Indian Journal of Science and Technology, *Vol 8(17), DOI: 10.17485/ijst/2015/v8i17/61362, August 2015*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Segmentation of Continuous Tamil Speech into Syllable like Units

V. Anantha Natarajan* and S. Jothilakshmi

Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar - 608 002, Tamil Nadu, India; friends_mpt2004@yahoo.com

Abstract

The present growth in the field of information and communication technologies has diverted the focus of many researchers towards the speech technologies. Speech technology comprises of many subfields like speech synthesis, speech recognition, speaker recognition, speech compression, speaker verification and Multimodal interaction. The basic units of the speech synthesis and speech recognition system are syllable, phoneme and word. This study mainly focuses on syllable segmentation or syllabification with the aim to further develop a speech synthesis tool in Tamil language for Human Computer Interaction [HCI]. The syllable boundaries are identified using the formant frequency, F1. The proposed syllable segmentation algorithm is applied and tested on a set of recorded continuous speech corpus. Initially, the continuous speech signal is divided into segments by removing the silence regions. The silence removal method used in this work depends on features such as signal energy and spectral centroid. After removing silence portion from the speech signals, the speech segments are further processed using Linear Predictive Coding (LPC) to extract the formant frequencies. Then the peaks in the formant frequencies are used as clue to mark the syllable boundaries in the speech. The proposed algorithm is producing an average accuracy of 89% in identifying syllable boundaries when it is compared with the hand labeled syllable boundaries.

Keywords: Linear Predictive Coding, Speech Recognition, Speech Synthesis, Syllabification and Formants

1. Introduction

Over the last few years an extensive number of researchers have been working hard to develop an more accurate and reliable Speech Recognition and Synthesis systems. In many of the researches in the area of speech recognition the syllables are considered as the basic units since it contains more temporal information than the phonemes. The syllable acts as a bridge between the lower level and the higher level representational tiers of language[1]. The syllable boundaries in the continuous speech signal are more precise and well defined than the phoneme boundaries so their segmentation can be done more accurately[2]. The number of possible syllables in most of the languages is huge which makes the storage and management process more difficult[3]. In some of the speech synthesis systems instead of storing the syllable like smaller speech units in the database the speech parameters like MFCC and formant frequencies are stored. During run time best parameters that exactly or closely match the given text will be used for the synthesis of the speech. Formant based speech synthesis for Amharic language was developed by NadewTademe[4]. Bhatti S., et al.[5], introduced a method which extracts formant frequencies from speech signal for speech recognition. In[6], a technique for building a syllable based continuous speech recognizer when un-annotated transcribed train data is available was presented.

Existing syllable segmentation approaches are based on minimum phase group delay. T. Nagarajan et al.[7], proposed a segmentation algorithm which splits the speech signal in to syllable-like units. The segmentation algorithm is based on the minimum phase signal derived from the short-term energy. R. Janakiram et al.[8], developed an syllable segmentation algorithm based on the Group Delay (GD) and vowel onset point detection. LokendraShastri, et al.[9], described an automatic

*Author for correspondence

method for finding the boundary of syllabic units in continuous speech. This segmentation algorithm utilizes a Temporal Flow Model (TFM) and modulation-filtered spectral features. An algorithm similar to the cepstral smoothing approach for formant extraction using homomorphic de-convolution which utilizes the properties of group delay function to give importance to formant peaks is proposed[10].

In this study formant based approach for syllable segmentation is proposed and tested using IIIT-H Indic speech databases recorded at Speech and Vision Lab, IIIT-H[15]. From the speech recordings the formants are extracted using LPC analysis. The identified peaks in the F1 formant frequency are used as a clue to detect the syllable boundaries. Formants can also be extracted using group delay functions.

This paper is structured as follows: In Section 2, the background concepts related to the field of Speech Synthesis and Recognition are discussed. The details of the proposed algorithm are presented in Section 3 and in Section 4 the results of the experiments conducted for this algorithm are detailed. Finally Section 5 concludes the paper.

