UNIVERSITY OF MIAMI

A COMPUTATIONAL SYSTEM FOR THE AUTOMATIC CREATION OF MUSIC PLAYLISTS FOR RHYTHMIC AUDITORY STIMULATION IN RECREATIONAL RUNNING

By

Eric J. Humphrey

A THESIS PROJECT

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Master of Science in Music Engineering Technology

Coral Gables, Florida

December 2009

UNIVERSITY OF MIAMI

A Thesis Project submitted in partial fulfillment of
the requirements for the degree of
Master of Science in Music Engineering Technology

A COMPUTATIONAL SYSTEM FOR THE AUTOMATIC CREATION OF MUSIC PLAYLISTS FOR RHYTHMIC AUDITORY STIMULATION IN RECREATIONAL RUNNING

Eric J. Humphrey

Approved:

Dr. Colby N. Leider Professor of Music Engineering	Dr. Edward Asmus Associate Dean of Graduate Studies
Dr. Shannon K. de l'Etoile Associate Professor of Music Therapy	Dr. Corey Cheng Associate Professor of Music Engineering

HUMPHREY, ERIC J. M.S., Music Engineering Technology
December 2009

A Computational System for the Automatic Creation of Music Playlists for Rhythmic Auditory Stimulation in Recreational Running

Abstract of a Master's Research Project at the University of Miami

Research Project supervised by Professor Colby N. Leider
Number of Pages in Text: [129]

A system is proposed to reliably compile suitable, personalized running playlists for rhythmic auditory stimulation with minimal human intervention. Previous work related to automatic music selection for accompanying rhythmic physical motion exhibits several significant shortcomings this project is designed to address, in particular, a disconnect between musical pulse and neuromuscular entrainment, a neglect of the actual musical content and the manual tabulation of a global track tempo. This system is developed such that a non-invasive cadence monitor, designed specifically for this application, is used to assess an individuals target physical activity tempo and natural variance. Additionally, an improved application-specific tempo induction algorithm is developed to computationally model human rhythm perception and map the tempo evolution of a digital music track, while simultaneously characterizing its rhythmic nature. A database of music is analyzed accordingly and the most suitable tracks for rhythmic auditory stimulation are identified and manipulated, for use as a pacing mechanism in free-field running. Preliminary results show significant advantages over previous

tempo induction algorithms, highly reliable activity cadence estimation, a decrease in RPE, and the observation that average track tempo alone is not sufficient for appropriate song selection. Further physical and psychological evaluation is to follow. Analysis and future areas of research are also discussed.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

1

Introduction

Digital multimedia has become an integral, and somewhat inescapable, aspect of modern life. Personal computers have seen a divergence in purpose, with desktop machines being relegated to work tasks and compact mobile devices being a natural fit for more personal, everyday applications. There is of course still a good deal of overlap, with laptop computers sitting squarely in the middle, but the trend itself is immediately recognizable. In short, personal handheld devices have been impeccable at acquiring, managing and playing content. Calling the iPhone, the first success of its kind, a "phone" is a substantial understatement of its capabilities; it can both play and record music and video, capture and edit pictures, take notes, play games, locate itself in the world, get directions, give directions, send and receive email, and, of course, make phone calls, to name but a handful of tasks. This device clearly builds upon the success of other personal media technology, namely its close kin, the iPod (Apple, 2009a). Ubiquitous computing has ushered in an age of other intelligent portable gadgets, including phones featuring the Android operating system (Gizmodo, 2009) and Microsoft's anxiously–anticipated tablet computer (Review, 2009).

Importantly, the iPod was not the first personal music player. Many knowledgeable parties will argue quite convincingly that it was not even the best when it first launched, nor possibly even today. One will be hard-pressed to find,

however, a single person that fails to recognize the manner in which the iPod and all of its success revolutionized the world of personal music players. The first widely used media for personal music was the cassette tape, which later gave way to the CD player. Sony's Walkman became the standard in personal music technology for the better part of the 1990's (Wikipedia, 2009b), due in large part to its portability. Other forms of personal music players lived in the background, such as the MiniDisc player, but they all suffered from the same shortcoming: the user was forced to carry all of the physical content they may wish to hear while on-the-go. MP3 players allowed users to carry a never-before-comprehendible amount of diverse content with minimal effort. Coincidentally, the growth of digital media players, like the iPod, happened to peak in stride with infamous music file-sharing services, like Napster, Kazaa and Limewire. Users were not only able to carry their entire digital libraries everywhere, but illegal file-sharing enabled an overwhelming percentage of music consumers to quickly expand their music collections well beyond the limits of what was considered a normal–sized collection.

In addition to facilitating the mobility of a large digital music library, users could now also easily make playlists, mixes, or simply shuffle their entire collection. This created the ability to spontaneously access content outside the container of an album while mobile, greatly diminishing the popularity of the mix tape (and its digital counterpart, the mix CD). Digital music players made it both easier and less wasteful to maintain music for certain activities. Rather

than accumulating "Driving Mix" CD after "Driving Mix" CD, which would invariably become outdated, users could create and maintain that same subset of their digital music library on their portable device. For this same reason, it became very common for people to begin exercising to music in greater numbers.

While exercising to music occurred in the time before MP3 players, this technological advancement had far-reaching effects on the entire world of exercise for years to come. In the past, aerobics have been strongly tied to music, with such examples as Jazzercise and "Sweatin' to the Oldies," hosted by Richard Simmons. Switching focus from the broad spectrum of exercise to a more specific example, taking music along with oneself while running was never really feasible, if at all enjoyable. Personal cassette players are far too bulky for many people to carry while running, and personal CD players suffer from the same physical form factor, in addition to being prone to skipping. As a result, listening to music while running generally precluded the use of a treadmill. Succinctly put, even though the personal music player was portable, it was hardly mobile. Returning to the current state of personal media players, a diminished size and increased memory capacity lend themselves naturally to extremely active mobile applications, such as running accompaniment. Apple, the world leader in personal music player sales (Telegraph, 2009) and digital music distribution (Apple, 2009c) as of this writing, manufactures several devices specifically developed for runners who listen to music. The Nike+ iPod system has yielded a product line of technology and accessories to enable and enhance the experience

of running with music, and it has generated a substantial online community, actively running and cataloging their performance (Jochelson and Fedigan, 2006).

There are a few key elements to derive from our cultural history, as related to exercise and personal music players. Exercise, and particularly running, is an integral component in the lives of a large number of people. This community increases considerably when factoring in the number of people that are, or would like to, start making exercise an important part of their lives as well. As it stands, over a million people already use the Nike+ iPod system (Apple, 2009b) to run with music, neglecting the many others that only listen to music on iPods or other miscellaneous personal media players. Merging these observations precipitates a series of related questions: How do people use music while running, and how do they find it? Is there optimal music to which an individual can run? Can "running music" be established as a specific use-genre, and is it extensible to recommendations? Ultimately, can this entire process be automated computationally? In addressing these questions, this work aims to enhance and improve the activity of running for novice and trained athletes alike by automatically providing rhythmic auditory stimuli through an individual's preferred music, meeting both physiological and psychological needs.

Foremost, it is necessary to introduce the fundamental elements upon which the proposed system is built, covering principles of musical rhythm and perception, aspects of motor neuroscience and kinesiology, and important components of computational music signal processing. Following this background

information, an in-depth synopsis of the relevant literature will be conducted, covering music-running systems, beat-induction algorithms, and music information retrieval research. An overview of the developed system, capable of automatic personalized playback generation, will be presented and followed by a step-by-step investigation of the sub-processes and the unique contributions made. A detailed testing and evaluation section will then outline the methodology and address the merits of the developed system, in terms of computational performance and human subject test data. This work will conclude with a comprehensive summary and discussion of the findings, and address identified areas of future work.

2

Background

Developing a system capable of generating user-specific music playlists to enhance the physical activity of running requires requisite knowledge from three distinct fields of study. Music perception plays a vital role in understanding how humans interpret sounds and form relationships between discrete events, forming rhythms, phrasings and songs. It is also necessary to establish a physiological foundation of the human body and the neurological processes that drive our motor skills. To tie these two separate camps together via a computational algorithm, machine-listening methods are introduced as a mathematical means to model human abilities and skills.

2.1 Music Perception

The human species' capacity for musical activity incorporates several processes, some of which are better understood than others. Biological mechanisms such as simple sound transduction have been thoroughly explored, while various faculties of the brain remain a mystery. As will be discussed, current scientific research has rather accurately traced acoustic perception from acoustic wave to encoded messages on the auditory nerve, but exploring the intricacies of perception beyond that have only recently begun to yield findings.

2.1.1 Sound Transduction

It is impossible, or at least ill advised, to build an algorithm that models the mechanisms of the human auditory system [HAS] without paying homage to millions of years of evolutional development, particularly when that very algorithm is intended to hear the way humans do. That being said, a review of the HAS is both pertinent and necessary.

All sound we as humans ever hear is transmitted to the brain as two encoded streams of impulses. To arrive at these impulse trains, the body performs preprocessing on the acoustic environment in which it resides. Sound is both collected and diffused by the pinnae and channeled through the auditory canal, where acoustic-mechanical transduction passes the signal to the eardrum. The bones of the inner ear – the malleus, incus, and stapes – then receive the signal from the eardrum and perform dynamic compression by loosening and contracting the muscles that control them. After passing through the inner ear, the signal reaches the cochlea, causing the fluid-filled organ to propagate the waves the length of the coiled structure. In a frequency-to-place transformation, somewhat analogous to a filterbank, the fluid-mechanical wave is encoded through electro-chemical transduction by hair cells lining the basilar membrane. Diagrams of the human ear and the cochlea are given in Figure 1 and Figure 2.

With a guided emphasis on the task of musical rhythm perception, the cochlea exhibits a variety of noteworthy nuances and behaviors. As mentioned, it is understood, albeit in the company of complementary theories, that the cochlea

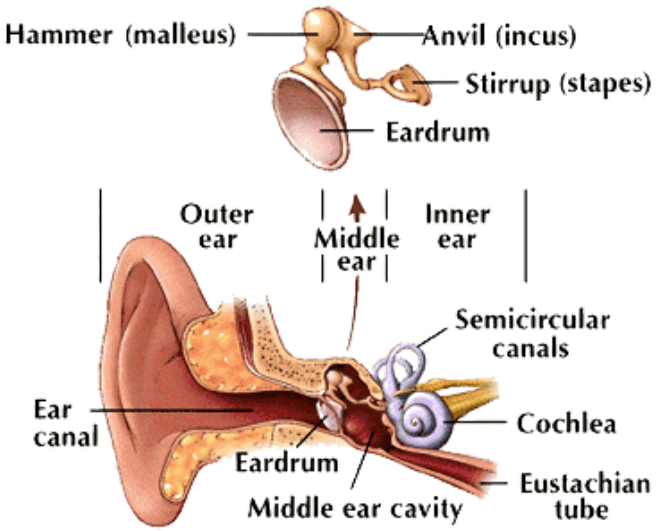

Figure 1. **The Human Ear** - Acoustic waves propagate through the ear canal, are transferred to physical energy by bones of the middle ear, and encoded as neural impulses by the inner ear. Diagram from (HearingCentral.com, 2009b)

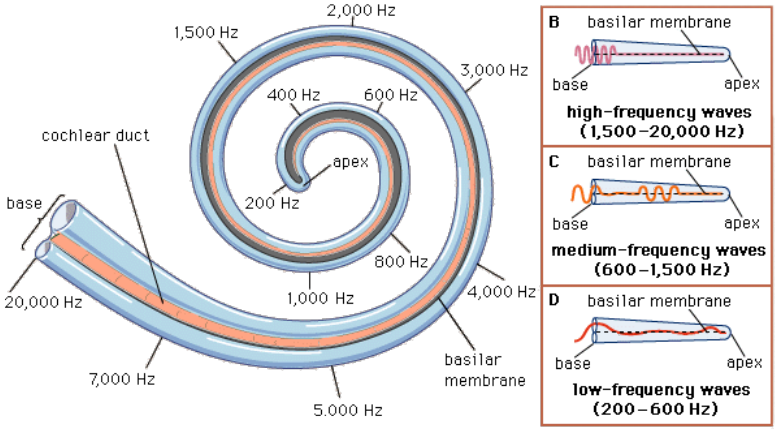

Figure 2. **Diagram of the Cochlea** - The cochlea performs spectral decomposition in a frequency-to-place mapping by exciting hair cells on the walls of the basilar membrane. Diagram from (HearingCentral.com, 2009a)

is comprised of a set of about 25 *critical bands* of increasing bandwidth, the frequency ranges of which are given in Table 1.

Humans perceive frequency logarithmically, and these widening critical bands can be mapped to the equivalent rectangular bandwidth [ERB] scale. Additionally, the cochlea exhibits both temporal and spectral masking behavior. Hair-cell activation in a critical band causes an electro-chemical reaction and, upon firing, requires some time to return to equilibrium, temporarily preventing the hair cell from excitation (Cook, 1999). Asymmetric spectral masking also occurs given the serial nature of the propagating fluid wave, causing stronger frequencies to mask weaker ones nearby.

Band	**Center Freq.**	**Bandwidth**	**Band**	**Center Freq.**	**Bandwidth**
1	50	– 100	14	2150	2000 – 2320
2	150	100– 200	15	2500	2320 – 2700
3	250	200– 300	16	2900	2700 – 3150
4	350	300– 400	17	3400	3150 – 3700
5	450	400– 510	18	4000	3700 – 4400
6	570	510– 630	19	4800	4400 – 5300
7	700	630– 770	20	5800	5300 – 6400
8	840	770– 920	21	7000	6400 – 7700
9	1000	920– 1080	22	8500	7700 – 9500
10	1175	1080– 1270	23	10500	9500 – 12000
11	1370	1270– 1480	24	13500	12000 – 15500
12	1600	1480– 1720	25	19500	15500 –
13	1850	1720 – 2000			

Table 1. **Critical Band Frequency Ranges** - Description of the filterbank-like nature of the cochlea.

2.1.2 Rhythm Fundamentals

Through the transduction of acoustic waves, the brain receives and interprets, in the context of rhythm audition, the most musically relevant events. Discerning structure, patterns and form are a uniquely human task, at which we as a species excel. The hierarchy of rhythm is built upon discrete perceived

musical events, referred to as onsets, accents or pulses. Attacks and transients are closely related concepts to onsets, and deserve distinct attention. An attack describes the time envelope of an acoustic sound. Transients are a type of signal behavior that may coincide with a musical onset. This is the case for percussive instruments, while bowed instruments tend to lack transients despite having a clearly perceived onset.

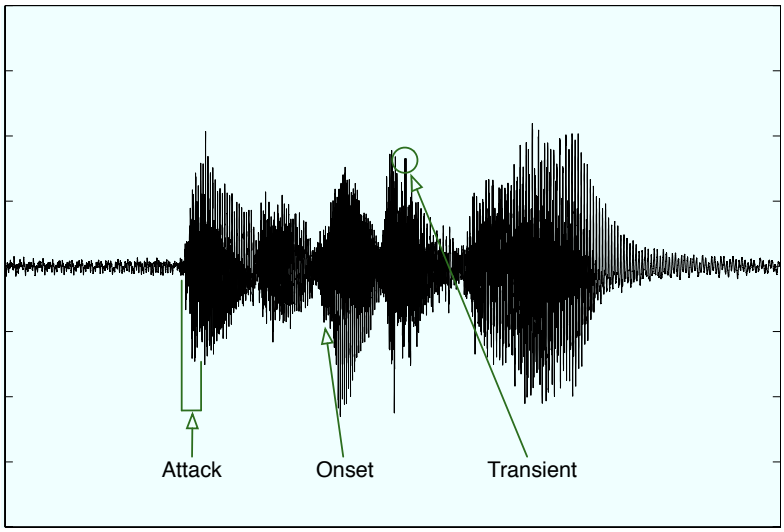

Figure 3. **An Attack, a Transient, and an Onset** - Waveform of a recorded cello playing five sequential notes, an instrument known for exhibiting particularly long attack periods and non-trivial onset annotation.

From these perceived onsets, the human brain identifies the structure by which the musical events are, or are not, related. This baseline structure forms the psychological concept of the beat, or internal clock, in music. Large succinctly states that "perceived beat is an inference from the acoustic stimulus, and functions as and expectation for when events are likely to occur in the future." (Large, 2000, p. 532). Therefore, making the leap from perceived onsets to a felt beat is the brain's transition from sensing to cognition. For different

levels of the felt beat, the most salient is defined as the *tactus* and the finest is defined as the *tatum*. Fractal patterns in common binary rhythms often result in the tactus being a superset of the tatum, as diagramed in Figure 4. Meter can then be broadly defined as a higher level of organization for the underlying beats.

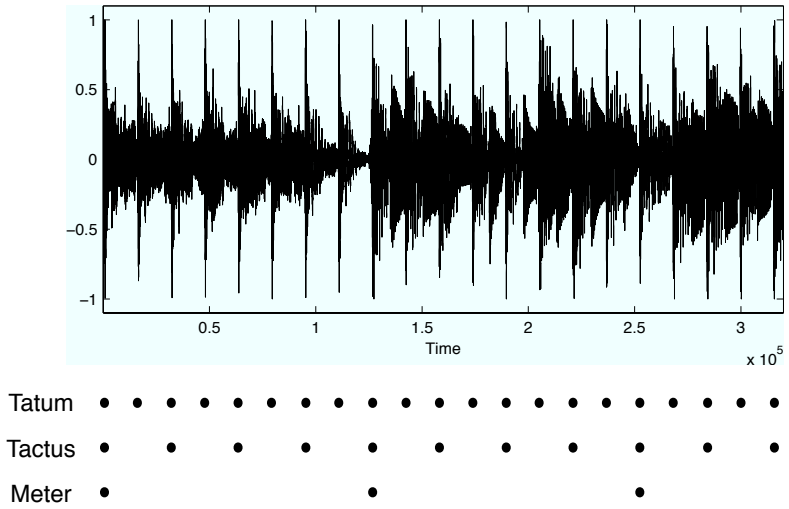

Figure 4. **Tatum, Tactus, and Meter** - For a clip of electronic music, the temporal grid can be inferred on visual inspection of the waveform.

For the purpose of the work here, it is necessary to understand, or at least model, exactly how the human brain infers a felt beat from onsets. One leading school of thought promotes the theory that human beat induction is achieved through the resonating, or entrainment, of oscillator banks in the brain as a interval-period based process (Thaut, 2008; Large and Kolen, 1994). To expand on this thought with an eye to the next topic, according to Thaut:

> Pulses are crucial components in music. They are inferred from and constructed upon time intervals, and ultimately depend on them for their continued existence. Pulses serve another important function in rhythm perception that has tremendous implications for the control of cognitive functions and motor performance: They establish anticipation and predictability, two components that have

> tremendous influence on the regulation of nonmusical temporal processes in perception, cognition, and motor control (Thaut, 2008, p. 8)

A thorough discussion of periodic temporal events is not complete without properly considering the corresponding spectral significance. Musical rhythm is understood equally well in terms of both beat period and frequency, where, in common time $\frac{4}{4}$, a quarter note at 60 beats per minute (BPM) is equivalent to a half note at 120 BPM. As a result, a human listener of moderate skill can easily feel a piece of music at various tempo octaves. Pitch audition, the ability to associate a pure tone of a given frequency with a perceived sound, also comprises octaves and can be compactly described by a chroma (class) and height. For example, in 12-tone equal temperament (12-TET), the standard intonation in almost all Western music, a C3 and a C5 have the same chroma (a C) but exist two octaves apart.

Interestingly, tempo can be described in similar terms, as a rhythmic excerpt will exhibit some fundamental tempo grid to which all sub-elements align. Kurth et al. illustrate this concept of tempo chroma as a cyclic beat spectrum (CBS), such that octave frequencies align on the unit circle (Kurth et al., 2006). The CBS can be easily expressed in terms of radians traversing a polar grid, such that for a tempo frequency ω, all frequencies of a chroma are defined by $\omega_k = (\omega + 2k\pi)$, where $[k = 0, \pm 1, \pm 2 \dots]$. The notion that musical tempi can be fractured into the two attributes of chroma and height is ultimately significant to this work. Expressed in these terms, the height of a

tempo is wholly subjective, whereas the chroma is an objective, matter-of-fact. For the purposes of auditory cuing and motor synchronization, only the chroma is relevant, and a computational algorithm can be developed to extract this feature. Tempo chroma is further depicted in Figure 5.

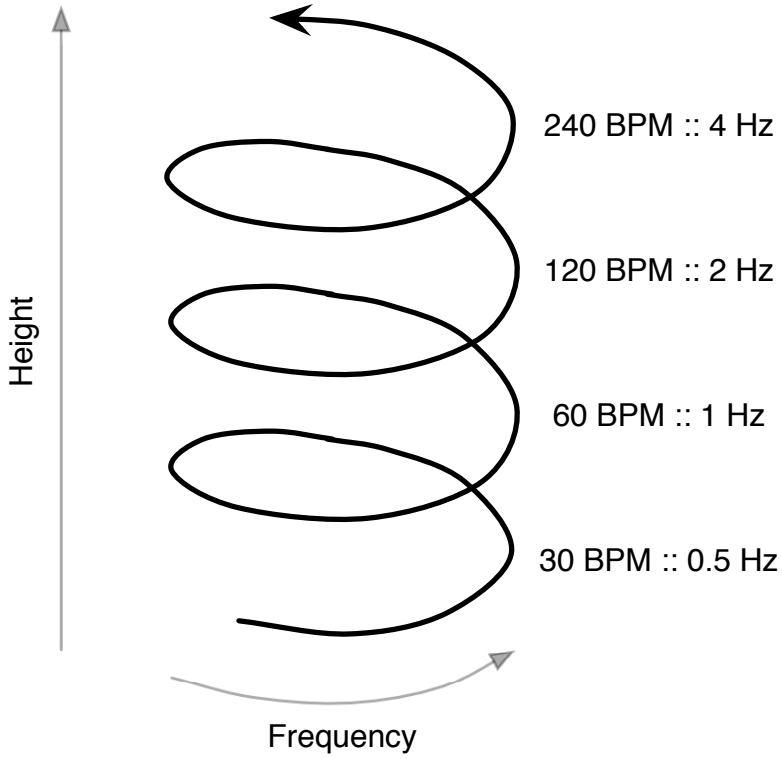

Figure 5. **Tempo Chroma** - Visualization of the relationship between tempo *octaves*.

2.2 Human Biology

As a mechanical system, the physical act of running depends on substantial coordination of neuromuscular activity that many take for granted, since automaticity — the ability to control motor functions without direct focus, e.g., walking and talking — is generally achieved early in life. This is particularly true for healthy individuals, where movement seldom requires conscious thought

or attention. Many recent advances in the understanding of the brain's timing mechanisms are the result of a multidisciplinary approach between neuroscience and music perception (Molinari et al., 2003).

2.2.1 Kinematics

The moving human body is intrinsically a dynamic mechanical system and, in the case of rhythmic motion (e.g., walking, running, dancing), is composed of multiple oscillations at different harmonically related frequencies. Like any oscillating system — electrical, mechanical or otherwise — the human body will invariably exhibit a fundamental operating frequency at which the system performs optimally, otherwise known as the resonant frequency. In the field of music therapy, where rhythmic cuing of physical movement is studied at great lengths, this is commonly referred to as a limit cycle. There are several factors that contribute to an individual's optimal cadence, or step frequency, including fitness, fatigue, gender, height and age (Auvinet et al., 2002). Additionally, running velocity is more a function of stride length than step frequency, which stays relatively constant independently, except for sprinting velocities (Dillman, 1975).

Non-invasive monitoring of these parameters is robustly achieved through the wearing of an accelerometer placed at the subject's center of gravity (Moe-Nilssen and Helbostad, 2004; Auvinet et al., 2002; Kavanagh and Menz, 2008). Unreliability of temporospatial gait parameter measurement may arise from ill-fitting or poorly attached accelerometric sensors,

so care must be taken to securely affix all sensors to the subject (Kavanagh and Menz, 2008). The only viable alternative to accelerometer-based solutions for cadence monitoring is some form of embedded, or surface-supported, pressure sensor. Typically, this style of sensor negates the prerequisite non-invasiveness, as the device must be mounted underfoot, in turn requiring some form of manipulation of the running shoe (e.g. Nike+ System (Apple, 2009b)).

2.2.2 Neuroscience and Rhythmic Auditory Stimulation

Through an increased understanding of the underlying mechanisms involved in the physiological response to music, our current knowledge supports the position that rhythm serves as a powerful external timing mechanism capable of entraining gait parameters and neuromuscular activity (Thaut et al., 1992). Tecchio et al. found evidence for the preconscious discrimination of temporal stimuli in the auditory cortex through the administration of pure tones and subsequent magnetoencephalographic (MEG) imaging of the brain, supporting the existence of a distinct timing neuronal network (Tecchio et al., 2000). The related work of Molinari et al. determined that temporal information is preconsciously processed external to the cerebellum by examining the responses of subjects with cerebellar disorders, such that time coding can be transferred directly into adjacent motor structures, entraining neural motor codes and allowing synchronization between auditory stimulus and motor response (Molinari et al., 2003). Other studies have shown that basic gait parameters,

such as velocity, step frequency, and swing symmetry, have been improved through the use of rhythmic cuing (Thaut et al., 1992).

The results of these and other clinical investigations intuitively complement the observation that, throughout human history, strenuous physical labor was often accompanied with rhythmic songs. As a prime example, sea shanties assumed many forms, tailored to facilitate different types of physical labor encountered on a sea-faring vessel. Short and long haul songs served to orchestrate the pulling of loads, capstans aided the raising of the anchor, and pumping songs accompanied the removal of water held in the ship's bilge (Wikipedia, 2009a).

