Automatic Continuous Speech Segmentation to Improve Tamil Text-to-Speech Synthesis

T.Jayasankar
Asst Professor/ECE,
Anna University of Technology, Trichy,

Dr R.Thangarajan
Professor/IT
Kongu Engg college, Erode

Dr.J.Arputha Vijaya Selvi
Professor/ECE
Kings College of Engg, Pudukkottai

ABSTRACT

Emerging growth of information and communication technologies has influenced the research trends to focus on speech technologies. This paper we have investigated the development of automatic Tamil speech segmentation system which would act as test bed as well as foundation for several speech applications involving Tamil language. This study proposes an algorithm for automatic segmentation of Tamil voiced speech. The calculations of absolute energy and zero crossing rates are used to process speech samples to accomplish the segmentation. The Algorithm are written and compiled using Matlab.

Keywords

Speech Segmentation, Tamil, Zero- crossing rate.

1. INTRODUCTION

He automatic continuous speech segmentation is essential for the development of Text –to –speech (TTS) system and Speech Recognition system [1].Recent speech synthesis and speech recognition system rely on large amount of segmented speech corpora to realize high quality synthetic speech. If it has to be done by humans, segmentation of speech is labor-intensive work that requires large cost and long time. For recent TTS systems requiring a speech corpus of 10 or more hours, hand segmenting the entire corpus is economically impractical. Therefore, a method is needed as an alternate to hand labeling which can segment the speech automatically segments the speech signals into any fundamental acoustic units.

Based upon a concatenative synthesizer whose unit inventory is generated by cutting speech segments from a database recorded by a target speaker [2][3][5]. There are typical three phases in the process of building a unit inventory:

1. Determine the synthesis units and derive the conversion between a phoneme string and a unit string.

2. Segmentation of each unit from spoken speech.

3. Selection of one (or a few) good unit instance when many are available in the corpus.

Manual segmentation and unit selection phases are typically very labor-intensive for concatenative synthesizers because it involves subjective judgement for thousands of synthesis units. They are often based on a trial and error process, which doesn't usually address the potential distortion at any concatenation point. Moreover, it is very difficult to optimize the segmentation and unit selection phases with the choice of synthesis units.

This study is focused on continuous Tamil number speech and sentence recognition with the intention to distinguish speech and non-speech segments. This study proposes an algorithm for automatic segmentation of Male and Female voiced speech. The calculations of short time energy and zero rate crossing are used to process speech samples to accomplish the segmentation.

The rest of the paper is organized as follows. Section 2 describes the nature of Tamil scripts. Section 3 discusses the method and implementation for *Tamil* speech segmentation. The experimental results of the evaluation are presented in Section 4 .Section 5 concludes this work.

2. NATURE OF TAMIL SCRIPTS

Tamil is a very ancient language with a rich heritage and literature. A character in Indian language scripts is close to a syllable and can be typically of the form: C∗VC∗, where C is a consonant and V is a vowel. There is fairly good correspondence between what is written and what is spoken. Typically there are about 35 consonant and 18 vowel characters. However, in Tamil there are fewer characters than many of the other Indian languages. There are 13 vowels and 18 consonants characters. Some of the consonants have more than one pronunciation and in effect there are 41 phones.[5]

3. METHODS AND IMPLEMENTATION

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 [7]. By combining with the ZCR, the detection process can be made very accurate [8]. The beginning and ending for each utterance can be detected.

