Grapheme Gaussian Model and Prosodic Syllable Based Tamil Speech Recognition System

Akila.A.Ganesh
Department of Computer Science,
D.J Academy for Managerial Excellence,
Coimbatore, India.
akila.ganesh.a@gmail.com

Chandra Ravichandran
Department of Computer Science,
Dr.SNS Rajalakshmi College of Arts & Science,
Coimbatore-32, India.
crcspeech@gmail.com

***Abstract* – Automatic Speech Recognition is an active field of research to identify speech patterns for providing the equivalent text. The challenges in automatic speech recognition start right from choosing the appropriate unit for the language being dealt with. Some of the units that could be used are word, phoneme, triphone, syllable, demisyllable, senone and morpheme. A syllable based Speech model is used which reduces the vocabulary size and also easier to align. This also suits Tamil, which is a syllable based language. In the model described in this paper, the connected word inputs are segmented into individual words using short term energy and the isolated words are further broken down into characters using Varied-Length Maximum Likelihood (VLML) algorithm. Gaussian Mixture Model (GMM), which is a speaker-independent model suitable for large sets of data, is used for classifying the characters for later pattern matching against the trained syllables. This paper introduces a new algorithm named VLML algorithm that is used for identifying the boundary of each character and explains the proposed process of speech recognition system for Tamil language.**

Keywords- Varied Length Maximum Likelihood (VLML); Formants, Prosodic Syllable; Gaussian Mixture Model (GMM); Hidden Markov Toolkit (HTK)

I.INTRODUCTION

Speech is a natural mode of communication for human and making an Automated Speech Production and Perception needs many resources due to its inherent complexity. The Automatic Speech Recognition (ASR) system has a unit of recognition. Many ASR systems have been developed with various units like phonemes, Triphones, syllable, Senone, demisyllable, Morpheme etc [1]. The proposed system uses a Syllable unit. The flow of the proposed model is specified in Fig 1. Each of the phases of the proposed model is discussed in the following section.

The speech input can be a connected word which is initially segmented into words. The segmentation is done using the Short Term Energy (STE). The signal is segmented into frames and STE of each frame is computed using the Eq. 1 where s(n) is the amplitude of each frame and n is the current frame of the N frames in the speech signal. The Frames with Short Term Energy less than a threshold value are found. The number of frames resulting with short term energy less than the threshold will be equal to the number of words in the connected word speech and each of those frames is the boundary of each word in the connected word speech.

$$E = \sum_{n=1}^{N} |s(n)^2| \quad (1)$$

The performance of the proposed model was identified by the accuracy, correctness.

II. OVERVIEW OF SYLLABLE BASED SPEECH RECOGNITION SYSTEMS

Speech has many field of research like speech recognition, Speech Synthesis, Speech Processing, Speech Enhancement etc. Speech Recognition is an important field of research for more than five decades. Rabiner and his team designed a Demisyllable based Isolated Word Recognition System [2] whose was able to represent any spoken word in a simple concatenative scheme. Syllable based speech recognition for Amharic [3] developed by Solomon used Hidden Markov Model and achieved a recognition accuracy of 90.43%.

An automatic segmentation and labeling Continuous speech signal into syllable like units was developed by Lakshmi [4] which used group delay based algorithm for automatic segmentation. The system was tested with inputs from Indian languages like Tamil and Telugu. A syllable analysis to build a dictation system in Telugu language [5] was done by N.Kalyani and K.V.N. Sunitha. They focused on creating speech database at syllable units and identifying the text with minimum training. A Novel acoustic modeling method for Chinese speech recognition based on Intra Syllable Dependent Phone set is proposed by Zhang and his team [6]. It extends the traditional phone set based on intra syllable information of Chinese phonetic knowledge.

The paper A Syllable Based Continuous Speech Recognizer for Tamil Language [7] presents a novel technique for building a syllable based continuous speech recognizer when unnnotated transcribed train data is available. A group delay based two level segmentation algorithm is proposed to extract accurate syllable units from the speech data. A rule based text segmentation algorithm is used to automatically annotate the text corresponding to the speech into syllable units.

