AUTOMATIC CHARACTERIZATION OF DIGITAL MUSIC FOR RHYTHMIC AUDITORY STIMULATION

Eric Humphrey
Music Engineering Technology Group
University of Miami
Coral Gables, FL 33124
humphrey.eric@gmail.com

ABSTRACT

A computational rhythm analysis system is proposed to characterize the suitability of musical recordings for rhythmic auditory stimulation, a neurologic music therapy technique that uses rhythm to entrain periodic physical motion. Current applications of RAS are limited by the general inability to take advantage of the enormous amount of digital music that exists today. The system aims to identify motor-rhythmic music for the entrainment of neuromuscular activity for rehabilitation and exercise, motivating the concept of musical "use-genres." This work builds upon prior research in meter and tempo analysis to establish a representation of rhythm chroma and alternatively describe beat spectra.

1. INTRODUCTION

Digital multimedia is now an integral, and somewhat inescapable, aspect of modern life. Personal handheld devices are designed to streamline the acquisition, management and playback of large volumes of content as cutting-edge computing devices approach ubiquity. This trend, in tandem with the commercial success of devices like the iPod and iPhone, has encouraged an environment where both content providers and end-consumers have access to enormous digital music collections. As a result, individuals are consuming and purveying more music than ever before and this realization introduces the classic logistical issue of content navigation; when a library becomes sufficiently large, more complex paradigms must be developed to facilitate the searching, indexing, and retrieval of its items.

Conventional music library systems employ metadata to organize the content maintained within them, but are typically limited to circumstantial information regarding each music track – such as the artist's name or the year it was produced – in addition to the somewhat amorphous attribute of genre. Understandably, stronger information concerning the specific nature of a track allows for more insightful and context-driven organizations or queries of a library.

The need for content-specific metadata introduces the challenge that someone, or something, must extract the relevant information necessary. One approach, like the one taken by the Music Genome Project, is to manually annotate a predetermined set of attributes by a diligent group of human listeners, a scheme with clear benefits and drawbacks. While this method is substantiated by the observation that no computational system has yet matched itsreliability, it simply takes a human listener far too much time to parse music. As an example, it would require about 68 years to listen to every track currently available in the iTunes Store, [1] which now contains some 12 million tracks.

Needless to say, the development of computational algorithms to extract meaningful information from digital music provides the ability to process content as fast as an implementing machine can manage. Many efforts over the last twenty years proceed to these ends in varying levels of scope and success. As mentioned however, no single solution has been able to rival the performance and versatility of even moderately skilled human listeners. It has been proposed previously that, in this period of continued research toward improved machine-listening technologies, algorithms are likely to perform best when developed for a specific application.

It is in this spirit that a computational system is proposed to characterize the suitability of musical recordings for rhythmic auditory stimulation, a neurologic music therapy technique that uses rhythm to entrain periodic physical motion. The remainder of the paper is structured as follows: Section II addresses the background of motor-rhythmic music as a use-genre and the physiological motivations; Section III briefly reviews relevant computational models of human rhythm perception and details the proposed system; Section IV explores the evaluation and visualization of the algorithm results; and Section V discusses the system behavior, observations, and directions of future work.

[1] With an average track duration of 3 minutes.

2. BACKGROUND

Music and motion share a long and intertwined relationship throughout human history. Dance comprised an integral role in many ancient civilizations for spiritual and social purposes and work song served to synchronize the physical labor of crews, as was common on sea-faring vessels. In modern times, physical exercise is often tightly coupled with music, ranging from joggers with personal media players to fitness classes.

Many individuals empirically find that music facilitates exercise, and recent advances in music therapy and neuroscience give this notion credence. Through an increased understanding of the underlying mechanisms involved in a human's physiological response to music, current knowledge supports the position that rhythm serves as a powerful external timing mechanism capable of entraining gait parameters and neuromuscular activity [1]. Building upon this principle, rhythmic auditory stimulation (RAS) is "a neurological technique using the physiological effects of auditory rhythm on the motor system to improve the control of movement in rehabilitation and therapy" [2].