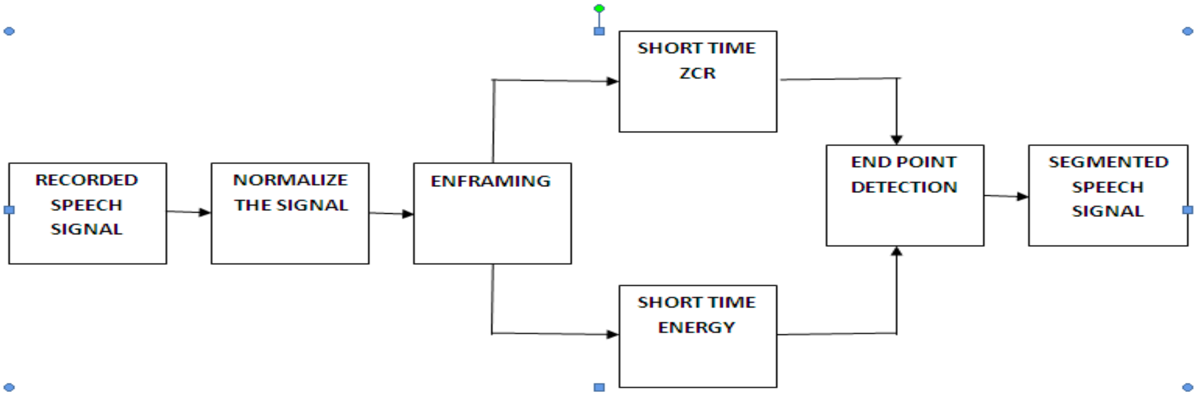

Fig 1.Automatic speech segmentation system

3.1 Enframing:

Enframe can be used to split a signal up into frames. It can optionally apply a window to each frame. Framing: decompose the speech signal into a series of overlapping frames In sense, the speech region has to be short enough so that it can reasonably be assumed to be stationary in that region: i.e., the signal characteristics (whether periodicity or noise-like appearance) are uniform in that region. Frame duration ranges are between 10 ~ 25 ms in the case of speech processing.

3.2 Short Time Energy:

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. And also the short time energy of the speech signal provides convenient representation that reflects these amplitude variations. Basic short time analysis functions useful for speech signals are the short time energy. This function to compute, and is useful for estimating properties of the excitation function in the model.

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m)\, h(n-m) \qquad (1)$$

Equation 1 defines the short time energy for a sampled signal where h(n-m) is a windowing function. For simplicity a rectangular windowing function is used as defined in equation 2.

$$H(n) = 1 \qquad 0 \le n \le N-1 \qquad (2)$$

$$= 0 \qquad \text{Otherwise}$$

Where, N in equation 2 is the length of the window i**n samples.**

3.3 Zero Crossing Rate:

The short time average zero crossing rate of a speech signal can be used in conjunction with the short time average energy (or magnitude) to discriminate between voiced speech, unvoiced speech and silence. The short time average crossing rate of a digitally sample speech signal is defined in digital Processing of Speech Signal.

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

Where

$$\text{sgn}[x(n)] = 1 \qquad x(n) \ge 0 \qquad (4)$$

$$= -1 \qquad x(n) < 0$$

and w(n) is the windowing function with a window size of N samples:

$$w = {}^{1}\!/_{2}\, N \qquad 0 \le n \le N-1 \qquad (5)$$

$$= 0 \qquad \text{Otherwise}$$

3.4 Algorithm for Speech Segmentation:

1. Taking frames from the recorded wave file.
2. Compute the Zero Crossing Rate and absolute energy for all frames taken from the recorded wave file.
3. Adjust the energy threshold for end point detection.
4. Find the upper level of end point for all words that are to be recorded.
5. Find the lower level of end point for all words that are to be recorded.

