in order to improve the intelligibility and overall perceptual quality of degraded speech signal using audio signal processing techniques. The enhancement of speech signal which is corrupted by noise is commonly performed using the short-time discrete Fourier transform domain [3]. The Bayesian algorithm for speech enhancement under a stochastic deterministic speech models which makes provision for the inclusion of *apriori* information by considering a non-zero mean [4]. The filters which allow explicit control of the tradeoff between noise reduction and speech distortion via the chosen rank of the signal subspace [5]. Paper [6] discuss regarding the measurement of enhancement considered a wide range of distortions introduced by four types of real-world noise at two signal-to-noise ratio levels by four classes of speech enhancement algorithms namely spectral subtractive, subspace, based on statistical-models, and Wiener algorithms.

The problem in designing the ASR system may also occur while selecting the frame size. In ASR system the windowing is done using the short-time frequency analysis. But in reality it has been conclude that human hearing is relatively insensitive to short-time phase distortion of the speech signal, so there is no apparent reason for the use of symmetric windows which give a linear phase response [7]. Paper [8] discuss on the paradigm of statistic in speech recognition for phonetic and phonological knowledge sources. They discuss on computational phonology and mathematical models like Bayesian analysis, statistical estimation theory, non-stationary time series, dynamic system theory and nonlinear function approximation theory.

Phonemes classification	English Sounds Alphabets
Vowels Front	/ IY /, / IH /, / EH /, / AE /
Vowels Mid	/ AA /, / ER /, / AH /, / AO /
Vowels Back	/ UW /, / UH /, / OW /
Diphthongs	/ AY /, / OY /, / AW /, / EY /
Semivowels – Liquids	/ W / , / L /
Semivowels – Glides	/ R / , / Y /
Consonants – Nasals	/ M / , / N /, / NG /
Consonants – Stops – Voiced	/ B /, / D /, / G /
Consonants – Stops – Unvoiced	/ P /, / T /, / K /
Consonants – Fricatives – Voiced	/ V /, / TH /, / Z /, / ZH /
Consonants – Fricatives – Unvoiced	/ F /, / THE /, / S /, / SH /
Consonants – Whisper	/ H /
Consonants – Affricates	/ JH /, / CH /

Table 2.1: Shows the details of English Phonemes Classifications and its symbols representation.

Phonetic Classification	Kannada Vowels Sounds
High Vowels	/ ii / , / uu /
Higher-mid vowels	/ e / , / o /
Low vowel	/ aa /
Rounded vowels	/ u / , / o /
Unrounded vowels	/ i /, / e / , / a /
Dipthong	/ ai /, / au /
Additional Vowel	/ R /

Table 2.2: Shows the Hindi, Kannada, Tamil and Telugu languages vowels sound classification and its representations of Symbols in our work.

Phonetic Classification		Kannada Consonants Sounds
Plosive	Unaspirated	/ p /, / b /, / t /, / d /, / k /, / g /, / T /, / D /
	Aspirated	/ ph /, / bh /, / th /, / dh /, / kh /, / gh /, / Th /, / Dh /
Affricates	Unaspirated	/ c /, / j /
	Aspirated	/ ch /, / jh /
Fricatives		/ h /, / s /, / sh /, / Sh /
Nasals		/ m /, / n /, / N /, / nY /, / nG /
Liquids		/ l /, / L /, / r /
Semi Vowels		/ v /, / y /

Table 2.3: Shows the Hindi, Kannada, Tamil and Telugu languages Consonants sound classifications and its representations of Symbols in our work.

In our experiment we adopted the speech enhancement technique based on statistical model for the degraded signals by additive noise as estimation matter which means the concepts of information and noise. In fact most of the work on the signal processing of random signal is concerned with the extraction of information from noisy observation, which leads to estimate clean speech signal from noisy signal. Signal classified into two categories like deterministic signal and random signal (stochastic) and in each class a signal may be continuous or discrete in time. The behaviour of deterministic signal can described in terms of a function of time, and the exact value of the signal can be calculated at any time and predict from the functional description and the past history of the signal. Random signal have unpredictable behave, it is not possible to formulate an equation that can predict the exact value of random process of a signal from its past history. The speech signal and noise are random process at least in part.

