A Time-Frequency Analysis of Music using Short-Time Fourier Transforms
Math 596 Project Summary – Spring 2016
Jarod Hart

1 Overview

The relationship between math and music has been well-established and well-explored over the years. Its origins go at least as far as the formulation and study of the wave equation in the early nineteenth century, which also coincides with the origins of Fourier analysis. Over the many, many years since then people continue to use Fourier analysis to understand the frequency content of audio signals, and associated time-frequency analysis to understand hidden structure in music.

The purpose of this project is to introduce some ideas of oscillation, wave superposition, pure-tone signal decomposition, and time-frequency analysis in a math and music context. The primary goal of the project is to implement an algorithm that performs a time-frequency analysis of an input digital signal by computing its short-time Fourier transform. We will use this to visualize the time-frequency content of the signal by creating a spectrogram image, which is the component wise squared magnitude values of the short-time Fourier transform. This way, our final product is a picture of time-frequency space with pixel intensity varying according to the frequency content of the given audio signal at each time. In ideal situations, we can even use this spectrogram to identify or create sheet music to accompany the given audio signal.

2 Mathematical and Programming Content

To complete this project, a background in the following topics is recommended.

- Complex numbers: Some familiarity with complex number is necessary for this project. For the most part, understanding the algebraic properties and some properties complex series is sufficient for this project.

- Calculus: A background in calculus I and II is needed for this project. There are a number of places where integrals, series, and several other topics from calculus come up.

- Linear algebra: Some familiarity with matrix operations and dot products is helpful for this project. However, it does not require a rigorous understanding linear independence, linear combinations, bases, subspaces, diagonalization, etc. In fact with a little extra effort, one can get by with very little knowledge of linear algebra in this project.

- Programming: This project requires a fair amount of programming ability. A thorough understanding of working with vectors/matrices/arrays, decision statements, and loops in essential to implement simulations. Some understanding of computer graphics is also necessary.

3 Primary Resources

For much of the mathematical content listed above, typical text books in the pertinent area are sufficient. Some additional resources on the topics as well as the relationship between math and music are the following (this is by no means a complete list).

4 Mathematical Description of the Project

This project involves viewing the relationship between math and music from two different perspectives. One is a theoretic, continuous time model of an audio signal, which we will denote by \( f(t) \) for \( t \in [0,L] \). The other is a digitally sampled audio signal, which we will denote by \( x = (x(1), \ldots, x(N)) \). We connect these two ideas by thinking of \( x(k) \) as sampled values from \( f(t) \) for sample points \( t_k \) that are computationally convenient; that is, we assume that a signal is truly described by some continuous time function \( f(t) \) for \( t \in [0,L] \), and that a digital signal \( x \) is obtained by sampling \( x(k) = f(t_k) \) for \( k = 1, 2, \ldots, N \) where \( t_k \in [(k-1)\Delta t, k\Delta t] \) and \( \Delta t = L/N \).

Suppose \( f(t) \) is a continuous time audio signal on the interval \([0,L]\). Then \( f \) is decomposed into its frequency content by forming its Fourier series:

\[
f(t) = \frac{a_0}{2} + \sum_{j=1}^{\infty} \left[ a_j \cos \left( \frac{2\pi j t}{L} \right) + b_j \sin \left( \frac{2\pi j t}{L} \right) \right],
\]

where

\[
a_j = \frac{2}{L} \int_{0}^{L} f(t) \cos \left( \frac{2\pi j t}{L} \right) \, dt \quad \text{for} \quad j = 0, 1, 2, \ldots
\]

\[
b_j = \frac{2}{L} \int_{0}^{L} f(t) \sin \left( \frac{2\pi j t}{L} \right) \, dt \quad \text{for} \quad j = 1, 2, \ldots
\]

To dispatch with several technical details, we do not concern ourselves with the sense in which these series converge or the conditions that should be imposed on \( f \) in order for this convergence to hold. These topics have been studied in-depth, and information on these rigorous details can be found in some of the references listed in the Primary Resource section. Using Euler’s formula \( e^{ix} = \cos(x) + i\sin(x) \), one can find the formulas \( \cos(x) = \frac{1}{2}(e^{ix} + e^{-ix}) \) and \( \sin(x) = \frac{1}{2i}(e^{ix} - e^{-ix}) \). Using this, we extend the Fourier series given above to the complex version of the Fourier series

\[
f(t) = \sum_{j=-\infty}^{\infty} c_j e^{2\pi ijt/L}, \quad \text{where} \quad c_j = \frac{1}{L} \int_{0}^{L} f(t) e^{-2\pi ijt/L} \, dt \quad \text{for} \quad j = 0, \pm 1, \pm 2, \ldots
\]

In other words, we can connect the sine/cosine and complex Fourier series via the equations \( c_0 = \frac{a_0}{2} \), \( c_j = \frac{a_j + ib_j}{2} \), and \( c_{-j} = \frac{a_j - ib_j}{2} \).

