ISSN 0146-4116, Automatic Control and Computer Sciences, 2018, Vol. 52, No. 1, pp. 25–31. © Allerton Press, Inc., 2018.

Syllable Segmentation of Tamil Speech Signals Using Vowel Onset Point and Spectral Transition Measure[1]

K. Geetha[a], * and R. Vadivel[b]

[a] *Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, 641046 India*
[b] *Department of Information Technology, Bharathiar University, Coimbatore, Tamilnadu, 641046 India*
**e-mail: geethakab@gmail.com*

Received November 28, 2016; in final form, November 28, 2017

Abstract—Segmentation plays vital role in speech recognition systems. An automatic segmentation of Tamil speech into syllable has been carried out using Vowel Onset Point (VOP) and Spectral Transition Measure (STM). VOP is a phonetic event used to identify the beginning point of the vowel in speech signals. Spectral Transition Measure is performed to find the significant spectral changes in speech utterances. The performance of the proposed syllable segmentation method is measured corresponding to manual segmentation and compared with the exiting syllable method using VOP and Vowel Offset Point (VOF). The result of the experiments shows the effectiveness of the proposed system.

Keywords: syllable segmentation, spectral transition measure, vowel onset point, vowel offset point, linear prediction cepstral coefficients

DOI: 10.3103/S0146411618010042

1. INTRODUCTION

Segmentation in speech processing is the process of dividing the continuous speech signal into a discrete non-overlapping sub word unit i.e. finding the sub word unit boundaries. Segmentation of speech signals into basic linguistic units has enormous applications in the fields of recognition, synthesis, labeling and transcription and coding. Speech segmentation is an important phase in continuous speech recognition, since it reduces the search space. In larger vocabulary tasks, automatic segmentation of speech utterances is preferable than manual segmentation which is a tedious and time consuming one. Most of the work carried out to segment the speech into units like word, sub word, syllable and phonemes using various segmentation approaches. Pronunciation duration is an important problem for speech segmentation, especially in languages like Tamil, which has vowels in both short and long forms.

A model [1] was designed to detect VOPs using Auto Associative Neural Network (AANN). The authors used five layered AANN with compression layer in middle and explored the distribution capturing of feature vectors. Similar work with little variation has been tried using AANN model to hypothesis the consonant and vowel regions and the detect VOPs in continuous speech. In their work, Consonant-Vowel (CV) unit is considered as a sub-word unit. As region around VOP contains most of the information of CV utterances, after detecting VOP, they segmented fixed duration segment around VOP and classified frequently occurring CV units in three Indian languages Hindi, Telugu and Tamil.

An effective speech segmentation method to split speech signals into CV and Consonant-Vowel-Consonant (CVC) unit of Tamil language is designed based the linguistically unconstraint approach which lead to design an efficient speech recognition/ synthesis model. In the proposed work, Hilbert envelope of the LP residual of the speech signal is calculated to identify the VOPs. Spectral transition measure speech utterances are made to identify the maximum spectral transition positions in the signal and the duration of the linguistic units analysed and found in the literature are considered to set the frame tolerance value.

This paper is organized as follows: In Section 2 significant features and works in speech segmentation of Tamil languages. Section 3 describes speech segmentation and its mathematical representation. Section 4 presents proposed framework with a clear sketch of the various stages in the work. Section 5 presents experimental results and comparative analysis of the proposed method with baseline method Vowel Onset

[1] The article is published in the original.

Table 1. Pronunciation duration of Tamil linguistic units

Linguistic unit	Time length of pronunciation in "maathirai"	Time length of pronunciation in milliseconds
Short vowel	1	250
Long vowel	2	500
Diphthong	1½	375
Consonants	½	125
Consonants + short vowel	1	250
Consonants + long vowel	2	500
Consonants + diphthong	1½	375

Table 2. Tamil linguistic units of the Sentence "அவள் படித்துக் கொண்டிருக்கிறாள்"

Linguistic unit	Segmented units
Word	அவள், படித்துக், கொண்டிருக்கிறாள்
Phoneme	அ, வ், அ, ள், ப், அ, ட் இ, த் த், உ, க், க், ஒ, ண், ட், இ, உ, ர், க், க், இ, ற், ஆ, ள்
Grapheme/character	அ, வ, ள், ப, டி, த், து, க், கொ, ண், டி, ரு, க், கி, றா, ள
Syllable	அ, வள், ப, டித், துக், கொண், டி, ருக், கி, றாள்

Point and Vowel Offset Point (VOP + VOF). Section 6 provides the conclusion and the scope for future work.

