J Sci.Univ.Kelaniya **6** (2011) : 35-44

AUTOMATIC SEGMENTATION OF SEPARATELY PRONOUNCED SINHALA WORDS INTO SYLLABLES

P. G. N. PRIYADARSHANI AND N. G. J. DIAS[*]

Department of Statistics & Computer Science, University of Kelaniya, Kelaniya, Sri Lanka.

ABSTRACT

Aligned corpora are widely used in various speech applications like automatic speech recognition, speech synthesis, as well as prosodic and phonetic research. The segmentation into syllables can be done manually or automatically. But it consumes significantly more time for a fully manual phonetic segmentation and practically it is a complicated task because in many cases it requires a large aligned speech corpus. If the manual syllabification is done by a group of individuals then the consistency is decreased because the analysis variations of the individuals. Consequently, there is a dire need for automatic syllabification and it is important because Sinhala language is syllable centric in nature.

A method for syllabification of acoustic signals of separately pronounced Sinhala words has been given. Detecting the syllable boundaries was achieved by two main phases and those phases have been described with examples.

Keywords: syllabification, oscillograph, vowel, consonant, envelope, threshold

INTRODUCTION

All speech can be analyzed in terms of increasingly smaller and more refined units such as utterances, phrases, words, syllables, morphemes, phonemes. A word can be divided into syllables and a syllable is a unit of speech that consists of at least one vowel. Consonants and vowels combine to make a syllable. In general, in Sinhala, syllable definition can be expressed as $C_0^n V C_0^n$ where C_0^n signifies 0 to n consonants and V signifies a vowel including two diphthongs ඖ /au/ and ඓ /aI/ (Kumara, 2009; රාජපක්ෂ, 1997). Further it can be considered a syllable as having onset, followed by a

[*] Corresponding author: Email: ngjdias@kln.ac.lk

vowel, and followed by a coda. For example ග් [g] is the onset of the syllable ගෑස් [gæ:s] and ස් [s] is the coda.

As a consequence of technology advances increasingly sophisticated tools have become available to use with speech and music signals and scientists have the facility to study sound waves more effectively. Such research has led to the development of speech and music synthesizers, speech transmission systems, continuous speech segmentation systems, and automatic speech recognition systems etc.

RELATED WORK

Every written word is parsed by fixed linguistic rules into its component syllables. Currently there are many reliable tools for this purpose (Kumara *et al.*, 2007; Weerasinghe *et al.*, 2005). But as an acoustic unit of speech, the exact phonemic boundaries of each syllable vary depending on the rate of speaking and rhythmic flow of pronunciation. Therefore, syllabic durations are difficult to obtain from the speech signal. For example, the duration of a syllable can vary depending on the speaker as well as the position of the syllable in a word such as absolute beginning, middle and absolute end. Consider the acoustic graphs of the words ආකල්ප [a:kalpa] and කෙකටිය [kekatiya] in Figure 1 and Figure 2. It is clear that words should be divided into syllables as /a:|kal|pa/ and /ke|ka|ti|ya/ according to the syllable rules but it is somewhat difficult to identify the definite boundaries.

Although syllables are somewhat difficult to read, as they still have consonants, the vowel sounds make up the majority of the syllables that produce the louder part of these signals. As a result, breaking the words into syllables is a good start. On the other hand it requires significantly more time to align the syllables manually (Goldman, 2010) even though it generates comparatively accurate corpus. Further, if the manual syllabification is done by a group of individuals then the consistency is decreased because the analysis variations of the individuals. Therefore, automatic syllabification is important.

