Modified Segmentation Algorithm Based on Short Term Energy & Zero Crossing Rate for Maithili Speech Signal

Sudhakar Kumar
Dept. of CSE,
MAKAUT,
West Bengal, Kolkata, India
Email: sudhakar06111990@gmail.com

Santanu Phadikar
Dept. of CSE,
MAKAUT,
West Bengal, Kolkata, India
Email: sphadikar@yahoo.com

Koushik Majumder
Dept. of CSE,
MAKAUT,
West Bengal, Kolkata, India
Email: koushikzone@yahoo.com

***Abstract*—Converting the voice signals (uttered in Maithili language) to text has lots of applications including speaker identification, voice mode interaction, bio-metric identification etc. Most of the research papers reported in literature, till date, concentrated mainly in English. Some works are also available for the languages Hindi, Bengali and Tamil. But working with regional languages like Maithili, Bhojpuri and Magahi etc remains untouched. So, in this research work one of the most popular regional language of Bihar state under India territory, Maithili language is selected for our case study. In order to segment a voice signal firstly the word boundary is detected from the voice signals of sentences using Short Term Energy (STE) and Zero Crossing Rate (ZCR). Since there exist a large number of words for the Maithili language; converting the words directly to a text is a complex and tedious job. Hence words are further dived into syllables, which are mostly unique in terms of signal and in manageable in size. Detection of rough syllable boundary from the word boundary has been performed by group delay algorithm. To make these syllable boundaries more accurate group delay algorithm has been modified by considering the differences of consecutive peaks in the negative derivative phase. Accuracy is quantified through error rate measured by taking the differences between the ground truth boundary of the signals (determine manually) and system detected boundary of the syllable signals. Experiment using the data set of 25 different sentences from 10 different speakers have been performed and an accuracy rate of 85.62% is computed using above algorithm.**

Keywords: Maithili language, group delay, speech segmentation, short term energy, speech recognition, voice to text, zero crossing rate.

I. Introduction

India is a multilingual country with twenty-two national languages [1] in which Maithili is one of the famous languages. Work has been proposed for development of voice to text in Maithili Language. Steps used for converting voice signal to text are acquisition of speech signal, pre-processing [3], framing and segmentation [4]. Among all these steps, speech segmentation is an important step as it determines the overall accuarcy by measuring starting and ending index of either words or syllables. Segmentation of Maithili words from sentences have been done by using energy based algorithm i.e. short term energy (STE) and zero crossing rate (ZCR) [5]. On the other hand, calculation of syllable boundary from the word boundary is based on group delay algorithm.

Studying literatures it is observed that entropy based segmentation algorithms provide the better result [6] in noisy environment compare to the energy based methods. But, the use of short term energy and Zero Crossing Rate under energy based segmentation methods boosts the performance if the energy difference is high between voiced and unvoiced part.

Due to the presence of high level voice portion with high energy in nucleus unit, strong fricatives and fixed valley pattern of onset and coda on left and right side of the nucleus in case of syllables of the acquired signals lead to select the energy based algorithm for segmentation. In the paper [2] syllable unit is identified for punjabi language using group delay algorithm. They have considered the differences of consecutive peaks for computing short term information present in negative derivative phase using group delay algorithm.

All methods of segmentation discussed so far, mainly focused on segmentation using word or syllable separately but not in combined. Also, existing group delay algorithm considers only positive peak for computation without considering the negative peaks, which is also present in all real speech data. thus to overcome these drawbacks of the above mentioned algorithms a novel method has been proposed. The proposed method determines the word boundaries using STE & ZCR and then determines the syllable boundaries using the modified group delay algorithm which consider both the positive as well as negative peaks of the signals for computation.

II. Proposed Work

Segmenting the voice signals (uttered in Maithili language) has been performed in several steps. In the first step, Maithili voice is acquired in the form of sentence from different speakers of age group (25-30) years. As these sentences are recorded in real time so, it must contain useless components like background noise and lower frequency components, which

978-1-5090-4291-3/16/$31.00 ©2016 IEEE

must have to be removed. Software named audacity 2.0.6 has been used for immediate removal of background noise. After observing this intermediate signal nature, it has been found that some noise of lower frequency components with un-desired frequencies still persist in the system. Thus to remove this noise, Pre-emphasis method [8] is applied. Here, amplitude and frequency properties of the acquired signal are being changed intentionally to reduce unwanted effect of the noise.