2. FEATURES AND RELATED WORK IN TAMIL LANGUAGE

Tamil is an ancient Indian language spoken widely in the southern state of India, Tamil Nadu. Tamil is a syllabic language. There are 18 consonants and 12 vowels of different categories are present in Tamil Language. The syllables are derived from a combination of consonants and vowels, a total of 216 different characters with unique written form. As the syllables are derived from short and long vowels, short and long syllable is present [2]. Time length of pronunciation of short and long vowel/syllable is measured in maatthirai (1 maatthirai = 0.25 s) and is given in Table 1. Duration of the character has a significant role in finding boundaries [3].

Tamil speech, similar to other extended speech, can be segmented into various linguistic units like word, phoneme, grapheme, syllable, prosodic syllable and morpheme. Researchers have tried methods to automate speech segmentation in Tamil Language. Tamil speech utterances can be segmented into various linguistic units like word, phoneme, grapheme and syllable is shown in Table 2. Segmentation of sentence into different linguistic units for a sample Tamil sentence "அவள் படித்துக் கொண்டிருக்கிறாள்" is tabulated.

Speech recognition for Tamil language has been carried out using prosodic syllables as sub-word units [4]. Two methodologies are proposed: first methodology, modeling syllable as an acoustic unit and for which Context Independent (CI) syllable models are trained and tested. The second methodology proposes integration of syllable information in the conventional tri-phone or Context Dependent (CD) phone modeling. The results were compared with two baseline models of a medium vocabulary CD phone based acoustic model and a small vocabulary CI word model. Results show that the syllable-based recognition produces better results.

An approach for segmenting continuous speech into smaller speech units was proposed and each speech unit was classified as consonant/vowel using the Formant frequencies in Tamil broadcast data [5].

Panda et al. [6] analyzed the performance of the syllable centric segmentation in three Indian languages Hindi, Bengali and Odia using vowel onset point and vowel offset point. Experiments are carried out to compare the method with the group delay based segmentation technique along with the manual

Segment	1					2						M	
Frame	1	2	...	B_1		...	B_2		...	B_{M-1}		...	N

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

segmentation technique. The results showed the effectiveness of the syllable centric segmentation technique in segmenting the syllable units from the original speech samples compared to the existing techniques. This method is used as the baseline method to analyse the proposed method.

3. SPEECH SEGMENTATION

Segmentation of speech signal or identification of the sub word unit boundary is a fundamental and crucial task since it has many important applications in speech and audio processing. Most of the speech recognition systems used phoneme as the basic unit for modeling. Since automatic phoneme segmentation is context dependent, to achieve higher accuracy, higher levels of speech units than phoneme, such as tri-phone, syllable are tried in many researches.

3.1. Mathematical Formulation of Segmentation

The problem of speech segmentation is described in Fig. 1. Let a speech utterance of S samples represented by $\{X_i\}_{i=1}^{N}$, is the vector sequence of N speech frames, where X_i is p dimensional parameter vector at frame 'i'. The segmentation problem is to find M consecutive segments in the N frame sequence. Let the boundaries of the segment be denoted by the sequence of integers $\{B_i\}_{i=1}^{M}$. The ith segment starts at frame $B_{i-1}+1$ and ends at frame B_i , where $B_0 = 0$ and $B_M = N$.

4. PROPOSED WORK

A detailed review of related research work in speech segmentation in various languages reveals that there are few works carried out in Tamil language and motivates for the proposed work. In this work, text independent unsupervised approach for syllable boundary identification is carried out using Tamil clean speech utterances. The outline of the proposed syllable segmentation method is presented in Fig. 2.