The impact of rhythmic auditory stimuli on movement can be summarized as three primary components. Sensory motor control provides priming and timing cues to individual in guiding a motor response. Motor programs are thought to be developed in the brain to control complex motor movement, where rhythmic stimuli encourage the creation of more efficient and fluid programs for cyclical movement. Also, RAS supports goal-directed movement where motion is cued by anticipation, a key musical element, rather than by explicit events like heel strikes.

Appropriate music to achieve RAS, best described as *motor-rhythmic*, must exhibit certain criteria: a strong beat-percept, regular meter, little to no tempo deviation, and maintain a tempo that encourages the desired entrainment frequency, referred to in the literature as an individual's resonant frequency or limit cycle. The ability to succinctly describe a class of musical content for a specific application motivates its distinction as a use-genre.

A fundamental problem faced in RAS-based research and applications is the inability to harness the abundance of available digital music as external entrainment stimuli, as no solution exists to characterize music for this purpose. It is for this reason that nearly all uses of RAS are confined to closely-monitored clinical settings that heavily rely on human supervision to provide, and sometimes compose, appropriate motor-rhythmic music. An automated system would not only facilitate the practice of RAS as a clinical rehabilitation technique, but also allow the integration of RAS methodologies on a significantly broader scale, such as exercise classes or personal fitness technologies.

Some previous systems attempt to link the rhythm, and more specifically the tempo, of music and physical motion in the form of running [3]. Each effort, however, incorporates the assumption that all content is accurately and sufficiently described by a single tempo value. Quickly considering the great diversity of musical content available, it is intuitive to conclude that this is inadequate. With these goals in mind, we seek to develop a system capable of quantifying the motor-rhythmic attributes of digital music content for use in applications of RAS.

3. PROPOSED SYSTEM

Computational rhythm analysis algorithms for digital music recordings have been extensively researched over the last twenty years. Early systems were developed to perform tempo extraction of individual tracks and excerpts to ascertain a single tempo value, and beat tracking to annotate the location of musical pulses in a recording, both achieving notable success. More recent efforts aim to improve upon these results by employing alternate mechanisms to fulfill various system tasks or seek to determine further information, such as meter [4] and beat spectrum [5]. A more thorough review of recent leading systems is provided in [6].

Being that human rhythm analysis remains the best performing system, explicit modeling of the human auditory system would appear to be a viable approach toward the development of a machine-listening algorithm for rhythmic analysis. By reducing the task of rhythm perception to the functional components of the overall biological process, each stage can be approximated computationally. At the most rudimentary level, human rhythm perception is achieved in a two-stage process of event perception and periodicity estimation.

The idea of determining meaningful events in music perception is admittedly a loaded topic. However, a semantic debate can be mostly avoided in considering that there are arguably three orthogonal dimensions in basic music perception: rhythmic, tonal and timbral. In the context of characterizing the suitability of music for RAS, the focus of meaningful events can – and should – be constrained primarily to rhythmic, or energy-based, events. Neglecting the other two dimensions serves to emphasize the importance of rhythmic content.

Periodicity estimation can be computationally achieved in a variety of different manners depending on performance concerns, such as causality and complexity. One common school of thought regarding human beat induction claims that the phenomena of felt-beat it is achieved through the resonating, or entrainment, of oscillator banks in the brain as an interval-period based process [2]. This is a particularly attractive option given the correlation between the oscillations of the human body as a dynamic mechanical system during movement and those of a mathematical model.