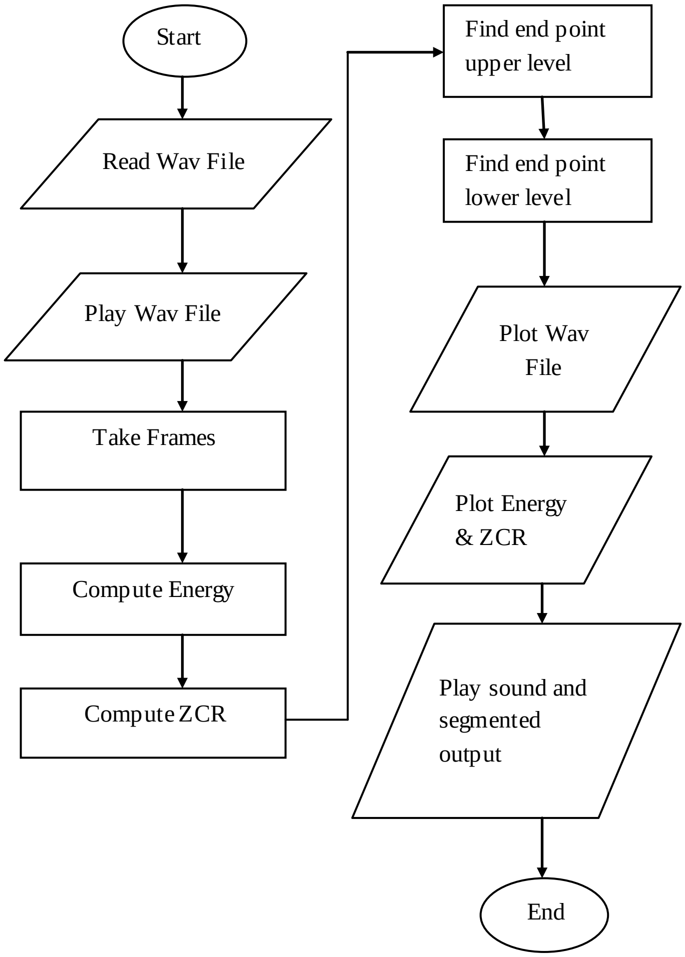

Fig 2: Flow Chart

5. RESULTS AND DISCUSSION

Technique has been implemented in Matlab. Various speech signals in Tamil have been recorded and segmented. Proposed method has been implemented and analyzed for different Tamil speech signals. Results have been shown fig 3&4 for signal where the boundary of syllables are marked automatically from that energy and zero-crossing rate.The wave file contains the following sentence

1.ஒன்று இரண்டு மூன்று நான்கு ஐந்து ஆறு ஏழு எட்டு ஒன்பது பத்து (Table-3)

2.இன்றே செய் அதை நன்றே செய் (Table 4)

6.CONCLUSION

It can be concluded that algorithm in this research segmentation is done automatically in the given speech signal which in turn represents the starting and end point of syllable respectively. The table 3 and 4 represents analyze accuracy for automatic segmentation of various male and female speakers. The algorithms manage to get 85% correct segmentation for speech recorded in a quiet.

7. REFERENCES

[1] R.Thangarajan, Natarajan A. M "Syllable Based Continuous Speech Recognition for Tamil", in South Asian language review, VOLXVIII No 1, 2008

[2] Nageshwara Rao,S. Thomas, T. Nagarajan and Hema A. Murthy,"Text-to-speech synthesis using syllable like units, proceedings of National Conference on Communication (NCC) 2005, pp. 227-280, IIT.

[3] Samuel Thomas, M. Nageshwara Rao, Hema A. Murthy and C.S. Ramalingam, "Natural Sounding TTS based on Syllable-like Units", proceedings of 14th European Signal Processing Conference, Florence, Italy, Sep2006.

[4] Jayasankar.T, Arputha Vijayaselvi .J "Realization of Tamil Syllables Text To Speech Transferring System using FPGA" in Tamil Internet 2010, Coimbatore.

[5] Venugopalakrishna.Y.R. et.al., " Design and Development of a Text-To-Speech Synthesizer for Indian Languages", pp. 259-262,proceedings of National Conference on Communication (NCC) 008,IIT-Bombay,February 2008.

[6] A.G.Ramakrishnan, "Thirukural text to speech synthesis system", proco.Tamil Internet 2001, Kuala Lumpur.

[7] Rabiner, L.R. and Sambur, M.R., 1975."An Algorithm for Determining the Endpoints of Isolated Utterances", BellSjjsl. Tech. J., Vol. 54, pp.297-3 1 5.

[8] Rabiner, L.R.and Schafer, R.W., 1978 Digital Processing of Speech Signals, Prentice-Hall Inc.