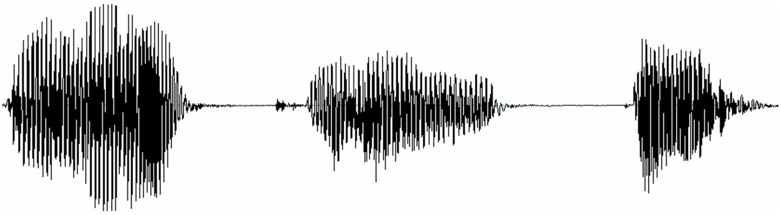

Figure 1: The acoustic graph of the Sinhala word /a:kalpa

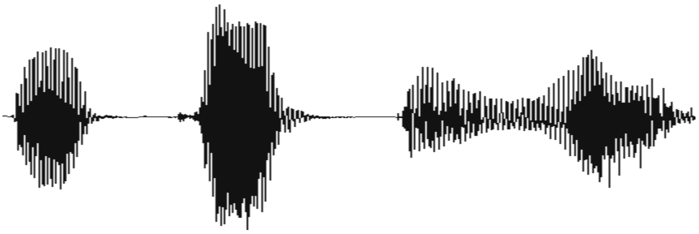

Figure 2: The acoustic graph of the Sinhala word /kekatiya/

In the past few decades, research on acoustic signal segmentation into units such as words, syllables, phonemes have been carried out (Goldman, 2010; Zhao and O'Shaughnessy, 2008; Kanda et al., 2008; Lewis and Tatham, 2001; Rahman et al., 2010; Bains and Sharma, 2010; Malfrere and Dutoit, 1997) for many languages in number of directions. For example, one approach is combining Text to Speech (TTS) with Dynamic Time Wrapping algorithm (DTW). In this case, synthetic speech is generated from the transcription and is compared with the corpus (Malfrere and Dutoit, 1997). The DTW finds the best temporal mapping between the two utterances using an acoustic feature representation. Segmenting acoustic signals using a Neural Network (Kanda et al., 2008) is another approach.

METHODOLOGY AND IMPLEMENTATION

Data Set

Frequently used 1150 words containing all the vowels and consonants of the Sinhala language were selected from the Ven. Pandit W. Sorata Nayaka Thera’s Sri Sumangala Sabdakosaya, a Sinhalese-Sinhalese Dictionary. Two speakers were selected in the recording of words whose native language is Sinhala and aged between 22-29 years. Each speaker was asked to utter words naturally and was recorded using the software Praat (Boersma and Weenink, 2010) with sampling frequency 16000Hz. We obtained two repetitions of each word from each speaker. The recordings were collected in a laboratory environment.

Detecting the Envelope

Initially we obtained the square of the acoustic signal as we needed to amplify the difference in the signal to identify the syllable boundaries. We observed that the words having clear low amplitude between the syllables can be separated into syllables directly using the oscillograph. But a higher proportion of words does not follow this pattern. Figure 3 shows the oscillograph of the word ගැඩවිලා [gædavila] while Figure 4 shows the square of the signal and it is clear that in the squared signal the syllable boundaries are significant than in the Figure 3. Thereafter, we examined the mean of the acoustic signal over a considerably small range (100 samples) with the aim of normalizing the signal. Now there is a somewhat clear definition of peaks. But in some cases all the peaks does not represent a syllable as well as some peaks relevant to the syllables are not significant. In other words, the correct threshold value should be decided for each acoustic signal to achieve a good accuracy. For example, with a threshold that is too low, noise may get picked up and with a too high threshold some syllables may be ignored.

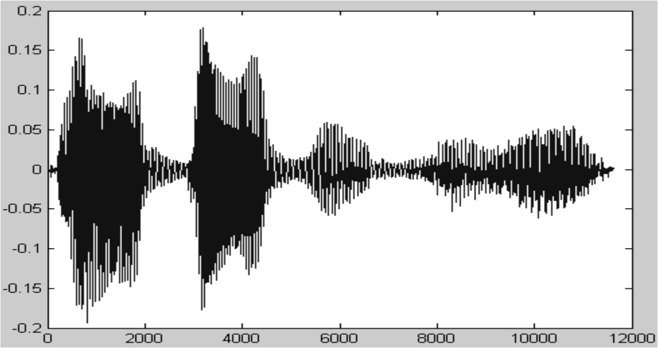

Figure 3: The acoustic graph of the Sinhala word /gedavila:/ taken from one of the speakers

Here we successfully determined the correct threshold value according to the number of syllables in the particular word. Based on that, we could develop the envelope that indicates the approximate boundaries for each word as shown in the Figure 4 and Figure 5. We choose the mid point between two syllables as the syllable boundary. The arrows indicate the separate positions (mid points) of the signal. Figure 6 shows the syllable boundaries of another word ගින්දැර [gindaral].

