Voiced and unvoiced separation in malay speech using zero crossing rate and energy

Rafizah Mohd Hanifa[1], Khalid Isa[2], Shamsul Mohamad[3], Shaharil Mohd Shah[4], Shelena Soosay Nathan[5], Rosni Ramle[6], Mazniha Berahim[7]
[1,5,6,7]Centre for Diploma Studies, Universiti Tun Hussein Onn Malaysia, Malaysia
[1,2,3,4]Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Malaysia

Article Info

Article history:

Received Jan 20, 2018
Revised Apr 27, 2019
Accepted Mei 17, 2019

Keywords:

Continuous Malay speech
Pre-processing signal
Short time energy
Speech analysis
Zero crossing rate

ABSTRACT

This paper contributes to the literature on voice-recognition in the context of non-English language. Specifically, it aims to validate the techniques used to present the basic characteristics of speech, viz. voiced and unvoiced, that need to be evaluated when analysing speech signals. Zero Crossing Rate (ZCR) and Short Time Energy (STE) are used in this paper to perform signal pre-processing of continuous Malay speech to separate the voiced and unvoiced parts. The study is based on non-real time data which was developed from a collection of audio speeches. The signal is assessed using ZCR and STE for comparison purposes. The results revealed that ZCR are low for voiced part and high for unvoiced part whereas the STE is high for voiced part and low for unvoiced part. Thus, these two techniques can be used effectively for separating voiced and unvoiced for continuous Malay speech.

Corresponding Author:

Rafizah Mohd Hanifa,
Centre for Diploma Studies,
Universiti Tun Hussein Onn Malaysia (UTHM),
Johor, Malaysia.
Email: ijeecs.iaes@gmail.com

1. INTRODUCTION

Speech technology has become popular as many applications today use speech as a medium to enhance our everyday life [1-2]. Speech recognition is different from voice recognition although it sounds similar [3]. Speech recognition is useful for people with a variety of disabilities such as those with physical disabilities who find typing difficult, painful or impossible and also those who have difficulties in recognizing words and in spelling such as those with dyslexia [4]. According to Abdullah et al. [5], the number of registered individuals with physical disabilities in Malaysia is 153,918 (33.1%) compared to speech disorder which is only 2,725 (0.59%). This shows that, there is still hope to help people with physical disabilities especially for those on smart wheelchair where they can use their own voice to ease their movements from one location to another. Most of the voice-recognition programs are in English [6] and there is a limited study conducted on other languages such as the Malay language.

Malay language is the national language of Malaysia and it is also one of the four official languages of Singapore. Besides these countries, Indonesia, Brunei and southern Thailand also used Malay language as a spoken language but with different dialects and accents [7]. Wu et al. [8] highlighted that the Malay language is a non-tonal language which means that it does not need lexical stress. A set of 37 phonemes are used as the phonemic representation in Malay language: six vowels, 27 consonants, three diphthongs and one for silence [9-10]. Vowels are divided into vowel backness and vowel height while diphthongs are grouped by vowel backness. On the other hand, consonants are grouped by manner and place of articulation. Interestingly, many consonants are pronounced nearly in the same way as in the English language. Syllabic or phonemic is the speech unit in many languages [11-13]. However, Malay language is an alphabetic language with salient syllabic structures [14]. From the phonological perspective, a syllable is made up of a consonant plus a vowel or a single vowel that follows the maximal onset and minimal coda [15]. Speech can be classified into voice and unvoiced [16-18]. Both classes have different characteristics in time and frequency domains which lead to different methods of processing [19-21].

This paper is organized in the following sequence. Section 2 presents a brief overview on the proposed algorithm that covers short time energy and zero crossings. Section 3 explains in depth the research methodology adopted in conducting the experiment. Section 4 discusses the results and finally, Section 5 draws the conclusion and avenues for further research.

2. TECHNIQUES FOR DETECTING VOICED AND UNVOICED SIGNALS

This work used the basic metrics calculation on speech such as energy and zero-crossings by considering the non-real time approach which implies that the signal is available for measures. In the following subsections, a brief explanation is given on the techniques used in this work which are the STE and ZCR.