III. PROSODIC SYLLABLE BASED TAMIL SPEECH RECOGNITION SYSTEM

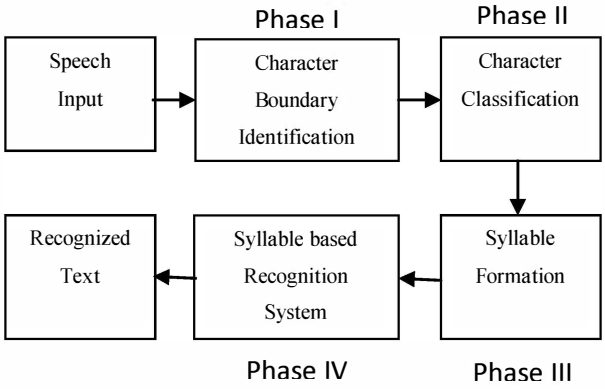

Fig 1: Flow of the Proposed Model

The proposed system has many phases as shown in Fig 1. The proposed system uses Matlab to implement the phase I to IV. The last phase is implemented using Hidden Markov Model Toolkit. The speech input is given to the phase 1 where the boundary of the character is identified and the character is classified in the next phase. Then the syllables are formed and given as input to the last phase of the proposed system.

The proposed system has taken into consideration of the existing algorithms of Speech Recognition and used the features of Speech signals like Formants, Short Term Energy to design the proposed model. The necessity of each phase is discussed and the implementation is compared with the existing whenever necessary in the following sections.

IV.CHARACTER BOUNDARY IDENTIFICATION

The steps carried out in the character boundary identification phase are shown in Fig 2.

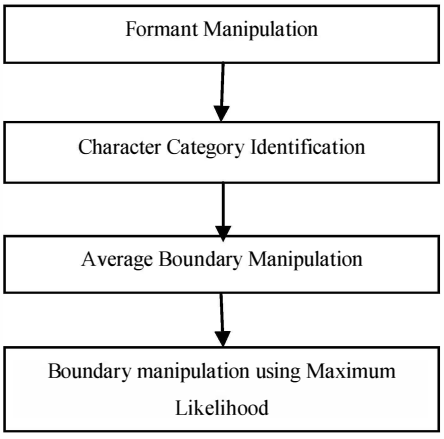

Fig 2: Character Boundary Identification

A. Formant Manipulation

Formants are defined as spectral peaks of the sound spectrum of the voice. It is the amplitude peak in the frequency spectrum of the sound [8]. The formants of a sample signal are shown in Fig 3. Spectrograms are used to visualize formants. The spectrum of phonemes can consist of several formants, but the first three are most important for recognition. Formants are present not only at vowels, but recognition of the vowels based on them is easier and gives better results.

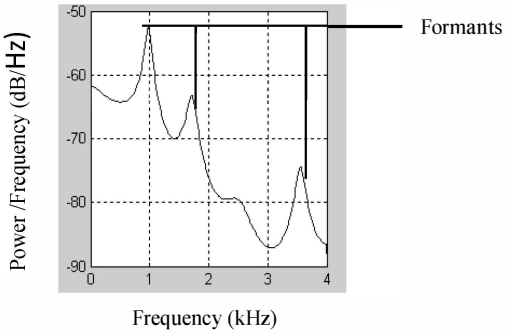

Fig 3: Sample Formants of a speech signal

TABLE I: FORMANT VALUES F0, F1 AND F2 OF SOME SAMPLE TAMIL CHARACTERS

Character	F0	F1	F2
அ	890	1370	2660
ச	845	1585	2390
ன்	370	2495	3030
ஸ்ரீ	735	1975	2835
ய	823	1769	2480

The spectrum of each character consists of many formants but the first three formants are considered to classify the character.

1) Procedure for Formant Manipulation:

Step 1: Perform Feature Extraction using Linear Predictive Coefficient (LPC).

Step 2: Find frequencies by root solving technique.

Step 3: Extract the frequencies which are greater than Zero (in Hertz).

Step 4: Sort the extracted frequencies in ascending order.

Step 5: Get the first three frequencies of the given speech signal of a character.