Here we have discussed about the large vocabulary speaker independent continuous speech recognition using phoneme and syllables (Syllable for Indian languages like Kannada, Hindi, Tamil and Telugu and Phoneme for English), HMM and Normal fit technique. It also provides the human speech production in briefly and the information regarding the phoneme of English and syllables of the major south Indian languages. In this paper we have used 'an4' speech database for English which has been designed by the Carnegie Mellon University (CMU), Kannada speech database designed by ourselves and we have also collected the speech corpus for Hindi, Kannada, Tamil and Telugu designed by organization called Linguistic Data Consortium for Indian languages (LDC-IL) which has formed by Central Institute of Indian Languages (CIIL) to collect the speech annotation samples for Indian languages. The above mentioned speech database are ruined through Hidden Markov Model toolkit (HTK) and Sphinx toolkit and compared results with the proposed model.

The remaining part of the paper is organized into six different sections; Section 2 deals with the human speech production. Section 3 discuss about speech database. Section 4 deals with proposed model. Section 5 deals with Experimental results. Section 6 deals with conclusion.

II. HUMAN SPEECH PRODUCTION

The human speech perception is quite different from the way computer programs work. We cannot analyse the sound spectrum in complicated mathematical ways as human brain does. The brain is able to, in a very successful way, distinguish interesting sound from noise. This noise could be of many kinds, it could be anything from noise of a big engine to a man speaking unknown language. Almost any kind of noise can be nearly ignored by our brain, enabling us to perceive the important information. Our brain recognizes the speech by representing finite number of sounds called

phonemes. The mechanism in the brain is to detect the acoustic features which lead to understand the meaning of the message in particular language.

In each language there are exact meaningful sound units called phoneme which are broadly classified into five major classes of sound [10]. The broad classifications of English phonemes are as shown in table 2.1. The table 2.2 and 2.3 shows the phonemes classification for major south Indian languages like Kannada, Tamil, Telugu and Hindi which has been used in this experiment. Among those classes the vowel sounds is the largest phoneme group. Vowels contain three subgroups defined by the tongue hump being along the front, central or back part of the palate. The source is quasi-periodic puffs of airflow through the vocal folds vibrating at achieved certain fundamental frequency. Here the term 'quasi' used because there is no perfect periodicity and it is never achieved. The consonants contain a number of sound subgroups as nasals, fricative, whisper and affricates. The nasals sounds are closest to the vowels sounds. As with the vowels, the source is quasi-periodic puffs of airflow through the vocal folds. Fricative consonants are specified in two classes voiced and unvoiced fricatives. The source of the unvoiced fricative is vocal tract relaxed and not vibrating. Noise is generated by turbulent airflow at some point of constriction along the oral tract. Voiced fricatives have similar noise source and system characteristic to unvoiced fricative, for fricatives the vocal tract is vibrate simultaneously with the noise generation at the constriction and the periodicity of the noise airflow is seen. As with in the fricative, plosives also has both voiced and unvoiced type. In the unvoiced plosives a burst is generated at the release of the build-up of pressure behind a total constriction in oral tract. The voiced same as the unvoiced plosives, but that vocal tract can also vibrate. Glides are formed by continually moving the articulator (mouth, tongue etc.) such as /i/, /r/ and /w/.