4.1. Preprocessing

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. Then, the signal is decomposed into a sequence of overlapping frames. The frame size of 20 ms and 10 ms frame shift were used for the segmentation approach considered. The resulting windowed frame is applied to auto correlation analysis, in which each frame of the windowed signal is auto correlated and provides $p + 1$ auto correlations for each frame, where p is the order of LPC analysis [7]. In LPC analysis step, each frame of $p + 1$ auto correlations are converted into LPC parameter set. Both LP residual and Linear Predictive Cepstral Coefficients (LPCC) are computed from the LPC. LP residual is used in computing the vowel onset point. LPCC features are used to find spectral changes and the process of extracting LPCC is depicted in Fig. 3.

4.2. Vowel Onset Point

Vowel onset point is a point at which the consonant region ends and vowel region begin in a CV utterance. Utterances of CV units consist of different speech production events like closure, burst, aspiration, transition and vowel. All CV units have a distinct VOP in their production, which is the significant property useful in CV unit segmentation or classification.

Different methods are found in the literature to find the VOP by utilizing various features and their combinations [8, 9]. The combined evidence plot of source excitation, spectral peak and modulation spectrum has been used to find VOP in the proposed method [10].

All utterances are processed by Linear Prediction (LP) analysis to extract the LP residual, which carries the excitation information. In LP analysis, the dependencies among adjacent samples are estimated and then removed from the speech signal to obtain the residual signal.

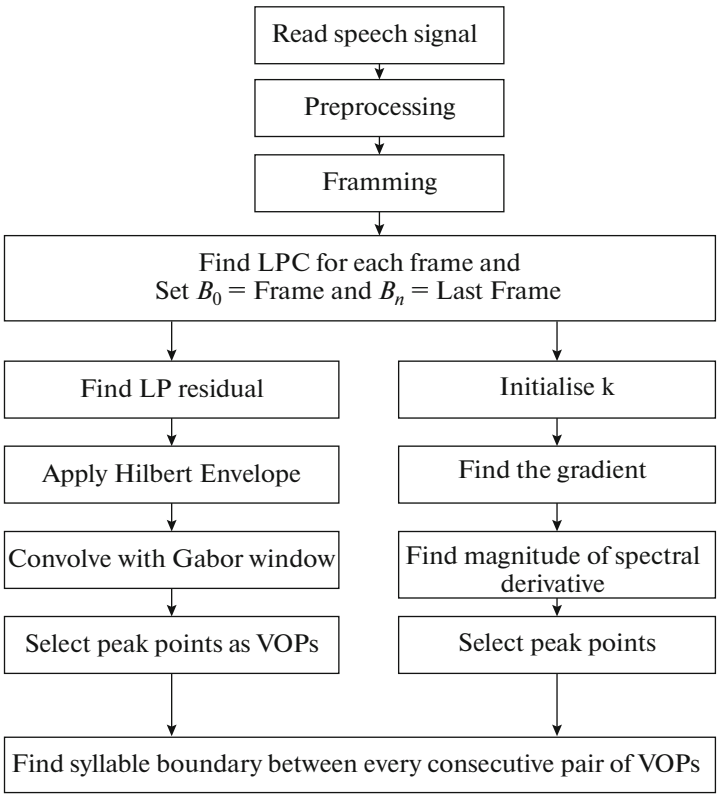

Fig. 2. Outline of the proposed syllable segmentation.

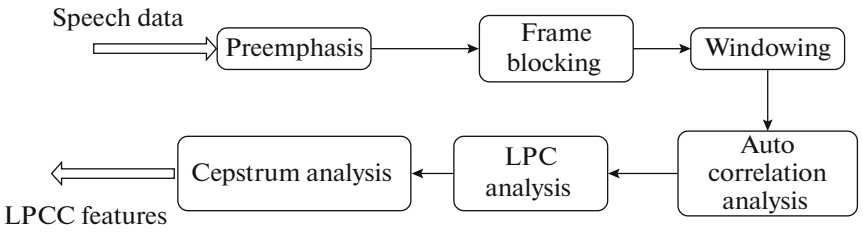

Fig. 3. LPCC features extraction.

