A Wavelet-Based Approach for Pitch Classifiers

José R. Ovalles Torres
Master of Engineering in Electrical Engineering
Luis M. Vicente, Ph.D
Electrical & Computer Engineering Department
Polytechnic University of Puerto Rico

Abstract *— A pitch is a human classification method which allows ordering sounds in discrete harmonics frequencies. Although many of the practical pitch detection methods uses the autocorrelation and cepstrum techniques to detect audio pitch, a wavelet-based classifier was developed using wavelets to create a unique signal coding for each pure pitch signals. A comparative study of how the wavelets coefficients vary on each of the seven proposed pure pitch signals were given in this article. Also mathematical and illustrated examples of this study were presented as well.*

Key Terms *— Autocorrelation, Classifier, KNN, Wavelets, Wavelets Transforms.*

INTRODUCTION

Fourier transform is a valuable tool for analyzing frequencies of a particular signal, however, we cannot tell at what instance a particular frequency occurred [1]. The Wavelet Transform (WT) is a technique for analyzing signals. It was developed as an alternative to the short time Fourier Transform (STFT) to overcome problems related to its frequency and time resolution properties [9]. The main focus of this project is to expose other ways to approach voice pitch recognition by implementing a wavelet-based voice pitch classifier.

AUDIO SIGNALS

Audio signals are generally referred to as signals that are audible to humans. Audio signals usually come from a sound source that vibrates in the audible frequency range. There are many ways to classify audio signals. An audio stream can be segmented into many categories such as silence, environmental sound, music, and speech [2]. Music normally has a wide range frequency distribution among the audible range of human, from 20Hz to 20k Hz. As we know the bandwidth of the speech signal is usually limited into 50 to 7k Hz. Thus, the spectral centroids of music signal are higher than that of the speech [4].

FOURIER

One of the most basic and ubiquitous methods in signal processing is the Fourier Transform (FT). It transforms a signal from time space into frequency space. The Discrete Fourier Transform (DFT) takes a signal, $z(n)$, with N samples and calculates the coefficients, $Z(m)$, corresponding to equally spaced frequencies starting with period equal to the length of the signal and ending with the sampling frequency fs [5].

$$Z(m) = \sum_{n=0}^{N} z(n) e^{\frac{(-2\pi i)(m)(n)}{N}} \qquad (1)$$

Where $N \in Z^{+}$ and $m,n \in \{0,1,2,\ldots,N\}$. The $Z(m)$ coefficient corresponds to the frequency $(m\,fs)$[5].

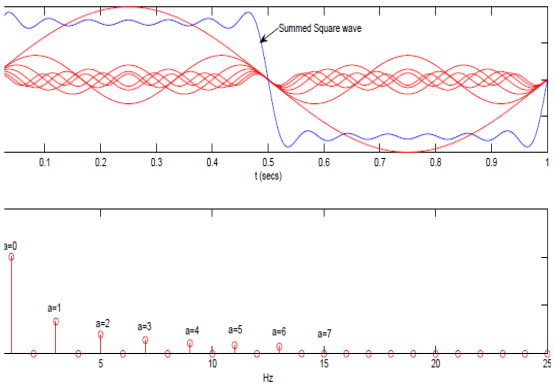

Figure 1
A Square Wave

In the square wave example of Figure 1.1, we can look at the coefficients of the DFT and see that

the appropriate frequencies are in the signal. However, this is only effective when the frequencies are evenly distributed across the signal. If we look at an example that is un-evenly distributed in time and frequency (Figure 1.2), we can see that there are peaks for the frequency components of the signal at ±16Hz and ±3Hz, but they are not isolated. The presence of the other frequencies in this spectrum is necessary to flatten out the quiet parts. If we constructed a signal based on these coefficients on the interval 0s to 12s, there would be two exact repetitions of the signal shown at the top of Figure 1.2. If we look at the highest peaks we can identify what the fundamental components are, though this is not always the case. Because we are in the frequency domain, we cannot identify where in time these frequencies are present based solely on the spectrum. If we could pick out the frequencies in a signal, it would not do us much good just to know that they existed at some point in a long signal. It would be more useful if we looked at shorter sections of a long signal, because using a FT on the shorter sections, we can identify which frequencies exist at the time and for the duration of each short section. [5]

