# Analysis of Fricatives, Stop Consonants and Nasals in the Automatic Segmentation of Speech Using the Group Delay Algorithm

Mohammed Musfir*, Raghava Krishnan K† and Hema A Murthy*
*Department of Computer Science and Engineering Indian Institute of Technology Madras,
†Department of Electrical Engineering Indian Institute of Technology Madras,

***Abstract*—Unit Selection based speech synthesis systems (USS) require accurate labeling of units. Accurate segmentation of speech waveforms manually is a laborious task. Syllable-based systems for Indian languages use a group delay based approach for semi-automatic segmentation of speech waveforms into syllables. This performance of the group delay based algorithm is poor when the syllables contain fricatives, nasals and unvoiced stop consonants. This paper proposes a modification to the algorithm that exploits the properties of these types of units to reduce errors. In particular, the ratio of energy in the high frequency bands to low frequency bands is used as a cue to segment the speech signal.**

## I. Introduction

The process of annotation of speech data involves segmentation of the speech signals into smaller units such as syllables or phonemes, and aligning the waveforms with the corresponding text. The task at hand is to develop a segmentation algorithm to build a speech corpora for the purpose of building speech synthesis systems for low resource Indian languages.

Generally, corpora for Indian Languages are built using the syllable as the basic unit [1]. This is because Indian languages are syllable timed and the syllable is the basic unit of sound production [2]. As the boundaries of syllables are characterised by regions of low energy, spectral mismatches between succesive units are not perceivable. The syllable also captures intra syllable co-articulation effectively. Due to the above mentioned reasons, it seems apt to use the syllable as the basic unit to build an inventory of speech units for Indian languages.

A syllable is defined as a canonical consonant (CV) unit or vowel consonant (VC) or even C*VC* where C is a consonant and V a vowel [3]. For the purpose of annotation of the speech data, the text is syllabified using a set of hand written rules. Conventionally, acoustic segmentation is performed automatically using a speech recognition system. Indian languages are low resource languages. Therefore, automatic acoustic segmentation usually is not accurate. One approach would be to manually annotate the data which is tedious, time-consuming and error prone owing to inconsistenices amongst annotators. Another approach would be to use signal processing techniques that would look for cues in the signal itself.

Various efforts have gone into segmenting stress timed languages at the phoneme level. Phoneme segmentation is obtained by using large amounts of data. [4] proposes a method where speech is segmented into its corresponding phonemes using the trigram model when only the orthographic transcription is available. Methods to segment speech into its corresponding phonemes have used statistical models such as Hidden Markov Models with a large amount of data to obtain accurate segmentation [5]. The syllable being a larger unit, the amount of data required to cover all the syllables in a language a sufficient number of times to obtain accurate syllable level segmentation of speech data would require a very large database. Therefore, signal processing techniques are employed which exploit acoustic cues of the speech signal to segment the waveform into its corresponding syllable units.

There have been various efforts in the past that have dealt with segmenting speech into syllabic units. [6] proposed a method for segmentation where Kullback-Leibler divergence is taken as the effective distance measure so as to detecct long term statistical differences in speech signals. The system detects changes in acoustic conditions and recognises previously observed conditions and thus pools the adaptation data accordingly.

[7] have proposed a syllable level segmentation technique based on a common syllable model for Japanese. The segment boundaries are determined by the optimal HMM state sequence. The training segments are classified into syllables with sustained vowels and devocalized[1] vowels.

[8] have proposed a short term energy based method for detecting syllable nuclei. The speech signal is first bandpass

978-1-4799-2361-8/14/$31.00 

[1]uttering with tense vocal chords

filtered and the short-term magnitude function is computed. This signal is further passed through a low pass filter to remove the ripples caused by transient phonemes. The peaks in the resulting energy contour are declared as syllable nuclei.

