Voiced/Unvoiced Detection using Short Term Processing

S. Nandhini
PG Scholar, Dept of ECE
National Engineering College
Kovilpatti, Tamilnadu, India.

A. Shenbagavalli, Ph.D.
Professor and Head of ECE
National Engineering College
Kovilpatti, Tamilnadu, India.

ABSTRACT
A new method for identifying voiced and unvoiced speech region is proposed. Voiced/unvoiced speech detection is needed to extract information from the speech signal and it is important in the area of speech analysis. Voiced and unvoiced speech region has been identified using Short Term Processing (STP) in this paper. Short Term Processing of speech has been performed by viewing the speech signal in frames, which has a size of 10-30ms. Short Term Processing has been performed in both time domain and frequency domain. Short Term Energy (STE), Short Term Zero Crossing Rate (STZCR) and Short Term Autocorrelation (STA) are computed from the time domain processing of speech. The spectral components in the speech signal are not apparent in the time domain. Hence, frequency domain representation is needed which is achieved using fourier transform. Conventional fourier representation is inadequate to provide information about the time varying nature of spectral components present in speech. So, short term version of fourier transform is needed, which is named as Short Term Fourier Transform (STFT).

Keywords
Voiced speech, Unvoiced speech, Short Term Energy, Short Term Zero Crossing Rate, Short Term Autocorrelation, Short Term Fourier Transform.

1. INTRODUCTION
Speech is an acoustic signal produced from the speech production system. In the human speech production system, air enters into the lungs during inhalation via articulators and speech organs. Articulators mainly include tongue, lips and jaw. The speech production organs include pharynx, glottis, vocal cord, larynx and trachea. This air is expelled out from the lungs through trachea and causes the vocal cord to vibrate. During this process, the air flow is chopped up into quasi-periodic pulses to generate the required excitation signal. Finally, the pulses are frequency shaped by the oral cavity and the nasal cavity to produce the desired speech signal.

The speech production system produces the output speech signal by responding to the input excitation signal. Depending upon the input excitation signal, the output speech signal from the speech production system is broadly divided into three classes which are discussed in section 2.

Identification of voiced/unvoiced speech mainly depends on the nature of the speech waveform, energy associated with the speech waveform [1], [2], [7], [8] number of zero crossings present in the speech waveform [1], [2], [5], [7], [8] and the correlation among the successive samples in the speech waveform [2], [5], [8].

Detection of voiced/unvoiced speech in the speech waveform is important in speech analysis system. In speech analysis, the voiced/unvoiced detection is usually performed in extracting information from the speech signals. But the voiced/unvoiced detection is critical, because it is essential to know whether the speech production system involves vibration of the vocal cords [2], [10]. The periodicity of the vocal tract vibration makes the voiced speech segment periodic and so distinguishable from the noise-like unvoiced speech segments [6].

This paper describes about the three classes of speech signal in section 2. Section 3 focuses on Short Term Processing of speech. Short Term Processing has been performed in both time domain and frequency domain. Parameters like Short Term Energy (STE), Short Term Zero Crossing Rate (STZCR) and Short Term Autocorrelation (STA) are computed from the time domain**.** Short Term Fourier Transform is computed from the frequency domain. The results are discussed in section 4.

2. THREE CLASSES OF SPEECH SIGNAL
The three classes are named as voiced speech, unvoiced speech and non-speech (silence). In the first class (voiced speech), the input excitation is nearly periodic in nature. In the second class (unvoiced speech), the input excitation is random noise-like nature. In the third class (non-speech), there is no excitation to the system.

2.1 Voiced Speech
When the input excitation to the speech production system is nearly periodic impulse sequence, then the output speech looks visually like a periodic signal and is termed as voiced speech.

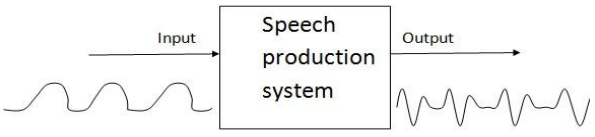

Fig 1: Block diagram representation of voiced speech production

During the production of voiced speech, the air breathe out of lungs through the trachea is interrupted periodically by the vibrating vocal cords. Due to this interruption, glottal wave is generated which excites the speech production system to produce voiced speech.