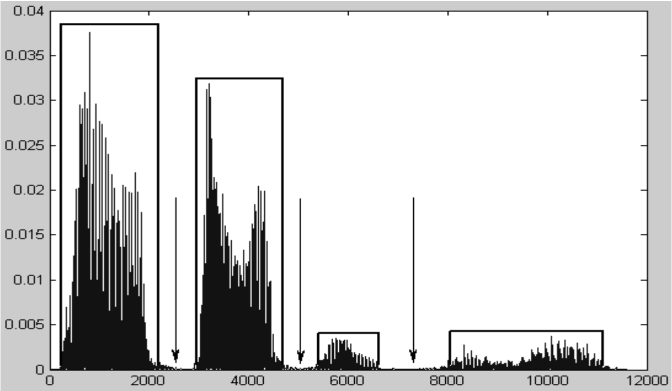

Figure 4: The squared acoustic graph of the Sinhala word /gædavila:/ in Figure 3 and the developed envelope

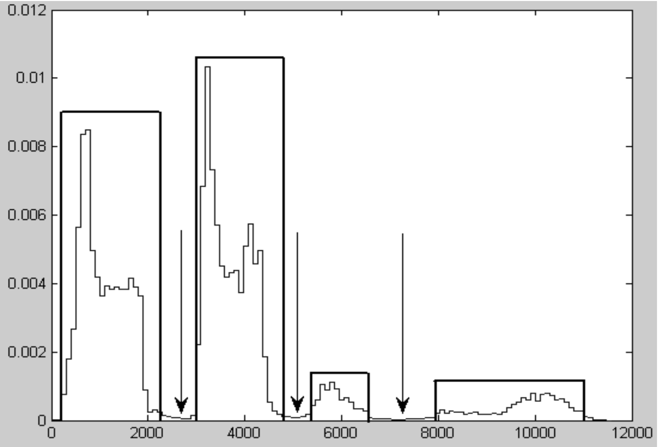

Figure 5: The squared mean acoustic graph of the Sinhala word /gædavila:/ in Figure 3 and the developed envelop.

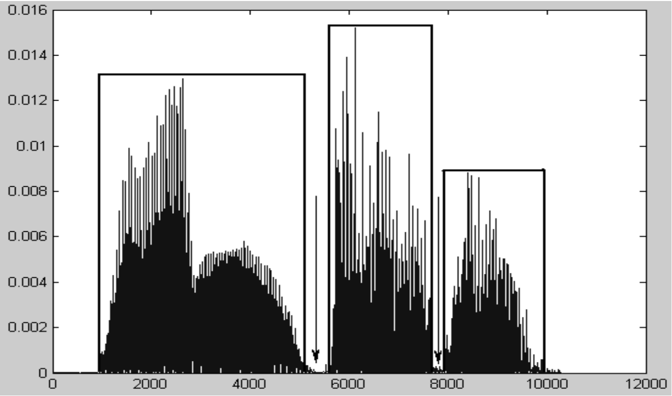

Figure 6: The squared acoustic graph and the developed envelope of the Sinhala word /gindara/.

We developed the relevant programs in MATLAB 7.0. Ultimately the identified syllables of a particular Sinhala word can be stored in separate files. For instance, Figure 7 and Figure 8 show the acoustic graph and squared acoustic graph of the Sinhala word /pedesa/ taken from one of the speakers. Separated syllables /pe/, /de/ and /sa/ of the word /pedesa/ are shown in Figure 9, Figure 10 and Figure 11 respectively and they are saved in separate files.