2.1. Short Time Energy

The amplitude of unvoiced segments is generally much lower than the amplitude of voiced segments. The short-time energy of the speech signal provides a convenient representation that reflects these amplitude variations. The short-time energy can be calculated by using (1) [22-24]:

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

It is important to have a short duration window to be responsive to rapid amplitude changes. Unfortunately, a window that is too short will not provide a sufficient average to produce a smooth energy function. The effect of window on the time-dependent energy representation can be illustrated by the properties of two representative windows, i.e., the rectangular window and can be referred to in (2):

$$\mathrm{h[n]} = \begin{cases} 1, \text{for } 0 \le \mathrm{n} \le \mathrm{N}-1 \\ 0, \text{otherwise} \end{cases} \quad (2)$$

And the Hamming window which can be referred to in (3):

$$\mathrm{h[n]} = \mathrm{f(x)} = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi \mathrm{n}}{\mathrm{N}-1}\right), \text{for } 0 \le \mathrm{n} \le \mathrm{N}-1 \\ 0, \text{otherwise} \end{cases} \quad (3)$$

Where N is the window length in the samples.

The rectangular window applies equal weight to all the samples in the interval, whereas the Hamming window gives more weight to the center of the window. If the window size, N is small, i.e., in the order of a pitch period or less, E(n) will fluctuate very rapidly depending on the exact details of the waveform. If N is too large, i.e., in the order of several pitch periods, E (n) will change very slowly. Thus, it will not adequately reflect the changing properties of the speech signal. This implies that no single value of N is entirely satisfactory.

2.2. Zero Crossings

As for discrete-time signals, a zero-crossing occurs if successive samples have different algebraic signs. The rate at which zero crossings occur is a simple measure of frequency content of a signal and this is true for narrowband signals. For example, a sinusoidal signal of frequency f_0, sampled at a rate of F_s, has F_s/f_0 samples per cycle of the sine wave. Each cycle has two zero crossings and for that reason, the long-time average rate of zero-crossings can be shown in (4):

$$\mathrm{Z} = \frac{2\mathrm{f}_0}{\mathrm{F}_s} \text{ Crossings per sample} \quad (4)$$

Thus, the average zero-crossing rate provides a reasonable way in estimating the frequency of a sine wave. The computation required is defined in (5) to (7):

$$Z_n = \sum_{m=-\infty}^{\infty} \left|\mathrm{sgn}\big(x[m] - \mathrm{sgn}(x[m-1])\big)\right| w[n-m] \tag{5}$$

Where

$$\mathrm{sgn}(x[n]) = \begin{cases} 1, \text{for } x[n] \geq 0 \\ -1, \text{otherwise} \end{cases} \tag{6}$$

And,

$$w[n] = \begin{cases} \frac{1}{2N}, \text{for } 0 \leq n \leq N-1 \\ 0, \text{otherwise} \end{cases} \tag{7}$$

The model for speech production suggests that the energy of voiced speech is concentrated below 3 to 4 kHz, while for unvoiced speech, the energy is found at higher frequencies. There is a strong correlation between zero-crossing rate and energy distribution with frequency. If the zero-crossing rate is high, the speech signal is unvoiced and if the zero-crossing rate is low, the speech signal is voiced.

3. RESEARCH METHODOLOGY

MATLAB 2014a is used in this work. MATLAB is chosen as it offers many advantages. It contains a variety of signal processing and statistical tools, which help users in generating a variety of signals and plotting them. MATLAB excels at numerical computations, especially when dealing with vectors or matrices of data [25-26].

In this work, four respondents read different texts in Malay language acquired from local news websites using the speech corpus developed from a collection of audio speeches collected by Tan et. al [9]. The creation of the corpora is necessary especially for low-resourced languages [27]. The methodology used in this work is shown in Figure 1.

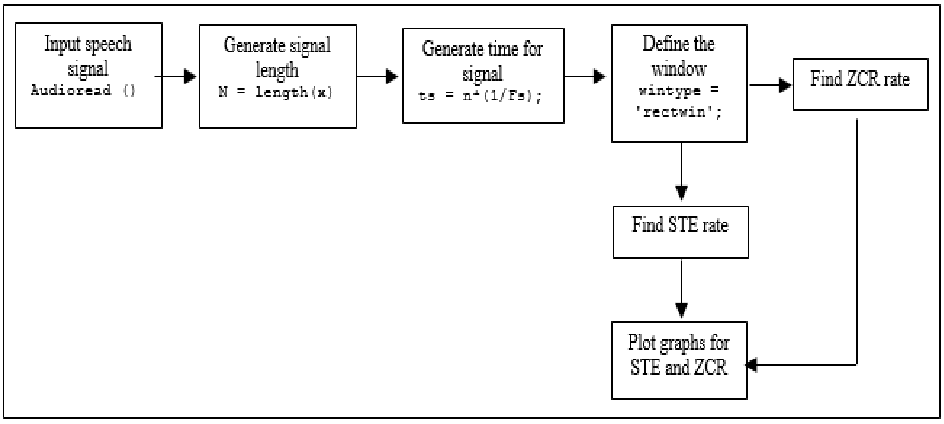


Figure 1. Block diagram for the reseaerch

The selected audio is read using the audioread () function. Then, the signal length is identified and the time taken for signal is generated by using the formula, ts = n * (1/Fs). In this work, the rectangular window is chosen because it applies equal weight to all the samples within the interval. The rate for ZCR and STE are then calculated. Lastly, the graphs for both ZCR and STE are plotted.