Voiced speech can be identified by the periodic nature of the speech waveform, relatively high energy associated with the speech waveform [7], [8], less number of zero crossings present in the speech waveform [3], [5], [7], [8] and more correlation among the successive samples in the speech waveform [8].

Due to the periodic nature of voiced speech, a fundamental frequency (pitch frequency) and its harmonics can be predicted in the spectrum of voiced speech. The presence of harmonic structure in the voiced spectrum is another distinguishing factor.

2.2 Unvoiced Speech

When the input excitation to the speech production system is random noise-like structure, then the output speech will also be random noise-like without any periodic nature and is termed as unvoiced Speech.

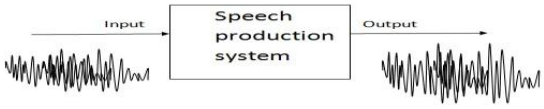

Fig 2: Block diagram representation of unvoiced speech production

During the production of unvoiced speech, the air breathe out of lungs through the trachea is not interrupted by the vibrating vocal cords. However, obstruction of air flow occurs scarcely from glottis somewhere along the length of vocal tract to produce total or partial closure. Due to this modification of air flow, stop or frication excitation occurs. This excites the vocal tract system to produce unvoiced speech.

Unvoiced speech can be identified by the non-periodic and noise-like nature of the speech waveform, relatively low energy associated with the speech waveform when compared to voiced speech [7], [8] more number of zero crossings present in the speech waveform when compared to voiced speech [3], [5], [7], [8] and relatively less correlation among successive samples in the speech waveform [8].

Due to the noise-like nature of unvoiced speech, there is no fundamental frequency (pitch frequency) and its harmonics in the spectrum of unvoiced speech. The absence of harmonic structure in the unvoiced spectrum is another distinguishing factor.

2.3 Non-Speech (Silence)

When there is no input excitation to the speech production system, then the output corresponds to no speech and is termed as non-speech. Non-speech (silence) is present between voiced and unvoiced speech. Hence, speech will not be intelligible without the presence of non-speech between voiced and unvoiced speech.

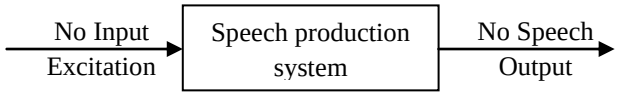

Fig 3: Block diagram representation of non-speech (silence) production

During the production of non-speech, there is no excitation supplied to the vocal tract and hence there is no speech output.

Non-speech is identified by the absence of any signal characteristics, lowest energy associated with the speech waveform when compared to unvoiced and voiced speech [8], relatively more number of zero crossings present in the speech waveform when compared to unvoiced speech [3], [8] and no correlation among successive samples in the speech waveform [8].

Due to the absence of speech signal, there is no output in the spectrum of non-speech. This is another distinguishing factor for non-speech.

3. SHORT TERM PROCESSING

Speech is produced from a time varying vocal tract system with time varying excitation. Due to this, the speech signal is non-stationary in nature. Speech signal is stationary when it is viewed in blocks of 10-30msec [8]. Short Term Processing divides the input speech signal into short analysis segments that are isolated and processed with fixed (non-time varying) properties. These short analysis segments called as analysis frames almost always overlap one another. Short Term Processing of speech is performed both in time domain and in frequency domain [4].

3.1 Time Domain Processing of Speech

Short Term Energy (STE), Short Term Zero Crossing Rate (STZCR) and Short Term Autocorrelation (STA) are computed from the time domain processing of speech.

3.1.1 Short Term Energy (STE)

Speech is time varying in nature. The energy associated with voiced speech is large when compared to unvoiced speech [7]. Silence speech will have least or negligible energy when compared to unvoiced speech [8]. Hence, Short Term Energy can be used for voiced, unvoiced and silence classification of speech.

Short Term Energy is derived from the following equation,

$$E_T = \sum_{m=-\infty}^{\infty} s^2(m) \quad (1)$$

where E_T is the total energy and $s(m)$ is the discrete time signal [4], [8].