The prediction of current sample as a linear combination of past p samples from the basis of linear predictive analysis when p is the order of prediction. The predicted sample $\hat{s}(n)$ can be written as

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k\, s(n-k), \tag{1}$$

where a_k are the linear prediction coefficients. $S(n)$ is the windowed speech sequence obtained by multiplying short time speech frame with hamming window. The difference between the actual sample $s(n)$ and the predicted sample $\hat{s}(n)$is the prediction error $e(n)$, which can be written as

$$e(n) = s(n) - \hat{s}(n). \tag{2}$$

Equation (2) can be written as

$$e(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k). \tag{3}$$

The LPC captures information about the vocal tract system. The information about the excitation source can be obtained from the speech signal by passing the signal through the inverse filter [11], the resulting signal is termed as LP residual $e(n)$.

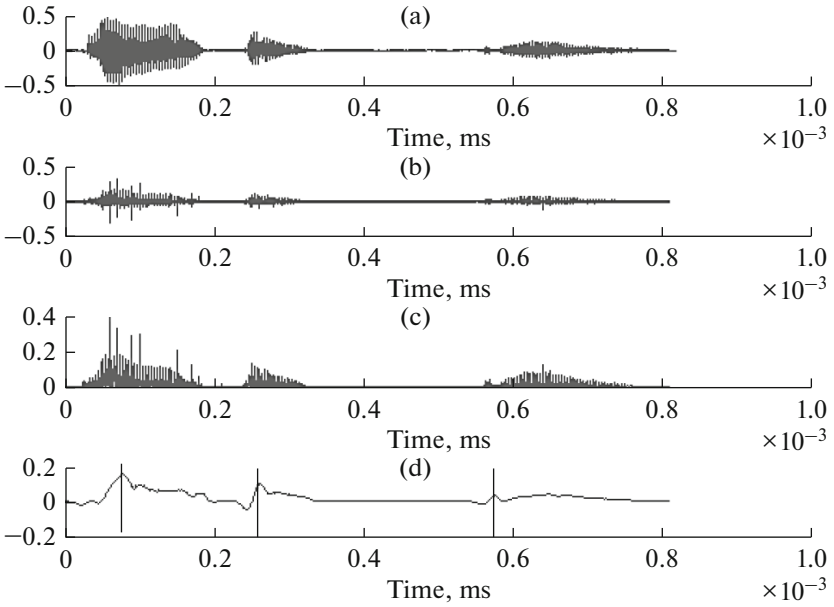

Fig. 4. (a) Speech signal of sample Tamil word 'thadukki'; (b) LP residual; (c) Hilbert envelope of LP residual; (d) VOP evidence plot with hypothesized VOP.

Hilbert transform of LP residual $e_h(n)$ is calculated using the residual $e(n)$. The analytical signal $x(n)$ is given by [12]

$$x(n) = e(n) + je_h(n), \tag{4}$$

where $e_h(n)$ is the Hilbert transform. The magnitude of the complex analytic signal of Eq. (4) is called the Hilbert envelope of the signal $h(n)$, which can be written as

$$h(n) = \sqrt{e^2(n) + e_h^2(n)}. \tag{5}$$

Significant change in the amplitude of the Hilbert envelope of the LP residual is the clue to identify VOPs. A modulated Gabor [13] window function in eqn.(6) is convolved with $h(n)$ to find the VOP evidence plot.

$$g(n) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{n^2 \cos(\omega n)}{2\sigma^2}}. \tag{6}$$

VOP detection for a sample Tamil word 'thadukki' is shown in the Fig. 4.

4.3. Spectral Transition Measure

The spectral transition measure employed in this study was the same as that proposed in [14]. The spectral gradient is given by the derivative of X_i denoted by $X'[i]$ can be viewed as a vector-valued distortion between infinitesimally small contiguous feature vector sequence. The derivative magnitude is expected to be large at the boundaries between consecutive quasi-stationary speech sounds, where the signal properties and hence the feature vector sequence will show abrupt changes. Thus, the frames corresponding to the most significant changes in the derivative magnitude will correspond to the segment boundaries.

