Robust Automatic Continuous Speech Segmentation for Indian Languages to Improve Speech to Speech Translation

J. Sangeetha
Department of CSE
Annamalai University
Chidambaram 608 002

S. Jothilakshmi
Department of CSE
Annamalai University
Chidambaram 608 002

ABSTRACT
This paper provides an analysis of phrase and word boundary detection in a background of noise, which occurs in the context of Automatic Recognition System (ASR) and Text-To-Speech (TTS) synthesis systems for Indian languages. ASR and TTS are the major components in Speech To Speech Translation (STST) system. Both are always need a speech signal to be segmented into some basic units like phrases, words, phonemes and syllables. Normal speech is a continuous sequence of sounds with no specific pause to indicate word boundaries. Hence to convert speech into corresponding text, it is necessary to identify the boundaries and phrases present in the continuous speech signal. In this work a robust algorithm for automatic continuous speech segmentation for Indian languages using short time energy and zero crossing rates has been proposed. This proposed method has been tested on various speakers' speech in four different Indian languages such as Tamil, Telugu, Hindi and Malayalam. The results shown to be computationally efficient for real time applications and it performs better than conventional methods for speech samples collected from noisy as well as noise free environment.

Index Terms—Automatic Segmentation, Indian languages, Short Time Energy, Zero Crossing Rate.

1. INTRODUCTION
The automatic segmentation of speech especially in real world noisy environment is a challenging problem. Most importantly, the efficiency achieved in automatic detection of speech boundaries largely determines the accuracy of the recognition as well as synthesis systems. Even minor improvement in speech boundary detection front-end greatly influences the overall system accuracy in the long run. For the isolated word recognition in a limited vocabulary scenario, this problem boils down to the determination of the correct isolated word boundary and the rejection of the speech artifacts such as breath, mouth and lip clicks etc. For the connected speech case, the problem is to get rid of intra-word silences and any other artifacts as mentioned in the previous case. For a continuous speech recognition engine, efficient automatic speech segmentation pre-processor can reduce the computational load and power consumption of the system.

There are several ways of classifying (labeling) events in speech. It is accepted convention to use a three-state representation in which states are (i) silence (*S*), where no speech is produced; (ii) unvoiced (*U*), in which the vocal cords [6] are not vibrating, so the resulting speech waveform is a periodic or random in nature and (iii) voiced (*V*), in which the vocal chords are tensed and therefore vibrate periodically when air flows from the lungs, so the resulting waveform is quasi-periodic [7]. 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 (like /*f*/ or /*th*/) from silence, or weak voiced sound (like /*v*/ or /*m*/) from unvoiced sounds or even silence.

However, it is usually not critical to segment the signal to a precision much less than several milliseconds; hence, small errors in boundary locations usually have no consequence for most applications. Since for most of the practical cases the unvoiced part has low energy content and thus silence (background noise) and unvoiced part is classified together as silence/unvoiced and is distinguished from voiced part. In this proposed work two widely accepted methods namely Short Time Energy (STE) [7], [8] and Zeros Crossing Rate (ZCR) [7], [8] have been used for segmentation process. To extract the features from the speech signal, the signal must be pre-processed and divided into successive windows or analysis frames.

This paper is organized as follows. In Section 2 we describe the theoretical background. Section 3 presents the algorithm for automatic segmentation along with a short discussion regarding computational complexity and performance measures. The results are presented in Section 4, and Section 5 concludes this paper.

2. THEORITICAL BACKROUND
Short-Time Energy and Zero-Crossing Rate
Two basic short-time analysis functions useful for speech signals [9] are the short-time energy and the short-time zero-crossing rate. These functions are simple to compute, and they are useful for estimating the properties of the excitation function in the model.