For Short Term Energy computation, speech is considered in terms of short analysis frames whose size typically ranges from 10-30 msec. Consider the samples in a frame of speech as m=0 to m=N-1, where N is the length of frame. Hence, equation (1) is written as,

$$E_T = \sum_{m=-\infty}^{-1} s^2(m) + \sum_{m=0}^{N-1} s^2(m) + \sum_{m=N}^{\infty} s^2(m) \quad (2)$$

The speech sample is zero outside the frame length. Hence, equation (2) is written as,

$$E_T = \sum_{m=0}^{N-1} s^2(m) \quad (3)$$

The relation in equation (3) gives the total energy present in the frame of speech from m=0 to m=N-1.

Short Term Energy is defined as the sum of squares of the samples in a frame and it is given by,

$$e(n) = \sum_{m=-\infty}^{\infty} [s_n(m)]^2 \quad (4)$$

After framing and windowing, the n^{th} frame speech becomes $s_n(m) = s(m).w(n-m)$ and hence Short Term Energy in equation (4) is given by [4], [7], [8],

$$e(n) = \sum_{m=-\infty}^{\infty} [s(m).w(n-m)]^2 \quad (5)$$

where $w(n)$ represents the windowing function of finite duration and n represents the frame shift or rate in number of samples. This frame shift is as small as one sample or as large as frame size.

3.1.2 Short Term Zero Crossing Rate (STZCR)

Zero Crossing Rate is defined as the number of times the zero axes is crossed per frame. If the number of zero crossings is more in a given signal, then the signal is changing rapidly and accordingly the signal may contain high frequency information which is termed as unvoiced speech. On the other hand, if the number of zero crossing is less, then the signal is changing slowly and accordingly the signal may contain low frequency information which is termed as voiced speech [7], [8]. Thus, Zero Crossing Rate gives indirect information about the frequency content of the signal. Zero Crossing Rate is computed using typical frame size of 10-30msec with half the frame size as frame shift.

Short Term Zero Crossing Rate is defined as the weighted average of number of times the speech signal changes sign within the time window [4] and it is given by [1], [4], [5], [7],

$$Z(n) = \sum_{m=-\infty}^{\infty} \left| sgn[s(m)] - sgn[s(m-1)] \right| . w(n-m) \quad (6)$$

where $sgn[s(m)] = 1 \; if \; s(m) \geq 0$

$= -1 \; if \; s(m) < 0,$

and $w(n) = \frac{1}{2N} if \; 0 \leq n \leq N-1$

$= 0 \; otherwise.$

If successive samples $s(m)$ and $s(m-1)$ have different algebraic signs then,

$\left| sgn[s(m)] - sgn[s(m-1)] \right|$ in (6) becomes equal to 1 and hence Zero Crossing Rate is counted.

If successive samples $s(m)$ and $s(m-1)$ have same algebraic signs then,

$\left| sgn[s(m)] - sgn[s(m-1)] \right|$ in (6) becomes equal to 0 and hence Zero Crossing Rate is not counted.

3.1.3 Short Term Autocorrelation (STA)

Correlation can be used for finding the similarity among one or two sequences. Autocorrelation is achieved by providing different time lag for the sequence and computing with the given sequence as reference. The autocorrelation $r_{XX}(k)$ of a stationary sequence is given by,

$$r_{XX}(k) = \sum_{m=-\infty}^{\infty} X(m).X(m+k) \quad (7)$$

Due to the non-stationary nature of speech, a short term version of the autocorrelation is needed. The Short Term Autocorrelation of a non-stationary sequence $s(m)$ is defined as the deterministic autocorrelation function of the sequence $s_n(m) = s(m).w(n-m)$ that is selected by the window shifted to time n and it is given by [4], [8],

$$r_{ss}(n,k) = \sum_{m=-\infty}^{\infty} [s(m).w(n-m)].[s(k+m).w(n-k-m)] \quad (8)$$

where $s_n(m) = s(m).w(n-m)$ is the windowed version of $s(m)$.

The nature of Short Term Autocorrelation is primarily different for voiced and unvoiced speech. The autocorrelation of voiced speech will have periodic waveform representing the pitch period [6], [9]. The autocorrelation of unvoiced speech will have random noise-like structure. Therefore, information from the autocorrelation sequence is used for discriminating voiced and unvoiced speech.

