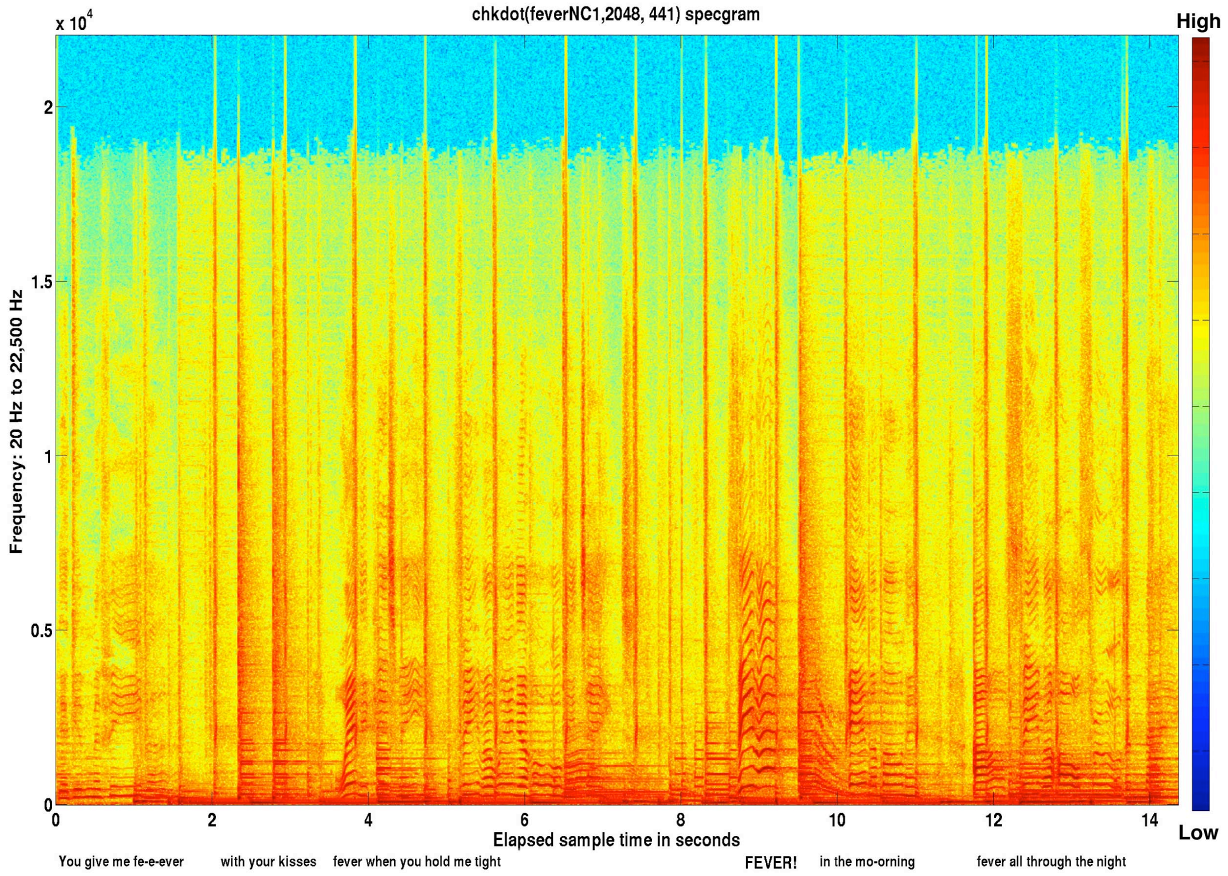
NEW SIGNAL PROCESSING TECHNIQUES FOR IMPROVED INFORMATION EXTRACTION FROM MUSIC AND AUDIO DATA



Ken Lindsay
Information Scientist
ken@tlafx.com
(650) 520-4536 ©, (541) 552-1509 (h)

Introduction

The purpose of this report is to convey new (and newly rediscovered) signal processing techniques which I believe are imminently practical in the field of speech recognition, music content analysis and search, and applications to other feature recognition tasks in one dimensional data sets. I also include some forward looking ideas which have not yet been implemented as far as I know, and if discussed in the literature, such talk is minimal.

