Automatic Pitch Detection Using Speech Segmentation and Autocorrelation

by

Estefany Carrillo and Bathiya Senevirathna

Supervised by

Dr. Uchechukwu Abanulo

Abstract

The beauty of human speech lies in the complexity of the different sounds that can be produced by a few tubes and muscles. This intricacy, however, makes speech processing using technology a challenging task. One defining characteristic of speech is its pitch, but it is oftentimes difficult to obtain this value because some segments of speech simply do not have a measureable pitch. This paper explores techniques in determining the pitch of a speaker in a noiseless utterance, starting with the classification of types of speech and automatic pitch detection.

1. Introduction

Speech can be classified into two general categories, voiced and unvoiced speech. A voiced sound is one in which the vocal cords of the speaker vibrate as the sound is made, and unvoiced sound is one where the vocal cords do not vibrate. A given speech utterance could contain a mix of different voiced and unvoiced segments depending upon the. Pitch detection relies on the periodic qualities of the sound waveform, therefore any attempt to determine pitch is only valid on voiced segments of an utterance. This factor necessitates a secondary step before pitch detection can take place: the voiced segments have to be identified. Therefore this paper presents a speech classification algorithm as part of pitch detection.

Figure 1 below is an energy plot of the word “Shout”

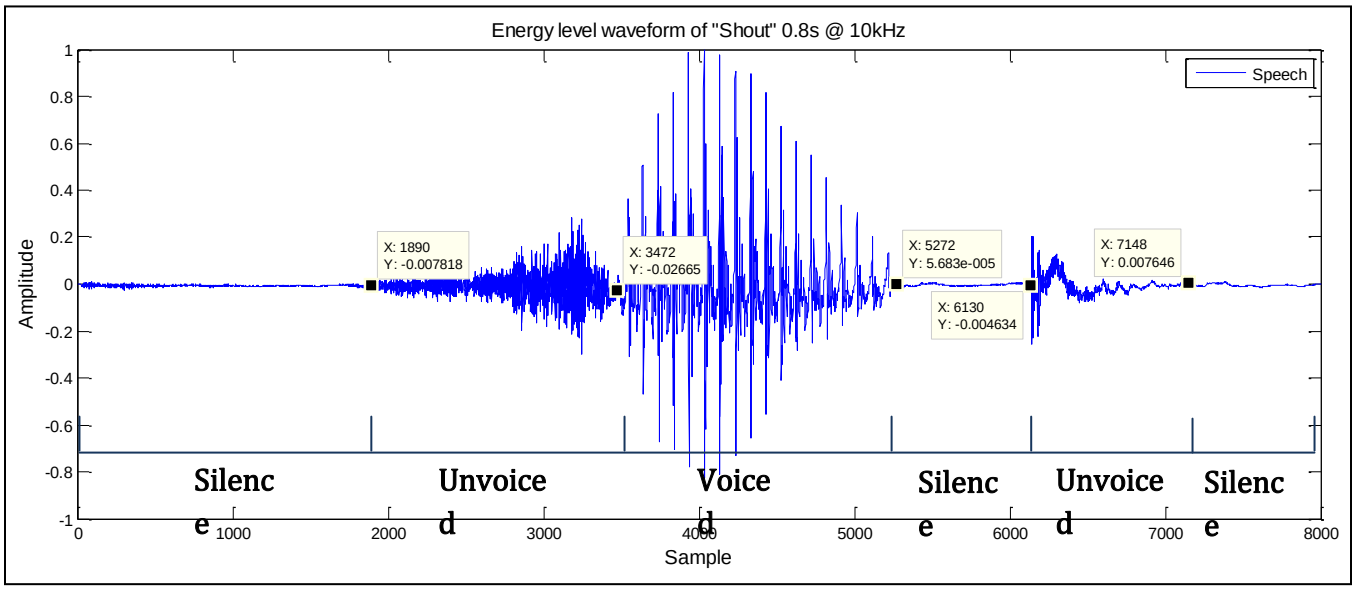

Figure 1: An energy plot of the word “shout” recorded at a sampling frequency of 10 kHz for 800 milliseconds. The waveform is annotated the location of boundaries of speech types.

For performance analysis, the ideal boundaries of voiced, unvoiced, and silence portions have been marked manually.

Table I below shows the speech type classifications for the word "Shout"

Lower sample boundary	Upper sample boundary	Classification
1	1890	Silence
1891	3472	Unvoiced
3473	5272	Voiced
5273	6118	Silence
6119	7155	Unvoiced
7156	8000	Silence

Table I: A listing of the locations of voiced, unvoiced, and silence segments corresponding to **Figure 1**.