3.2 Frequency Domain Processing of Speech

The short term time domain analysis is useful for computing the time domain features like energy, zero crossing rate and autocorrelation at the gross level. The different frequency or spectral components that are present in the speech signal are not directly apparent in the time domain. Hence, frequency domain representation is needed which is achieved using fourier transform. The conventional fourier representation is inadequate to provide information about the time varying nature of spectral information present in speech. Hence, short term version of fourier transform commonly known as Short Term Fourier Transform (STFT) is needed [4].

3.2.1 Discrete Time Fourier Transform (DTFT)

Discrete Time Fourier Transform (DTFT) is used to obtain the frequency domain representation [4]. If $X(w)$ is the Discrete Time Fourier Transform of $x(n)$, then the DTFT is given by,

$$X(w) = \sum_{n=-\infty}^{\infty} x(n) e^{-j\omega n} \quad (9)$$

Where $x(n)$ is a discrete time signal and it is given by,

$$x(n) = \frac{1}{2\Pi} \int_{-\Pi}^{\Pi} X(w) \, e^{j\omega n} \, d\omega \quad (10)$$

As $X(w)$ is continuous function of frequency, it cannot be computed. To make it as a discrete version of the DTFT termed as Discrete Fourier Transform (DFT), we have to uniformly sample $X(w)$ as,

$$X(w) = X(w_k) \quad (11)$$

Where $w_k = \frac{2\Pi}{N} k$, k = 0, 1... (N-1), where N is the number of samples of $X(w)$.

In speech, large magnitude frequency components dominate in the linear magnitude spectrum to give a false picture of small magnitude frequency components. To overcome this, the logarithmic version of the magnitude spectrum is taken. Although large and small magnitude frequency components are visible in log magnitude spectrum, it does not tell when a particular frequency component is present. Hence, the timing information is completely lost. Speech is a non-stationery signal where the frequency components change with time.

Therefore, a representation that will give time varying spectral information is needed for speech.

3.2.2 Short Term Fourier Transform (STFT)

To obtain the time varying spectral information, Short Term Fourier Transform approach is employed. DTFT is computed using DFT for a block size of 20ms and this process is repeated for all the blocks of speech signal. All the spectra computed are stacked together as a function of time and frequency to observe the time varying spectra. This leads to Short Term Fourier Transform which is given by [4], [8],

$$X(w,n) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-j\omega m} \qquad (12)$$

where $w(n)$ is the window function for Short Term Processing, $[x(m).w(n-m)]$ represents the windowed segment at the time instant n.

In STFT, the spectral amplitude and phase are function of both frequency and time whereas it is the only function of frequency in DTFT. STFT magnitude spectra is a three dimensional (3D) plot of spectral amplitude versus time and frequency. By fixing the time n at a particular value $n = n_0$ and observing STFT gives spectrum of that segment of speech. Alternatively, by fixing the frequency w at a particular value $w = w_0$ and then observing STFT gives time varying nature of that particular frequency component. The spectral amplitude associated with a frequency component varies as a function of time.

4. RESULTS AND DISCUSSIONS

A clean speech signal without any noise added to it is taken as input.

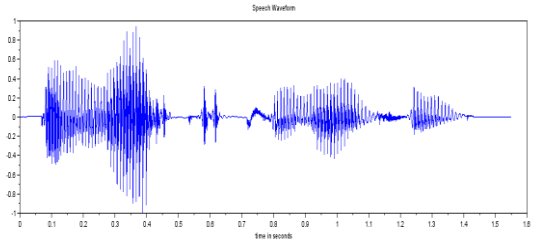

Fig 4: Input speech waveform

Fig. 4 shows the input speech waveform of a clean speech signal. Short Term Energy, Short Term Zero Crossing Rate, Short Term Autocorrelation are computed from time domain processing of speech. Short Term Fourier transform is computed from frequency domain processing of speech which overcomes Discrete Time Fourier Transform.