In the field of neurologic music therapy, an entire method of sensorimotor rehabilitation is founded on this principle. One of the foremost researchers in the field, Michael Thaut, defines Rhythmic Auditory Stimulation (RAS) as "a neurological technique using the physiological effects of auditory rhythm on the motor system to improve the control of movement in rehabilitation and therapy" (Thaut, 2008, p.139). There exists a large, and growing, body of work exploring the usage of RAS for the rehabilitation of various gait disorders stemming from a variety of physical, neurological, and age-related causes (de l'Etoile, 2008; Thaut et al., 1992; Thaut et al., 1997; Thaut et al., 1996; Hurt et al., 1998). Much of the previous work related to RAS has focused primarily on clinical rehabilitation studies of subjects with significant gait deficiencies, with a common objective of aiding the recovery of healthy walking abilities. According to Thaut,

RAS can be used in two distinct ways (Thaut, 2008, p.140):

> 1. As an immediate entrainment stimulus providing rhythmic cues during the movement. For example, individuals may listen to a metronome or rhythmic music tape while walking to enhance their walking tempo, balance, and control of muscles and limbs.
> 2. As a facilitating stimulus for training; patients train with RAS for a certain period of time in order to achieve more functional gait patterns, which they then transfer to walking without rhythmic facilitation.

Exploring the impact of rhythmic auditory stimuli on movement, there are three elements to address (Hurt et al., 1998). 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 — a golf swing, for example — that can be recalled and executed, and rhythmic stimuli encourages the creation of more efficient and fluid programs for cyclical movement. Considerably the most meaningful concept is that of *goal directed movement*, where motion is defined by anticipation rather than an explicit event such as a heel strike. Evidence has found that movement is further guided by interval adaptation rather than phase entrainment, the significance of which can be seen in Figure 6.

Despite the discrepancy in subject and goal, both uses of RAS are equally applicable physical therapy techniques for healthy subjects during running. In this new scenario, the objective of the rhythmic motor cuing is to facilitate the continued development of the relevant musculature, cardiovascular stamina, and physical endurance necessary for improved physical performance. It is important,

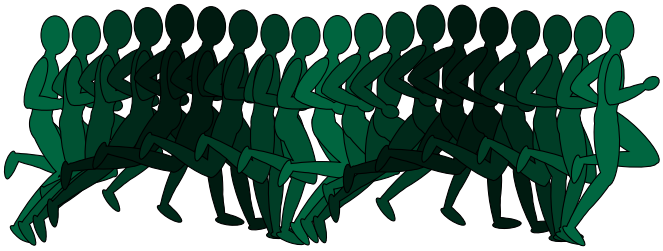

Figure 6. **Phases of Running** - Though marked by significant discrete events, complex motor coordination is controlled over a continuous interval.

Thaut states, "RAS frequencies need to be set initially at the current natural or intrinsic frequency of the person's movement." Only upon achieving stability or, in the case of healthy subjects, comfort, "new limit cycles can be gradually entrained through a stepwise entrainment process" (Thaut, 2008, p.143).

2.2.3 Ratings of Perceived Exertion

Complementary to the technique of rhythmic cuing and motor synchronization, music accompaniment during sub-maximal exercise can positively affect a person's psychological state. Studies over the last twenty years have investigated the link between attentional manipulation in both active and passive forms, finding that music can reduce one's rating of perceived exertion (RPE). In the words of Mohammadzadeh et al., an RPE is defined as "the subjective intensity of effort, strain discomfort and/or the fatigue that is experienced during an exercise" (Mohammadzadeh et al., 2008), and it is influenced by the subject's environment and conditions. It is important to note that the most marked occurrences of lowered RPEs occur in subject demographics that identify themselves as novice, or non-, runners, as opposed to

habitual or expert runners. This discrepancy is resolved by the theory that subjects at different fitness levels employ alternate psychological coping mechanisms, where beginners use music to consciously dissociate from the exercise and experts instead focus on it (Mohammadzadeh et al., 2008). Being a subjective metric, RPEs are generally measured using the Borg scale, shown in Table 2, and are usually taken several times during an activity.

Rating	**Description**
20	maximal exertion
19	extremely hard
18	
17	very hard
16	
15	hard (heavy)
14	
13	somewhat hard
12	
11	light
10	
9	very light
8	
7	extremely light
6	no exertion at all

Table 2. **Borg Scale for Ratings of Perceived Exertion** - Numerical values associated with an individual's perception of exercise intensity.

The validity of music to reduce a subject's RPE is not uncontested, with an opposing body of work finding no statistical correlation between the two (Mohammadzadeh et al., 2008). However, one of the more convincing studies in support of lowered RPE via music listening takes into account a subject's musical preference, a crucial and evasive variable to constrain. Boutcher and Trenske theorize that music acts as a distraction, and may heighten emotional arousal due to association with past experiences (Wininger and Pargman, 2003). This realization is exceedingly important in the context of a computational algorithm that attempts to model an individual's listening abilities. Even the best

machine-listening algorithm, or any human for that matter, can only infer information from the song at hand if it does not have memory, and does not have or maintain any a priori knowledge of the individual the algorithm (or person) is modeling. An individual's music is important to them because it is his or her music, and any music psychology study exploring the existence of heightened affective arousal or stimulation must consider this fact accordingly.

2.3 Music Signal Processing

The transition to ubiquitous digital audio over the last twenty years has concurrently generated interest and effort in adapting conventional digital signal processing (DSP) theory to specific audio and music techniques. Entire fields of research, such as automatic music transcription (AMT), have stemmed from the ability to directly process musical information creating a wide spectrum of derivative technologies. A full review is given in (Klapuri and Davy, 2006). Pertinent signal processing methods are addressed to establish foundational elements upon which the proposed work will build.

2.3.1 Auditory Modeling and Biomimetic Design

All philosophy aside, humans and computers are machines; for obvious reasons, the latter are far better understood than the former. While a great deal concerning the inner workings of the human mind is still a mystery — although slowly unraveling — it is best to focus on the elements that are understood when modeling the more complex system with a simpler one. Biomimetic technology imitates processes and adaptations found in nature to serve some function. The

aperture of a camera is a prime example of a biomimetic technology, designed with cues from the human eye, and is illustrated in Figure 7. There are a number of benefits to biomimetic technology, but two in particular apply to this work. First, a system that will process and analyze audio to arrive at a result similar to a human counterpart would likely do well to operate in a manner similar to that of a human. Second, the evolutionary process has spent a great deal of time refining the biological mechanisms that constitute the human species, and the result is very adept at performing a wide variety of tasks.

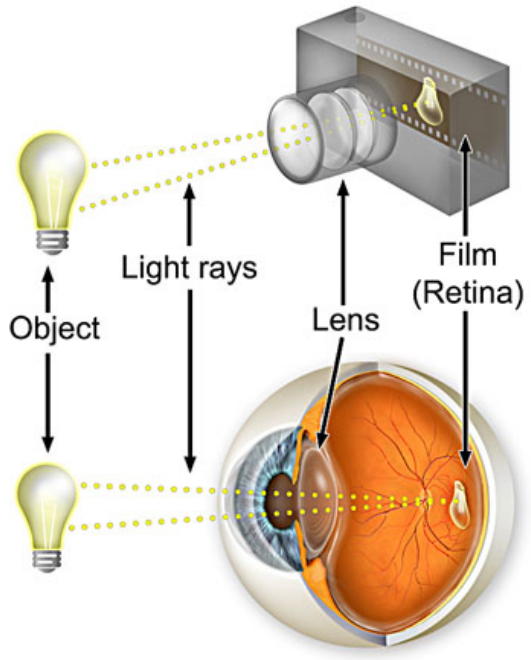

Figure 7. **The Human Eye & the Camera** - Design cues from nature manifested in human progress.(JirehDesign.com, 2009)

Therefore, a biomimetic approach is adopted in modeling the HAS and the manner in which humans perceive sound and, as a cognitive abstraction, rhythm. Human beat audition is coarsely performed through observing a musical

signal, detecting meaningful events, and establishing a temporal relationship between them. Acoustic observation spans the steps from sound reaching the outer ear to arriving at the auditory nerve, so these stages are modeled accordingly. The cochlea, for example, is approximated as a frequency-to-place filterbank. Onset extraction is performed as a function of the hair cells encoding information to the auditory nerve as a series of pulses. Involuntary pattern recognition of temporal relationships between pulses is modeled computationally to arrive at an estimated percept of tempo and beat structure.

2.3.2 Maximally-Decimated Filterbank Design

Decomposition schemes for signal analysis serve to break down an input signal into an arbitrary number of subband components. A Quadrature mirror filterbank (QMF) is a special class of filters designed to meet specific requirements that divide an input into two halfband signals. By splitting the frequency content in half, the sampling rate of both output signals can also be halved with no loss of information. Subsampling the output sequences by a factor of two will expand the spectral content of both, such that the signals are critically sampled, or maximally decimated. The outputs of the analysis bank can be recombined via filtering with a pair of matched synthesis filters to produce the original input signal, thereby satisfying perfect reconstruction conditions. Any aliased frequency content is negated through the proper, complementary design of a QMF analysis/synthesis system, in addition to phase and amplitude distortions. Certain wavelet families, such as Haar and Daubechies, meet perfect

reconstruction criteria, and prove beneficial in the perceptual compression of various media.

Maximally decimated filterbanks provide particular computational advantages over other designs and implementations. Successive filtering and subsampling of an input signal produces no redundant information and effectively narrows the passband of the cascaded filtering operation. Elaborating, for an input signal of length N, a halfband decomposition and downsampling by two will produce two signals of length $\frac{N}{2}$. Repeating this process an arbitrary number of times, symmetrically or otherwise, will maintain at most a total of N samples across all output channels (barring, of course, subsampling of a non-factor ratio). For frame-based halfband decomposition schemes, it is intuitive that the selection of a power-of-two frame length is considered good practice. Decomposition via a maximally decimated filterbank can be conceptualized as partitioning the spectra of a signal into fractal components, a diagram of which is presented in Figure 8. Multilevel filterbank decomposition can employ application-specific tree structures to optimally perform the task at hand, achieving comparable flexibility to direct implementations of bandpass filterbanks. Significantly more information can be found in (Vaidyanathan, 1993).

2.3.3 The Phase Vocoder

In addition to computational music analysis, the digital representation of musical signals presents the opportunity for subsequent processing and manipulation for a multitude of ends. Realizing the fact that, for a given digital

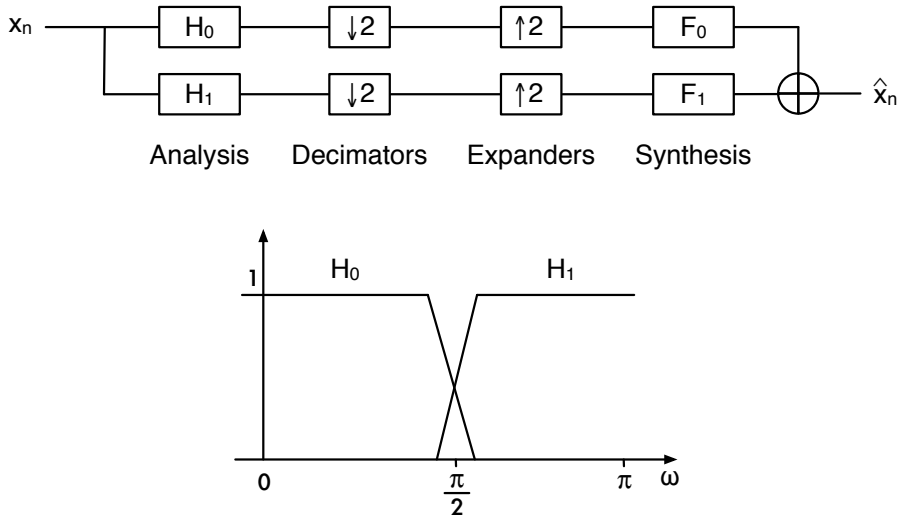

Figure 8. **Single–Level QMF Structure** - Filter diagram (top) and corresponding magnitude response (bottom).

music library (DML) of arbitrary size, the subset of content that is inherently both suitable for rhythmic auditory stimulation *and* exhibits the necessary tempo chroma may be extremely limited or even null. To overcome this potential lack of content, it is desirable to adjust the tempo of motor-rhythmic music to expand the resulting subset.

Given its long research history, there is an undeniable interest in the development of a high–fidelity time–scale modification algorithm for digital audio. Direct time manipulation of a digital waveform, such as sample rate adjustment, causes (generally) undesirable pitch shifting as time and frequency are, at least naively, inextricably linked. There is primarily one time domain algorithm, synchronous overlap-add (SOLA), aimed at avoiding undesirable distortions. SOLA operates by splicing an input signal into small frames, overlapping them in time accordingly, and reconstituting the signal. A variation

of this, pitch synchronous overlap-add, attempts to improve subjective ratings by overlapping adjacent frames at optimal points of alignment (Zöler, 2002).

A more common approach to time-scale modifications is taken by manipulating the time-frequency representation of a digital audio signal. The phase vocoder algorithm operates in an analysis-modification-synthesis paradigm, transforming an input signal into a two-dimensional spectrogram, followed by appropriate processing and inverted back to the time domain (Dolson, 1986). Phase vocoders are prone to producing artifacts in the final output, and much emphasis is placed on the nature of the mid-stage modifications. Described as phasiness, perceptible errors result from phase information blurring occurring during processing, and are particularly present for significant time expansions. PVC is an example of a publically available phase vocoder software package for Unix machines. Among a wide variety of functionality, PVC allows the command line and shell script analysis and subsequent time warping of digital audio files (Koonce, 2009).

3

Previous Work

Based upon the outlined merits of previous neuromuscular and physiological research, a review of work related to the development of an automated system is conducted. It is a logical first step to explore those systems that explicitly link music selection and exercise through computational means. Comprehensive automation also necessitates a review of modern machine–listening techniques in the area of algorithmic beat induction and its subcomponents. Feature extraction and classification — two principal elements of music information retrieval — are additionally pertinent topics to the task at hand.

3.1 Music-Running Systems

Over the last several years, many groups have invested effort in the development of a musical running accompaniment system that harnesses the recognized benefits of music listening during exercise. The first identifiable mention of a music playback system for running accompaniment was outlined as early as 1995, evolving over the course of several years and spawning a few publications. Researchers at the Fraunhofer Institute developed the StepMan music system, described by its creators as "matching music to your moves" (Bieber et al., 1995). Biometric sensors non-invasively monitor user performance parameters, and adjust the system, in real-time, accordingly. Time-warping

without introducing frequency distortion is outlined as a significant contribution of the system, and the developed algorithm of Time-Domain Harmonic Scaling is discussed in the context of an implemented phase vocoder. The authors propose that the system can operate in both a master- or slave-clock environment, allowing the music to synchronize to the runner or vice versa. Step frequency is measured using an accelerometric sensor, and an online step detection algorithm calculates the runner's current cadence. Additionally, StepMan can purportedly adjust playback speed as a function of heart rate or step frequency, with a bidirectional accelerometer attached at the thigh. The authors go on to discuss the inclusion of additional biometric sensors, where the correlation between observed physical output and musical feature is more abstract in nature (pulse is too high, slow down the tempo). Absent from the discussion of the system, however, is a mechanism by which the StepMan technology gains awareness of the actual music content, the fundamental component on which the entire system rests. (Bieber et al., 1995; Bieber and Diener, 2005).

Most subsequent systems adopt a similar approach to this initial predecessor. In 2005, an adaptive music playback algorithm was developed titled IM4Sports (Interactive Music for Sports), with the goal of establishing a personalized process to vary the playback speed of music based on both performance and a training program. Music selection is implemented as a heuristic retrieval search, based on age/gender parameters and training programs, with a priori content metadata. Suitability of a track for playback is

influenced notably more by a penalty function that observes user-invoked track skipping than the actual musical content (Wijnalda, 2005).

Apple has produced two relevant bodies of work relating running and music accompaniment. The Nike+ iPod running system, mentioned previously, is a commercially successful exercise system that integrates biometric sensing into a conventional personal music player. Users are able to select a single "Powersong" from the device's local music library that can be easily queued for instant playback. This action is, however, the extent of linking actual musical content and an individual's physical activity. In the confines of the Nike+ iPod system, however, music and running otherwise exist independently of each other, and the system can be used without music playback little impact. The two primary functions of the system — performance tracking and music playback — could exist in two distinct enclosures with no loss of functionality (Apple, 2009b). Apple has shown interest in the development of another system, notably similar to the work done at the Fraunhofer Institute. Through a series of patent applications, first filed in 2005 after the publication of the StepMan system and culminating in an issued patent in April of 2009, Apple outlines a system that provides equivalent functionality of previous systems. Titled "Music Synchronization arrangement," there are, given the detail in the intellectual property documentation, no meaningful distinguishing characteristics from previously addressed research (Bowen, 2009).

In addition to these projects, other similar systems have successively come

to light. Microsoft Research produced a white paper on their MPTrain system in 2006, detailing the hardware and software environment to establish a user-adaptive music playback environment. Distinguishing features of the MPTrain system include the user's ability to define a training program of heart-rate stress over time and its capacity to "learn the mapping between musical features (volume, beat, or energy), the user's current exercise level (running speed or gait), and the user's current physiological response (heart-rate)." It is important to note, however, that the authors design the system to monitor speed in number of steps per minute, a sign of misappropriated cause and effect that will be addressed shortly (Oliver and Flores-Mangas, 2006).

Greg Elliot's PersonalSoundtrack, also published in 2006, employs an accelerometric sensor and a laptop to communicate with an iPod music player, creating another context-aware music playback system. The only markedly distinct aspect of PersonalSoundtrack is the approach by which it is developed, coming more from a user-experience angle than a physiological or algorithmic one. The primary emphasis is on how the user interacts with the system and the subject experience one has through the use of a system capable of automatically adapting to exhibited human behavior (Elliot and Tomlinson, 2006).

A more recent project, published in late 2008, focused again on the development of an automated system to select music for playback based on a runner's step frequency, while also attempting to assess the psychological impact such a system has on users. In establishing a knowledge background for the

project, the authors reference an earlier publication (Ahmaniemi, 2007) that found, in the course of a study, no significant relationship between a song's tempo and a runner's step frequency. After conducting their own similar experiment, they concluded "people tend to feel pleased when BPM is near their SPM" (Masahiro et al., 2008). In the context of rhythmic auditory stimulation, this result is attributed to the system operating more efficiently.

Given the degree of similarity between these previous systems, they can effectively be discussed as a single entity. One of the more meaningful observations of previous work centralizes around the lack of emphasis on the fundamental importance of musical tempo and step frequency. There are two schemas in which music can be synchronized with an individual's physical movements. Adopting computer science terminology of digital communication protocols, synchronization between two or more time-sensitive processes can easily be achieved with one process acting as the "master clock" and all other processes assuming the "slave clock" role, taking timing cues from the master. This analogy exactly describes the motor entrainment effects of RAS detailed previously. Therefore, in the time-sensitive process of rhythmic motor activity (e.g., running), rhythm can serve as a master clock while the human timing mechanism acts as a slave, or vice versa. Every previous system operates on the latter, based on the misguided premise that forcing music to synchronize to a runner is desirable. As a result, this method of operation fails to realize and harness the benefits of RAS in a physical training environment.

This is compounded by the observation that some studies (Masahiro et al., 2008; Oliver and Flores-Mangas, 2006) use a treadmill for subject running trials. Rhythmic auditory stimulation facilitates muscular synchronization in physical activity by providing a subconscious and psychological external timing mechanism. A treadmill, however, also acts as an external timing mechanism, but takes precedence given the physical interaction with the runner. Succinctly, rhythmic music and a treadmill both provide the runner with competing clock sources, where the psychological and physiological timing mechanism (music) will yield out of necessity to the physical timing mechanism (treadmill), enforcing a fixed relationship between velocity and step frequency of the runner. Similar behavior results when the tempo of a music stimulus is too distant from the individual's resonant frequency to allow synchronization. Rather than maintain a cadence that is uncomfortable or impossible, the runner disregards the musical stimuli entirely.

Additionally, and of equal importance, the previous work outlined above neglects to account for a semantic audio analysis mechanism. Each system (Bieber et al., 1995; Bieber and Diener, 2005; Masahiro et al., 2008; Oliver and Flores-Mangas, 2006; Wijnalda, 2005; Elliot and Tomlinson, 2006; Bowen, 2009; Ahmaniemi, 2007) operates on the assumption a global track tempo is previously ascribed to every track of interest, and the manner in which this is achieved is of no consequence to the system. Utilizing a global tempo for playback selection for running accompaniment, regardless of the derivation

method (human or computational), relies heavily on the assumption that the tempo of that track is sufficiently constant throughout its entirety. MPTrain makes mention of maintaining a time-dependent tempo vector to describe each track in the system's DML (at a sampling rate of 0.05Hz, or every 20 seconds), but does not consider this information in the selection of music. The argument is made here that the suitability of a track for running accompaniment is dependent on not only the tempo of that track, but also the stationarity of tempo and perceived beat strength. Tempo, a time-dependent percept, can varying in the course of a performance, and the degree of modulation defines its stationarity, or lack-thereof. Beat strength describes the subjective percept of felt beat, where electronic dance music tends to stylistically exhibit a strong beat versus that of rhythmically ambiguous avant-garde music. These characteristics converge to form a definition of motor rhythmic music, otherwise described as optimal music for use in RAS. It then becomes necessary to develop a mechanism capable of objectively quantifying these parameters and therefore identify songs that meet this classification.

3.2 Tempo Induction

Computational rhythmic analysis of musical signals falls under the umbrella of automatic music transcription and is composed of two overlapping subtopics. Beat–tracking is defined as identifying and marking the locations of beats in a musical excerpt, whereas tempo extraction, alternatively, aims to identify the tempo marking. As a result, both tasks are not without their

respective complexities. Beat–tracking must resolve perceptual issues, such as distinguishing between weak and strong beats or phase modulations. The idea of rhythmic phase modulation addresses the variety of reasons a beat may not properly align with the previously established temporal grid: the tempo could be changing or a recorded attack may have been performed slightly off-beat, intentionally or in error. Tempo extraction does not necessarily require a thorough analysis of the beat information of a piece to arrive at a tempo characterization, but the compact nature of the result natively increases the significance of performance errors. Additionally, whereas accurate beat–tracking will encode tempo modulations, tempo extraction cannot adequately describe this temporal variation with a scalar meter marking, or how a human would perceive this change, if at all.

The emphasis of this work predicates a specific exploration of previous work in the area of tempo extraction. However, for several reasons that will be expanded upon shortly, information describing the perception of tempo modulations is also vital to meet the goals of the system. Beat–tracking systems can, and have, implemented back-end processing stages to arrive at global tempo estimations. A third derivative approach, referred to as tempo induction, synthesizes these two related areas by establishing time-dependent perceptual tempo estimation.

Longuet-Higgins and Steedman made initial attempts at algorithmic beat detection as early as the 1970s for symbolic representations of Bach fugues

(Gouyon et al., 2005). Other score-based processing systems were developed throughout the next 20 years, employing heuristic and signal processing methods to extract tempo information. Comb filterbanks for periodicity estimation in rhythmic analysis first made an appearance in the work of Miller et al. in 1992 (Scheirer, 1998), and other oscillator mechanisms were explored throughout the middle of the 1990s (Scheirer, 1998; Scheirer, 1997; Large, 2000; Large and Kolen, 1994). Up until this point, all tempo extraction research was being conducted on symbolic, MIDI-style inputs (Allen and Dannenberg, 1995). Goto and Muraoka, in 1994 (Goto and Muraoka, 1994), and Scheirer, in 1998 (Scheirer, 1998), were among the first to develop successful beat–tracking algorithms designed to parse digital acoustic input data, generalizing computational machine rhythm. Ideas outlined in these seminal works encouraged the evolution of advanced methods, establishing a cross-pollination of system elements that has produced a diverse field of descendant beat–tracking and tempo induction algorithms.

In the process of performing rhythmic analysis, covering the topics of beat–tracking, tempo extraction and the like, modern algorithms assume a common two-step approach: establish a driving function and perform periodicity estimation. Systems developed early in the history of rhythmic analysis (Scheirer, 1998) showed that signal decomposition proves to be an essential preprocessing step in the derivation of a driving function, and can to some degree be considered a third, initial stage of the task at hand. Each stage is functionally

independent from the other, reducing an end-to-end algorithm to the optimization of separable components. Table 3 shows the mechanisms employed for the various stages from a sizable sample space of research. A genealogical review is presented, focusing on the genetic components and their first appearance in the pool, rather than an investigation of any specific "children."

Decomposition	**Driving Function**	**Periodicity Estimation**
Multichannel Filterbank	Spectral Flux	Autocorrelation Function
FFT–Banding	HAS-Modeling	Spectral Sum and Product
Wavelet	Local Maxima, Linear Regression	Dynamic Programming
None	Phase Difference (Group Delay)	Inter–Minima Interval
	Complex Spectral Flux	Heuristics (Rule–Based)
	Self–Similarity Matrix	Comb Filterbank
	Mid–Level Features	Temporal IOI Clustering
		Multiple Agents
		Hidden Markov Models
		Adaptive Oscillators
		Phase Locked Loop
		Particle Filtering
		Discrete Fourier Transform

Table 3. **Rhythmic–Analysis Algorithm Components** - Commonly used techniques for the significant system elements.