Given the large discrepancy between human audio processing capabilities and the abilities of computers, new techniques for software should produce substantial advances in automated audio cognition. In my recent thesis research I analyzed the nature of swing rhythm in music. The image on the title page is a spectrogram of Natalie Cole singing *Fever* in the 2004 recording by Ray Charles. Useful details of both rhythm and vocal are clearly visible. In many other cases, the complexity of the musical clip was beyond the practical analysis limits of the standard STFT spectrogram approach, and I wished for, conceptualized and discussed at some length, my ideas for improving Fourier analysis. My advisor and others were skeptical, at least until I started to uncover prior work similar to my ideas.

Improvements to Standard FFT Usage

Fulop and Fitz (2006) published a new(ly rediscovered) approach to the spectrogram which utilizes the phase information in the complex form of the standard FFT to reassign time and frequency results to new locations, and so extract better details. Their results are an impressive improvement on the traditional approach, and computationally cheap, being based on finite difference methods applied to the FFT results. They give details about three forms of the algorithmic approach, all of which are basically equivalent in terms of compute cost and quality of results. Oddly, they do not go beyond first order finite differences. I believe their approach is just a beginning, and that a variety of useful heuristics can be found which enhance the results for particular applications, e.g. choosing FFT length, windowing functions, and overlap (for STFT) that tease out features in particular frequency ranges, due to "artifacts" caused by the inherent limits present in the parameter choices – artifacts which might be theoretically offensive, but may be very useful in the real world. Step size(s) choice or using higher orders in the finite difference scheme may also reveal useful tricks.

Fulop & Fitz give a good summary of the history of the reassigned spectrogram, going back some fifty years, although the main breakthrough came in the 1970's. They also touch on other practical considerations such as the time/frequency information resolution tradeoff limits that are analogous to the Heisenberg uncertainty principle. I am curious what aspects of some of my other ideas can be uncovered in prior work. One of my main complaints in my research was the waste of throwing away all phase information in the FFT, so I find the Fulop & Fitz technique very appealing.

Alternatives to Harmonic Fourier Series

Traditional Fourier Analysis is based on using sines and cosines whose frequencies are related by the set of integers, the so-called harmonics, to generate a set of basis functions for estimating objects in the function space of interest. Of course, most spaces have infinitely many sets of basis functions, and Fourier series are only one of many possible choices of basis sets for audio work. From the common talk however, one would think that not only is harmonic analysis the *only* choice, but also that it is somehow true and accurate despite obvious limitations. The audio space that sources the stream we listen to, or analyze in the computer, is more complicated than the model used in standard Fourier Analysis processing. At its most basic, sound is a 4 dimensional system – 3D plus time. There are also important effects in human audio perception such as echoes and phase shifts from the shape of the pinna (outer ear), binaural hearing, and movements of the head, all of which add dimensionality to the analysis. This complexity is an important aspect of hearing, clear to anyone who uses their ears in a critical, self aware manner. While theoretical constructs from math, DSP, information theory or experimental audiology research are useful tools for analyzing this information, I think it is important to go with the primary experience rather than to let abstractions unduly color the understanding of sound.

The FFT has a very attractive aspect – efficiency of the Cooley-Tukey algorithm. This is also known as the Danielson-Lancszos algorithm, or Runge-Konig algorithm. This algorithm has also been rediscovered several times. Press et al. (2002) trace the first description of this approach to Gauss in 1805. The efficiency gain is well known. What is less well known is the mechanics of the approach. This involves factoring the commonalities of the complex exponential functions that are duals of the sines and cosines of the real valued form. I've discussed my idea with my math advisors and got the "Sure, sounds like that should work if you can figure it out" response. Given a sparse, non linear spaced set of integers to determine the basis frequencies, the trick would be to factor the common exponential coefficients that combine to determine the various frequencies as is done in Cooley-Tukey. Such a technique may already exist, buried in deep theoretical math articles. A quadratic frequency distribution would be the first obvious approach. This could speed up an already quick algorithm by omitting calculation of redundant frequencies (basis vectors).

The primary benefits of a good non-harmonic Fourier scheme are efficiency and precision of detail. The standard decomposition of an audio signal by linearly spaced harmonics causes 3/4 of the frequencies extracted to be in the 5000 Hz and up range. Most of the important information for tasks like pitch detection and speech recognition are below 5 KHz. Omitting many of the high frequency (redundant) harmonics would speed up the algorithm many times, if the same trick of factoring complex exponentials could be used.

