A New and Improved Spectrogram.

Neil J Boucher[1], Michihiro Jinnai [2], and Hollis Taylor [3]
[1] SoundID, Maleny, QLD, 4552, Australia
[2] Kagawa National College of Technology, Takamatsu, Kagawa, Japan
[3] Postdoctoral Fellow, Muséum national d'Historie naturelle, Paris

Australian Institute of Physics Conference, Melbourne, December 2010

Abstract Summary

A study of the FFT-based spectrogram including one based on a 4-dimensional graphical display that includes phase. The study concludes that the FT and FFT are not suitable for most spectral analysis problems.

Keywords-component; spectrogram; sonogram; FFT; sound analysis; acoustics.

I. A New And Improved Spectrogram

The idea of using the Fourier Transform (FT) to display acoustic waveforms dates back to a few decades after Fourier first promoted the technique. In the book, "Acoustics", W.E.Donkin, Macmillan and Co, London, 1870, Donkin, the mathematician and musician, elegantly covers the application of the FT to sound analysis.

With the availability of affordable computing and the FFT in the 1960s the use of the spectrogram (or sonogram) became widespread. Since then surprisingly little accommodation has been made to the ever-increasing power of the PC, and the spectrogram has almost remained frozen the 1960's mindset.

As a result of some long-term studies of the FFT and its limitations, for the purposes of understanding how these manifest themselves in our sound-recognition software, we uncovered the shortcomings of the spectrogram as it is widely used today. We are addressing those limitations in the recognition software, but realized that an improved spectrogram was worth pursuing in its own right.

II. Method

The spectrogram of the 1960s has nowhere to display the phase information and merely discards it. We propose to display the phase as a fourth graphical dimension. Thus, as per the conventional spectrogram, we display the time as the "x" axis, the frequency as the "y" axis, and the intensity as the colour. We then incorporate a fourth dimension to turn the graphical image into a solid graphical object that displays the phase information.

Next we realized that the traditional way of selecting a bin size for the graphics must be equivalent of optimizing the display for one frequency only. We explored the idea of a variable bin size optimized for the frequency.

We also discovered that people find the full-colour graphic hard to interpret. A low-level signal (say 60 dB down on the peak level) may be displayed as (for example) a deep blue, but, that colour is as bright as the red which might represent a signal 60 dB stronger. Our solution to this is a sliding grey scale that replaces the colour scheme at a user-determined level.

We demonstrate a new spectrogram that fully displays the spectrum to a precision approaching the theoretical limits of the mathematical technique. We also show that it has significant implications for the way that the spectrogram should be interpreted.

III. Introduction

The objective of this study was essentially to improve the spectrogram from its widely used, but seriously flawed, FFT implementation. The first line of exploration was to consider the role of the phase of the FFT, a factor which is conveniently ignored in the conventional spectrogram. Additionally the magnitude of the spectrum is used without acknowledgment that its relevance is questionable unless the phase is considered. A surprise result was that far from elucidating the way forward for the FFT transform, the study showed that the FFT was fatally flawed and that no amount of tweaking could fix it (including adding in the phase information).

Most importantly, this study is not about the accuracy of the FFT per se, but about the suitability of its use as a signal analysis tool, and in particular of its application to the spectrogram. The FFT does have some known anomalies, but these are not at issue here.

This study led on to the search for a more appropriate transform for the spectrogram, and in particular the LPC transform.

IV. The Maths

The FT can be defined as

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x)\, e^{-2\pi i x \xi}\, dx,$$

for every real number ξ.

And noting that

$$e^{2\pi i\theta} = \cos 2\pi\theta + i \sin 2\pi\theta$$

and so the FT becomes an expansion in sine and cosine terms.

V. Precision Studies

We ran some tests of various waveforms using the FT, the FFT and the Fourier Integral. As was expected, the more precise methods drastically increased the processing time but made almost no difference to the end result. Similarly, using higher sample rates made no real difference.