2. Background Concepts

Syllables are considered as the basic unit of organization for a sequence of sounds. A syllable contains a vowel with consonants in the initial and the final margins. One of the main reason for choosing the syllable as the basic unit is that the Tamil language is syllable-centric. When a syllable database is constructed using the proposed algorithm it can be used for developing both speech recognition system and synthesis system. In Figure 1 the role of the syllable database in both the systems is shown. The data storage space required to store the syllable database will be high. Hence instead of storing the syllable like speech units, the reflection coefficients and the residual signal of the speech units can be stored. The residual signal and reflection coefficients will require less amount of storage space than the original speech signal. Speech recognition by analyzing the residual signal can be implemented easily and more accurately.

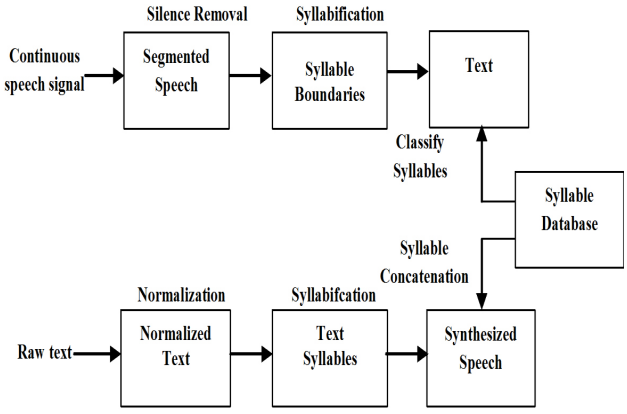

Figure 1. Block Diagram of Syllable Based Speech Recognition & Synthesis System.

3. Proposed Algorithm

The proposed method for the segmentation of speech in to syllable like units based on formant frequency consists of three steps. The first step being the pre-processing in which silence region from the input signal is removed. The second step is the formant extraction using Linear Predictive Coding analysis and the third step is identifying the formant peaks in the voiced region of the signal to mark the syllable boundaries.

3.1 Pre-Processing

The computational speed of any signal processing application can be increased by using a dimensionality reduction technique. Silence removal can be considered as a one of the efficient dimensionality reduction technique in speech signal processing. Therefore in speech analysis it is needed to first apply a silence removal method. Before proceeding the silence removal process the speech signal is windowed using hamming window of length 50ms. As described in[13] by Theodorous Giannakopoulous, the two acoustic features namely the signal energy and the spectral centroid are used for silence removal from the speech signal. Initially the two feature sequences are extracted from the given input speech signal and thresholds are determined for each sequence. Speech segments are detected based on the simple thresholding technique.

Signal Energy of the i^{th} frame is defined using the formula

$$E(i)=\frac{1}{N}\sum_{n=1}^{N}\left|x_i(n)\right|^2$$, N is the length of the sample.

The spectral centroid, C_i represents the spectrums' center of gravity.

$C_i = \frac{\sum_{k=1}^{N} (k+1)\, x_i(k)}{\sum_{k=1}^{N} x_i(k)} . x_i(k)$, where $X_i(k)$ is the DFT of the i^{th} frame.

The energy and the spectral centroid will be low in the silent region of the speech signal. The Figure 2 shows the results obtained after pre-processing stage using the proposed algorithm. In Figure 2(a) the short time energy of the original and the filtered signal is plotted and in Figure 2(b) the spectral centroid of the original and the filtered signal is plotted and in Figure 2(c) the original wave form is shown.

3.2 Formant Analysis using LPC Parameter

Linear predictive analysis of speech for formant extraction poses many merits and some demerits. There are two methods for estimating formants from the predictor parameters. The widely used and the method used in this study for formant analysis is factoring the predictor polynomial and based on the roots obtained, formants are extracted. The other method is to obtain the spectrum and choose the formants by a peak picking method. The advantage of using the linear predictive method of formant analysis is that the formant centers the frequency and its bandwidth can be calculated accurately by factoring the predictor polynomial[14]. Initially the order of the predictor '*p*' is found using the formula given below.

p = round(fs/1000)+2, where fs is the sampling frequency in Hz.

In linear prediction an estimate $\hat{x}(n)$ for a sample value is given as a linear combination of previous sample values.

$$\hat{x}(n) = \sum_{i=1}^{p} a_i\, x(n-i)$$

After performing the LPC analysis on the speech signal, to identify the missing fundamental frequency auto correlation is performed on the speech signal. Autocorrelation is performed on a signal by taking cross correlation of the corresponding signal with itself.