The Root Solving is a technique which uses the Feature Vector extracted from LPC as coefficients of the polynomial and computes the root of the polynomial. If LPC vector L contains N elements, then the polynomial is represented as L(1)*X^(N-1)+...+L(N-1)*X+L(N).

B. Character Category Identification

1) Identifying Threshold Length of the Character Categories:

The utterance of all the characters in Tamil language are recorded with Audacity and each character is uttered five times. The length of characters according to their character

category is considered to find the mean length of each character category [1]. The mean value found is taken as the threshold length of the particular category. Table I specifies the Threshold length of each category.

2) Identifying the Formant Limits: The first three formant values of all characters uttered five times are found using the procedure specified in the previous section. The maximum formant value for each character category is found by manipulating the mean of formant values F0, F1 and F2 for the multiple utterances of each character. After finding the mean value, the Maximum Formant values for each category is found and considered as formant limits. Table II shows the F0, F1 and F2 Formant Limits of each category [1].

C. *Average Boundary Manipulation*

The threshold length and the formant limit are manipulated for the character set of Tamil language. The number of segments in the given speech can be identified by identifying the number of characters in the input speech. The length and boundaries of each character in the input word are approximately computed using the threshold length and formant limit values. The procedure for finding the length and boundary is shown below [1].

Step 1: Find the sample data for the given input word.

Step 2: The sample data of length equal to that of average length of a short vowel is used to compute the Formants F1, F2 and F3.

Step 3: The formants are compared with the average formants of each Tamil character.

Step 4: If the formant values match the range of any Tamil character, then the appropriate length of the short vowel is found with approximate boundary values.

Step 5: If no match found then the process from step 2 to 4 is repeated for long vowel, vowel consonant and consonant.

Step 6: The segment boundary for each character is found by repeating the process from step 2 to 5 with the different ranges of the sample data.

D. Boundary Manipulation

The Boundary of the character segments of the given input word is found using Varied Length Maximum Likelihood Algorithm (VLML). The VLML is an extension of the Maximum Likelihood Algorithm (ML) [9]. The maximum likelihood Algorithm is a feature vector based segmentation approach. The spectral property similarity within a segment is checked using ML Algorithm. The segment is a collection of frames and frames are the feature vector extracted using Feature Extraction Procedures like Mel Frequency Cepstral Coefficient (MFCC), Linear Predictive Coefficient (LPC). The segmentation in ML is done based on the spectral distortion within a segment which is the deviation on the spectral properties. The frame with minimum Intra segment distortion may correspond to the boundary of the segment. The spectral distortion is measured in terms of intra segment distortion (δ) and generalized centroid (μ). The intra segment distortion (δ) is given by Eq. (2)

$$\delta(i,j) = \sum_{n=i}^{j} d(x_n, \mu) \quad (2)$$

where i and j are the start and end frames of the segment, d is the positive distortion measure which is the Euclidian distance between x_n and μ, x_n is the nth frame in the signal and μ is the centroid of each segment which is calculated using the Eq. 3 where μ' represents the generalized centroid of all segments .

$$\mu = arg \min_{\mu'} \frac{1}{j-i} \sum_{n=i+1}^{j} d(x_n, \mu') \quad (3)$$

The segments given as input to the VLML algorithm is of varied size whose boundaries are found using the algorithm specified in the previous section. In the traditional Maximum Likelihood [1], the segments are assumed to be equal size. The varied length maximum likelihood works as discussed below [1].

The centroid of the all the segments (μ') is manipulated by taking the mean of all the segments. The centroid (μ) and intra segment distortion (δ) are found using Eq 2 and Eq 3 for each segment. The Cumulative Intra Segment Distortion (D_{intra}) for all frames is found using the Eq. 4. The Backtracking pointer is manipulated using Eq 5.

$$D_{intra}(i,j) = \min_{(i-1)\le k<j} \{D_{intra}(i-1,k) + \delta(k+1,j)\} \quad (4)$$

$$\Psi(i,j) = \arg \min_{(i-1)\le k\le j} \{D_{intra}(i-1,k) + \delta(k+1,j)\} \quad (5)$$