The successive difference estimate of $X'[i]$ can be noisy because of errors in the parameter estimation process. The gradient is therefore calculated by fitting of a straight line to each of the vector elements, within a defined time window (in the minimum mean square sense). The gradient (for mth dimension) is thus estimated as

Table 3. Specification used in creating dataset

Language	Tamil
Speech type	Text read speech
Approach	Unsupervised
Recording conditions	Room environment
Number of speaker	2 male speakers
	2 female speakers
Age group	25–30
Region	Native

$$\Delta x_m(i) = \frac{\sum_{k=-K}^{K} k\, h_k\, x_m(i+k)}{\sum_{k=-K}^{K} h_k}, \tag{7}$$

where h_k is a symmetric window of length $(2K+1)$. The magnitude of the spectral derivative is then given by

$$x'[i] = \sum\nolimits_{m=1}^{p} \left(\Delta x_m(i)\right)^2. \tag{8}$$

Thus, $x'[n]$ tends to exhibit peaks at the boundaries between speech sounds corresponding to changing vocal tract configurations. Thus, its peaks correspond to the points of non-stationarity of the signal [15]. The problem of finding segment boundaries, between two stationary segments, thus reduces to a peak-picking procedure on $x'[i]$. Thus, STM uses heterogeneity between successive segments as a measure for segmentation.

4.4. Syllable Segmentation

The point at which VOP occur are identified using excitation information. The proposed syllable segmentation method works based on vowel onset point events and spectral transition. The frames in which the maximum spectral transitions happen are identified using spectral transition measure for the given speech signal. The procedure for proposed segmentation method has been presented as follows.

Step 1: Compute VOP using excitation information.

Step 2: Compute10 Linear Prediction Cepstral Coefficients (LPCC) for every frame of 20 ms with 10 ms overlap.

Step 3: Find the frames F in which VOP occurs.

Step 4: Set boundary B_0 = frame 1 and B_n = last frame of the signal.

Step 5: For every consecutive pair of VOPs.

Step 6: Select an appropriate frame from F, which has more spectral transition as the boundary between the two consecutive pair of VOPs.

Step 7: Repeat the steps 5 through step 6 for all remaining VOPs.

5. EXPERIMENTAL RESULTS

Since no open Tamil speech data sets are available, selective passages from Tamil News magazine read by 4 speakers are recorded and considered as data set. 100 Tamil speech utterances consisting of 25 unique Tamil words containing only CV and CVC patterns are separated and considered in this experiment. Data are recorded with the help of a unidirectional microphone using a recording tool audacity in a normal room with minimum external noise. The sampling rate used for recording is 16 kHz. The description about the data used is also given in Table 3. The syllable boundaries are manually identified by analysing the speech and used in the performance analysis of automatic segmentation.

For each speech utterance, the frames in which maximum spectral transition occurs are found using STM and the frames in which VOPs occurs are identified using excitation information. STM frames, VOP frames information, the proposed segmentation method identify the hypothesized syllable boundary frames.

The proposed method is compared with the baseline method Vowel Onset Point and Vowel Offset Point(VOP + VOF) [6]. To evaluate the performance of the methods, both methods are evaluated with respect to the manually segmented syllable boundaries of Tamil utterances in terms of percentage of match

Table 4. Percentage of Match of syllable segmentation method with different tolerance

No of words	Number of actual boundaries (excluding first and last boundary)	Method	% Match				
			deviation 10 ms	deviation 20 ms	deviation 30 ms	deviation 40 ms	deviation 50 ms
100	560	VOP + STM	58.8	65.3	77.3	79.3	81.7
		VOP + VOF	55.6	64.1	75.2	78.2	81

with the tolerance value ranges of 10 to 50 ms. In the experiment, the speech signal is blocked into frames of 20 ms duration with 10 ms overlapping. The average duration of the CV and CVC units of Tamils language is 375 and 625 ms respectively. The performance of the proposed segmentation method with tolerance value in the range of 10 to 50 ms is presented in Table 4. It is noted that over segmentation and under segmentation is based on the number of VOPs extracted.

