Segmentation of Continuous Punjabi Speech Signal into Syllables

Er. Amanpreet Kaur and Er. Tarandeep Singh

***Abstract*—Speech Recognization and Synthesis systems always need a speech signal to be segmented into some basic units like words, phonemes or syllables. These basic units are searched while implementing any segmentation or Recognization process. One of the methods for segmentation is by hand labeling speech based on linguistic interpretation of what was spoken. This is the approach taken in most acoustic-phonetic recognization system. But as this method is time consuming, error-prone and even the results are not re-producible. Therefore, a method is needed as an alternate to hand labeling which can segment the speech automatically into basic units without visual inspection of speech signal. So, here a procedure has been implemented that automatically segments the speech signals into syllable like units. At the end, the new Automatic Segmentation technique is also compared with the available techniques. Finally, the deviation between manual and automatic segmentation has been calculated for the onset and offset values for the syllable boundaries.**

***Index Terms*—Automatic Segmentation, Manual Segmentation, Short Term Energy, Punjabi Syllables.**

I. Introduction

Speech Recognization system requires segmentation of Speech waveform into fundamental acoustic units [7]. Segmentation is a process of decomposing the speech signal into smaller units. Segmentation is the very basic step in any voiced activated systems like speech recognization system and speech synthesis system. The set of fundamental acoustic units into which the speech waveform can be segmented are words, phonemes or syllables.

II. Basic units for Segmentation

Most of time word is considered to be the most natural unit of speech. Every word in Punjabi or any language has its well defined boundaries. But there are other problems that arise by using word as a speech unit. Each word has to be trained individually and there any sharing of parameters cannot be possible among words. Therefore, it is essential to have a very large training set so that all words in the vocabulary are adequately trained. In addition to these, more memory is also required as the number of words grow which in turn increases the problem of memory management. So choosing word as a basic unit for segmentation is not a good choice. Another option for the segmentation unit is phoneme. There are about 50 phonemes in a language.

Amanpreet Kaur, Assistant Professor, Department of Computer Science and Engg, RIMT-IET, Mandi Gobindgarh. Email id: taranaman@gmail.com
Tarandeep Singh, Director, Terminus Education Consultants, Sirhind.
Email id: trndeep@gmail.com

So it is easy to train with a training set of reasonable size. It is a well known fact that the same phone in different words has different realizations [6]. The realization of a phone is strongly affected by its adjacent phones or in other words, phones are highly context dependent. Therefore, the acoustic variability of basic phonetic units due to context is sufficiently large and is not well understood in many languages. Thus, it can be observed that there is overgeneralization in phone models while word models lack in generalization [7]. So it is clear from the discussion that we need the segmentation unit that is in between word & phonemes i.e. third fundamental unit syllables. Combination of phonemes gives rise to next higher unit called syllables which is one of the most important units of a language. A syllable must have a vowel called its nucleus, whereas presence of consonant is optional [4].

III. Punjabi Syllables

A syllable is a unit of organization for a sequence of speech sounds. 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. A syllable is typically made up of a syllable nuclear (most often a vowel) with optional initial and final margins (typically, consonants) [4]. For example, in Punjabi, the word ਮੰਗਲਵਾਰ is composed of two syllables: ਮੰਗਲ and ਵਾਰ. In Punjabi seven types of syllables are recognized [5]. These syllable types are: V, VC, CV, VCC, CVC, CCVC and CVCC; where V and C represent vowel and consonant respectively.

IV. Segmentation

For the purpose of Recognization and synthesis, speech often needs to be segmented into basic Phonetic Units. Segmentation is a process where a speech signal is decomposed into smaller acoustic units like words, syllables and phonemes. There are two ways that set of sub word units can be created. The first is by hand labeling speech based on linguistic interpretation of what was spoken. This is the approach taken in most acoustic-phonetic Recognization system. The alternative to hand segmentation of speech into acoustic phonetic units is to devise an automatic procedure which can provide consistent identification of sub word units in speech signal.