3.3 Identifying Peaks in the F1-Formant Frequency

Identifying and analyzing peaks in a given time-series is important in many applications, because peaks are useful topological features of a time-series[11]. By using a thresholding technique the local extrema in the frequency vector is found which check and ensure whether the peaks are comparatively larger (or smaller) than the data around it. The disadvantage with the derivative based peak finding algorithms is that for noisy signal many spurious peaks are found. To eliminate the detection of spurious peaks as a local extrema, the feature vector is smoothed using an average filter. Figure 3 shows the plots of the raw frequency vector and the smoothed frequency vector with identified peaks. From Figure 3 it is very clear that by smoothening the frequency vector, the spurious peaks are eliminated whose presence will give an adverse effect on the output of the proposed syllable segmentation algorithm.

More complex peak detection methods consume large time for large data sets, require constant user interaction and they produce highly variable results. In this work the local peaks in the frequency vector is find by utilizing the alternating nature of derivatives and the user defined threshold.

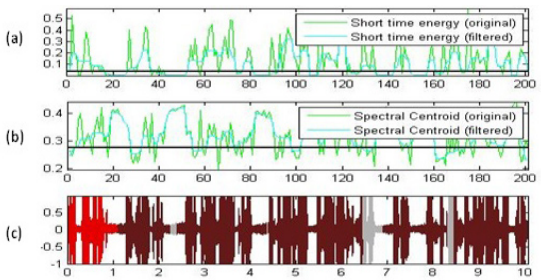

Figure 2. Waveform showing the Silent regions after Pre-processing.

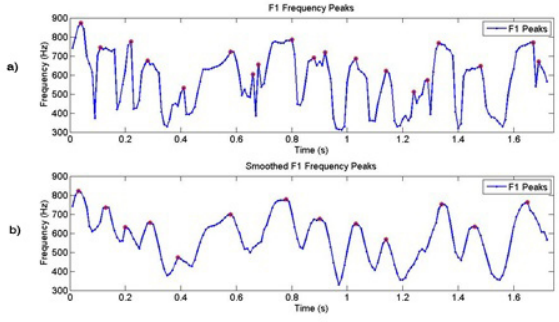

Figure 3. (a) Raw F1 Frequency with peaks (b) Smoothed F1 Frequency with peaks.

Table 1. Performance of Proposed Syllable Segmentation Algorithm

Sl no	Duration in secs	Number of Hand labeled Syllables	No. of Syllables detected		Accuracy in %	
			Proposed algorithm	Harma Syllable Segmentation Algorithm	Proposed algorithm	Harma Syllable Segmentation Algorithm
1	10	13	13	9	100	69.2
2	10	19	18	14	94.7	73.6
3	50	72	64	55	88.8	76.3
4	100	112	98	78	87.5	69.6

4. Results and Discussions

To evaluate the proposed method the IIT-H Indic speech database was considered. The speech corpuses were recorded from different Tamil speakers at a moderate speaking rate. The proposed algorithm was implemented in MATLAB and run on the entire set of test data. The algorithm outputs a set of detected syllable boundaries, '***Y***'. To find the detection accuracy the detected boundaries were compared with the hand-labeled boundaries '***X***'. The difference between a detected boundary 'y_i' and a reference boundary 'x_i' of 20ms is considered to be a match. The results of the proposed algorithm are also compared with the results obtained by the Harma syllable segmentation algorithm[12].

The Figure 4 represents the results produced when a sample speech file is tested with the proposed segmentation algorithm. The long dashed lines represent the syllable boundaries identified using the proposed algorithm and the dark solid lines mark the hand labeled syllable boundaries. While comparing the hand labeled boundaries and detected boundaries in the Figure 4, all the boundary lines are matching exactly except the sixth boundary line. Since the difference between the detected and hand labeled boundary is more than 20ms it treated as an error. The actual number of syllables detected by the hand labeled boundary lines is 13 in this sample speech signal and the number is same when using the proposed algorithm. When the same sample speech file is tested with the Harma syllable segmentation algorithm it identifies only 9 syllables. The proposed algorithm is evaluated with different sample speech sample files and the results obtained are tabulated in Table 1.