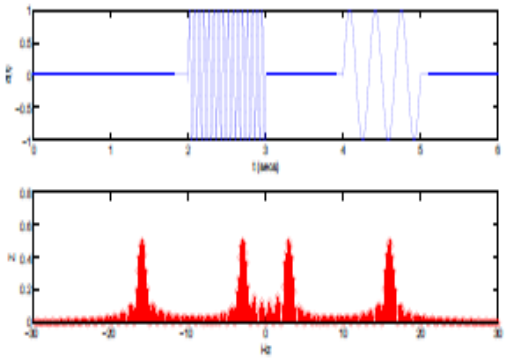

Figure 2
A 16Hz Sine Wave on the Interval 2-3s and a 3Hz Sine Wave on the Intervals 4-5s

INTRODUCTION TO PITCH DETECTION

Pitch detection is an essential task in a variety of speech processing applications [3]. The period of a Pitch is a fundamental parameter in the analysis process of any physical model. A pitch detector is basically an algorithm that determines the fundamental pitch period of an input musical signal and when it happened. Pitch Detection of musical signals is not a trivial task due to some difficulties such as the transient attacks in low and high frequencies [4].

AUTOCORRELATION FUNCTION FOR PITCH DETECTION

The autocorrelation function could be used as a time domain pitch detector. For non-stationary signals, short-time autocorrelation function is defined for the following function:

$$ph(m)=\frac{1}{N}\sum_{n=0}^{N-m-1}[f(n+l)w(n+l)][f(n+m+l)w(n+m+l)], \quad (2)$$

$$0 \le m \le M_0 - 1,$$

Where $w(n)$ is a window function, *N* the frame size, l is the index of the starting frame, *m* is the autocorrelation parameter and M_0 is the total number of points to be computed in the autocorrelation function [4]. When we obtain the highest peak on the autocorrelation function at *m*=0, we also determine the average power of the input signal.

DISCRETE WAVELET TRANSFORM

The Wavelet Series is just a sampled version of **Continuous Wavelet Transform (**CWT) and its computation may consume significant amount of time and resources, depending on the resolution required. The Discrete Wavelet Transform (DWT), which is based on sub-band coding, is found to yield a fast computation of Wavelet Transform. It is easy to implement and reduces the computation time and resources required. [8]

The foundations of DWT go back to 1976 when techniques to decompose discrete time signals were devised. Similar work was done in speech signal coding which was named as sub-band coding. In 1983, a technique similar to sub-band coding was developed which was named pyramidal coding. Later many improvements were made to

these coding schemes which resulted in efficient multiresolution analysis schemes. [7]

In CWT, the signals are analyzed using a set of basis functions which relate to each other by simple scaling and translation. In the case of DWT, a time-scale representation of the digital signal is obtained using digital filtering techniques. The signal to be analyzed is passed through filters with different cutoff frequencies at different scales. [7]

USING DISCRETE WAVELET TRANSFORM FOR PITCH DETECTION

A transform method is just another way to represent an arbitrary signal without changing its original information content. When we talk about the Wavelet Transform we are representing the arbitrary signal by its time-frequency domain. One of the benefits that the Wavelet Transform method has is that it can analyze non-stationary signals and use a multi-resolution technique to analyze different frequency with different resolution simultaneously. Another benefit that the Wavelet Transform analysis has is that it uses finite energy signals called wavelets to analyze a wave signal. The DWT (Discrete Wavelet Transform) is a computational method of the CWT that was developed due to the significant consumption of time and resources when computing the CWT method.