VI. A four-dimensional approach

We began this study by assuming that more could be extracted from the FFT if the phase was considered. To incorporate the phase we would need a four-dimensional graphic and such a graphic can be seen in Figure 1 below.

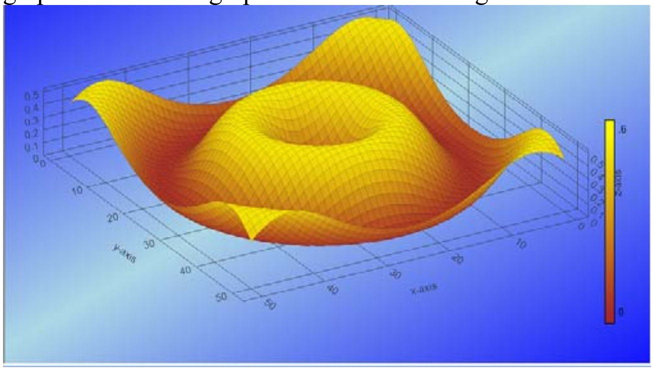


Figure 1 *A four-dimensional graphic*

The graphic as can be seen has a 3-D shape and it can be rotated and viewed from any angle as seen in Figure 2. Here the graphic has been inverted. The colour of the skin of the object can be controlled also (here it is seen in different shades of yellow, but any colour scheme can be used), and this gives a 4th dimension.

It was the intention that this graphic would form the basis of a new spectrogram,

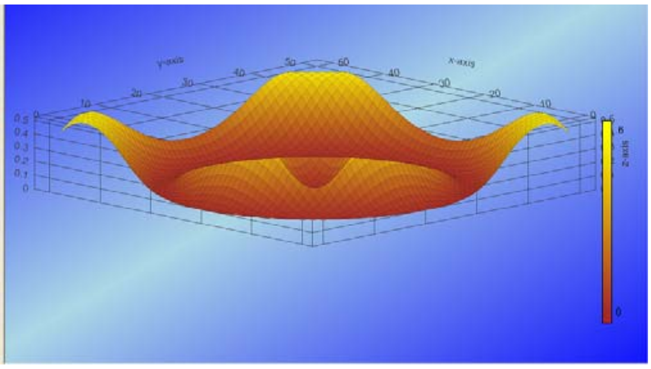


Figure 2 *The 4-D graphic rotated.*

VII. Phase Studies

So what do real sounds actually look like in the time and frequency domains? In Figure 3 we see a human female voice saying the number “one”. Notice that it can be reasonably characterized as a string of modulated pulses that vary only gradually in frequency. In fact the first half of this sound is basically a 200 Hz tone (more precisely 211 Hz, but there is a lot of variability in speech, so we will round it down to 200 Hz) with some second harmonic at 400 Hz.

Figure 4 is a Double-Eyed Fig Parrot and its call as seen in the diagram is a modulated sine wave at about 6900 Hz.

The FT is basically blind to the low-frequency modulation (which is generally lower in frequency than any practical bin period). Both voice and bird calls, can to a first approximation be seen as a string of sine wave pulses. So for example a 22 kHz signal sampled at 44.1 kHz using bin sizes of 256 points has a bin period of 256/44100=0.0058 seconds which corresponds to a frequency of 172 Hz, being the lowest frequency that can be resolved.