2. Speech Classification Algorithm

The speech classification algorithm presented here is based on the energy levels of the speech waveforms and the recognizing different sound types based on certain threshold values. Two methods for calculating the scaled energies are presented here, one that is based directly on the amplitude of the energy, and another that takes the periodic characteristic of the wave into consideration. The segmentation process is outlined as follows:

1. Divide the sound wave into frames of 10ms
2. Assign a scaled energy value to each sample. Two methods are presented in this paper:
    a. Using the average value of the squares of each energy point
    b. Using the periodicity of the wave

    The equations used in each method are discussed in detail in section 1.3.
3. Using the resulting scaled energy levels, obtain two thresholds by which classification is conducted:

    silence and unvoiced thresholds.
4. Cycle through each frame and assign them speech classifications in a marker array
    a. If the scaled energy of the frame is less than the silence threshold, mark as 0
    b. If the scaled energy of the frame is less than the unvoiced threshold, mark as 0.5

c. All other frames, mark as 1

5. Using the markers the voiced, unvoiced, and silence portions of the sound can be indentified

2.1 Scaled Energy Level Calculations

The amount of samples in a given length of sound is defined by the sampling rate at which the sound was recorded, given in samples per second. The sound analyzed in this section has been recorded at a rate of 10 kHz. Even at this relatively low rate, however, the computational time needed to analyze each and every sample would be large. Therefore the sound waveform was segmented into frames 10ms in length. This size reduces computational time without sacrificing too much quality. A scaled energy value was then assigned to each frame that would be used to classify the frame as voiced, unvoiced, or silence.

2.1.1 Scaled Energy: Average Value Method

The first method involves a direct analysis of the energy of each sample. The Scaled Energy (SE_1) is calculated using the given equation:

$$SE_1 = \frac{a \times \sum(Energy)^2}{Frame\ Size} = \frac{a \times \sum_{i=l}^{u}(E_i^2)}{u-l} \quad (1)$$

Where *a* is a scaling factor, *E* is energy amplitude for each sample, and *u* and *l* are the upper and lower indices of the frame. *E* is first squared in order to factor out the sign of the amplitude, since the waveform has both positive and negative values.

2.1.2 Scaled Energy: Peak-to-Average Ratio

The second method takes the characteristics of voiced and unvoiced waveforms into consideration. Voiced segments appear to be more rarefied; the points are dense in the low energy levels, but have very high peaks. Unvoiced segments on the other hand are dense in low energies without much variation. This method uses these characteristics by obtaining a ratio between the maximum energy and the average energy of a sample. This value would be high for the voiced portions and low for other, "noisier" segments. This ratio R is calculated using the following expression:

$$R = \left| \frac{b \times Maximum\ Energy}{Average\ of\ Absolute\ Energies} \right| = \left| \frac{b \times E_{max}}{[\sum_{i=l}^{u}|E_i|/(u-l)]} \right| \quad (2)$$

Where *b* is a scaling factor. Once R is obtained, it can be used to weight the average scaled energy as shown in the following equation:

$$SE_2 = (c \times R) + \left(d \times \frac{\sum_{i=l}^{u}|E_i|}{u-l}\right) \quad (3)$$

Where *c* and *d* are scaling factors.

2.2 Testing & Results

Graphs of the scaled energy values were required to obtain silence and unvoiced threshold levels.

Figure 2 below plots the original sound and the two scaled energy assignments SE_1 and SE_2 using scaling factors of 5, 0.1, 1, and 2 for *a, b, c,* and *d* respectively.

