Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal

Bachu R.G., Kopparthi S., Adapa B., Barkana B.D.

Electrical Engineering Department

School of Engineering, University of Bridgeport

Abstract

In speech analysis, the voiced-unvoiced decision is usually performed in extracting the information from the speech signals. In this paper, we performed two methods to separate the voiced- unvoiced parts of speech from a speech signal. These are zero crossing rate (ZCR) and energy. In here, we evaluated the results by dividing the speech sample into some segments and used the zero crossing rate and energy calculations to separate the voiced and unvoiced parts of speech. The results suggest that zero crossing rates are low for voiced part and high for unvoiced part where as the energy is high for voiced part and low for unvoiced part. Therefore, these methods are proved more effective in separation of voiced and unvoiced speech.

1. Introduction

Speech can be divided into numerous voiced and unvoiced regions. The classification of speech signal into voiced, unvoiced provides a preliminary acoustic segmentation for speech processing applications, such as speech synthesis, speech enhancement, and speech recognition.

"Voiced speech consists of more or less constant frequency tones of some duration, made when vowels are spoken. It is produced when periodic pulses of air generated by the vibrating glottis resonate through the vocal tract, at frequencies dependent on the vocal tract shape. About two-thirds of speech is voiced and this type of speech is also what is most important for intelligibility. Unvoiced speech is non-periodic, random-like sounds, caused by air passing through a narrow constriction of the vocal tract as when consonants are spoken. Voiced speech, because of its periodic nature, can be identified, and extracted [1]".

In recent years considerable efforts has been spent by researchers in solving the problem of classifying speech into voiced/unvoiced parts [2-8]. A pattern recognition approach and statistical and non statistical techniques has been applied for deciding whether the given segment of a speech signal should be classified as voiced speech or unvoiced speech [2,3,5, and 7]. Qi and Hunt classified voiced and unvoiced speech using non-parametric methods based on multi-layer feed forward network [4]. Acoustical features and pattern recognition techniques were used to separate the speech segments into voiced/unvoiced [8]. The method we used in this work is a simple and fast approach and may overcome the problem of classifying the speech into voiced/unvoiced using zero-crossing rate and energy of a speech signal. The methods that are used in this study are presented in the second part. The results are given in the third part.

2. Method

In our design, we combined zero crossings rate and energy calculation. Zero-crossing rate is an important parameter for voiced/unvoiced classification. It is also often used as a part of the front-end processing in

automatic speech recognition system. The zero crossing count is an indicator of the frequency at which the energy is concentrated in the signal spectrum. Voiced speech is produced because of excitation of vocal tract by the periodic flow of air at the glottis and usually shows a low zero-crossing count [9], whereas the unvoiced speech is produced by the constriction of the vocal tract narrow enough to cause turbulent airflow which results in noise and shows high zero-crossing count.

Energy of a speech is another parameter for classifying the voiced/unvoiced parts. The voiced part of the speech has high energy because of its periodicity and the unvoiced part of speech has low energy. The analysis for classifying the voiced/unvoiced parts of speech has been illustrated in the block diagram in Fig.1