3.2.1 Decomposition Schema

While it is widely accepted that the human auditory system can be represented as a multichannel filterbank, the role of subband decomposition in computational rhythmic analysis systems is much more varied. Scheirer was the first to popularize the observation that the percept of rhythm is maintained when amplitude modulating white noise with the envelopes of as few as four subbands of an audio waveform (Scheirer, 1998). It is even proposed in this same work that subband components could be derived using time-domain bandpass filterbanks or grouped STFT bands, the former of which is actually implemented. The conclusion drawn from this observation motivates the notion that, if subband

decomposition aids in the human induction of rhythm, it will likely aid in the machine induction of rhythm as well. Tzanetakis, Essl and Cook introduced wavelet decomposition techniques for music transcription applications in 2001, incorporating elements of the system in genre classification work published the following year (Tzanetakis et al., 2001). Fourier transform banding, where the bins of an STFT frame are reduced to a collapsed representation by a given weighting function, first appeared in the beat–tracking system proposed by Goto and Muraoka as seven octave bands (Goto and Muraoka, 1996). More recent implementations have used triangular weighting functions to achieve a better estimate of tonal content (accounting for the spread of inexact bin alignment) (Klapuri, 2003; Harper and Jernigan, 2004; Klapuri et al., 2006; Ellis, 2007; Antonopoulos et al., 2007b; Antonopoulos et al., 2007a), in addition to arbitrary banding of FFT-bins (Hainsworth and Macleod, 2003;
Holzapfel and Stylianou, 2008). Various filterbank designs have been implemented, taking cues from the wavelet transform by employing pyramidal decompositions (Duxbury et al., 2004; Duxbury et al., 2002), HAS modeling via critical bandwidth approximation (Klapuri, 1999), or direct bandpass implementations. Linear phase response is often compromised in favor of low-complexity IIR bandpass filters, as seen in (Scheirer, 1997; Scheirer, 1998; Seppanen, 2001; Alonso et al., 2003b; Uhle and Herre, 2003).

3.2.2 Driving Function

Human audition can be accurately described as an iterative, real-time reduction of data on multiple orders of magnitude. A driving function, which encompasses onset detection, extraction, and the general emphasis of musically meaningful events, serves as a stage in the funneling of information from periphery to event encoding. Regardless of the domain, it is intuitive to represent rhythmic sound as a stream of information at a rate corresponding to the temporal accuracy of the HAS, from which periodicities can be inferred. However, several competing approaches have been undertaken in the arrival at a suitable function to drive a periodicity estimation mechanism.

Goto and Muraoka have, in the course of several published works, explored the viability of local gradient maxima detection in tandem with higher level features, such as drum detection (Goto and Muraoka, 1994; Goto and Muraoka, 1997a; Goto and Muraoka, 1995), chord changes (Goto and Muraoka, 1997b), and merged information (Goto and Muraoka, 1996). Gradient maxima provide can an advantage over decoupled one-dimensional subband maxima in identifying two-dimensional peaks in the time-frequency representation, although this effect is significantly diminished in wide-band, polyphonic recordings. Multiple early algorithms by Dixon, which would influence the landscape of acoustic rhythmic analysis, aimed to derive a detection function directly from a high-pass filtered version of the acoustic input via linear regression (Dixon, 1997; Dixon, 1999; Dixon, 2000; Dixon, 2001;

Dixon et al., 2002), employed subsequently by (Santos et al., 2003; Hainsworth and Macleod, 2003; Scaringella and Zoia, 2004). Human auditory system modeling, first employed by Scheirer (Scheirer, 1997; Scheirer, 1998) and followed by several derivative works (Seppanen, 2001; Tzanetakis and Cook, 2002; Alonso et al., 2003a; Klapuri, 2003; Alonso et al., 2003b; Uhle and Herre, 2003; Ellis, 2007; Harper and Jernigan, 2004; Klapuri et al., 2006), attempts to mimic the prolific human ability of rhythm induction. The technique, generally, follows the transduction of sound from arrival at the outer ear to auditory nerve encoding incorporating varying levels of detail in psychoacoustic properties, being, namely, the Haas precedence, frequency decomposition, and compression. Spectral flux, alternatively described as the difference in energy between adjacent STFT frames, characterizes both the energy envelope and the tonal content of an input acoustic waveform, and is summed across channels to form a singleton vector. This technique is a decoupled form of the gradient method used by Goto and Muraoka, employed by the algorithms detailed in (Jensen and Andersen, 2003a; Jensen and Andersen, 2003b; Laroche, 2003; Alonso et al., 2004; Davies et al., 2005; Davies and Plumbley, 2005a; Peeters, 2005; Kurth et al., 2006; Jochelson and Fedigan, 2006; Peeters, 2007; Chen et al., 2007; Alonso et al., 2007a; Alonso et al., 2007b; Dixon, 2007). Other less common driving mechanisms include self-similarity matrix processing (Foote and Uchihashi, 2001; Pikrakis et al., 2004;

Pikrakis and Theodoridis, 2007; Antonopoulos et al., 2007b; Antonopoulos et al., 2007a), complex spectral difference (Davies and Brossier, 2005; Davies and Plumbley, 2007), group delay (Sethares et al., 2005; Holzapfel and Stylianou, 2008), and mid-level feature extraction (Jensen and Andersen, 2003a; Jensen and Andersen, 2003b; Sethares et al., 2005), all of which have been explored more recently.

An important distinction must be made with regard to the derivation of a driving function. Some algorithms make a deliberate effort to detect onsets by applying criteria, such as a threshold, to some intermediary signal, which incurs both advantages and shortcomings. Applying detection criteria to a driving function produces a sparse vector, compactly described by its non-zero elements, thereby accelerating periodicity estimation and eliminating noise that can accumulate in successive stages. These benefits come at the expense of potential false negatives (missed onsets that are) and positives (passed onsets that are not), the importance of which depends entirely on the periodicity-estimation mechanism.

3.2.3 Periodicity Estimation

Once arriving at a driving function of rhythmically meaningful events, computational processing evolves from a perceptual to a cognitive task. Several, in particular early, beat induction algorithms have made use of multiple agent architectures that concurrently maintain individual estimates of the instantaneous tactus (Allen and Dannenberg, 1995; Goto and Muraoka, 1994;

Goto and Muraoka, 1997a; Goto and Muraoka, 1995; Goto and Muraoka, 1996; Goto and Muraoka, 1997b; Dixon, 1997; Dixon, 2000; Dixon and Cambouropoulos, 2000; Cemgil et al., 2000; Meudic, 2002; Dixon et al., 2002; Santos et al., 2003; Scaringella and Zoia, 2004; Harper and Jernigan, 2004; Dixon, 2007). These models typically drive a probability-based model of perceptual transitions, e.g. hidden Markov models, to account for discontinuous tempo modulations or corrections, in the event of an incorrect lock. Concisely, each beat agent maintains a hypothesis about its own reliability, and adapts temporally to information as it is received. In addition to the task of properly identifying beat periods accurately, the robustness of this mechanism depends heavily on the system's capacity to resolve competing tempo estimates.

A more perceptually motivated model is described by a bank of comb filters, otherwise referred to as oscillators or resonators, encouraged by psychoacoustic principles of rhythmic entrainment. Early work by Scheirer detailed the first successful implementation of a comb filters for rhythmic analysis (Scheirer, 1997; Scheirer, 1998), with subsequent adoption in a variety of different algorithms (Klapuri, 2003; Davies and Plumbley, 2005a; Davies and Plumbley, 2005b; Davies and Brossier, 2005; Davies et al., 2005; Klapuri et al., 2006; Kurth et al., 2006; Davies and Plumbley, 2007; Holzapfel and Stylianou, 2008). The comb filter difference equation is defined as the sum of a current input and a scaled, delayed output, causing reinforcement for

integer multiples, and factors, of the delay length. Comb filterbanks exhibit the unique behavior that a wide range of harmonics and partials are also reinforced as a direct result of the difference equation, mimicking multi-resolution aspects of human rhythm induction, but suffer from complexity considerations. Pulse train sampling rate and the number of distinct comb filters has a direct impact on the computational intensity and memory requirements of the implementation.

In a similar manner, the autocorrelation function is commonly used for periodicity estimation in beat–tracking systems. Natively a measure of self-similarity, the autocorrelation function will peak when a delayed version of the input signal also exhibits a large degree of similarity, as in the case of an equally spaced pulse train. While present in many of the aforementioned multiple agent systems, autocorrelation is employed in the systems detailed in (Foote and Uchihashi, 2001; Goto, 2001; Tzanetakis and Cook, 2002; Alonso et al., 2003a; Uhle and Herre, 2003; Alonso et al., 2004; Davies and Plumbley, 2005a; Davies and Plumbley, 2005b; Davies et al., 2005; Klapuri, 2003; Ellis, 2007; Peeters, 2007; Alonso et al., 2007a; Alonso et al., 2007b; Davies and Plumbley, 2007). While computationally more efficient than a comb filterbank approach, the autocorrelation necessitates the decision of several performance-influencing parameters. Developing a temporal estimate of tempo perception requires a windowing of the autocorrelation function, while the time duration encapsulating an equal number of beats is tempo dependent. Additionally, unlike the comb filterbank, rhythmic partials are

absent from the output of the autocorrelation function, while integer harmonics are passed with some regularity, depending on the length of the window and the signal content contained within it.

Less common periodicity estimation mechanisms include inter-onset interval clustering (Dixon, 1999; Seppanen, 2001; Dixon, 2001; Jensen and Andersen, 2003a; Jensen and Andersen, 2003b; Uhle and Herre, 2003), particle filtering (Cemgil and Kappen, 2002; Cemgil and Kappen, 2003; Hainsworth and Macleod, 2003; Sethares et al., 2005), spectral sum/product (Alonso et al., 2003b; Alonso et al., 2004; Jochelson and Fedigan, 2006; Alonso et al., 2007a; Alonso et al., 2007b), and dynamic programming (Laroche, 2003; Ellis, 2007). Inter-onset interval clustering, a time-dependent histogram method of identifying the most prevalent beat intervals, has, to some degree, fallen out of practice, while interest in particle filtering, a variant of linear prediction, has recently increased.

3.2.4 Alpha Systems

Of the work outlined, encompassing a good deal of iterative experimentation, there are a few state-of-the-art algorithms that perform significantly better than others in the field. Fortunately, select authors have collaborated in the past (Gouyon et al., 2005; McKinney et al., 2007) to directly compare various system designs, finding that Davies' (Davies and Plumbley, 2005b) and Klapuri's (Klapuri, 2003) algorithms perform the best at tempo extraction, approaching mean P-scores of 0.8 and 0.83,

respectively. Interestingly enough, both exhibit a degree of processing parallelism — multiple driving functions and subband decomposition, respectively — and a comb filterbank in the periodicity estimation stage (albeit employed as an ACF parsing mechanism is the former system). Though some work has underscored the importance of a suitable driving function (Davies and Plumbley, 2005b), this is the primary difference between Klapuri's algorithm and Scheirer's as compared in (Gouyon et al., 2005). The primary source of error in these alpha systems can be broadly attributed to harmonic errors (e.g., 2:1, 3:2, 4:3), the occurrence of which is possibly mitigated through the derivation of a cleaner driving function or an improved periodicity-estimation stage.

3.3 Areas of Improvement

Toward the development of a system connecting running cadence and musical tempo, with additional focus on track selection, previous research efforts exhibit shortcomings at different system levels. For clarity, there are two main stages of a running-music system: the runner in a music listening environment, and the content comprising that music listening environment. All prior research investigating the linkage between music and running, typically with the goal of music synchronization to human motion, focuses solely on the higher system level, creating a variety of functional disconnects.

Primarily, the significance of the human body as a dynamic mechanical system is effectively neglected, in particular the relationship between step frequency, stride length, and velocity. On occasion, the correlate between music

tempo and running velocity has been, in the past, attributed to the energy, or valence, of the music listened to during exercise (Davies et al., 2005; Gao and Lee, 2004). This operates on the notion that playing faster music will encourage a runner to run faster, and vice versa, by encouraging an increase in step frequency. In reality, except for sprinting, velocity is more a function of stride length than step frequency for moderate exercise, where step frequency actually increases with a decreased velocity (shorter stride length amounts to choppier steps). Almost more importantly, though, is the observation that there is actually a narrow step frequency bandwidth an individual can comfortably run within, determined by such context-dependent parameters as fatigue, physical dimensions, and fitness. Having addressed this topic, one case concluded that there is no correlation between cadence and music tempo because music at various tempi did not influence the step frequency of runners (Ahmaniemi, 2007). This conclusion is inaccurate, as the tempi of the various musical excerpts were well outside the suitable range for movement synchronization, shown in Figure 9. The tempo failed to influence runners because it did not fall within their resonant bandwidth. Developing a system that attempts to link physical movement with music must focus on the motion associated with rhythmic exercise.

That being stated, regardless of the manner in which music and running are algorithmically linked, any such system requires some form of semantic description of every track in a given DML. Previous music-running systems attend to this need by providing human annotated content to the music selection

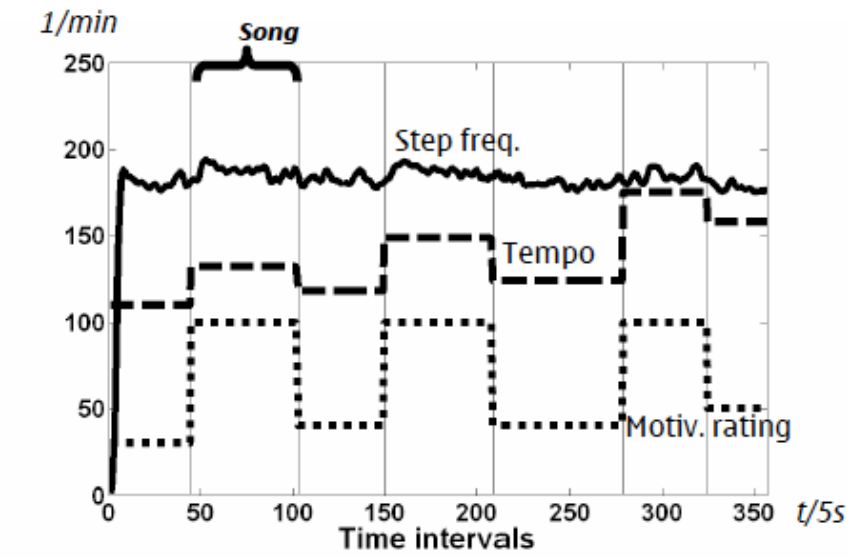

Figure 9. **Motivational Rating and Tempo** - A prior investigation into a possible connection between subjective rating and musical tempo, from (Ahmaniemi, 2007).

algorithm, a suboptimal solution for a few pressing reasons. Human annotation of musical content, while arguably the most robust means feasible, is somewhat impractical for implementation outside of the research environment. Even in the event that the manpower can be attained, be it through crowd sourcing or a single concerted effort like the Music Genome project (Westergren, 2009), a task as simple as the manual tabulation of a global track tempo estimate would take a daunting amount of time. This example lives outside of the debate of whether or not a global tempo estimate is sufficient information for a running-music system. For example, it is trivial to conceive two drastically different excerpts that have the same average tempo, or a case where a recorded piece of music has multiple movements at different tempi. The more common scenario is the presence of tempo modulation in a recording, either as a result of an expressive performance

or unintentional fluctuation, highlights the inadequacy of a global tempo estimate. As a result, all prior efforts in the development of a music-running system stand to greatly benefit from the advent of a computational rhythmic analysis component capable of describing music tracks in multiple relevant dimensions.

Shifting focus to the development of a machine-listening algorithm, this work necessitates computational rhythmic-analysis mechanism for the task of quantifying the "runability" of a track for running accompaniment and, more specifically, motor entrainment, which must be achieved with reliability and temporal precision. Previous algorithms place an emphasis on identifying the tactus of a track rather than the class of beat frequency, a more crucial descriptor for the target goal of RAS. As discussed earlier, the argument has been made that timing and rhythm is a quintessential musical component, and the tempo variance and evolution is integral in describing the motor–rhythmic nature of a music track. The general consensus in the research community is that application–specific rhythmic analysis algorithms will consequently perform better at that task than a generalized approach. Modern tempo extraction algorithms conventionally strive to accurately characterize all music, while no system has been explicitly designed to accurately parse content for a specific task. Therefore, a machine-listening algorithm is proposed with goals that fall between rudimentary tempo extraction and genre classification, to characterize the fundamental beat frequency class, perceptual salience, strength, and

temporal evolution (modulation). Ancillary design goals include improved accuracy, precision, and reduced computational complexity over other state-of-the-art systems where possible.

4

Proposed System

Prior neurobiological and music therapy research validate the use of rhythmic auditory stimulation as a rehabilitation technique. It is a logical evolution of thought that healthy individuals may also benefit by entraining body movement during cyclical physical activity to appropriate rhythmic cues. Concurrently, advances in automatic music transcription and music information retrieval have generated a solid foundation of digital music signal processing methods and techniques. Here, these two somewhat distinct fields are merged in the development of a system to automatically identify and create motor-rhythmic music playlists matching an individual's natural resonant frequency.

This work addresses a specific problem that sits at the intersection of several fields. The motivating force is the physical and psychological research that has led to an increased understanding of the benefits of music-accompanied motion. Upon internalizing this knowledge, an overriding question naturally evolves: "What music best suits this task?" Framed in this context, a system capable of identifying and retrieving music for suitable running accompaniment is merely a search process. As a first step, it is necessary to define the specific criteria with which to determine the relevance of an item in the search space to the query. Once the query can be accurately represented, the search itself must be subsequently conducted.

At the highest level of abstraction, the system executes in three distinct stages. It is first necessary to determine an individual's current natural resonant frequency. Once identified, that person's digital music library is parsed, returning the most appropriate tracks for RAS. These resulting tracks are then casted to the previously identified user-specific resonant frequency, which can be used as an accompanying playlist during subsequent runs. Key elements of system design include the capability for potential wide-scale deployment via automation and wearable sensor technology, an improved application-specific tempo-induction algorithm, and phase vocoder tempo correction to create unique tracks for personalized RAS-playlists.

4.1 Overview

As stated, the system operates in a three-stage process. A wearable accelerometer-based sensor is designed to non-invasively measure an individual's physical activity during unconstrained, free-field running. Data collected from this sensor during a run is processed offline by an algorithm developed as a derivative of the primary music analysis stage. Though it precedes a thorough dissection of the music-analysis algorithm, these two analysis mechanisms are extremely similar in implementation and discussion of the kinematic analysis will only address the differences. Proper processing of this kinematic data produces the resonant frequency of the running event used to drive a subsequent stage of the system.

Building upon previous work in tempo induction, a digital music analysis

algorithm is developed to identify the time-dependent tempo and beat strength of tracks in a DML. Decomposition of an input waveform is achieved through the implementation of a 22-band maximally decimated filterbank, with bandwidths closely approximating those of the cochlea in the HAS. Quasi-symbolic onset trains are extracted from each subband component, summed across all channels, and post-processed by a sliding window motivated by the Haas precedence effect. This final onset vector is fed into a bank of modified comb filters to act as a means of periodicity estimation, behaving similarly to the manner in which humans perceive motor-rhythms. The corresponding track tempo feature vectors are extracted and accumulated in a separate database for fast indexing and retrieval for current and future queries.

Original user cadence data is used in conjunction with the feature vectors to seed the music-retrieval algorithm. Appropriate motor-rhythmic music is identified as the maximization of beat strength and tempo stationarity by the maintained database of tempo feature vectors and compiled as a subset of a given DML. This subset of tracks is adjusted to match the appropriate cadence via phase vocoder time warping. An individual uses the resulting time-modified playlist during the next running event, and continued execution of this system loop may conceivably produce similar performance gains to those seen in previous RAS studies. High–level system architecture is shown in Figure 10.

An important distinguishing characteristic of this system from previous work is the use of a short-time static, macro-delay feedback. Implementation of

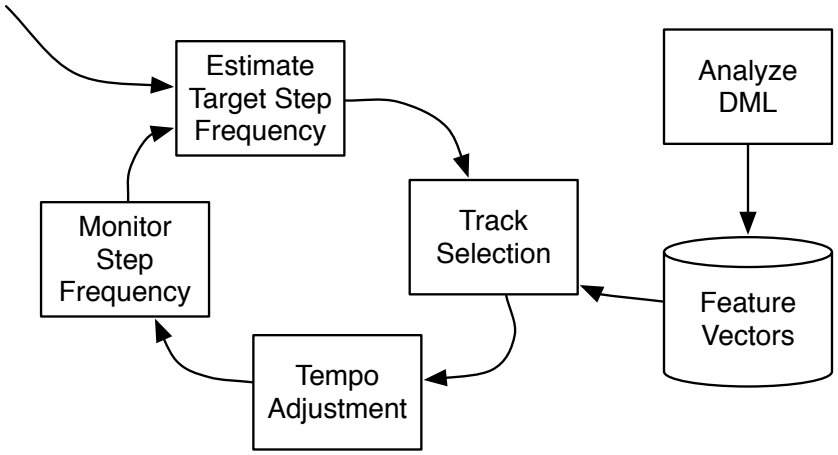

Figure 10. **System Architecture** - High-level process diagram of the proposed system.

RAS requires cadence stability, and it is undesirable for cadence parameters to significantly change on a minute-to-minute basis (Thaut, 2008). The primary goal of music accompaniment is to entrain movement for long-term advancement and increase of physical performance, in the form of endurance, stamina, and velocity. At its heart, the proposed work is a MIR system with music therapy motivations, and its surface benefits of enjoyment and entertainment, while important, are secondary. The stance is taken that continual adjustment of a target entrainment tempo conflicts with the goals of a RAS-based system, and the effective performance sampling rate occurs at large intervals. To these ends, physical performance is charted at the time scale of discrete running events, and the system operates in a sample-process-synthesize-and-repeat methodology. Both kinematic data and musical content are processed in what is perceived by the end user as an offline process, as they are acquired.

4.2 Kinematic Analysis

The ability of the proposed system to characterize an individual's natural running behavior across a variety of environments is crucial to establishing its validity. As mentioned previously, this cannot be accurately achieved by observing treadmill running, due to the influence of external forces on the relevant physical and psychological timing mechanisms. Cadence estimation of treadmill running is also somewhat trivial, and can be disregarded in the design of a sensor mechanism. It is important to note that most previous work investigating the linkage between running and music through the relationship of cadence and tempo use footstep detection to determine inter-onset intervals and, as its reciprocal, cadence. Direct heel strike detection is, however, a suboptimal path in the design of a biometric sensor in order to meet the criteria of non-invasiveness. Monitoring impacts directly requires that a sensor must exist between an individual's foot and the contact medium (track/pavement/grass), requiring the manipulation of the ground (impossible) or the runner's shoe (impractical). Following reliable data acquisition, the kinematic data can be analyzed to arrive at, among other characteristic behavior, the resonant frequency of the running event.

4.2.1 The Navi Cadence Monitor

Toward the design of a truly environment-agnostic sensor, it is more generalized to monitor the motion of an object and process the resulting signal to determine periodicity, or rhythm, of that movement. Abstracting the kinematic

measurement mechanism comes with the added benefit of gathering more complete information about the entire body. For example, previous medical research has to date investigated the reliability of using accelerometers to characterize gait disorders (Moe-Nilssen and Helbostad, 2004; Kavanagh and Menz, 2008). Activity monitors, often used in energy expenditure research, directly measure acceleration components, but provide no insight or analysis into the raw measured data. Existing literature reaches a common agreement that the lower back, an area closely approximating the center of gravity for a human body, is the optimal placement for a single-sensor system.

Synthesizing these observations, a wearable activity monitor, *Navi*, is designed and assembled employing a tri-axial accelerometer capable of sensing force in each orthogonal direction (Humphrey and Leider, 2009). A resultant magnitude vector is calculated from each dimensional component to resolve orientation issues at the expense of additional direction information. Opting for this loss of information is decided in the interest of increasing the recording lifetime of the monitor by a factor of three. For the purposes of this work, however, a loss of information is an acceptable compromise. Acceleration is sampled at a frequency of approximately 250Hz with 10-bit precision, and logged to a 2GB microSD card as a human-readable ASCII text file of integer values.

The initial prototype, shown in Figure 11, used a Microchip PIC24FJ32 microcontroller to handle handshaking between a digital H48C accelerometer and a Logomatic v.2 data logger, developed by SparkFun, Inc. The final prototypes

used in the testing phase alternatively incorporate a ratiometric tri-axial accelerometer, reducing the footprint of the device and improving reliability. These sensors can measure upwards of 6G of force in each direction, providing greater dynamic range for measuring kinematic data. At the current data rate, the Navi is capable of recording data for over 200 hours, surpassing a full discharge cycle of the attached 100mAh Lithium-polymer battery. An example signal measured by the device is shown in Figure 12.

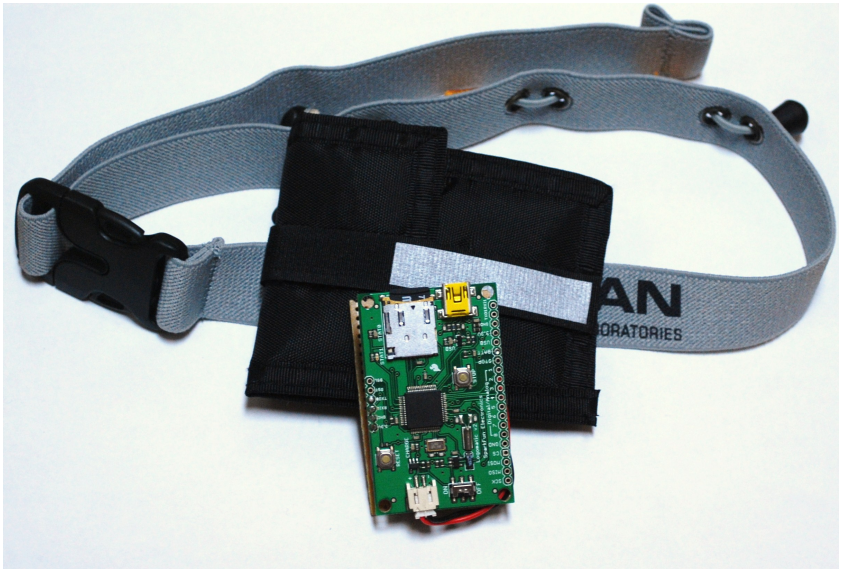

Figure 11. **Navi Cadence Monitor** - The developed prototype and race-belt used in subject testing.