III. Speech Signal Database

In our experiment, we have used Standard English speech database designed by Carnegie Mellon University (CMU) called 'an4' speech corpus. They have designed the speech corpus by collecting the personal information of the speakers, such as name, address, telephone number, date of birth etc. In addition to these, subjects also spoke randomly generated sequences of words containing control words. All data are sampled at 16 kHz, 16-bit linear sampling. All recordings were made with a close talking microphone. In this database all audio files are RAW files. In this database the directory an4_clstk used for training which consist of the training data has 74 sub-directories, one for each speaker. Here there are 53 male speakers and 21 female speakers. The total number of utterances is

Languages	Total Speakers	Region of Recording	Type of Speech
Hindi	650	Uttar Pradesh and Bihar	Standard, Bhojpuri & Magahi
Kannada	642	North-East (Hyderabad Karnataka), North-West (Mumbai Karnataka) and Canara.	Non standard
Tamil	450	Tamil Nadu	Standard
Telugu	56	Andhra Pradesh	Standard

Table 3.1: Shows the details of LDC-II, CIIL Designed Speech database. Here we have mentioned only languages which have been used in this

948. The directory an4test_clstk is used for the testing purpose which consist the test data has 10 sub-directories, one for each speaker. Here 3 of them are female and 7 are male. The total number of utterances is 130.

Secondly, we have designed the Kannada speech corpus by ourselves. Here 250 unique sentences were randomly selected from the Hampi text corpus and recorded form native and non-native Kannada speakers. Here 15 male and 15 female speakers are selected with age distribution of 15-20, 21-50 and 51+ aged. We have built the Kannada speech corpus which consist of total number of utterances is 7500. Here we have used only 1500 utterances for training and 300 utterances for testing purpose. These signals were recorded at a little noisy environment with the help of Mat Lab recording program designed by us for the purpose of recording with the help of mini microphone at the sampling rate of 16 KHz, 16 bps, mono channel.

Thirdly, we have collected the speech corpus for Hindi, Kannada, Tamil and Telugu languages. Those speech corpuses are designed by the organization formed in the name of Linguistic Data Consortium for Indian languages (LDC-IL) by Central Institute for Indian Languages (CIIL). The designed speech database are consist of voice of the native speakers of that particular languages and who were selected according to age distribution (16-20, 21-50, 51+), Gender, Dialectical Regions and environment (home, office and public place). Each speaker recorded a news text in a noisy environment through recorder having an inbuilt microphone. The recordings are in stereo recording and the extracted channels are also included in the specific files. The table 3.1 shows the details of the speech corpus designed from the organization called LDC-II, CIIL for Hindi, Kannada, Tamil and Telugu languages.

IV. Proposed Model

The proposed model works for speaker independent and in the offline mode. So all speech

signals are pre-recorded and stored in speech database and then passed through algorithm for training or testing the unknown signal. The designed algorithm is capable in the recognition of continuous speech of English and above mentioned Indian languages that to only the trained set of data. Our algorithm first checks the sampling rate of speech signal and automatically converts it into 16 K samples/second, 16 bit resample rate, then it passed for proposed model. The proposed algorithm consists of five steps as follows.

First step is to Pre-processing the sampled signals: In this stage analog speech signal is sampled and quantized at the rate of 16K samples/second. S(n) is the digitalized value. Then using equation 4.1 the DC component is removed from digitalized sample values

$$S(n) = S(n) - \text{mean}(S) \quad (4.1)$$

Then first order (low-pass) pre-emphasis network equation 4.2 is applied to compensate for the speech spectral fall-off at higher frequencies and approximates the inverse of the mouth transmission frequency response.

$$\hat{s}(n) = S(n) - \tilde{a} * S(n-1) \quad (4.2)$$

Here we have used the constant value $\tilde{a} = 0.95$ [11][12]. Then standardization is done to entire set of values to have standards amplitude by applying the equation 4.3. This process will increases or decreases the amplitude of speech signal.

$$S(n) = \hat{s}(n) - \max(|s|) \quad (4.3)$$

The second step is Detection of Voiced region in the speech signal, also called segmentation of speech signal. To solve this problem, we have applied dynamic threshold approach; here algorithm is designed for automatic segmentation of continuous speech signal into phoneme, syllable and sub-word [13] [16]. Here we have combined the short time energy and magnitude of frame. The dynamic threshold of short time energy is computed by using equation 4.4 and dynamic threshold of magnitude of frame is computed by using equation 4.5. Dynamic threshold for each frame is detected by combining equation 4.6 and 4.7. Lastly, it is checked for voiced region in those frames using that frame threshold. This is achieved by following steps [11]