The boundary of the last segment M is the last frame N and the boundaries (b_s) of the other segments are found using Eq.6 where s ranges from M-1 to 1.

$$b_s = \Psi(s+1, b_{s+1}) \quad (6)$$

V.CHARACTER CLASSIFICATION

A. Dynamic Time Warping

Dynamic Time Warping (DTW) is a pattern matching algorithm [10] used to find the matching reference pattern for the input pattern. The algorithm uses Euclidean distance to measure the distance between the input pattern and the reference pattern. The DTW is a Speaker dependent algorithm which means that the reference should be created for each recognizing unit to be recognized by the speech recognition system for every user separately. The DTW algorithm is suitable for Speaker Verification Applications and simple command and control systems.

B. Gaussian Mixture Model

Gaussian Mixtures are combinations of Gaussian or normal distributions. A mixture of Gaussians can be written as a weighted sum of Gaussian densities [11] as specified in Eq 10. They are used to model complex multi-dimensional

distributions. Fig 5 and Fig 6 shows a sample three single Gaussians and their one dimensional Gaussian mixture Probability Density Function. The mean (μ) and variance (σ)

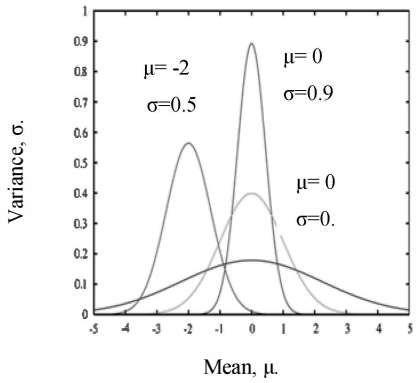

Fig 4: Gaussian Distribution using different Mean and Variance

are the parameters of the Gaussian distribution. They are manipulated using the formula specified in Eq.7 and Eq.8. The mean and variance are calculated for each Gaussian. The Gaussian or Normal distribution with different mean and variance are shown graphically in Fig 4.The dimension d is the number of components which indicates the number of utterances of each word to be trained in the model.

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x(n) \tag{7}$$

Where x(n) is the feature vector of each frame in the N frames of the speech signal.

$$\sigma = \frac{\sum_{n=1}^{N} x(n)^2}{N} - \left(\frac{\sum_{n=1}^{N} x(n)}{N}\right)^2 \tag{8}$$

C. Process in Gaussian Mixture Model

Step 1: Get the Number of Components K.
Step 2: Find the Mean (μ) and Variance (σ) for each mixture of each component using Eq.7 and Eq.8.
Step 3: Find the Gaussian Mixture (g) for each component using Eq.9.
Step 4: Compute the Gaussian Mixture Model (G) with K components using Eq.10.
Step 5: Repeat step 1 to 4 for all reference data to be trained.
Step 6: Compute the log likelihood using posterior function for the test data and each GMM.
Step 7: The reference pattern with maximum likelihood is the classified character for the test data.

The parameters of a probability density function G [11] are the number of Gaussians K, their weighting factors wk, and the mean vector μ_k and covariance matrix $\sum_k$ of each Gaussian function is found using Eq.7 and Eq.8. The single dimension Gaussian (g(x)) is calculated using the Eq.9 where x is the current frame and the Gaussian mixture PDF with K Gaussians is given using Eq.10.

$$g_{(\mu,\Sigma)}(x) = \frac{1}{\sqrt{2\pi}\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \tag{9}$$