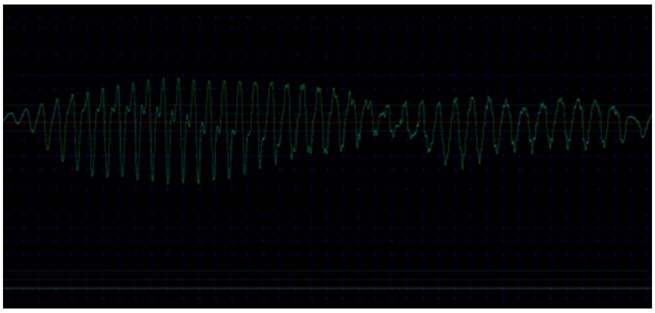


Figure 3 *Human female saying "one"*

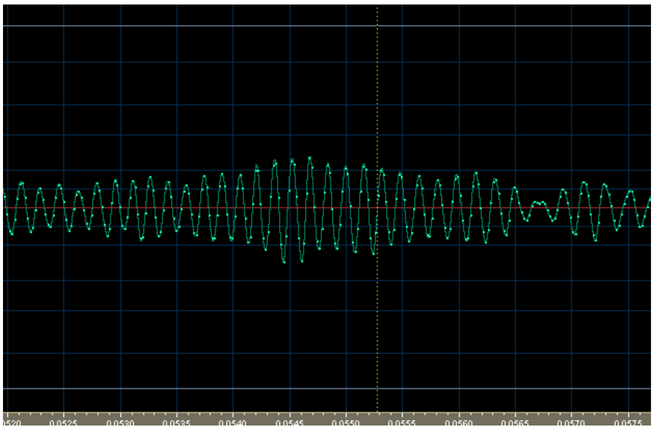


Figure 4 *A Double-Eyed Fig Parrot call.*

The contribution of the phase was studied by using (at first) a simple signal. The signal was simply a 1 kHz pulse seen in Figure 5.

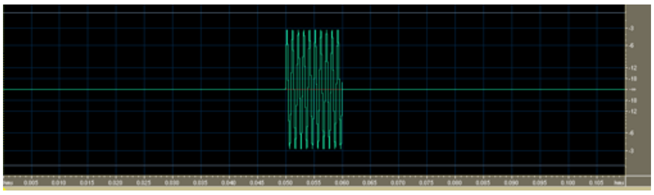


Figure 5 *A 1 kHz pulse.*

The pulse was first looked at using 128 bands of resolution (from a 44.1 kHz sample rate WAV file). A frequency domain representation of the magnitude and phase of this signal is shown in Figure 6. The black line is the magnitude of the signal while the green gives the phase in radians.

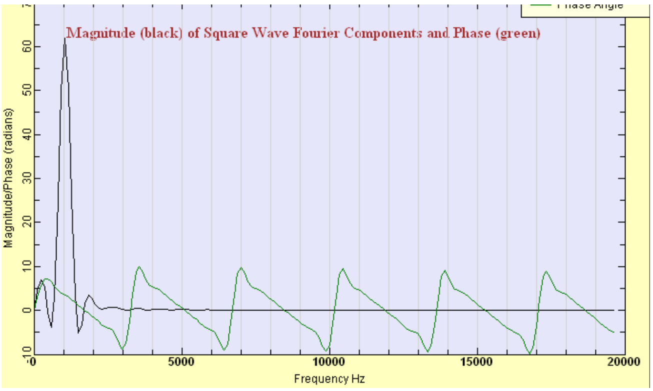


Figure 6 *The Magnitude and phase of the pulse at 128 frequency bands in the frequency domain.*

Next, this was compared with the same signal but looked at in 1024 bands as in Figure 7.

