ORIENTAL JOURNAL OF
COMPUTER SCIENCE & TECHNOLOGY
An International Open Free Access, Peer Reviewed Research Journal
Published By: **Oriental Scientific Publishing Co., India.**
www.computerscijournal.org

ISSN: 0974-6471
March 2017,
Vol. 10, No. (1):
Pgs. 114-119

Phoneme Segmentation of Tamil Speech Signals Using Spectral Transition Measure

K GEETHA* and DR. R. VADIVEL

Department of Computer Science, D.J. Academy for Managerial Excellence,
Coimbatore, 641032, India.
Department of Information Technology, Bharathiar University, India.
*Corresponding author E-mail: email: geethakab@gmail.com

http://dx.doi.org/10.13005/ojcst/10.01.15

(Received: February 22, 2017; Accepted: March 02, 2017)

ABSTRACT

Process of identifying the end points of the acoustic units of the speech signal is called speech segmentation. Speech recognition systems can be designed using sub-word unit like phoneme. A Phoneme is the smallest unit of the language. It is context dependent and tedious to find the boundary. Automated phoneme segmentation is carried in researches using Short term Energy, Convex hull, Formant, Spectral Transition Measure (STM), Group Delay Functions, Bayesian Information Criterion, etc. In this research work, STM is used to find the phoneme boundary of Tamil speech utterances. Tamil spoken word dataset was prepared with 30 words uttered by 4 native speakers with a high quality microphone. The performance of the segmentation is analysed and results are presented.

Keywords: Speech Recognition, Speech Segmentation, Spectral Transition Measure (STM), Phoneme Segmentation.

INTRODUCTION

In Natural Languages, speech is considered as the sequential link of phonemes. The automatic segmentation of speech using only the phoneme sequence is an important task, especially if manually pre-segmented sentences are not available for training. The availability of segmented speech databases is useful for many purposes, mainly for the training of phoneme-based speech recognizers[1]. Such an automatic segmentation can be used as the primary input data to train other more powerful systems like those based on Hidden Markov Models or Artificial Neural Networks[2].

In linguistics, a phoneme is defined as the minimal information bearing distinct unit[3]. In the acoustic realization, it is uncertain to define phoneme boundaries. In the short-time representation, the speech signal is considered stationary and the voiced segments quasi-periodic. In statistical phoneme segmentation, it is assumed that the properties of the speech signal change when transition occurs from one phoneme to the next phoneme. But in reality, the transitions occur smoothly due to the influence of the adjacent phonemes.

Automatic Speech Segmentation (ASS) methods can be classified into two categories. In the first category the basic acoustic unit that the Automatic Speech Recognition (ASR) can handle will be kept as transcription of the speech. Segmentation algorithms can use this transcription to find the sub word boundaries and the number of subunits in it. In the second case, there will be no linguistic knowledge prior. This type of Speech segmentation algorithms also called as blind segmentation algorithms in which the number of sub word units and the boundaries found based on the acoustic cues only[4].

In Modern ASR approaches, the concatenation principle is used to represent words by its successive phonemes. Since phonemes are context dependent, context dependent model such as triphone, demiphone are also proposed in which the fundamental unit is phoneme and the words are represented in the pronunciation lexicon as concatenations of phonemes[5].

Related Work

Sharma and Mammone[6] designed a Level Building Dynamic Programming (LBDP) based speech segmentation, a dynamic programming based algorithm to optimally locate the sub-word boundaries by minimizing distortion metric. They have proposed a novel blind speech segmentation procedure to determine the optimal number of sub-word units present in the given speech sample as well as the boundary locations based on acoustic cues, without any linguistic knowledge.

Odette Scharenborg[7] *et al*. investigated the fundamental problems in unsupervised segmentation algorithms. The authors have compared phoneme segment obtained using only the acoustic information derived from the signal with a reference segment created by human annotators. From the experiments, it is concluded that the acoustic change is a fairly good indicator of segment boundaries and proved that the errors are related to segment duration, sequences of similar segments, and inherently dynamic phones. To improve the unsupervised automatic speech segmentation, instead of using one-stage bottom-up segmentation method, it is suggested to propose two-stage segmentation methods which uses both bottom –up data extracted from speech signal and automatically derived top-down information.

Zió³ko, Bartosz[8], *et al*. proposed a new phoneme segmentation method based on the analysis of Discrete Wavelet Transform(DWT) spectra. The values of power envelopes and their first derivatives for six frequency sub bands were used segmentation of Polish Language[9]. Specific scenarios which suits for phoneme boundaries are searched first and then the discrete times with such events are recorded and graded using a distribution-like event function. This event function represents the change of the energy distribution in the frequency domain. Finally, the decision on localization of boundaries is taken from the analysis of the event function. Boundaries are extracted using information from all sub bands. This method was tested with a small set of Polish hand segmented words and tested on another large corpus containing 16,425 utterances and recall and precision measure used to measure the quality of speech segmentation with F-score equal to 72.49%.