[9] have proposed the group-delay based algorithm that uses short-term energy as the feature for syllabification. The short term energy is used to derive the minimum phase signal which is treated as the magnitude spectrum. The minimum phase equivalent function of the short term energy and its group delay function are computed [10]. The group delay algorithm resolves boundaries very well for syllables that are purely made up of voiced segments while precision reduces for syllables containing unvoiced segments. This paper proposes that, the issues faced in the resolution of boundaries at the unvoiced segments could be solved by considering high frequency to low frequency energy ratios. The regions where the group delay segmentation algorithm fails can be predicted from the text information available. Thus, the analysis of the high frequency to low frequency energy ratios is therefore required only in regions where the group delay based segmentation approach is known to fail.

### A. Group Delay based segmentation

The group delay function resolves peaks and valleys of the spectrum effectively if the given signal is minimum phase. The peaks and valleys of the STE function correspond to voiced and unvoiced parts of the waveform. In the STE function, at the syllable level, energy is high in the voiced region which is the region of the vowel and tapers off towards the end of the syllable. Fluctuations are created in the energy contour if the syllable ends with a consonant or the energy contour tapers off smoothly if the syllable ends with a vowel. Local minima at the start and end of the syllable can be obtained by smoothing these fluctuations if present. The algorithm is summarized below as given in[9].

- Compute STE function $E(m)$ $m = 1, 2, 3...M$ using overlapping windows.
- Compute the $N^{th}$ order FFT
- Invert $E(m)^{\gamma}$ where $\gamma = 0.001$ after appending $N/2 - M$ $E_{min}$ to the sequence
- Construct the symmetric part of the sequence by lateral inversion. The resulting $E(K)$ is the magniude spectrum of an arbitrary signal
- Inverse DFT of the $E(K)$ is computed to get the root cepsrum and the causal portion of the same is the minimum phase signal[9].
- Group delay function of the windowed causal sequence of the above signal is calculated, to give $E_{gd}(K)$. The size of window applied is

$$N_c = \frac{Length\ of\ short\ term\ energy\ function(STE)}{Window\ scale\ factor\ (WSF)} \quad (1)$$

Length of STE is the number of samples in the STE function. The WSF is a scale factor with which the root cepstrum is truncated.

- The positive peaks in the minimum phase group delay function $(E_{gd}(K))$ are the regions of low energy in the STE function. The positive peaks are identified if $(E_{gd}(K))$ is positive and

$$E_{gd}(K-1) < E_{gd}(K) < E_{gd}(K+1) \quad (2)$$

## II. Parameters to be Adjusted to Improve Segmentation Accuracy

The segmentation algorithm implemented on a Hindi utterance is shown in Figure 1. The text denoted in red font are the erroneously segmented regions. The algorithm performs well when the energy minima in the STE function are distinct. The region of interest is where the minima in the STE function corresponding to unvoiced frames is not sufficient to obtain accurate segmentation. Performance of group delay segmentation is based on various parameters such as the length of the syllable, the window scale factor which in turn decides the cepstral scale factor, presence of silences and the nature of the phoneme in the unvoiced region[11]. Section III discusses various issues with the method and the attempts made to resolve them.

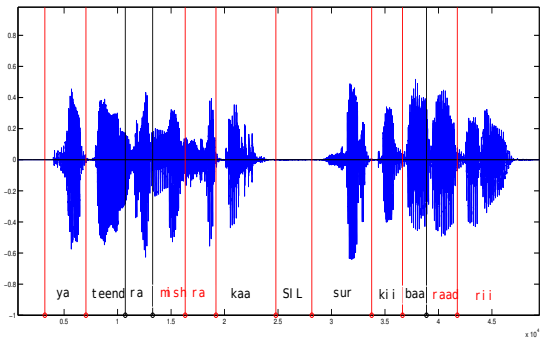


Fig. 1: Segmentation using group delay algorithm

### A. Window Scale Factor (WSF)

WSF plays an important role in determining the scaling factor for truncating the cepstrum as in Equation 1. The $N_c$ determines the resolution of segmentation which increases when the WSF is low. Very high resolution will cause the boundaries to be marked for CV/VC at the C-V or V-C transition point, which is not the desired boundary. The values of WSF are chosen with respect to the language and speaker after analysing the recorded speech data and performing durational analysis on it. Analysis conducted on the language Hindi shows that 41% of the syllables lie in the range of 150ms to 200ms. Figure 2 gives an insight into the duration of syllables. It can be inferred

that about 90% of the syllables vary in duration between 125-270 ms.

