Computer Music Analysis *

David Gerhard
School of Computing Science
Simon Fraser University
Burnaby, BC
V5A 1S6
email: dbg@cs.sfu.ca

Simon Fraser University, School of Computing Science
Technical Report CMPT TR 97-13

Keywords: Automatic Music Transcription, Musical Grammars,
Computer Music, Psychoacoustics.

July 11, 2002

Abstract

Computer music analysis is investigated, with specific reference to the current research fields of automatic music transcription, human music perception, pitch determination, note and stream segmentation, score generation, time-frequency analysis techniques, and musical grammars. Human music perception is investigated from two perspectives: the computational model perspective desires an algorithm that perceives the same things that humans do, regardless of how the program accomplishes this, and the physiological model perspective desires an algorithm that models exactly how humans perceive what they perceive.

1 Introduction

A great deal of work in computer music deals with synthesis—using computers to compose and perform music. An interesting example of this type of work is Russell Ovens' thesis, titled "An Object-Oriented Constraint Satisfaction System Applied to Music Composition" [Oven88]. In contrast, this report deals with analysis—using computers to analyze performed and recorded music. The interest in this field comes from the realization that decoding a musical representation into sound is fairly straightforward, but translating an arbitrary sound into a score is a much more difficult task. This is the problem of Automatic Music Transcription (AMT) that has been looked at since the late 1970's. AMT consists of translating an unknown and arbitrary audio signal into a fully notated piece of musical score. A subset of this problem is monophonic music transcription, where a single melody line played on a single instrument

*This research is partially supported by The Natural Sciences and Engineering Research Council of Canada and by a grant from The BC Advanced Systems Institute.

in controlled conditions is translated into a single note sequence, often stored in a MIDI track[1]. Monophonic music transcription was solved in 1986 with the publication of Martin Piszczalski's Ph.D. thesis, which will be discussed in Section 3.0.1 on page 6.

AMT is a subset of Music Perception, a field of research attempting to model the way humans hear music. Psychological studies have shown that many of the processing elements in the human auditory perceptual system do the same thing as processing elements in the human visual perceptual system, and researchers have recently begun to draw from past research in computer vision to develop computer audition. Not every processing element is transferable, however, and the differences between vision and audition are clear. There are millions of each of the four types of vision sensors, L, M, and S cones and rods, while there are only two auditory sensors, the left and right ears. Some work has been done on using visual processing techniques to aid in audition (see Section 2.1 on page 4 and Section 4.2.2 on page 13), but there is more to be gained from cautiously observing the connections between these two fields.

Many research areas relate to computer music and AMT. Most arise from breaking AMT down into more manageable sub-problems, while some sprouted from other topics and studies. Six areas of current research work will be presented in this report:

- Transcription
  - Music perception
  - Pitch determination
  - Segmentation
  - Score generation
- Related Topics
  - Time-frequency analysis
  - Musical grammars

Part I

Transcription

The ultimate goal of much of the computer music analysis research has been the development of a system which would take an audio file as input and produce a full score as output. This is a task that a well-trained human can perform, but not in real time. The person, unless extremely gifted, requires access to the same audio file many times, paying attention to a different instrument each time. A monophonic melody line would perhaps require a single pass, while a full symphony might not be fully transcribable even with repeated auditions. Work presented in [Tang95] suggests that if two instruments of similar timbres play "parallel" musical passages at the same time, these instruments will be inseperable. An example of this is the string section of an orchestra, which is often heard as a single melody line when all the instruments are playing together.

Some researchers have decided to work on a complementary problem, that of extracting errors from performed music [Sche95]. Instead of listening to the music and writing the score, the computer listens to the music, follows the score, and identifies the differences between what the ideal music should be and what the performed music is. These differences could be due to expression in the performance, such as vibrato, or to errors in the performance, such as incorrect notes. A computer system that could do this would be analogous to a novice music student who knows how to read music and can listen to a piece and follow along in the music, but cannot yet transcribe a piece. Such a system gives us a stepping stone to the full problem of transcription.

[1] MIDI will be discussed in Section 6.1 on page 15.

2 Music Perception

There have been two schools of thought concerning automatic music transcription, one revolving around computational models and one revolving around psychological models.

The psychological model researchers take the "clear-box" approach, assuming that the ultimate goal of music transcription research is to figure out how humans hear, perceive and understand music. A working system is not as important as little bits of system that accurately (as close as we can tell) model the human perceptual system. This attitude is valuable because by modeling a working system – the human auditory system, we can develop a better artificial system, and gain philosophical insight into how it is that we hear things.

In contrast, the computational model researchers take the "black-box" approach, assuming that a music transcription system is acceptable as long as it transcribes music. They are more concerned with making a working machine and less concerned with modeling the human perceptual system. This attitude is valuable because in making a working system we can then work backward and say "How is this like or unlike the way we hear?" The problem is that the use of self-evolving techniques like neural nets and genetic algorithms limits our ability as humans to understand what the computer is doing.

