IJCSC International Journal of Computer Science & Communication Vol. 1, No. 2, July-December 2010, pp. 267-270

Automatic Segmentation of Wave File

Nishi Sharma[1] & Parminder Singh[2]

[1]Department of Computer Science & Engg., B.G.I.E.T. Sangrur,
[2]Department of Computer Science & Engg., G.N.D.E.C. Ludhiana
Email: [1]naisha_sharma@yahoo.com, [2]parminder2u@rediffmail.com

ABSTRACT

This paper presents an ASS (Automatic Speech Segmentation) Technique to segment spontaneous speech into syllable like units. In the development of a syllable-centric ASS system, segmentation of the acoustic signal into syllabic units is an important stage. In this paper we focus on the identifying minimum unit of speech to be considered while training any speech recognition system. There are systems developed for continuous speech recognition in few Indian languages like Hindi and Tamil. This paper gives the statistical details of syllables in Punjabi and its use in minimizing the search space during recognition of speech. In this paper we describe how most of the segmentation can be done automatically. The results are plotted for frequency of syllables and the number of syllables in each word. We propose automatic segmentation method for syllable repetition in read speech for objective assessment of stuttered disfluencies which uses a new approach and has three stages comprising of feature extraction, Rule Matching and segmentation.
Keywords: Segmentation, Syllables, ASS.

1. INTRODUCTION

Syllables are very important unit of language, without the complete knowledge of syllables there is no possibility to linguistic the speech. Syllables are the combination of consonants and vowels. If we have a syllable then there must be a vowel present in that syllable. In previous papers [1], [2] we have argued that for natural sounding speech the syllable should be the preferred unit. Automatic Speech Segmentation of sound file starts with decomposing the speech into various fundamental units. These units can be words, phonemes, syllables. All Indian languages are syllable-centric. [3] described various choices of unit size word, syllable, and phone. Perceptual tests conducted to evaluate the quality of the synthesizers with different unit size indicate that the syllable synthesizer performs better than the phone or word. This characteristic gives us a clue to separate syllabic nuclei from consonants. [5] proposed a technique for automatic segmentation of speech into syllabic units. This technique was reasonably successful in achieving its aims and, indeed, our own algorithm makes use of a similar criterion in identifying possible boundaries. However, it cannot be used in isolation because the minima are phonetically determined and more closely allied to the phonological description of syllables rather than our morphemic definition. The use of syllables as units in automatic speech recognition has gained in popularity over the last few years and a number of algorithms have been proposed for their automatic identification [7], [8], [9]. However, for speech recognition purposes the boundary of the syllable does not have to be precisely defined and so these algorithms are not so useful for speech synthesis. We propose automatic segmentation method for syllable repetition in read speech for objective assessment of stuttered disfluencies which uses a new approach and has three stages comprising of feature extraction, Rule Matching and segmentation.

2. BACKGROUND

In digital world, a lot of the information is available but accessible to a few who can read or understand a particular language. Various solutions in the form of natural interfaces are provided by language technologies so that digital content can reach to the masses and facilitate the exchange of information across different people speaking different languages. These technologies play a crucial role in multi-lingual societies such as India which has about 1652 dialects/native languages. Seamless integration of speech recognition, machine translation and speech synthesis systems could facilitate the exchange of information between two people speaking two different languages. Our overall goal is to develop speech segmentation systems which can be further used for speech synthesis.

3. SYLLABLE DEFINITION

A word can be divided into syllables. Each syllable is a sound that can be said without interruption and are usually a vowel which can have consonants before and/or after it. Syllable Example: "*SOMVAAR*" have two syllables; "*SOM*" and "*VAAR*". Before embarking on the task of automatic syllable detection one has to decide

what constitutes a syllable in the first place. There is no universal agreement on a rigorous definition of the syllable but one which has wide acceptance, for English, is the following [4]:

- Syllables can be expressed in the form $C^{3}_{4}VC^{3}_{4}$ where Cn0 signifies 0 to n consonants and V signifies a vowel.

A much wider discussion of what constitutes a syllable can be found in [6]. This definition still allows one some freedom in deciding where the syllabic boundaries should be because the definition is abstract and the marking of the syllables is done on the physical waveform. For example do we express "*SOMVAAR*" (in Punjabi) as "*SOM*" and "*VAAR*" or "*SO*" and "*MVAAR*". We have decided that for engineering purposes a morphemic decomposition of words into syllables is to be preferred whenever possible with segmentation occurring on a phonological basis otherwise. The reason for this is that since syllables will be used for constructing new words then it is most likely that these words will be built-up on a morphemic basis. With the above definition of a syllable, a syllable boundary can be one of the following types:

- V-V, V-C, C-V, C-C: In practice we found that the number of rules is significantly reduced because in many cases they as so similar as to be not worthy of separate description.

