Vol.4,No.1, July, 2015
106 - 112,ISSN:0975-5446 [VNSGU JOURNAL OF SCIENCE AND TECHNOLOGY]

Segmentation of Gujarati words from Continuous spoken Gujarati speech signal

PATEL Bharat C
T. & M. T. College of Information Science, Surat
patelbharat99@yahoo.co.in

DESAI Apurva A.
Veer Narmad South Gujarat University, Surat
aadesai@vnsgu.ac.in

Abstract

One of the most difficult tasks in continuous speech processing is to define boundary of the phonetic units present in the signal. Phones are strongly co-articulated and there are no clear borders among them, so the link between the linguistic and the acoustic segmentation is not simple to define. In this paper, we presented a proposed model which uses speech features such as Short-Term Energy (STE), Zero-Crossing Rate (ZCR) and peaks, and develop an algorithm to segment continuous Gujarati speech signal into word or sub-words. The experiments are carried out on continuous Gujarati speech signal and obtained results are presented.

Keywords: Speech Segmentation, Continuous Gujarati Speech Signal, Zero-Crossing Rate, Short-Term Energy

1. Introduction

Speech is the most efficient way through which human being can communicate with one another. Speech signal carries silence, unvoiced and voiced part. The silent part of speech signal contains no speech; unvoiced speech signal is produced when vocal cords are not vibrated so that resulting speech waveform is aperiodic or random in nature. The voiced speech signal is generated when an air flows from the lungs; the vocal cords are tensed and therefore vibrate periodically so the resulting speech waveform is quasi-periodic [1]. It should be clear that the segmentation of the waveform into well defined regions of silence, unvoiced and voiced signals is not exact; it is often difficult to distinguish a weak, unvoiced sound from silence, or a weak voiced sound from unvoiced sound or even silence.

1.1 Segmentation

Automatic segmentation of speech has been a subject of study for over 3-decades and it plays major role in many speech processing and Automatic Speech Processing (ASP) application. Speech segmentation is the process of dividing input speech signal into a sequence of segments with similar characteristics. There are two approaches to segment a speech signal: Manual segmentation, also known as blind segmentation and Automatic segmentation.

In a first approach, segmentation does not require any prior knowledge of the signal being processed. Here, the boundary is detected manually. Moreover, this approach has some limitation as follow:

- ✓ It is time consuming
- ✓ It required highly trained human annotators. So it is difficult and costly to obtain.
- ✓ No two annotators can identify the boundaries exactly same.

These problems can be overcome by a second approach of speech segmentation called automatic segmentation, but it requires prior knowledge to process the input speech signal. There are several methods available for automatic speech segmentation such as zero crossing rate, short term energy, minimum phase group delay method, wavelet etc. The segmentation algorithm of continuous speech signal is required in various applications like: speech recognition, speech synthesis, speech enhancement, speech corpus collection, speaker verification and in the research field of natural language processing.

1.2 Related work

There are many works done in segmentation of speech signal. Authors use different segmentation method in their work. This section introduces method used by different authors to segment a speech signal and the result obtained in their proposed work.

T.Nagarajan et. al. [2] used a minimum phase group delay based approach to segment spontaneous speech into syllable like units. They analyzed over 5000 speech dialogs in Tamil language. The duration of the speech signal varies from 0.5 sec to 25 sec and obtained good result.

Sharma and Kaur [3] described the concept of automatic segmentation of continuous speech signal of Punjabi language. They found that group delay function is better representation of STE for syllable boundary detection.

Rahman and Bhuiyan [4] presented several dynamic thresholding approaches for segmenting continuous Bangla speech sentence into words or sub-words. They proposed three efficient methods for speech segmentation: two of them are usually used in pattern classification (i.e. k-means and Fuzzy c-means (FCM) clustering) and one of them is used

in image segmentation (i.e. Otsu's thresholding method). They achieved an average segmentation accuracy of approximately 94%.