These two fields of research rarely work together and are more often at odds with each other. Each does have valuable insights to gain from the other, and an interdisciplinary approach, using results from both fields, would be more likely to succeed.

Aside: Computational Models. A point of dispute in interdisciplinary research has often been the notion of a computational model for explaining human psychological phenomena. Computer scientists and engineers have been using computational models for many years to explain natural phenomena, but when we start going inside the mind, it hits a little closer to home.

When a scientific theory tries to explain some natural phenomenon, the goal is to be able to predict that phenomenon in the future. If we can set up a model that will predict the way the world works, and we use a computer algorithm to do this prediction, we have a computational model. The argument is that these algorithms do not do the same thing that is happening in the world, even though they predict what is happening in the world. Kinematics models do not take into account quantum effects in their calculations, and since they do not model the way the world really works, they are not valuable, some would say. These models *do* explain and predict motion accurately within a given domain and whether or not they model the complete nature of the universe, they do explain and predict natural phenomena.

This is less easy to accept when dealing with our own minds. We want to know the underlying processes that are going on in the brain, so how useful is a theory, even if it is good at predicting the way we work, if we don't know how close it is to our underlying processes? Wouldn't it be better to develop a theory of the mind from the opposite viewpoint and say that a model is valid if it simulates the way we work from the inside – if it concurs with how the majority of people react to a certain stimulus? This is very difficult because psychological testing cannot isolate individual processes in the brain, it can only observe input and output of the brain as a whole, and that is a system we cannot hope to model at present.

The other problem with this is one of pragmatics. Ocham's razor says that when theories explain something equally well, the simpler theory is more likely to be correct. A simpler theory is more likely to be easier to program as well, and so simpler theories tend to come out of computational models. Of course, we must be careful that the theories we compare do in fact explain the phenomena equally well, and we must be aware that the simplest computational model we get is not necessarily the same processing that goes on in the mind.

This brings another advantage of computational models, their testability. Psychological models can be tested, but it is much more difficult to control such experiments, because all systems in the brain are working together at the same time. In computational models, only the processing we are interested in is present, and that bit of processing can be tested with completely repeatable results.

2.1 Auditory Scene Analysis

Albert Bregman's landmark book in 1990 [Breg90] presented a new perspective in human music perception. Until then, much work had been done in the organization of human visual perception, but little had been done on the auditory side of things, and what little there was concentrated on general concepts like loudness and pitch. Bregman realized that there must be processes going on in our brains that determine how we hear sounds, how we differentiate between sounds, and how we use sound to build a "picture" of the world around us. The term he used for this picture is "the auditory scene".

The classic problem in auditory scene analysis is the "cocktail party" situation, where you are in a room with many conversations going on, some louder than the one you are engaged in, and there is background noise such as music, clanking glasses, and pouring drinks. Amid all this cacophony, humans can readily filter out what is unimportant and pay attention to the conversation at hand. Humans can track a single auditory stream, such as a person speaking, through frequency changes and amplitude changes. The noise around may be much louder than your conversation, and still you have little trouble understanding what the other person is saying. For a recent attempt to solve this problem, see [GrBl95].

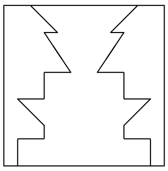


Figure 1: The Face-Vase Illusion.

An analogy that shows just how much processing is done in the auditory system is the lake analogy. Imagine digging two short trenches up from the shore of a lake, and then stretching handkerchiefs across the trenches. The human auditory system is then like determining how many boats are on the lake, what kind of engines are running in them, which direction they are going, which one is closer, if any large objects have been recently thrown in the lake, and almost anything else, merely from observing the motion of the handkerchiefs. When we bring the problem out to our conscious awareness, it seems impossible, and yet we do this all the time every day without thinking about it.

Bregman shows that there are many phenomena going on in the processing of auditory signals that are similar to those in visual perception. *Exclusive allocation* indicates that properties belong to only one event. When it is not clear which event that property applies to, the system breaks down and illusions are perceived. The most common visual example of this is the famous "face-vase" illusion, Figure 1, where background and foreground are ambiguous, and it is not clear whether the boundary belongs to the vase or the two faces. This phenomenon occurs in audition as well. In certain circumstances, musical notes can be ambiguous. Depending on what follows a suspended chord, the chord can be perceived as both major and minor, until the ambiguity is removed by resolving the chord.