Kuo *et. al*[10] presented an improved HMM/SVM method for a two stage phoneme segmentation framework. The first stage performs hidden Markov model (HMM) forced alignment according to the Minimum Boundary Error (MBE) criterion and the second stage uses the support vector machine (SVM) method to refine the hypothesized phoneme boundaries derived by HMM-based forced alignment. The designed to align phoneme boundary based on MBE-trained HMMs and explicit phoneme duration models. They tested their method with TIMIT database and MATBN Mandarin Chinese database.

Mousmita Sarma *et.al*[11] described an Artificial Neural Network (ANN) based algorithm for the segmentation and recognition of the vowel phonemes of Assamese language from the words containing vowels. Self-Organizing Map (SOM) used to train and segmentation was done to segment the word into its constituent phonemes.

Probabilistic Neural Network (PNN) trained with clean vowel phonemes was used to recognize the vowel segment. The experimental speech samples were recorded from five female speakers and five male speakers. In the authors proposed method, the first formant frequency of all the Assamese vowels was predetermined by estimating pole or formant location from the linear prediction (LP) model of the vocal tract. The proposed algorithm showed a high recognition performance in comparison to the conventional Discrete Wavelet Transform (DWT) based segmentation**.**

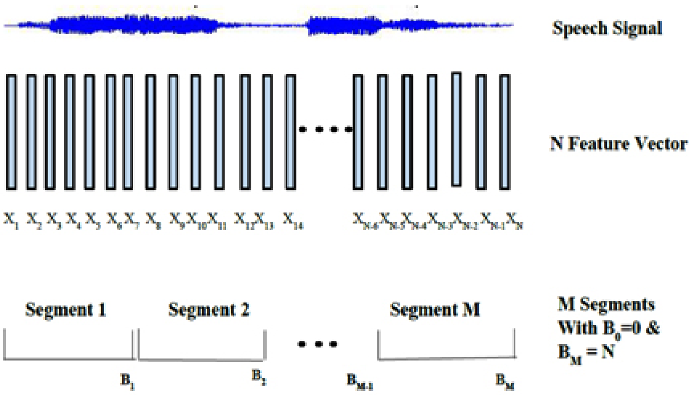

Fig. 1: Segmentation of $\{X_i\}^{N}_{i=1}$, into M segments

Mathematical Formulation of Segmentation

The problem of speech segmentation is described in Fig.1. Let $\{X_i\}_{i=1}^{N}$, denote the sequence of mel cepstrum vectors calculated for each frame of every Tamil uttered word from the data set, where N is the number of speech frames and X_i is p dimensional parameter vector at frame 'i'. The objective of the segmentation problem is to divide the sequence X into M non-overlapping consecutive segments where each sub sequence corresponds to a phoneme. Let the boundaries of the segment be denoted by the sequence of integers $\{B_i\}_{i=1}^{M}$. The i^{th} segment starts at frame B_{i-1} +1 and ends at frame B_i; where B_0=0 and B_M=N.

METHOD

Compared to generative methods based on HMMs, phonemic segmentation methods based on spectral distortion measures are independent of linguistic constraints and computationally inexpensive[12,13]. So, an attempt to segment the Tamil speech utterances into phoneme segments is made using spectral transition measure and the outline of the method is given in Fig. 2.

Data Set

Tamil speech utterances consisting of 30 unique Tamil words constituting 172 phonemes uttered by 4 native speakers are recorded with the help of a unidirectional microphone and considered as data set. Data are recorded using a recording tool audacity in a normal room with

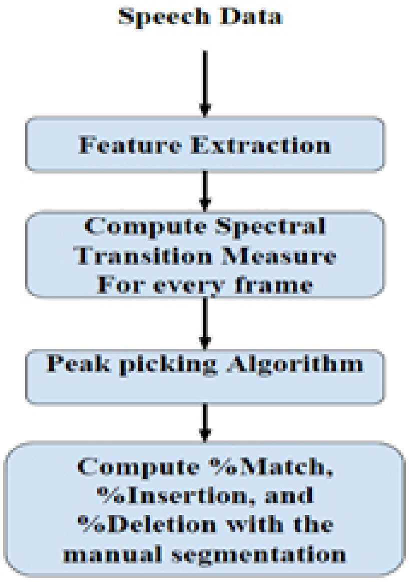
Speech Data
Feature Extraction
Compute Spectral Transition Measure For every frame
Peak picking Algorithm
Compute %Match, %Insertion, and %Deletion with the manual segmentation

Fig. 2: Outline of the Phoneme Segmentation using STM

minimum external noise. The sampling rate used for recording is 16 kHz. The description about the data used is also given in Table 1.