Prasad et. al. [5] proposed a novel approach for segmenting the speech signal into syllable-like units. They propose a group delay based approach for processing the short-term energy to determine segment boundaries. The performance of this technique is tested on both continuous speech utterances and connected digit sequences. The error segment boundary is ≤ 20% of syllable duration for 70% of the syllables.

Malcangi [6] demonstrated soft computing method for segmentation of a speech into phonetic units. The author used fuzzy decision logic to infer phonetic unit separation point and prove that this approach is more effective in separating phonetic units.

Rahman and Bhuiyan [7] used feature extraction approaches for segmenting Bangla speech sentences into word or sub-words. They proposed a method based on two speech feature namely time domain and frequency domain features and achieved segmentation accuracy of 96%.

Jayasankar et. al [8] proposed an algorithm for automatic segmentation of Tamil voiced speech. They use absolute energy and zero crossing rate to process speech samples to accomplish the segmentation.

Chaudhury et. al [9] presented a word separation algorithm for Real Time Speech. The algorithm is developed by considering prosodic features with energy. The proposed algorithm correctly detects word boundaries of 98%.

This paper is organized as follows. In section 2 we describe the methodology used to separate voiced/unvoiced region from the speech signal. Section 3 presents the implementation details of proposed model and an algorithm. Result and conclusion are discussed in section 4 and 5 respectively.

2. Methodology

Automatic speech recognition requires analyzing features of speech signal. In order to segment continuous speech signal, it requires to check whether the speech signal is voiced or unvoiced. There are several methods available to detect voiced/unvoiced part from speech signal. In proposed model, we combined three parameters of speech, i.e. ZCR, STE and peak, are used to determine the voiced or unvoiced region of speech signal.

2.1 The Short-time energy

The amplitude of the speech signal varies with time. Generally, the amplitude of unvoiced speech segment is much lower than the amplitude of voiced segment. The short-time energy is defined in Equation 1.

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

Where w[n-m] is a windowing function. The choice of the window determines the nature of the short-time energy representation. In our model, we used Hamming window which is defined in equation 2.

$$W(n) = 1 \quad 0 \le n \le N\text{-}1 \qquad = 0 \quad \text{otherwise} \quad (2)$$

2.2 The Short-time Zero Crossing Rate

Zero-crossing rate is the measure of number of times in a given time interval/frame that the amplitude of the speech signals passes through a value of zero. The short-time zero-crossing rate is defined in equation 3.

$$Z_n = \sum_{m=-\infty}^{\infty} 0.5\ |sgn\{x[m]\} - sgn\{x[m-1]\}|w[n-m] \quad (3)$$

Where $sgn\{x[m]\} = \begin{cases} 1, x[m] \ge 0 \\ -1, x[m] < 0 \end{cases}$

A reasonable generalization is stated by different authors in [3,6,8,10] that if the zero-crossing rate is high then the speech signal is unvoiced and if the zero-crossing rate is low then the speech signal is voiced. Conversely, the energy of unvoiced speech signal is much lower than the energy of voiced speech signal.

3. Implementation

To develop a proposed model, we use MATLAB as a programming environment. The speech signal has more than 10000 samples with sampling frequency rate 8000Hz. It is analyzed by non-overlapped frame of size 500 samples. To locate the voiced region from continuous speech signal, we have taken threshold value for parameters zero-crossing rate, maximum amplitude and peak is <100, ≥0.1 and >20 respectively.

The steps involved in the segmentation of a continuous speech signal are as follow:

1. Read the continuous speech signal stored in a wave file.
2. Compute length of speech signal read in step1.
3. Set frame length of specific size and compute number of frame of specified frame size for entire speech signal.
4. Repeat steps (a) to (d) for each frame of speech signal till end of speech signal is encountered.
   - (a) Locate the current frame of the speech signal.
   - (b) Compute maximum amplitude, number of zero-crossing rate and number of peaks in a current frame.

(c) Determine the start and end boundary of voiced region on the basis of value of different parameters computed in step (b) with their corresponding threshold value.
(d) Switch to next frame.

4. Experimental Results

To evaluate the performance of proposed model, we have tested many speech signal of Gujarati sentences. One of the speech signal of Gujarati sentence "**એક રાજા હતો**" used in this study is given in Figure 1.