$$Thr_{STE} = \left(\left[\frac{\sum_{i=1}^{n} STE}{n}\right] - [\min(STE) * 0.5]\right) + \min(STE) \quad (4.4)$$

$$Thr_{msf} = \left(\left[\frac{\sum_{i=1}^{n} msf}{n}\right] - [\min(msf) * 0.6]\right) + \min(msf) \quad (4.5)$$

$$if\ (STE \geq Thr_{STE}) then\ marked\ as\ voiced_{STE} = 1 \quad (4.6)$$

$$if\ (msf > Thr_{msf})\ then\ marked\ as\ voiced_{msf} = 1 \quad (4.7)$$

$$if\ (Voiced_{STE} * Voiced_{msf} = 1)\ then\ that\ frame\ contains\ voice,\ otherwise\ its\ unvoiced\ frame$$

Where STE is Short Time Energy, msf is the Magnitude of Frame, n is number of samples in the frame, Thr_{STE} is dynamic threshold value computed using short time energy and Thr_{msf} is dynamic threshold value computed using magnitude of frame.

Third stage is Feature Extraction: Here we have selected the voiced region of signal and then frame blocking has done for *N* samples with adjacent frames spaced *M* samples apart. Typical values for *N* and *M* correspond to frames of 20 ms duration with adjacent frames overlap by 6.5 ms. A hamming window is applied to each frame using frame same size. Next, the autocorrelation is applied to that part of signal. LPC method is applied to detect LPC coefficients. The LPC coefficients are converted into Real Cepstrum Coefficients. Here the output data will be of the size p*L, where p is the LPC order and it will be constant and L is the number of frames in that particular voice segmented region. So it varies. Here 16^{th} order LPC is applied.

The Fourth stage is Speech model building: In this stage the real cepstrum coefficients are in dimension of p*L matrices. This matrix will be passed into k-means algorithm by keeping k=3 and output values are passed into 3 state Baum–Welch algorithm and each syllable or sub-word is trained. The Baum-Welch re-estimation procedure is the stochastic constraints of the HMM parameters [12] [14], namely the equations are 4.8, 4.9 and 4.10 as follows.

$$\sum_{i=1\ldots N} \bar{\pi}_i = 1 \quad (4.8)$$

$$\sum_{j=1\ldots N} \bar{A}_{ij} = 1 \quad , \ 1 \leq i \leq N \quad (4.9)$$

$$\sum_{k=1\ldots M} \bar{B}_j(k) = 1 \quad , \quad 1 \leq j \leq N \quad (4.10)$$

Are automatically incorporated at each iteration. The parameter estimation problem as a constrained optimization of P(O | λ). Based on a standard Lagrange optimization setup using Lagrange multipliers, P is maximized by using equation 4.11, 4.12 and 4.13

$$\pi_i = \frac{\pi_i(\partial P/\partial \pi_i)}{\sum_{k=1\ldots N} \pi_k(\partial P/\partial \pi_k)} \qquad (4.11)$$

$$A_{ij} = \frac{A_{ij}(\partial P/\partial A_{ij})}{\sum_{k=1\ldots N} A_{ik}(\partial P/\partial A_{ik})} \qquad (4.12)$$

$$B_j(k) = \frac{B_j(k)(\partial P/\partial B_j(k))}{\sum_{l=1\ldots M} B_j(l)(\partial P/\partial B_j(l))} \qquad (4.13)$$

Then Normal fit technique is applied for 3 consecutive HMM parameter λ(A, B, pi) and Normal fit parameters are computed. Her the trained three consecutive λ(A, B, pi) are considered has sample data. So, we will be having a sample ($x_1 \ldots x_n$), for this a normal parameter N($\hat{\mu}$, $\hat{\sigma}^2$) is computed by using the equation 4.14

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \hat{\sigma}^2 \sim \frac{\sigma^2}{n} . X^2_{n-1} \qquad (4.14)$$

The output of normal parameter N($\hat{\mu}$, $\hat{\sigma}^2$) is labelled. The labelled (using transcript Roman letters) $\hat{\mu}$ and $\hat{\sigma}^2$ value will be classified according to acoustic classes and then stored. Those data are the representatives of phoneme or syllables or sub-words in that particular class. In Language model we have designed bi-syllable and tri-syllable language model for words and sentences respectively.

