Filter Bank Energy Based Malayalam Speech Segmentation and Recognition

Primekumar K.P *primekumar@yahoo.co.in*
Department of Computer Science
Cochin University of Science and Technology
Kochi, 682022, India

Sumam Mary Idiculla *sumam@cusat.ac.in*
Department of Computer Science
Cochin University of Science and Technology
Kochi, 682022, India

Abstract

Even though speech recognition technologies have made substantial progress, LVSR and vocabulary independent systems have not yet attained sufficient accuracy levels. For vocabulary independent speech recognition systems, segmentation of speech signal in to its constituent units such as phonemes, syllables is necessary. This paper presents a method of segmentation of spoken Malayalam words in to its constituent syllables and analyses the classification accuracy using PNN and HMM. Variations in peak filter bank energy is used for modeling criteria for segmentation. Mel Frequency Cepstral Coefficients (MFCC) and energy in each frame is used to extract the resultant feature vector in the feature extraction stage. A semi-automatic method is used for labeling the speech segments in the training phase. The system is trained using 30 samples of 26 syllables semi automatically segmented from fifty words collected from a male and female and tested on another set of fifty words containing 4720 syllables gives maximum accuracy of 74.7% and 66.77% for male and female respectively.

Keywords: Speech Segmentation, Filter Bank Energy, MFCC, Probabilistic Neural Networks, Hidden Markov Models.

1. INTRODUCTION

Nowadays many of the research related to speech recognition are focused on segmentation of speech signal in to its constituent units such as phonemes, syllables and sub syllable units. Segmentation of speech signals in to its constituent units is necessary in order to achieve sufficient accuracy levels for vocabulary independent and large vocabulary speech recognition systems. Coarticulation effects, wide variations in speaking styles and presence of background noise, makes the speech segmentation task more complex. There are different speech segmentation methods such as blind, supervised, hierarchical and Non-hierarchical based on the method of finding segmentation points. In this work blind segmentation of Malayalam words in to syllable like units based on peak filter bank energy variation is presented.

Most of the literatures relating to Malayalam speech recognition are based on small vocabulary and word level recognition using features such as LPC, MFCC and DWT [1,2]. Large vocabulary speaker independent speech recognition systems are mostly based on HMM [3, 4]. In order to achieve sufficient accuracy levels for vocabulary independent and LVSR systems, segmentation of speech signal in to its constituent units such as phonemes, syllables or sub-syllable units is necessary.

There exist different types of speech segmentation methods based on sub word units such as syllables, sub syllable, phonemes and graphemes [5,6,7]. In blind speech segmentation methods

variation in spectral properties such as spectral centroid, spectral flux, power spectral density etc is most commonly used for finding segmentation points [8][9]. Group delay based segmentation is presented in [10].Syllable based systems claims to have superior performance than phoneme or triphone based systems [11]. This paper presents a method of automatic segmentation of spoken Malayalam words in to its constituent syllables and analyses the classification accuracy of these units using PNN and HMM. Syllables in Malayalam may be any combination of (C)(C)(C)V(V)(C), where bracketed expressions are optional. In this work, we have selected words consisting of syllables in C, V and CV combination only. Selection of sub-word units such as syllables, sub syllables or phonemes is as per the requirement of the system. Malayalam is a syllable based language and the variation of spectral property of the speech signal such as peak filter bank energy is more predominant for transitions between syllables. So we have selected syllable based modeling for word recognition system. Segments obtained after segmentation process consists mostly syllable like units. Proposed segmentation algorithm uses variations in peak filter bank energy between the neighboring frames in order to find the segmentation points. This method neither requires previous knowledge about the language nor any type of training for segmentation. A total of 26 syllables selected from a set of fifty commonly used words from general conversation were considered in this work. Such systems can be extended to build vocabulary independent speech recognition systems.

Among different methods such as LPC, PLP, MFCC we have used MFCC features extraction method as it is dominating as the standard choices of feature extraction methods [12]. PNN and HMM is used as the classifier. During training each of the syllables constituting the word is segmented, labeled semi automatically and is given as the input to the system. During recognition automatic segmentation algorithm is used for the segmentation of speech signals in to individual units and these units are given as the input of recognition module. The speech recognition system consists of mainly five stages namely Pre-processing and frame blocking of speech signal, Segmentation, Feature extraction Training and recognition. This paper is arranged as seven sections. Section 2 deals with Pre-processing and frame blocking Section 3 explains the Segmentation Section 4 deals with MFCC feature extraction Section 5 deals with Training/ Recognition Section 6 gives the detailed performance analysis and results; conclusion is given in section 7. Block diagram of the overall system using HMM and PNN is shown in Figure1.

