Technical Implementation Using MATLAB

Audio Voice/Unvoice Segmentation and Pitch finding

Introduction:

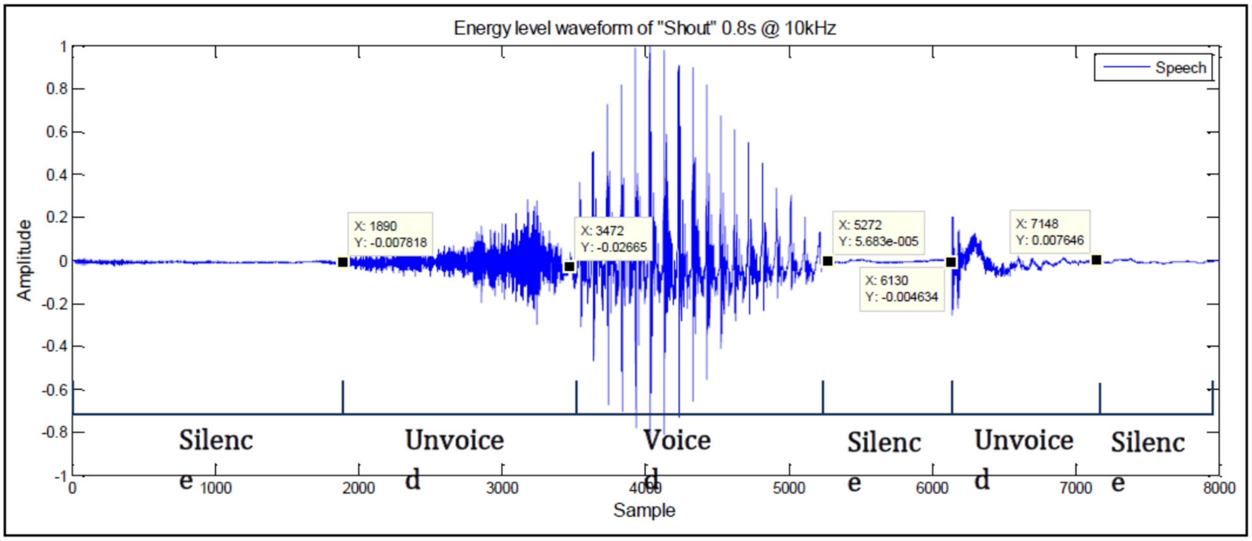
Speech can be divided into numerous voiced and unvoiced regions. The classification of speech signal into voiced, unvoiced provides a preliminary acoustic segmentation for speech processing applications, such as speech synthesis, speech enhancement, and speech recognition.

"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."

There are different methods which can classify voiced/unvoiced parts of speech but their accuracy is not 100% :

1. Zero Crossing Rate Method.
2. Average Energy.
3. Moments & Difference.
4. Voice Variation.
5. Change Rate.
6. Mel-frequency cepstral coefficients (MFCC).
7. Linear Predictive Coding(LPC).
8. LPC feature with Linear Discriminant Analysis(LDA) and Gaussian mixture model(GMM) Classifier**.**

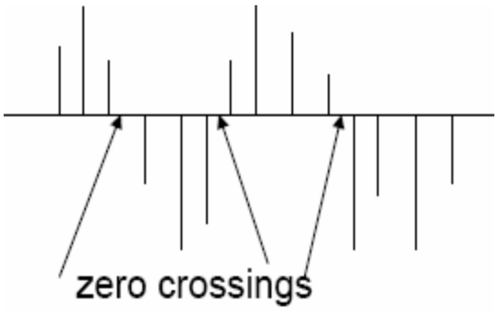
Energy plot of the word 'Shout'



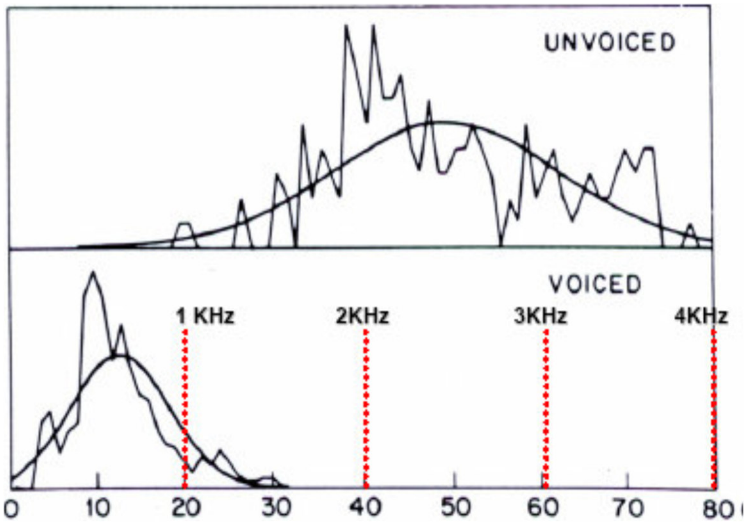
Zero Crossing Rate:

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, 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.