With the Fourier series written in this way, one can readily interpret \( |a_j|^2 + |b_j|^2 \), or equivalently \( |c_j|^2 + |c_{-j}|^2 \), as the frequency content of \( f(t) \) at frequency corresponding to \( j \). To be a little more precise, if we assume that \( t \) and \( L \) are measured in seconds, then the frequency corresponding to \( j \) is \( j/L \) Hertz (cycles per second). This is our mechanism for measuring the frequency content of a signal, as least for a theoretic, continuous time domain audio signal \( f(t) \) in this situation.

Next we make a connection between the frequency content of the theoretic, continuous time domain signal \( f(t) \) and the digitally sampled signal \( x \). Suppose we have sampled the signal \( f(t) \) at a rate of \( r \) samples per second on an interval \([0,L]\) to obtain a digital signal \( x(k) = f(t_k) \) for some values \( t_k \in [(k-1)\Delta t, k\Delta t] \) for \( k = 1, 2, \ldots, N \), where \( N = r \cdot L \) and \( \Delta t = L/N = 1/r \). Typical sample rates \( r \) for audio signals can be easily found in online documentation. For example, .wav files typically range from sample rates of 4,800 to 176,400 samples per second (4.8 to 176.4 kilobits per second), and .mp3 files range from 8,000 to 320,000 (8 to 320 kilobits per second). In these examples, \( r \) is the reciprocal of the samples per second, i.e., \( r = \frac{1}{4,800}, r = \frac{1}{176,400} \), etc. Recall that the frequency content of \( f \) at frequency \( j/L \) can be measured by computing \( |c_j|^2 + |c_{-j}|^2 \). We connect this notion of frequency to the sampled

signal $x$ by approximating the $c_{\pm j}$ coefficients using Riemann sums, which ultimately leads us to the discrete Fourier transform (hence this approach is a way to motivate/formulate the definition of the discrete Fourier transform). This Riemann sum approach to connecting continuous time and discrete notions of frequency can be found in the SIAM Review paper by Alm and Walker cited in the Primary Resources section. Let $\Delta t = L/N$ and $t_k = (k-1)\Delta t$. Using a Riemann sum approximation and setting $x(k) = f(t_k)$, we have

$$c_j = \frac{1}{L} \int_0^L f(t)e^{-2\pi ijt/N} dt \approx \frac{1}{N} \sum_{k=1}^N x(k)e^{-2\pi ijk/N} = \frac{1}{N} \sum_{k=1}^N x(k)e^{-2\pi i(k-1)n/N} = \tilde{x}(j)$$

as long as $N$ is sufficiently large. The last line here, we take as the definition of the Discrete Fourier Transform $\tilde{x}(j)$ of the signal $x(j)$. Hence from this point forward, for $j \in \mathbb{Z}$, but it $N$-periodic in the sense that $\tilde{x}(j+N) = \tilde{x}(j)$ for all $j \in \mathbb{Z}$. Hence we simply consider the values of $\tilde{x}(j)$ for $j = 1, 2, 3, \ldots, N$. In this way, $\tilde{x}(j)$ is an $N$-vector of complex numbers that carries the frequency information that we need about our signal $f(t)$ and $x(k)$.

There are some technical details here that must be addressed in order to appropriately interpret the frequency information given to us by $\tilde{x}(j)$. Consider the following for integers $j$ in the range $1 \leq j \leq N/2$,

$$\tilde{x}(N-j) = \frac{1}{N} \sum_{k=1}^N x(k)e^{-\frac{2\pi i(N-j)(k-1)}{N}} = \frac{1}{N} \sum_{k=1}^N x(k)e^{2\pi i(k-1)\frac{N-j}{N}} = \frac{1}{N} \sum_{k=1}^N x(k)e^{2\pi i(k-1)} \approx c_{-j}.$$ 

Then for $1 \leq j \leq N/2$, we measure the content of $x$ at a frequency of $j/L$ Hertz by computing

$$|\tilde{x}(j)|^2 + |\tilde{x}(N-j)|^2 \approx |c_j|^2 + |c_{-j}|^2.$$ 

Hence from this point forward, for $j = 1, 2, \ldots, N/2$ we interpret $|\tilde{x}(j)|^2 + |\tilde{x}(N-j)|^2$ as the $j/L$ Hertz frequency content of a signal $x$ that is $N$ samples long sampled at a rate (in samples per second) of $r = L/N$, $L$ being the length of the signal in seconds.