4.1 Short Term Energy Computation

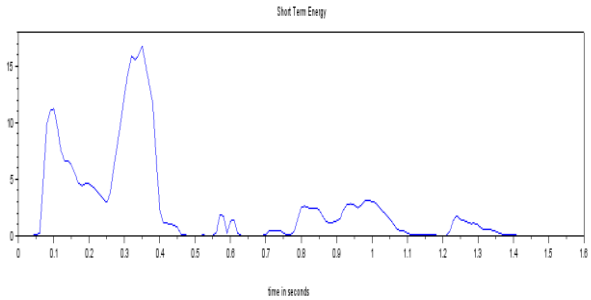

Fig 5: Short Term Energy contour for the speech signal with 20ms frame size

Fig. 5 depicts that waveform region with high amplitude and high energy is detected as voiced speech and the waveform region with low amplitude and low energy is detected as unvoiced speech.

4.2 Short Term Zero Crossing Rate Computation

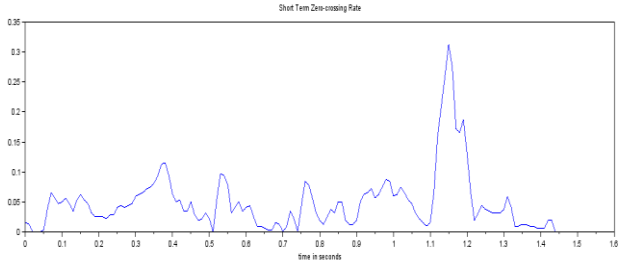

Fig 6: Short Term Zero Crossing Rate contour for the speech signal with 20ms frame size

Fig. 6 depicts that the waveform region with more number of zero crossings is detected as unvoiced speech and the waveform region with less number of zero crossings is detected as voiced speech.

4.3 Short Term Autocorrelation Computation

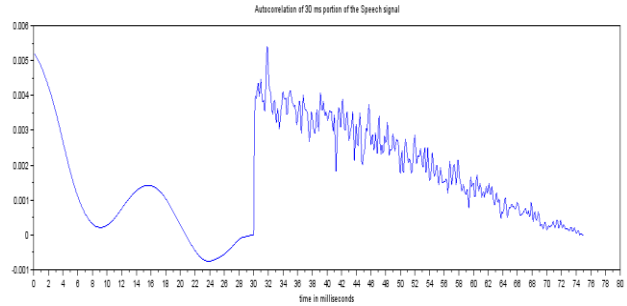

Fig 7: Short Term Autocorrelation for the speech signal with 30ms time duration

Fig. 7 depicts that the speech signal waveform within 30ms duration is periodic and hence it is detected as voiced speech where as the speech signal waveform after 30ms duration is random noise-like structure and hence it is detected as unvoiced speech.

4.4 Discrete Time Fourier Transform Computation

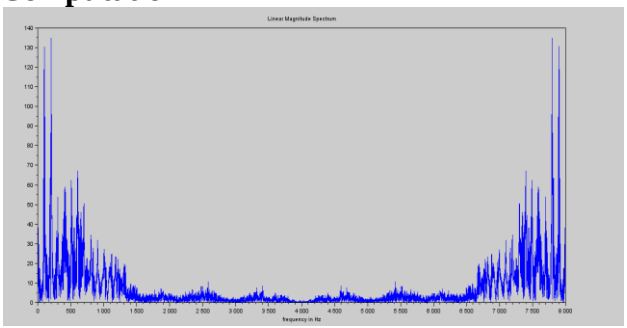

Fig 8: Linear magnitude DTFT spectrum of a speech signal

Fig. 8 shows that large magnitude frequency component dominates the small magnitude frequency component. Hence, linear magnitude DTFT spectrum cannot be used for spectral analysis.

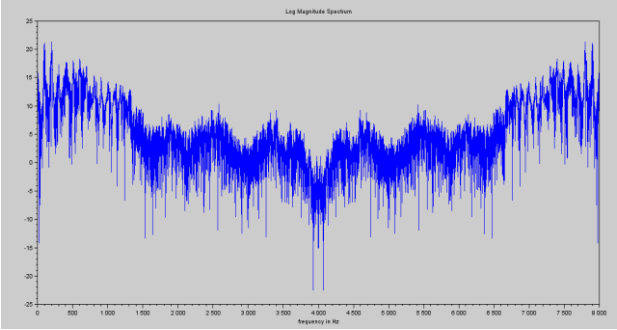

Fig 9: Log magnitude DTFT spectrum of a speech signal

Fig. 9 shows both large magnitude and small magnitude frequency component. However, the log magnitude DTFT spectrum does not tell when a particular frequency component is present. Hence, timing information is completely lost.