Coincidentally, these are essentially the main system components presented by Scheirer in [7] and Klapuri et al in [4]. Building upon the work outlined therein, the proposed system proceeds in the following manner: an input signal is first decomposed into twenty-two subband components via a maximally-decimated filterbank closely approximating the critical bands of the cochlea and rhythmic events are derived for each. These onset events are reduced to a single stream of pulses and periodicity estimation is performed using a bank of modified comb-filter oscillators. The resulting beat spectra is transformed into

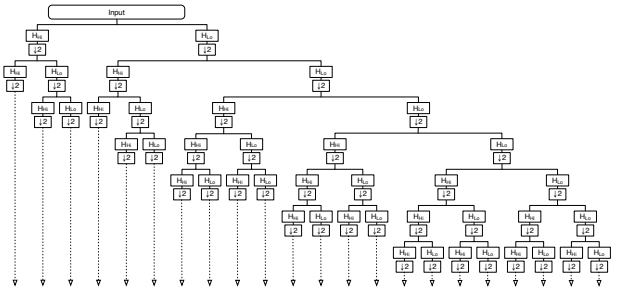

Figure 1. A perceptually-motivated dyadic filterbank for the decomposition of an input audio waveform.

Band	Range (Hz)	Band	Range (Hz)
1	0 – 125	12	1750 – 2000
2	125 – 250	13	2000 – 2500
3	250 – 375	14	2500 – 3000
4	375 – 500	15	3000 – 3500
5	500 – 625	16	3500 – 4000
6	625 – 750	17	4000 – 5000
7	750 – 875	18	5000 – 6000
8	875 – 1000	19	6000 – 8000
9	1000 – 1250	20	8000 – 10000
10	1250 – 1500	21	10000 – 12000
11	1500 – 1750	22	12000 – 16000

Table 1. Frequency ranges for the resulting subband components.

rhythm chroma over time, from which global features are calculated to compactly describe the entirety of a music track.

3.1 Cochlear Modeling

At this point in time, it is commonly held that the human auditory system is reasonably understood so far as the point where electrical signals are encoded and transmitted to the brain via the auditory nerve. Most stages prior to neural processing though, such as diffraction of the pinnae or dynamic compression from the bones of the inner ear, are not overly integral to the perception of rhythm. However, the cochlea does perform a coarse frequency decomposition as transduction occurs across the critical bands of the organ. Scheirer observed that the perception of rhythm is maintained when amplitude modulating white noise with the envelopes of as few as four subbands of an audio waveform [7]. Therefore, it is proposed that monitoring the fluctuation of energy in each critical band serves as a reasonable approximation of preconscious observation of meaningful rhythmic events.

Motivated in part by the system developed by Tzanetakis et al [8], a multi-resolution time-domain filterbank is used to decompose an input waveform into twenty-two subbands. Whereas wavelet processing implements complimentary half-band filters and a true pyramidal structure, the filterbank divides frequency content similarly to the cochlea, the ranges of which are listed in Table 1 and diagramed in Figure 1.

It is important to note that, given the cascaded nature of the structure, non-linear phase distortion introduced by IIR filters is unacceptable and errors will propagate differently in each band. This is particularly troublesome in the context of a system developed to analyze the temporal relationship between events. Therefore, half-band FIR filters of Daubechies' coefficients are chosen, and appropriate all-pass filters are designed to flatten the group delay at each successive level to ensure alignment of the resulting subband components. The accumulative delay and complexity of the filterbank decomposition is mainly dependent on the length of the Daubechies' filter shape selected (N = 40 in our experiments), though the impact of using different filter lengths on performance has yet to be explored.

3.2 Rhythm Event Detection

Following decomposition, each subband signal is processed identically to identify rhythm event candidates. Consistent with [7] and [4], subband envelopes are calculated by half-wave rectifying and low-pass filtering each subband waveform with a half-Hanning window, defined by Equations 1 and 2.

$$X_{HWR_k}[n] = \max(X_k[n], 0) \tag{1}$$

$$E_k[n] = \sum_{i=0}^{N_k-1} X_{HWR_k}[n] * W_k[i-n] \tag{2}$$

Subband envelopes are then uniformly down-sampled to 250 Hz, influenced by the temporal resolution of the human auditory system, and compression is applied to the resulting signals according to Equation 3. Event candidates are calculated by filtering the subband envelopes with the Canny operator defined in Equation 4, commonly used in digital image processing for edge detection and first applied to audio processing in [9]. The frequency response of the Canny operator is more desirable than that of a first-order differentiator, being band-limited in nature and serving to attenuate high-frequency content.