It is also observed that, speech signal is always continuous in nature. Thus to bring discontinuity and making the signal stationary, experimental speech signal is divided into several small number of frames for small period of time called framing. Post framing, segmentation is applied at different levels to extract the word and syllable from framed signals.

First level derives the word boundary. It's very common that, there is a gap containing silence portion between the two consecutive words in a sentence. Energy of silence portion is very less and of voiced portion is very high. Based on these energy difference, Short Term Energy (STE) and Zero Crossing Rate (ZCR) methods have been used for extracting the word from framed signals of sentences. Since there exists a large number of words in Maithili language. Converting the words directly to a text is a complex and tedious job.

Hence in the second level words are further divided into syllables, which are mostly unique in terms of signal and in manageable in size. Approximate syllable boundary from the word boundary has been identified by group delay algorithm. Thus to increase the performance of the system, threshold peak value is calculated under negative derivative phase. After comparing this threshold value with the phase difference of consecutive peaks, more accurate starting and ending indices of syllable are detected. Flow chart of proposed method is shown in fig. 1.

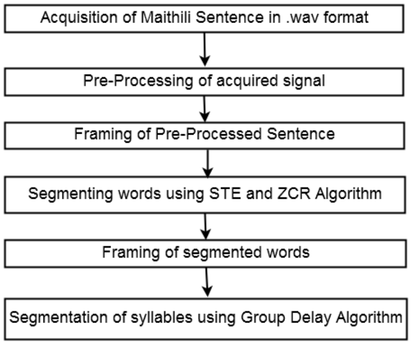
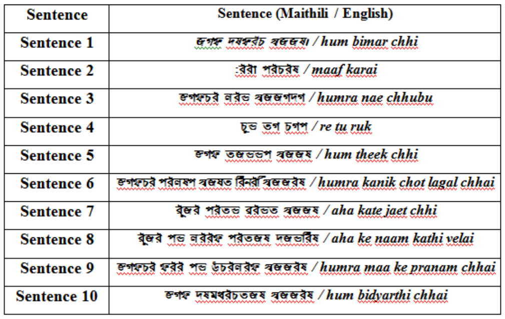

Acquisition of Maithili Sentence in .wav format
Pre-Processing of acquired signal
Framing of Pre-Processed Sentence
Segmenting words using STE and ZCR Algorithm
Framing of segmented words
Segmentation of syllables using Group Delay Algorithm

Fig. 1. Block Diagram of Proposed Segmented Method

A. Acquisition of Maithili Sentence

For the proposed work, the speech sample of Maithili sentences from age group 25-30 yrs is recorded. This has been done using Samsung GALAXY CORE 2 SM-G355H microphone in .wav format using audacity 2.0.6 sound recording software. Although 25 different Maithili sentences are recorded from 10 different speakers, five male and five female but for the simplicity, only 10 sentences are shown in Fig. 2.



Sentence	Sentence (Maithili / English)
Sentence 1	ভগক্ষ দবক্ষরঁচ ব্বজ্জবা / *hum bimar chhi*
Sentence 2	:রঁরা পরঁচরঁষ / *maaf karai*
Sentence 3	ভগক্ষচরঁ নরঁভ ব্বজ্জগদগ / *humra nae chhubu*
Sentence 4	চুভ তগ চগপ / *re tu ruk*
Sentence 5	ভগক্ষ তজ্জভভপ ব্বজ্জষ / *hum theek chhi*
Sentence 6	ভগক্ষচরঁ পরঁনষপ ব্বজষত রিঁনরিঁ ব্বজ্জরঁষ / *humra kanik chot lagal chhai*
Sentence 7	রঁুজরঁ পরঁতভ ররঁভত ব্বজ্জষ / *aha kate jaet chhi*
Sentence 8	রঁুজরঁ পভ নরঁরঁক্ষ পরঁতজ্জষ দজ্জভরিঁষ / *aha ke naam kathi velai*
Sentence 9	ভগক্ষচরঁ ক্ষরঁরঁ পভ উচরঁনরঁক্ষ ব্বজ্জরঁষ / *humra maa ke pranam chhai*
Sentence 10	ভগক্ষ দষমধরঁচতজ্জষ ব্বজ্জরঁষ / *hum bidyarthi chhai*