The short-time energy is defined as

$$E_{\bar{n}} = \sum_{m=-\infty}^{\infty} (x[m]\omega[\hat{n}-m])^2 = \sum_{m=-\infty}^{\infty} x^2[m]\,\omega^2[\hat{n}-m] \quad (1)$$

As shown in the Eqn (1), it is often possible to express short-time analysis operators as a convolution or linear filtering operation. In this case, $E_{\bar{n}} = x^2[n] * h_e[n]|_{n=\bar{n}}$ where the impulse response of the linear filter is $h_e[n] = w^2[n]$.

Similarly, the short-time zero crossing rate is defined as the weighted average of the number of times the speech signal changes sign within the time window. Representing this operator in terms of linear filtering leads to

$$Z_{\hat{n}} = \sum_{m=-\infty}^{\infty} 0.5 \left| \begin{matrix} \text{sgn}\{x[m]\} \\ -\text{sgn}\{x[m-1]\} \end{matrix} \right| \omega[\hat{n} - m], \qquad (2)$$

Where

$$\text{sgn}\{x\} = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases} \qquad (3)$$

Since $0.5|\text{sgn}\{x[m]\} - \text{sgn}\{x[m-1]\}|$ is equal to 1 if $x[m]$ and $x[m-1]$ have different algebraic signs and 0 if they have the same sign, it follows that $Z_{\hat{n}}$ in the Eqn (2) is a weighted sum of all the instances of alternating sign (zero-crossing) that fall within the support region of the shifted window $w[\hat{n} - m]$.

Fig.1 shows an example of the short-time energy and zero-crossing rate for a segment of speech with a transition from unvoiced to voiced speech. In both cases, the window is a Hamming window (two examples shown) of duration 25 ms (equivalent to 401 samples at a 16 kHz sampling rate). Thus, both the short-time energy and the short-time zero-crossing rate are output of a low pass filter.

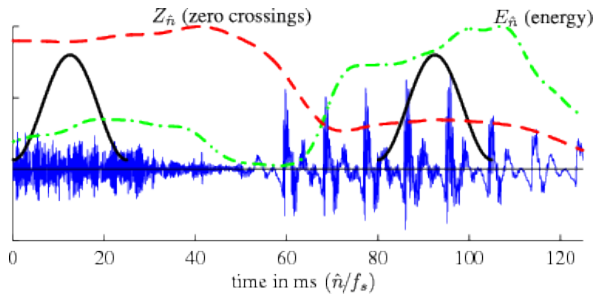

Fig.1 Section of speech waveform with short-time energy and zero-crossing rate superimposed.

Note that during the unvoiced interval, the zero-crossing rate is relatively high compared to the zero-crossing rate in the voiced interval. Conversely, the energy is relatively low in the unvoiced region compared to the energy in the voiced region. Note also that there is a small shift of the two curves relative to events in the time waveform. This is due to the time delay of M samples (equivalent to 12.5 ms) added to make the analysis window filter causal.

The short-time energy and short-time zero-crossing rate are important because they abstract valuable information about the speech signal, and they are simple to compute. The short-time energy is an indication of the amplitude of the signal in the interval around time $\hat{n}$. From our model, we expect unvoiced regions to have lower short-time energy than voiced regions. Similarly, the short-time zero-crossing rate is a crude frequency analyzer. Voiced signals have a high frequency (HF) falloff due to the low pass nature of the glottal pulses, while unvoiced sounds have much more HF energy. Thus, the short-time energy and short-time zero-crossing rate can be the basis for an algorithm for making a decision as to whether the speech signal is voiced or unvoiced at any particular time $\hat{n}$. A complete algorithm would involve measurements of the statistical distributions of the energy and zero-crossing rate for both voiced and unvoiced speech segments (and also background noise distributions). These distributions can be used to derive thresholds used in voiced/unvoiced decision.

3. IMPLEMENTATION

A. Proposed Method

In speech signal processing, two basic parameters are zero crossing rate (ZCR) and short time energy. The energy parameter has been used in endpoint detection since the 1970's. By combining with the ZCR, the detection processes can be made very accurate. The beginning and ending for each utterance can be detected.

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 this system will be speech and output will be the segmented speech and waveform representing the boundaries.