Figure 2 below is energy plot of the word "Shout" with its scaled energy levels

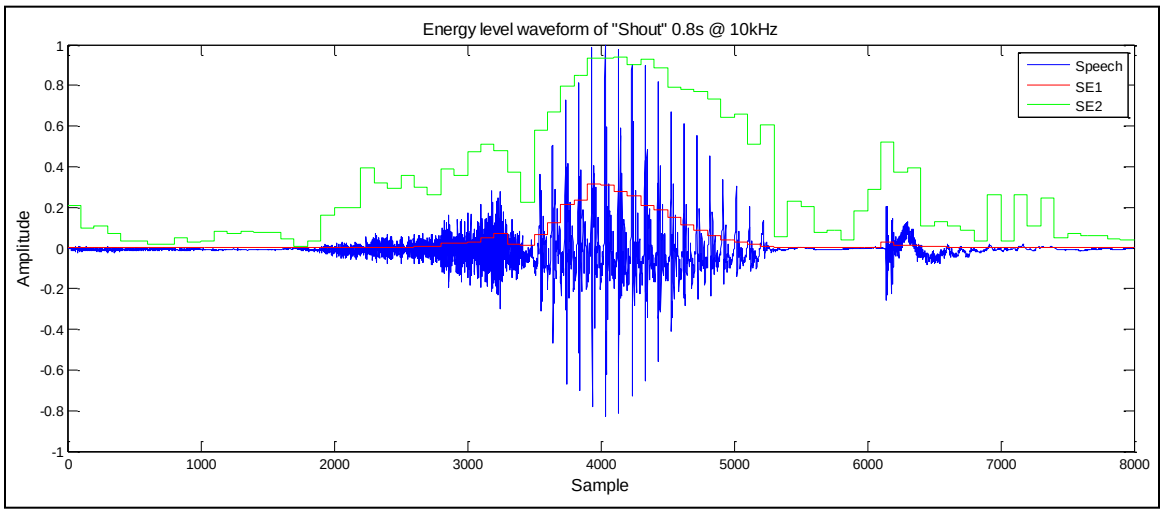

Figure 2: An energy plot of the word "shout" that displays the scaled energy levels calculated using the two methods discussed previously. SE1 (red) corresponds to the Average Value Method; SE2 (green) corresponds to the Peak-to-Average Ratio Method.

It can be seen that both SE_1 and SE_2 show similar patterns but differ in a few key areas. Both curves highlight large amplitude portions but SE_2 is more sensitive to periodic portions of the sound wave. Silence and unvoiced thresholds were then obtained using this plot: 0.02 and 0.10 respectively for SE_1 and 0.13 and 0.50 respectively for SE_2. Using these thresholds, the program automatically assigns a mark to each frame: 0 for silence, 0.5 for unvoiced, and 1 for voiced.

Figure 3 below shows the original sound wave, and the ideal, SE_1, and SE_2 classifications.

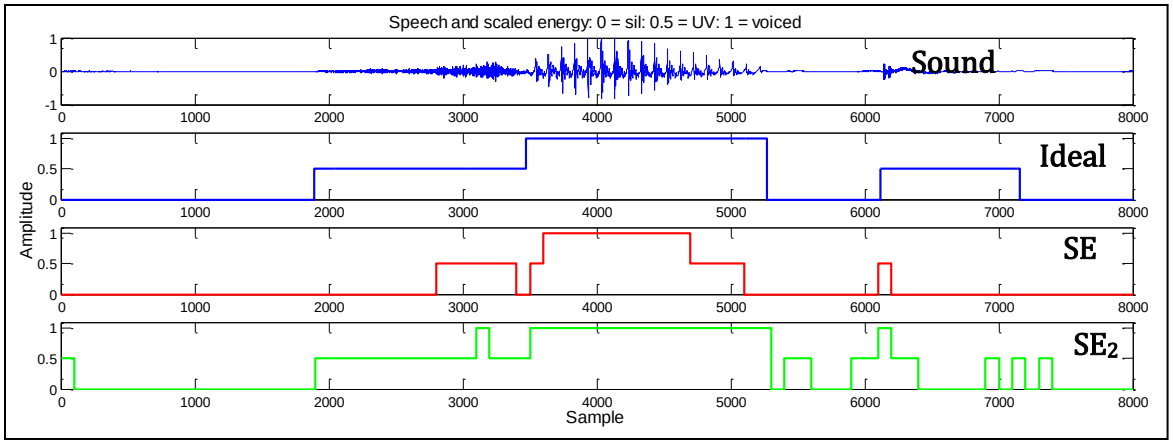

Figure 3: An energy plot of the word "shout" and its speech classifications through three methods: manual (2nd panel), average value method; SE2 (3rd panel), and peak-to-average ratio (4th panel).

In order to gauge the performance of each scaling method, the classifications acquired by each algorithm were compared to the expected values found manually. The percentage of correct markers was calculated for each algorithm: SE_1 had an accuracy rate of 66.81%, and SE_2 had a rate of 81.11%.

3. Manual Pitch Detection

Once a voiced segment is identified, its pitch can be obtained through straightforward analysis of the sound waveform. A zoomed in energy plot of a voiced segment yields a distinctly periodic waveform.