Up to this point, we are able to measure the frequency content of a signal, but there is no sense of time in these measurements; we may know that a frequency is present in a signal, but we don’t know where it occurs in time. To address this, we consider the short-time Fourier transform (or windowed Fourier transform). For the purposes of this project, we will proceed primarily with the discrete version of this theory. Of course, there is a continuous domain analog, but we will forego those details here.

Let $\omega(t)$ be a function on $\mathbb{R}$ supported inside an interval $[-1, 1]$. Further assume that $\omega(t)$ is non-negative and $\omega(t) \approx 1$ for $t$ near 0 (we will leave this imprecise here). A typical example of to use for $\omega(t)$ is

$$\omega(t) = \frac{1}{2} \left( 1 + \cos(\pi t) \right).$$ 

However several otherwise will do. One could also choose $\omega(t)$ to be identically 1, but this would have other drawbacks. In particular, this introduces discontinuities to the function $f(t)$, which causes problems with the frequency decomposition described above; roughly speaking, these discontinuities can interpreted as artificial high frequency content added to the signal, which results in distortion of time-frequency analysis results. Softening the edges of these sub signals with a window function that vanishes at the endpoints of its support helps to dissipate this distortion.

Let $f(t)$ be a signal depending on $t \in [0, L]$ and a window function $\omega(t)$ supported on $[-1, 1]$. The purpose of this function $\omega(t)$ is so that we can window our signal $f(t)$ to form time-localized sub signals for analysis. For $s > 0$ and $\tau \in \mathbb{R}$, define $\omega_s^\tau(t) = \omega(\frac{t-s}{s})$. It follows that $\omega_s^\tau(t)$ is supported inside of $[\tau-s, \tau+s]$ and that $\omega_s^\tau(t) \approx 1$ for $t$ “close to” $\tau$ (here “close to” should be taken to be at the scale of $s$; for example $|1 - \omega_s^\tau(t)| < |t - \tau|^p$ for some power $p > 0$ is a reasonable interpretation in this scenario). Then we have two parameters, $s$ and $\tau$, to control how we window our function to obtain sub signals of the form $\omega_s^\tau(t) \cdot f(t)$.

To discretize this process, fix a parameter $n < N$. As a matter of preprocessing, we pad the signal $x$ with $n/2$ zeros on both the front and the back ends. Call this signal $\tilde{x}$, which is now of length $N + n$. To construct our
discrete short-times Fourier transform, we define a window function defined on \( k = 1, 2, \ldots, n \). To do this, define
\[
w(k) = \omega_{n/2}^{k-n}(k) = \omega(\frac{2k-n}{n})\text{ for } k = 1, 2, 3, \ldots, n.
\] Now for each \( j = 1, 2, 3, \ldots, N \), define the windowed sub signal \( x_j \) of \( x \) by
\[
x_j(k) = w(j) \cdot \tilde{x}(j + k - 1)
\] for \( k = 1, 2, 3, \ldots, n \). Recall \( \tilde{x} \) is of length \( N + n \), and so \( x_j \) is well defined for all \( k = 1, 2, 3, \ldots, n \). Finally for \( k = 1, 2, 3, \ldots, N \) and \( j = 1, 2, 3, \ldots, n \), define the short-times Fourier transform \( S(j, k) = \tilde{x}_j(k) \).

If we pause for a moment to assess what we have constructed, we will see how the time-frequency analysis is coming into the picture and how to proceed with the appropriate inferences about frequency content. Since \( x \) is a vector of length \( n \), it follows that \( S(j, \cdot) = \tilde{x}_j \) is a vector of length \( n \). In particular, \( A(j, k) = |S(j, k)|^2 + |S(j, n-k)|^2 \) is the frequency content of \( x \) at scale \( k \) for \( 1 \leq k \leq n/2 \). Finally, to construct the spectrogram, make an \( N \times n/2 \) grayscale image with pixel intensity \( A(j, k) \) for \( j = 1, 2, 3, \ldots, N \) and \( k = 1, 2, 3, \ldots, n/2 \). This creates a picture, called a spectrogram, with frequency on the vertical axis, times on the horizontal axis, and time-localized frequency content indicated by pixel intensity.

Finally, we give a description of how to track the precise frequency ranges within the spectrogram. Suppose we are given a signal of length \( N \) that is sampled at rate \( r \) samples per second and a sub signal length \( n < N \). Construct the window function \( w \), the sub signals \( x_j \) for \( j = 1, 2, \ldots, N \), and the amplitudes \( A(j, k) \) as above. It follows that the original signal is \( L = N/r \) seconds long, and that each sub signal \( x_j \) is \( \ell = n/r \) seconds long. Then for \( j = 1, 2, 3, \ldots, n/2 \), the amplitude \( A(j, k) \) is a measurement of the amount of \( j/\ell = r \cdot j/n \) Hertz frequency localized at time \( t_k = (k-1)\Delta t = (k-1)L/N \) for \( k = 1, 2, 3, \ldots, N \).