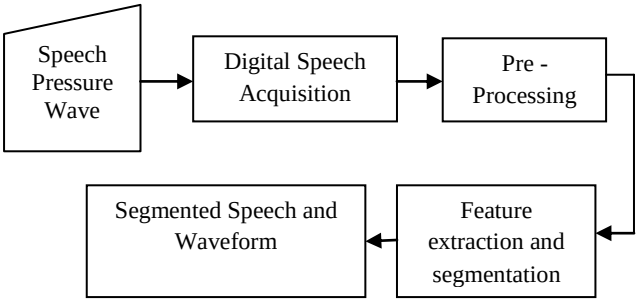

Fig.2 Block Diagram of the Proposed Speech Segmentation Algorithm

Automatic speech segmentation system has three major steps:

a) Digital speech acquisition
b) Signal Pre-processing
c) Feature Extraction and Segmentation

a. Digital Speech Acquisition

Digital speech acquisition is acquiring of the analog speech signal which is pressure wave through the microphone and obtaining digital representation of speech signal. Speech capturing or speech recording is the first step of implementation. For the proposed algorithm the sampling frequency is 8 KHz; sample size is 8 bits, and mono channel is used.

b. Signal Pre-processing

It is very crucial to pre-process the speech signal in the applications where silence or background noise is completely undesirable.

1. Pre-emphasis

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.

2. 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. Continuous speech signal has been blocked into *N* samples, with adjacent frames being separated by *M* (*M*<*N*). In our work, after the Pre-emphasis, filtered samples have been converted into frames, having frame size of 25 msec. Each frame overlaps by10 msec.

3. 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 in this work.

c. Feature extraction and Segmentation

The short time energy and zero crossing rates are used to process speech samples to accomplish the proposed segmentation method. Following procedure has been used for automatically marking the boundaries in sound file.

a) Short term energy and zero crossing rates are computed for the preprocessed frames.
b) Some threshold value which is dynamically generated has been taken and signals 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 boundary.
d) After getting the starting location, the 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 boundary. Once an end point has been detected, we can precede analyzing signal from end point of first one looking for the starting position of next one.

3. B. PERFOMANCE MEASURES

Percentage of correctness regarding extraction of voiced sample from a speech signal is defined as follows:

$$Hit\ rate = \frac{No.correctly\ identified\ words}{No.of\ word\ boundaries\ in\ utterance}$$

$$False\ alarm\ rate = \frac{No\ of\ erroneous\ word\ boundaries\ identified}{No.of\ word\ boundaries\ in\ utterance}$$

4. RESULTS AND DISCUSSION

The proposed technique has been implemented in Matlab 2012a.Various speakers' speech in four different Indian languages such as Tamil, Telugu, Hindi and Malayalam have been recorded and segmented. Proposed method has been implemented and analyzed for four different languages' speech signals.Table 1 summarizes the results showing percentage of correctness in detection of word boundaries for four Indian languages. Figure 3 & 4 shows the boundaries of signal that are marked automatically from the energy and zero crossing rate.

5. CONCLUSION

In this paper we have presented an alternate method for robust automatic segmentation for Indian languages speech. It can be concluded that the algorithm in this paper performs the segmentation of word and sentence boundaries automatically.

The performance of the proposed system has been analyzed for various speech signals belongs to four different Indian languages such as Tamil, English, Hindi and Malayalam.