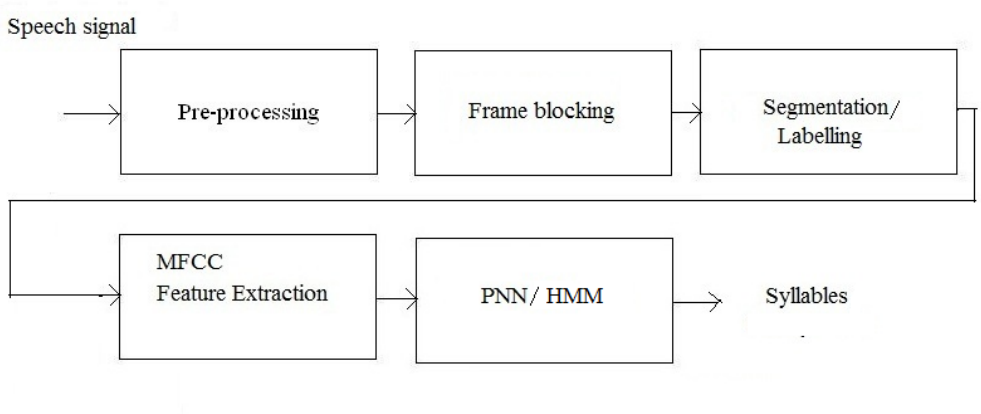

FIGURE 1: Block diagram of the system using PNN/HMM.

2. PREPROCESSING AND FRAME BLOCKING

Preprocessing is done in order to eliminate the noise and improve the quality of the acquired speech signal. The steps involved in the preprocessing are 1. Pre-emphasis, 2.End-point detection 3.Noise Suppression. Pre emphasis is to be performed in order to reduce the dynamic range of the speech signal. Fixed coefficient high pass filter is applied for achieving the same [15]. Speech signal corresponding to spoken word is separated from the background noise analyzing the zero crossing rate and energy in fixed time. If the zero crossing rate and logarithm of energy is greater than a threshold those positions belongs to the speech signal. The threshold

is set empirically after observing the surrounding noise. Spectral subtraction method is used for eliminating inherent noise present in the speech signal. Then consecutive frames having duration of 20ms in every 10ms is extracted. Each of the frame is multiplied by hamming window then normalized and is given as the input of speech segmentation stage.

3. SPEECH SEGMENTATION

Speech segmentation is the process of dividing the speech signal in to constituent basic units such as phonemes or syllables, words etc. The proposed automatic speech segmentation algorithm, divides speech signal in to syllables in Malayalam. Spectral properties of the speech based on STFT are most commonly used for segmentation. The frame is extracted using method described above and variations in peak filter bank energy of the neighboring frames are used for finding the segmentation points as described below.

3.1 Peak Filter Bank Energy

Peak filter bank energy variations can be used to locate probable segmentation points. For each filter 'l', the sum of product of DFT magnitude vector and corresponding triangular sub band filter channel gain function is calculated to get energy coefficient of the filter as shown below [14].

$$E_{l,n} = \sum_{k=1}^{M} (X_n[k] \cdot F_l[k])^2, \quad n = 1,2.....W,$$

$$\text{and } l = 1,2.....L$$

Where '$l = 1,2.....L$' is the filter number index**,** here 24 mel spaced triangular filter banks were used. **'** M ' is the number of frequency bins, 'n' is the frame index out of 'W' frames considered. '$E_{l,n}$' is the energy coefficient of the corresponding filter. 'X_n[k]' is the DFT magnitude vector corresponding to bin 'k' in DFT spectrum and 'F_l' is the corresponding filter channel gain function. Peak filterbank energy for 'n^{th}' frame, '$E_{p,n}$' is found as