It is observed that setting the WSF within the range 3 - 5 resulted in the number of boundaries detected being almost equal to the actual number of syllables. Errors occur when long syllables are segmented multiple times and short syllables are not resolved. Cues from the spectral energy ratios can be used to address the issue.

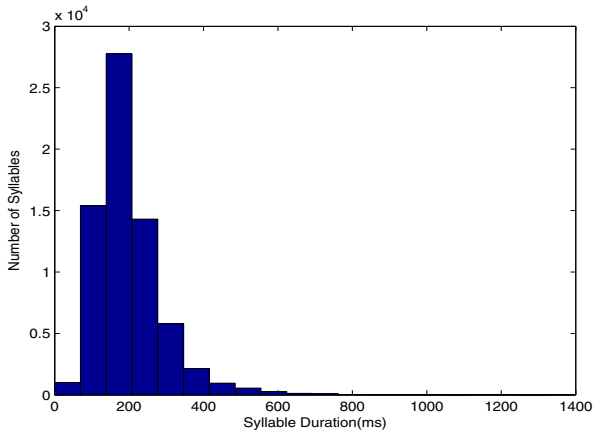


Fig. 2: Syllable Duration Analysis

### B. Thresholds of Silence and Voiced Regions

The silence threshold is the threshold of the duration of a silence below which it will be considered to be an unvoiced segment. The voiced region threshold is the threshold of duration of a voiced segment above which it will be considered to be multiple voiced segments.

## III. Factors Affecting the Accuracy of Segmentation

### A. Long - Silences

Long silences may be misinterpreted as multiple boundaries by the group delay algorithm. It is therefore necessary to remove long silences [9]. This has been achieved by performing Voice Activity Detection (VAD) with consideration being given to the energy variation and spectral flatness of the region.

### B. Detection of Fricatives

Analysis of spectrograms of syllables with fricatives shows that the energy of the high frequency region of the spectrum [4000 - 8000Hz] will be higher than the energy in the low frequency region of the spectrum [0 - 4000Hz]. It is possible to predict the occurrence of a fricative in the utterance from the text available. This could be exploited to locate the frames in the waveform where a fricative is expected. The ratio of the high frequency energy to low frequency energy is calculated for a few consecutive frames in this region. The boundary is moved to the nearest frame where there is a sudden decrease in the ratio.

### C. Detection of Stop Consonants

The energy bursts in the stop consonants often occur after a small region of inactivity which often could be mistaken as a silence. This should be avoided by setting proper thresholds for the duration of the silence and voiced segments. The accuracy can be further enhanced by looking at the high frequency to low frequency energy ratios in the region of the bursts. A sudden decrease in the ratio indicates the end of the stop consonant.

## IV. Results and Discussion

The experiments were conducted on recorded Hindi data which consisted of read utterances. The syllable durations were estimated from 4 hours of available recording. The range of durations over which the syllables are distributed has been tabulated in Table I. As it is evident from the table, most of the

TABLE I: Syllable Duration Analysis

| Duration in (ms) | Percentage of Syllables |
|---|---|
| $< 50$ | 1.47 |
| $50 - 150$ | 22.5 |
| $150 - 200$ | 41 |
| $200 - 270$ | 20.8 |
| $270 - 300$ | 8.5 |
| $350 - 400$ | 3.1 |
| $> 400$ | 2.63 |

syllables lie within the range of 125ms to 250ms, the WSF was tuned to satisfy the segmentation of majority of the syllables. The errors occuring due to longer duration of syllables were eliminated by thresholding the silences and voiced regions appropriately. In the following sections we discuss the use of the *hi-lo* (high frequency energy to low frequency energy) ratios to correct the syllable boundaries when they contain fricatives and stop consonants.