4. METHODOLOGY

4.1. Speech Material

The speech samples consisted of 50 sound files of the one Punjabi word ("*SOMVAAR*" and "*SULTAANPUR*") by different speakers in similar conditions and using similar equipments.

- Test 1: about 3 seconds of one isolated word.
- Test 2: about 10-30 seconds of sentence.

4.2. Automatic Syllable Detection Method

The detection scheme used for assessment is divided into four steps as shown in Figure 1:

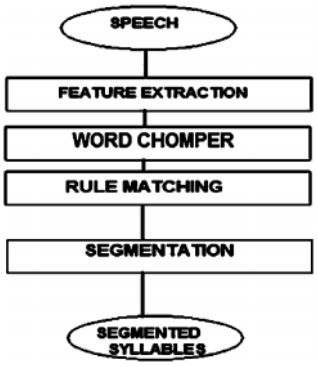

Fig. 1: Block Diagram of Automatic Detection Method

a. *Feature Extraction:* A common first step in feature extraction is frequency or spectral analysis. The signal processing techniques aim to extract features that are related to identify the characteristics. The speech signal is analyzed in successive narrow time windows of 10-30msec width, for its frequency content with 2msec offset. For each and every window we obtain the intensity of several bands on the frequency scale using feature extraction algorithm. The features in figure 2 which are extracted in this work are length, number of channels, sample rate, Data Length, Bits per Sample.

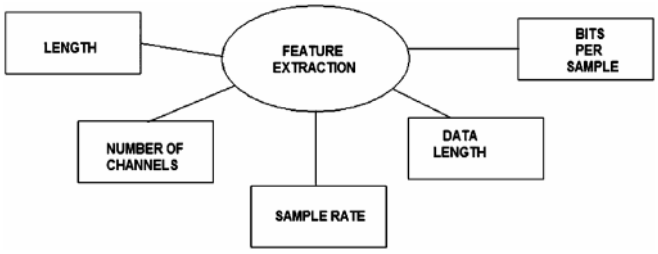

Fig. 2: Block Diagram of Feature Extraction Method

b. *Word Chomper:* The feature Extraction is subjected to word chomper. The length extracted in step 1 is chopped according to bits per sample that is 8 bits per sample.

c. *Rule Matching:* The default byte ordering assumed for WAVE data files is little-endian. Files written using the big-endian byte ordering scheme have the identifier RIFX instead of RIFF. The sample data must end on an even byte boundary. 8-bit samples are stored as unsigned bytes, ranging from 0 to 255. 16-bit samples are stored as 2's-complement signed integers, ranging from -32768 to 32767. We tested the Decision logic using rules to take a decision whether a sequence of five or more consecutive 127 is repeated or not. If such a sequence is found then this region is known as silence region. Afterwards, this region is being stripped off.

d. *Segmentation:* Now, Rule matching is subjected to segmentation of Syllables. Before stripping off the silence region, we store the starting and ending point in an array. In figure 3, word '*SOMVAAR*' is segmented into two syllables from where the pitch is almost equivalent to zero. Further the two segmented wave file are generated and stored separately. Phonetics gives no exact specification of syllables. The characteristic feature of the syllable is the dynamical transient part consonant vowel or consonant –vowel –consonant. The feeling of syllable boundaries, although usually very strong, is subjective and often not unique. For

Automatic segmentations of syllable many methods are available, which uses signal extremes, first Autoregressive (AR) coefficient, etc [10].

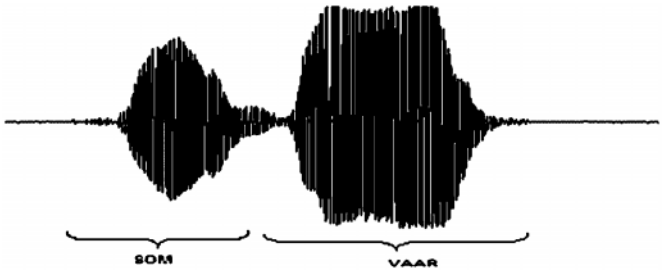

Fig. 3: Wave Form of Word '*SOMVAAR*'

The provision of both a phonetically defined syllable boundary together with precise knowledge of voiced and unvoiced sections of the waveform make the task of automatically marking the syllable boundaries in the waveform considerably easier.

5. WORKING

The first step of this system is input of wave/sound file. This is done with the help of sound forge. The basic attributes of the wave file are:

- 16KHz of Sample Rate;
- Mono Channels;
- 8 Bits Per Sample.