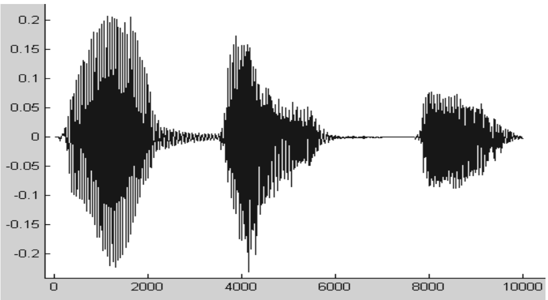

Figure 7:The acoustic graph of the Sinhala word /pedesa/ taken from one of the speakers

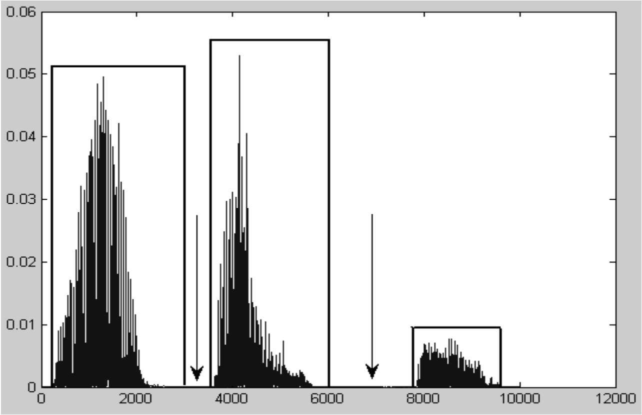

Figure 8: The squared acoustic graph of the Sinhala word /pedesa/ and the developed envelope

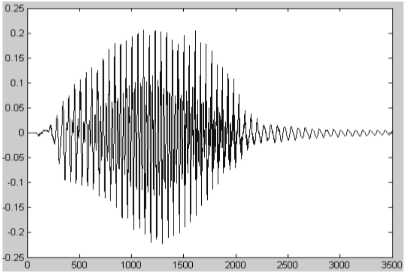

Figure 9: The separated syllable /pe/ from the word /pedesa/ in Figure 8

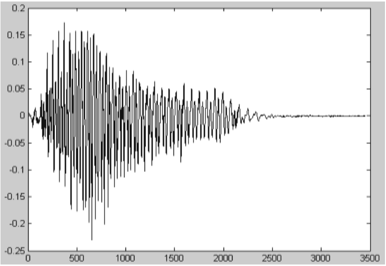

Figure 10: The separated syllable /de/ from the word /pedesa/ in Figure 8.

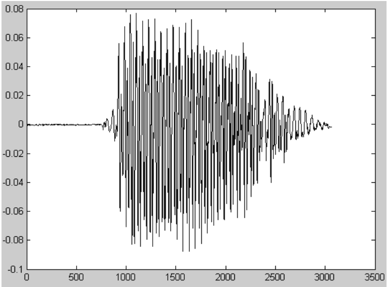

Figure 11: The separated syllable /sa/ from the word /pedesa/ in Figure 8.

We did the experiment on 1150 individual words, including the repetitions and the total number of words was 4600. We successfully separated 3795 words into syllables. In other words we could determine the correct threshold value along with the correct envelope for 3795 words out of 4600 words. The rest of the words were belonged to either *rejection envelope* or *no envelope* category. In the case of *rejection envelope*, even though the envelope was generated and the acoustic signal of the word separated into components equal to the number of syllables of the particular word, the separated positions were incorrect. In the case of *no envelope*, the program was not succeeded to determine an envelope as required. The Table 1 summarizes the outcome of our methodology.

Table 1: The identification rate for selected words

	Successful envelop	**Not Successful**	
		rejection envelop	**no envelop**
Percentage	82.5%	8.0%	9.5%

EVALUATION

Our results indicate a higher reliability on large number of words because the experiment exhibited 82.5% accuracy compared with human-tagged syllables. It was clear that if there is a clear gap between two syllables then our methodology works very well. For instance, the Sinhala word ආකල්ප [a:kalpa] in Figure 1 and පෙදෙස /pedesa/ in Figure 7 can be divided into syllables accurately. Further more, as long as the inter syllable energy is lower than the energy of each syllable of the signal such as in the words කෙකටිය [kekatiya] and ගැඩවිලා [gædavila] in Figure 2 and Figure 3, it is possible to separate the acoustic signal into their syllables.