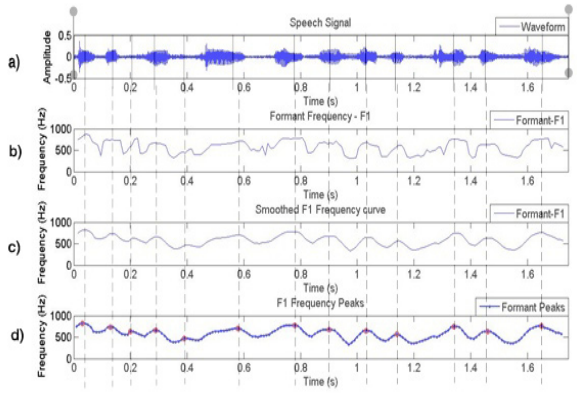

Figure 4. (a) Waveform of the original sample speech signal (b) Formant Frequency F1 vs Time (c) Smoothed F1 Frequency curve (d) Smoothed F1 curve with Peaks.

The accuracy of the syllable segmentation algorithm is calculated by finding the ration between the number syllable boundaries detected correctly by the algorithm to the actual number of syllables handled labeled in the speech corpus. The tolerable difference between the detected syllable boundary and the hand labeled syllable boundary is taken as 20 ms. If the detected boundary is greater or lesser than the hand labeled boundary by 20 ms then the detected syllable boundary is considered as an error.

$$Detection\ Accuracy = \frac{Number\ of\ Syllables\ detected\ by\ the\ algorithm}{Number\ of\ syllables\ hand\ labeled}$$

5. Conclusion

A novel and simple technique for syllable segmentation of continuous speech signal using formants is proposed. Once the formants peaks are identified the speech is segmented in to various smaller units to constitute a syllable database. The syllable database can be used for both speech recognition and speech synthesis. The proposed algorithm is simple to implement and performs better than other syllable segmentation algorithms mentioned in the literature. This research has been initiated with a motive to further extend this algorithm to build a concatenation based speech synthesis system for Tamil Language.

To build such a concatenation based speech synthesis using syllables, a syllable database has to be constructed. Moreover this database can also be used to develop a syllable based continuous speech recognizer also.

6. References

1. Fant G. Acoustic Theory of Speech Production. The Hague, Netherlands: Mouton and Co.; 1960.
2. Hugo M, Neto JP. The use of syllable segmentation information in continuous speech recognition hybrid systems applied to the Portuguese language. ISCA. 2000; 927–30.
3. Villing R, Timoney J, Ward T, Costello J. Automatic blind syllable segmentation for continuous speech. ISSC; 2004. p. 41–6.
4. Nadew T. Formant based speech synthesis: Synthesizing Amharic vowels. VDM verlag; 2009. ISBN: 3639178947.
5. Ali A, Bhatti S, Mian MS. Formants Based Analysis for Speech Recognition, Engineering of Intelligent Systems. IEEE International Conference; 2006. p. 1–3.
6. Lakshmi A, Murthy HA. A syllable based continuous speech recognizer for Tamil. Interspeech; 2006. p. 1878–81.
7. Nagarajan T, Murthy HA, Hegde RM. Segementation of speech in to syllable like units. EUROSPEECH. Geneva. 2003.
8. Janakiram R, Kumar CJ, Murthy HA. Robust syllable segmentation its application to syllable centric continuous speech recognition. Proceedings of National Conference on Communications; Chennai, India. 2010. p. 276–80.
9. Shastri L, Chang S, Greenberg S. Syllable Detection and Segmentation Using temporal Flow Neural Networks. Proceedings of the Fourteenth International Congress of Phonetic Sciences; 1999.
10. Murthy HA, Yegnanarayana B. Formant Extraction from group delay function. Speech Communication - Elsevier; 1991 Aug; 10(3):209–21.
11. Palshikar G. Simple algorithms for peak detection in time-series. TRDDC Technical Report' 09.
12. Harma A. Automatic Recognition of Bird Species Based on Sinusoidal Modeling of Syllables. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003); 2003 Apr. p. 535–8.
13. Theodorous G. A method for silence removal and segmentation of speech signals. Computational Intelligence Laboratory (CIL), Institute of Informatics and telecommunications, NSCR Demokritos; Greece. 2009.
14. Rabiner. Digital processing of speech signals. Pearson Education India; 1979.
15. Kishore P, Kumar EN, Venkatesh K, Rajendran S, Black AW. The IIIT-H Indic Speech Databases. Proceedings of Interspeech; Portland, Oregon, USA. 2012.