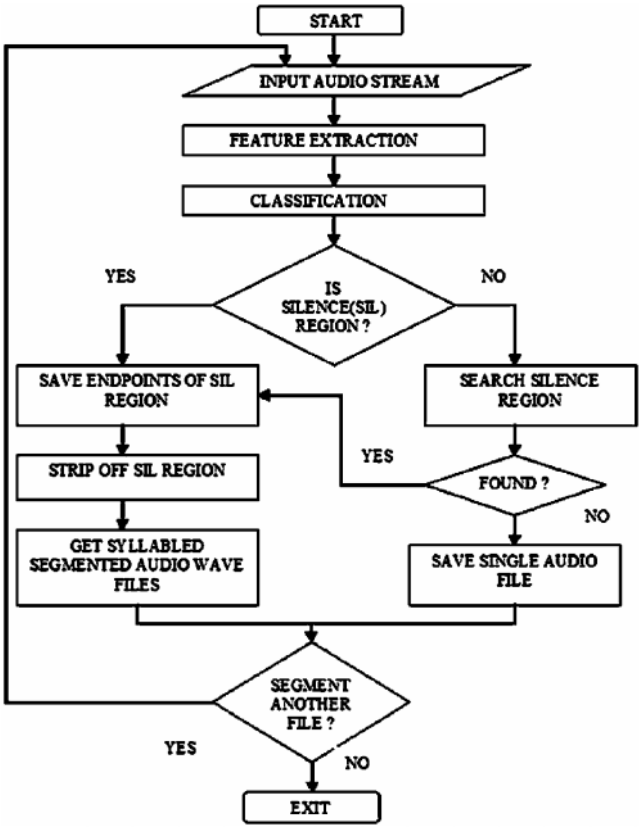

Fig. 4: The Flowchart of Automatic Syllable Segmentation (ASS) Algorithm

The figure 4 shows flowchart of this work which takes the audio wave file as input. Then features are extracted and classified to find silence and non-silence regions. If silence region is found then extract the endpoints of that region and strip it off. Using these endpoints, extract different syllable segmented audio wave files. If such region is not found, then segmentation is not possible.

6. CONCLUSIONS

Rules for the automatic segmentation of words into syllables have been derived based on a word chopper. The proposed algorithm has been produced for the implementation of these rules which utilizes the silence region. Although not capable of doing a complete automatic segmentation of all words we believe that about 90% of the waveform mark up can now be automated allowing much faster derivation of synthetic voices. In this paper, a novel approach for segmenting the speech signal into syllabic units is presented. The advantage of segmentation prior to labeling in speech is that it can be independent of the task. Simple isolated syllable models are built from the segmented data. Once syllable sequences are available, appropriate post-processing can be done to build systems for specific task.

7. FUTURE SCOPE OF WORK

The results of this work point to following directions of research that are likely to be needed to further improve accuracy of this work. We can also create a speech database at syllable units and this works can also be enhanced to a full fledge TTS system. This work can be modified according to the particular application.

REFERENCES

[1] Lewis, E. and Tatham, M., "SPRUCE – a New Text-to-speech Synthesis System", *Proceedings of Eurospeech* '91. *ESCA, Genova,* pp 1235-1238, 1991.

[2] Tatham, M. and Lewis, E., "Syllable Reconstruction in Concatenated Waveform Speech Synthesis", *Proceedings of the XIVth International Congress of Phonetic Sciences,* pp 2303-2306, 1999.

[3] S P Kishore and Alan W Black, "Unit Size in Unit Selection Speech Synthesis" *Proceedings of EUROSPEECH 2003*, Geneva, pp 1317 - 1320.

[4] Gimson, A.C., *An Introduction to the Pronunciation of English.* First Edition. Arnold, London, 1962.

[5] Mermelstein, P., "Automatic Segmentation of Speech into Syllabic Units", *J. Acoust. Soc. Amer.,* **58(4)**, 880-883, 1975.

[6] Van der Hulst, H. and Ritter, N., (eds.) *The Syllable: Views and Facts,* Walter de Gruyter, Berlin, 1999.

[7] Green, P. D., Kew, N. R. and Miller, D. A, "Speech Representations in the SYLK Recognition Project", *Visual Representation of Speech Signals,* Cooke, M. P. Beet, S. W. and Crawford M. D., Editors, Chapter 26, pp 265-272. John Wiley, 1993.

[8] Reichl, W. and Ruske, G., "Syllable Segmentation of Continuous Speech with Artificial Neural Networks", *Proceedings of Eurospeech* '93, *ESCA, Berlin*. pp 1771-1774, 1993.

[9] Wu, S.-L., Shire, M. L., Greenberg, S. and Morgan, N., "Integrating Syllable Boundary Information into Speech Recognition", *Proceedings of the International Conference Acoustics, Speech and Signal Processing (ICASSP* '97), **2**, 987-990, 1997.

[10] W. Reichl and G. Ruske, "Syllable Segmentation of Continuous Speech with Artificial Neural Networks," *In Processing of Eurospeech*, Berlin, **3**, pp. 1771-1774, 1993.