5 Summary of Results

In order to demonstrate the action of the Discrete Fourier Transform and the spectrogram, first consider the pulse signal and the component-wise magnitude of its Discrete Fourier Transform pictured below.

On the left is a signal that has a high frequency pulse localized around \( t = 200 \) and a low frequency pulse localized around \( t = 800 \). On the right is a plot of the absolute value of the Discrete Fourier Transform of the pulse signal. It is easy to observe the two spikes in the DFT of the pulse signal corresponding to the two pulses in the original signal with different frequencies. The location of the spikes on the horizontal access correspond to the frequencies of the corresponding pulse; the left spike is the low frequency pulse, and the right spike is the high frequency pulse. Since the high frequency pulse persists for a longer time in the original signal and the amplitude of the pulses are similar, there is “more” high frequency information in the signal, and the high frequency spike is higher. As can be seen with these two graphs, one can identify the frequencies that are present in the signal, but no information about time localization is provided.

We can use the spectrogram to identify time-localized frequency content. Below is a picture of the pulse signal put on top of its spectrogram.
In the spectrogram, the horizontal axis represents time and the vertical axis represents frequency. White pixels represent higher intensity, and black pixels represent lower intensity. One can think of vertical slices of the spectrogram as the component-wise magnitude of the Discrete Fourier Transform of the windowed signal and can be interpreted much like that two-pulse example above. The spectrogram initially gives the vertical axis a reciprocal scale. We reversed the orientation of the vertical axis in the picture above to provide a picture where high/low frequency can be interpreted in the natural way (with high frequencies towards the top and low frequencies towards the bottom). The picture above was also modified to zoom in around the significant frequency content. As a result the meaning of the scale on the vertical axis was lost; so do not interpret the scale on the vertical axis as physical frequency in any way. It is now easy to identify the time localized behavior of the two pulses in the signal. This provides a sort of proof of concept of the spectrogram. Now we are ready to discuss some more complicated signals.

Below we show the first few seconds of the Happy Birthday Song and its spectrogram.

Here we can identify the notes of the song by the areas of high pixel intensity. This spectrogram is relatively low resolution, so it is a little difficult to identify exactly the musical properties here, but looking closely one can identify fundamental frequencies and overtones corresponding to the notes of the song. The sheet music can also be associated to the time-frequency profile given by the spectrogram. For example, the boxed region in the spectrogram corresponds to the boxed region on the sheet music. It is not hard to follow along on the rest of the spectrogram corresponding to the sheet music. This sheet music was downloaded from [http://www.8notes.com/](http://www.8notes.com/), a website that provides free sheet music and lessons for several musical instruments. Modifications to the spectrogram were made, just like the case of the two-pulse spectrogram. So the meaning of the numbers on the vertical axis cannot be directly interpreted as physical frequency values. With more care, these attributes can be determined by using the method for computing the frequency in Hertz described in the last paragraph of the previous section.
The following is a spectrogram of approximately the first 15 seconds of the Moonlight Sonata by Ludwig van Beethoven. The audio clip was downloaded from [archive.org](http://archive.org) and is also available at the [author's webpage](http://author's webpage).

6 Possible Extensions

One possible extension of this analysis is using wavelet analysis in place of short-time Fourier analysis. The closest analog to the work in this project is most closely related to continuous wavelet transforms. One could also use discrete wavelet transforms to do a similar analysis. Another direction of research that could be pursued is to use the spectrogram to develop an algorithm that automatically identifies the attack of a note, its fundamental frequency, and determines the duration of the note. In this way, one can automate an algorithm to extract sheet music for a given digital audio signal.

7 Note From the Author

This is a student project from the Math and Biomedical Research course, taught by the current author Jarod Hart, offered at the University of Kansas in the Spring of 2016. Some modification and additions were made to the original project for this summary. The course is supported by the Initiative for Maximizing Student Development (IMSD) through an NIH grant NIH-NIGMS 5R25GM062232. The PIs of this IMSD grant are Professors Estela Gavosto (Mathematics Department) and James Orr (Biology Department). We are happy to share these project ideas, and welcome those who are interested to use them. We'd love to hear about your results and extensions related to these projects, and in some cases, will provide some support for the projects. Please contact Jarod Hart at jvhart@ku.edu with any typos, errors, questions, or comments about this project summary.