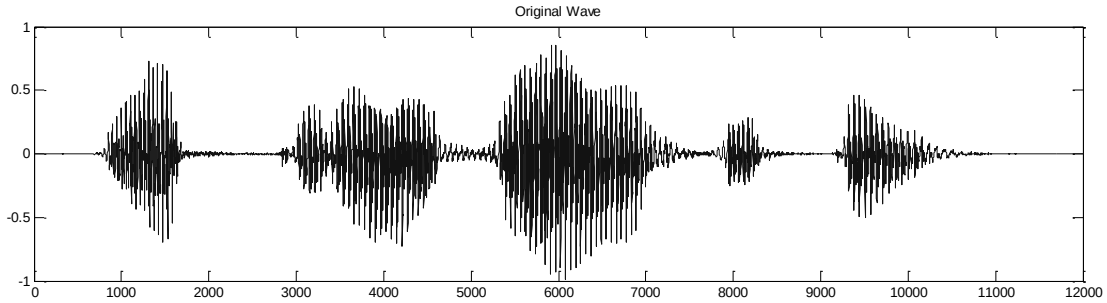

Figure 1: Original speech signal of Gujarati sentence "**એક રાજા હતો**".

The algorithm proposed in section 3 is applied on a speech signal. This speech signal is analyzed frame-by-frame manner into a non-overlapping frame of samples. It is processed into frame by frame until the entire speech signal is covered. Figure 2 shows the boundary of voiced region is marked automatically by our proposed algorithm.

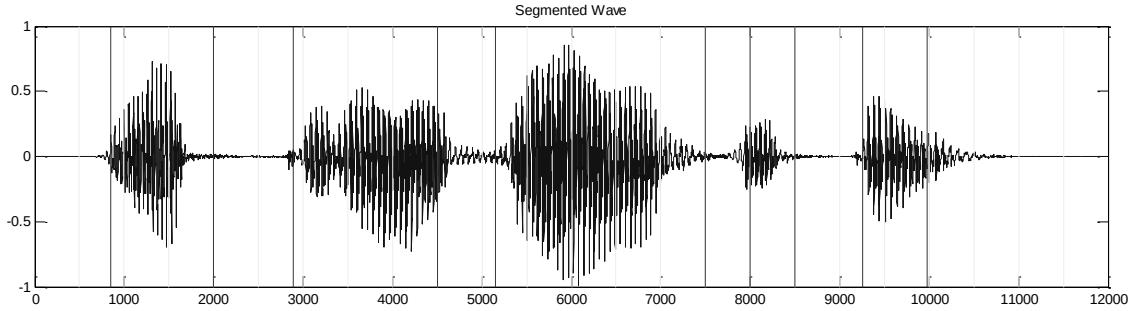

Figure 2: Segmented speech signal of Gujarati sentence "**એક રાજા હતો**"

During the analysis of speech signal, we set the frame size as 500 samples. Table 1 displays frame number, frame range, number of zero-crossing rate, number of peaks and maximum value of amplitude in a particular frame of speech signal. Last column of the table shows silent, voiced, unvoiced or voiced/unvoiced status of speech signal in each individual frames. This decision is taken on the basis of three features obtained from a particular frame of speech signal. Silent status denotes no sound, unvoiced status specifies that there is no sound or noise in a frame, voiced status specifies that a frame comprises of sound and voiced/unvoiced status denotes that there is some extent sound or no sound/noise in a frame.