V. Short Time Analysis of Speech

Because of the slowly varying nature of the speech signal, it is common to process speech in blocks (also called “frames”)

ISBN: 978-988-17012-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

over which the properties of the speech waveform can be assumed to remain relatively constant. This leads to the basic principle of short-time analysis, which is represented in a general form by the "(1),"

$$X_n = \sum_{m=-\infty}^{+\infty} T\{x[m]w[n-m]\}, \quad (1)$$

where X_n represents the short-time analysis parameter (or vector of parameters) at analysis time n. The operator T { } defines the nature of the short-time analysis function (linear or non-linear transformation on speech), and w [n − m] represents a time shifted window sequence, whose purpose is to select a segment of the sequence x[m] in the neighborhood of sample m = n. The infinite limits in Equation imply summation over all nonzero values of the windowed segment $x_n[m] = x[m]$ w [n − m]; i.e. for all m in the region of support of the window. The short time energy measurement of a speech signal can be used to determine voiced vs. unvoiced speech. Short time energy can also be used to detect the transition from unvoiced to voiced speech and vice versa. The energy of voiced speech is much greater than the energy of unvoiced speech. Basic short-time analysis functions useful for speech signals are the short-time energy. This functions is simple to compute, and is useful for estimating properties of the excitation function in the model.

VII. Proposed work

Automatic Speech Recognition (ASR) deals with automatic conversion of acoustic signals of an utterance into text transcription. Speech recognition requires segmentation of speech waveform into fundamental acoustic units. Automatic speech segmentation is important for continuous speech recognition because it reduces the search space effectively in automatic speech recognition. Moreover, the signal segmentation technique is useful in automatic speech marks and labels. However, for automatic speech recognition (ASR), it is difficult to segment the speech input reliably into useful sub-units. Punjabi is not mature enough and hence most current research is trying to improve the accuracy of Automatic Speech Segmentation system in these languages. Basic units for segmentation are Words, Phonemes and Syllables. Different problems like large training set and more memory management arise by using word as a speech unit. In phonemes, it is difficult to find a direct correspondence between a speech segment and a phoneme [2]. Thus, it can be observed that there is overgeneralization in phone models while word models lack in generalization. Therefore a higher level of linguistic organization, namely, syllable, is a better linguistic unit for segmentation. Now taking syllable as a unit, a method for segmentation is needed. One of the methods for segmenting speech into syllable like units is manually by visually examining the waveform of the acoustic signal with the aid of graphics displays of the energy contour or spectrogram. However, this process is extremely tedious and time consuming. The decisions are subject to human errors such as mechanical errors committed by hands or mis - interpretation of a spectral or waveform display. So another method which is very convenient is automatic segmentation. There are two main approaches for automatic segmentation of the acoustic signal. One of them transfers segmentation data from an utterance of identical content, which has already been segmented, onto the utterance requiring segmentation.

This method uses a reference waveform which may be hand segmented and labeled natural speech of a reference speaker [8], or may be synthetically generated utterances from a known phonetic string, in which case, no manual segmenting and labeling is needed [1]. Unless a sufficiently high-performance text to speech system can generate the reference utterance the task of hand segmentation and labeling is still required for the generation of reference material. The other approach does not have to compare with reference waveform. The speech is segmented automatically into sub words units which are defined acoustically, but not necessarily phonetically [3].

VIII. Implementation

The speech segmentation system has been implemented by using Multimedia API in Windows environment. So, to implement such types of application we need multimedia kit (sound blaster card, microphone and speaker) and some programming languages to program the sound blaster card and implementation of speech technology algorithms. So, we have been used tool MATLAB Kit for developing any multimedia based application.

From the signal processing point of view, speech can be characterized in terms of the signal carrying message information. The waveform could be one of the representations of speech, and this kind of signal has been most useful in practical applications. Input for the system will be speech and output will be the waveform representing the boundaries of syllables.

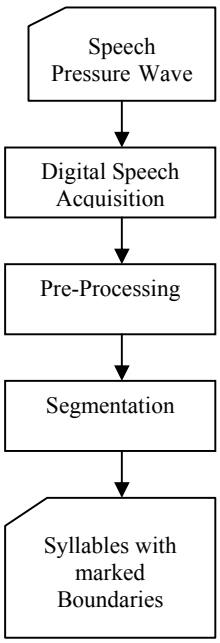

Fig. 1: Flow chart for Automatic Speech Segmentation

Automatic speech segmentation system has three major steps:

a) Digital speech acquisition
b) Signal Preprocessing
c) Segmentation

ISBN: 978-988-17012-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Signal preprocessing is to preprocess the signal to make it available for next process. After preprocessing, boundaries of the syllables will be marked automatically by recognizing the valleys in the waveform.