Preprocessing and Feature Extraction

The signal, sampled at 16 kHz, is decomposed into a sequence of overlapping frames. The frame size of 25 ms and 10ms frame shift were used for the segmentation approach considered. The input speech data are pre emphasized with co-efficient of 0.97 using a first order digital filter. The samples are weighted by a Hamming window for avoiding spectral distortions.

The windowed frame obtained after hamming is used to extract the twelve Mel Frequency Cepstral Coefficients (MFCC). Usually the zero order coefficient that represent the total energy and for our experiment vector excluding the total energy is used for further process. The short-time fourier transform analysis is then performed to compute the magnitude spectrum. Filter bank design with triangular filters uniformly spaced on the mel scale between 300 Hz to 3400 Hz as lower and upper frequency limits is followed. The filter bank is applied to the magnitude spectrum values to produce Filter Bank Energies (FBEs) 20 per frame. Log-compressed FBEs are then de-correlated using the Discrete Cosine Transform (DCT) to produce cepstral coefficients. The co-efficients obtained are then rescaled to have the similar magnitudes achieved through liftering with the value of 22 as L. The steps involved in the MFCC feature extraction are shown in Fig. 3.

Table 1: Specification used in Creating Dataset

Description	Feature
Language	Tamil
Speech type	Text independent read speech
Approach	Unsupervised
Recording Conditions	Room Environment
Number of Speaker	2 male speakers
2 female speakers	
Age group	25-30
Region	Native
Number of Words used for testing	30

Spectral Transition Measure (STM)

Mitchell *et al.*[13] introduced Spectral Variation Function (SVF), calculated as the angle between two normalized cepstral vectors for phoneme segmentation. The spectral transition measure employed in this study was the same as that proposed in[14]. Dusan and Rabiner[14] detected the maximum spectral transition frames as phoneme boundaries, where spectral transition represents the magnitude of the spectral rate of change. This spectral transition measure (STM), at frame m, can be computed as a mean squared value as in Eq. 1.

$$\mathrm{STM(m)} = (\textstyle\sum_{i=1}^{D} a_i^2(m))/D) \qquad \ldots(1)$$

where D is the dimension of the spectral feature vector which is 12 coefficients in this experiment without gain term. The regression coefficient or the rate of change of the spectral feature is computed using eqn. 2.

$$a_i(m) = \left(\textstyle\sum_{n=-I}^{I} \mathrm{MFCC}(n+m) * n / (\textstyle\sum_{n=-I}^{I} n^2\right) \qquad \ldots(2)$$

where n represents the frame index and I defines the number of frames to be included in both side of the current frame to form a symmetric window for computing regression coefficients. In this experiment, I with 1,2,3 is used.

Phoneme Boundary Detection

Phoneme boundaries detection can be defined as a bi-step process: a peak picking method and a post-processing method for removing local boundaries. In the proposed method, all the peaks in the spectral transition measure are computed for every frame. Then, from the STM values of all frames, the locations of all peaks which proceeded with negative region are identified. They are referred

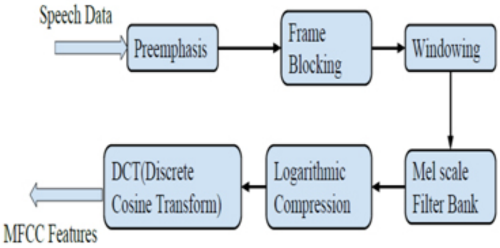

Fig. 3: MFCC Features Extraction

Table 2: Performance of the Speech Segmentation using STM

No of words	Number of Actual boundaries	Frame Tolerance	% Match	% Insertion
30	688	±10ms	72.1	27.9
		±20ms	75.8	24.2
		±30ms	79.4	20.6

as valleys. Peak picking method is applied to select prominent peaks with deep valleys. Peaks which are closer to other peaks within 40ms to 60ms are considered for elimination. After choosing the peaks, the frames in which the peaks occur are identified and used to perform phoneme segmentation. Let, M be the number of segments and the boundaries be $\{B_i\}_{i=1}^{M}$. Then the M-1 most significant peaks are to be obtained i.e.$\{B_i\}_{i=1}^{M-1}$ and $B_M = N$.