4. RESULTS AND DISCUSSION

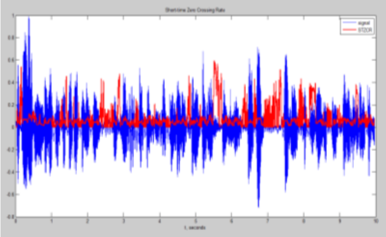
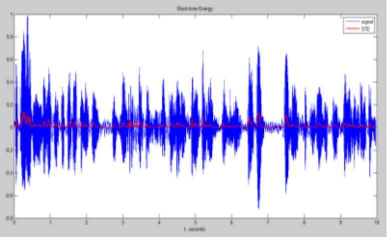
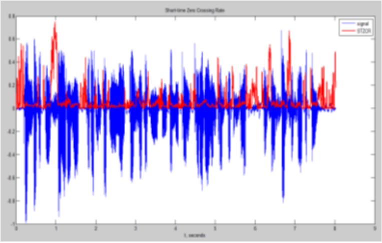
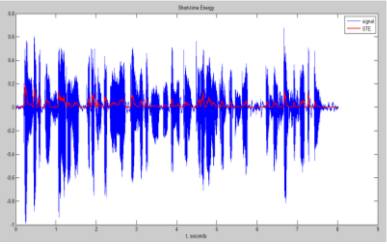
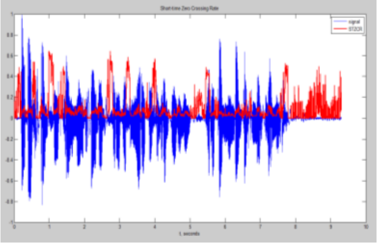
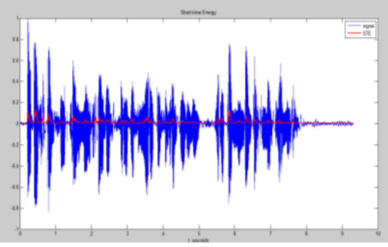
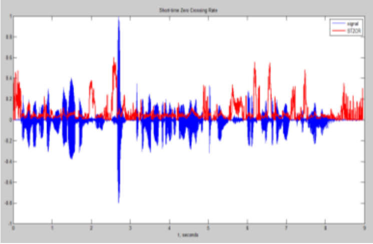
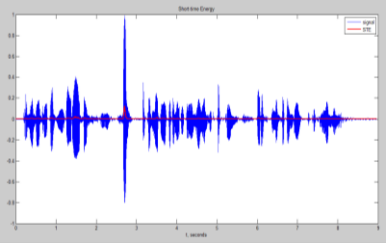
The demographic profiles of four respondents involved in this work are shown in Table 1. The reason for choosing different ethnic groups is to provide evidence that the two proposed techniques could detect the voiced and unvoiced in continuous Malay speech.

Table 1. Demographic Profiles

Gender	Ethnic	Speech Uttered
Female	Chinese	*Perkara ini berlaku kepada pelakon Eizlan Yusuf yang berlakon dalam filem Panas yang menampilkan terma seksualiti dalam masyarakat*
Female	Indian	*Pun begitu, Sarayala tidak sempat menunaikan hajatnya untuk makan kari ayam campur ubat, pisang dan sambal petai*
Male	Malay	*Pada Sukan SEA Chiengmai di Thailand 1995, snoker menyumbang enam pingat emas*
Male	Others	*Kami diberitahu Erra dan Yusri sah akan hadir pada hari tayangan yang ditetapkan pada 10 September pukul 9 malam*

Table 2 shows the results of uttered speech from four respondents' as stated in Table 1. The results reveal that for ZCR, the voiced part has lower energy as compared to STE which has higher energy for the voiced part and vice versa for the unvoiced part.