Apparent motion occurs in audition as it does in vision. When a series of lights are flashed on and off in a particular sequence, it seems like there is a single light traveling along the line. If the lights are flashed too slow or they are too far apart, the illusion breaks down, and the individual lights are seen turning on and off. In audition, a similar kind of streaming occurs, in two dimensions. If a series of notes are of a similar frequency, they will tend to stream together, even if there are notes of dissimilar frequencies interspersed. A sequence that goes “Low-High-Low-High...” will be perceived as two streams, one high and one low, in two circumstances. If the tempo is fast enough, the notes seem closer together and clump into streams. Also, if the difference between the “Low” and the “High” frequencies is large enough, the streaming will also occur. If the tempo is slow and the frequencies do not differ by much, however, the sequence will be perceived as one stream going up and down in rapid succession.

There are more examples of the link between vision and audition in Bregman’s book, as well as further suggestions for a model of human auditory perception. Most of the book explains experiments and results that reinforce his theories.

2.2 Axiomatization of Music Perception

In 1995, Andranick Tanguiane presented the progress of a model of human music perception that he had been working on for many years [Tang95]. The model attempts to explain human music perception in terms of a small set of well-defined axioms. The axioms that Tanguiane presents are:

Axiom 1 (Logarithmic Pitch) The frequency axis is logarithmically scaled.

Axiom 2 (Insensitivity to the phase of the signal) Only discrete power spectra are considered.

Axiom 3 (Grouping Principle) Data can be grouped with respect to structural identity.

Axiom 4 (Simplicity Principle) Data are represented in the least complex way in the sense of Kolmogorov (least memory stored).

Axiom 1 stems from the well-recognized fact that a note with a frequency twice that of some reference note is an octave higher than the reference note. An octave above “A” at 440 Hz would be “A” at 880 Hz, and then “A” at 1760 Hz and so on. This is a logarithmic scale, and has been observed and documented exhaustively.

Axiom 2 is a bit harder to recognize. Two different time signals will produce the same auditory response if the power frequency spectra are the same. The phase spectrum of a signal indicates where in the cycle each individual sinusoid starts. As long as the sinusoidal components are the same, the audio stimulus will sound the same whether the individual components are in phase or not.

The third axiom is an attempt to describe the fact that humans group audio events in the same gestalt manner as other perceptual constructs. Researchers in musical grammars have identified this, and there is more material on this phenomenon in Section 2.1 on page 4.

The last axiom suggests that while we may hear very complicated rhythms and harmonies in a musical passage, this is done in mental processing, and the mental representation used is that which uses the least memory.

Tanguiane argues that all of these axioms are necessary because without any one of them, it would be impossible for humans to recognize chords as collections tones. The fact that we are able to perceive complex tones as individual acoustical units, claims Tanguiane, argues for the completeness of the axiom set. This perception is only possible when the musical streams are not parallel.

2.3 Discussion

Research has approached the problem of music perception and transcription from two different directions. In order to fully understand music enough to develop a computational model, we must recognize the psychological processing that is going on. Early work in automatic music transcription was not concerned with the psychological phenomena, but with getting some sort of information from one domain to another. This was a valuable place to start, but in its present state, the study of music perception and transcription has much to gain from the study of the psychology of audition.

3 Beginnings

In the mid 70's, a few brave researchers tried to tackle the whole transcription problem by insisting on a limited domain. Piszczalski and Galler presented a system that tried to transcribe musical sounds to a musical score [PiGa77]. The intent was to take the presented system and develop it toward a full transcription system. In order to make the system functional, they required the input audio to be monophonic from a flute or a recorder, producing frequency spectra which are easily analyzed.

3.0.1 Piszczalski and Galler

At this early point in the research, Piszczalski and Galler recognized the importance of breaking the problem down into stages. They divided their system into three components, working bottom-up. First, a signal processing component identifies the amplitudes and starting and stopping times of the component frequencies. The second stage takes this information and formulates note hypotheses for the identified intervals. Finally, this information is analyzed in the context of notation to identify beam groups, measures, etc., and a score is produced.

A windowed Fourier transform is used to extract frequency information. Three stages of heuristics are used to identify note starts and stops. The system first recognizes fundamental frequencies below or above the range of hearing as areas of silence. More difficult boundaries are then identified by fluctuations in the lower harmonics and abrupt changes in fundamental frequency amplitude. The second phase averages the frequencies within each perceived "note" to determine the pitch which is then compared to a base frequency. In the final phase, a new base frequency is determined from fluctuations in the notes, and grouping, beaming and other notational tasks are performed.

Since this early paper, Piszczalski has proposed a computational model of music transcription in his 1986 Ph.D. Thesis [Pisz86]. The thesis describes the progress and completion of the system started in 1977. There is also a thorough history of AMT up to 1986. This thesis is a good place to start, for those just getting in to computer music research. It is now more than 10 years old and much has been accomplished since then in some areas, but monophonic music transcription remains today more or less where Piszczalski left it in 1986.