DWT can be decomposed simultaneously by using a low-pass and a high-pass filter. The outcomes of both filters are the approximation coefficients (low-pass filter) and the detail coefficients (high-pass filter). We can describe this procedure by using the following equations:

$$y_{low}(n) = \sum_{k=-\infty}^{\infty} x[k]g[2n-k], \quad (3)$$

where $g(n)$ *is a low pass filter*

$$y_{high}(n) = \sum_{k=-\infty}^{\infty} x[k]h[2n-k] \quad (4)$$

where $h(n)$ *is a high pass filter*

By performing the DWT to a signal, the output signals are down sampled by half the sampling rate, therefore, in order to reconstruct the original signal we need to up sample the filter signals as is show in the diagram below: [1]

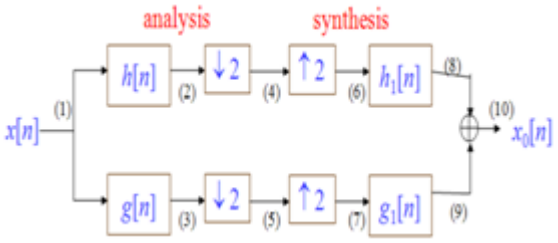

Figure 3
The Schematic Diagram to Realize Discrete Wavelet Transform

UNDERSTANDING THE MALLAT ALGORITHM

The DWT is computed by successive lowpass and highpass filtering of the discrete time domain signal as shown in figure 1.3. This kind of daisy chain filtering is called the Mallat algorithm or the Mallat-tree decomposition. The low pass filter is denoted by G0 while the high pass filter is denoted by H0. At each level, the high pass filter produces detail information d[n], while the low pass filter associated with the scaling function produces approximations a[n]. [7]

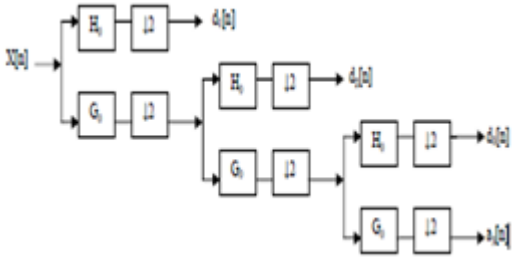

Figure 4
Three-Level Wavelet Decomposition Tree

At each decomposition level, the half band filters produce signals spanning only half the frequency band. This doubles the frequency resolution as the uncertainty in frequency is reduced by half. In accordance with Nyquist's rule if the original signal has a highest frequency of ω, which requires a sampling frequency of 2ω radians, then it now has a highest frequency of ω/2 radians. It can now be sampled at a frequency of ω radians thus discarding half the samples with no loss of

information. This decimation by 2 halves the time resolution as the entire signal is now represented by only half the number of samples. Thus, while the half band low pass filtering removes half of the frequencies and thus halves the resolution, the decimation by 2 doubles the scale. [7]

With this approach, the time resolution becomes arbitrarily good at high frequencies, while the frequency resolution becomes arbitrarily good at low frequencies. The filtering and decimation process is continued until the desired level is reached. The maximum number of levels depends on the length of the signal. The DWT of the original signal is then obtained by concatenating all the coefficients, a[n] and d[n], starting from the last level of decomposition. [7]

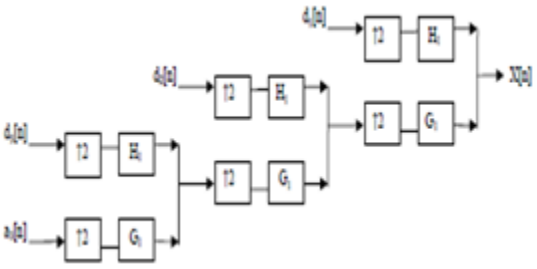

Figure 5
Three-Level Wavelet Reconstruction Tree.

Figure 5 shows the reconstruction of the original signal from the wavelet coefficients. Basically, the reconstruction is the reverse process of decomposition. The approximation and detail coefficients at every level are up sampled by two, passed through the low pass and high pass synthesis filters and then added. This process is continued through the same number of levels as in the decomposition process to obtain and the original signal. The Mallat algorithm works equally well if the analysis filters, G_0 and H_0, are exchanged with the synthesis filters, G_{11}. [7].