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. 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.



Definition of zero-crossings rate



Distribution of zero-crossing for voice/unvoiced speech

A definition of Zero Crossings Rate is:

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

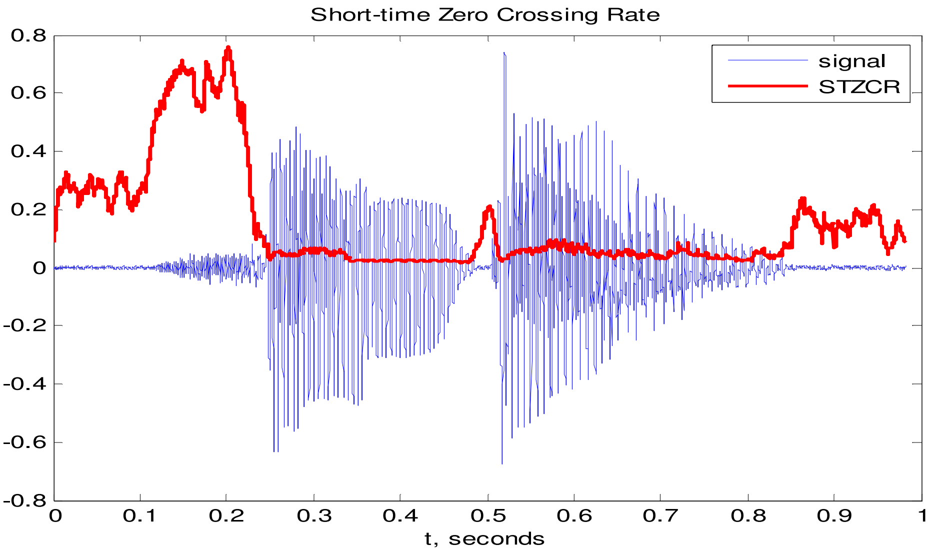
where

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

and

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

The model for speech production suggests that the energy of voiced speech is concentrated below about 3kHz 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.



Average Energy:

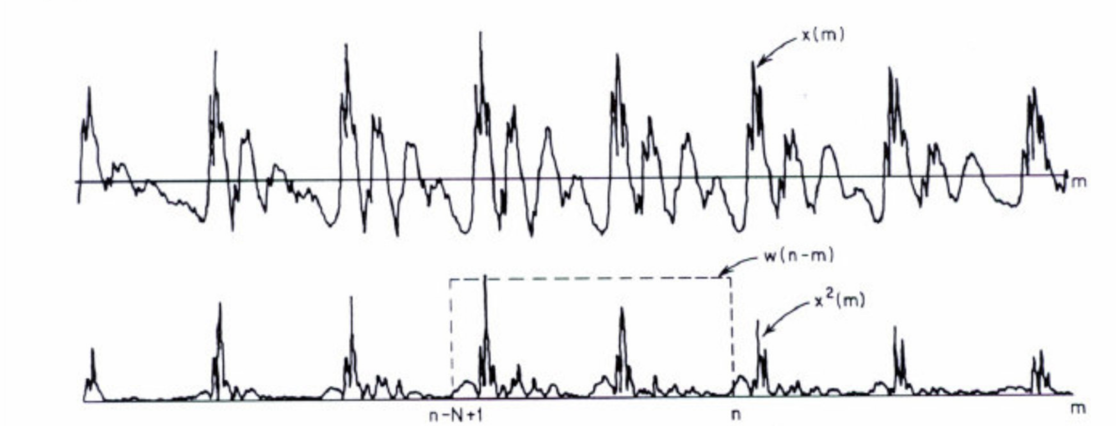
The amplitude of the speech signal varies with time. Generally, the amplitude of unvoiced speech segments is much lower than the amplitude of voiced segments. The energy of the speech signal provides a representation that reflects these amplitude variations. Short-time energy can define as:

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

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.