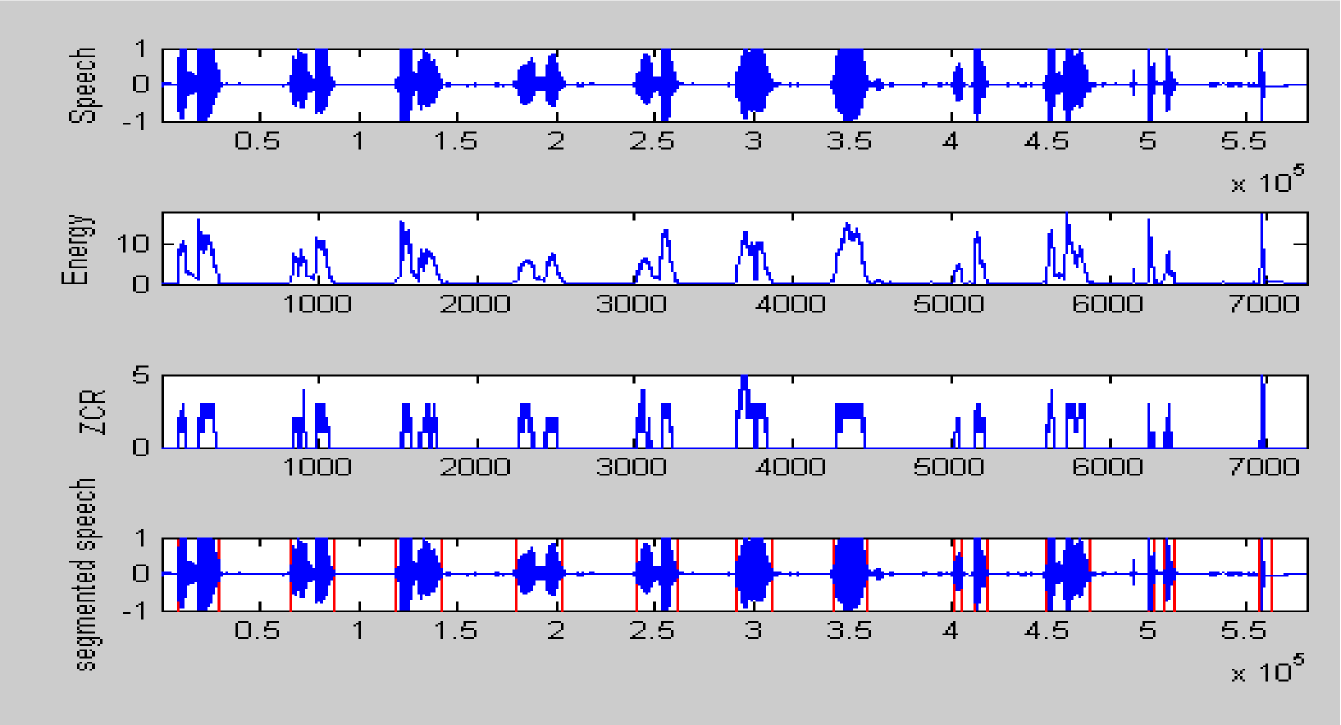

Fig 3: The wave form, Energy and Zero-crossing rate for continuous Tamil number spoken in wave file recorded from speaker # 1.