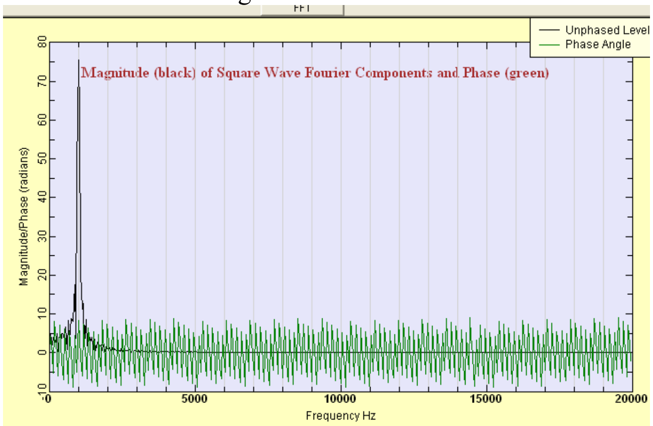


Figure 7 *The same signal at 1024 bands in the frequency domain.*

We see that at a higher resolution the phase information gets far more complex. Zooming in on the signal near the 1 kHz region we see in Figure 8 that there are significant phase transitions near the 1 kHz region. However, while the graphic tells us a lot about how the FFT behaves, it fails to reveal anything very useful about the underlying 1 kHz pulse. At this stage we began to suspect that the FFT inherently could not be coaxed to reveal much about the original signal.

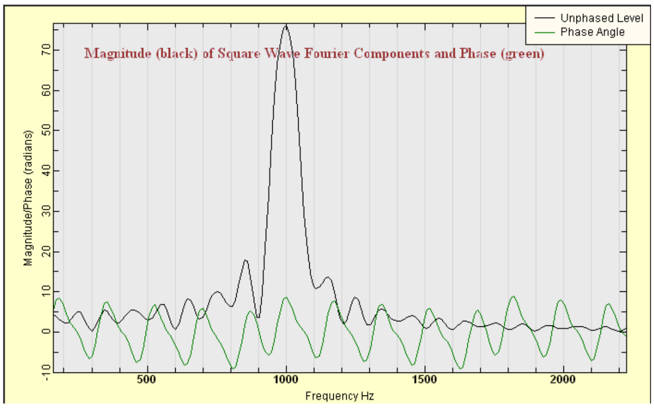


Figure 8 *A 1024 band frequency domain zoomed view in the frequency domain.*

VIII. Behaviour of the FFT with the 1 kHz signal

In this section we used a commercial FFT spectrogram program to examine how the bin sizes (number of bands) determines the spectrogram's nature. A question often asked is "what bin size is best?", and that question, as we shall see, is best answered "it depends".

We begin by looking at a very small bin size that only has 16 bands. Such a setting would rarely be used, but it does help to see the effect of using a small number of bins.

As Figure 9 shows, having a small number of bins smears the frequency over the bins. This results in a huge uncertainty in the frequency, but it does define the duration of the signal with reasonable precision. This needs to be kept in mind as we shall see later that the FFT can deliver greater certainty in frequency only at the expense of less certainty in the duration of the pulse. In fact, 16 bands gives the 10 millisecond pulse width as 10.05 seconds and the bandwidth as 0-5,000 Hz (an error >100%).

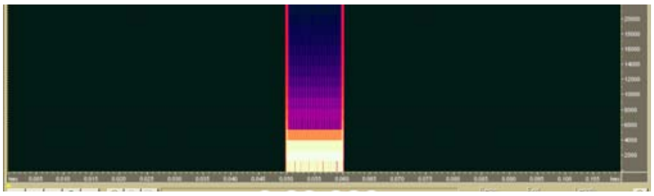


Figure 9 *The FFT with 16 frequency bands.*

Moving on to a more practical bin size of 128 gives the result of Figure 10. This shows a flaring at the discontinuity in the pulse as the FFT compensates by introducing a complex array of components that simulate this step. The compensating components extend to the 20 kHz level which is the highest level that is calculated. Additionally the duration is now seen to be 13 milliseconds (error of +30%) and the 6 dB bandwidth 500-1500 Hz (error of +/-50%). The red and purple "horned" structures are all artifacts so far as the original signal is concerned.