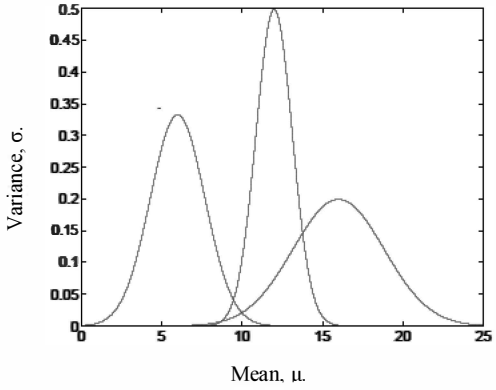

Fig 5: Three single sample Gaussians

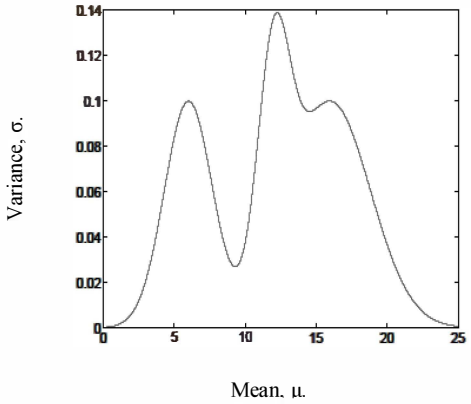

Fig 6: One dimensional Gaussian Mixture PDF

$$G(x) = \sum_{m=1}^{K} w_m \; g_{(\mu,\Sigma)}(x) \tag{10}$$

$$where\ all\ w_m \geq 0\ and \sum_{m=1}^{K} w_m = 1$$

The Gaussian Mixture Model is a statistical Model with the d dimension Gaussian with parameters K mixtures, and mean μ_K and covariance $\sum_K$ for each mixture [11]. In the training phase the model is created for all reference patterns. In the testing phase the log likelihood is computed using Posterior function for the test data with each reference pattern and the reference data with maximum log likelihood is classified as the character of the test data.

D. Performance Comparison of GMM and DTW

The GMM is a parametric model which consumes less time for manipulating parameters for a large set of data. It is easy to implement and it is speaker independent. The GMM follows a Probabilistic Framework and it is computationally efficient.

DTW is cost minimization matching technique and works well for small set of data.DTW is a speaker dependent system which means need actual training for each speaker and each recognition unit.

As GMM performs better in large set of data than DTW and it is speaker independent system, it is suitable for many real time speech applications. In the proposed model, after the character segmentation is performed to the input word using Varied Length Maximum Likelihood Algorithm, it was given to both DTW and GMM to classify the character. The Accuracy of GMM was 77% and the performance of DTW was poor because most of the character classification was wrong. So the GMM Algorithm was used for the character classification in the proposed model.

VI. SYLLABLE FORMATION

A. Prosodic syllable Pattern Rules

TABLE II: Prosodic Pattern Rules

P. No	Pattern	Description
1	SV+LV+C(s)	Syllable with a Short Vowel (SV) followed by a Long Vowel (LV) followed by any number of Consonants (C)
2	SV+LV	Syllable with a Short Vowel (SV) followed by a Long Vowel (LV)
3	SV+SV+C(s)	Syllable with a Short Vowel (SV) followed by a Short Vowel (SV) followed by any number of Consonants (C)
4	SV+SV	Syllable with a Short Vowel (SV) followed by a Short Vowel (SV)
5	SV+C(s)	Syllable with a Short Vowel (SV) followed by any number of Consonants (C)
6	LV+C(s)	Syllable with a Long Vowel (LV) followed any number of Consonants (C)
7	LV	Syllable with only one Long Vowel (LV)
8	SV	Syllable with only one Short Vowel (SV)
9	D	Syllable with only one Diphthong (D)

The syllable can be formed from the individual characters classified with GMM using the prosodic pattern rules as specified in Table II [1]. The pattern rules consists mainly eight rules and one for diphthong. The pattern rule is checked with the input from the pattern 1 to 9 in the same order and if any pattern suits to the input the syllable is formed. Diphthong will be considered as a separate syllable. The Input word can be mono syllabic, bi syllabic or tri syllabic. It can have up to five syllables.

B. *Formation of Syllable*

The syllable is formed by using the pattern rules specified in Table II. The characters classified in the previous section are given as input as a sequence of characters to this phase. The character can be either a Vowel or Consonants. Tamil language consists of 12 Vowels and 18 Consonants. These consonants and vowels combines to form 216 vowel consonants. The 12 vowels consists of 5 Short Vowels, 5 Long Vowels and 2 Diphthongs which is shown in Fig 7. In the 216 vowel consonants, 90 alphabets are Short Vowels Consonants, 90 alphabets are Long Vowel Consonants and 36 alphabets are diphthong consonants.