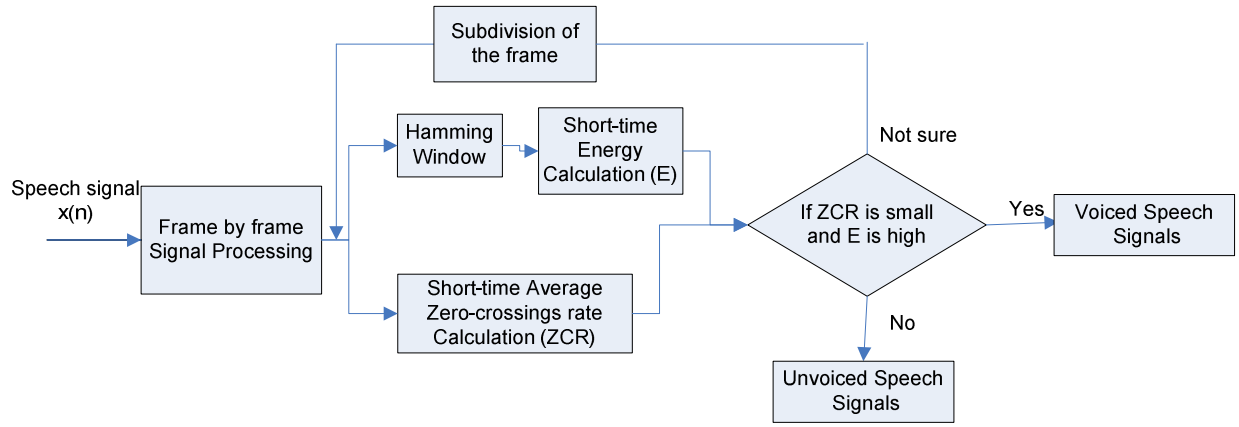

Fig.1: Block diagram of the voiced/unvoiced classification.

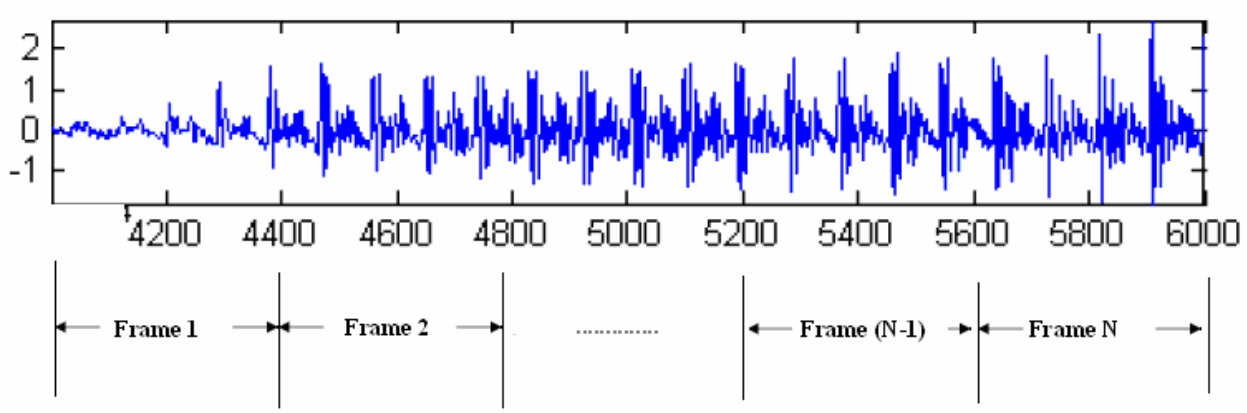

Fig. 2: Frame-by–frame processing of speech signal.

At the first stage, speech signal is divided into intervals in frame by frame without overlapping. It is given with Fig.2.

2.1. Zero-Crossings Rate

In the context of discrete-time signals, a zero crossing is said to occur if successive samples have different algebraic signs. The rate at which zero crossings occur is a simple measure of the frequency content of a signal. Zero-crossing rate is a measure of number of times in a given time interval/frame that the amplitude of the speech signals passes through a value of zero, Fig3 and Fig.4. Speech signals are broadband signals and interpretation of average zero-crossing rate is therefore much less precise.

However, rough estimates of spectral properties can be obtained using a representation based on the short-time average zero-crossing rate [11].