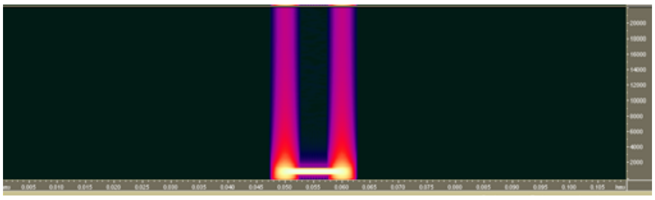


Figure 10 *The FFT with 128 frequency bands.*

Then at 256 bands the duration error increases even more as the frequency resolution improves as seen in Figure 11. The duration is now 18 milliseconds (error 80%) and the frequency bandwidth 800-1220 Hz (error +/-20%)

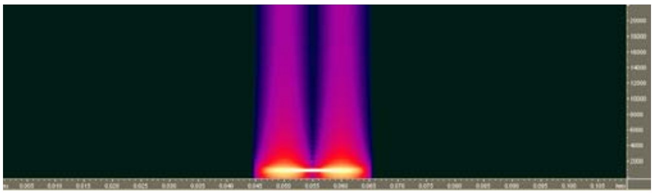


Figure 11 *The FFT with 256 frequency bands.*

Finally at 1024 bands (Figure 12) the frequency is beginning to look sharp while the duration is smeared over a long period of time, in fact over 53 milliseconds (+530% error) and now the bandwidth is 888 Hz to 1098 Hz.= (error +/- 9%)

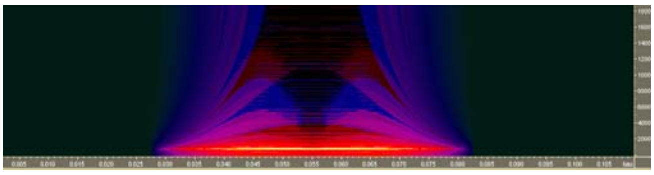


Figure 12 *The FFT with 1024 frequency bands.*

All of this can now be plotted as error % vs. bin size as seen in Figure 13.

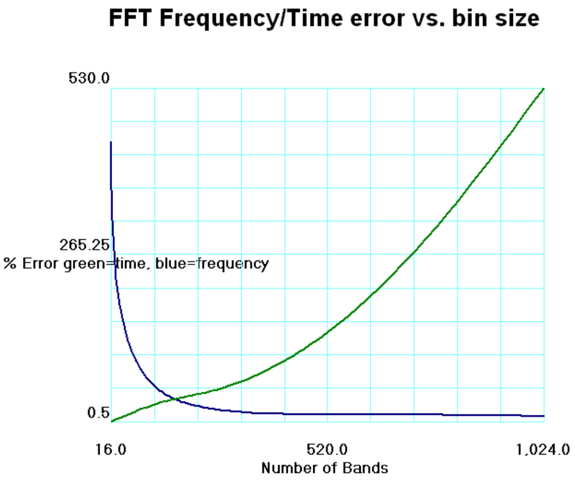


Figure 13 *The error in frequency and duration vs bandwidth.*

A few approximations can be gleaned from this.

First, if

F_e = frequency error percentage (+/-)

T_e = duration error percentage (+)