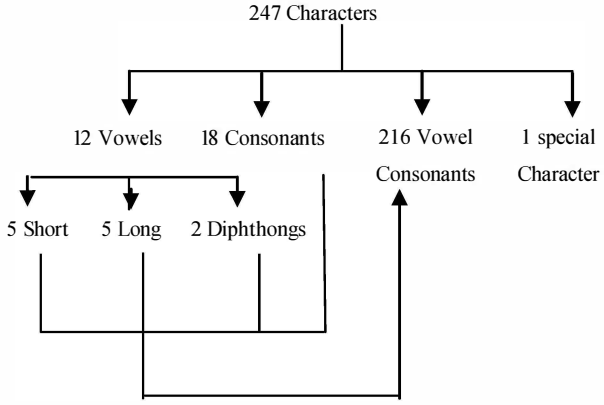

Fig 7: Character Chart for Tamil Language

The short vowel consonant is treated as short vowel , long vowel consonant is treated as long vowel and diphthong consonants is treated as diphthong [12] for applying the syllable pattern rules. The sequence of characters classified using the Character Classification stage is taken as input in this stage and syllable is formed. Initially each character in the input is segregated as Short Vowel, Long Vowel, Consonants and diphthong. The next step will be combine consonants followed by a short vowel as a single unit, consonants followed by Long vowel as a unit and consonants followed by diphthong as a unit. The last step will be to form the syllable using the pattern rules in Table II. The order of pattern rule is followed strictly. The segment of sound signal of each character whose boundary was found using varied length Maximum Likelihood (VLML) are concatenated to form a syllable. To form each syllable sound segment, the pattern described in Table II is used.

VII. SYLLABLE RECOGNITION

A. Data Collection and Implementation

The data used for training and testing was recorded using Audacity with project rate of 8000Hz. A single input channel microphone is used for getting the input. The data was recorded in a noiseless room environment. The data used for training are from the agriculture application. Each input data is uttered 5 times by 4 speakers (2 female and 2 male). The character set of Tamil language was recorded with the same setup as specified above with a single speaker. The phases of the proposed model are implemented using MATLAB 7.10.

B. Hidden Markov Model Toolkit

Hidden Markov Model Toolkit (HTK) is a toolkit for building and manipulating Hidden Markov Model (HMM). It is written in Ansi-C and runs mainly in a UNIX based system. It can run also in other operating system. It consists of library

modules and set of tools. It is mainly used in speech recognition applications. The library modules of HTK provide an effective programming environment for implementing new algorithms and the tools of HTK allow building a range of recognizers quickly and effectively. The HTK can be used to build a recognition system with a very large vocabulary size easier. HTK has been used to build word spotting systems, Speaker separation system and face recognition systems.

C. *Training Phase using Syllables*

The syllables are trained using HMM Toolkit [13]. Initially a Grammar is constructed which defines the constraints that the Speech Recognition Engine (SRE) can expect when an input is given. The list of the syllables used for training the system is listed in the grammar. This grammar is parsed as a word network, which the SRE can understand. The word network is defined using a low level notation called Standard Lattice Format. After constructing the Grammar, the next step will be to build a dictionary. The Dictionary consists of the list of syllables and corresponding output to be given by SRE when the input is recognized. The third step will be creating a transcription file to label the training data. The 39 MFCC Feature Vectors are extracted after creating the transcription files. The fifth step in training is to create an acoustic model. The acoustic model contains a statistical representation of the distinct sound that make up each syllable in the grammar. The syllables are represented in a statistical format which is used by the SRE to recognize an input in the testing phase.

The syllable segment sounds which come as an output from the previous phase syllable formation are used for training the system. Each distinct syllable is trained for 5 times to create enough training data for each distinct syllable. The number of utterances of each syllable can be increased to improve the accuracy.

D. *Testing Phase*

In the testing phase, a speech signal is given to the Character boundary phase and the character is classified in the next phase. The syllable is formed and given as input to HTK. The Speech recognition engine uses the acoustical model and finds the maximum likelihood using the Viterbi algorithm. The performance is measured using accuracy [13] with the formula given in Eq 11.

$$\text{Accuracy Percentage} = \frac{N - D - S - I}{N} \times 100 \qquad (11)$$

Where N is the number of test utterances, D, S and I are the number of deletions, substitution and insertion in the result.