PITCH TRACKING

We can use Fast Lifting Wavelet Transform which uses Haar wavelet as mother wavelet for pitch tracking. We can regard fast lifting wavelet transform with Haar wavelet as a signal passing through a low pass filter and then down sampling to generate a rough component of this signal and passing through a high pass filter and then down sampling to generate a detail component. The equation of fast lifting wavelet transform with Haar wavelet is [6]

$$d_0 [n] = x [2n +1] \quad (5)$$

$$a_0 [n] = x [2n] \quad (6)$$

$$d_1 [n] = d [n] - a [n] \quad (7)$$

$$a_1 [n] = a [n] + d [n] \quad (8)$$

In the above equation, x[n] is the input signal, a_1 [n] is the first rough component, and d1 [n] is the first detail component [6].

After we separate two components from the input signal, we can retain the rough components and abandon the detail components. And then repeat the above step using the rough components until the underlying periodic waveform is shown. [6]

IMPLEMENTING DWT USING HAAR WAVELETS FOR PITCH DETECTION

In order to achieved a voice pitch classifier using Haar wavelets we started defining which harmonic pure sine waves were going to be used for the approach of the voice pitch classifier method. We simulated the classifier by using one-dimensional Discrete Wavelet Transform in order to obtain the coefficients for the classifier and assume the following constant values for the sampling frequency, signal duration, and sampling time.

$$f_s = 44100\ hz$$

$$duration = 0.1\ sec$$

$$t_d = \frac{1}{f_s} = \frac{1}{44100} \approx 0.1221\ x\ 10^{-3}\ sec$$

$$Samples = 4410\ samples$$

We selected the first seven basic notes on the fourth octave with are: C4, D4, E4, F4, G4, A4, and

B4. The frequency measurements for these basic pure sine waves were the following:

Pure Sine Wave American Standard Terms	Frequency
C4	261.63 Hz
D4	293.66 Hz
E4	329.63 Hz
F4	349.23 Hz
G4	392.00 Hz
A4	440.00 Hz
B4	493.88 Hz

Figure 6
Pure Sine Wave Frequencies

After generating all the pure sine signals we perform a Fast Lift Wavelet Transform using Haar Wavelets to obtain the wavelet coefficient from order 1 to 3 of each pure sine signal. In the figure below all the approximation output signals were concatenated three times for each pure sine signal.

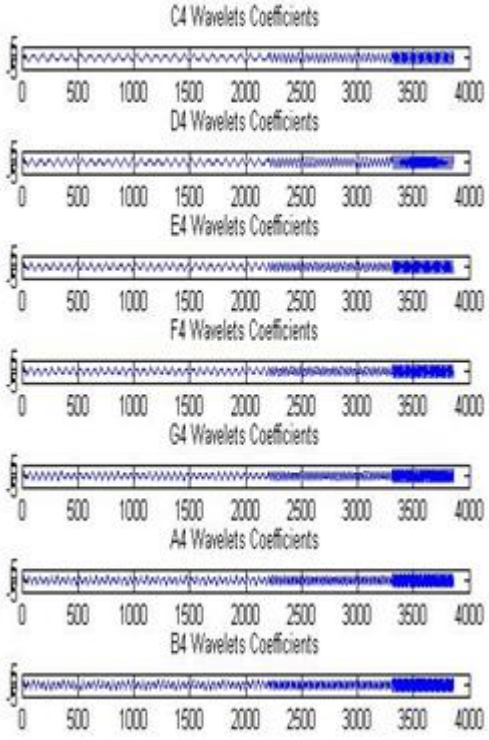

Figure 7
Level 3 Wavelet Coefficients (C4-B4)

We can see that in the figure above that for each approximation level output signal the result signal is down sampled by half its frequency rate. These approximation coefficients from each pure sine signal were used as the training set for the pitch classifier.