4.2.2 Resonant Frequency Estimation

Over the course of a specific running event, there exists a resonant, or fundamental, frequency of motion that an individual naturally gravitates toward. Harmonic analysis of the resultant acceleration signal is necessary to characterize the spectral content of this cyclical activity signal and identify its fundamental.

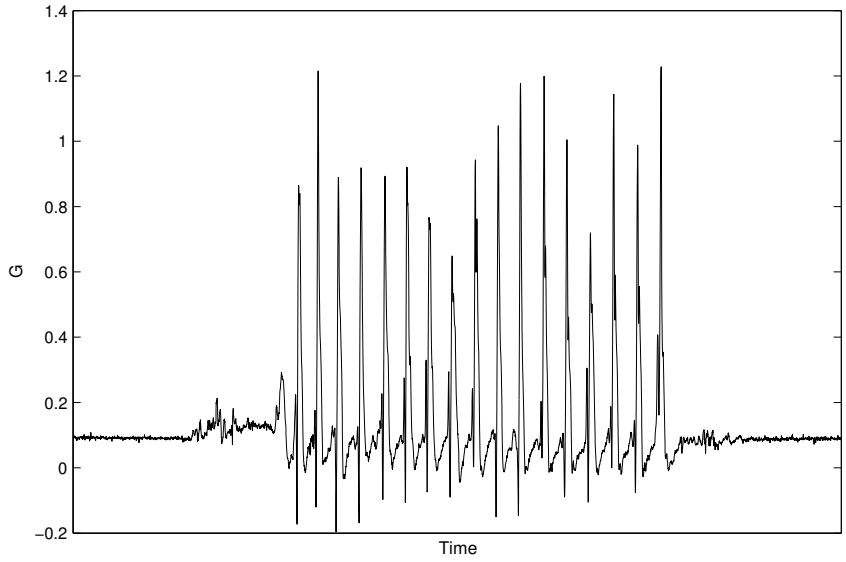

Figure 12. **Kinematic Data** - Visualization of the signal measured by the Navi activity monitor.

As will be addressed in far greater detail shortly, kinematic data processing employs a similar, derivative system of the one used in the analysis of the DML. There are very tangible benefits to this approach, particularly the uniformity in the periodicity estimation process. Resulting tempo curves from both the kinematic data and the corresponding musical content can be directly compared to observe the presence of motor entrainment in the course of running events. Estimation nuances that manifest in the analysis of one data type will naturally be present in the other. Furthermore, there is some evidence for a correlate between physical motion and musical tempo in a conceptual sense (Friberg and Sundberg, 1999), and the momentum of music and movement can be observed with similar mechanisms and models.

The two systems deviate in pre-processing approaches, with kinematic data being arguably a less complex waveform. There is little need for

decomposition of the kinematic signal since it is recorded at 250Hz, initially demonstrating a defined envelope. Being a resultant acceleration vector, it is desirable to negate the effects of gravity by removing the DC offset (steady-state value). Additionally, the waveform can be simplified further by only concerning the analysis with increases in acceleration over that of the constant pull of gravity. In other words, an acceleration sensor can only record magnitude values less than 1G (the force of gravity) in a free-fall, such as in the period between the apex of a runner's stride and the next heel strike. After removing the gravity-bias, half-wave rectification (HWR) serves to emphasize meaningful events in the kinematic data. These events can be coarsely described as onsets, in keeping with the parlance of common tempo induction methods, and extracted by filtering the waveform with a 50 ms Canny Filter, followed by HWR. This produces a vector of onsets candidates, on which periodicity estimation can be directly performed. In the interest of identifying only the fundamental spectral components of movement, a 100ms non-causal sliding window (50ms forward, 50ms backward) is applied to the onset candidate vector, allowing only the strongest peaks to survive over the length of the window as onsets. Realistically, two related factors motivate the application of a peak-picking sliding window. First, physical limitations cap step frequencies below 5Hz (300BPM), and is accounted for in the length of the window. As a result, the fundamental frequency will also be well below this frequency limitation.

Once an onset vector is obtained, periodicity estimation is then achieved

utilizing a modified comb filterbank (MCFB). A thorough discussion of this stage is reserved for the next section to maintain continuity of the more conventional tempo-induction context. Nonetheless, harmonic analysis via the MCFB produces two time-dependent vectors, representing instantaneous rhythmic frequency (tempo, ω) and beat strength (salience, β). Resonant frequency is determined from these two vectors as follows: a threshold is applied to beat strength as a percentage, μ, of its mean and used to logically index the tempo vector. This compact representation of time-dependent tempo comprises the strongest frequencies present in the course of the run, which only occur after periods of relative stability. A histogram of this compact tempo representation implements weighted voting of harmonic content. Peak-picking arrives at a single resonant frequency, the reliability of which can be described by the spread about the peak and its relative energy (percentage). This process is explained in greater detail in the following section.

4.3 Computational Rhythmic Analysis

Building primarily upon the work of (Tzanetakis and Cook, 2002; Tzanetakis et al., 2001; Scheirer, 1998; Klapuri et al., 2006), an application-specific tempo-tracking algorithm is developed. Whereas conventional tempo extraction aims to identify a singular, global tempo and beat-tracking aims to identify the locations of individual beat events, a hybridized approach is taken to track the evolution of tempo through the course of a one-dimensional waveform with the goals of establishing temporal significance of tempo.

Expanding, in identifying and selecting music for RAS, it is not enough to know that the desired tempo is present in a track, but also the strength of this percept and its variation, or more accurately, the lack thereof, over time. Good design of an algorithm suited to this task should directly incorporate perceptual models of tempo salience and characterize the temporal evolution of tempo.

The developed algorithm incorporates four stages, inspired by the modern understanding of human rhythm perception and psychoacoustics, illustrated in Figure 13. An input digital acoustic waveform is decomposed into twenty-two maximally decimated subband components, closely approximating the critical bandwidths of the cochlea. Onsets are extracted from each subband waveform via further computational modeling of the human auditory system and merged to generate an onset candidate pulse train. Candidates are deemed onsets through a sliding-window process examining relative intensity and proximity, again inspired by psychoacoustic knowledge. This single onset train is subsequently measured for periodicity relationships through a modified comb filterbank, mentioned previously, generating a two-dimensional mapping of time versus oscillator period. From this tempo map, a rhythmic feature vector is created to characterize the behavior of the track.

4.3.1 Decomposition

In one of the first reliable automatic beat-tracking algorithms to analyze acoustic data directly, Scheirer correctly observed that an initial frequency decomposition stage is crucial to a computational model of rhythm perception. It

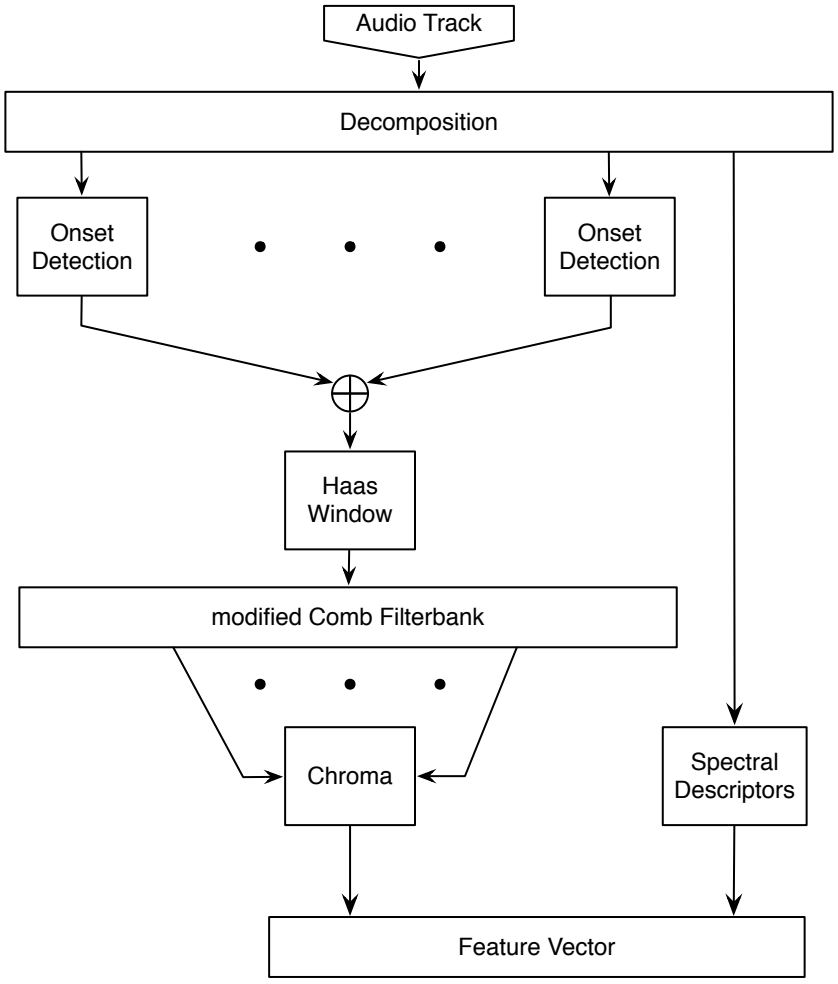

Figure 13. **Rhythmic Analysis Diagram** - An illustrated view of a computational model of human rhythm perception.

is postulated through empirical evidence however that system performance is not particularly sensitive to the design and implementation of the filterbank. Intuitively, this seems contradictory to the highly redundant nature of the human auditory system. When perceiving rhythm, human listeners take advantage of many other intrinsic abilities of the HAS, including stream segregation, timbral recognition and pitch tracking. These highly parallel processes provide further information to the human listener that aid in discerning the tatum and tactus elements in a complex polyphonic signal. Simply put, the ideal computational model of rhythm perception incorporates several other aspects of machine listening. The argument can be made that a rhythm-analysis algorithm lacking similar faculties of a human counterpart cannot perform as robustly. In lieu of a more thorough preprocessing front-end, frequency decomposition serves as a coarse, yet computationally efficient, mechanism to parse wideband rhythmic information.

A brief conceptual example demonstrates the importance of this preprocessing stage to overall machine rhythm perception. Consider an acoustic waveform of a monophonic sinusoidal melody that periodically alternates between two tones, a whole-step (in 12-TET) apart. Assuming that the amplitude of the voicing stays constant for both tones, the temporal envelope of the waveform is also constant, regardless of whether or not the transition between tones is continuous or not. Should both of these tones fall inside the same filterbank band, they will be merged into the same musical event. Whereas

a human listener would easily discern between the two tones and perceive a rhythm, a computer model using only a filterbank for segregation will not. Alternatively, tonal or timbral onsets are not necessarily indicative of motor-rhythmic music, but rather transient energy that can be traced to the transduction of acoustic events in the cochlea.

With this in mind, a higher-resolution filterbank is presented to decompose the acoustic waveform into a more accurate representation of motor rhythmic music perception. Approximating the critical bandwidths of the cochlea, a multi-level dyadic filterbank is designed to produce twenty-two maximally decimated channels, as shown in Figure 14.

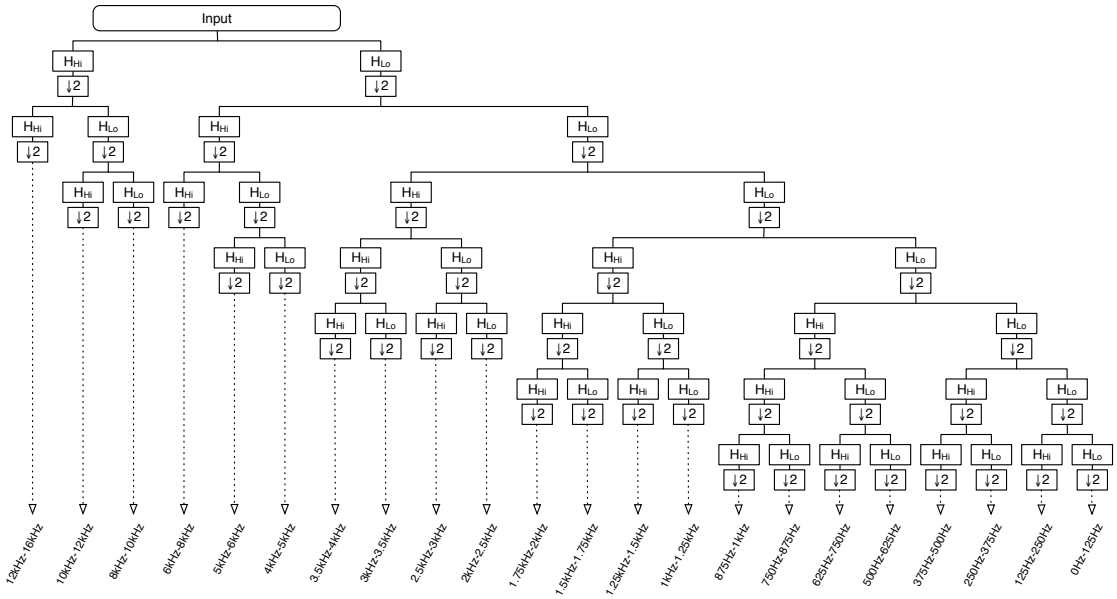

Figure 14. **A Perceptually–Motivated Dyadic Filterbank** - Diagram of the multi-rate decomposition of an input audio waveform using two complementary half-band filters.

There are several noteworthy advantages in this signal-decomposition approach. While the desire for a high-resolution filterbank that models human

perception is in no way a new development in machine listening, massively multi-channel implementations such as the gamma-tone filterbank are computationally expensive. For an increase in the number of channels, gamma tone decomposition produces a linearly increasing amount of information as each band maintains the same time resolution as the original waveform. This information redundancy subsequently increases the computational intensity of the decomposition and therefore the execution time. Alternatively, a dyadic filterbank cascades pairs of half-band filters and successive downsampling operations, such that the amount of input and output information are equal. The designed implementation is similar to pyramidal wavelet decomposition, with the significant difference that, at purposeful branches, the output of the highpass half-band filter is also decomposed. For an input sampling rate of 32kHz, seven decomposition levels are required to closely approximate the critical bands of the cochela. Selection of 32kHz as the sampling rate is not an arbitrary choice. Mathematically, it decomposes very cleanly by half-band divisions, with 256 as its largest power-of-two factor. Therefore, when processed in a conventional frame-by-frame basis, an input vector with a power-of-two length greater than 256 will decompose into vectors that are also powers of two, shown in Figure 15.

There are important issues that must be addressed with regard to half-band filter design. A decomposition of this method requires reasonable selectivity of the half-band filters. Downsampling will incur some degree of aliasing, and if the stop-band attenuation of these filters is not severe enough,

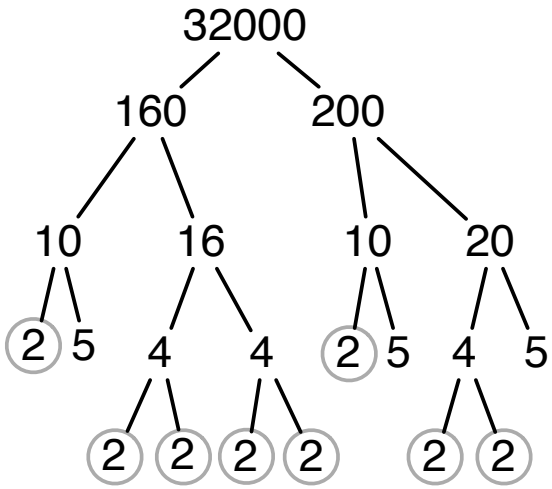

Figure 15. **Sampling Rate Factorization** - A sampling rate can cleanly support a number of decimations equal to the number of twos it has as factors. A sampling rate of 32kHz can be decimated 8 times.

aliased frequency content will propagate through the entire filter structure due to its cascaded nature. It is undesirable, however, to use an IIR filter to achieve these ends. Filter design forces the trade-offs between transition width passband to stopband and its other characteristics, namely significant phase distortion, which can cause errors that, again, consequently propagate. It is essential to keep in mind that the purpose of the filterbank is to extract the onsets of frequency subbands for an input waveform. Phase errors and therefore, by definition, group delay, accrue through successive levels and, left uncorrected, become non-trivial at the output.

For this reason, FIR half-band filters are designed using the db40 (N = 80) Daubechies' coefficients as a basis. The low-pass coefficients are invertible, producing the corresponding high-pass filter, and normalized to unity gain. It is also important to note that the low-pass coefficients impose a group delay

approximately equal to half the filter length. As previously mentioned, special attention must be given to phase handling so that transient behavior is optimally preserved in the decomposition process. The group delay difference between filter outputs is initially adjusted by convolving the high-pass coefficients with a half-length delay, reducing group delay discrepancies to roughly three samples across all frequencies. To correct this final phase distortion, an all-pass filter is designed for each corresponding filter shape to yield an effectively constant (sub-sample precision) group delay. The magnitude and group delay responses of the half-band filter pair are shown in Figures 16 and 17, respectively.

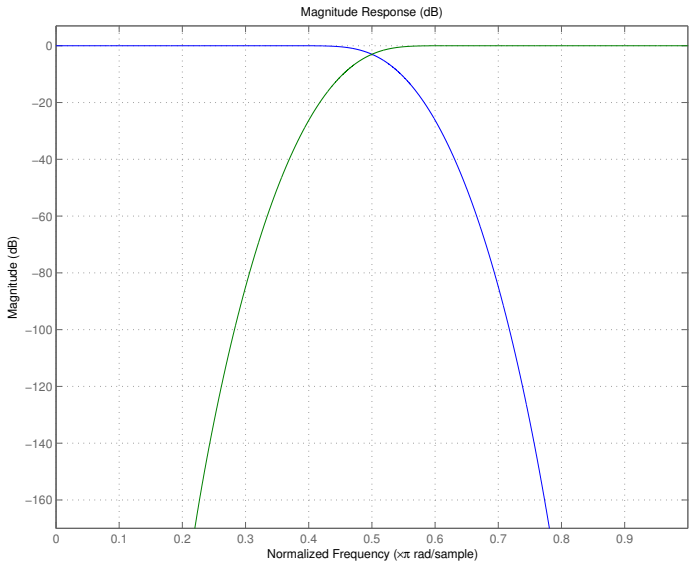

Figure 16. **Magnitude Response** - Half–band Filter Responses for the Low (blue) and High (green) filters.

4.3.2 Onset Detection

Following tested time-domain onset detection strategies in (Scheirer, 1998; Tzanetakis et al., 2001; Klapuri et al., 2006), the human auditory system is

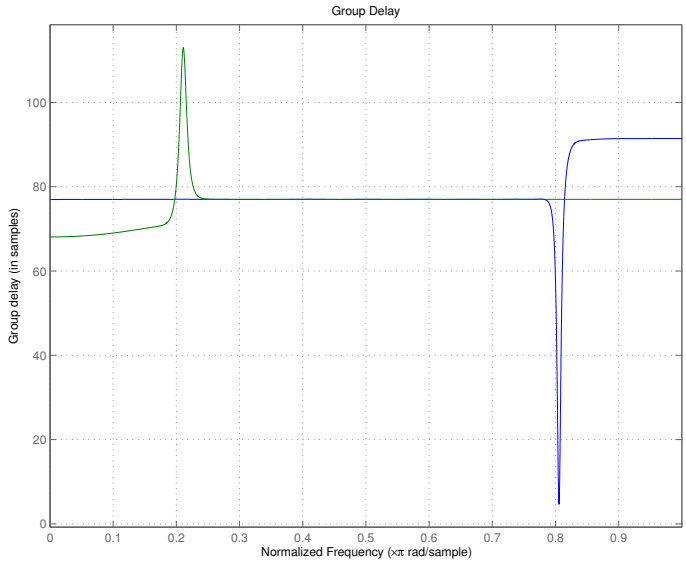

Figure 17. **Group Delay** - Half–band Filter Responses for the Low (blue) and High (green) filters. Note that the phase distortion in each falls within the stopband region of the corresponding magnitude response.

modeled to extract onsets from the decomposed subbands. Computational rhythm perception algorithms must extract musical pulse information before estimating the periodicity of an audio waveform. Complications arise from the fact that musical pulse information is derived from all three domains of music rhythmic, tonal, and timbral often requiring a different tailored system optimized for each task (pitch tracking benefits from tempo induction, etc). In lieu of a more complex mechanism to parse audio, generating symbolic information, onset detection is performed as a coarse approximation to this end. Importantly, while the distinction between pulse detection and onset detection must be made, the end application of this particular tempo extraction transforms this shortcoming into functional element of the system. Optimal motor

entrainment accompaniment, or motor-rhythmic music, exhibits, by definition, strong pulses that coincide with note onsets; music the algorithm aims to identify is the music that will produce the best results. Excerpts that maintain a weak beat or no beat at all are not desired as a final output, and this style of analysis will reflect as much.

The basic onset detection process is diagramed in Figure 18 and draws from modern onset detection methods. Each decomposed subband, $X_k[n]$ for the k^{th} subband, is half-wave rectified and low-pass filtered, modeling the transduction mechanism of the human ear. A 175 millisecond half-Hanning window, $W_k[n]$, is used to smooth the HWR subband signals, from Eq. 1, producing the amplitude envelope of each subband, $E_k[n]$ via Eq. 2. Multi-resolution decomposition results in seven distinct sampling rates, and therefore seven distinct half-Hanning window lengths. This discrepancy is resolved after low-pass filtering via decimating to a common sampling rate of 250Hz. Transient behavior is accentuated by performing μ-law compression on the extracted envelope signal, producing $E_{C_k}[n]$ in Eq. 3. Motivated by the system in (Alonso et al., 2004), the Canny operator, commonly used for edge detection in image processing defined in Eq. 4, is used in place of a 1st-order differentiator to extract transient behavior of the compressed, downsampled envelope, denoted by $C_k[n]$ in Eq. 5. As a penultimate step, the Canny-filtered envelope is half-wave rectified, resulting in the signal, $O_{C_k}[n]$, referred to as a vector of onset *candidates*, a term coined in (Klapuri et al., 2006).

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

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

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

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

$$C_k[n] = \sum_{i=-L}^{L-1} E_{C_k}[n] * C[i-n] \tag{5}$$

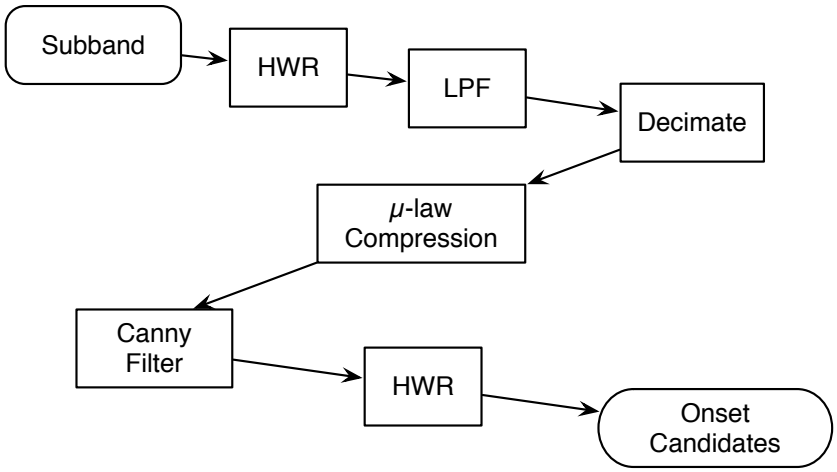

Figure 18. **Onset Detection** - Onset Detection is performed on each decomposed subband before being merged against channels.

The finite impulse response half-Hanning window exhibits a desirable low-pass filtering characteristic that can adapt to transient behavior with minimal phase distortion, performs comparatively to the HAS, and exhibits a

cut-off frequency at roughly 10Hz. It is also worthwhile to highlight that frequencies in the range of 0.5-5Hz, or 30-300BPM, are perceived as being rhythmic in nature (Thaut, 2008) with a perceptual gravitation toward beat intervals of 100BPM. Whereas the decomposition stage aimed to preserve phase information in the best manner possible, the non-linear phase behavior of this low-pass filtering operation is acceptable as a psychoacoustic model. While the rhythmic onset behavior of interest resides entirely in the frequency range below 20Hz, the fundamental tempo, over-sampling of each decimated envelope signal at 250Hz allows for the necessary time resolution to achieve accurate temporal localization of onset events. Rhythmic "partials" and phase information are integral components in complex rhythms with human temporal perception accurate to about 5 milliseconds, but narrowing the meaningful spectra facilitates periodicity estimation.

The Canny operator is a difference-of-Gaussians filter that offers the added benefit of high-frequency attenuation over a simple differentiation operation. Given the necessary over-sampling of subband envelopes, it is intuitive that a single zero high-pass filter is a sub-optimal choice to detect transient behavior. High frequency noise is passed with minimal attenuation and can lead to accumulated error in subsequent processing. Onset information in the amplitude envelope really only lives in the frequency content between 0.5-5 Hz, and it is ideally this range that we aim to extract with minimal distortion in both time and amplitude. A length-13 Canny operator with a Gaussian kernel

sigma value of 3, expressed in Eq. 4, achieves these goals quite well. Periodicity estimation mechanisms are sensitive to noise, and additional efforts must be made to generate clean onset pulse trains.