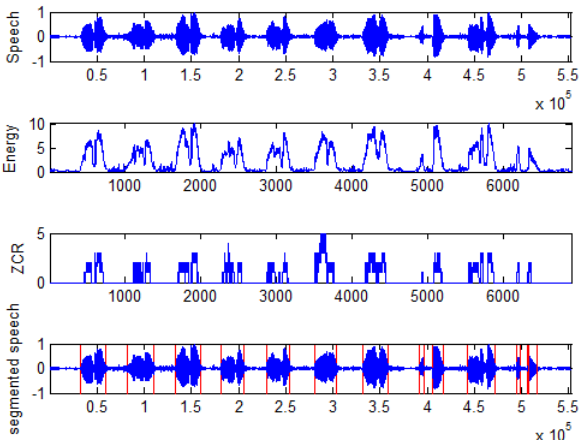

Fig.4 Word boundaries detection

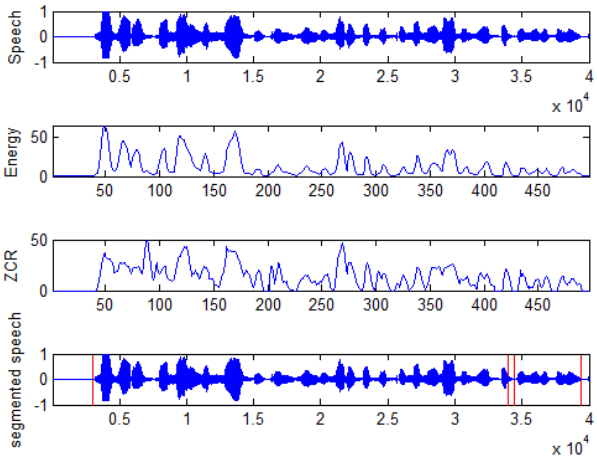

Fig.4 Phrase boundaries detection

Table 1: Performance measures

Sl. No	Language	Total No. of words present	No. of Correctly Identified	Hit Rate	False Alarm Rate
1	Tamil	50	47	92%	8%
2	Telugu	50	46	92%	8%
3	Hindi	50	45	90%	10%
4	Malayalam	50	43	86%	14%

The experimental result shows that the proposed method exhibits the better performance for word and phrase boundary detection of Indian languages speech. This is an alternate method for segmentation of speech signals for speech to speech translation system which tends to produce the better results in the noisy as well as noiseless environment.

6. REFERENCES

[1] Jayasankar.T, Dr. R. Thangarajan, Dr.Arputha Vijayaselvi .J " Automatic continuous speech segmentation to improve Tamil text to speech a synthesis system", International Journal of Computer Applications (0975 – 8887) Volume 25 No.1, July 2011.

[2] Er. Amanpreet Kaur and Er.Tarandeep Singh, "Segmentation of Continuous PunjabiSpeech signal into syllables", Proceedings of the World Congress on Engineering and Computer Science 2010 Vol IWCECS 2010, October 20-22, 2010, SanFrancisco,USA.

[3] G.Lakshmisarada, A.Lakshmi, Hema A Moorthy and T.Nagarajan, "Automatic transcription of continuous speech into syllable-like units for Indian languages", Sadhana Vol. 34, Part 2, April 2009, pp. 221–233. © Printed in India.

[4] T.Nagarajan, H.A.Murthy "Subband –Based Group Delay Segmentation Spontaneous Speech into Syllable like Units" EURASIP JOURNAL on Applied signalprocessing 2004.

[5] Deller J. R. Jr., Hansen J. L. H. and Proakis J. G.: "Discrete Time Processing of Speech Signals", IEEE Press, NJ, 2000.

[6] K. Ishizaka and J.L Flanagan, "Synthesis of voiced Sounds from a Two-Mass Model of the Vocal Chords," Bell System Technical J., 50(6): 1233-1268, July-Aug., 1972.

[7] Atal, B.; Rabiner, L., "A pattern recognition approach to voiced-unvoiced-silence Classification with applications to speech recognition" Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on , Volume: 24 , Issue: 3 , Jun 1976, Pages: 201 - 212.

[8] G. Childers, M. Hand, J. M. Larar, "Silent and Voiced/Unvoiced/ Mixed Excitation(Four-Way), Classification of Speech", IEEE Transaction on ASSP, Vol-37, No-11, pp. 1771-74, Nov 1989. 9.

[9] http://www.nowpublishers.com/product.aspx?product=SIG&doi=2000000001&section=x1-56r1