3.0.2 Moorer

In 1977, James Moorer presented a paper [Moor77] on the work he accomplished since then, which would turn out to be the first attempt at polyphonic music transcription. Piszczalski's work required the sound input to be a single instrument and a single melody line, while Moorer's new system allowed for two instruments playing together. Restrictions on the input to this system were tighter than Piszczalski's work: there could only be two instruments, they must both be of a type such that the pitches are voiced (no percussive instruments) and piecewise constant (no vibrato or glissando) and the notes being played together must be such that the fundamental frequencies and harmonics of the notes do not overlap. This means that no note pairs are allowed where the fundamental frequencies are small whole number ratios of each other.

Moorer claims that these restrictions, apart from the last one, are merely for convenience and can easily be removed in a larger system. He recognized that the fundamental frequency restriction is difficult to overcome, as most common musical intervals are on whole number frequency ratios. The other restrictions he identified have been shown to be more difficult to overcome than he first thought. Research into percussive musical transcription has shown that it is sufficiently different from voiced transcription to merit independent work, and glissando and vibrato have proved to be difficult problems. Allowing more than two instruments in Moorer's system would require a very deep restriction on the frequencies.

These restrictions cannot be lifted without a redesign of the system, because notes on the octave, or even a major third apart have fundamental frequencies that are whole number multiples of each other, and thus cannot be separated by his method.

Moorer uses an autocorrelation function instead of a Fourier transform to determine the period of the signal, which is used to determine the pitch ratio of the two notes being played[2]. The signal is then separated into noise segments and harmonic segments, by assigning a quality measure to each part of the signal. In the harmonic portions, note hypotheses are formed based on fundamental frequencies and their integer multiples. Once hypotheses of the notes are confirmed, the notes are grouped into two melody lines (which do not cross) by first finding areas where the two notes overlap completely, and then filling in the gaps by heuristics and trial and error. Moorer then uses an off-the-shelf program to create a score out of the two melody lines.

3.1 Breakdown

The early attempts at automatic music transcription have shown that the problem must be limited and restricted to be solved, and the partial solutions cannot easily be expanded to a full transcription system. Many researchers have chosen to break the problem down into more manageable sub-problems. Each researcher has her own ideas as to how the problem should be decomposed, and three prominent ones are presented here.

3.1.1 Piszczalski and Galler

In [PiGa77], discussed above, Piszczalski and Galler propose three components, breaking the process down temporally as well as computationally.

1. (low-level) Determine the fundamental frequency of the signal at every point, as well as where the frequencies start and stop.
2. (intermediate-level) Infer musical notes from frequencies determined in stage 1.
3. (high-level) Add notational information such as key signature and time signature, as well as bar lines and accidentals.

Piszczalski's 1986 thesis proposes a larger and more specific breakdown, in terms of the intermediate representations. In this breakdown, there are eight representations, suggesting seven processing stages. The proposed data representations are, from lowest level to highest level:

- Time waveform: the original continuous series of analog air pressures representing the sound.

[2] Pitch and Period are not necessarily interchangeable. For a discussion, see Section 4.1.2 on page 10.

- Sampled signal: a series of discrete voltages representing the time waveform at every sample time.

- Digital spectrogram: the sinusoid spectrum of the signal at each time window.

- Partials: the estimation of the frequency position of each peak in the digital spectrogram.

- Pitch candidates: possibilities for the pitch of each time frame, derived from the partials.

- Pitch and amplitude contours: a description of how the pitch and amplitude vary with time.

- Average pitch and note duration: discrete acoustic events which are not yet assigned a particular chromatic note value.

- Note sequence: the final representation.

The first four representations fit into stage 1 of the above breakup, The next three fit into stage 2, and the last one fits into stage 3. In his system, Piszczalski uses pre-programmed notation software to take the note sequence and create a graphical score. The problem of inferring high-level characteristics, such as the key signature, from the note sequence has been researched and will be covered in Section 6.2 on page 16.

3.1.2 Moorer

James Moorer's proposed breakdown is similar to that of Piszczalski and Galler, in that it separates frequency determination (pitch detection) from note determination, but it assigns a separate processing segment to the identification of note boundaries, which Piszczalski and Galler group together with pitch determination. Moorer uses a pre-programmed score generation tool to do the final notation. Indeed, this is a part of automatic music transcription which has been, for the most part, solved since the time of Moorer and his colleagues, and there exist many commercial software programs today that will translate a MIDI signal (essentially a note sequence) into a score. For more on MIDI see Section 6.1 on page 15.

3.1.3 Tanguiane

In 1988, Andranick Tanguiane published a paper on recognition of chords[Tang88]. Before this, chord recognition had been approached from the perspective of polyphonic music—break down the chord into its component notes. Tanguiane didn't believe that humans actually did this dissection for individual chords, and so his work on music recognition concentrated on the subproblems that were parts of music. He did work on chord recognition, separate from rhythm recognition, separate again from melody recognition. The division between the rhythm component and the tone component has been identified in later work on musical grammars, and will be discussed in Section 8 on page 18.