Table 2. Results Of ZCR and STE For Continuous Speech Uttered by Different Ethnic Groups

Gender	Ethnic	ZCR	STE
Female	Chinese		
Female	Indian		
Male	Malay		
Male	Others		

One word from the continuous speech uttered by a Chinese female has been selected as shown by the black dotted box in Figure 2 for ZCR and Figure 3 for STE. Figure 4 shows the extracted segment for the word "perkara" using PRAAT into its syllables which are "per", "ka" and "ra". The highlighted yellow color is the unvoiced signal. As can be seen from the figure, the yellow part is high for ZCR and low for STE.

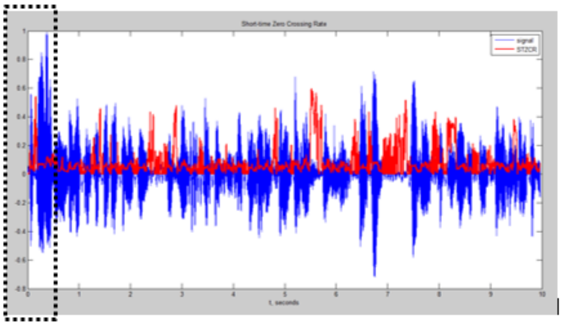


Figure 2. The dotted box contains the word “perkara” from ZCR calculation

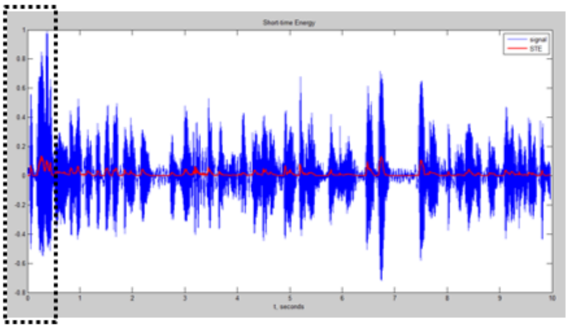


Figure 3. The dotted box contains the word “perkara” from STE calculation

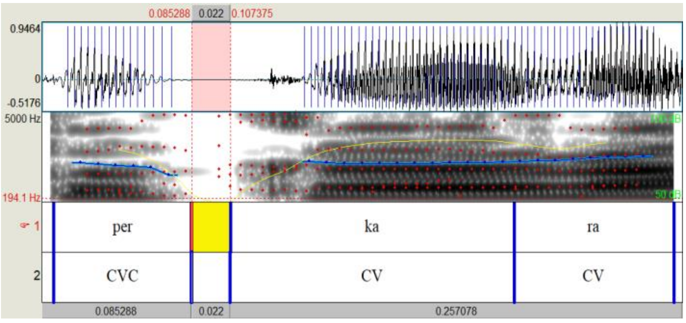


Figure 4. PRAAT is used to extract the word “perkara”

5. CONCLUSION

The two proposed techniques serve as a reasonable tool for the purpose of analysing Malay speech audio signals to determine the voiced and unvoiced signal. This particular stage is important to remove any noise before feeding the required speech signal to feature extraction. Future work may be considered in terms of studying the variations in the mean pitch and intensity of the Malay uttered speech by each ethnic group due to differences in their mother tongue.

ACKNOWLEDGEMENT

The authors would like to thank Universiti Tun Hussein Onn Malaysia (UTHM) for funding this research under the Tier-1 Research Grant (U926).

REFERENCES

[1] Preeti Saini and Parneet Kaur, “Automatic Speech Recognition: A Review,” *International Journal of Engineering Trends and Technology*, vol. 4, no. 2, 2013.

[2] Palsania Muzaffar and V.C. Kotak, “A Novel Approach for Smart Home Using Microsoft Speech Recognition,” *International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)*, vol. 3, no. 4, 2014.

[3] Thompson L., “Key differences between speech recognition and voice recognition,” 2017, [Online], Available: http://www.streetdirectory.com/travel_guide/139545/technology/keydifferences_between_speech_recognition_and_voice_recognition.html.

[4] R.M. Hanifa, K. Isa, S. Mohamad, “Malay speech recognition for different ethnic speakers: An exploratory study,” *2017 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, 2017

[5] W. Abdullah, and W. Arnidawati, “Supported employment: persons with learning difficulties in Doctoral Malaysia,” Ph.D. dissertation, University of Warwick, 2013.

[6] M. Farchi, K. Tahiry, S. Mounir, B. Mounir, A. Mouhsen, “Energy distribution in formant bands for arabic vowels”, *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 2, pp. 1163-1167, Apr 2019.

[7] A.E-L. Yousif, and Z.M. Don, “Text-to-Speech Conversion of Standard Malay,” *International Journal of Speech Technology*, Kluwer Academic Publishers, vol. 3, pp. 129-146, 2000.