The Fifth stage is Recognition part / Testing Unknown Signal: Initially, for the unknown speech signals, HMM parameters are computed and passed into normal fit method. Subsequently, the outputted sigma hat value is identified and then matched with trained set of data by retaining threshold values. The outputted phoneme or syllables or sub-words are matched with the bi-syllable and tri-syllable language model. The concatenation of outputted phoneme, syllables and sub-words are done for sentence building. On this basis decision is taken has recognized sentence by checking for top ranked.

V. Experimental Results

The experimentation is done on recognition of continuous speaker independent speech signals. Here we have tested for Indian languages like Hindi, Kannada, Tamil and Telugu speech and English speech using HTK, Sphinx toolkit and compared with the proposed model for same speech database.

Languages	WRR	WER
Hindi	85.01	14.99
Kannada	85.78	14.22
Tamil	84.87	15.13
Telugu	84.98	15.02
Average	85.16	14.84

Table 5: Provides the results comparison for different languages from proposed model.

Technique	Database	Accuracy	Error Rate
HTK Toolkit	an4 English	84.11%	15.89%
	Kannada	80.81%	19.19%
Sphinx3 Toolkit	an4 English	84.57%	15.43%
	Kannada	80.94%	19.06%
Proposed Model	an4 English	85.01%	14.99%
	Kannada	85.04%	14.96%

Table 6: Providing the results comparison for HTK Toolkit, Sphinx Toolkit and proposed model.

The proposed model is written in mat lab and ruined on Intel Core i5 processor speed of 2.67 GHz and RAM of 3 GB.

Here firstly experimented for automatic speech signal segmentation into syllables and sub-words. The recognition of syllables or sub-words is done for the selected voiced part of signal and not on the frame of the signal. The experiment for identifying the voiced part or unvoiced part in the signal we have used the combination of short time energy and magnitude of signal and computed the dynamic threshold. We have tested this method on isolated words and continuous speech signals of Hindi, Kannada, Tamil, Telugu and English languages. In our experiment the error occurs only in the segmenting the noised speech signal and speaker having more breathing air-pressure noise occurring during the time of speaking or reading. To the continuous speech signal segmentation the noise from breathing is more problem than the noise from external environment. In the experiment of automatic segmentation of speech signal success rate of individually uttered of words / sentences in experiments is excellent and has reached average accuracy rate of 95.42% and miss rate of about 4.58%.

Secondly our experiment is to recognition of speech signal irrespective of speaker. Table 5 shows the compared experimental results by proposed model for the Hindi, Kannada, Tamil and Telugu languages speech database which has designed by an organization called LDC-II, CIIL. The table 6 shows the average accuracy rate and error rate occur from the HTK toolkit, Sphinx3 toolkit and proposed model for an4 and Kannada speech database.

VI. CONCLUSIONS

Our experiment shows that dynamic threshold computation using combination of frame magnitude and short time energy provides excellent accuracy rate in segmenting the speech signals into syllables and sub-words. Then the proposed model clearly showed that normal fit value can be stored has representative of phoneme, syllable and sub-word. The normal fit computes the mean hat and sigma hat using the square root of the unbiased estimator of the variance. The sigma hat is the maximum likelihood value computed from those of trained HMM, so that sigma hat value can be stored has a representative of syllable, which is capable in providing good recognition rate. This experiment shows that storing normal fit values reduces the memory size than storing the HMM parameters. The proposed models can used to design Automatic continuous speaker independent speech recognition for small, medium and large vocabulary for any Indian languages.