High resolution in the low frequencies of a Fourier decomposition requires longer FFTs than low resolution. Adequate resolution for tracking a melody would require lengthy FFT windows that dilute the signal strength as it is spread over too much time, blurring frequency resolution, but in a much different way from commonly known FFT window resolution problems. A non-harmonic scheme would map the high frequencies to far fewer basis functions, and still maintain sufficient resolution to recognize high frequency features. Whether non-harmonic analysis would also help overcome the time/frequency limits in the lower frequencies is less clear, since these are primarily based on the window length of the FFT. But no doubt other useful heuristics can and will be developed, e.g. low pass filtering used with spline based approximation such as Chebyshev polynomials, or estimation of low frequencies by counting zero crossings.

The use of non-harmonically related Fourier series in control theory goes back to the 1930's and 1940's by Bellman and others. This technique is used for generating the waveforms needed rather than analyzing an unknown signal, which is our current interest. Wavelet analysis is a related construct that can be useful *if* you can find the appropriate wavelet family, which I found difficult in practice. Moreover, it is likely than many different wavelets would be needed to solve the general audio parsing problem. Last year I explored empirically in MATLAB the topic of both harmonic and non-harmonic Fourier analysis with interesting, though fairly obvious, results. Today, Google favored me better than last year, and I found references to work being done in the non-harmonic Fourier *analysis* side of things. These seem to be entirely in the theoretical math domain. Whether there is a non-harmonic version of the Cooley-Tukey algorithm remains to be seen.

Practical Software Applications

My work to date has shown simple and practical techniques for characterizing rhythm in musical recordings, without relying on meta information like sheet music, MIDI files, or explicit analysis by humans. This sort of approach is critical for fully automating music search. My thesis research involved a lot of hand work, but extending this to automated analysis is mostly a question of developing a catalog of information features for search, rather than any great fundamental technical breakthrough. One major reason for this is the relative simplicity of searching for percussion sounds in musical recordings. These musical events are typified by strong and sudden onsets, and these onset events are easily mapped to time locations which makes extraction of rhythm quite straightforward. While I chose simple audio mixes for quick and easy analysis, the same techniques can be applied to more complex musical samples, making general rhythm extraction and comparison practical. I am sure there will be some situations where the simple approach proves inadequate, but more sophisticated techniques such as Blind Signal Separation and Independent Components Analysis should extend the practical limits in many cases.

There are numerous other cases in music where more advanced techniques would be needed. Such work can be found in the literature, but it typically has a disconnected academic quality, and does not appear ready for professional real world use. To take an example, an original recording of a Beatles song and an *elevator music* version of the same tune would match in many ways: same key, same tempo, same chord progression, same harmonic structure etc. A human would not be fooled, and returning *The 101 Strings* version of *Yesterday* would probably just annoy a user and cause them to use a different search engine. Clearly better techniques are needed for the general task of music search.

Here is an example of a more difficult analysis task. Figure 1 is a spectrogram of Bob Marley singing *Stir it up*, "quench me darling, when I'm thirsty." Contrasted with Natalie Cole's strong and punchy vocal image in *Fever,* Marley sings in a plaintive sensitive style. The interleaving of the vocal frequencies is visually clear, and undoubtedly, the presence of emotional content in vocal music can be traced to features like these. A practical feature extraction and recognition scheme would allow music searching for emotional similarity in music without relying on meta information explicitly entered by human experts. Similarly, vocal harmonies and complex instrumentation could be analyzed for search without relying on human supplied meta information. In this example, the percussive rhythmic note events stand out clearly, although they are sometimes masked by the vocal signal.