$$E_{C_k}[n] = \frac{\log_{10}(1+\mu * E_k[n])}{\log_{10}(1+\mu)} \tag{3}$$

$$C[n] = \frac{-n}{\sigma^2}\exp(-n^{\frac{2}{2\sigma^2}}), \quad \text{where } n = [-L, L] \tag{4}$$

At this stage, event candidates effectively represent the activation potential of their respective critical bands in the cochlea. Though there are multiple hair cell transduction theories concerning the significance of place and rate on pitch perception, the fact remains that temporal masking is caused by the necessary restoration time inherent to the chemical reaction associated with neural encoding. Known as the precedence effect, sounds occurring within a 50 millisecond window–about 10 milliseconds before and 40 milliseconds behind–are perceived as a single event. This phenomena is modeled by a sliding window to eliminate imperceptible or unlikely event candidates.

Rhythm event detection concludes with the summation of subband events to a single train of pulses and a zero-order hold to reduce the effective frequency of the pulses.

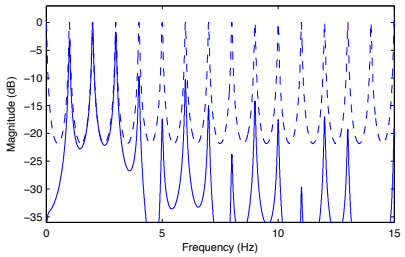

Figure 2. Magnitude response of a typical comb-filter (dashed line) and cascaded with a Canny filter (solid line).

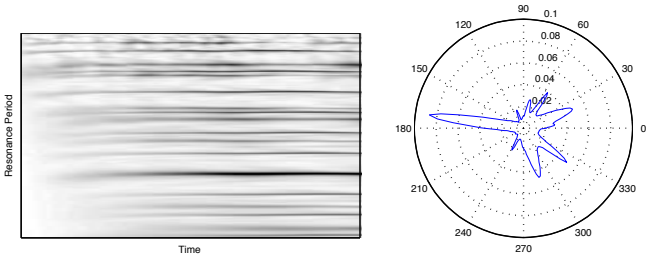

Figure 3. Example of a tempogram and chroma for *bonus5.wav*, from the MIREX practice data set.

A single-sample pulse is the half-wave rectified counterpart to a single period of the highest frequency that can be represented by the current sampling rate. Rhythmic frequency content, such as the tactus or felt-beat, typically exists on the range of .25–4 Hz (or 30–240 BPM), with tatum and metrical levels falling just above and below that range, respectively. Therefore, a zero-order hold of 50 ms is applied to band-limit the signal, constraining frequency content to 20Hz while maintaining the temporal accuracy necessary.

3.3 Periodicity Estimation

In continuing with modeling preconscious rhythm audition, periodicity estimation is performed using a set of tuned comb-filters spanning the frequency range of interest. This method was pioneered as a computational model of rhythm induction by Scheirer in [7], and has since been incorporated in a variety of derivative works due to reliability and modest computational complexity. Importantly, modifications are introduced here to improve performance and tailor the model to better suit the target application.

Unlike previous systems that aim to set a constant resonance half-life across each oscillator, we propose that perceived resonance of a pulse train is dependent not on time but the number of pulses observed. It seems intuitive that a 40 BPM click track at 40BPM should take longer to perceive at the same strength as one at 180 BPM. Though a more perceptually-motivated method may better capture this nuance, the value of α is set at 0.825 to require a period of regularity before resonating, while maintaining the capacity to track modulated tempi.