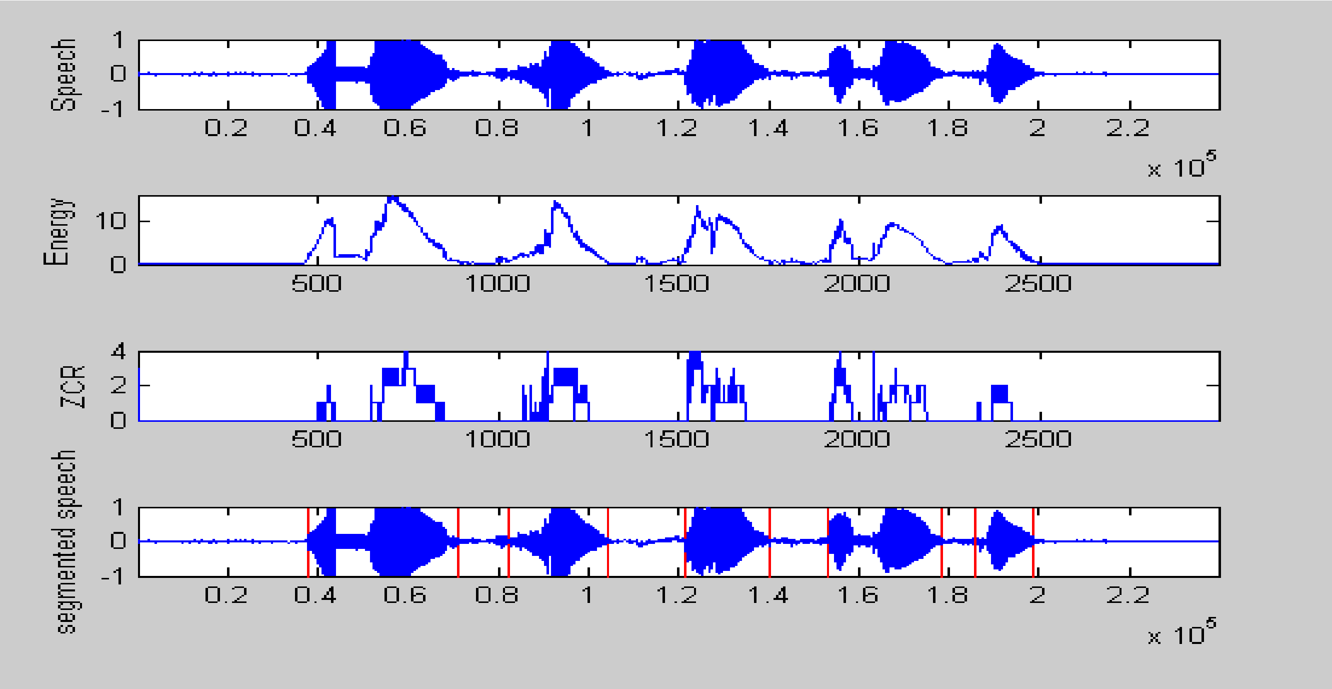

Fig 4: The wave form, Energy and Zero-crossing rate for continuous Tamil number spoken in wave file recorded from speaker # 1.

Segment	Original	Male speaker			Female speaker		
		1	2	3	1	2	3
1	ஒன்று	ஒன்று	ஒன்று	ஒன்று	ஒன்று	ஒன்று	ஒன்று
2	இரண்டு	இரண்டு	இரண்டு	இர	இரண்டு	இரண்டு	இரண்டு
3	மூன்று	மூன்று	மூன்று	ண்டு	மூன்று	மூன்று	மூன்று
4	நான்கு	நான்கு	நான்கு	மூன்று	நான்கு	நான்கு	நான்கு
5	ஐந்து	ஐந்து	ஐந்து	நான்கு	ஐந்து	ஐந்து	ஐந்து
6	ஆறு	ஆறு	ஆறு	ஐந்து	ஆறு	ஆறு	ஆறு
7	ஏழு	ஏழு	ஏழு	ஆறு	ஏழு	ஏழு	ஏழு
8	எட்டு	எட்	எட்	ஏழு	எட்	எட்	எட்
9	ஒன்பது	டு	டு	எட்	டு	டு	டு
10	பத்து	ஒன்பது	ஒன்பது	டு	ஒன்பது	ஒன்பது	ஒன்பது
11		பத்	பத்	ஒன்பது	பத்	பத்	பத்
12		து	து	பத்	து	து	து
13				து			
Total segment correct		**8/10**	**8/10**	**7/10**	**8/10**	**8/10**	**8/10**
Correct over group		**76.66 %**			**80 %**		
Overall accuracy		**78.33 %**					

Table 3: Segmentation of Recorded Tamil number Wave File from 4 Male & 4 Female Speakers

Segment	Original	Male Speaker				Female Speaker			
		1	2	3	4	1	2	3	4
1	இன்றே	இன்றே	இன்றே	இன்றே	இன்றே	இன்றே	இன்றே	இன்றே	இன்றே
2	செய்	செய்	செய்	செய்	செய்	செய்	செய்	செய்	செய்
3	அதை	அதை	அதை	அதை	அதை	அதை	அ	அ	அதை
4	நன்றே	நன்றே	நன்றே	நன்றே	நன்றே	நன்றே	தை	தை	நன்றே
5	செய்	செய்	செய்	செய்	செய்	செய்	நன்றே	நன்றே	
6							செய்	செய்	
Total Segment Correct		5/5	5/5	5/5	5/5	5/5	4/5	4/5	4/5
Correct Over Group		100 %				85 %			
Overall Accuracy		92.5 %							

Table 4: Segmentation of Recorded Tamil sentence Wave File from 4 Male & 4 Female Speakers