3.2 An Adopted Breakdown

For the purposes of this report, AMT will be broken down into the following sub-categories:

1. Pitch determination: identification of the pitch of the note or notes in a piece of music. Work has been done on instantaneous pitch detection as well as pitch tracking.

2. Segmentation: breaking the music into parts. This includes identification of note boundaries, separation of chords into notes, and dividing the musical information into rhythm and tone information.

3. Score generation: taking the segmented information and producing a score. Depending on how much processing was done in the segmentation section, this

could be as simple as taking fully defined note combinations and sequences and printing out the corresponding score. It could also include identification of key and time signature.

4 Pitch Determination

Pitch Determination has been called Pitch Extraction and Fundamental Frequency Identification, among a variety of other titles, however pitch and frequency are not exactly the same thing, as will be discussed later. The task is this: given an audio signal, what is the musical pitch associated with the signal at any given time. This problem has been applied to speech recognition as well, since some languages such as Chinese rely on pitch as well as phonemes to convey information. Indeed, spoken English relies somewhat on pitch to convey emotional or insinuated information. A sentence whose pitch increases at the end is interpreted as a question.

In monophonic music, the note being played has a pitch, and that pitch is related to the fundamental frequency of the quasi-periodic signal that is the musical tone. In polyphonic music, there are many pitches acting at once, and so a pitch detector may identify one of those pitches or a pitch that represents the combination of tones but is not present in any of them separately. While pitch is indispensable information for transcription, more features must be considered when polyphonic music is being transcribed.

Pitch following and spectrographic analysis deal with the continuous time-varying pitch across time. As with instantaneous pitch determination, many varied algorithms exist for pitch tracking. Some of these are modified image processing algorithms, since a time-varying spectrum has three dimensions (frequency, time and amplitude) and thus can be considered an image, with time corresponding to width, frequency corresponding to height, and amplitude corresponding to pixel value.

Pitch determination techniques have been understood for many years, and while improvements to the common algorithms have been made, few new techniques have been identified.

4.1 Instantaneous Pitch Techniques

Detecting the pitch of a signal is not as easy as detecting the period of oscillation. Depending on the instrument, the fundamental frequency may not be the pitch, or the lowest frequency component may not have the highest amplitude.

4.1.1 Period Detectors

Natural music signals are pseudo-periodic, and can be modeled by a strictly periodic signal time-warped by an invertible function[ChSM93]. They repeat, but each cycle is not exactly the same as the previous, and the cycles tend to change in a smooth way over time. It is still meaningful to discuss the period of such a signal, because while each cycle is not the exact duplicate of the previous, they differ only by a small amount (within a musical note) and the distance from one peak to the next can be considered one cycle. The period of a pseudo-periodic signal is how often the signal “repeats” itself in a given time window.

Period detectors seek to estimate exactly how fast a pseudo-periodic signal is repeating itself. The period of the signal is then used to estimate the pitch, through more complex techniques described below.

Fourier Analysis. The “old standard” when discussing the frequency of a signal. A signal is decomposed into component sinusoids, each of a particular frequency and amplitude. If enough sinusoids are used, the signal can be reconstructed within a given error limit. The problem is that the discrete Fourier transform centers sinusoids around a given base frequency. The exact period of the signal must then be inferred by examining the Fourier components. The common algorithm that is used to calculate the fourier spectrum is called the fast fourier transform (FFT).

Chirp Z Transform. A method presented in [Pisz86], it uses a charge-coupled device (CCT) as an input. The output is a frequency spectrum of the incoming signal, and in theory should be identical to the output of a Fourier transform on the same signal. The method is extremely fast when implemented in hardware, but performs much slower than the FFT when simulated in software. The CCD acts as a delay line, creating a variable filter. The filter coefficients in the delay line can be set up to produce a spectrum, or used as a more general filter bank.

Cepstrum Analysis. This technique uses the Fourier transform described above, with another layer of processing. The log magnitude of the Fourier coefficients is taken, and then inverse Fourier-transformed. The result is a large peak at the frequency of the original signal, in theory. This technique sometimes needs tweaking as well.

Filter Banks. Similar to Fourier analysis, this technique uses small bandpass filters to determine how much of each frequency band is in the signal. By varying the center frequency of the filters, one can accurately determine the frequency that passes the largest component and therefore the period of the signal. This is the most psychologically faithful model, because the inner ear acts as a bank of filters, providing output to the brain through a number of orthogonal frequency channels.

Autocorrelation. A communication systems technique, this consists of seeing how similar a signal is to itself at each point. The process can be visualized as follows: take a copy of the signal and hold it up to the original. Move it along, and at each point make a measurement of how similar the signals are. There will be a spike at "0", meaning that the signals are exactly the same when there is no time difference, but after a little movement, there should be another spike where one cycle lines up with the previous cycle. The period can be determined by the location of the first spike from 0.