The period of the segment can be calculated by finding the time difference of two successive peaks.

Figure 4 below is an energy-time plot of the speech wave

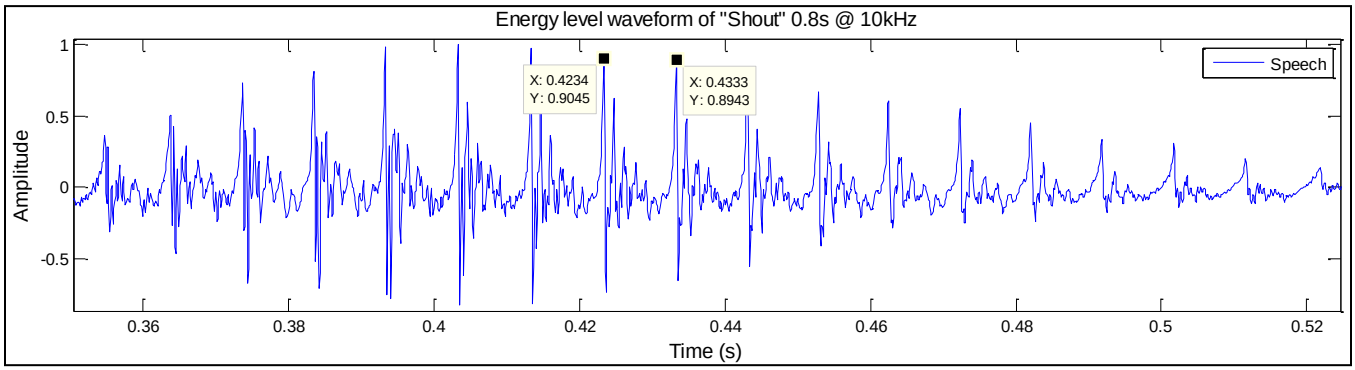

Figure 4: An energy vs. time plot of the word "shout". The plot is focused on an identified voiced segment of speech with two peaks annotated for pitch calculation.

The pitch of the voice can then be calculated as follows:

$$f = \frac{1}{T} = \frac{1}{(0.4333) - (0.4234)} = \boxed{101.01\ Hz}$$

4. Automatic Pitch Detection

The main method presented in this paper to determine pitch automatically is the process of autocorrelation. It is a method that is based on the periodic characteristic of the wave, regardless of amplitude. The procedure for estimating the pitch of the signal using the autocorrelation method is as follows:

1. Divide the sound into frames of 10ms
2. Overlay a frame on itself, shift it along the time axis, and multiply the signals together
3. If two frames are almost the same, the plot of the area overlap after performing cross-correlation will show a distinctive peak
4. The lag of this peak is considered the period for that frame, and the pitch can then be calculated

By pre-processing the sound file using the speech classification methods discussed above, it is possible to discard inaccurate pitch estimations from unvoiced segments.

Figure 5 below shows the pitch contour plot of the word "Shout"

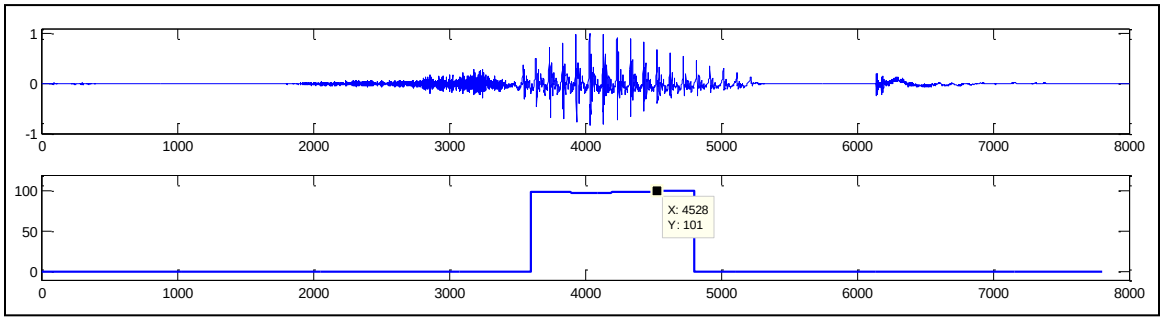

Figure 5: An energy plot of word "shout" with a corresponding pitch plot obtained using the autocorrelation method. The algorithm assigns a zero pitch to all unvoiced segments specified by a manually entered unvoiced threshold. An annotation of the identified pitch is shown as 101 Hz.