Beat spectra is computed over time for each delay lag T, as defined by the comb-filter difference equation in Equation 5, varied linearly from 50–500 samples, inversely spanning the range of 30–300 BPM. Each comb-filter is also cascaded with a band-pass filter – the Canny operator – to augment the frequency response of the periodicity estimation stage. As shown in Figure 2, this attenuates the steady-state behavior of the comb-filter effectively lowering the noise floor, while additionally suppressing resonance of frequency content in the range of pitch perception over 20Hz. The Canny filter is also corrected by a scalar multiplier to achieve a passband gain of 0 dB.

$$y_k[n] = (1-\alpha) * x[n] + \alpha * y_k[n - T_k] \qquad (5)$$

Instantaneous tempo is calculated by low-pass filtering the energies of each oscillator over time. Scheirer previously described this process of determining the energy in the delay line over the length of the resonance period, and is analogous to computing an unweighted-average. A Hanning window of length W_k, set corresponding to the delay lag of its respective comb-filter channel and given in Equation 6, serves as an estimation of "resonance memory." This time-frequency representation is referred to as a *tempogram* and estimates perceived tempo strength over time, an example of which is shown in Figure 3.

$$R_k[n] = \frac{1}{W_k} \sum_{i=0}^{T_k-1} w_k[i] * (y_k[n-i])^2 \qquad (6)$$

3.4 Chroma Transformation

As observed by Kurth et al [5], the duality of pitch and rhythm allows the representation of beat spectra in terms of chroma. In the same way that all pitches can be described as having a height and class, various metrical levels exhibit a similar relationship. Octave errors, a typical issue faced in tempo extraction, are mitigated by eliminating the subjective aspect of rhythm and reducing the task to a purely objective one. Fundamental tempo class is especially important to RAS-applications, and is the ultimate focus of the system.

Rhythm chroma is computed by first transforming beat spectra to a function of frequency, rather than period, scaled by the base-2 logarithm and referenced to 30 BPM. Three tempo octaves (30–60, 60–120, and 120–240 BPM) are collapsed by summing beat spectra with identical chroma, as detailed in Equation 7. Understanding this representation is facilitated by plotting amplitude as a function of $\log_2$ tempo class in the polar coordinate system, shown in Figure 3, such that the harmonic structure of a given input becomes readily apparent.

For clarity, rhythm chroma consists of a radial amplitude and an angular frequency, referred to as a class and measured in units of degrees or radians. The transformation from tempo, in BPM, to class, in normalized radians, is defined by Equation 8. This is a many-to-one mapping, and is not singularly invertible. Visualizing rhythm chroma in this alternative manner allows for deeper insight into the nature of musical content and the extraction of novel features, and will be discussed in greater detail shortly.

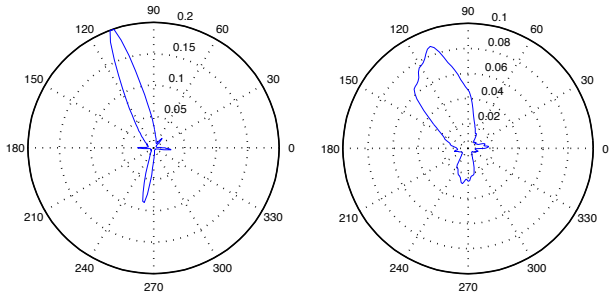

Figure 4. Chroma diagrams for a 148 BPM click track, before and after tempo automation. Note the difference in scale and amplitude of the fundamental.

$$\Psi_n[\omega] = \frac{1}{L}\sum_{i=0}^{L-1} R_n[\omega + 2\pi * k] \tag{7}$$

$$\omega_{class} = \log_2 \frac{BPM}{BPM_{reference}} \tag{8}$$

3.5 Feature Vector Representation

A single rhythm chroma is obtained for a track by summing over time and normalizing by the length. Several key features of interest are emphasized by producing a global chroma, though this set presented is not intended to be exhaustive by any means. Beat strength is effectively described by the amplitude of the largest lobe, and fundamental tempo class is given by the angle of this peak. Other lobes are actually subharmonics of the fundamental, and provide further information about the rhythmic composition. It is important to note that the radius and angle of all harmonics, the fundamental as well as the partials, are significant, as they describe what is best referred to as rhythmic timbre. Amplitude ratios between the fundamental and the various partials serve as a metric of beat salience– the clarity of the prevailing rhythmic percept– as well as a confidence interval regarding system reliability.