$$E_{p,n} = \max(E_{l,n})_{\forall l}$$

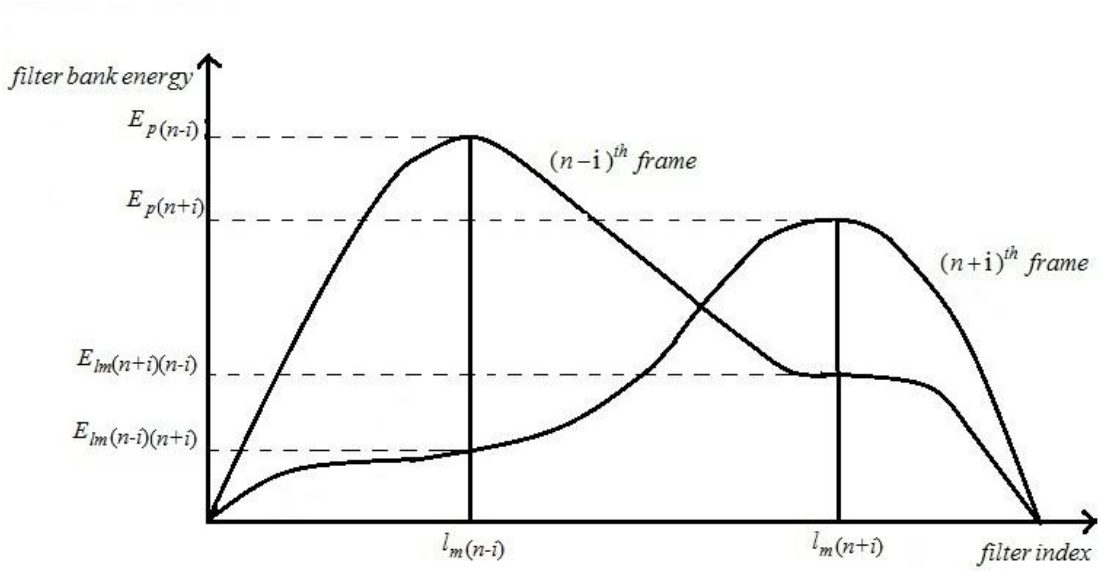

FIGURE2: Filter bank energy corresponding to $n-i^{th}$ and $n+i^{th}$ frame **.**

After extracting peak filterbank energies, function which is proportional to the difference between peak filter bank energies of the neighboring frames is calculated using following equation.

$$f = \sum_{i=1}^{N} \frac{\sqrt{(E_{p(n-i)} - E_{lm(n-i),(n+i)})^2 + (E_{p(n+i)} - E_{lm(n+i),(n-i)})^2}}{2i^2 \min(E_{n-i}, E_{n+i})}$$

Where '$E_{p(n-i)}$' and '$E_{p(n+i)}$' are the peak filter bank energies of $n-i^{th}$ and $n+i^{th}$ frame respectively. '$E_{lm(n-i),(n+i)}$' and '$E_{lm(n+i),(n-i)}$' are the filter bank energies of $lm(n-i)^{th}$ and $lm(n+i)^{th}$ filter for '$(n+i)^{th}$' and '$(n-i)^{th}$' frame respectively. Where '$lm(n+i)$' and '$lm(n-i)$' is the filterbank index corresponding to peak filter bank energy of $(n+i)^{th}$ and $(n-i)^{th}$ frame as shown in figure2. After extracting filter bank energy differences, a peak based method is used in order to detect segmentation points. Median filtering is applied on this resultant feature sequence and dominant positive peak points are selected as segmentation points as shown in Figure3. The segments having duration less than 30ms is merged with the adjacent one based on similarity. Those segments having peak values less than 15% of the maximum value of the speech signal is discarded.

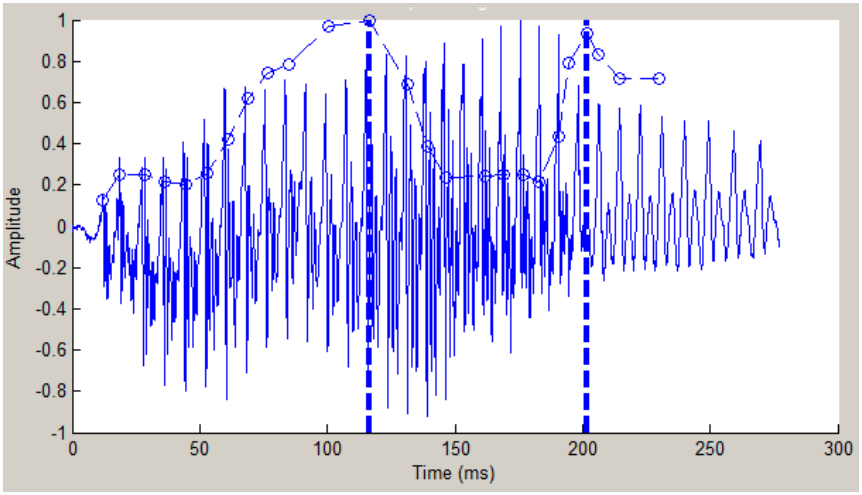

FIGURE 3: Segmentation result of word "LiR".

