Asian Journal of Information Technology 4 (12): 1127-1130, 2005

Automatic Segmentation of Bangla Speech: A New Approach

[1]Syed Akhter Hossain, [2]Md Lutfar Rahman and [3]Md Farruk Ahmed,
[1]Department of Computer Science and Engineering, East West University, 43 Mohakhali C/A Dhaka 1212,
[2,3]Department of Computer Science and Engineering, North South University, Dhaka, Bangladesh

Abstract: Automatic segmentation of Bangla isolated utterances is investigated in this study based on zero crossing and frame energy based segmentation of containing both vowel and consonant phonemes of both male and female speakers. This study proposed an algorithm for automatic segmentation of sections of speech samples based on whether they are silence, voiced or unvoiced speech. The segmentation is accomplished based on the processing of the speech samples and calculation of zero crossing and short-term energy functions. The thresholds are set based on the maximum likely hood for more accurate labeling the parts of speech. The study has shown good accuracy of lassification in the algorithms implemented over all speech samples. Such an effective speech segmentation would play a central role in voice activated application development.

Key words: Speech processing, zero crossing, energy, voiced, unvoiced, phoneme

INTRODUCTION

Speech is generated by the compression of the lung volume causing air flow which may be made audible if set into vibration by the activity of the larynx. This sound source can then be made into intelligible speech by various modifications of the supra-laryngeal vocal tract[1-4].

Computationally, speech production can be viewed as a filtering operation in which a sound source excites a vocal tract filter. The source is periodic, resulting in voiced speech or aperiodic, resulting in unvoiced speech as shown in (Fig.1)[5-7].

The voicing source occurs at the larynx at the base of the vocal tract, where airflow can be interrupted periodically by the vocal folds. The velum, tongue, jaw, teeth and lips are known as the *articulator*s. These provide the finer adjustments to generate speech. The excitation used to generate speech can be classified into *voiced, unvoiced, mixed, plosive, whisper* and *silence*. Any combination of one or more can be blended to produce a particular type of sound. A *phoneme* describes the linguistic meaning conveyed by a particular speech sound[8].

Vowels are associated with well-defined formant frequencies, which have provided the dominant approach to acoustic characterization of these vowels. According to Bangla Linguistics, there are eight classified cardinal vowels grouped into categories of frontal and back vowels and one central or neutral vowel /Av/. The frontal vowels are B, G, G¨v and back vowels are A,I,J and D respectively[9].

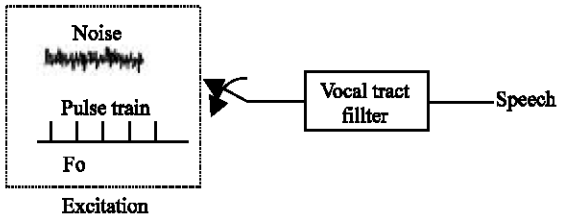

Fig. 1: Generation of voiced and unvoiced speech

Consonants differ from vowels in that they had more energy in the high frequency region compared to the low frequency region. Stop consonants can be divided in three classes viz. labials, alveolars and velars, each having a distinct release burst spectrum shape.

Bangla linguistics classifies consonants based on the manner of articulation. The different classes are Glottal or Laryngeal: n , Velar: K L M N O, Dorso Alveolar: P Q R S, Post Alveolar: k, Alveolar-Retroflex: U V W X p o, Alveolar: i j k l m h b, Dental: Z V ` a, Labial: c d e f g, Labio-Dental: d f [9].

Analysis technique: A classification of speech into voiced or unvoiced sounds provides a useful basis for subsequent processing, for example fundamental frequency estimation, formant extraction or syllable marking. A three-way classification into silence/unvoiced/voiced extends the possible range of further processing to tasks such as stop consonant identification and endpoint detection for isolated Bangla utterances[10,11].