Fig. 2. TEN DIFFERENT MAITHILI SENTENCE

Fig. 3 shows speech signal of sentence 1 uttered by speaker1.

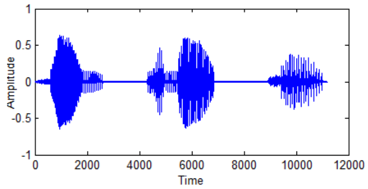


Fig. 3. Speech Signal for sentence 1

B. Pre Processing of Acquired Signal

In real world environment speech signal is analog in nature. To convert it into discrete form, speech signal is sampled at 8 KHz sampling rate. This recorded discrete signal contains background noise with some lower frequency components which are useless for speech recognition process. Software named audacity 2.0.6 has been used for removing background noise. Pre-emphasis filter [8] is applied on the processed signal to remove the some of the existing background noises. Formula for implementing this filter is defined by equation (1):

$$y\{n\} = x\{n\} - a * x\{n-1\} \qquad (1)$$

Where x{n}: input speech signal Y{n}: output signal. 'a' is a constant and the value lies in between 0.9 to 0.97. Value of a has been taken as 0.97 as it attenuates low frequencies in comparison to other values. After applying the pre-emphasis filter, the pre-processed signal corresponding to signals shown in fig. 3. is shown in fig. 4.

C. Framing of the Pre-Processed sentence

Speech Signal is time varying in nature. So, to analyze the time varying signal as stationary for small period of time, preprocessed speech signal is divided into nonoverlapping frames of 20ms time durations. Frame no. 7 of the pre-processed of fig. 4 has been illustrated in fig. 5.

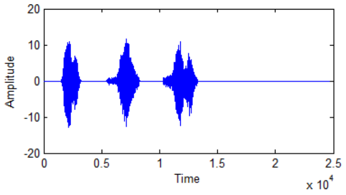


Fig. 4. Prepcrocessed Speech Signal for sentence 1

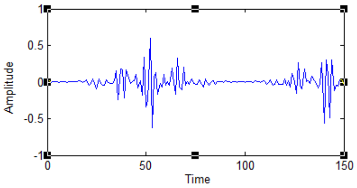


Fig. 5. Framed Speech Signal for sentence 1

D. Segmentation words using STE and ZCR Algorithm

To determine the ZCR of a frame, firstly the changes of sign in the amplitude value with in the frame is computed. After that a threshold value is computed by taking the average of ZCR of each of the frames of a sentences. Now the ZCR value of the each individual frame is compared with respect to the computed threshold value. If the ZCR of the frame is greater than the threshold value, it indicates that the frame is belongs to the voiced portion of the signal. Similarly, STE of each of the frame is computed using the equation (2). and threshold value for a sentences is also computed as the average of STE of all the frames of that sentence. The short-term energy equation can be represented as :

$$E_n = \sum_{n=-\infty}^{+\infty} \{x(n) * w(N)\}^2 \qquad (2)$$

Where E = Energy of the speech signal, x(n) = segmented speech, n=frame no, w(N) = hamming window and N = length of each frame.

Zero crossing is combined with short-term energy to determine the word boundaries of speech signal. ZCR plot for sentence 1 is shown in fig. 6.

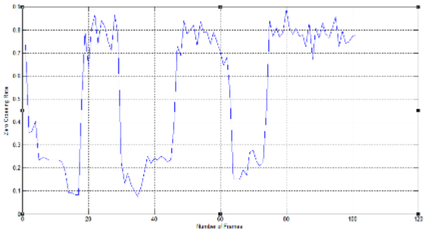


Fig. 6. ZCR for sentence 1

All the frames of a sentence whose values found greater than Threshold ZCR and Threshold STE are selected. Now, only those frames are selected which are consecutive in nature at least up to four numbers of frames. In this way, the boundary of words is obtained, which are further fed as an input for determining the syllable boundary.