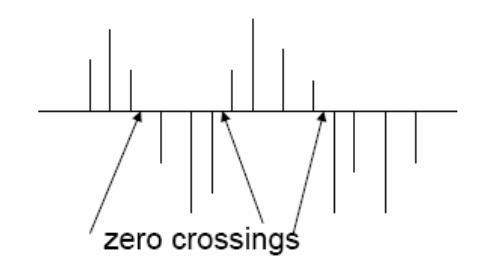

Fig. 3: Definition of zero-crossings rate

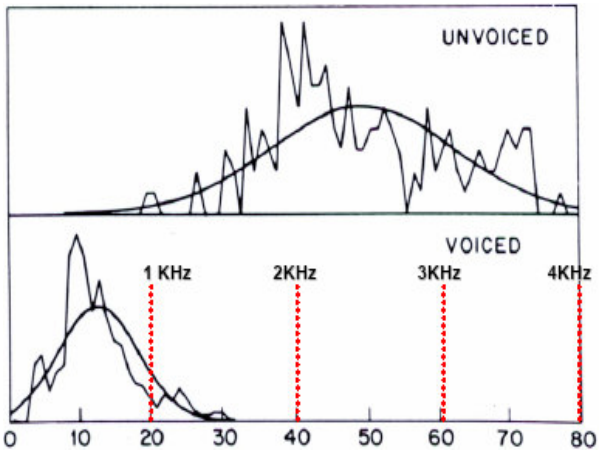

Fig. 4: Distribution of zero-crossings for unvoiced and voiced speech [11].

A definition for zero-crossings rate is:

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

where

$$\text{sgn}[x(n)] = \begin{cases} 1, x(n) \geq 0 \\ -1, x(n) < 0 \end{cases} \tag{2}$$

and

$$w(n) = \begin{cases} \dfrac{1}{2N} \, for, 0 \leq n \leq N-1 \\ 0 \, for, otherwise \end{cases} \tag{3}$$

The model for speech production suggests that the energy of voiced speech is concentrated below about 3 kHz because of the spectrum fall of introduced by the glottal wave, whereas for unvoiced speech, most of the energy is found at higher frequencies. Since high frequencies imply high zero crossing rates, and low frequencies imply low zero-crossing rates, there is a strong correlation between zero-crossing rate and energy distribution with frequency. A reasonable generalization is that if the zero-crossing rate is high, the speech signal is unvoiced, while if the zero-crossing rate is low, the speech signal is voiced [11].

2.2. Short-Time Energy

The amplitude of the speech signal varies with time. Generally, the amplitude of unvoiced speech segments is much lower than the amplitude of voiced segments. The energy of the speech signal provides a representation that reflects these amplitude variations. Short-time energy can define as:

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 \quad (4)$$

The choice of the window determines the nature of the short-time energy representation. In our model, we used Hamming window. The hamming window gives much greater attenuation outside the bandpass than the comparable rectangular window.

$$h(n) = 0.54 - 0.46\cos(2\pi n/(N-1)),\ 0 \le n \le N-1 \quad (5)$$

$$h(n) = 0,\ otherwise$$

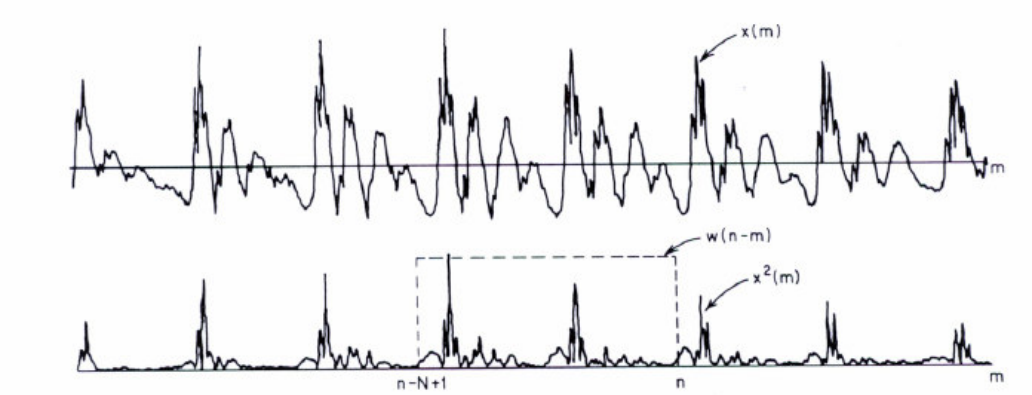

Fig. 5: Computation of Short-Time Energy [11].