Corresponding Author: Syed Akhter Hossain, Department of Computer Science and Engineering, East West University, 43 Mohakhali C/A Dhaka 1212, Bangladesh, Tel: +880 2 9882308 Fax: +880 2 8812336

The notion of zero-crossings is defined to be the number of times in a sound sample that the amplitude of the sound wave changes sign. For a 10ms sample of clean speech, the zero-crossing rate is approximately 12 for voiced speech and 50 for unvoiced speech based on the window length used for the segmentation. For clean speech the zero-crossing rate should also be useful for detecting regions of silence, as the zero-crossing rate should be zero for the silence[12,13].

Unfortunately, very few sound samples are recordings of perfectly clean speech. This means that often there is some level of background noise, that interferes with the speech which result in a silent regions having quite a high zero-crossings rate as the signal changes from just one side of zero amplitude to the other and back again. For this reason a tolerance threshold is included in the function that calculates zero crossings to try and alleviate this problem. The thresholds work by removing any zero-crossings, which do not both start and end a certain amount from the zero value. In this study we have used a threshold for zero crossing which is 10% of maximum zero crossing rate.

Short-term energy allows us to calculate the amount of energy in a sound at a specific instance in time and is defined in equation 1.

$$E_n = \sum_{m=n+N-1}^{n} \left(x(m)w(n-m)\right)^2 \quad (1)$$

Unfortunately, unlike zero-crossings there are no standard values of short-term energy for specific window sizes. Short-term energy is purely dependent upon the energy in the signal, which changes depending on how the sound was recorded. For example, if a person is recorded saying the same phrase twice, one while whispering and once while shouting, then the short-term energy values will be vastly different, although the zero crossing values should be roughly the same. It is required to inspect the recorded speech files to determine at what level to make the distinction between voiced and unvoiced speech. The short-term energy is higher for voiced than un-voiced speech and should also be zero or close to zero for silent regions in clean recording of clean speech. In a similar way to zero-crossings, the short-term energy is calculated using a 10ms non-overlapping rectangular window. The threshold of short-term energy in the study has been set to 30% of the maximum energy.

The proposed algorithm for the automatic labeling of the speech into voiced, unvoiced and silence segment is shown in the (Fig.2).

As shown in the (Fig. 2), the isolated Bangla utterance KZ containing velar phoneme /K/ is loaded into memory, down sampled to 10KHz, normalized and filtered to reduce

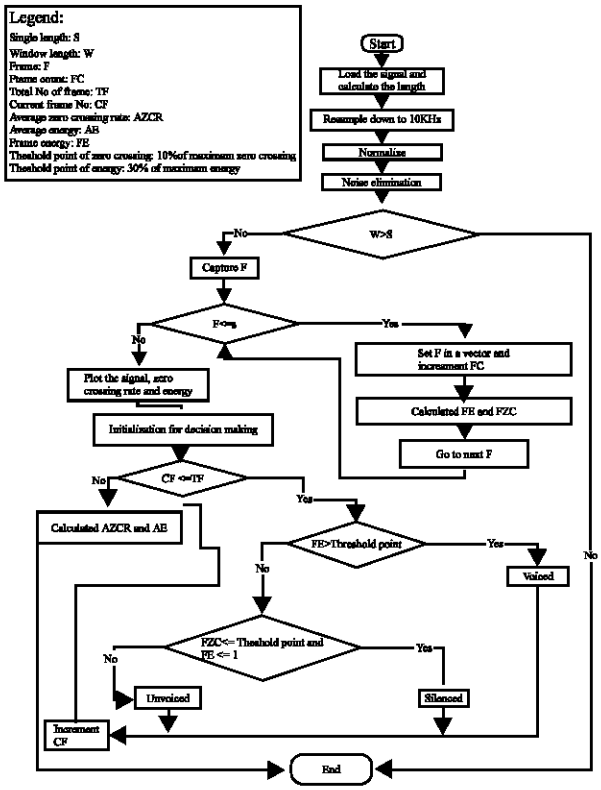

Fig. 2: Flow chart of the automatic labeling into voiced/unvoiced/silence

the line noise frequency and the frequency greater than 5KHz. The speech is then windowed into frames using rectangular window of 10msec. Each frame is then processed for the zero crossing rate and short time energy calculation. The threshold is set based on average zero crossing rate and average frame energy calculated during analysis of the speech. The threshold for the decision of the region is then applied on the analysis frame for the classification. The proposed algorithm is implemented and tested using Matlab.