[8] Z-Z. Wu, E.S. Chng, and H. Li, "Development of HMM-based Malay Text-to-Speech System," *Proceedings of the Second APSIPA Annual Summit and Conference*, Biopolis, Singapore, 2010, pp. 494-497.

[9] T.P. Tan, B. Ranaivo-Malançon, "Malay Grapheme to Phoneme Tool for Automatic Speech Recognition," *Third International Workshop on Malay and Indonesian Language Engineering*, Singapore, 2009.

[10] Hay Mar Htun, Theingi Zin, Hla Myo Tun, "Text To Speech Conversion Using Different Speech Synthesis," *International Journal of Scientific & Technology Research,* vol. 4, no. 7, 2015.

[11] Ruvan Weerasinghe, Asanka Wasala and Kumudu Gamage, "A Rule Based Syllabification Algorithm for Sinhala," *Lecture Notes in Computer Science, Natural Language Processing (IJCNLP 2005),* vol. 3651, pp. 438-449, 2005.

[12] Maimaitimin Saimaiti, Zhiwei Feng, "A Syllabification Algorithm and Syllable Statistics of Written Uyghur," [Online], Available: http://corpus.bham.ac.uk/corplingproceedings07.

[13] N.H. Samsudin and T.E. Kong. "A Simple Malay Speech Synthesizer Using Syllable Concatenation Approach," In *proceeding of MMU International Symposium on Information and Communications Technologies 2004 (M2USIC 2004)*, 2004.

[14] L.W. Lee, H.M. Low, and A.R. Mohamed, "A Comparative Analysis of Word Structures in Malay and English Children's Stories," *Pertanika Journal of Social Science and Humanities,* vol. 21, no. 1, pp. 67-84, 2013.

[15] H. Musa, A.R. Kadir, A. Azman, and M.T. Abdullah, "Syllabification Algorithm based on Syllable Rules Matching for Malay Language," *Applied Computer and Applied Computational Science*, 2011.

[16] M. Wolfel and J. McDonough, *Distant Speech Recognition*, John Wiley & Sons, Ltd, Publication, 2011.

[17] Gamit, M. R., Dhameliya, K., and Bhatt, N. S., "Classification Techniques for Speech Recognition: A Review," *International Journal of Emerging Technology and Advanced Engineering*, vol. 5, no. 2, 2015. Available: http://www.ijetae.com.

[18] S.K. Gaikwad, B.W. Gawali and P. Yannawar, "A Review on Speech Recognition Technique," *International Journal of Computer Applications*, vol. 10, no. 3, 2010.

[19] Jaber Marvan, "Voice Activity Detection Method and Apparatus for voiced/unvoiced decision and Pitch Estimation in a Noisy speech feature extraction," 08/23/2007, United States Patent 20070198251.

[20] D.S.Shete, S.B. Patil, S.B. Patil, "Zero crossing rate and Energy of the Speech Signal of Devanagari Script," *IOSR Journal of VLSI and Signal Processing (IOSR-JVSP)*, vol. 4, no. 1, ver. I, Jan. 2014.

[21] R. Bachu, S. Kopparthi, B. Adapa, and B. Barkana, "Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy," *Advanced Techniques in Computing Sciences and Software Engineering, Springer*, Dordrecht, 2010.

[22] Rabiner L. R., and Schafer R. W., *Digital Processing of Speech Signals,* Pearson; US edition, Englewood Cliffs, New Jersey, Prentice Hall, Sep 1978.

[23] Ranganadh Narayanam, "Voiced and Unvoiced Separation in Speech Auditory Brainstem Responses Of Human Subjects Using Zero Crossing Rate (ZCR) And Energy Of The Speech Signal," *International Journal Of Engineering Sciences & Research Technology*, 2017.

[24] Anu Priya Sharma, "Implementation of ZCR and STE techniques for the detection of the voiced and unvoiced signals in Continuous Punjabi Speech," *International Journal of Emerging Trends in Science and Technology*, 2016.

[25] Rafizah Mohd Hanifa, Khalid Isa and Shamsul Mohamad, "Silence Removal from Isolated Malay Words Using Framing and Windowing Method," *AIP Conference Proceedings 2016*, 2018.

[26] "Matlab: Creating Graphical User Interfaces," 2015.

[27] Aye Nyein Mon, Win Pa Pa, Ye Kyaw Thu, "UCSY-SC1: A Myambar Speech Corpus for Automatic Speech Recognition," *International Journal of Electrical and Computer Engineering (IJECE),* vol. 20, no. 4, pp. 937-949, 2017.