### A. Fricatives

In Figure 3, the syllable /mish/, in the highlighted region, is truncated in the middle of the phoneme /sh/. This occurs because the dips in the STE function are not distinct enough for the group delay function to resolve them. From the spectrogram in Figure 3 we can see that the energy of the higher frequencies in the spectrum is higher than that of the lower frequencies in the spectrum. The end of the syllable /mish/ can be identified as the point where the ratio of the energies of the high frequency region of the spectrum to the low frequency region of the spectrum decrease drastically. The correction made is highlighted in Figure 4.

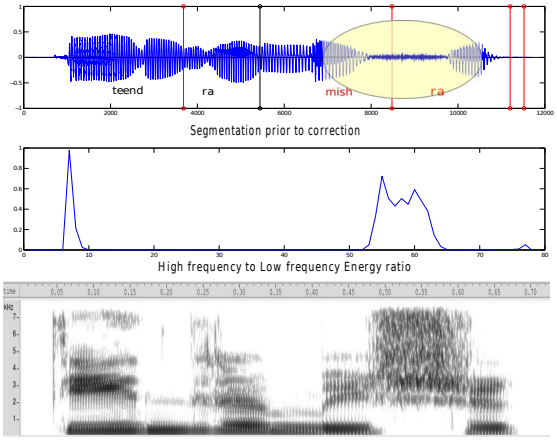


Fig. 3: Group Delay Segmentation on an utterance with a fricative

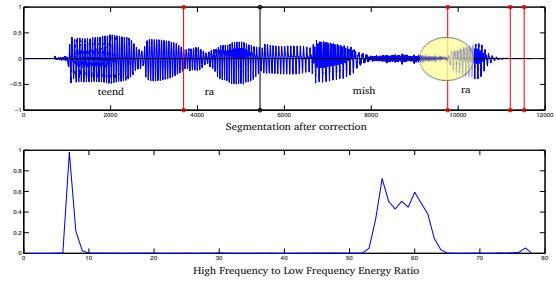


Fig. 4: Corrected segmentation of utterance with fricative after spectral energy ratio analysis

A similar problem is resolved in the utterance /tharaf/ /sa/ depicted in Figures 5 and 6. The yellow shaded portion is the region in the waveform where the unvoiced consonant occurs. In this case, the algorithm inserts a syllable boundary in the middle of the utterance /tharaf/ between the phones /a/ and /f/. The energy ratios of high frequencies to low frequencies were analysed with cues from the spectrogram and the segmentation was thus corrected as shown in the Figure 6.