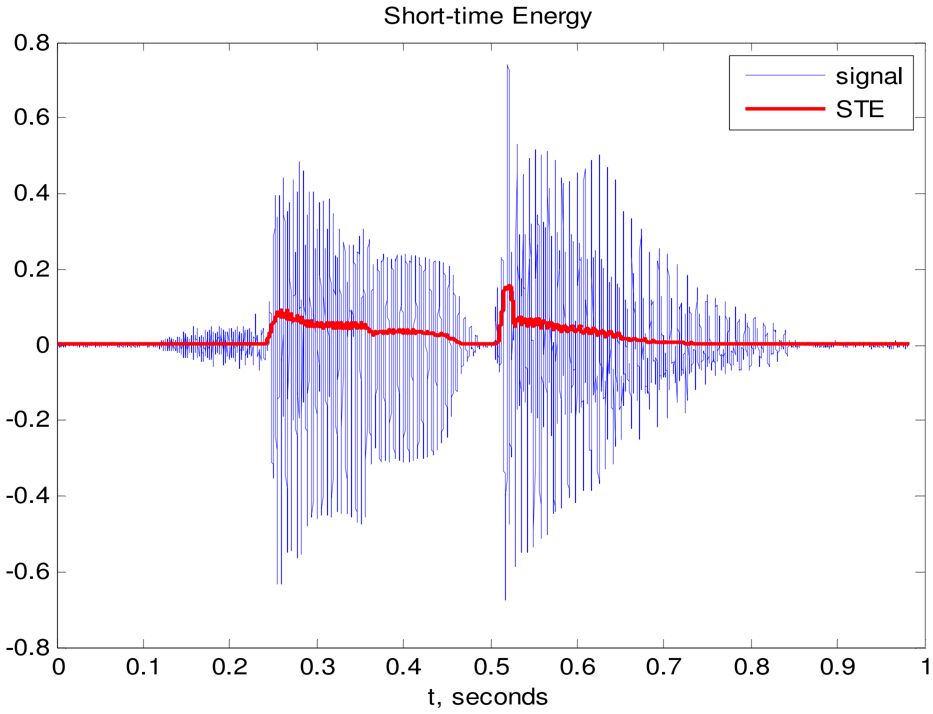
$$h(n) = 0.54 - 0.46\cos(2\pi n/(N-1)),\ 0 \le n \le N-1$$

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

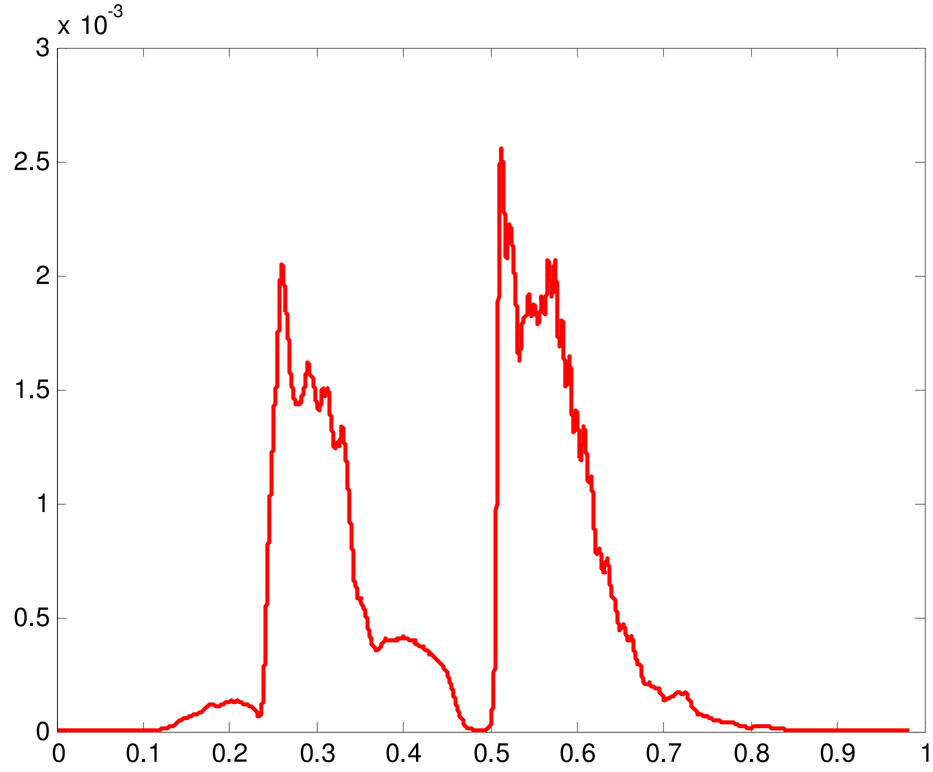


Computation of Short-Time Energy

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, n E will fluctuate very rapidly depending on the exact details of the waveform. If N is too large, n E will change very slowly and thus will not adequately reflect the changing properties of the speech signal.