RESULTS AND DISCUSSION

The Bangla speech samples KZ, KU, Lc, LU and other velar speech were analyzed according to the processing flow chart shown in the Figure 2 and automatic labeling of speech along with the frame processing decisions were recorded as shown in the Fig. 3 and 4 and in the Table 1 for the classification of the speech into Voiced / Unvoiced and Silence region.

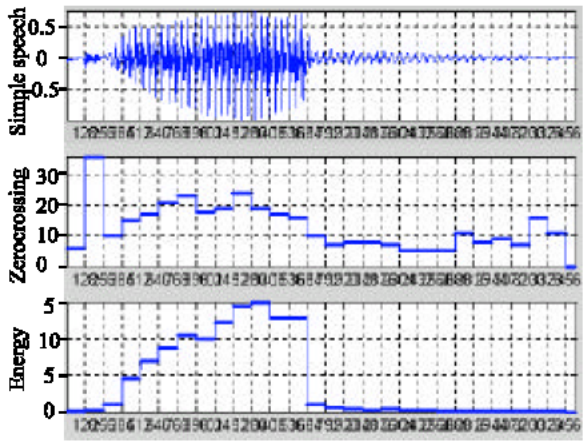

Fig. 3: The waveform, zero crossing and short-term energy for the word KZ spoken by the female speaker

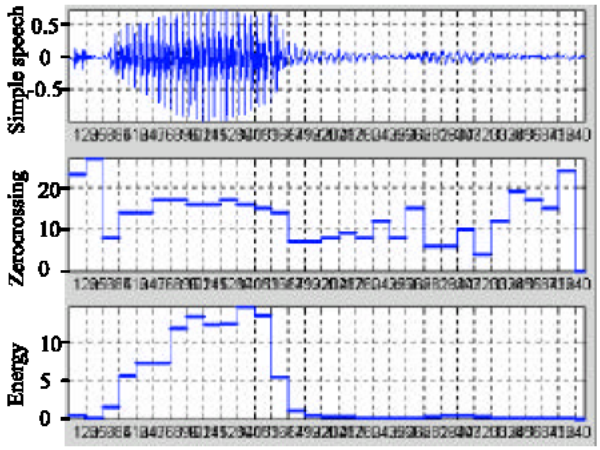

Fig. 4: The waveform, zero crossing and short-term energy for the word KZ spoken by the female speaker

Voiced speech can be distinguished from unvoiced speech as it has a much greater
amplitude displacement, when the speech is viewed as a waveform as shown in (Fig. 3 and 4).

To label the segmented speech frame, the zero-crossing function and short-term energy function were applied to the frame. These functions are complementary, since zero crossings are high when the speech is unvoiced, but low short-term energy at this point, the vice versa is true for voiced speech and both are approximately zero for silence. Cut off values are used to identify if a particular window of a frame is of the type 'Silence', 'Unvoiced' or 'Voiced'. The problem is that there are situations where the Zero Crossing and Average energy