E. Framing of Segmented words

Now the extracted words are further divided into several small frames of 20ms time durations. This frame size has been fixed as 20 ms after several manual testing over the speech signal so, frame size in terms of count of amplitude = (8000/20)=160.

F. Segmentation of syllable using group delay algorithm

1) After framing, STE is calculated using equation (2) further for each word with frame size 20 ms as shown in fig. 7.

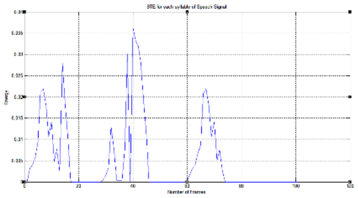


Fig. 7. STE of each syllable of sentence 1

2) Inverse first fourier transform is applied on reciprocal of all framed signal with their calculated STE energy values. These steps are performed to calculate the phase required for subsequent steps of group delay algorithm. This step not only limits the frequency range but also generates real and imaginary components.

3) Magnitude of complex to real content in frequency domain is calculated and the intermediate result is stored for calculation of negative derivate of phases of all frames as shown in below fig. 8.

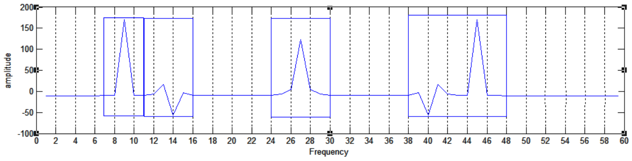


Fig. 8. Negative derivative of phases of sentence 1

4) Threshold phase is calculated by averaging all positive peaks and excluding all local maximum values.

5) Finally, differences are taken between consecutive peaks and compare with the Threshold peak if found greater then move on to the next frame and the process is repeated until reach to the boundary of the syllable. The same process is repeated for each and every word to get the boundary associated with each and every syllable corresponds to that word as shown in Fig. 9.

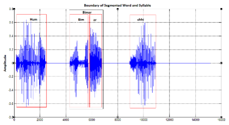


Fig. 9. Segmented Word and syllable of Speech signal for sentence 1

All the above experiments have been implemented in MATLAB R2013a.

III. Result and Discussion

The experiment is carried out using the data set of 25 different Maithili sentences (selected randomly) collected from 10 different speakers of different age group and different sex. Then the boundary of the syllables of each of the words of each sentences is computed manually in terms of frame no, which has been considered as a ground truth speech signal . Then the syllables of each word of each sentences is computed using the proposed method. Sample boundary of each syllables corresponding to the sentence 1 of fig. 2. is shown in TABLE I.

TABLE I
Segmented Index for each syllable within sentence 1

	हम	बीम	आर	छी
Expected Syllable Index	2-17	29-37	38-47	60-73
Index based on group Algorithm	2-17	29-37	38-45	60-73

Performance has been measured by taking the differences between the ground truth boundary of the signals (Expected index) and system detected boundary (Index based on algorithm). Formula for measuring the performance is given by equation (3) :

$$P = [\{100 - \sum_{i=1}^{n}\{G_i(s) - A_i(s)\} + \{G_i(e) - A_i(e)\}/d\}/n * 100] \quad (3)$$

where d = $\{G_i(e) - G_i(s)\}$; P = accuracy percentage; i = starting count of syllables; n = number of syllables; $G_i(s)$ = starting index of ground truth syllable i; $A_i(s)$ = starting index of system detected boundary of syllable i; $G_i(e)$ = ending index of ground truth syllable i; $A_i(e)$ = ending index of system detected boundary of syllable i;
P (sentence 1) = {100 - (0/16+0/9+2/8+0/14)/.04} = 93.80 %
Although calculation are applied over 25 number of sentences but for simplicity, tabulation is done for for ten Maithili sentences only as shown in TABLE II.