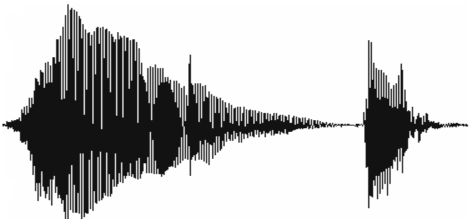

Figure 12: The acoustic graph of the Sinhala word /olma:da/ taken from one of the speakers

But when the syllables are overlapped the syllabification process becomes somewhat complicated and our methodology was not successful in tackling the problem. For instance, the word ඔල්මාද /olma:da/ represented in Figure 12 should be divided into /ol|ma:|da/ according to the syllable rules. But the first two syllables are overlapped and difficult to separate. As well, if the energy of a particular syllable is relatively lower than or equal to the energy of a inter syllable period of the signal then it was difficult to segment the word accurately.

CONCLUSION AND FUTURE EXTENTIONS

Our objective was to develop a method for syllabification of the acoustic signals of Sinhala words and we have achieved considerably high precision in the outcome. It was clear that the wrong pronunciation creates problems in the process of syllabification. In this work the informants were not trained to pronounce the words correctly, and therefore we have collected the pronunciation of words in their natural speech. As an effect, this decreased the accuracy of our methodology to some extent. Further, the two occurrences of the same word generated by the same speaker were behaved differently due to the variation in the pronunciation. It is better if all the syllable rules can be implemented in the program to solve the above mentioned drawbacks to some extent. Here all the data were recorded in a laboratory environment, but when it is implemented in a real world situation, the noise will be the main obstacle to achieve the necessary accuracy.

REFERENCES

Bains A. K. and N. Sharma 2010. Automatic syllable segmentation for Indian languages. Proceedings of ISCET, pp 170-171.

Boersma P. and D. Weenink 2010. Praat: doing phonetics by computer. Phonetic Sciences, University of Amsterdam, The Netherlands. http://www.fon.hum.uva.nl/praat.

Goldman, J. P. 2010. EasyAlign: a friendly automatic phonetic alignment tool under Praat, latlcui.unige.ch/phonetique/easyalign/ easyalign_unpublished.php

Kanda H., T. Ogata, K. Komatani and H. G. Okuno 2008. Segmenting acoustic signal with articulatory movement using Recurrent Neural Network for phoneme acquisition. International Conference on Intelligent Robots and Systems, pp 1712-1717.

Kumara, K H. 2009. Text- to-speech synthesis for Sinhala language. M.Phil. Thesis, University of Kelaniya, Sri Lanka.

Kumara K. H., N. G.J. Dias and H. Sirisena 2007. Automatic segmentation of given set of Sinhala text into syllables for speech synthesis. Journal of Science of University of Kelaniya 3: 53-62.

Lewis E. and M. Tatham 2001. Automatic segmentation of recorded speech into syllables for speech synthesis. Proceedings of Eurospeech, pp 1703-1707.

Malfrere F. and T. Dutoit 1997. High-quality speech synthesis for phonetic speech segmentation. Proceedings of Eurospeech.

Rahman Md. M., Md. F. Khan and A. M. Mohammad 2010. Speech recognition front-end for segmenting and clustering continuous Bangla speech. Daffodil International University Journal of Science and Technology 5: 67-72.

Weerasinghe R., A. Wasala and K. Gamage 2005. A rule based syllabification algorithm for Sinhala. Proceedings of 2nd International Joint Conference on Natural Language Processing, pp 438-449.

Zhao, X. and D. O'Shaughnessy 2008. A new hybrid approach for automatic speech signal segmentation using silence signal detection, energy convex hull, and spectral variation, Canadian Conference on Electrical and Computer Engineering, pp 000145-000148.

රාජපක්ෂ ආර්.එම්.ඩබ්ලිව් 1997. භාෂණ සිංහල ස්වර ශබ්ද, දර්ශක ප්‍රකාශන, කැලණිය.