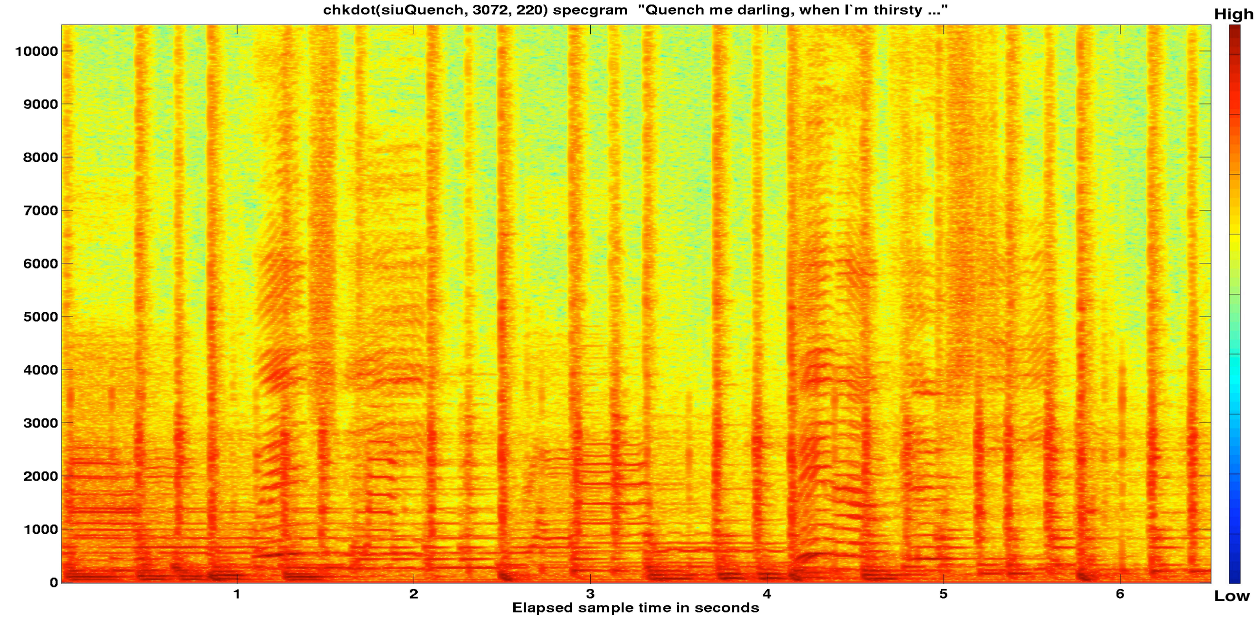


Figure 1. Bob Marley sings *Stir it up* (1974)

New Paradigms Based on Human Auditory Perception

The higher frequency range of the human ear uses hardware (or wetware) which is fairly similar to Fourier analysis – the frequency sensing cells in the cochlea respond to input signals and discriminate frequency content something like Fourier series. However, the low frequency response of the hearing system is quite different from Fourier analysis. Instead the basilar membrane flexes into frequency specific shapes in response to incoming signals, and the nerve pathways pick up displacement dynamics and delay information in order to extract very fine resolution in the 100 to 2000 Hz range. The 2000 Hz cutoff is due to limits of neuron recovery time, which would not be a problem in a computer algorithm. Below 200 Hz there is an additional mechanism based on detecting "beats" between component frequencies, although some researchers discount this third mechanism. Nonetheless, it is useful to learn from Nature to devise methods which can help bypass the inherent time/frequency limits of standard windowed FFTs.

Conclusions

Music analysis for searching will soon develop to use a Google like interface where we copy and paste a few sample audio clips into a search field, and quickly see a list of songs which have characteristics like those present in the samples. There are numerous players in music search technology, but they typically rely heavily on human expertise and manual labor in cataloging and categorizing musical pieces. If it can be made practical, an automated approach would be better. I believe it *will* be practical within a few years, and am keen on being part of that development process. My interests are not academic. I am very much a pragmatic developer and researcher. While I am glad that there are highly skilled theoreticians in the world, my interests lie in creating practical applications which will be used extensively by real people to help resolve their real world needs. Similarly, speech recognition is quite impressive as it stands now, but clearly falls far short of even the most basic of human skills. Developing and using better DSP and pattern recognition techniques are crucial for bringing these tasks into the 21st century.

References

Fulop, Sean. A. & Fitz, Kelly (2006). ***Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications.*** Journal of the Acoustical Society of America, 119:1 pp. 360 - 370. January 2006.

Press, William H., Teukolsky, Saul A., Vetterling, William T., & Flannery, Brian P. (2002) ***Numerical Recipes in c++: The Art of Scientific Computing, 2nd edition.*** Cambridge University Press. Cambridge, UK.

Young, Robert M. ***An Introduction to Nonharmonic Fourier Series, revised 1st ed.*** Academic Press. San Diego, San Francisco. 2001.