TABLE II
Accuracy percentage for ten different sentences by five different speakers in Maithili language

Sentence	S1	S2	S 3	S4	S5	Avg
Sentence 1	87.67	90.00	80.91	92.31	93.80	**88.93**
Sentence 2	81.82	85.71	88.24	83.33	88.89	**85.60**
Sentence 3	77.38	72.78	79.24	83.33	84.44	**79.44**
Sentence 4	92.31	91.67	92.86	93.75	86.67	**91.45**
Sentence 5	87.50	88.57	88.07	87.45	83.77	**87.07**
Sentence 6	87.50	87.55	87.60	83.33	87.82	**86.76**
Sentence 7	85.00	84.52	84.62	83.81	88.75	**85.34**
Sentence 8	84.23	84.15	80.91	74.13	80.81	**80.84**
Sentence 9	83.07	79.22	83.22	83.31	93.33	**84.43**
Sentence 10	82.35	86.27	90.19	84.62	88.07	**86.30**
Average	**84.88**	**85.04**	**85.58**	**84.94**	**87.63**	**85.62** **85.62**

Line graphs are plotted for accuracy percentage as shown in fig. 10.

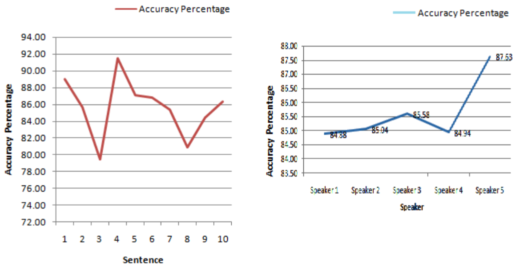


Fig. 10. Accuracy Rate

Conclusion

This paper proposed an intelligent segmentation algorithm for determining word and syllable boundary together for Maithili utterances. Performance evaluation is done through accuracy rate obtained as 85.62%. Proposed segmentation algorithm for syllable also shows the improvement over existing group delay algorithm. This algorithm will be helpful for Automatic voice to text conversion of Maithili utterances which can be further transformed for developing some Android Applications for the use of Maithili community. Other than Maithili, this speech recognition system can be further extended to include other regional languages say Bhojpuri as well. In future, feature extraction could be included for bringing more accuracy to the system.

Acknowledgment

First of all, I would like to give a thank to my faculty members and also to fellow members for their support in providing input data. I am also thankful to West Bengal University TEQIP II program and DST FIST program with reference number FIST/ETI/296/2011.

References

[1] https://en.wikipedia.org/wiki/Languages_of_India, 2016 August 13 [13:15 IST]

[2] Sharma, Anupriya, and Amanpreet Kaur. ”Automatic Segmentation of Punjabi Speech Signal using Group Delay.” Global Journal of Computer Science and Technology 13.12 (2013).

[3] Keerio, Ayaz, et al. ”On Preprocessing of Speech Signals.” World Academy of Science, Engineering and Technology 47 (2008): 317-323

[4] Kalamani, M., Dr S. Valarmathy, S. Anitha, and R. Mohan. ”Review of Speech Segmentation Algorithms for Speech Recognition.” International Journal of Advanced Research in Electronics and Communication Engineering 3, Vol. 11, IJAREC 2013.

[5] Jalil, Madiha, FaranAwais Butt, and Anuj Malik. ”Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals.” In Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE), 2013 International Conference on, pp. 208-212. IEEE, 2013.

[6] Waheed, Khurram, Kim Weaver, and Fathi M. Salam. ”A robust algorithm for detecting speech segments using an entropic contrast.” Circuits and Systems, 2002.MWSCAS-2002.The 2002 45th Midwest Symposium on.Vol. 3.IEEE, 2002

[7] Jia, Chuan, and Bo Xu. ”An improved entropy-based endpoint detection algorithm.” In International Symposium on Chinese Spoken Language Processing. 2002.

[8] Paul, Anup Kumar, Dipankar Das, and Md Mustafa Kamal. ”Bangla speech recognition system using lpc and ann.” In Advances in Pattern Recognition, 2009. ICAPR’09. Seventh International Conference on, pp. 171-174. IEEE, 2009.