The problem with most of these techniques is that they assume a base frequency, and all higher components are multiples of the first. Thus, if the frequency of the signal does not lie exactly on the frequency of one of the components for example on one of the frequency channels in a bank of filters, then the result is a mere approximation, and not an exact value for the period of the signal.

4.1.2 Pitch from Period

A common assumption is that the pitch of a signal is directly confessed by its period. For simple signals such as sinusoids, this is correct in that the tone we hear is directly related to how fast the sinusoid cycles. In natural music, however, many factors influence the period of the signal apart from the actual pitch of the tone within the signal. Such factors include the instrument being played, reverberation, and background noise. The difference between period and pitch is this: a periodic signal at 440 Hz has a pitch of "A", but a period of about 0.00227 seconds.

A technique proposed in [Pisz86] to extract the pitch from the period consists of formulating hypotheses and then scoring them and selecting the highest score as the fundamental frequency of the note. Candidates are selected by comparing pairs of frequency components to see if they represent a small whole number ratio with respect to other frequency components. All pairs of partials are processed in this way, and the result is a measure of pitch strength versus fundamental frequency.

4.1.3 Recent Research

Xavier Rodet has been doing work with music transcription and speech recognition since before 1987. His fundamental frequency estimation work has been done with Boris Doval, and they have used techniques such as Hidden Markov Models and Neural Nets. They have worked with frequency tracking as well as estimation.

In [DoRo93], Doval and Rodet propose a system for the estimation of the fundamental frequency of a signal, based on a probabilistic model of pseudo-periodic signals previously proposed by them in [DoRo91]. They consider pitch and fundamental frequency to be two different entities which sometimes hold the same value.

The problem they address is the misnaming of fundamental frequency by computer when a human can easily identify it. There are cases where a human observer has difficulty identifying the fundamental frequency of a signal, and in such cases they do not expect the algorithm to perform well. The set of partials that the algorithm observes is estimated by a Fourier transform, and consists of signal partials (making up the pseudo-periodic component) and noise partials (representing room noise or periodic signals not part of the main signal being observed).

Doval and Rodet's probabilistic model consists of a number of random variables including a fundamental frequency, an amplitude envelope, the presence or absence of specific harmonics, the probability density of specific partials, and the number and probability of other partials and noise partials. Partials are first classified as harmonic or not, and then a likelihood for each fundamental frequency is calculated based on the classification of the corresponding partials. The fundamental frequency with maximum likelihood is chosen as the fundamental frequency for that time frame. This paper also presented work on frequency tracking, see Section 4.2 on page 12.

In 1994 Quirós and Enríquez published a work on loose harmonic matching for pitch estimation [QuEn94]. Their paper describes a pitch-to-MIDI converter which searches for evenly spaced harmonics in the spectrum. While the system by itself works well, the authors present a pre-processor and a post-processor to improve performance. The pre-processor minimizes noise and abhorrent frequencies, and the processor uses fuzzy neural nets to determine the pitch from the fundamental frequency.

The system uses a "center of gravity" type measurement to more accurately determine the location of the spectral peaks. Since the original signal is only pseudo-periodic, an estimation of the spectrum of the signal is used based on a given candidate frequency. The error between the spectrum estimation and the true spectrum will be minimal where the candidate frequency is most likely to be correct.

Ray Meddis and Lowel O'Mard presented a system for extracting pitch that attempts to do the same thing as the ears do [MeOM95]. Their system observes the auditory input at many frequency bands simultaneously, the same way that the inner ear transforms the sound wave into frequency bands using a filter bank. The information that is present in each of these bands can then be compared, and the pitch extracted. This is a useful method because it allows auditory events to be segmented in terms of their pitch, using onset characteristics. Two channels whose input begins at the same time are likely to be recognizing the same source, and so information from both channels can be used to identify the pitch.

4.1.4 Multi-Pitch Estimation for Speech

Dan Chazan, Yoram Stettiner and David Malah presented a paper on multi-pitch estimation [ChSM93]. The goal of their work was to segment a signal containing multiple speakers into individuals using the pitch of each speaker as a hook. They represent the signal using a sum of quasiperiodic signals, with a separate warping function for each quasiperiodic signal, or speaker.

It is unclear if this work can be extended to music recognition, because only the separation of the speakers was the goal. Octave errors were not considered, and the actual pitch of the signal was secondary to the signal separation. Work could be done to augment the separation procedure with a more robust or more accurate pitch estimation algorithm. The idea of a multi-pitch estimator is attractive to researchers in automatic music transcription, as such a system would be able to track and measure the overlapping pitches of polyphonic music.