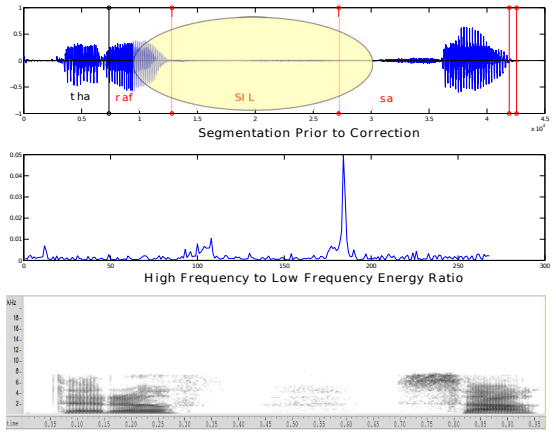


Fig. 5: Group delay segmentation on utterance /tharaf/ /sa/

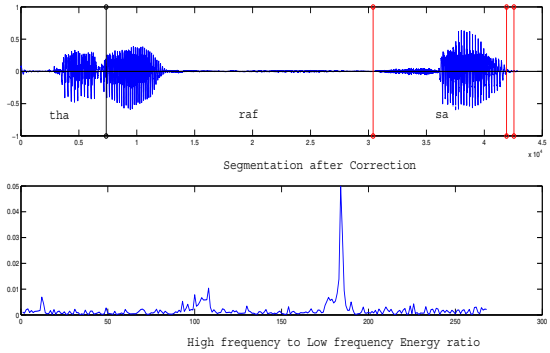


Fig. 6: Corrected segmentation of utterance /tharaf/ /sa/

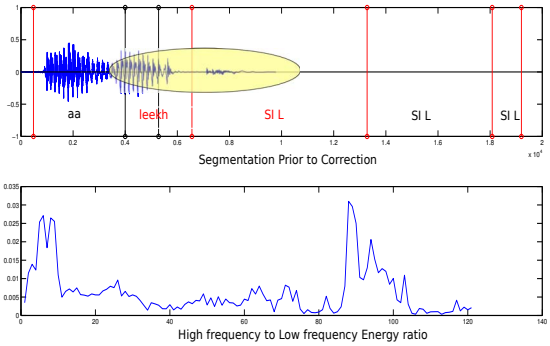


Fig. 7: Group Delay segmentation of utterance /Aleekh/

## B. Stop Consonants

Here the issue would be the inability to resolve the energy burst and thus mistaking the region of inactivity as a silence.The high frequency to low frequency energy ratio drops drastically at the end of the burst, which marks the end of the syllable. The Figures 7 and 8 show the waveform segmented using the group delay algorithm and the correction made to it on the basis of the hi-lo energy ratios respectively. The shaded portion indicates the wrong segments which were removed after analyzing the high frequency to low frequency energy ratios.

The basic problem here is that the short burst is also considered as a syllable. As the text is available, this ripple in the STE is eliminated by ensuring that peaks which are not very prominent in the group delay function are elimiated.

## C. Nasals

The syllables ending with nasals are also often wrongly segmented. This arises due to the coarticulation arising due to the phoneme following the nasal. This leads to the introduction of a syllable which is different from that of the one expected as per the text. This can be illustrated by the Hindi utterances /hum/ and /aap/, which ideally are two syllables. The pronunciation of these syllables though is /humaap/ thus

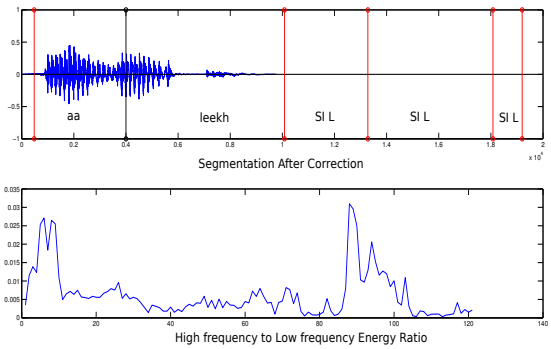


Fig. 8: Corrected segmentation of utterance /Aleekh/

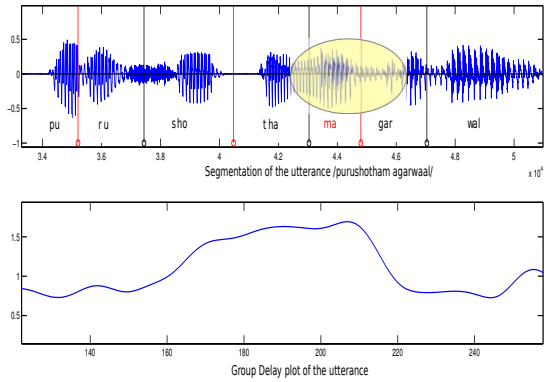


Fig. 9: Coarticulation in nasals

causing the waveform to be segmented as /ha/, /maa/ and /p/. Thus the syllable /maa/ is introduced as an additional segment. An example where the introduction of a syllable occurs is shown in Figure 9 where the /m/ of /purushotham/ and /a/ of aggarwal, coarticulate to form /ma/ in the shaded region. This problem can be resolved using dictionary based methods, where syllabification of the text has to be modified using cues from the acoustic signal. The text should therefore be syllabified according to the pronunciation, and the appropriate syllables have to be selected during synthesis depending on the context i.e, the phoneme preceding the nasal and the phoneme following the nasal.