Table 1: Labeling of word KZ

Segment	Zero Crossing	Avg. Energy	Frame Label
1	6	0.0614606	Un-Voiced
2	36	0.193423	Un-Voiced
3	10	0.998897	Un-Voiced
4	15	4.52236	Un-Voiced
5	17	7.03101	Voiced
6	21	8.85623	Voiced
7	23	10.5658	Voiced
8	18	10.1076	Voiced
9	19	12.3104	Voiced
10	24	14.6559	Voiced
11	19	15.1493	Voiced
12	17	13.0183	Voiced
13	16	12.9169	Voiced
14	10	0.992654	Un-Voiced
15	7	0.496244	Un-Voiced
16	8	0.360856	Un-Voiced
17	8	0.312882	Un-Voiced
18	7	0.331334	Un-Voiced
19	5	0.279137	Un-Voiced
20	5	0.264912	Un-Voiced
21	5	0.126168	Un-Voiced
22	11	0.108972	Un-Voiced
23	8	0.0894793	Silence
24	9	0.052611	Silence
25	7	0.0387826	Silence
26	16	0.00920721	Silence
27	11	0.025796	Silence

function differ in values partially caused by the background noise. This causes the cut off for silence to be raised, as it may not be quite zero due to noise being interpreted as speech by the functions whereas under clean speech both zero-crossings and short-term energy should be zero for silent regions. Based on the observations we decided that if the results of the functions didn't match then if the short-term energy implies voiced speech and the zero crossings implies silence, then the result should be voiced speech. This is because zero crossings have a low value for silence and voiced speech, therefore there is more chance of an error between these values, but the short-term energy is only ever high when voiced speech occurs as reflected in Table 1.

CONCLUSIONS

The parts of speech labeling produced using the algorithms outlined in this study are reasonably accurate of the order of 85% for well recorded, fairly clean speech but are not nearly as accurate for quiet recordings of speech.

The accuracy, of the algorithms outlined in this study, could be improved in two ways. Firstly more time could be spent on tweaking the cut-off values used by the algorithms to label the different parts of speech. The problem with this, however, is that if the values are fine tuned for one speech sample it is unlikely that they will be as accurate on other speech samples.

REFERENCES

1. Rabiner, L.R. and R.W. Schafer, 1978. Digital Processing of Speech Signals. Trentice-Hall Inc, Englewood Cliffs.
2. John, R. Deller, G. John, H.L. Proakis and H. John, 1993. Discrete-Time Processing of Speech Signals, Macmillan Publishing Company.
3. Gold, B. and N. Morgan, 2000. Speech and audio signal processing, Wiley.
4. Akhter H.S., M.L. Rahman and F. Ahmed, 2003. Vowel Space Identification of Bangla Speech, Dhaka Univ. J. Sci., 51: 31-38.
5. Hossain, S.A., M. Faruk Ahmed, M. Huq, A. Khan, M.A. Sobhan and M. Lutfar Rahman, 2002. Analysis by Synthesis of Bangla V owels, 5 th ICCIT Proceeding, pp: 272-276.
6. Akhter Hossain, S., M.A. Sobhan and M. Huq, A. Khan, 2001. Acoustic Vowel Space of Bangla Speech, ICCIT 2001 Proceeding, pp. 312-316
7. Akhter Hossain, S. and M. Abdus Sobhan, 1997. Fundamental Frequency Tracking of Bangla Voiced Speech –1 st NCCIS Proceeding, pp: 302-306
8. Fant, G., 1960. Acoustic Theory of Speech Production, S-Gravenhage, The Netherlands: Mounton and Co.
9. Abdul Hai, M., 2000. Dhvani Vijnan O Bangla Dhvani-Tattwa, Mullick Brothers.
10. Blumstein, S. and K. Stevens, 1980. Perceptual invariance and onset spectra for stop consonants in different vowel environments, J. Acoust. Soc. Am., 67: 648- 662.
11. Boll, S., 1979. Suppression of acoustic noise in speech using spectral subtraction, IEEE Trans. Acoust., Speech, Signal Process., 27: 113-120.
12. Berouti, M., R. Schwartz and J. Makhoul, 1979. Enhancement of speech corrupted by acoustic noise, Proc. IEEE Int. Conf. on Acoust., Speech, Signal Procs., pp: 208- 211.
13. Blumstein, S. and K. Stevens, 1979. Acoustic invariance in speech production, J. Acoust. Soc. Am., 66: 1001-1017.