4.1.5 Discussion

There is work currently being done on pitch detection and frequency detection techniques, but most of this work is merely applying new numerical or computational techniques to the original algorithms. No really new ideas seem pending, and the work being done now consists of increasing the speed of the existing algorithms.

If a technique that could accurately determine the fundamental frequency without requiring an estimation from the period or the spectrum could be found, it would change this field of research considerably. As it stands, the prevalent techniques are estimators, and require checking a number of candidates for the most likely frequency.

Frequency estimators and pitch detectors work well only on monophonic music. Once a signal has two or more instruments playing at once, determining the pitch from the frequency becomes much more difficult, and monophonic techniques such as spectrum peak detection fail. Stronger techniques such as multi-resolution analysis must be used here, and these topics will be discussed in Section 7 on page 17.

4.2 Pitch Tracking

Determining the instantaneous frequency or pitch of a signal may be a more difficult problem than needs to be solved. No time frame is independent of its neighbors, and for pseudo-periodic signals within a single note, very little change occurs from one time frame to the next. Tracking algorithms use the knowledge acquired in the last frame to help estimate the pitch or frequency in the present frame.

4.2.1 From the Spectrogram

Most of the pitch tracking techniques that are in use or under development today stem from pitch determination techniques, and these use the spectrogram as a basis. Individual time frames are linked together and information is passed from one to the next, creating a pitch contour. Windowing techniques smooth the transition from one frame to the next, and interpolation means that not every time frame needs to be analyzed. Index frames may be considered, and frames between these key frames should be processed only if changes occur between the key frames. These frames must be close enough together not to miss any rapid changes.

While improvements have been made on this idea (see [DoNa94]), the basic premise remains the same. Use the frequency obtained in the last frame as an initial approximation for the frequency in the present frame.

In [DoRo93] presented earlier, a section on fundamental frequency tracking is presented where the authors suggest the use of Hidden Markov Models. Their justification is that their fundamental frequency model is probabilistic. A discrete-time continuous-state HMM is used, with the optimal state sequence being found by the Viterbi algorithm. In their model, a state corresponds to an interval of the histogram. The conclusion that they come to is that it is possible to use HMMs on a probabilistic model to track the frequency across time frames. HMMs are also used in [DeGR93], where partials are tracked instead of the fundamental frequency, and the ultimate goal is sound synthesis. A natural sound is analyzed using Fourier methods, and noise is stripped. The partials are identified and tracked, and a synthetic sound is generated. This application is not directly related to music transcription, rather music compression, however the tracking of sound partials instead of fundamental frequency could prove a useful tool.

4.2.2 From Image Processing

The time-varying spectrogram can be considered an image, and thus image processing techniques can be applied. This analogy has its roots in psychology, where the similarity between visual and audio processing has been observed in human perception. This is discussed further in Section 2 on page 3.

In the spectrum of a single time frame, the pitch is represented as a spike or a peak for most of the algorithms mentioned above. If the spectra from consecutive frames were lined up forming the third dimension of time, the result would be a ridge representing the time-varying pitch. Edge following and ridge following techniques are common in image processing, and could be applied to the time-varying spectra to track the pitch. The reader is referred to [GoWo92] for a treatment of these algorithms in image processing. During the course of this research, no papers were found indicating the application of these techniques to pitch tracking. This may be a field worthy of exploration.

5 Segmentation

There are two types of segmentation in music transcription. A polyphonic music piece is segmented into parallel pitch streams, and each pitch stream is segmented into sequential acoustic events, or notes. If there are five instruments playing concurrently, then five different notes should be identified for each time frame. For convenience, we will refer to the note-by-note segmentation in time simply as segmentation, and we will refer to instrument melody line segmentation as separation. Thus, we separate the polyphonic music into monophonic melody streams, and then we segment these melody streams into notes.

Separation is the difference between monophonic and polyphonic music transcription. If a reliable separation system existed, then one could simply separate the polyphonic music into monophonic lines and use monophonic techniques. Research has been done on source separation using microphone arrays, identifying the different sources by the delay between microphones, however it is possible to segment polyphonic sound even when all of the sound comes from one source. This happens when we hear the flute part or the oboe part of a symphony stored on CD and played through a set of speakers. For this reason, microphone array systems will not be presented in this report, however arrays consisting of exactly two microphones could be considered physiologically correct, since the human system is binaural.

5.1 Piszczalski

The note segmentation section in Piszczalski's thesis takes the pitch sequence generated by the previous section as input. Several heuristics are used to determine note boundaries. The system begins with the boundaries easiest to perceive, and if unresolved segments exist, moves on to more computationally complex algorithms.

The first heuristic for note boundaries is silence. This is perceived by the machine as a period of time where the associated amplitude of the pitch falls below a certain threshold. Silence indicates the beginning or ending of a note, depending on whether the pitch amplitude is falling into the silence or rising out of the silence.