3.2 Labeling of Speech Segments

Labeling of the segments is required only in the training phase. Manual labeling of the syllables is very time consuming therefore a semi-automatic method is used for labeling the speech segments. After segmentation process each of the speech segments have to be labeled with corresponding syllable associated with them in order to prepare the training data set. In order to achieve this, MFCC features are extracted from each of the segments and only coefficients at the first, centre and last portion of the segment are concatenated to get a resultant vector of fixed dimension representing that segment. The resultant vectors of all of the segments are clustered using k-means algorithm and each of the clusters are assigned a symbol corresponding to that syllable by manual intervention. During this process misclassification in labeling is corrected manually. These set of labeled segments are used as the database for training.

4. FEATURE EXTRACTION

After segmentation and labeling, features in the speech segments are to be extracted so as to recognize individual syllables or sub syllables. Among different feature extraction methods,

MFCC is used in the feature extraction stage as it is dominating as the standard choices of feature extraction methods [15].

4.1 Mel Frequency Cepstral Coefficients (MFCC)

After segmentation process, a short-time DFT is performed on each windowed signal. Then mel-filter bank which simulates cohela of human auditory system is applied on the short time DFT. Then natural log of the filtered bank output is calculated and DCT of these logarithmic values gives MFCC coefficients [12]. Twenty four MFCC coefficients are considered for feature extraction. The energy of the signal in each window is also added to improve the performance of the system. First and second derivatives of these coefficients are concatenated to get the 39 dimensional feature vector. These are given as the input to the recognition module.

5. TRAINING/RECOGNITION

Performance of the system is analysed using the statistical classifier HMM and neural network based Probabilistic network

5.1 Training/Recognition using HMM

After feature extraction process, in the training phase, the features of the whole training dataset are clustered using K-means algorithm choosing Euclidian distance as the clustering criteria. Here 240 clusters are used and each cluster centre is assigned a unique symbol. Then the feature vectors are converted to a sequence of symbols by assigning each of the temporal features, the symbol corresponding to the nearest cluster centre. Then these sequences of symbols are fed as input to Discrete HMM, in which the observations are discrete symbols. Hidden Markov models are doubly stochastic process with an underlying stochastic process, corresponding to states that are hidden, but the state changes are observed through another set of stochastic process. It explores the relationships between consecutive observations of the pattern to be classified [4, 15]. The HMM is trained using Baum–Welch algorithm. Here we have used left-to-right HMM model. The number of states used is 3. In the recognition phase the feature vector is vector quantized and is converted in to a sequence of symbols. Then the decoding of this sequence of symbols is done using Viterbi algorithm.

5.2. Training/Recognition using PNN

Probabilistic Neural Networks is based on the theory of Bayesian Classification and these network estimates the probability density functions from the set of training samples [16]. The advantage of PNNs is that learning is several times faster than most of the other neural networks and HMM. They are inherently parallel in structure and guaranteed to converge to an optimal classifier as the size of the training set increases. The principle behind PNNs is to create the probability density functions of each of the classes from the training samples and during classification these probability densities are used along with Bayesian decision rule to find the most probable class corresponding to the unknown vector. The probability density functions are estimated using

$$f_i(x) = \frac{1}{(2\Pi)^{d/2}\,\sigma^d N} \sum_{j=1}^{N} e\left(\frac{-(x - x_{ij})^T (x - x_{ij})}{2\sigma^2} \right)$$

Where 'i' is the class number, 'j','the pattern number, 'x_{ij}', jth vector from class 'i', 'x' is the test vector, 'N' is the number of training vectors in class 'i', 'd' is the dimension of vector 'x'and 'σ'is the smoothing parameter.

Three separate PNNs are trained using features extracted from frames having 15, 20 and 25ms duration. MFCC features extracted from a set of two adjacent frames with in a segment are concatenated and are given as the input of PNNs. During recognition each of the two frames are classified by these PNNs and a voting scheme is used to find the syllable corresponding to that

segment. Finally deductions of the following form are done to get the final syllable from individual syllable units.

d [+ : = dl
(C) + (V) = (CV)

6. PERFORMANCE ANALYSIS

The Accuracy of the system was analyzed in speaker dependent mode. The words used were common words in Malayalam language and is collected from a male and female person. Training set consisting of 30 samples of 26 syllables and semi automatically segmented from 50 Malayalam words collected from a male and female person. Test set is entirely different set consisting of 20 samples of 50 words having a total of 4720 syllables, collected from the same male and female person. In the speaker dependent mode, maximum accuracy of 74.7% and 66.77% is obtained for male and female respectively. The accuracy obtained is comparable with spectral transition measure based speech segmentation systems. There is no major work reported on the segmentation of Malayalam spoken words in to syllables. Accuracy (*A*) is found using the following equation.

$$A = \frac{N - D - S - I}{N}$$

Where 'I' is the number of insertions, 'D' is number of deletions and 'S' is the number of substitutions and 'N' is the total number of syllables.