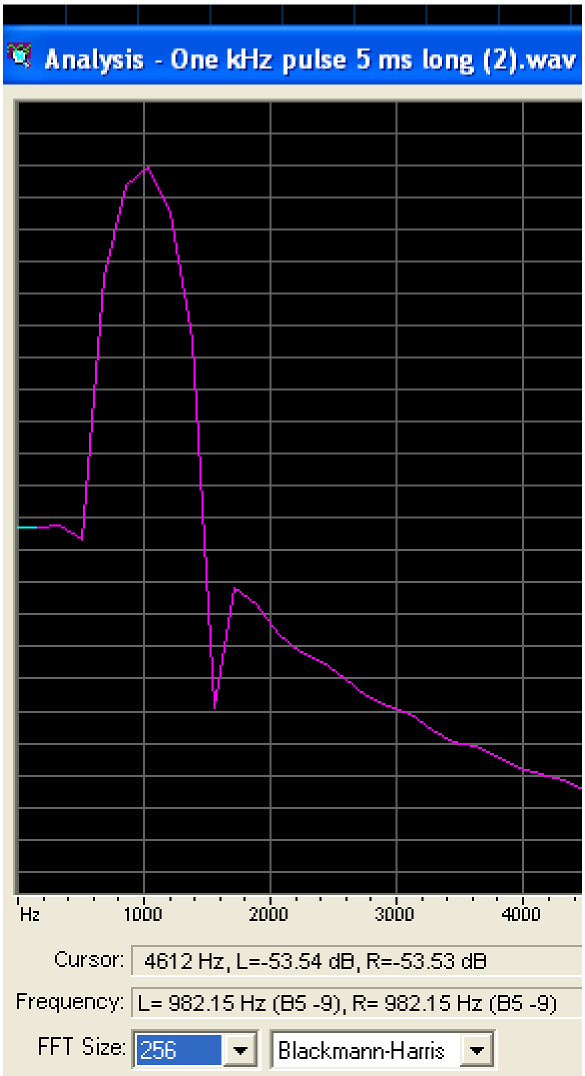
Then

$240/T_e + 22 = F_e$ (near enough)

And where B = bin size.

$F_e = 0.000255B^2 + 0.26B - 4.9$ (near enough)

These equations are presented merely as approximations because they do not relate directly to the FFT methodology, but rather to the FFT as applied to the particular waveform under study. The essential thing is that the frequency accuracy and the pulse duration accuracy are inversely related. Consequently there is no such thing as an optimum setting. Ergo, the FFT cannot, in principle, extract the true underlying nature of even a relatively simple waveform when used as a spectrogram.



Cursor: 4612 Hz, L=-53.54 dB, R=-53.53 dB

Frequency: L= 982.15 Hz (B5 -9), R= 982.15 Hz (B5 -9)

FFT Size: 256 Blackmann-Harris

Figure 14 *The pulse when analysis in the frequency domain is seen as broadband signal.*

The 1 kHz original signal is “smeared” over a wide frequency range. This smearing gives the impression that the original signal was broadband whereas in fact it was rather clean. In Figure 14 each vertical division represents 3 dB. This smearing is worse at smaller bin sizes. However, in general, the peak frequency is correct, and this may be a clue to a way of finding a useful role for the FFT.

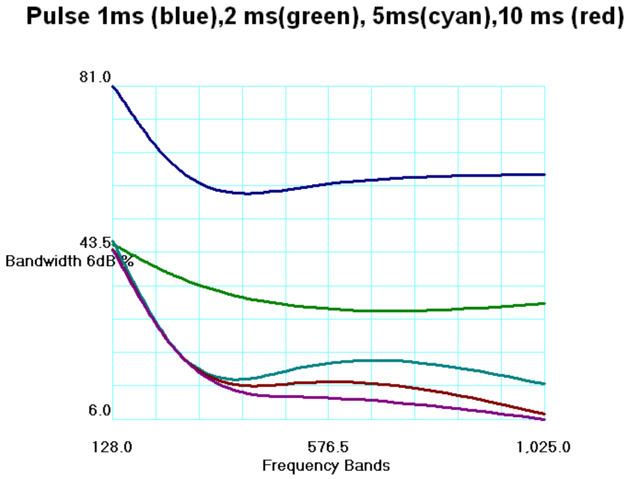


Figure 15 The frequency spread vs. bin size at the 6 dB point.

Of course, the study so far applies only to the 10 ms pulse seen as the red line in Figure 15. Now we consider what happens if the pulse is of different lengths. As the signal is 1 kHz, each millisecond of duration represents a full cycle. The purple line, seen at the bottom of Figure 11 is for a one second pulse, showing that 10 cycles is a reasonable approximation to the continuous signal case.

The single and two cycle pulses (one and two ms) show a very large spread of the spectrum at all bin sizes. What is more the spread is not much affected by the actual choice of bin size.

The Uncertainty Principle

It turns out that this spread is generalized and leads to the Uncertainty Principle which says that the more we concentrate the function $f(x)$ the greater will be the spread of the transform function $f(\xi.)$.