An added benefit of averaging the rhythm chroma is found in the fact that frequency modulations of the fundamental chroma manifest as a widening of the primary lobe. Due to the behavior of comb-filter resonance, tempo deviations will inherently attenuate the amplitude of the fundamental. From these observations, optimal music for RAS will exhibit a large, narrow and clearly-defined fundamental with smaller, though still clearly-defined, partials.

4. EVALUATION

Since there are, to our knowledge, no previous attempts to mathematically quantify the motor-rhythmic attributes of musical content, system behavior is explored for a small set of content defined as ground-truths. Initially, we examine the responses for a constant-tempo click track and a frequency-modulated version of itself. For familiarity, select content from the MIREX tempo tracking practice data is then processed by the proposed system.

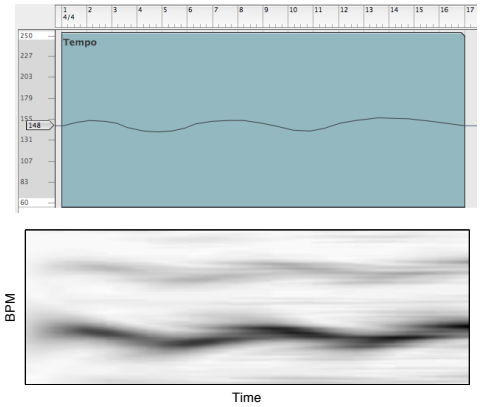

Figure 5. Image of the tempo automation used to modulate the tempo of the click track, and the corresponding chromagram after analysis.

The prominent role of metronomes and click tracks in past RAS research is indicative of the fact that they are the most basic form of motor-rhythmic stimuli. A thirty-second audio click track was created using a sampled clave in Propellorhead's Reason software and the tempo was set at 148 BPM. The software also offers the capability of tempo automation and allowed for the creation of a second, frequency-modulated click track to model an expressive performance. As shown in Figure 4, the constant-tempo click track produces a chroma with clearly defined fundamental and several smaller subharmonics, while the chroma lobes of the frequency-modulated click track are smeared and roughly half the amplitude. While salient, given the ratio of the significant peaks, the widening of the lobes is a direct result of the tempo variance in over time. Importantly, a chromagram is shown above the tempo automation curve used to modulate the tempo of the click track in Figure 5. Though the chromagram incurs some delay in tracking the modulation of the click track, the system is able to follow the tempo throughout.

Though informative and worthwhile examples to consider, click tracks are not the primary focus of this system and it is necessary to also examine the chroma of real music data. For ease of access and familiarity within the research community, musical content is selected from practice data available on the MIREX website [10]. The set of excerpts contains a variety of different styles, but there are two tracks in particular – *train8.wav* and *train12.wav* – that serve as prime examples of what is and what is not motor-rhythmic music.

Figure 6 shows the chroma for the two separate tracks. It is evident from the diagram that *train8.wav*, an electronic piece by Aphex Twin, is significantly more motor-rhythmic than *train12.wav*, an orchestral performance of a composition by J. S. Bach, with a beat strength nearly 40 times greater in amplitude. Despite the lack of harmonic definition in the chroma of the orchestral track, this system is capable of identifying the correct fundamental class for both excerpts according to metadata provided.

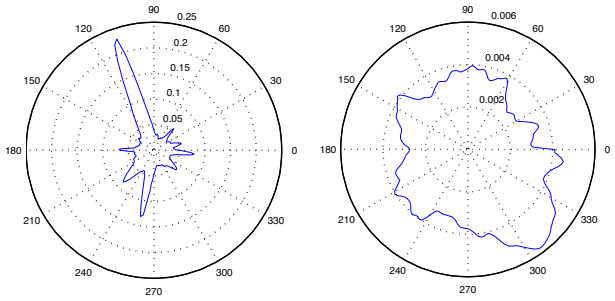

Figure 6. Instances of good (left) and poor (right) motor-rhythmic music.