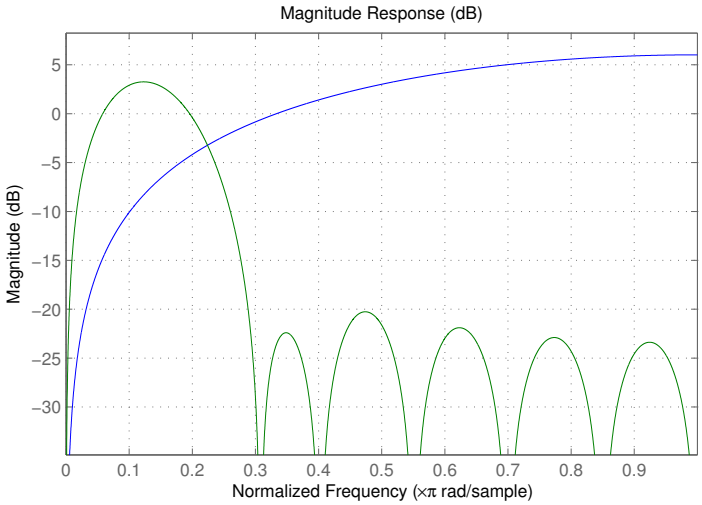

Figure 19. **Onset Extraction Filters — Magnitude Response** - Characterization of the conventional 1^{st} order differentiator (blue) and the implemented Canny filter (green). Note the attenuation of high frequency noise for the Canny filter.

In a continuing effort to de-noise the generated pulse train of onsets, a sliding window is implemented to model the precedence effect and arrive a final, comprehensive onset vector, $O[n]$. HAS research has shown that sounds occurring within close temporal proximity (10ms forward, 40ms backward) are perceived as a single sound. Time-domain masking is achieved by passing only the most intense peak over the 50ms window, followed by stretching the resulting impulses with a zero-order hold filter. Unsuppressed, this small-signal noise can accumulate over time, raising the noise floor of mCFB output, thereby impeding the reliability of the following periodicity estimation stage. Concisely, each step in the onset detection process aims at producing an increasingly pure

impulse-like signal. However, these pure pulses represent frequency content at the Nyquist rate, in this case being a 125Hz signal. Returning to a previous statement, higher temporal resolution is maintained to preserve phase information only, and these transient impulses are only representative of lower frequency content. This lower frequency content is synthesized by elongating each pulse over a 50 millisecond window, while maintaining the same time resolution. By this approach, the subband onset candidate vectors $O_{C_k}[n]$ are summed across all channels, emphasizing the compensation of phase distortion to produce a resultant candidate vector, $O_C[n]$. A final onset vector is obtained by filtering $O_C[n]$ with the sliding window algorithm ultimately arriving at the extracted onset pulse train, $O[n]$. Figure 20 shows the extracted onsets $O[n]$ superimposed on the corresponding input waveform $X[n]$.

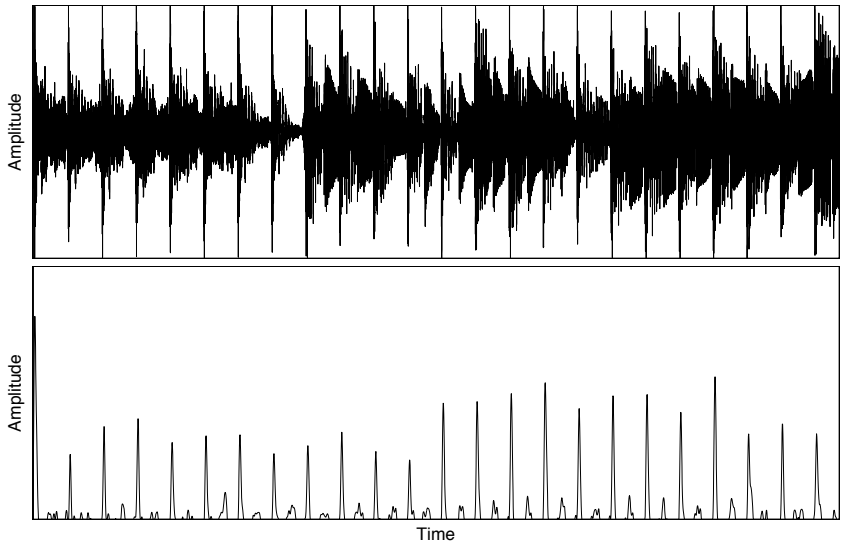

Figure 20. **Audio and Detected Onsets** - For a given input waveform (top), an onset signal (bottom) is calculated accordingly and plotted on the same time scale.

4.3.3 Periodicity Estimation

Spectral analysis comprises a significant portion of digital signal processing theory, and there exist, with their respective nuances and shortcomings, a wide variety of thoroughly understood techniques to perform this task. Interestingly, banks of comb filters, or oscillators, naturally lend themselves to the task at hand with regard to both biomimetic design and algorithmic behavior. Previous studies of human beat perception show that weakly coupled oscillator banks have significant merit as a computational model of rhythm perception for neuromuscular entrainment. In this manner, oscillator banks perform, in addition to tempo induction, beat strength and variation quantification on a pulse train input. Steady, salient pulses will generate positive reinforcement in the comb filter bank and a clearly defined output. Pulses lacking interval consistency, or rather exhibiting temporal deviation, regardless of perceptual salience, will not generate a strong time-dependent oscillator bank output. This mechanism thereby natively identifies motor-rhythmic music by distinguishing inputs that generate the most intense responses.

With respect to computational theory, oscillator banks are better suited to produce a time-dependent characterization. Fourier analysis provides a more global synopsis of a signal, and autocorrelation assumes an arbitrary degree of signal stationarity. Scheirer also makes the insightful observation that, though the autocorrelation function does share some similar behavior with an oscillator bank, the ACF fails to accurately describe the rhythmic harmonic structure in a

manner consistent with the rhythm-perception abilities of a human listener. With minimal effort, a listener, and particularly a musician, can hop between tempo octaves (cut-time or double-time) and, depending on skill level, small integer ratios, such as a 3:2 (dotted quarter over a felt four). While these percepts should be some degree weaker than the salient tactus, they should nonetheless be present in a computational estimation of rhythm as a means to quantify the existence and complexity of polyrhythms.

At a sampling rate of 250Hz, a bank of 575 parallel comb filters, with singularly increasing delay lags from 0.4-5Hz (24 to 300 BPM), is utilized in the periodicity analysis of the cumulative onset pulse train. Increasing the period interval linearly causes an exponentially widening set of frequency-tuned comb filters. Whereas Scheirer implemented a bank of log-spaced oscillators, the efficiency earned as a result is now compromised for increased resolution. Importantly though, as will be soon addressed, tempo estimation via decoupled parallel oscillators offers various opportunities for significant optimization.

The previously described onset vector, $O[n]$, is fed into the comb filterbank (CFB) identically across parallel channels. For a bank of K oscillators, each channel k is uniquely defined by the difference equation given in Eq. 10, where α is a scalar gain values and T_k corresponds to the time lag of the filter. Memory requirements and computational complexity, in a direct form implementation of the comb filter, increases linearly with T_k. As seen in the magnitude response plot, a comb filter passes all octaves of a fundamental

resonance equally, in addition to the DC-component, both of which are undesirable behaviors. Human perception does not consider all octaves of rhythmic information with equal weight, exhibiting a preference for tempi around 100BPM (1.66Hz), and noise in the input signal will accumulate in each comb filter, respectively. Considering a pure pulse train, every integer multiple causes the comb filter to resonate at an equal energy. For example, a pulse train with a period of T will pass through its corresponding-delay comb filter as two pulse trains with a period of $2T$, 180 degrees out of phase. While the human perception of rhythm is fractal by definition, a click track with a tactus and tatum of 150 BPM will be perceptually salient at 75, 150, and 300 BPM, likely varying in intensity. Weighted perception curves have been explored in tempo induction algorithms with notable success in correcting octave errors (Klapuri et al., 2006).

$$y_k[n] = (1 - \alpha) * x[n] + \alpha * y_k[n - T_k] \tag{6}$$

Rather than incorporating this fact as a post-processing heuristic, the comb filter bank itself is directly modified. The bandpass nature of Canny filter suits this task well, and is implemented in cascade with the comb filter difference equation, Eq. 7, with a magnitude response diagramed in Figure 21. Suppression of the DC-component is apparent in the magnitude response of the filter, and harmonic content in excess of human rhythm perception are effectively suppressed. Benefits to this approach are readily apparent in comparing Figures 22 and 23. Having obtained more desirable behavior from the modified base

oscillator, the observation is made that the numerator and denominator coefficients of the difference equation are separable. The FIR Canny filter is independent of the channel k on which it operates, and can be extracted as preprocessing step, prior to the now modified CFB (mCFB), while maintaining the same effect on the output.

$$y_k[n] = (1-\alpha) \sum_{i=-L}^{L-1} x[n] * C[i-n] + \alpha * y_k[n-T_k] \tag{7}$$

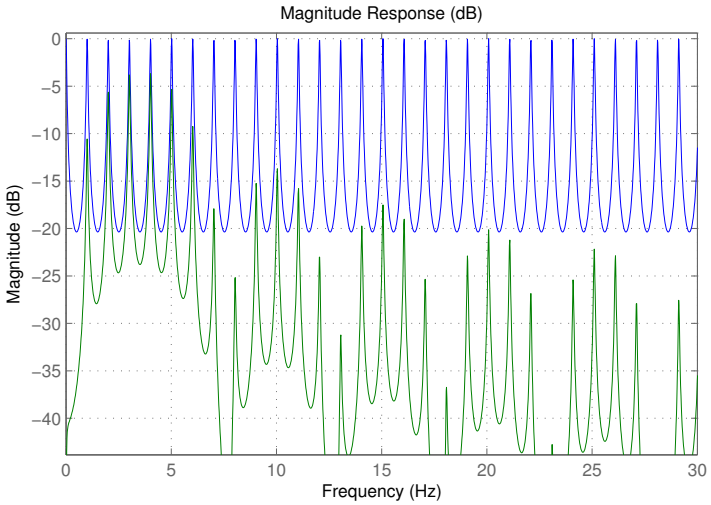

Figure 21. **Comb Filter Magnitude Responses** - Spectral Behavior of the original (blue) comb filter and the Canny-augmented (green) comb filter.

Instantaneous tempo estimation is performed by directly summing the energies of the comb filter delay states across channels in the mCFB. Specifically, an oscillator's output at time t is squared to generate an energy signal, and summed over a variable length window corresponding to the lag time of the oscillator. This filtering process can be conceptualized as a set of one-dimensional normalized averaging filters of increasing length, and is defined in Eq. 8.

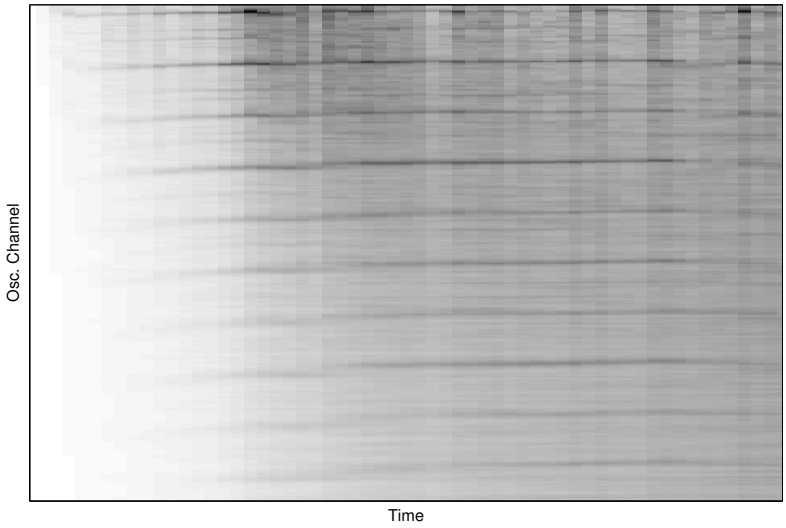

Figure 22. **Comb Filter Tempogram** - Periodicity estimation across K oscillator channels based on the original comb filter equation, 10

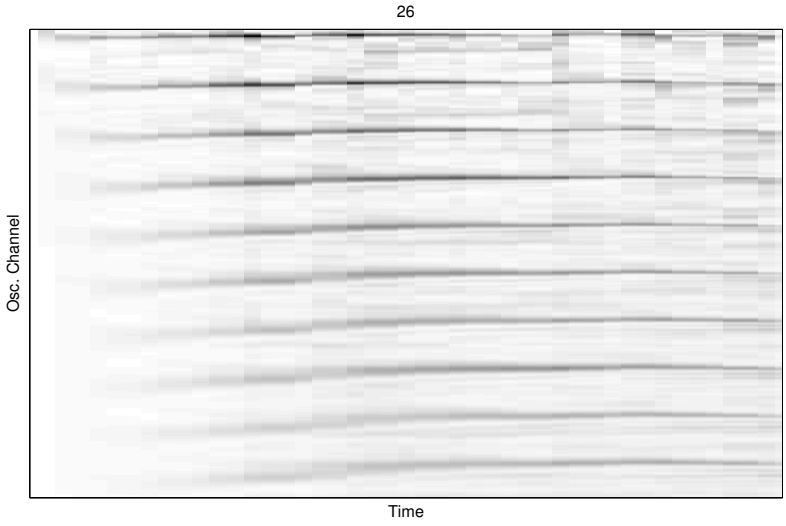

Figure 23. **modified Comb Filter Tempogram** - Periodicity estimation across K oscillator channels for the same input as Figure 22, based on the Canny-augmented comb filter equation, Eq. 7. Note the increased clarity in prevalent harmonic structure.

Effectively, the energy maintained by a resonator is calculated as a function of time at a resolution of 250Hz, producing a two-dimensional rhythmic resonance map approximating the perception of tempo evolution spanning the range of tuned oscillator frequencies. At each point in time, local maxima are calculated across the frame of oscillators after and collapsed to a normalized $\log_2$ frequency spectrum as a sum over 3 octaves [25:50, 50:100, 100:200]. The transformation from lag period T_k to normalized frequency ω_k is simply the inverse of the vector, followed by linear interpolation, and the frequency wrapping function is defined by Eq. 9. Here the partials, or peaks, of the frequency spectra are reduced to a frequency-normalized tempo chroma, characterizing the temporal structure of the initial audio input by the location (frequency) and salience (height) of rhythmic partials over L octaves. To continue with the tonal correlate, this system is presently concerned less with the height (octave) of the felt tempo and more so with the chroma (fundamental class), and attempts to resolve tactus ambiguities in other ways.

$$R_k[n] = \frac{1}{T_k} \sum_{i=0}^{T_k-1} (y_k[n - T_k])^2 \tag{8}$$

$$\Psi_n[\omega] = \frac{1}{L} \sum_{i=0}^{L-1} R_n[\omega + 2\pi * k] \tag{9}$$

As a final stage of rhythmic characterization, partials are collapsed in time as a summed, weighted histogram, defined by summing $\Psi_{n,\omega}$ in time, normalized to its length. A comprehensive histogram of partials generated by the mCFB

compactly described the temporal rhythmic evolution and content of an audio input. Frequency partials of relatively large amplitude only result in the event that they are consistent for a substantial duration, causing reinforcement. Therefore, the presence of a partial in the collapsed chroma vector implies that the frequency was sufficiently consistent to cause its respective oscillator to resonate. In this manner, analysis of the final chroma vector conveys information about the organization of rhythmic events globally, such as meter, common polyrhythms, interval complexity, and tempo modulations.

To expand upon the point further, the tempogram and chroma, defined as the graphical representations of $R_k[n]$ and $\Psi[\omega]$ respectively, are presented for four different types of audio input. Figures 24 and 25 demonstrate the parsing of a simple, digitally–sampled drum stick click track. The static nature of the rhythmic pattern is clearly seen in both the tempogram and the chroma. A tempogram is a time-frequency representation of tempo perception, where signal amplitude corresponds to the strength of a beat frequency at a given instant. Returning to an earlier concept, chroma is defined as the reduction of a wide-band tempogram to a single $\log_2$-spaced octave and collapsed in time. Wrapping frequency to a single octave produces a normalized frequency range and is visualized in polar coordinates, where the angle corresponds to fundamental normalized frequency and the radius defines the strength of the beat percept at that frequency.

Figures 26 and 27 represent a rock and roll track, which was likely not

recorded to using a metronome. Tempo regularity is apparent, with minor fluctuations. A blurring of the primary lobe in the corresponding chroma indicates this. Figures 28 and 29 characterize a piece of electronic music, and was likely recorded with the aid of some computerized clock source given the rigidity of tempo. Finally, Figures 30 and 31 show the rhythmic analysis of a classical solo piano piece, the performance of which is extremely expressive. In particular, the corresponding chroma gives a strong indication that there is no strong, salient beat percept.

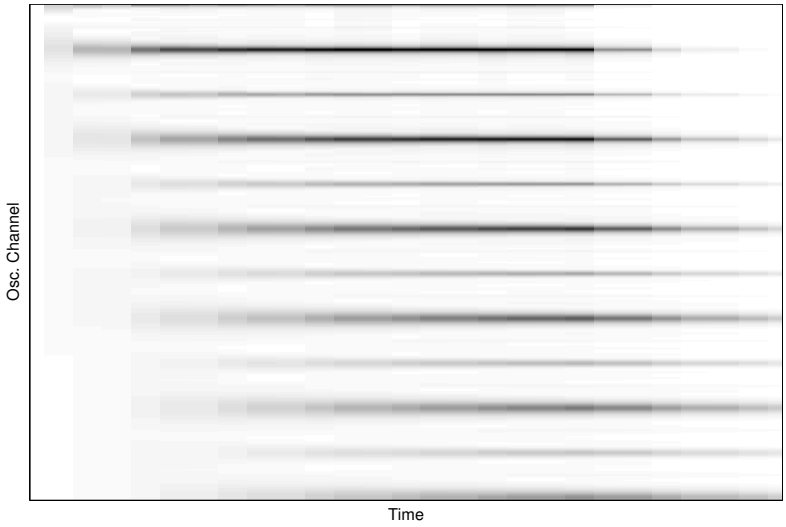

Figure 24. **Click Track — Tempogram** - Synthesized Drum Stick Click Track

4.3.4 Track Characterization

The primary motivating factor in the development of a robust yet efficient tempo-extraction algorithm is the need to process volumes of acoustic musical content. Despite some instances and applications for which it may be desired or necessary to determine these features in real time, the musical content considered

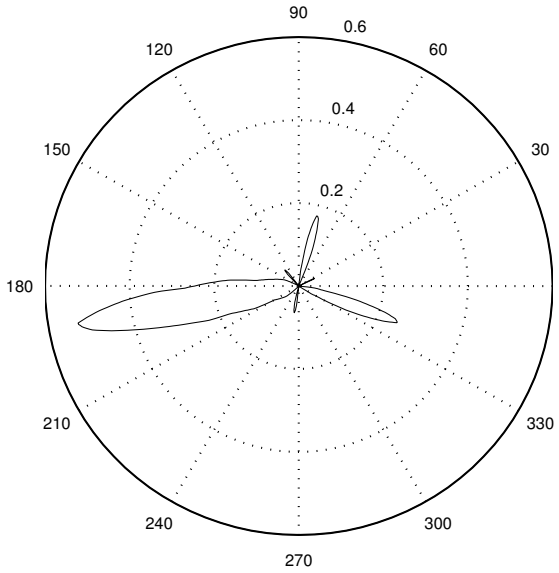

Figure 25. **Click Track — Chroma** - Synthesized Drum Stick Click Track

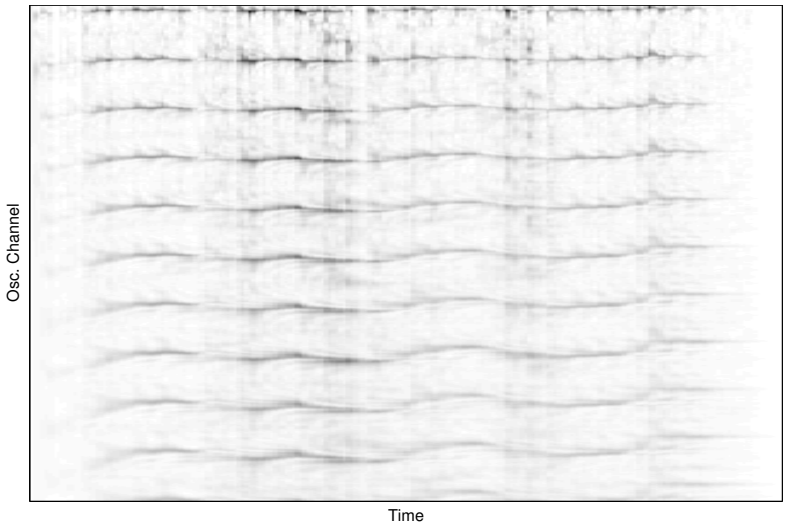

Figure 26. **Rock Track — Tempogram** - Heretics, by Andrew Bird (Bird, 2007)

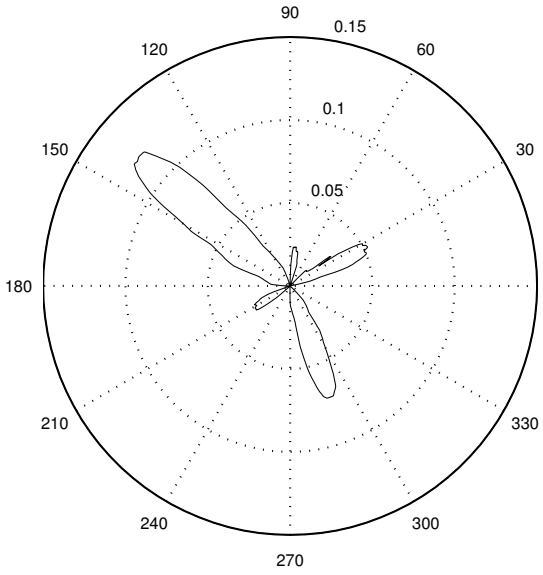

Figure 27. **Rock Track — Chroma** - Heretics, by Andrew Bird (Bird, 2007)

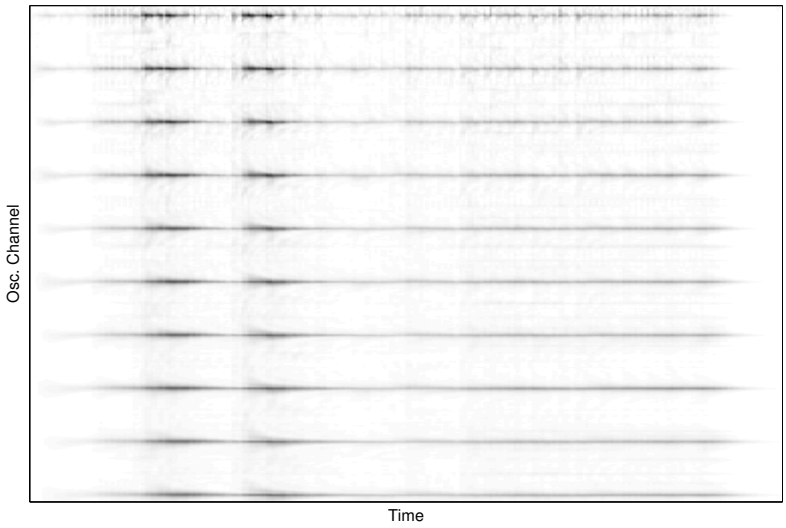

Figure 28. **Electronic Track — Tempogram** - Wraith Pinned to the Mist and Other Games, by Of Montreal (of Montreal, 2005)

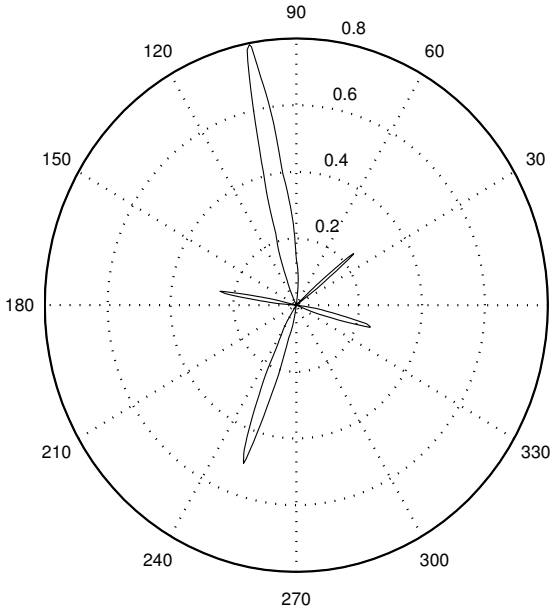

Figure 29. **Electronic Track — Chroma** - Wraith Pinned to the Mist and Other Games, by Of Montreal (of Montreal, 2005)

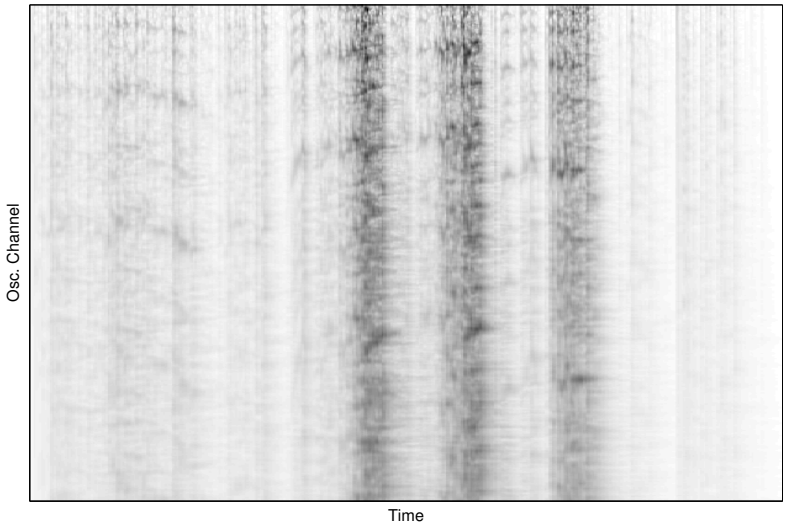

Figure 30. **Classical Track — Tempogram** - Raindrop Prelude, by F. Chopin (Chopin, 1997).