Merging ZCR and Average Energy



Voice & Unvoice Segmentation

Set two thresholds and then segment all the three parts of speech signal i.e Voice, Unvoice, Silence.
The two techniques **zero crossing rate (ZCR) and average energy** are used for classification. The speech sample divided into some segments and used the zero crossing rate and energy calculations to separate the voiced ,unvoiced and silent 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.

Concept: Combining the both Zero crossing rate and average energy

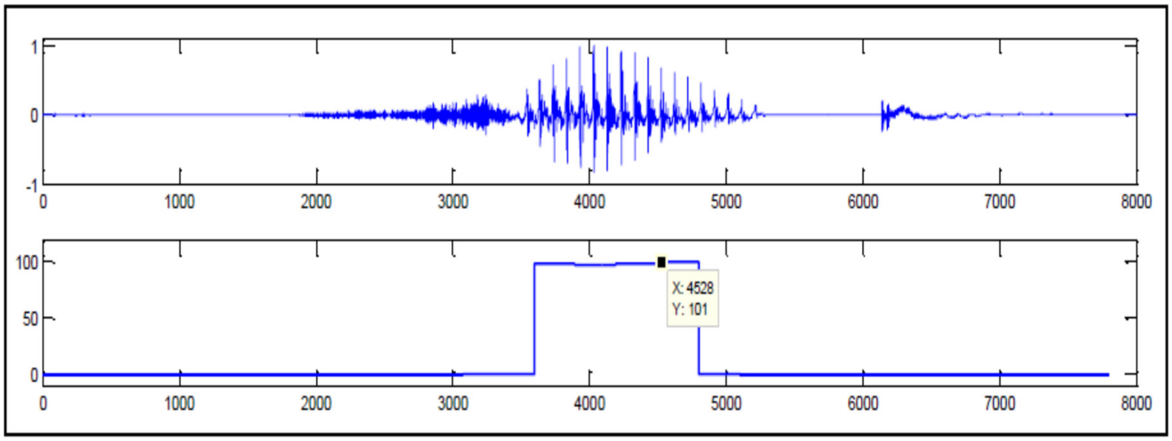
SN	Audio File name	Silent Sample	Un-Voice Sample	Voice Sample	Mean	Variance
1	testt1.wav	94424	3808	32839	0.8800	6.7138* 10-006
2	test2.wav	95980	5965	29126	0.1456	1.1112*10-006
3	test3.wav	43391	37470	29816	0.0718	5.4801*10-007
4	test4.wav	11043	37071	17421	0.0182	2.7726*10-007
5	test5.wav	43391	37470	50210	0.0718	5.4801*10-007
6	test6.wav	23461	56303	51307	0.0260	1.9872*10-007
7	test7.wav	94424	3808	32839	0.8800	6.7138*10-006
8	test8.wav	6825	1944	7614	0.0109	6.6821*10-007
9	test9.wav	11043	37071	17421	0.0182	2.7726*10-007
10	test10.wav	6825	1944	7614	0.0109	6.6821*10-007

Pitch Estimation

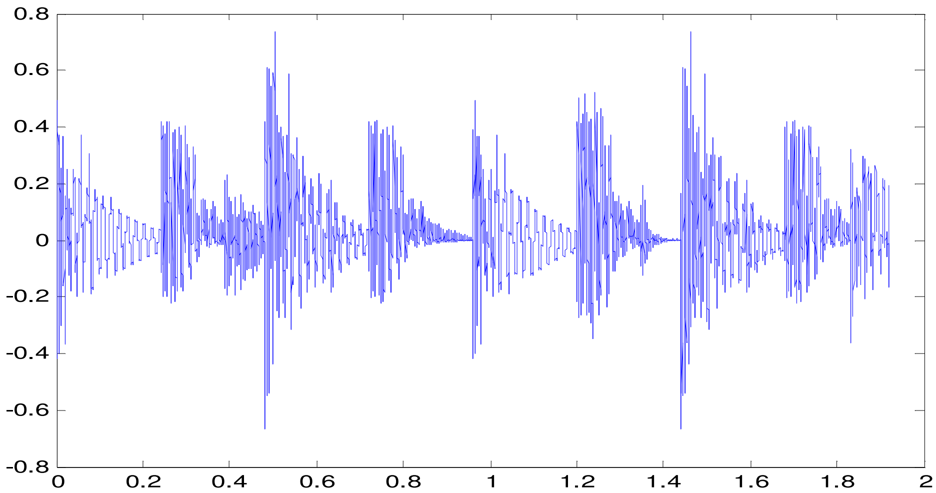
Pitch:

Pitch (fundamental frequency) provides information in speech that very useful in speech signal processing which can help for voice identification. It also helpful for speech iditifcaion, speaker identification and emotion detection. The pitch is more concern with fundamental frequency of speech. The pitch is in high frequency (voice) part and we know the Voice part is having the high correlation factor compare to unvoiced part. So the correlation may be used to estimate to find out the pitch in given speech segment.
Some of the pitch estimation are Modified Autocorrelation Method (AUTOC) , Cepstrum Method (CEP) and Average Magnitude Difference Function (AMDF) , offer a logical algorithm that perform well on average, but some time fails . Such pitch detection algorithms are not sufficient if the purpose of the application is to analyze the behavior of the pitch contour.

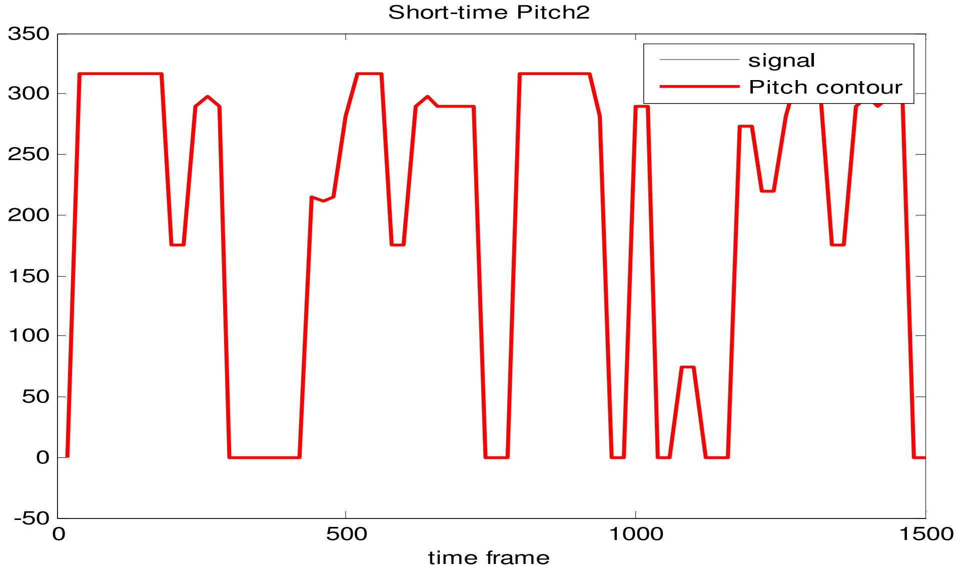
Pitch contour plot of the word 'Shout'



Audio Signal:



Pitch Contour:



Pitch Estimation male/female:

Serial No	File	Pitch	Male (Pitch range:60-170) Female (Pitch range:170-430) Higher Pitch (>400) Lower Pitch (<60)
1	./temp/test01.wav	135.5932	MALE Voice
2	./temp/test02.wav	275.8621	FEMALE Voice
3	./temp/test03.wav	150.9434	MALE Voice
4	./temp/test04.wav	119.4030	MALE Voice
5	./temp/test05.wav	500.00	No Male/Female Voice(High Pitch)
6	./temp/test06.wav	235.2941	FEMALE Voice
7	./temp/test07.wav	156.8627	MALE Voice
8	./temp/test08.wav	205.1282	FEMALE Voice
9	./temp/test09.wav	228.5714	FEMALE Voice
10	./temp/pitch_test1_440.wav	440.1198	Test pitch function
11	./temp/ftest11.wav	296.2963	FEMALE Voice
12	./temp/ftest12.wav	380.9524	FEMALE Voice

M-Male Pitch: Male (Pitch range:60-170)
F –Female Pitch: Female (Pitch range:170-430)
H-High Pitch: Higher Pitch (>430)

Result

These voice and unvoiced estimation is based on combining zero crossing rates, average energy and the Threshold value. The pitch estimation has been implemented with statistical correlation algorithm.