6. CONCLUSION

Segmentation of speech signals into linguistic units has enormous applications in the fields of recognition, synthesis, labeling and transcription and coding. In the proposed work, Tamil speech segmentation task has been carried out to identify the boundaries of syllable in the given speech utterance. A novel syllable segmentation method is proposed which uses the VOPs as anchor points to identify the position of consonant-vowel unit and spectral difference are used find the syllable boundary. The performance of the proposed method is evaluated using a range of frame tolerance to find the percentage of the match with handmade segmentation and compared with the exiting VOP + VOF method. The proposed method provides reasonable results for segmentation of syllables of Tamil language. In future, the work can be extended with enhanced alignment technique to increase the percentage of match with actual boundaries. Segmentation with prosodic syllable or morpheme-based segmentation in Tamil language may also be proposed.

REFERENCES

1. Gangashetty, S.V., Sekhar, C.C., and Yegnanarayana, B., Spotting multilingual consonant-vowel units of speech using neural network models, *International Conference on Nonlinear Analyses and Algorithms for Speech Processing; Lect. Notes Comput. Sci.,* 2005, vol. 3817, pp. 303–317.
2. Karunakaran, K., Jeya, V., and Mozhyiyal, *Kavitha pathipagam,* India, 1997.
3. Sridhar Krishna, N. and Murthy, H.A., Duration modeling of Indian languages Hindi and Telugu, *SSW5,* 2004, pp. 197–202.
4. Thangarajan, R., Natarajan, A.M., and Selvam, M., Syllable modeling in continuous speech recognition for Tamil language, *Int. J. Speech Technol.,* 2009, vol. 12, pp. 47–57.
5. Anantha Natarajan, V. and Jothilakshmi, S., Segmentation of continuous speech into consonant and vowel units using formant frequencies, *Int. J. Comput. Appl.,* 2012, vol. 56, no. 15.
6. Soumya Priyadarsini Panda and Ajit Kumar Nayak, Automatic speech segmentation in syllable centric speech recognition system, *Int. J. Speech Technol.,* 2016, vol. 19, no. 1, pp. 9–18.
7. Lawrence Rabiner and Biing-Hwang Juang, *Fundamentals of Speech Recognition,* Prentice Hall, 1993, pp. 102–108.
8. Vuppala, A.K., Rao, K.S., and Chakrabarti, S., Improved vowel onset point detection using epoch intervals, *AEU—Int. J. Electron. Commun.,* 2012, vol. 66, no. 8, pp. 697–700.
9. Vuppala, A.K., Yadav, J., Chakrabarti, S., and Rao, K.S., Vowel onset point detection for low bit rate coded speech, *IEEE Trans. Audio Speech, Lang. Process,* 2012, vol. 20, no. 6, pp. 1894–1903.
10. Prasanna, S.R.M., Reddy, B.V.S., and Krishnamoorthy, P., Vowel onset point detection using source, spectral peaks, and modulation spectrum energies, *IEEE Trans. Audio Speech Lang. Process.,* 2009, vol. 17, no. 4, pp. 556–565.
11. Linear Prediction Analysis. iitg.vlab.co.in/?sub=59&brch=164&sim=616&cnt=1108.
12. Prasanna, S.R.M. and Yegnanarayana, B., Detection of vowel onset point events using excitation information, *Ninth European Conference on Speech Communication and Technology,* 2005.
13. Gabor, D., Theory of communication, *J. Inst. Electr. Eng., Part 1,* 1947, vol. 94, no. 73, p. 58.
14. SaiJayram, A.K.V., Ramasubramanian, V. and Sreenivas, T.V., Robust parameters for automatic segmentation of speech, *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, USA,* 2002, pp. I-513–I-516.
15. Dusan, S. and Rabiner, L., On the relation between maximum spectral transition positions and phone boundaries, *Proceedings of 9th International Conference on Spoken Language Processing,* 2006, pp. 645–648.