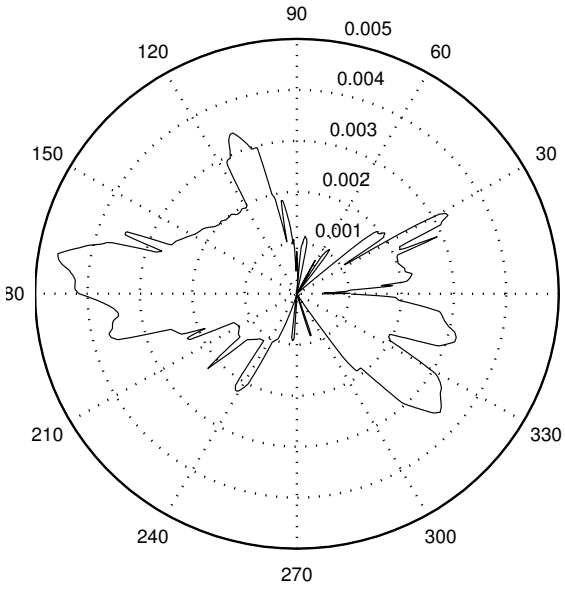

Figure 31. **Classical Track — Chroma** - Raindrop Prelude, by F. Chopin (Chopin, 1997).

in this context is static in the sense that, once acquired, the user does not typically modify it. The goal becomes the reduction of a singular musical track to a compact representation of its rhythmic behavior. A rhythmic feature vector is synthesized from the comprehensive analysis stage outlined previously, with the end application here requiring a more complete rhythmic description of each music track.

There are, precipitating from the analysis algorithm, a few key components that are sufficient to represent a musical track to meet the needs of this system. The two most important scalar values produced are the normalized fundamental beat frequency and strength of the fundamental. In terms of this wrapped harmonic representation, the frequency of the strongest partial

corresponds to the global tempo class of the track. Fundamental height provides a great deal of information about the rhythmic nature in the track, as relatively large values will be produced by strong, consistent onsets. Deviation in attack times will cause neighboring oscillators to resonate, scattering energy in the comb filterbank, while weaker onsets will obviously generate weaker resonance.

Three other descriptors are calculated from the description of the rhythmic harmonic spectra. A salience metric is created as the ratio of strengths between the fundamental and the next strongest partial. Motor music is composed of fractal rhythmic patterns that reinforce the fundamental frequency as integer octaves and sub-harmonics. For example, sixteenth, eighth, quarter, half, and whole notes are all integer multiples of a fundamental tactus, and the presence of which will overlap in a compact harmonic spectra. Syncopated note values, even as a simple as a dotted quarter, will begin to reinforce the 3:2 partial of the fundamental. Therefore, regardless of the specific frequency of this second-most salient rhythmic harmonic, this ratio of strengths quantifies the fractal nature of the track's beat.

Similarly, the spread of energy about the fundamental, described as the width of the primary lobe, characterizes the variation of the salient tempo percept in time. Simply put, the narrower the width of the primary lobe, the less continuous tempo modulation present throughout the entirety of the track. Local tempo modulation is calculated in a similar manner to audio spectral spread, as outlined in the MPEG-7 specification. In essence, this represents the second

moment of inertia about the fundamental. An important distinction must be made here where discontinuous, step-wise tempo modulations are not characterized by primary lobe spread and are instead passed as entirely different harmonics. Resolution of these distinct transitions in tempo is achievable through appropriate track segmentation via mechanisms such as temporal reduction of the over-riding tempo percept or timbral vectors.

Complementary to rhythmic information, it is equally important to coarsely quantify the temporal evolution of valence and activity in a track. Spectral power and flatness measures are derived from the twenty-two band decomposition scheme at the front end of the analysis, and collapsed to two one-dimensional timbral vectors. Spectral power serves as both a segmentation cue in the playlist generation stage, as well as a redundancy check in the verification that a track, or at least a portion of it, is suitable for running accompaniment. Segmentation of the track is performed by finding the areas of spectral transition regions agreed upon by the spectral power and flatness signals, which are determined by canny filtering and subsequent peak picking. Mean values are assigned between these segmentation boundaries for the spectral power signal and define sound intensity as a piece-wise function, allowing for a compact representation of song segments and power.

4.4 Playlist Generation

After arriving at a set of feature vectors for the plurality of tracks in the DML, they can be filtered, collapsed to a singular dimension and ranked for

suitability, defined here as a metric of "runability." Motivated primarily by a desire to minimize psychological discomfort and confusion, a subset of potential playlist candidates is created by first considering only a limited range of normalized fundamental beat frequencies (FBF). Biometric data acquired in the kinematic analysis is used to identify a target step frequency (TSF) and is used here to define the limits of the range by considering only FBFs both higher and lower than a normalized scalar frequency, set in the system to be 0.15 and 0.25, respectively. This subset of tracks is then segmented according to the spectral power vector to identify the optimally stimulative part of the track, and warped in time with phase vocoder techniques to cast all the clips to a uniform tempo. For playback continuity, the modified output tracks are faded in and out at a length of 600ms.

4.4.1 Ranking

As a naïve, generalized implementation, track candidates with suitable FBFs are directly ranked rather than classified into distinct groups, motivated by a few noteworthy elements. Automatic playlist generation, in the context of this project at least, emphasizes tempo accuracy as a mechanism to encourage motor entrainment. Semantic psychological characterizations, such as assessments of valence or arousal, are not quantified in an overly complex way, as the issue exists as a research endeavor in its own right, where the degree of emotional arousal evoked by a track may not even be due to a low-level feature of the acoustic signal. Additionally, limiting a user's library to a subset based on tempo

range exhibits, in practice, highly variable performance. In other words, the suitability of a generated playlist for running accompaniment is obviously tied directly to the intrinsic suitability of the content sample space provided, where suitability is described by both the acceptable tempo range defined about the TSF and the user's subjective rating of the music. Establishing an acceptable target range of beat frequencies can, in the event that there is minimal intersection between the distribution of track FBFs in a sample space and said target range, produce a significantly reduced, or even null, subset of content.

Initial ranking methods involved computing the Euclidean distance of a playlist candidate as the magnitude of its corresponding feature vector. Track feature vectors are composed of scalar values on the range of [0,1]. Euclidean distance favors feature vectors with the largest individual elements, valuing each independent feature equally. Alternative ranking mechanisms include calculating the geometric mean of the vector, punishing vectors with any number of independent features approximately equal to zero. Feature weighting was not explored, on the observation that ranking did not change significantly across methods.

4.4.2 Tempo Adjustment

Rhythmic analysis provides a quantitative characterization of tempo modulation in the course of a track, allowing the distinction to be made between heavily expressive and more metronome-like music. While the tempo-induction mechanism developed excels at establishing the suitability of a track for the end

application of RAS, it cannot provide a time-accurate estimation of instantaneous tempo. Oscillator-bank resonance – and therefore instantaneous tempo – are driven by both pulse interval and amplitude, causing the group delay of the system to depend on the input signal. Though it may be feasible to define a transformation to account for this group delay, the assumption is made that the most suitable tracks for RAS will natively exhibit highly constant tempo curves. Static-tempo music is sufficiently described by a global FBF, or it is segmented and ascribed a local FBF. In either event, the operational assumption of a constant tempo value allows for the direct adjustment of a track from its original time scale to one corresponding to the TBF.

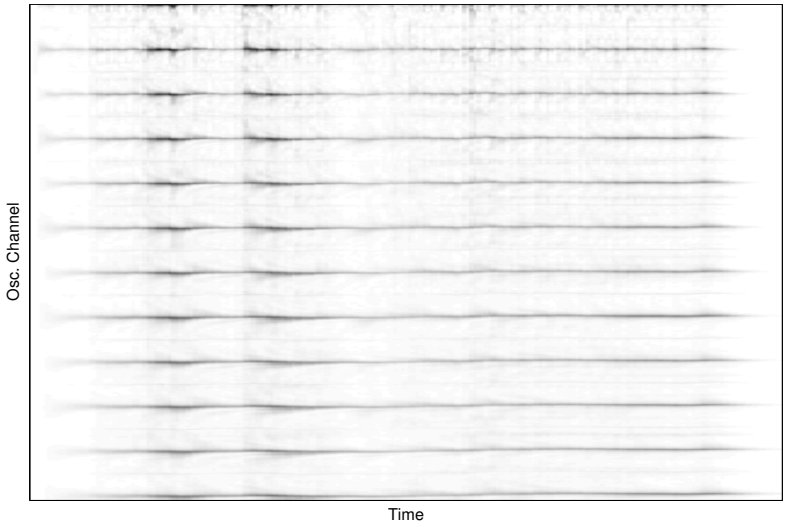

Figure 32. **Adjusted Electronic Track — Tempogram** - After tempo estimation and modification to 145BPM

For all ranked tracks, clips are segmented accordingly and phase vocoder techniques are used to effectively cast the tracks to a static tempo. Referencing a previous discussion of the rhythmic harmonic phasor representation, uniformly

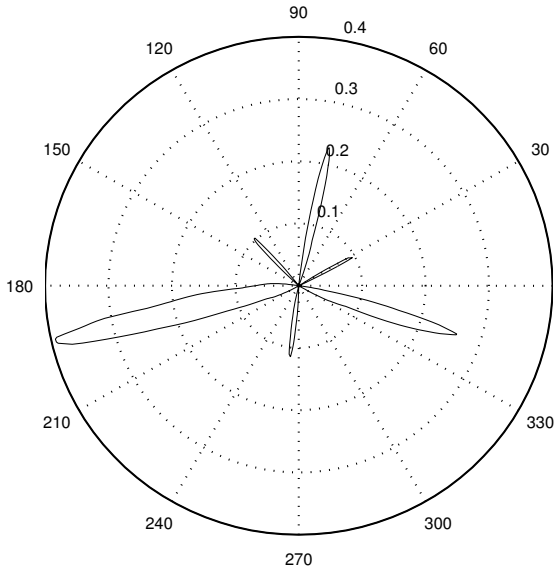

Figure 33. **Adjusted Electronic Track — Chroma** - After tempo estimation and modification to 145BPM

phase vocoding a track to a different time scale amounts to rotating the harmonic structure some number of degrees. Tempo correction then, in a future system, would reduce the widths of these spectral lobes, and, having been analyzed again, increase their respective amplitudes.

5

Evaluation

Having designed and implemented a complete end-to-end system, performance must be evaluated across multiple dimensions. Returning to the initial discussion of music-running systems, there are two discrete levels inherent to the task. Assuming the synthesis approach, musical content must first be semantically characterized and indexed, followed by suitable content selection and retrieval. The former stage is a matter of computational accuracy and, the latter, one of human cognition and reaction. Therefore, separate approaches are undertaken to evaluate these fundamentally different components.

Performance characterization of the developed rhythmic-analysis algorithm can be conducted objectively by applying traditional statistics methods. Resonant frequency measurement, fundamental beat–frequency estimation, and playlist tempo modulation can be explicitly described in terms of error rates and tolerances. Computational complexity and memory requirements are constant, measurable parameters that result directly from the system implementation. Other facets of system performance are not so clearly defined, but rather exist on a sliding, occasionally subjective, scale. This is true of beat strength and salience, both of which the algorithm is designed to assess.

Measuring the impact of a RAS-targeted playlist on an individual is significantly more complicated, due mainly to the introduction of the human

element. In light of this, feedback from test subjects can be measured in the two domains of psychological response and physical performance evaluation. Past research has shown that ratings of perceived exertion, as well as other psychological variables, can be affected by music listening conditions during sub-maximal exercise. An extension of this investigation seeks to identify any correlation between musicality, running affinity and the influence of resonant-frequency-targeted playlists. An evaluation methodology is presented, and an analysis of variance is calculated for the observed subject data across the different demographics.

5.1 Computational Performance

The developed system is responsible for automatically performing two separate analysis tasks. Biometric data are measured and parsed to extract a fundamental frequency of movement. Separately, but in a similar manner, musical content is parsed and characterized by a compact feature vector detailing, among other parameters, fundamental beat frequency. Both of these tasks are benchmarked through experimentation and described in terms of statistical accuracy.

5.1.1 Kinematic Analysis Accuracy

Without exaggeration, the merits of the entire system rest on its capacity to accurately identify the operational frequency of movement. A straightforward experiment is developed to validate whether the Navi cadence monitor behaves reliably. Using a digital metronome, the monitor is shaken in time with a

click-track for approximately 10 seconds over a range of tempi from 100 to 200 beats per minute, at 20BPM intervals. Data collected are processed and a resulting tempogram is generated, as shown in Figure 34. Represented as wrapped spectra, the peaks of each tempo interval are shown in Figure 35, corresponding to their respective frequency within a tolerance of a single BPM. The observant reader will notice that 100BPM kinematic frequency has aliased to slightly less than 200BPM as a result of information reduction that occurs when generating a single octave chroma.

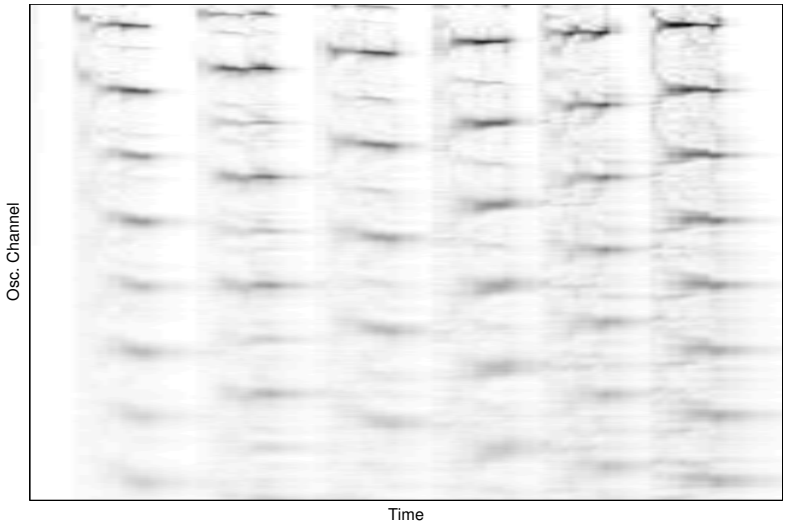

Figure 34. **Kinematic Data — Tempogram** - Validation of kinematic movement estimation accuracy.

5.1.2 Computational Complexity

One of the motivating forces behind the development of a computational rhythmic analysis algorithm is the observation that human completion of the same task is exceedingly time-intensive. Even simplifying the task to the determination of a binary trait – such as answering the question, "Is this entire

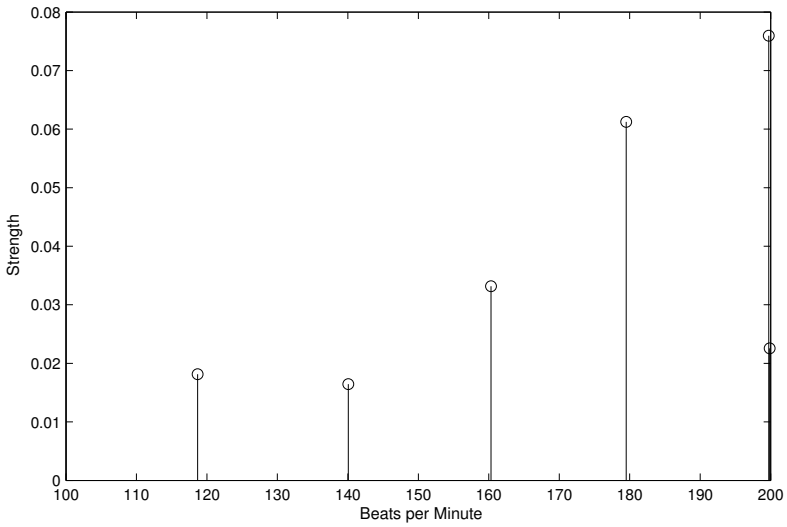

Figure 35. **Kinematic Data — Peaks** - Frequencies of motion for each stage of the recorded kinematic data.

song in the key of C Major?" – a human requires at least one thorough listening attempt in real-time to arrive at a decision. For the more continuous task of charting tempo evolution in time, it may require significantly more attempts to produce an accurate result. It is clearly meaningful to determine the degree of efficiency earned as a result of automating the process. The conventional metric of computational complexity is defined as the number of operations performed before returning a result for a given function. Since the overall task is dependent on the duration of content, however, complexity is expressed here in terms of operations per second of input audio.

Table 4 outlines the approximate theoretical number of required operations per stage of the algorithm at an input sampling rate of 32kHz. Halfband decomposition alone dwarfs the remainder of the algorithm, which is otherwise relatively efficient. A comparable (theoretical) FFT implementation

would require about 8000 operations, or roughly 4 MIPS to generate the same resolution subband envelopes. These numbers are hypothetical, however, as the FFT implementation in the MATLAB environment actually requires significantly more time to process than the proposed filterbank. Additionally, the complexity of the halfband decomposition is, as presented, the absolute worst-case scenario, and can be optimized as performance criteria are relaxed. High order FIR filters used can be turned down, or phase distortion can be compromised in favor of low-order elliptical IIR filters. As an example, replacing the current Daubechies coefficients with 8th-order half-band IIR filters reduces the computational complexity over an order of magnitude.

Stage	**Operations per Second**
Decomposition	104.5 MIPS
Onset Detection	0.2 MIPS
Haas Modeling	0.01 MIPS
Comb Filterbank	0.32 MIPS
Chroma Estimation	0.003
Total	*105 MIPS*

Table 4. **Computational Complexity** - Values are estimated as Millions of Instructions per Second (MIPS).

5.1.3 Rhythmic Analysis Accuracy

Multiple previous research endeavors (Goto and Muraoka, 2001; Gouyon et al., 2005; McKinney et al., 2007; Dixon, 2007; Gouyon and Meudic, 2003) have explored the topic of beat tracking and tempo-induction evaluation best practices. Admittedly, a scarce supply of accurately annotated reference content renders testing a non-trivial task, a testament to the utility of a robust computational algorithm. Cross-system comparisons are also difficult to draw, because individual systems are generally

tested with locally available content. To mitigate these issues, a sample space of over 1100 drum loops with BPM tags, ranging from 4 to 22 seconds in duration, used in a previous accuracy experiment (Gouyon et al., 2005), are also used in the evaluation of the proposed algorithm.

For the entire sample space, 74.2% of the audio clips are correctly identified according to the corresponding fundamental beat frequency class. When the sample space is limited to content at least 10 seconds in duration, the accuracy score increases to 81.75%, shown in Figure 36 as a ratio of the estimated and correct tempi, according to preassigned tags. The error tolerance is set to 4% in keeping with modern evaluation techniques. With respect evaluation with this set of drum tracks, errors typically fall into a few specific categories. The rhythmic content of several tracks is metrically ambiguous, such that the tatum may correspond to different tactus levels with no loss of accuracy. Alternatively, some tracks do not provide enough onset information for the algorithm to properly lock to a tempo.

Empirical evaluation was also conducted on the music content supplied by the human subjects, discussed in the next section. Manual annotation of tempi was performed for roughly 80 of the more than 2000 tracks, with the only observed inaccuracies arising from music that lacks a steady, discernable tempo. An example of such a track is shown in Figures 37 and 38. Note that there are visible areas of tempo reinforcement in the tempogram, despite erratic portions of the track, and the most consistent rhythmic part of the song is found after the

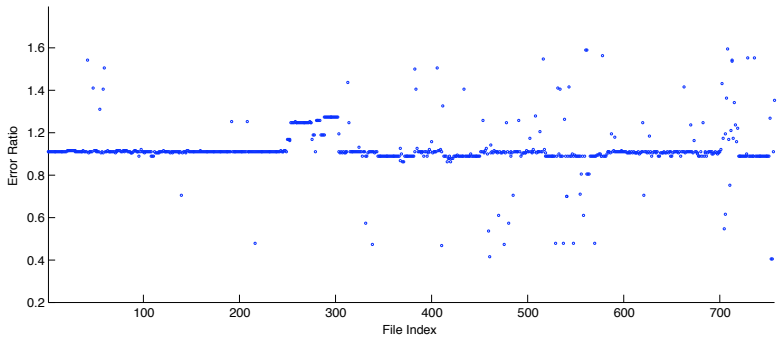

Figure 36. **Tempo Estimation Accuracy** - Results of an exploration of tempo estimation accuracy for a set of over 700 drum loops. Correct scores have an error ratio approximately equal to one, where the ratio of the estimate to actual is similar. Data representation in this manner serves to highlight small integer ratio errors, e,g, 3:2.

200th second, corresponding to the bridge. This track does not truly qualify as an error, but the width of the primary lobe, shown in the polar spectra, signifies considerable continuous tempo modulation in the track.

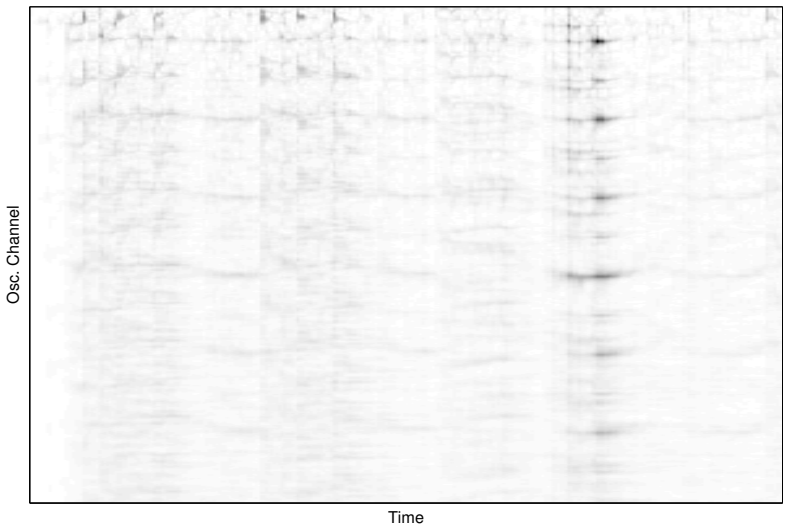

Figure 37. **Expressive Hard Rock Track — Tempogram** - F.C.P.R.E.M.I.X., by the Fall of Troy **??**

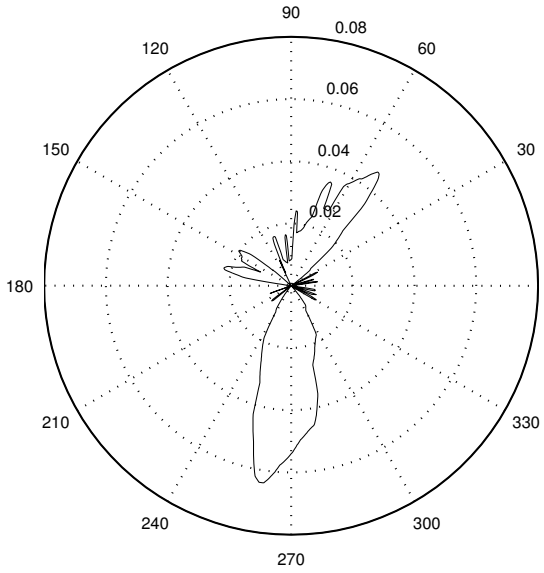

Figure 38. **Expressive Hard Rock Track — Chroma** - F.C.P.R.E.M.I.X., by the Fall of Troy **??**

5.1.4 Comparative Results

A final metric is investigated to determine how the performance of the developed algorithm compares to other modern tempo-extraction systems. Directly pitting the performance of other algorithms against the one proposed here is unfair, however, because the developed rhythmic–analysis algorithm makes no claims about the tactus of an excerpt and can be considered the easier task. Instead, the algorithm is tasked with parsing content used in the evaluation of tempo extraction algorithms as part of the 2006 and 2007 MIREX (MIREX, 2009).

Unfortunately, for the sake of parity in future evaluations, only a reduced subset is made publically available as a training database. Twenty tempo annotated musical excerpts are made freely available on the MIREX 2006

website, in addition to six bonus tracks which, lacking tempo information, were annotated manually. In processing this provided sample space, the developed algorithm convincingly identified the fundamental beat frequency of all but one of the twenty-six musical excerpts, including an excerpt with a meter of $\frac{7}{8}$, shown below. It can be argued that the correct tatum was actually identified for the one excerpt, but the music, a classical piece with no percussion, is too expressive to judge appropriately. Emphasis is placed on tactus estimation in both the 2006 (Gouyon et al., 2005) and 2007 (McKinney et al., 2007) MIREX, so performance comparisons can at best be inferred from the data available. Evaluation metric discrepancies notwithstanding, the developed algorithm performs at least as well as other state-of-the-art algorithms.

5.2 Human Subjects Testing

A unique aspect of the proposed system is its end application as human exercise accompaniment. Unlike binary task systems that either perform an action "right" or "wrong," human interactions and responses are, by nature, much more difficult to quantify. It is to the benefit of the experiment to minimize the complexity of that which is being evaluated, and constrain as many environment variables as feasibly possible. Beyond the process of characterizing content, discussed previously, the focus shifts toward understanding how, if at all, the selected content influences various human parameters.

Running, as a means of exercise, directly impacts two of the three dimensions of human health physical and mental although a case could be made

for its social implications. Motivated as a RAS-system, the hypothesis is formed that, since motor entrainment can guide and facilitate neuromuscular activity and recovery, it can be used to encourage and enhance the physical training of healthy motor systems. Possible observable manifestations of this may occur as increased performance levels or movement stability. Psychological impact may also be present, though, and necessitates statistical analysis of subject feedback to identify trends.

5.2.1 Methodology

A sample space of participants was recruited for the study representing four groups based on two binary traits musician/non-musician and runner/non-runner. Each subject was classified for each trait, resulting in four group subclasses: [1] musician/runner, [2] non-musician/runner, [3] musician/non-runner and [4] non-musician/non-runner. Subjects were characterized as musicians based upon current or previous study and/or performance history. A subject's running trait was determined based upon both self-perception and recent habits. All participants were over the age of 18 with no known motor or gait disorders, capable of jogging or running for 20 minutes without greater than normal strain, provided a minimum of 60 digital music tracks and understood English. Volunteers were recruited through the use of flyers and direct contact on the University of Miami, Coral Gables campus. Volunteers were accepted to participate in the study upon meeting the outlined criteria and the corresponding subclass requiring additional subjects to retain

test parity.