4.5 Short Term Fourier Transform Computation

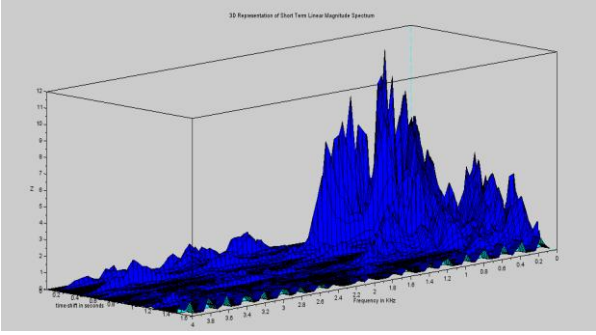

Fig 10: 3D plot of Short Term Fourier Transform

Fig. 10 shows time varying frequency components. Here, low frequency region with high amplitude and energy is detected as voiced speech and high frequency region with low amplitude and energy is detected as unvoiced speech.

5. CONCLUSION AND FUTURE WORK

Detection of voiced and unvoiced speech region is important in speech analysis. Voiced/unvoiced identification is usually performed to extract information from the speech signal. Short Term Processing (STP) was performed to detect the voiced and unvoiced regions of speech. Short Term Processing divides the input speech signal into short analysis frames of size 10-30ms with a frame shift of 10ms. Short Term processing was computed both in time domain and in frequency domain. Short Term Energy (STE), Short Term Zero Crossing Rate (STZCR), Short Term Autocorrelation (STA) was computed from the time domain. In the time domain, the spectral components of the speech signal are not apparent. Hence, frequency domain representation was needed and achieved using fourier transform. Hence, Short Term Fourier Transform (STFT) was computed to provide information about the time varying nature of spectral components present in the speech signal. This classification of speech into voiced and unvoiced regions can be used for applications like speaker identification and verification, speech coding, speech synthesis and speech enhancement in future.

6. REFERENCES

[1] A.E.Mahdi and E.Jafer, "Two-Feature Voiced/Unvoiced Classifier Using Wavelet Transform", The Open Electrical and Electronic Engineering Journal, No.2, pp.8-13, 2008.

[2] Bishnu S.Atal and Lawrence R.Rabiner, "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.24, No.3, 2003, pp. 201-212.

[3] D.G.Childers, M.Hahn and J.N.Larar "Silent and Voiced/Unvoiced/Mixed Excitation (Four-way) Classification of Speech", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.37, No.11, November 1989.

[4] Lawrence R.Rabiner and Ronald W.Schafer, "Introduction to Digital Speech Processing", Foundations and Trends in Signal Processing, Vol.1, No.33-53, 2007.

[5] Mojtaba Radmard, Mahdi Hadavi and Mohammad Mahdi Nayebi, "A New Method of Voiced/Unvoiced Classification Based on Clustering", Journal of Signal and Information Processing, 2011, Vol.2, pp.336-347.

[6] N.Dhananjaya and B.Yegnanarayana, "Voiced/ Non-voiced Detection Based on Robustness of Voiced Epochs", IEEE Signal Processing Letters, Vol.17, No.3, March 2010.

[7] R.G.Bachu, S.Kopparthi, B.Adapa and B.D.Barkana, "Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy", Advanced Techniques in Computing Sciences and Software Engineering, pp.279-282, 2010.

[8] Ronald W.Schafer and Lawrence R.Rabiner, "Digital Representations of Speech Signals", Proceedings of the IEEE, Vol.63, No.4, April 1975.

[9] S.Ahmadi and A.S.Spanias, "Cepstrum Based Pitch Detection Using a New Statistical V/UV Classification Algorithm", IEEE Transactions on Speech and audio Processing, Vol.7, No.3, pp.333-338, 2002.

[10] Yingyong Qi and Bobby R.Hunt, "Voiced-Unvoiced-Silence Classifications of Speech Using Hybrid Features and a Network Classifier", IEEE Transactions on Speech and Audio Processing, Vol.1, No.2, pp. 250-255, 2002.