The attenuation of this window is independent of the window duration. Increasing the length, N, decreases the bandwidth, Fig 5. If N is too small, E_n will fluctuate very rapidly depending on the exact details of the waveform. If N is too large, E_n will change very slowly and thus will not adequately reflect the changing properties of the speech signal [11].

3. Results

MATLAB 7.0 is used for our calculations. We chose MATLAB as our programming environment 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.

One of the speech signal used in this study is given with Fig.6. Proposed voiced/unvoiced classification algorithm uses short-time zero-crossings rate and energy of the speech signal. The results of voiced/unvoiced decision using our model are presented in Table 1.

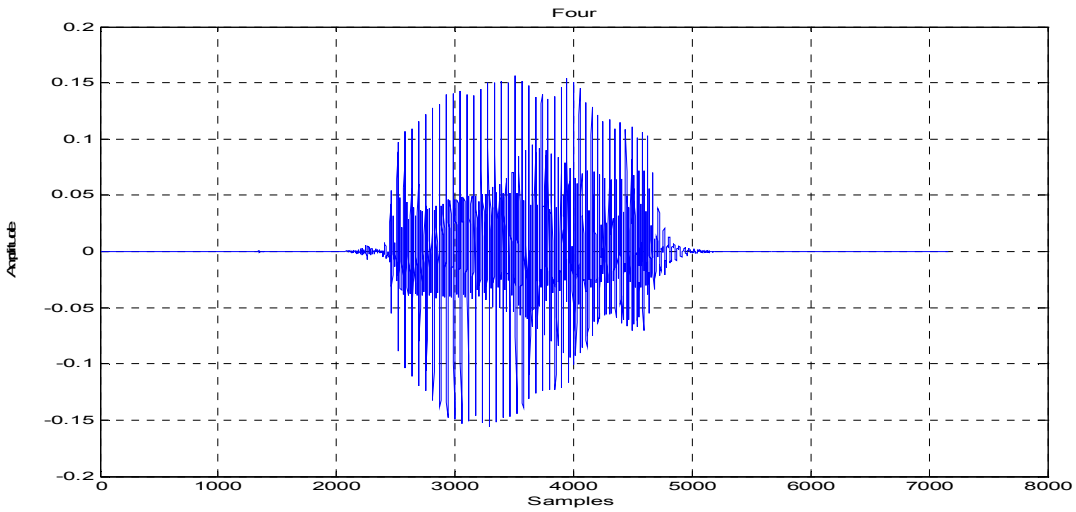

Fig.6: Original speech signal for word "four."

Table 1: Voiced/unvoiced decisions for the word "four" using the model.

Frames For word "four", Sampling frequency fs=8000Hz		**ZCR**	**Energy (J)**	**Decision**
Frame-1 (400 Samples)		152	0.0018	Unvoiced
Frame-2	Frame-2 1(200 Samples)	52	0.0543	Unvoiced
	Frame-22(200 Samples)	19	21.1189	Voiced
Frame-3 (400 Samples)		41	186.6628	Voiced
Frame-4 (400 Samples)		41	230.5772	Voiced
Frame-5 (400 Samples)		43	252.98	Voiced
Frame -6(400 Samples)		56	193.70	Voiced
Frame-7	Frame-71(200 Samples)	31	27.2842	Voiced
	Frame-72(200 Samples)	30	25.960	Voiced
Frame-8	Frame-811(100 Samples)	24	3.4214	Voiced
	Frame-812(100 Samples)	11	0.4765	Unvoiced
	Frame-82(200 Samples)	19	0.166	Unvoiced
Frame-9 (400 Samples)		89	0.0054	Unvoiced

In the frame-by-frame processing stage, the speech signal is segmented into a non-overlapping frame of samples. It is processed into frame by frame until the entire speech signal is covered. Table 1 includes the voiced/unvoiced decisions for word "four." It has 3600 samples with 8000Hz sampling rate. At the beginning, we set the frame size as 400 samples. At the end of the algorithm if the decision is not clear, energy and zero-crossing rate is recalculated by dividing the related frame size into two frames. This phenomenon can be seen for Frame 2, 7, and 8 in the Table 1.