[Table 1: voiced/unvoiced decision obtained using proposed model]

Frame Number	Frame Range	No. of Zero-crossing rate	No. of Peaks	Max. amplitude	Status
1	1-500	136	0	0.0052	Unvoiced
2	501-1000	77	10	0.3637	Voiced/unvoiced
3	1001-1500	97	113	0.7267	Voiced
4	1501-2000	108	30	0.6516	Voiced/unvoiced
5	2001-2500	154	100	0.0157	Unvoiced
6	2501-3000	106	10	0.0567	Voiced/unvoiced
7	3001-3500	50	30	0.3827	Voiced
8	3501-4000	74	41	0.5239	Voiced
9	4001-4500	80	44	0.4389	Voiced
10	4501-5000	45	12	0.3309	Voiced/unvoiced
11	5001-5500	100	29	0.4617	Voiced/unvoiced
12	5501-6000	73	57	0.8560	Voiced
13	6001-6500	66	50	0.8116	Voiced
14	6501-7000	58	57	0.5392	Voiced
15	7001-7500	43	15	0.2224	Voiced/unvoiced
16	7501-8000	72	8	0.2328	Voiced/unvoiced
17	8001-8500	51	19	0.2810	Voiced/unvoiced
18	8501-9000	88	0	0.0181	Unvoiced
19	9001-9500	105	16	0.4641	Voiced/unvoiced
20	9501-10000	53	30	0.4306	Voiced
21	10001-10500	31	13	0.1824	Voiced/unvoiced
22	10501-11000	51	0	0.0302	Unvoiced
23	11001-11500	0	0	0	Silent
24	11501-12000	0	0	0	Silent

Voiced/unvoiced region of the frame is further divided into sub frames and analyze the features of speech signal till boundary of voiced region within the frame is encountered. The proposed algorithm also computes the length of voiced segment and then determines whether the segment is of specified length or not. If it is then the voiced segment is accepted and starts and end boundary of the speech signal is determined otherwise that segment is discarded automatically by the proposed model. This process is repeated for entire speech signal.

5. Conclusion

We have presented an approach for separating the voiced /unvoiced region of continuous speech in a simple and efficient way. The algorithm shows good results in classifying the speech as we segmented speech into many frames. In our future study, we plan to improve our results for voiced/unvoiced discrimination in noise.

Acknowledgement

We acknowledge UGC for Special Assistance Program (SAP) for Natural Language Processing and Data Mining (file number is F.3-48/2011) under which this work has been done.

References

[1] Atal B., Rabiner L., "A pattern recognition approach to voiced-unvoiced-silence Classification with applications to speech recognition", Acoustics Speech and Signal Processing IEEE Transactions (1976) 201 - 212.

[2] T.Nagarajan, Hema A. Murthy, Rajesh M. Hegde, "Segmentation of speech into syllable-like units", EuroSpeech (2003) 2893-2896.

[3] Anupriya Sharma, Amanpreet Kaur, "Automatic Segmentation of Punjabi Speech Signal using Group Delay", Global Journal of Computer Science and Technology Software & Data Engineering (2013) 7-10.

[4] Md. Mijanur Rahman, Md. Al-Amin Bhuiyan, "Dynamic Thresholding on Speech Segmentation", International Journal of Research in Engineering and Technology (2013) 404-411.

[5] V. Kamakshi Prasad, T. Nagarajan, Hema A. Murthy, "Automatic segmentation of continuous speech using minimum phase group delay functions", Speech Communication 42 (2004) 429–446.

[6] M.Malcangi, "Softcomputing approach to segmentation of speech in phonetic units", International Journal of Computers and Communications (2009) 41-48.

[7] Md. Mijanur Rahman, Md. Al-Amin Bhuiyan, "Continuous Bangla Speech Segmentation using Short-term Speech Features Extraction Approaches", International Journal of Advanced Computer Science and Applications (2012) 131-138.

[8] T.Jayasankar, R.Thangarajan, J.Arputha Vijaya Selvi, "Automatic Continuous Speech Segmentation to Improve Tamil Text-to-Speech Synthesis", International Journal of Computer Applications (2011) 31-36.

[9] Nipa Chaudhury, Md. Abdus Sattar, Anup Kanti Bishwas, "Separating Words from Continuous Bangla Speech", Global Journal of Computer Science and Technology, 172-175.

[10] J. Sangeetha, S. Jothilakshmi, "Robust Automatic Continuous Speech Segmentation for Indian Languages to Improve Speech to Speech Translation", International Journal of Computer Applications (2012) 13-16.