The determination of the boundaries using *hi-lo* energy ratio plots works for those syllables ending with consonant or a stop consonant. The technique was tested on a set of 40 sentences.

| | |
|---|---|
| Total Number of Syllables | 972 |
| Wrong Syllables | 180 |
| Fricatives Corrected | 35 |
| Stop Consonants | 25 |

The accuracy of the labelling tool could be improved from 81.5% to 87%

## V. Conclusion

The segmentation algorithm based on the group delay function performs to a high degree of accuracy when the energy minima in the STE function are distinct. At the regions of fricatives and unvoiced stop consonants, the dips in the STE do not seem to be significant enough. The ratio of high frequency energy to low frequency energy have been identified to be an added cue in the regions of fricatives and stop consonants whereby the syllable boundaries are shifted to the point where the ratio decreases drastically. In the case of nasals, the coarticulation between the nasal and the phonemes preceeding and following it cause errors in segmentation. Therefore, the syllabification of text has to be modified according to the context in which the nasal is present. The issues arising due to semivowels could also be resolved by analysing pitch and formant frequencies along with the vowel onset/offset points which could be an extension to this work.

## Acknowledgment

The authors would like to thank DeitY for funding the project, *Development of Text to Speech Systems for Indian Languages* (11(7)2011-HCC(TDIL)).

## References

[1] S. P. Kishore and A. W. Black, "Unit size in unit selection speech synthesis," in *Proceedings of EUROSPEECH*, 2003, pp. 1317–1320.

[2] Osamu Fujimura , "Syllable as a unit for speech recognition," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 23, 1975, pp. 82 – 87.

[3] W. Su-Lin, M. L. Shire, S. Greenberg, and N. Morgan, "Integrating syllable boundary information into speech recognition," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 2, Munich, April 1997, pp. 987–990.

[4] A. Ljolje and M. D. Riley, "Automatic segmentation and labeling of speech," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 1, Apr. 1991, pp. 473–476.

[5] S. Young and G. Evermann, "The htk book."

[6] Matthew, A. S., Uday Jain, Bhiksha Raj, and Richard, M. S., "Automatic segmentation, classification and clustering of broadcast news audio," in *In Proc. of the DARPA speech recog. workshop*, Chantilly, VA, Feb. 1997, pp. 97 – 99.

[7] S. Nakagawa and Y. Hashimoto, "A method for continuous speech segmentation using hmm," in *Pattern Recognition, 1988., 9th International Conference on*, 1988, pp. 960–962 vol.2.

[8] H. R. Pfitzinger, S. Burger, and S. Heid, "Syllable detection in read and spontaneous speech," in *Proceedings of Int. Conf. Spoken Language Processing*, 1996, pp. 1261–1264.

[9] T. Nagarajan, V. K. Prasad, and H. A. Murthy, "The minimum phase signal derived from the magnitude spectrum and its applications to speech segmentation," in *SPCOM*, July 2001, pp. 95–101.

[10] V. K. Prasad, "Segmentation and Recognition of Continuous Speech," PhD Dissertation, Indian Institute of Technology Madras, Department of Computer Science and Engg., Madras, India, May 2002.

[11] Nagarajan, T., “Implicit Systems for Spoken Language Identification,” PhD Dissertation, Indian Institute of Technology Madras, Department of Computer Science and Engg., Madras, India, May 2004.