Experiment and Discussions

Segmentation method using the spectral transition measure is experimented with the frame tolerance with ±(10ms<"30ms). The dataset developed contains 688 boundaries excluding BM. The optimal tolerance is obtained as the value for which the peak-picking procedure gives the number of segments same as that of the actual number of segments in the manual segmentation of Tamil word considered. The window length parameter L is assigned with values of 1, 2 and 3 and optimized for a maximum segmentation match with manually segmented data and the number of boundaries detected within a tolerance window ±(10ms to 30ms). The automatic segment boundaries with parameter value of l as 2 and the tolerance ±(10ms to 30ms) is shown in Table 2. The performance of phoneme segmentation with respect to the manually segmented Tamil utterances using three measures: percentage of match (%M), percentage of insertion (%I) and percentage of deletion (%D) within a tolerance window ±(10ms to 30ms).

$$\%\text{Match} = \frac{\text{number of boundaries detected correctly}}{\text{total number of ground truth boundaries}} * 100$$

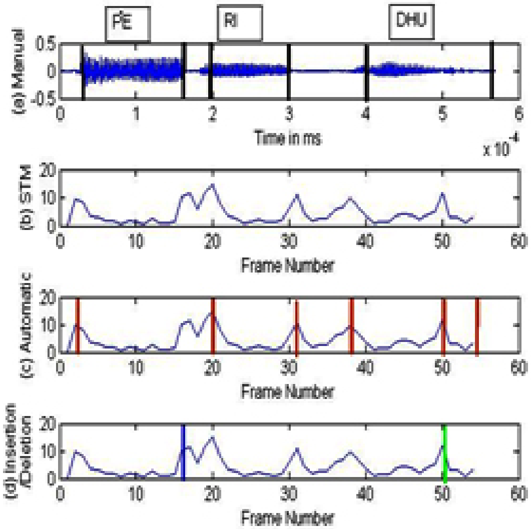

Fig. 4: (a) Speech signal of sample Tamil word 'PERIDHU' with manual boundary (b) Spectral Transition Measure of the word (c) Boundaries found using automated STM (d) Boundary found as Insertion and Deletion.

Percentage of Insertion(%I) gives the percentage of segments obtained by the automatic segmentation without corresponding manual segment within the tolerance window. Percentage of Deletion (%D) gives the percentage of segments obtained by the manual method without any corresponding automatically segmented boundary within the tolerance window.

CONCLUSION

In this paper, the phoneme boundaries of Tamil speech utterances are found spectral transition measure. The performance of the segmentation is analysed in terms of percentage of matches with the manual boundary. It is suggested to have better alignment techniques in future to get the better results. Furthermore larger linguistic units of language than phoneme may also be proposed.

REFERENCES

1. Gómez, Jon Ander, and María José Castro, Automatic segmentation of speech at the phonetic level. Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR). Springer, Berlin Heidelberg, 2002.
2. R. Rabiner, and B. H. Juang, *Fundamentals of Speech Recognition*, (Prentice-Hall International, 1993).
3. Thangarajan, R.,Natarajan, A.M. and Selvam, M., Word and Triphone Based Approach in Continuous Speech Recognition for Tamil Language, *WSEAS Transaction on Signal Procesing*,ISSN:1790-5022,**4**(3) , pp 76-85, 2008.
4. Qiao, Yu, and Nobuaki Minematsu, Metric learning for unsupervised phoneme segmentation, *INTERSPEECH,* 2008.
5. Dusan, Sorin, and Lawrence R. Rabiner,On integrating insights from human speech perception into automatic speech recognition, *INTERSPEEC,* 2005.
6. Sharma, Manish, and Richard Mammone. "Blind" speech segmentation: automatic segmentation of speech without linguistic knowledge." *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*. **2**. IEEE, 1996.
7. Scharenborg, Odette, Vincent Wan, and Mirjam Ernestus, Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries, *The Journal of the Acoustical Society of America*, **127**(2): 1084-1095, (2010).
8. B. Zio³ko, S. Manandhar, and R. C. Wilson, Phoneme segmentation of speech, *In ´ Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06*), **4**, pages 282–285, 2006.
9. Zió³ko, Bartosz, *et al*, Phoneme segmentation based on wavelet spectra analysis, *Archives of Acoustics,* **36**(1): 29-47, (2011).
10. Kuo, Jen-Wei, Hung-Yi Lo, and Hsin-Min Wang, Improved HMM/SVM methods for automatic phoneme segmentation, *INTERSPEECH,* 2007.
11. Sarma, Mousmita, and Kandarpa Kumar Sarma. "Segmentation and classification of vowel phonemes of assamese speech using a hybrid neural framework." *Applied Computational Intelligence and Soft Computing* 2012: **28**, (2012).
12. Almpanidis, George, and Constantine Kotropoulos, Phonemic segmentation using the generalised Gamma distribution and small sample Bayesian information criterion, *Speech Communication* **50**(1), 38-55, (2008).
13. Qiao, Yu, Dean Luo, and Nobuaki Minematsu, A study on unsupervised phoneme segmentation and its application to automatic evaluation of shadowed utterances, *Technical report*, 2012.
14. Dusan, Sorin, and Lawrence R. Rabiner, On the relation between maximum spectral transition positions and phone boundaries, INTERSPEECH. 2006.