RESULTS AND DISCUSSION

We simulated a pitch detection classifier in matlab in order to study the efficiency of the classifier. For the following simulation we used the same approximation coefficients of the training set and the approximation coefficients of different instruments pitch audio wave clips show in the figure 8.

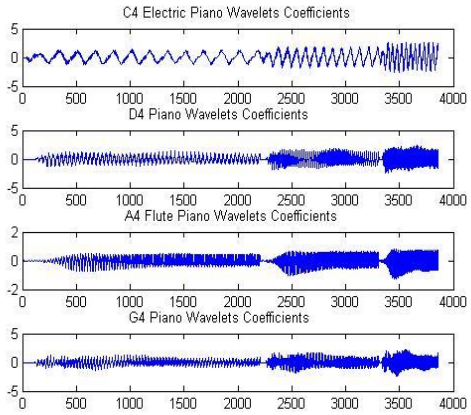

Figure 8
Wavelets Coefficients for Different Piano Sounds

After performing the classification process by using matlab *knn* classify function with Euclidean comparison, I proceed to make a comparison table that show the different outputs of the classifier depending on the level of approximation coefficient used. As result of the experiment I found that by maintaining the same sample duration and changing the approximation coefficient levels, the best results were gathered when the approximation coefficients level was 3. The table 1 shows the results.

Table 1
Efficiency Results by Level of Decomposition

Expect Results	LVL 1	LVL 2	LVL 3	LVL 4	LVL 5	LVL 6	LVL 7
1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7
1	1	1	1	1	1	1	1
2	7	7	7	7	7	7	7
6	1	1	6	7	7	7	7
5	5	5	5	5	7	7	7
Efficiency	**81 %**	**81 %**	**90 %**	**81 %**	**72 %**	**72 %**	**72 %**

CONCLUSION

In this paper, the results of the experiment concluded that the most likely solution of implementing the pitch detection using the wavelet transforms was to use the Fast Lifting wavelet transform, which is similar to the Mallat decomposition tree algorithm. However the best result was obtained by using a level 3 decomposition tree filtering. The objectives of creating a classifier for mid tones (C4-B4) were almost achieved; however few adjustments will be needed in order to have a 99.99% of reliability that the algorithm will classify without errors. The next step of the classifier will be to detect multiple pitch sounds in the same or different time intervals and be able to classify them and determine at which time frame was the pitch tone detected. By achieving the next step of the classifier, it will be easier to implement the midi converter and achieved the main objective of the project.

ACKNOWLEDGEMENT

I will like to thank God and my mentor Dr. Luis Vicente for helping me to complete this article. I hope to be able to work with Dr. Luis Vicente in other future projects.

REFERENCES

[1] Chun-Lin, L., "A tutorial of Wavelet Transforms", February 2010, pp 1.

[2] Yen, J., "Wavelet for Acoustics", n.d.., pp 1.

[3] Ahmadi, Sassan; Spanias, Andreas, "Ceptrum-Based Pitch Detection Using a New Statistical V/UV Classification Algorithm", May 1999, pp 1.

[4] Fitch, J. & Shabana, W., "A Wavelet-Based Pitch Detector for Musical Signals", 1999, pp 1.

[5] McCullough, J., "Using Wavelets for Monophonic Pitch", September 2005, pp 3.

[6] Wen-Chun Shih, "Time Frequency Analysis and Wavelet Transform Tutorial, Wavelet for Music Signals Analysis", May 2006, pp 10.

[7] Er. Manpreet Kaur, Er. Gagndeep Kaur, "A Survey on Implementation of Discrete Wavelet Transform for Image Denoising", June 2013, pp 3-5.

[8] Salma, F. JP. & Kumar, J., "Content-Aware Efficient Image Compression Using SPIHT for Arbitrary Resolution Display Devices", 2013, pp 2.

[9] Tzanetakis, G., "Audio Analysis using the Discrete Wavelet Transform", January 2000, pp 1.