ACKNOWLEDGMENT

The authors would like to thank for Department of Computer Science and Application, Bharathiar University, Coimbatore for giving an opportunity to pursuing part-time PhD degree. Author also like to thanks Principal and friends of Government College for Women (Autonomous), Mandya for cooperating while pursuing Ph.D degree. Authors would like to thanks for all our friends, reviewers and Editorial staff for their help during preparation of this paper.

REFERENCES

[1] Douglas O Shaughnessy, Speech Communications: Human and Machine, Universities Press (India) Private Limited, Hyderabad, Reprinted on 2004.

[2] Sabato Marco Siniscalchi et Al., "Hermitian Polynomial for Speaker Adaptation of Connectionist Speech Recognition Systems", IEEE Transactions on Audio, Speech, And Language Processing, Vol. 21, NO. 10, October 2013, page No 2151-2161.

[3] Martin Krawczyk and Timo Gerkmann, "STFT Phase Reconstruction in Voiced Speech for an Improved Single-Channel Speech Enhancement", IEEE Transactions on Audio, Speech and Language Processing, Vol. 22, No. 12, December 2014, Pg. 1931-1940.

[4] Matthew McCallum et al., "Stochastic-Deterministic MMSE STFT Speech Enhancement with General *A Priori* Information", IEEE Transactions on Audio, Speech and Language Processing, Vol. 21, No. 7, July 2013, Pg. 1445-1457.

[5] Jesper Rindom Jensen et al., "A Class of Optimal Rectangular Filtering Matrices for Single-Channel Signal Enhancement in the Time Domain", IEEE Transactions on Audio, Speech and Language Processing, Vol. 21, No. 12, December 2013, Pg. 2595-2606.

[6] Yi Hu and Philipos C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement", IEEE Transactions on Audio, Speech and Language Processing, Vol. 16, No. 1, January 2008, Pg. 229-238.

[7] Robert Rozman and Dusan M. Kodek, "Using asymmetric windows in automatic speech recognition", Speech Communication 49 (2007), page no 268–276.

[8] Li Deng, "A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition", Speech Communication 24 (1998), page no. 299 to 323.

[9] Yi Hu and Philipos C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement", IEEE Transactions on Audio, Speech and Language Processing, Vol. 16, No. 1, January 2008, Pg. 229-238.

[10] Patricia Scanlon and Daniel P.W. Ellis, "Using Broad Phonetic Group Experts for Improved Speech Recognition", IEEE transaction on Audio, Speech and Language processing, VOL 15, No. 3, March 2007.

[11] Hemakumar G. and Punitha P., "Large Vocabulary Isolated Word Recognition Using Syllable, HMM And Normal Fit", published by International Journal of Scientific & Engineering Research, Volume 5, Issue 9, Sept-2014, Pg. No: 34-37, ISSN: 2229-5518.

[12] Hemakumar G. and Punitha P., "Large Vocabulary Speech Recognition: Speaker Dependent and Speaker Independent", Springer - Advances in Intelligent and Soft Computing, Vol 339, Pg. No 73-80, Mandal et al (Eds): Information Systems Design and Intelligent Applications.

[13] V. Kamakshi Prasad et al., "Continuous Speech Recognition Using Automatically Segmented Data as Syllabic Units", Published at ICSP'02 Proceedings, 0-7803-7488-6/02 © 2002 IEEE, Page No.235-238.

[14] Lalit R.Bahl, et al, "Estimating Hidden Markov Model Parameters So as to maximize speech recognition Accuracy", IEEE Transactions on Audio, Speech and Language processingvol.1,no.1, 1993.

[15] Nam Soo Kim et al., "On estimating Robust probability Distribution in HMM based speech recognition", IEEE Transactions on Audio, Speech and Language processing, vol.3, no.4, 1995.

[16] Thangarajan R., Natarajan A. M. and Selvam M. "Syllable modeling in continuous speech recognition for Tamil language", International Journal for Speech Technology, vol. 12, pp.47 -57 2009.

[17] R. K. Aggarwal et al (2011), "Using Gaussian Mixtures for Hindi Speech Recognition System", International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 4, No. 4, December, 2011, page no 157-170.