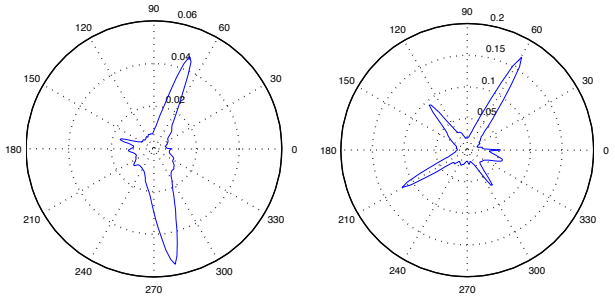

Figure 7. Chroma representations for non-binary meter tracks performed in $6/8$ (left) and $7/8$ (right).

5. DISCUSSION

Content analysis algorithms for the computation of feature-specific metadata will no doubt play a vital role in the future as digital music libraries continue to increase in volume seemingly without bound. The system presented here details one such application of a relatively straightforward use-genre that extends previous machine listening efforts. The task of characterizing music for RAS benefits greatly from the circumstances of the context in which it is used, wherein the most relevant attributes of motor-rhythmic music are objectively quantifiable.

Furthermore, representing the global rhythm in terms of chroma allows for a compact description of the temporal structure of music. Succinctly stated, the degree of tempo variation inherent in a track influences both the width and height of the chroma partials. Any music track can be reasonably approximated as a set of rhythmic partials with corresponding amplitudes, angles, and widths.

5.1 Future Work

One of the more interesting observations to result from this work is the realization that the harmonic structure of rhythm chroma may provide information about the meter and other time segmentations. Figure 7 shows the global chroma of two tracks of note from the MIREX practice data set: *train5.wav* and *bonus3.wav*. These tracks are of particular interest as they are not binary meter; the former is $6/8$ and the latter is $7/8$. The chroma of *train5.wav* is really only comprised of a fundamental and a closely-competing subharmonic at a difference angle of about 150°. Alternatively, *bonus3.wav* is comprised of a variety of subharmonics, but the partial located 70° from the fundamental is not even remotely present in any other chroma representations observed. More work is necessary to determine the true depth of the information contained within these data.

6. REFERENCES

[1] M. Thaut, G. McIntosh, S. Prassas and R. Rice, "Effect of Rhythmic Auditory Cuing on Temporal Stride Parameters and EMG Patterns in Normal Gait." *Journal of Neurologic Rehabilitation*, Vol. 4, No. 6, pp. 185–190, 1992.

[2] M. Thaut, *Rhythm, Music, and the Brain: Scientific Foundations and Clinical Applications.* Routledge, 2008.

[3] N. Masahiro, H. Takaesu, H. Demachi, M. Oono and H. Saito, "Development of an Automatic Music Selection System Based on Runner's Step Frequency." *Proc of the 9th Int Conf on MIR*, pp. 193–198, 2008.

[4] A. Klapuri, A. Eronen and J. Astola, "Analysis of the Meter of Acoustic Musical Signals." *IEEE-TSAP*, 2006.

[5] F. Kurth, T. Gehrmann and M. Muller, "The Cyclic Beat Spectrum: Tempo-related Audio Features for Time-scale Invariant Audio Identification." *Proc of the 7th Int Conf on MIR*, pp. 35–40, 2006.

[6] M. McKinney, D. Moleants, M. Davies and A. Klapuri, "Evaluation of Audio Beat Tracking and Music Tempo Extraction Algorithms." *New Music Research*, 2007.

[7] E. Scheirer. "Tempo and Beat Analysis of Acoustic Musical Signals." *Journal Acoustical Society of America*, 1998.

[8] G. Tzanetakis and P. Cook. "Musical Genre Classification of Audio Signals." *IEEE-TSAP*, Vol. 10. No. 5. pp. 293–302, 2002.

[9] L. Lu, D. Liu and H. J. Zhang. "Automatic Mood Detection and Tracking of Music Audio Signals", *IEEE-TSAP*, Vol. 14, No. 1, pp. 5–18, 2006.

[10] MIREX Website, [Online]. http://www.music-ir.org/mirex/2006/index.php/Audio_Tempo_Extraction.