As can be seen, the pitch of the voiced portion of the sound was found to be 101Hz, which is almost the same as the pitch calculated manually.

5. Automatic Speaker Identification using Pitch

To further analyze the performance of the autocorrelation method, sample recordings of four females and ten males were obtained. Each individual recorded the sentence: "This is an exciting project" twice. One sample from the pool was chosen as a reference and was compared to the other samples. The algorithm sought to match the reference speaker with his/her other sample in the test pool.

Figure 6 below shows the speech and pitch contour plots of the reference sample an

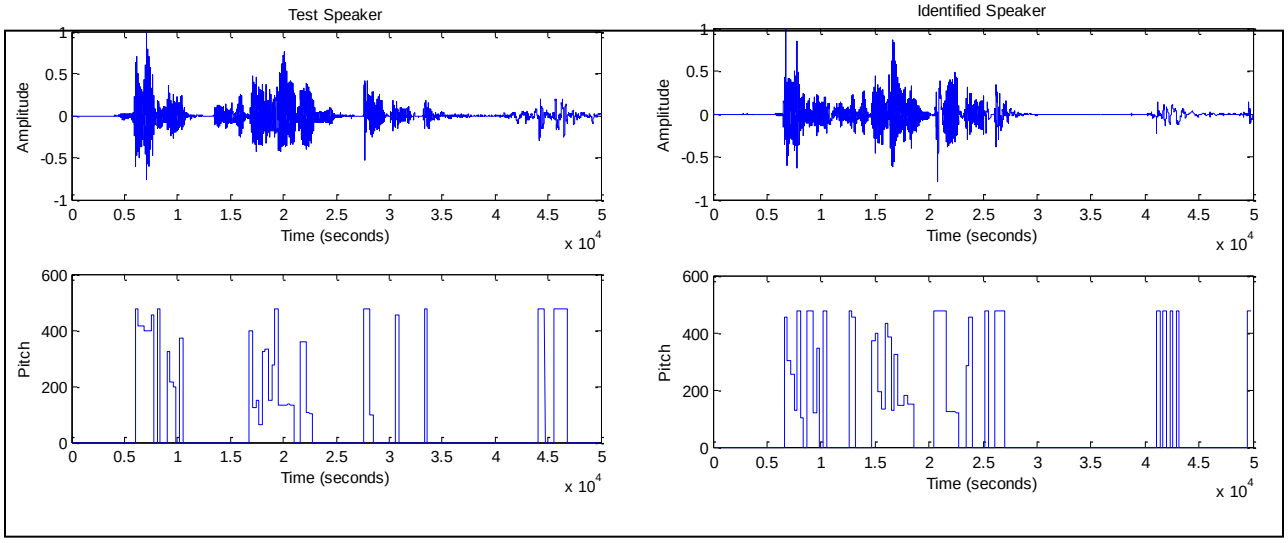

Figure 6: An energy plot of "This is an exciting project" with a corresponding pitch plot obtained using the autocorrelation method. The two left panels are from the reference sample and the right panel is derived from the sample that was identified as the least discrepant in terms of pitch.

As can be seen the pitch amplitudes matched fairly well. The sample identified as the best match was in fact from the same speaker as the reference sample.

6. Conclusion

1. It is evident from the presented results that a speech classification technique based on Peak to Average Ratio performs better than a system that simply looked at energy amplitudes. This is particularly important in instances when the intensity of the sound varies through the speech utterance. This technique, however, does have a drawback for the very same reason that it works well: it is sensitive to all periodic patterns in the sound wave. This was evident in the scaled energy graph in Figure 2. The first 2000 samples clearly contain silence; SE_1 demonstrates this fact well with

a near-zero scaled energy. SE_2 on the other hand is irregular for this same section of the wave. This was because the waveform in this segment was rarefied, albeit on a smaller energy scale. The challenge is to vary the scalar factors b, c, and d in order to weight the R factor adequately for the entire utterance.

2. The autocorrelation method of pitch detection was quite accurate in determining the pitch of a speaker *using a pre-processed sound signal*. It was necessary to first remove, or at least identify, the unvoiced portions of the sound because they are not periodic and would therefore give an erroneous result. One method of facing this issue without have to find voiced portions first is to set an acceptable range for pitch. Human speech is fairly limited in terms of pitch. Thus it would be possible to only recognize pitch values within a certain range, 70Hz to 220Hz for males, for example.