In particular

Suppose $f(x)$ is an integrable and square-integrable function. Without loss of generality, assume that $f(x)$ is normalized:

$$\int_{-\infty}^{\infty} |f(x)|^2\,dx = 1.$$

It follows from the Plancherel theorem that $\hat{f}(\xi)$ is also normalized.

The spread around $x = 0$ may be measured by the *dispersion about zero* (Pinsky 2002) defined by

$$D_0(f) = \int_{-\infty}^{\infty} x^2|f(x)|^2\,dx.$$

In probability terms, this is the second moment of $|f(x)|^2$ about zero.

The Uncertainty Principle states that, if $f(x)$ is absolutely continuous and the functions $x\cdot f(x)$ and $f'(x)$ are square integrable, then

$$D_0(f)D_0(\hat{f}) \geq \frac{1}{16\pi^2} \text{ (Pinsky 2002)}$$

With suitable manipulation this relationship can be seen as an instance of the Heisenberg Uncertainty Principle as the wave function of the position and momentum can be considered as Fourier pairs.

IX. Conclusions and Further Work

The FT (and so the FFT) is fundamentally flawed as a transform to extract the underlying frequency components of signal for two important reasons. Firstly it is not a unique frequency transform of the signal and its original purpose was to find the structure of infinitely long periodic signals. Secondly it will always have the uncertainty principle to contend with that means that any attempt to concentrate the signal will spread the spectrum of the transform and vice-versa.

To truly extract the underlying signal structure we need to abandon the FT/FFT and look to other options such as the LPC transform, wavelet or chirplet transform and possibly, but less promisingly, the De Groot Fourier transform. However none of these is without some problems.

At present we are studying the possibilities and limitations of the LPC transform. That the LPC is a closer approximation to the actual signal can be seen by comparing the FFT representation of the word “one” (Figure 16), which was otherwise determined to be mainly a 200 Hz tone with a small second harmonic component in the first half of the sound, with the LPC version in Figure 17, which clearly shows just the two main components of the call.

The additional “harmonics” in Figure 16 are largely artifacts of the FFT transform and have led to many a wrong interpretation of sound structure.

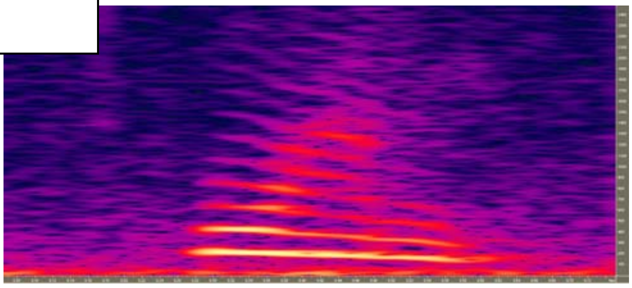


Figure 16 *The FFT representation of the word "one"*

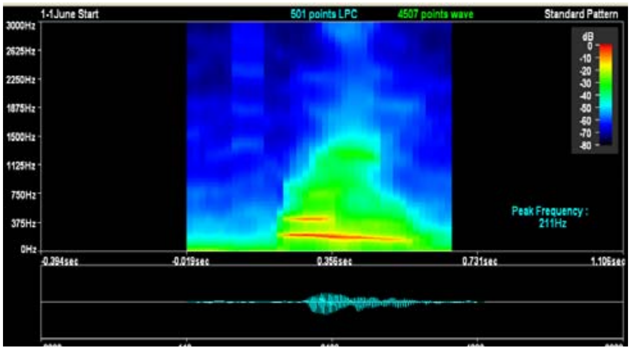


Figure 17 *The LPC representation of the word "one".*

X. References

- Pinsky, Mark (2002), *Introduction to Fourier Analysis and Wavelets*, Brooks/Cole, ISBN 0-534-37660-6
- Morrison, Norman (1994).*Introduction to Fourier Analysis,* John Wiley, ISBN 0-471-01737-X
- Wikipedia. *Fourier Transform*.