4. Conclusion

We have presented an approach for separating the voiced /unvoiced part of speech in a simple and efficient way. The algorithm shows good results in classifying the speech as we segmented speech into many frames. In our future study, we plan to improve our results for voiced/unvoiced discrimination in noise.

References:

[1] Jong Kwan Lee, Chang D. Yoo, "Wavelet speech enhancement based on voiced/unvoiced decision", Korea Advanced Institute of Science and Technology The 32nd International Congress and Exposition on Noise Control Engineering, Jeju International Convention Center, Seogwipo, Korea,August 25-28, 2003.

[2] B. Atal, and L. Rabiner, "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition," *IEEE Trans. On ASSP*, vol. ASSP-24, pp. 201-212, 1976.

[3] S. Ahmadi, and A.S. Spanias, "Cepstrum-Based Pitch Detection using a New Statistical V/UV Classification Algorithm," *IEEE Trans. Speech Audio Processing*, vol. 7 No. 3, pp. 333-338, 1999.

[4] Y. Qi, and B.R. Hunt, "Voiced-Unvoiced-Silence Classifications of Speech using Hybrid Features and a Network Classifier," *IEEE Trans. Speech Audio Processing*, vol. 1 No. 2, pp. 250-255, 1993.

[5] L. Siegel, "A Procedure for using Pattern Classification Techniques to obtain a Voiced/Unvoiced Classifier", *IEEE Trans. on ASSP*, vol. ASSP-27, pp. 83- 88, 1979.

[6] T.L. Burrows, "Speech Processing with Linear and Neural Network Models", Ph.D. thesis, Cambridge University Engineering Department, U.K., 1996.

[7] D.G. Childers, M. Hahn, and J.N. Larar, "Silent and Voiced/Unvoiced/Mixed Excitation (Four-Way) Classification of Speech," *IEEE Trans. on ASSP*, vol. 37 No. 11, pp. 1771-1774, 1989.

[8] Jashmin K. Shah, Ananth N. Iyer, Brett Y. Smolenski, and Robert E. Yantorno "Robust voiced/unvoiced classification using novel features and Gaussian Mixture model", Speech Processing Lab., ECE Dept., Temple University, 1947 N 12th St., Philadelphia, PA 19122-6077, USA.

[9] 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.

[10] Thomas F. Quatieri, Discrete-Time Speech Signal Processing: Principles and Practice, MIT Lincoln Laboratory, Lexington, Massachusetts, Prentice Hall, ISBN-13:9780132429429.

[11] Rabiner, L. R., and Schafer, R. W., Digital Processing of Speech Signals, Englewood Cliffs, New Jersey, Prentice Hall, 512-ISBN-13:9780132136037, 1978.

Short Biographies of the Authors:

Rajesh G. Bachu is Graduate Assistant in Electrical Engineering at the University of Bridgeport, Bridgeport, CT. He is pursuing his Masters of Science, Electrical Engineering at the University of Bridgeport, CT.

Kopparthi S. is a graduate student in Electrical Engineering at the University of Bridgeport, Bridgeport, CT. He is pursuing his Masters of Science, Electrical Engineering at the University of Bridgeport, CT.

Adapa B is a graduate student in Electrical Engineering at the University of Bridgeport, Bridgeport, CT. He is pursuing his Masters of Science, Electrical Engineering at the University of Bridgeport, CT.

Buket D. Barkana is a Visiting Assistant Professor in the Department of Electrical Engineering at the University of Bridgeport, Bridgeport, CT. She earned her Ph.D. from Eskisehir Osmangazi University, Turkey in 2005. Her interests include Power Electronics Device Design, Speech Signal Processing.