Before beginning the study, participants completed a demographic questionnaire to provide information regarding age, musical training and athletic ability. Each participant was asked to provide sufficient "music of interest" (approximately 100+ songs), to be used in the two music listening trials. Music of interest was defined for participants as "music you would be interested in listening to while running." Over the course of three separate meetings, each participant then completed running trials on the indoor track at the Wellness Center on the Coral Gables campus. For each trial, the participant was outfitted with the Navi cadence monitor, which can be easily worn as a belt. The sensor was positioned over the lower vertebrae on the back of the participant for consistency of data collection, although cadence estimation is equally accurate for arbitrary placement. Participants were instructed to run or jog comfortably for either twenty-two laps, equal to two miles, or until fatigued. Only one participant (0x10) voluntarily concluded each trial before completing the full two miles. At the conclusion of the each trial, the participant was given a post-exercise questionnaire evaluating the psychological aspects of the exercise, included in the Appendix.

The three trials were conducted in the following manner. An initial running trial was conducted without music to meet two goals. First, this provides an opportunity to assess a subject's psychological state when they are forced to associate with the physical activity, serving as a control for the other trials.

Free-field running without music also allows for an untainted estimate of the subject's resonant step frequency. The second trial introduces music listening in the form of randomly ordered excerpts provided by the participant, ranging in length from 30-60 seconds. A random music condition acts as an auxiliary control on the impact of music listening, to assess whether music alone has an influence. A final trial is conducted with music excerpts adjusted to match the previously identified resonant frequency of the participant. In both music listening conditions, subjects are allowed to skip tracks they find unsuitable.

5.2.2 Results

Eighteen subjects were recruited and fifteen finished the entirety of the study. All demographics were represented equally except for the non-runner/non-musician (nRnM) type, which proved rather difficult to recruit for the study. A one-way analysis of variance (ANOVA) was performed on the data collected in a variety of approaches in an attempt to identify statistically significant behavior. Variance is considered across binary traits of musician (M), non-musician ($\tilde{\text{M}}$), runner (R), and non-runner ($\tilde{\text{R}}$) rather than the 4 specific types, given the under-representation of nRnM subjects for three scenarios. First, the variance of data across all three trial conditions was analyzed. Two other ANOVAs are explored for meaningful data, in the comparison of music/no-music and random/RAS music listening conditions. The query statements presented to the subjects are given in Table 5.

Each ANOVA calculated offers a variety of noteworthy findings. Across all

Number	Query
Q1	That was a pretty intense workout.
Q2	I enjoyed the exercise.
Q3	I felt comfortable during the run.
Q4	I am extremely fatigued now.
Q5	I was very motivated during the exercise.
Q6	I feel like I could have kept going.

Table 5. **Subject Queries** - Subjects were directed to rate the accuracy of each statement from "Strongly Disagree" (1) to "Strongly Agree" (7).

conditions, motivation (Table 6, Q5) does not meaningfully differ between subjects. Also, the actual time required to run two miles varies consistently for all subject types. Ratings of enjoyment (Table 6, Q2) do not change statistically for non-runners, but they do for runners. As can be expected, runners' perception of their ability to have kept going (Table 6, Q6) does not vary across listening conditions.

	Global	M	$\tilde{\mathbf{M}}$	R	$\tilde{\mathbf{R}}$
Time, Est.	0.1578	0.4969	0.2855	0.3795	0.2376
Time, Act.	0.3579	0.5941	0.4561	0.4510	0.3946
Borg	0.8266	0.8879	0.9130	0.5563	0.7427
Q1	0.4207	0.5787	0.6860	0.4015	0.8451
Q2	0.2431	0.1639	0.6603	0.9717	0.0611
Q3	0.4737	0.2105	0.6013	0.3447	0.2653
Q4	0.9122	0.9296	0.6302	0.5771	0.5141
Q5	0.0063	0.0437	0.1289	0.3037	0.0083
Q6	0.4572	0.2788	0.5559	0.1691	0.2394

Table 6. **Subjective Ratings, All Conditions** - One-way ANOVA of subject feedback across all three exercise conditions.

The data also encourages the notion that random music and RAS-playlists have a measurable impact on subjects. Musicians and non-runners appear to be particularly sensitive to different music conditions, as both the perception of completion time and actual completion time exhibit more variation when considered separate from the no-music condition (Table 7, Time Est. and Act.). Again, the difference in music has minimal impact on a runner's perception of their exertion and fatigue (Table 7, Q3 & Q6), but musicians reported a change in

comfort (Q3), fatigue (Q4), motivation (Q5) and an ability to "keep going" (Q6).

	Global	**M**	$\tilde{\mathbf{M}}$	**R**	$\tilde{\mathbf{R}}$
Time, Est.	0.4880	0.7430	0.4716	0.4901	0.7011
Time, Act.	0.8759	0.8813	0.5444	0.5238	0.8116
Borg	0.7908	0.9080	0.8059	0.3614	0.5109
Q1	0.3510	0.5951	0.4807	0.2746	0.8918
Q2	0.2352	0.3984	0.4105	0.8525	0.0296
Q3	0.4309	0.8448	0.3409	0.1501	0.2959
Q4	0.8916	0.8597	1.0000	0.6113	0.3875
Q5	0.8616	1.0000	0.5995	0.8290	0.5995
Q6	0.4836	0.7894	0.3575	0.0512	0.4442

Table 7. **Subjective Ratings, Random vs. RAS Music** - One-way ANOVA of subject feedback across random and generated RAS music listening exercise conditions.

To determine the influence of music listening during running, a one-way ANOVA is calculated between the no-music data and each music listening condition, averaging the resulting variances. The p-scores calculated for the binary music listening condition approaches a crossover between the two previous data tables. All subjects, except for runners, reported minimal change in motivation regardless of the music listening environment (Table 8, Q5). Interestingly, runners, and not musicians, exhibited a high degree of variance for ratings of enjoyment between music and no-music conditions.

	Global	**M**	$\tilde{\mathbf{M}}$	**R**	$\tilde{\mathbf{R}}$
Time, Est.	0.1505	0.3306	0.2531	0.3253	0.1633
Time, Act.	0.2249	0.3930	0.3738	0.3752	0.2339
Borg	0.5929	0.6763	0.7677	0.6336	0.8205
Q1	0.4964	0.4605	0.6995	0.6147	0.6351
Q2	0.3700	0.1544	0.6506	0.9196	0.2619
Q3	0.4374	0.1310	0.7238	0.4699	0.3027
Q4	0.7263	0.7739	0.4017	0.4483	0.6615
Q5	0.0114	0.0402	0.1468	0.2252	0.0165
Q6	0.3966	0.1906	0.6788	0.3587	0.2357

Table 8. **Subjective Ratings, Music vs. No Music** - One-way ANOVA of subject feedback across music and non-music listening exercise conditions.

In addition to questionnaire data, subjects were engaged in open discussion about the study, and the personal experience of running with a

pre-adjusted constant tempo playlist. Feedback for the implementation evaluated was somewhat mixed, with the system working extremely well for some and not as well for others. In the cases that a generated playlist was met with dissatisfaction, participants reported that the tempo was within their resonant bandwidth, but too fast to comfortably keep pace with for the duration of the two-mile run.

About half the subjects did report positive experiences with the generated playlists, and the recorded kinematic data for one such subject is shown below for the first and third trial. Completion times of the two trials were 16:35 and 16:00, while maintaining an average step frequency of approximately 164 and 161 SPM, respectively. It is significant to note that despite a faster completion time, the subject's step frequency also decreased. As a result, the subject's stride length increased from an average of 3.87 feet to 4.10 feet. This behavior is significant for a few reasons, which will be discussed in greater detail shortly, but it is crucial to observe that successful (as reported by the subject) motor entrainment can encourage a longer stride for at a more optimal frequency, thereby increasing performance. It is somewhat surprising that there is more spurious behavior in the data gathered from the third, RAS trial, but its overall cadence curve is much flatter, indicated by the curvatures of main lobe apex shown in the polar spectra. The upper tempogram from the first trial has a steady upward slope, seen clearly in the lower harmonics (around the 500th oscillator channel). This behavior is detailed in Figures 39, 40, 41, and 42.

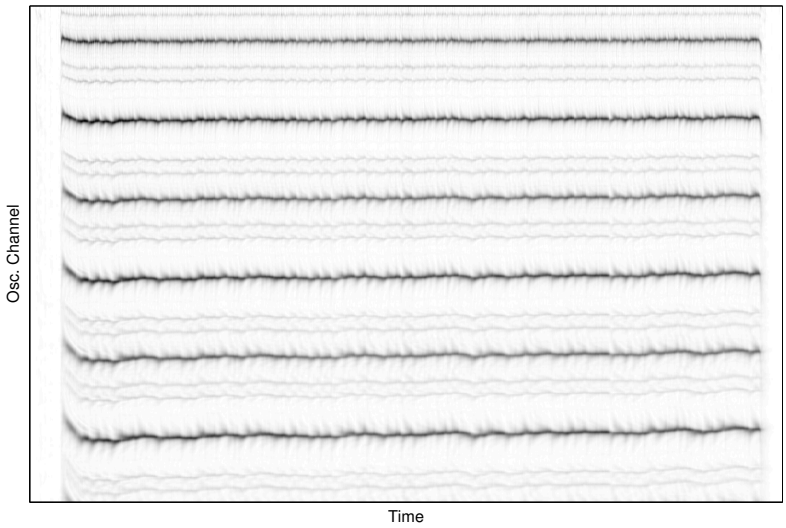

Figure 39. **Kinematic Tempogram, no Music** - Analysis of cadence data during a no-music running trial.

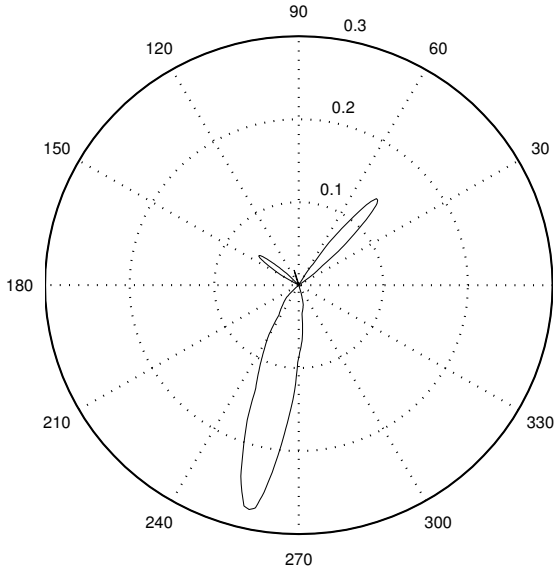

Figure 40. **Kinematic Chroma, no Music** - Cadence chroma during a no-music running trial.

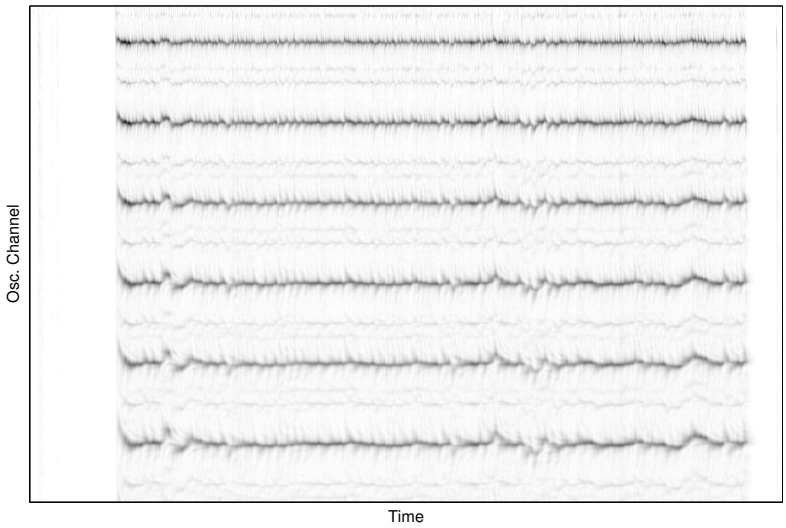

Figure 41. **Kinematic Tempogram, RAS–Music** - Analysis of cadence data during a generated RAS–music running trial.

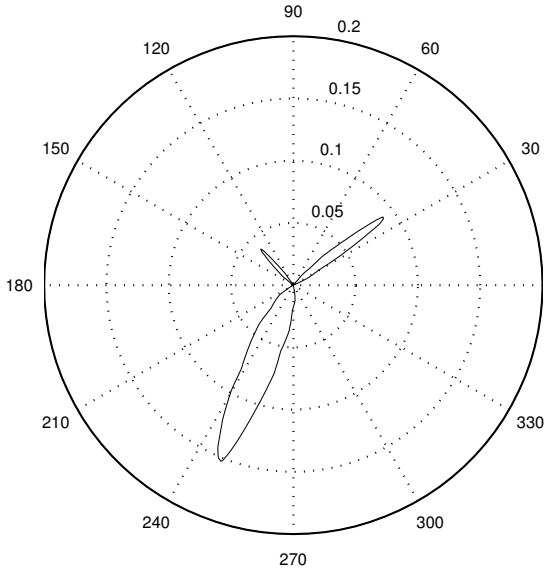

Figure 42. **Kinematic Chroma, RAS–Music** - Cadence chroma during a generated RAS–music running trial.

6

Discussion

The design, implementation and evaluation of a fully automated system for running accompaniment selection and retrieval invariably warrant a full discussion of intermediary conclusions, observations and insights. It is not only worthwhile to assess the system's strengths and weakness, but address why it succeeds when it does, as well as where it fails. System behavior extends beyond the algorithmic vantage point of accuracy, and motivates further discussion of the system as an intelligent user interface fostering human interaction. This topic naturally evolves into thoughts and considerations for future work.

6.1 Observations

As previously discussed, the developed system operates in a two-stage progression from the semantic analysis of content to the selection, retrieval and subsequent processing of the optimal subset for running accompaniment. Semantic analysis of audio, particularly in the context of suitability for rhythmic motor entrainment, presents a variety of fundamental issues. Above all else, the rhythmic-analysis algorithm proposed here was purposefully designed for the task of quantifying the appropriateness of a track for running accompaniment and, more specifically, motor entrainment. Acoustic information is really parsed as signal energy envelopes in the rhythmic dimension of music. Specific scenarios can be contrived from which the algorithm will be unable to reliably extract

rhythmic information, such as a low-frequency-modulated sine tone. However, the case can be made that this is not the optimal style of music for running accompaniment.

For the purposes of the algorithm, it is sufficient to extract energy-based onsets as a model of cochlear excitation. Partial signal decomposition into approximately critical bandwidths provides a quantifiable measure of acoustic stimulation, which is often the extent of an individual's locus of attention during physical activity. Other domains would account for different types of onsets, like variations in timbral contour or tonal patterns, and would no doubt improve the tempo extraction accuracy. The goal, however, is not necessarily the identification of a fundamental beat frequency for all possible musical input, but rather those input that are suitable for the intended application. Robust tempo extraction is an extremely difficult task to computationally model, as the human auditory system is redundant on many levels and benefits greatly from accumulated experiences. Operating on the principle that the development of a perfect system is not feasible, the approach is taken to develop a system that approaches optimal accuracy for the content of interest, and therefore content that "breaks" the system is of little consequence.

That being said, while the rhythmic-analysis algorithm is highly accurate in the estimation of fundamental beat frequency, an individual's perception of track suitability for running accompaniment would appear to be heavily tied to a model of valence and activity. Fundamental beat frequency class may not be

sufficient in the creation of playlists targeted at motor entrainment. One study participant in particular brought this issue to light, because a significant portion of the music she provided could be broadly described as "relaxation" music. As a style of music, the grouping empirically exhibits a tactus between 70 and 90 BPM, which effectively aliases to 140–180 BPM in terms of tempo chroma. This becomes a bit of a perceptual problem, as step frequencies naturally fall into this range, particularly toward the upper bound for females. The observation is made that females and shorter males tend to have significantly higher resonant step frequencies, but subjects did not typically provide a great deal of content with a tactus at that range. In the absence of music at this metrical level, tempo chroma defaults to the lower octave, is generally perceived as less stimulative, and therefore deemed less suitable as running accompaniment. One possible resolution is the inclusion of a back-end tactus estimation module, or it may be sufficient to describe a track by a fundamental beat frequency and an emotional arousal descriptor. Since the single-most important track feature to the performance of the system is the accurate estimation of tempo chroma, direct tactus identification from unwrapped beat spectra runs the risk of selecting partials or other miscellaneous harmonics.

The tempo distribution of the subject–provided content is shown with fundamental beat frequency mapped to the octave range of 100–200 BPM for local content for clarity. A few mentionable details can be inferred from this distribution. The second clustering around 165 BPM in the tempo distribution of

local content can be attributed to the aliasing of content with a tactus in the range of 75–100 BPM. Also, several spikes are apparent in the local tempo distribution, corresponding to decade tempo values, the most prominent at 120, 130 and 160 BPM, indicating a preference for "aesthetic" tempi. Also, even-valued tempi have a higher occurrence, given that a meter marking in a lower octave will produce a doubled harmonic. As a rule, it seems that tempo chroma is somewhat evenly distributed with moderate preference to the 100–130 BPM range.

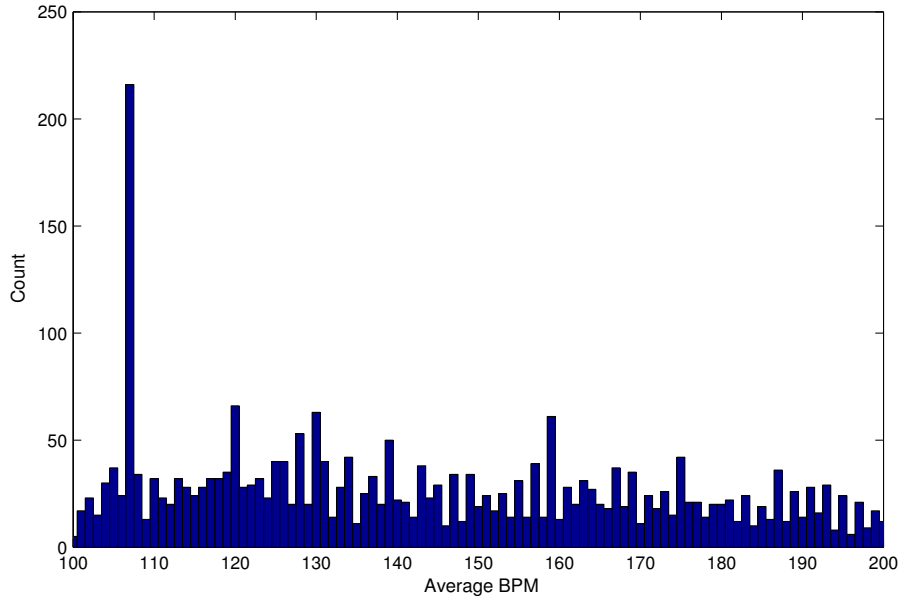

Figure 43. **Distribution of Track Tempi** - Integer-valued BPM histogram of all subject-provided content.

Segmentation is performed on selected tracks as a means, at least initially, of simply providing more content for the subject to evaluate. In the course of a 20-minute running event, an individual would be able to listen to roughly seven 3-minute tracks. Alternatively, segmenting tracks to a range of less then a minute

triples that number, allowing for an increase in the number of judgments a subject can make about the system. The realization was quickly made that segmented excerpts should not be the least ideal part of a song, if not the optimal part. It is desirable to bypass song introductions, truncate meandering endings, and generally identify an interesting segment of the audio. In agreement with intuition, this is hardly a trivial task. However, all subjects reported a preference for musical excerpts in the context of running.

Somewhat more interesting is the insight afforded through subject interaction with the system. A few subjects reported sub-optimal or even negative experiences with system-generated constant tempo playlists. Feedback concerning poor subject reactions can be grouped into two main categories, being either "wrong tempo" or "occasional bad track." Content matched to the wrong tempo may result from a few different causes that can be explored in the future. Early in the subject evaluation, there were admittedly some issues with attributing measured kinematic data to the appropriate subject due to the file-naming convention of the developed Navi activity monitor, occurring only a few times. Obviously, associating kinematic data to the wrong subject will likely produce erroneous results, and continued evaluation can elucidate the cause. Also, the subject evaluation stage was the first opportunity to expand on the small data set used in the development of the system, and parameter tuning and adjustment is likely necessary to generalize performance. In either event, the occurrence of a universally wrong tempo is a single misstep in the process of the

system, and will probably be resolved through iterative system development.

The occurrence of a "bad track," however, is a different type of issue altogether. There are, in terms of this application, a few basic reasons an individual will deem a track to be unsuitable. A track may not be at the right tempo, where right is defined as a frequency suitable for motor entrainment. An individual may perceive the song as not being suitable running accompaniment (a vocal solo, for example). Alternatively, an individual may simply not want to listen to a given track. Preference, with specific emphasis on the last example, is beyond the scope of this work, and cannot be held against the system. Perceptual suitability, which has been briefly addressed, is, to some degree, the responsibility of the developed system. Beat strength and regularity are quantified, in addition to estimates of signal intensity and spectral flatness, but the felt metrical level is not explicitly described. Tempo chroma for motor entrainment is clearly the crux of the system, and the occurrence of errors in this domain is exceedingly significant.

One subject in particular (0x08) reported that, for 19 tracks, she determined three were unsuitable for running accompaniment citing that they were "slow", and were skipped. In the questionnaire administered after completing the third trial, the rating of music suitability dropped from a 6 in the random music condition to 4, on a scale of 1:7. Upon closer inspection of this set of tracks, it was found that every track maintained a tempo in the range of 165:170 BPM, centered about 168 BPM (the target step frequency, mapped to

the appropriate tempo octave). With this knowledge, a few explanations can be offered to rationalize this data.

One straightforward answer is that the subject's resonant frequency was, in reality, closer to 170 BPM than 168 BPM, as some, but not all, tracks were found to be suitable. However, should this be the case, it is apparent that an individual's resonant frequency bandwidth is very narrow, on the range of 3 BPM. As will be discussed shortly, other subjects reported that the music generated was too fast, so this bandwidth may be even narrower, or highly subject dependent.

Another potential reason stems from the varied tactus level of the selected content. Without knowing exactly which tracks were perceived to be slower, it can only be hypothesized that two different music excerpts at different tempo octaves – but equal in chroma – played consecutively are perceptually different. Perceptual continuity of music that transitions between regular and double-time segments may lessen the impact, but this discrepancy may be amplified by the cognitive awareness that they are different songs. This hypothetical auditory illusion is illustrated by the classic Müller–Lyer optical illusion, shown in Figure 44. Both line segments are of equal length, but the directionality of the arrowheads gives the impression that the top line segment is shorter. Human perception is amazingly susceptible to sensory illusions, and this may be yet another example.

Regardless of the actual cause of an individual finding a track to be

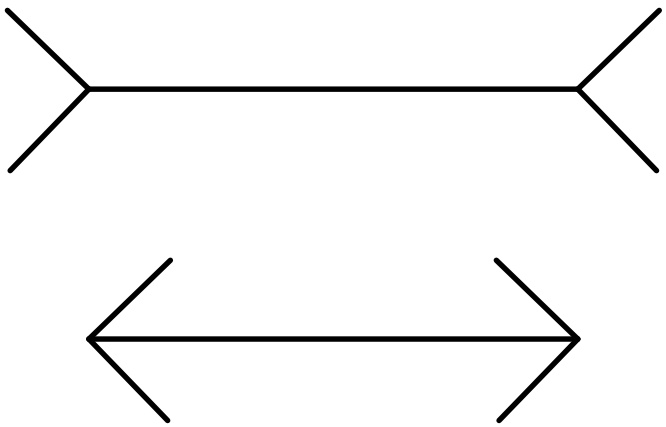

Figure 44. **The Müller–Lyer Effect** - An optical illusion; both center lines are of equal length, but are perceived differently due to surrounding information.

unsuitable for running accompaniment, the fact remains that subjects were considerably more sensitive to the tempo of music in the third trial than the second, diagramed in Figure 45. Subjects were encouraged to skip any tracks they found unsuitable at the time of the trial, which occurred significantly more in the third trial than the second. The two interpretations of this data are that either the generated playlists were drastically worse than the random condition, or that subjects became increasingly aware of the musical content when it was adapted to their environment. The former scenario can be safely ruled out based on responses from subjects and secondary review of generated content.

The conclusion to be drawn from this trend is that, when provided a music-listening environment that is approximately well suited to running accompaniment, subjects actively seek to optimize the experience. McCloud defines this general phenomenon as "amplification through simplification," albeit initially in the context of illustrations, but the application of the principle is valid

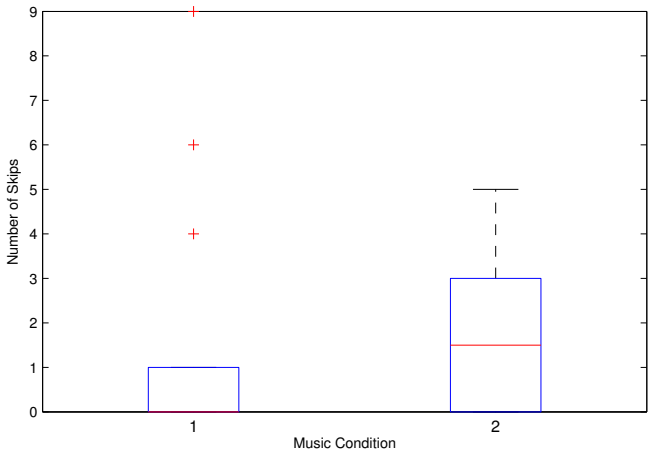

Figure 45. **Subject Track Skips vs. Condition** - Save for a select few, subjects skipped significantly more songs in the generated-RAS music listening condition.

nonetheless **??**. Concisely, through a reduction of detail in our periphery, humans are able to focus more intently on the information that remains, thereby amplifying our awareness of it. Most subjects place little value on finding a more suitable track in the random music condition because the evaluation criteria of what makes a "good" track is considerably broader. Musical content in the generated condition can be evaluated on a simpler set of criteria, allowing optimization to be perceived as an attainable, worthwhile endeavor.