VIII. CONCLUSION

The Syllable based Speech Recognition System was trained with 20 words related to agriculture discipline which was uttered for 2 times by 2 speakers. The connected words were segmented into words correctly. It was checked manually by hearing the segmented word sound in Matlab. The segmented word was again classified with each character's boundary identification using the Varied Length Maximum Likelihood Algorithm. The accuracy of the boundary detection was comparatively better in Varied Length Maximum Likelihood than by traditional Maximum Likelihood which is shown graphically in the Fig 8.

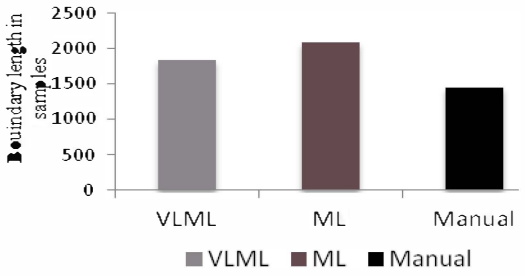

Fig 8: Graphical Representation of the Boundary points found for the character ி using VLML, ML and Manual Method

After the boundary was found using VLML, the characters are classified using GMM with 77% accuracy as few Tamil characters which has similar acoustic properties are substituted. After the classification of characters, the Syllables are identified using statistical approach and the accuracy was 70%. Accuracy was measured using the formula given in Eq 11.

REFERENCES

[1] Akila.A.Ganesh and Chandra Ravichandran, Syllable Based Continuous Speech Recognizer with Varied Length Maximum Likelihood Character Segmentation, IEEE Conference ICACCI 2013, pp.935-940, August 2013.

[2] L.R. Rabiner et al, Demisyllable Based Isolated Word Recognition System, IEEE Transactions on Acoustics, speech and Signal Processing, Vol ASSP-31 No 3, pp. 713-726, June 1983.

[3] Solomon T.A and Wolfgang.M, Syllable Based Speech Recognition for Amharic, 5th Workshop on Important Unresolved Matters, pp. 33-40, 2007.

[4] G.Lakshmi et al, Automatic Transcription of Continuous Speech into Syllable like units for Indian languages, Sadhana Vol 34, Part 2, pp. 221-233, April 2009.

[5] N.Kalyani and K.V.N.Sunitha, Syllable Analysis to build a dictation system in Telugu language, IJCSIS Vol.6, No.3, pp. 171-176, 2009

[6] Zhang Jiyong et al, Intra-Syllable dependent Phonetic Modeling for Chinese Speech Recognition, International Symposium on Chinese Spoken Language Processing, pp. 1-4, 2009

[7] Lakshmi A, Hema A Murthy, A Syllable Based Continuous Speech Recognizer for Tamil, Interspeech ICSLP, pp. 1878-1881, 2006.

[8] Ladefoged, Peter, Vowels and Consonants: An Introduction to the Sounds of Language, Maldern, MA: Blackwell, pp. 40, 2001.

[9] Anindya Sarkar and T.V. Sreenivas, Automatic Speech Segmentation Using Average Level Crossing Rate Information, Proc. of ICASSP, Philadelphia, pp-I-397-I-400, March 2005.

[10] Dr.E.Chandra and A.Akila, An Overview of Speech Recognition and Speech Synthesis Algorithms, Int.J.Computer Technology & Applications, Vol.3 (4), pp. 1426-1430, July 2012.

[11] L.Rabiner and B.H.Juang, Fundamental of Speech recognition, 1st ed., Pearson Education, 1993.

[12] R.Thangarajan and A.M.Natarajan, Syllable Based Continuous Speech Recognition for Tamil, South Asian Language Review, Vol XVIII NO.1, January 2008.

[13] A. Akila and E. Chandra, Isolated Tamil Word Speech Recognition System Using HTK, International Journal of Computer Science Research and Application, Vol.3, No 2, pp.30-38, , 2013.