A. Digital Speech Acquisition

Digital speech acquisition is acquiring of the analog speech signal which is pressure wave through the microphone which gives digital representation of speech signal. Speech capturing or speech recording is the first step of implementation, i.e. how to capture the speech from speaker mouth to computer. Sound editing software "Sonic Foundry Sound Forge 5.0b" has been used for recording the speech. Recording has been done by native female speaker of Punjabi. The sampling frequency is 16 KHz; sample size is 8 bits, and mono channels are used.

B. Signal Preprocessing

It is very crucial to Pre-Process the Speech Signal in the applications where silence or background noise is completely undesirable. It is important in applications like Speech and techniques from speech signal where most of the voiced part contains Speech or Speaker specific attributes.

1. Eliminate the Background Noise

Background noise elimination is the first step in the signal processing. By this process, background noise is removed from the data so that only speech samples are the input to the further processing.

2. Pre-emphasis filtering

Pre-emphasis of the speech signal at higher frequencies is a preprocessing step employed in various speech processing applications. Pre-emphasis of the speech signal is achieved by the first ordering differencing of the speech signal. Although possessing relevant information, high frequency formants have smaller amplitude with respect to low frequency formants. A pre-emphasis of high frequencies is therefore required to obtain similar amplitude for all formants. This is usually obtained by filtering the speech signal with a first order FIR (Finite Impulse Response) filter, known as pre-emphasis filter. A first-order digital network processes the digitized speech signal to spectrally flatten the signal whose transfer function in the z-domain is given by "(2),"

$$H_{pre}(z) = 1 + a_{pre} Z^{-1} \quad \ldots\ldots \quad (2)$$

In time domain, Equation (2) can be written as

$$\tilde{S}(n) = S(n) + a_{pre}{}^{*} S(n-1) \quad \ldots\ldots \quad (3)$$

Where $\tilde{S}$ the output of the preprocessing stage and value of a_{pre} is –0.95 and n is sample number. A typical range of values for a_{pre} is [-1.0, -0.4]. Values close to –1.0 that can be efficiently implemented in a fixed-point hardware, such as –1 or – (1-1/6), are most common in speech recognition.

3. Framing

In most processing tools, it is not appropriate to consider a speech signal as a whole for conducting calculations. A speech signal is often separated into a number of segments called frames. This process of separation is known as framing. Continuous speech signal has been blocked into N samples, with adjacent frames being separated by M (M<N). After the pre-emphasis, filtered samples have been converted into frames, having frame size of 20 msec. Each frame overlaps by 10 msec.

4. Windowing

The window, w (n), determines the portion of the speech signal that is to be processed by zeroing out the signal outside the region of interest. To reduce the edge effect of each frame segment windowing is done. Rectangular window has been used.

C. Segmenting Signal into syllables

Following procedure has been used for automatically marking the boundaries of syllables in sound file.

a) Short term energy of a preprocessed signal has been computed by using the "(4),"

$$E_n = \sum_{m=n-N+1}^{n} x^2[m] w(n-m) \quad \ldots\ldots \quad (4)$$

Where w(n-m) is a windowing function, N is the length of window in samples and n is the sample number.

b) Some threshold value has been taken and signal having value less than this threshold value has been changed to zero as signal having syllable will have a data value more than threshold value.

c) Then signal has been checked for value not equal to zero and greater than some particular value and that point will be marked as starting location of the syllable.

d) After getting the starting location of syllable, then zero values of signal has been checked and if there are suitable numbers of continuous zeros then it has been defined as the end of syllable. Once end point, we can proceed analyzing signal from end point of first syllable looking for the starting position of next syllable.

IX. Results and Discussion

Technique has been implemented in Matlab 7.8. Various speech signals in Punjabi have been recorded and segmented. Proposed method has been implemented and analyzed for different Punjabi speech signals. Results have been shown for signal where the boundaries of syllables are marked automatically. The wave file contains the following sentence:

1. ਭਾਰਤ ਤੇ ਪਾਕਿਸਤਾਨ ਵੱਲੋਂ ਗੱਲਬਾਤ ਮੁੜ ਸ਼ੁਰੂ ਕਰਨ ਦਾ ਐਲਾਨ

ISBN: 978-988-17012-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

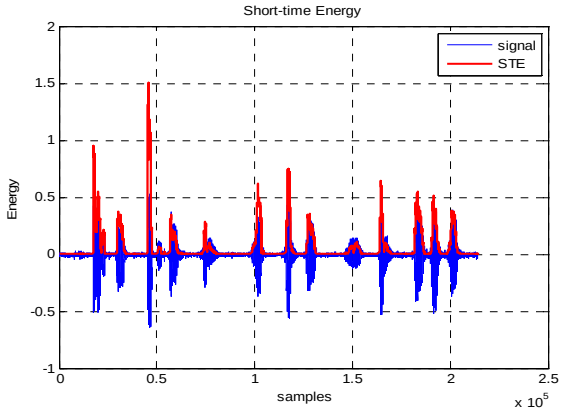

Fig. 2: Short Term Energy Waveform of speech signal sentence 1

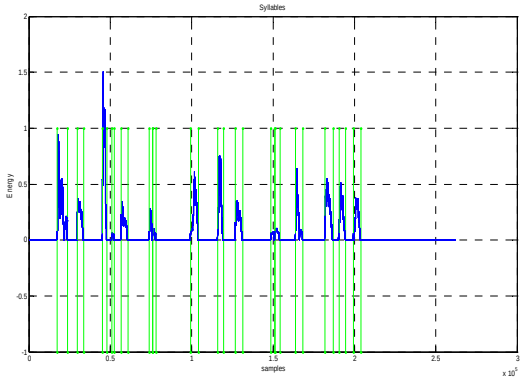

Fig. 3: Waveform with marked Syllable Boundaries for sentence 1

A .Comparison with Traditional Manual Segmentation

Results of proposed automatic approach have been compared with results of Manual segmentation for starting and end point of each syllable so that the deviation between two approaches can be found. Manual segmentation has been done by visually inspecting the original waveform and by marking the boundaries of Punjabi syllables manually. But in this approach segmentation is done automatically or acoustically based on stress and silence in the given utterance which in turn represents the starting and end point of syllable respectively. The following table represents the difference between Automatic and Manual segmentation for starting as well as endpoints of the some of the syllables in a sentence. It has also been observed that boundaries of syllables marked by automatic technique are very much accurate and the difference between two (Automatic and Manual) techniques is very much negligible.

TABLE I
COMPARISON OF AUTOMATIC AND MANUAL RESULTS OBTAINED FOR SOME OF SYLLABLES IN THE TARGET SENTENCE

Sentence	Auto-Segmentation		Manual-Segmentation		Deviation	
	Onset (in samples)	Offset (in samples)	Onset (in samples)	Offset (in samples)	Onset (in samples)	Offset (in samples)
ਭਾਰਤ	17537	23590	17490	23530	47	60
ਤੇ	29643	33614	29584	33562	59	52
ਪਾ	45172	48071	45145	48013	27	58
ਕਿਸ	51019	52564	50977	52489	42	75
ਤਾਨ	56601	60938	56575	60885	26	53

REFERENCES

[1] Bridle J.S. and Chamberlian R.M.,"Automatic Labelling of Speech using Synthesis by Rule and Non Linear Time Alignment," in Speech Communication, 1983, pp 187-189.

[2] Prasad V. K., Nagarajan T, Murthy H A ," Automatic segmentation of continuous speech using minimum phase group delay functions," In Speech Communication 42 , 2004, pp 1883–1886.

[3] Shuping Ran and J. Bruce Miller, "Exploring the Phonetic Structure of Speech Signal using Multi Layer Perceptrons " in Proceeding of SST , 1990, pp 22-27 .

[4] Singh Parminder, Gurpreet Singh Lehal, "Syllable Based Text-To-Speech Synthesizer for Punjabi", in 10th Punjab Science Congress, Feb 2007.

[5] Singh Prem, "Sidhantik Bhasha Vigeyan", Patiala: Madan Publications, pp. 391.

[6] T. Nagarajan, Hema A. Murthy, and, Rajesh M. Hegde, "Segmentation of speech into syllable-like units", in Proc. EUROSPEECH-03, Geneva, Switzerland, Sep. 2003, pp.2893-2896

[7] Thangarajan R, Natarajan A.M.,"Syllable Based Continuous Speech Recognition for Tamil," in South Asian Language Review VOL.XVIII. No. 1, 2008.

[8] Wagner W, "Automatic Labelling of Continuous Speech with a Given Phonetic Transcription using Dynamic Programming Algorithm", IEEE Conf. on Acoustics, Speech and Signal Processing, 1981, pp 1156-1159

ISBN: 978-988-17012-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)