External to the developed system, several general observations are made about the global environment of running and music listening. Contrary to intuition, a runner's step frequency actually increases with fatigue when maintaining the effort to run, as the stride length shortens causing choppier steps. A motor-entrainment system for physical training purposes would likely function better by employing a pulse-width modulated target cadence program

rather than a continuous, frequency-modulated contour. This is especially true depending on the acceptable resonant frequency bandwidth. Training programs could be implemented as phases or stages, rather than the often–suggested continuous periods of step frequency transitions. Additionally, the significant benefit of motor entrainment for runners is realized by keeping the step frequency constant and preventing it from rising with fatigue, which in turn encourages a greater stride length.

As a final comment, the general population is generally oblivious to the concept and potential implications of time stretching audio signals. Stretching and scaling of audio is performed quite regularly in television broadcasts of movies and other content that must be manipulated to fit within the framework of traditional programming schedules. It is equally common for radio stations to play music at a faster sampling rate than recorded, the motivations of which are no doubt insidious. Though detectable, particularly by individuals with perfect or absolute pitch, any artifacts that result are generally not noticeable and, except for those with pitch memory, definitely not detrimental. Time stretching performed as part of the algorithm begins to enter an ambiguous realm of human perception and cognition. Based on empirical, informal listening tests, the range of acceptable time-scale modifications was set as a ratio above and below the target frequency, based on the assumption that individuals will not be comfortable listening to familiar music or, even more so, favorite music if it is drastically different than what they know. This range also helps to prevent

phase–vocoding artifacts that result from substantial time–scale modifications.

Interestingly, the only subjects who reported hearing any phase vocoding artifacts were subjects currently pursuing degrees in Music Engineering Technology at the University of Miami. Even participants with considerable musical backgrounds did not notice any artifacts until it was indicated that the audio quality might be degraded. This has serious implications for deployable music-running systems that integrate some form of audio manipulation, indicating that individuals are even more indiscriminate listeners while exercising than initially assumed.

6.2 Future Work

The proposed system performs adequately well for the outlined goals as an initial implementation. There are a few key areas integral to overall performance that necessitate a continued research effort, as well as derivative topics that have materialized in the course of the work. Optimization is the most immediate priority should the algorithm venture beyond the bounds of a research environment. Various channels exist for optimizing the implementation for speed without sacrificing accuracy. Changes to the system essentially fall under the categories of design modification or restructuring for parallel computing. The proposed algorithm design is admittedly a rigorous implementation, specifically with regard to the decomposition stage. Relaxing filter parameters will ease the computational burden of the system, and it would be prudent to chart rhythmic analysis accuracy as a function of decomposition complexity. The use of

sequential oscillators for periodicity estimation could also be explored for a better implementation using fewer log-spaced oscillators.

Without impeding system performance by altering system design, there are additionally areas that could be optimized by directly restructuring the implementation architecture. Performing onset detection reduces the vector to a one-dimensional sparse matrix. The difference equation of a comb filter, given again in Eq. 10, can be represented by the piecewise function given in Eq. 11 for a sparse matrix input. A null input to a comb filter is merely a remapping of a past output to the current one, and therefore multiply-accumulate operations only need to be performed at the points when the input is nonzero. Alternatively, optimizing the algorithm for parallel computing architectures could attain potentially additive performance speed increases. For the system topology shown in Figure 46, the number of processing objects before and after parallelization is given at each stage. Assuming no limit to parallel cores, this reduces the number of processing objects from 650 to 10. For the current – non-optimized – processing speed of about 0.23 sec/second of audio, this rate would see an estimated reduction to under $10 msec/second$, or processing a 3 minute song in under 2 seconds.

$$y_k[n] = (1 - \alpha) * x[n] + \alpha * y_k[n - T_k] \tag{10}$$

$$y_k[n] = \begin{cases} \alpha * y_k[n - T_k], & \text{for } x[n] = 0, \\ (1 - \alpha) * x[n] + \alpha * y_k[n - T_k], & \text{elsewhere.} \end{cases} \tag{11}$$

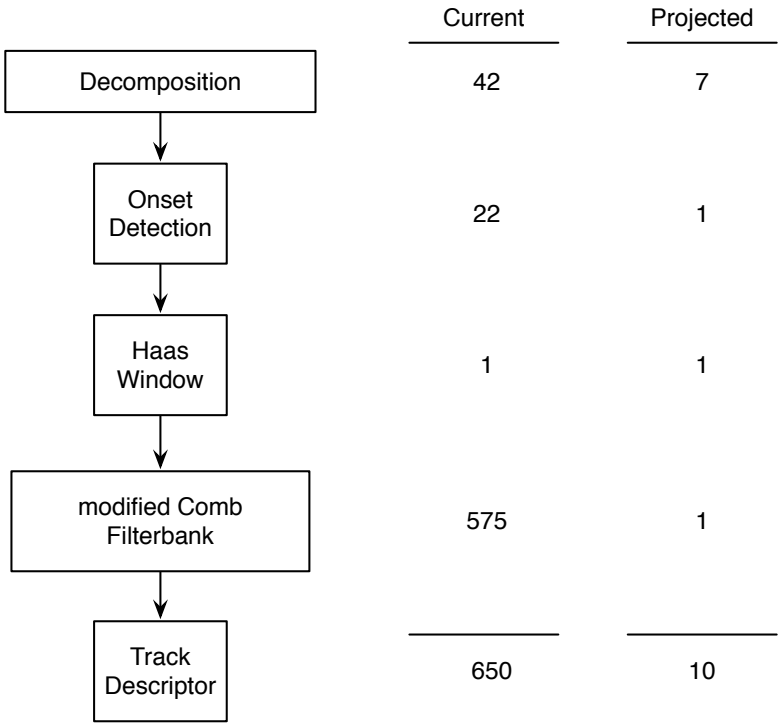

Figure 46. **System Processing Objects** - Significant computational reduction can be obtained through the parallelization of the proposed system. Values are shown in terms of high-level operations.

Specific system improvements include the capacity for tempo correction in the form of demodulation. A hypothetical implementation would front-end the system to first correct a track, should it require tempo demodulation, before performing rhythmic analysis to describe beat strength, fundamental beat frequency, and so forth. It is also necessary to motivate a discussion on human perception and preference with respect to determining the allowable limits of global tempo modification. There may be common agreement between

individuals on the extent of acceptable time companding for listening purposes, and future research would serve to evaluate this notion.

As is often the case with expert systems, which a semantic audio analysis algorithm could be considered, graceful degradation is often difficult to achieve. A true deployment "into the wild" would really first require the advent of a mechanism to identify and catch potential errors. The reality is that a system with 100% accuracy all of the time does not exist. Humans occasionally, and sometimes often, err at various semantic audio analysis tasks, but the robustness of the human annotation system lies in its ability to recognize when it has made a mistake. Whether self-correcting or defaulted to the hypothetical user, any real world computational machine listening algorithm necessitates the research and development of an accuracy estimation mechanism.

Part of the solution to this previous problem may lie in another research topic that has precipitated from this research. Representing beat spectra in polar coordinate form opens the door for an investigation into how various rhythms and meter are described by their harmonic arrangement. Typically, in computational rhythmic analysis, tempo octaves partials and harmonics are often seen as competing information in the primary task of identify the tactus. It may be possible, however, to infer more information from beat spectra than just the components. Conceivably, ratios between partials should give some indication of salient percepts, but it is not too far of a stretch that a ternary meter would produce different spectra than conventional binary meter. Much in the same way

pitch chroma is used to find transitions in phrasing, so too could rhythmic chroma. The amount of intrinsic information within beat spectra is a matter that requires focused research.

By the same token, the application of other signal-processing techniques, namely audio processing algorithms, for kinematic data analysis should be explored. Music analysis techniques lend themselves naturally to biometric data parsing of rhythmic events that may even be enhanced or improved in the presence of music. The information that can be mined from something as simple as a runner's acceleration waveform has only begun to be investigated. It may be possible to fingerprint and diagnose the motion of a wide array of physical activity that is not necessarily musical in nature. Activity monitors and step counters now regularly employ accelerometers, but they are equally applicable in measuring dancers, rowers, cyclists, and so on.

Returning to the topic of content analysis, there are many motivating factors into the exploration of the musical elements that make better running music. At this juncture, it is safe to say that appropriate tempo synchrony is a necessary, but not necessarily sufficient, condition in the qualification of a track as good running music. This motivates the question then, what does? Is it a matter of tactus, or simply psychological affect? Or, worse, is it something entirely referential that only a user-aware system could attempt to quantify? These are huge research questions that music informatics as a relative new discipline struggles with across a variety of different applications.

Of course, the natural evolution from adequately parsing a single user's DML to select and retrieve suitable running music is to extend the concept to the recommendation of music they do not already have, like, or even know exists. Recommendation systems invariably traverse broad areas of technology and philosophy, and constantly beg the question of whether or not the "problem" is even solvable. Theoretically, assuming a fixed amount of discrete content and a user who has the goal of finding an optimal sequence, a perfect, personalized recommendation system could conceivably guide that user from one piece of content to the next. However, if the user becomes aware of the perfect recommendation algorithm, does that render it imperfect?

On that note, without hesitation or apprehension, recommendation systems will qualify as future, unfinished work for a good while into the future.

LIST OF REFERENCES

Ahmaniemi, T. (2007). Influence of tempo and subjective rating of music in step frequency of running. *Proceedings of the 8th International Conference on Music Information Retrieval*.

Allen, P. and Dannenberg, R. (1995). Tracking musical beats in real time. *IJCAI*.

Alonso, M., Badeau, R., David, B., and Richard, G. (2003a). Musical tempo estimation using noise subspace projections. *WASPAA*.

Alonso, M., David, B., and Richard, G. (2003b). A study of tempo tracking algorithms from polyphonic music signals. *COST 276 Workshop*.

Alonso, M., David, B., and Richard, G. (2004). Tempo and beat estimation of musical signals. *Proceedings of the ISMIR*.

Alonso, M., David, B., and Richard, G. (2007a). Tempo estimation for audio recordings. *New Music Research*.

Alonso, M., Richard, G., and David, B. (2007b). Accurate tempo estimation based on harmonic + noise decomposition. *EURASIP*.

Antonopoulos, I., Pikrakis, A., and Theodoridis, S. (2007a). Self-similarity analysis applied on tempo induction from music recordings. *New Music Research*.

Antonopoulos, I., Pikrakis, A., and Theodoridis, S. (2007b). A tempo extraction algorithm for raw audio recordings. *Avaliable Online*.

Apple (2009a). Apple - iphone - mobile phone, ipod, and internet device. [Online]. Available: http://www.apple.com/iphone/.

Apple (2009b). Apple - nike + ipod. [Online]. Available: http://www.apple.com/ipod/nike/.

Apple (2009c). itunes store top music retailer in the us. [Online]. Available: http://www.apple.com/pr/library/2008/04/03itunes.html.

Auvinet, B., Berrut, G., Touzard, C., Moutel, L., Collet, N., Chaleil, D., and Barrey, E. (2002). Reference data for normal subjects obtained with an accelerometric device. *Gait and Posture*, 16:124–134.

Bieber, G. and Diener, H. (2005). Stepman — a new kind of music interaction. *HCI International*.

Bieber, G., Kirchner, B., and Diener, H. (1995). Stepman —- matching music to your moves. *MST News*, pages 31–32.

Bird, A. (2007). Heretics. In *Armchair Apocrypha*. Fat Possum Records.

Bowen, A. (2009). Music synchronization arrangement. Patent 7,521,623, Apple Inc.

Cemgil, A. and Kappen, B. (2002). Tempo tracking and rhythm quantisation by sequential monte carlo. *Advanes in Neural Information Processing Systems*.

Cemgil, A. and Kappen, B. (2003). Monte carlo methods for tempo tracking and rhythm quantization. *Artificial Intelligence Res.*

Cemgil, A., Kappen, B., Desain, P., and Honing, H. (2000). On tempo tracking: Tempogram representation and kalman filtering. *New Music Research*, 29(4):259–273.

Chen, H., Hsiao, M., Tsai, W., Lee, S., and Yu, J. (2007). A tempo analysis system for automatic music accompaniment. *Proceedings of the ICME*.

Chopin, F. (1997). Prelude for piano no. 15 in d flat major, op. 28/15, ct. 180. In *Chopin: Music for Piano*. Point Classics.

Cook, P., editor (1999). *Music, Cognition, and Computerized Sound: An Introduction to Psychoacoustics*. MIT Press.

Davies, M. and Brossier, P. (2005). Fast implementations for perceptual tempo extraction. *MIREX*.

Davies, M., Brossier, P., and Plumbley, M. (2005). Beat tracking towards automatic musical accompaniment. *AES Conference Proceedings*.

Davies, M. and Plumbley, M. (2005a). Beat tracking with a two-state model. *Proceedings of the ICASSP*.

Davies, M. and Plumbley, M. (2005b). Comparing mid-level representations for audio based beat tracking. *DMRN Summer Conference*.

Davies, M. and Plumbley, M. (2007). Context-dependent beat tracking of musical audio. *IEEE-TSAP*.

de l'Etoile, S. (2008). The effect of rhythmic auditory stimulation on the gait parameters of patients with incomplete spinal cord injury: An exploratory pilot study. *International Journal of Rehabilitation Research*, 31:155–157.

Dillman, C. (1975). Kinematic analyses of running. *Exercise and Sport Sciences Reviews*, 3(1):193–218.

Dixon, S. (1997). Beat induction and rhythm recognition. *Lecture Notes in Computer Science*.

Dixon, S. (1999). A beat tracking system for audio signals. *Mathematical and Computational Methods in Music, Diderot Forum on Mathematics & Music*.

Dixon, S. (2000). A lightweight multi-agent musical beat tracking system. *AAAI Workshop on AI and Music*.

Dixon, S. (2001). Automatic extraction of tempo and beat from expressive performances. *New Music Research*.

Dixon, S. (2007). Evaluation of the audio beat tracking system beatroot. *New Music Research*.

Dixon, S. and Cambouropoulos, E. (2000). Beat tracking with musical knowledge. *ECAI*.

Dixon, S., Goebl, W., and Widmer, G. (2002). Real-time tracking and visualisation of musical expression. *Proceedings of the Conference on Musical and Artificial Intelligence*.

Dolson, M. (1986). The phase vocoder: A tutorial. *Computer Music Journal*, pages 14–27.

Duxbury, C., Bello, J. P., Sandler, M., and Davies, M. (2004). A comparison between fixed and multiresolution analysis for onset detection in musical signals. In *Proc. of the 7th Int. Conference on Digital Audio Effects*.

Duxbury, C., Sandler, M., and Davies, M. (2002). A hybrid approach to musical note onset detection. *Proc. of the 5th Int. Conference on Digital Audio Effects*, pages 33–38.

Elliot, G. and Tomlinson, B. (2006). Personalsoundtrack: Context-aware playlists that adapt to user pace. *CHI*.

Ellis, D. (2007). Beat tracking with dynamic programming. *New Music Research*.

Foote, J. and Uchihashi, S. (2001). The beat spectrum: A new approach to rhythmic analysis. *ICME*.

Friberg, A. and Sundberg, J. (1999). Does music performance allude to locomoton? a model of final ritardandi derived from measurements of stopping runners. *J. Acoust. Soc. Am.*, 105(3):1469–1484.

Gao, S. and Lee, C. (2004). An adaptive learning approach to music tempo and beat analysis. *Proceedings of the ICASSP*.

Gizmodo (2009). Courier: First details of microsoft's secret tablet. [Online]. Available: http: //gizmodo.com/5365299/courier-first-details-of-microsofts-secret-tablet.

Goto, M. (2001). An audio-based real-time beat tracking system for music with or without drum sounds. *New Music Research*.

Goto, M. and Muraoka, Y. (1994). A beat tracking system for acoustic signals of music. *ACM-ICM*.

Goto, M. and Muraoka, Y. (1995). A real-time beat tracking system for audio signals. *ICMC*.

Goto, M. and Muraoka, Y. (1996). Beat tracking based on multiple-agent architecture: A real-time beat tracking system for audio signals. *Proceedings of the Conference on Multiagent Systems*.

Goto, M. and Muraoka, Y. (1997a). Music understanding at the beat level – real-time beat tracking for audio signals. *IJCAI*.

Goto, M. and Muraoka, Y. (1997b). Real-time rhythm tracking for drumless audio signals - chord change detection for musical signals. *IJCAI*.

Goto, M. and Muraoka, Y. (2001). Issues in evaluating beat tracking systems. *ICMC*.

Gouyon, F., Klapuri, A., Dixon, S., Alonso, M., Tzanetakis, G., Uhle, C., and Cano, P. (2005). An experimental comparison of audio tempo induction algorithms. *IEEE-TSAP*.

Gouyon, F. and Meudic, B. (2003). Towards rhythmic content processing of musical signals: Fostering complementary approaches. *New Music Research*.

Hainsworth, S. and Macleod, M. (2003). Beat tracking with particle filtering algorithms. *WASPAA*.

Harper, R. and Jernigan, M. (2004). Self-adjusting beat detection and predicition in music. *Proceedings of the ICASSP*.

HearingCentral.com (2009a). The cochlea. [Online]. Available: http://www.hearingcentral.com/images/cochlea1.gif.

HearingCentral.com (2009b). The ear. [Online]. Available: http://www.hearingcentral.com/Images/ear.gif.

Holzapfel, A. and Stylianou, Y. (2008). Beat tracking using group delay based onset detection. *Proceedings of the 9th International Conference on Music Information Retrieval*.

Humphrey, E. and Leider, C. (2009). The navi activity monitor: On using kinematic data to humanize computer music. In *Proceedings of the 2009 International Conference on New Interfaces for Musical Expression*. New York: Association of Computing Machinery.

Hurt, C., Rice, R., McIntosh, G., and Thaut, M. (1998). Rhythmic auditory stimulation in gait training for patients with traumatic brain injury. *Journal of Music Therapy*, 35(4):228–241.

Jensen, K. and Andersen, T. (2003a). Beat estimation on the beat. *WASPAA*.

Jensen, K. and Andersen, T. (2003b). Real-time beat estimation using feature extraction. *Computer Music Modeling and Retrieval Symposium*.

JirehDesign.com (2009). Camera analogy. [Online]. Available: http://www.maculacenter.com/images/illustrations/eyeAnatomyCamera.jpg.

Jochelson, D. and Fedigan, S. (2006). Design of an automatic beat-matching algorithm for portable media devices. *AES Conference Proceedings*.

Kavanagh, J. and Menz, H. (2008). Accelerometry: A technique for quantifying movement patterns during walking. *Gait and Posture*, 28:1–15.

Klapuri, A. (1999). Sound onset detection by applying psychoacoustic knowledge. *ICCASP*, 6:3089–3092.

Klapuri, A. (2003). Musical meter estimation and music transcription. *Cambridge Music Processing Colloquium*.

Klapuri, A. and Davy, M., editors (2006). *Signal Processing Techniques for Music Transcription*. Springer.

Klapuri, A., Eronen, A., and Astola, J. (2006). Analysis of the meter of acoustic musical signals. *IEEE-TSAP*.

Koonce, P. (2009). Pvc. [Online]. Available: http://silvertone.princeton.edu/winham/PPSK/pvc.osx.html.

Kurth, F., Gehrmann, T., and Muller, M. (2006). The cyclic beat spectrum: Tempo-related audio features for time-scale invariant audio identification. *Proceedings of the 7th International Conference on Music Information Retrieval*, pages 35–40.

Large, E. (2000). On synchronizing movements to music. *Human Movement Science*, 19:527–566.

Large, E. and Kolen, J. (1994). Resonance and the perception of musical meter. *Connection Science*, 6(1):177–208.

Laroche, J. (2003). Efficient tempo and beat tracking in audio recordings. *Audio Engineering Society*.

Masahiro, N., Takaesu, H., Demachi, H., Oono, M., and Saito, H. (2008). Development of an automatic music selection system based on runner's step frequency. *Proceedings of the 9th International Conference on Music Information Retrieval*, pages 193–198.

McKinney, M., Moleants, D., Davies, M., and Klapuri, A. (2007). Evaluation of audio beat tracking and music tempo extraction algorithms. *New Music Research*.

Meudic, B. (2002). A causal algorithm for beat tracking. *Proceedings of the ISMIR Convention*.

MIREX (2009). Evaluation database. [Online]. Available: http://www.music-ir.org/evaluation/MIREX/data/2006/tempo/.

Moe-Nilssen, R. and Helbostad, J. (2004). Estimation of gait cycle characteristics by trunk accelerometry. *Journal of Biomechanics*, 37:121–126.

Mohammadzadeh, H., Tartibiyan, B., and Ahmadi, A. (2008). The effects of music on the perceived exertion rate and performance of trained and untrained individuals during progressive exercise. *Physical Education and Sport*, 6(1):67–74.

Molinari, M., Leggio, M., Martin, M. D., Cerasa, A., and Thaut, M. (2003). Neurobiology of rhythmic motor entrainment. *Annals of the NY Academy of Sciences*, pages 313–321.

of Montreal (2005). Wraith pinned to the mist and other games. In *The Sunlandic Twins*. Polyvinyl Records.

Oliver, N. and Flores-Mangas, F. (2006). Mptrain: A mobile, music and physiology-based personal trainer. *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*, pages 21–28.

Peeters, G. (2005). Time-variable tempo detection and beat marking. *ICMC*.

Peeters, G. (2007). Template-based estimation of time-varying tempo. *EURASIP*.

Pikrakis, A., Antonopoulos, I., and Theodoridis, S. (2004). Music meter and tempo tracking from raw polyphonic audio. *Proceedings of the ISMIR Convention*.

Pikrakis, A. and Theodoridis, S. (2007). An application of empirical mode decomposition on tempo induction from music recordings. *Proceedings of the 8th International Conference on Music Information Retrieval*.

Review, N. B. (2009). Motorola droid: hands-free gps features in first google android 2.0 phone. [Online]. Available: http://www.nbr.co.nz/article/motorola-droid-critics-verdict-first-google-android-20-phone-114267.

Santos, J., Umali, E., and Garcia, I. (2003). A multi-agent algorithm for real-time automatic beat and tempo synchronization. *Avaliable Online*.

Scaringella, N. and Zoia, G. (2004). A real-time beat tracker for unrestricted audio signals. *Avaliable Online*.

Scheirer, E. (1997). Pulse tracking with a pitch tracker. *WASPAA*.

Scheirer, E. (1998). Tempo and beat analysis of acoustic musical signals. *Acoustical Society of America*.

Seppanen, J. (2001). Tatum grid analysis of musical signals. *WASPAA*.

Sethares, W., Morris, R., and Sethares, J. (2005). Beat tracking of musical performances using low level audio features. *IEEE-TSAP*.

Tecchio, F., Salustri, C., Thaut, M., Pasqualetti, P., and Rossini, P. (2000). Conscious and preconscious adaptation to rhythmic auditory stimuli: A magnetoencephalographic study of human brain responses. *Experimental Brain Research*, 135:222–230.

Telegraph (2009). Apple reasserts itself as market leader with ipod event. [Online]. Available: http://www.telegraph.co.uk/technology/apple/6166340/Apple-reasserts-itself-as-market-leader-with-iPod-event.html.

Thaut, M. (2008). *Rhythm, Music, and the Brain: Scientific Foundations and Clinical Applications*. Routledge.

Thaut, M., McIntosh, G., Prassas, S., and Rice, R. (1992). Effect of rhythmic auditory cuing on temporal stride parameters and emg patterns in normal gait. *Journal of Neurologic Rehabilitation*, 6(4):185–190.

Thaut, M., McIntosh, G., and Rice, R. (1997). Rhythmic facilitation of gait training in hemiparetic stroke rehabilition. *Journal of Neurological Sciences*, 151:207–212.

Thaut, M., McIntosh, G., Rice, R., Miller, R., Rathburn, J., and Brault, J. (1996). Long-term effects of rhythmic auditory stimulation on gain in patients with parkinson's disease. *Movement Disorders*, 11:193–200.

Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals. *IEEE-TSAP*, 10(5):293–302.

Tzanetakis, G., Essl, G., and Cook, P. (2001). Audio analysis using the discrete wavelet transform. In *Proc. Conf. in Acoustics and Music Theory Applications*.

Uhle, C. and Herre, J. (2003). Estimation of tempo, microtime, and time signature from percussive music. *DAFx*.

Vaidyanathan, P. (1993). *Multirate Systems and Filterbanks*. Prentice Hall.

Westergren, T. (2009). The music genome project. [Online]. Available: http://www.pandora.com/mgp.shtml.

Wijnalda, G. (2005). *Interactive music for sports*. PhD thesis, Vrije University.

Wikipedia (2009a). Sea shanty. [Online]. Available: http://en.wikipedia.org/wiki/Sea_shanty.

Wikipedia (2009b). Walkman. [Online]. Available: http://en.wikipedia.org/wiki/Walkman.

Wininger, S. and Pargman, D. (2003). Assessment of factors associated with exercise enjoyment. *Journal of Music Therapy*, 40(1):57–73.

Zöler, U. (2002). *Digital Audio Effects*. John Wiley & Sons.

Appendix