Number of	Male	Female
Syllables (N)	4720	4720
Insertions (I)	263	346
Deletions (D)	508	610
Substitutions (S)	423	690

TABLE1. Performance parameters using HMM

Number of	Male	Female
Syllables (N)	4720	4720
Insertions (I)	400	421
Deletions (D)	406	677
Substitutions (S)	453	470

TABLE2: Performance Parameters using PNN.

7. CONCLUSION

This paper analyses the performance of segmentation of Malayalam spoken words in to its constituent syllables using variations in peak filter bank energy. Some of the insertions are due to the noise present in the acquired speech signal. Often substitutions occurred between similarly sounding syllables such as short and long vowels and in places of incorrect segmentation. It is found that for small vocabulary speaker dependent applications the accuracy of PNN is slightly better than that of HMM. Training time required for PNN is very much lesser than that of HMM. The recognition accuracy of the recognition module can be improved by increasing the number of training samples. Peak filter bank energy along with other spectral features that can distinguish between syllables, in hierarchical manner can improve the performance of the system.

8. REFERENCES

[1] Krishnan, V.R ; V. Jayakumar A, Anto P.B (2008) ,"Speech Recognition of isolated Malayalam words using wavelet features and Artificial Neural network, *Fourth IEEE*

International symposium on Electronic Design, Test and Applications, 2008 volume Issue 23-25 Jan, 2008. Page(s) 240 – 243.

[2] Cinikurian and Kannan Balakrishnan, "Continuous Speech Recognition System for Malayalam Language using PLP Cepstral Coefficient, IJCBR, Vol3, Issue1, Jan2012.

[3] S. Young, "A review of large vocabulary continues speech recognition,"Proc.IEEE Sig. Processing. Mag. September1996, 45-57

[4] Lawrence R. Rabiner. "A tutorial on HMMs and selected applications in speech recognition". Proceedings of IEEE, Vol77, No2, Feb1989.

[5] Rudi Villing, Joseph Timoney, Tomas Ward and John Costello, Automatic Blind Syllable Segmentation for Continuous Speech, ISSC 2004, Belfast.

[6] K.F. Chow, Tan Lee and P.C Ching, "Sub syllable Acoustic Modelling for Cantonese Speech Recognition"

[7] Kaichiro Hatazaki, Yasuhiro Komori, Takeshi Kawabata and Kiyohiro Shikano, "Phoneme segmentation using spectrogram reading knowledge", IEEE,1989.

[8] Md. Mijanur Rahman, Md Al-Amin Bhuiyan, "Continuous Bangla Speech Segmentation using Short-term Speech Features Extraction Approaches",IJACSA, Vol3, No11, 2012.

[9] Dzmitry Pekar and Siarhei Tsikhanenka, "Speech segmentation algorithm based on an analysis of the normalized Power Spectral Density", 2010

[10] Prasad, V.K nagarajan T and Murthy H.A "Automatic segmentation of continuous speech using minimum phase group delay functions", Vol.42, Apr2004, pp 1883-1886.

[11] Aravind Ganapathiraju, Jonathan Hamaker, Joseph Picone, Mark Ordowski and George R Dddington, "Syllable –Based Large Vocabulary Continuous Speech Recognition", IEEE Transactions on Speech and Audio Processing, Vol9, No4, May2001.

[12] Fu-Hua, Richard M Stern, Xuedong Huang,Alejandro Acero, "Efficient cepstral normalization for robust speech recognition, human language technology", 1993

[13] Sergios Theodoridis and Konstantinos Koutroumbas, "Pattern Recognition", Fourth Edition

[14] Marko Kos, Matej Grasic, Zdravko Kacic, " Online Speech/Music Segmentation Based on the Variance Mean of Filter Bank Energy" ,2009

[15] Lawrence R. Rabiner , Biing Hwang Juans."Fundamentals of speech recognition", Pearson Education.

[16] D.f specht, Probabilistic Neural Networks, neural Networks, Vol3,pp109- 118,1990.