The next heuristic is pitch change. If the perceived pitch changes rapidly from one time frame to the next, it is likely that there is a note boundary there. Piszczalski's system uses a logarithmic scale independent of absolute tuning, with a change of one half of a chromatic step over 50 milliseconds indicating a note boundary.

These two heuristics are assumed to identify the majority of note boundaries. Other algorithms are put in place to prevent inaccurate boundary identifications. Octave pitch jumps are subjected to further scrutiny because they are often the result of instrument harmonics rather than note changes. Other scrutinizing heuristics include the rejection of frequency glitches and amplitude crevices, where the pitch or the amplitude change sufficiently to register a note boundary but then rapidly change back to their original level.

The next step is to decide on the pitch and duration for each note. Time frames with the same pitch are grouped together, and a region growing algorithm is used to pick up any stray time frames containing pitch. Abhorrent pitches in these frames are associated with the proceeding note and the frequency is ignored. The pitch of the note is then determined by averaging the pitch of all time frames in the note, and the duration is determined by counting the number of time frames in the note and finding the closest appropriate note duration (half, quarter etc.). Piszczalski's claim is that the system generates less than ten percent false positives or false negatives.

5.2 Smith

In 1994, Leslie Smith presented a paper to the Journal of New Music Research, discussing his work on sound segmentation, inspired by physiological research and auditory scene analysis [Smit94]. The system is not confined to the separation of musical notes, and uses onset and offset filters, searching for the beginnings and endings of sounds. The system works on a single audio stream, which corresponds to monophonic music.

Smith's system is based on a model of the human audio system and in particular, the cochlea. The implications are that while the system is more difficult to develop, the final goal is less a working system and more an understanding of how the human system works.

The first stage of Smith's system is to filter the sound and acquire the spectra. This closely models the human process. It is known that the cochlea is an organ that converts time waveforms into frequency waveforms.

One might ask why not use a model of the human system as a first stage in any computational pitch perception algorithm. The reason is that the cochlea uses 32 widely spaced frequency channels [Smit94]. The processing necessary to go from 32 channels to an exact pitch is very complicated, more so than the algorithms that approximate a pitch from the hundreds of channels in a Fourier spectrum. Until we know how the brain interprets the information in these channels, pitch extraction might as well use the more information available in modern spectrographic techniques.

The output of the 32 filter bands are summed to give an approximation of the total signal energy on the auditory nerve, and this signal is used to do the onset/offset filtering. Theories have been stated that human onset/offset perception is based on frequency and amplitude, either excitatory or inhibitory. Smith's simplification is to interpret all the frequencies at once, and even he considers this too simple. It is evident, however, that until we know how the brain interprets the different frequencies to produce a single onset/offset signal, this simplification is acceptable, and instructive.

The onset/offset filters themselves are drawn from image processing, and use a convolution function across the incoming signal. This requires some memory of the signal, but psychological studies have shown that human audio perception does rely on memory more than the instantaneous value of the pressure on the eardrum. The beginning of a sound is identified when the output of this convolution rises above a certain threshold, but the end of the sound is more difficult to judge. Sounds that end sharply are easy, but as a sound drifts off, the boundary is less obvious. Smith suggests placing the end of the sound at the next appropriate sound beginning, However this disagrees with Bregman's theory that a boundary can correspond to exactly one sound, and is ambiguous if applied to more than one sound.

Smith's work is intended to model the human perceptual system and to be useful on any sound. He mentions music often, because it is a sound that is commonly separated into events (notes) but the work is not directly applicable to monophonic music segmentation yet. Further study on the frequency dependent nature of the onset/offset filters of the human could lead to much more accurate segmentation procedures, as well as a deeper understanding of our own perceptual systems.

5.3 Neural Oscillators

A number of papers have recently been presented using neural nets for segmentation, specifically [Wang95], [NGIO95], and [BrCo95], as well as others in that volume. The neural net model commonly used for this task is the neural oscillator. The hypothesis is that neural oscillators are one of the structures in the brain that help us to pay attention to only one stream of audition, when there is much auditory noise going on. The model is built from single oscillators, consisting of a feedback loop between an excitatory neuron and an inhibitory neuron. The oscillator output quickly alternates between high values and low values.

The inputs to the oscillator network are the frequency channels that are employed within the ear. Delay is introduced within the lines connecting the inputs to the oscillator network, and throughout the network itself.

When an example stream of tones "High-Low-High-Low..." is presented to the network, the high tones trigger one set of frequency channels, and the low tones trigger another set of channels. If the tones are temporally close enough together, the oscillators do not have time to relax back to the original state from the previous high input and are triggered again, thus following the auditory stream. If the time between pulses is long enough, then the oscillators relax from the high tone and are excited by the low tone, making the stream seem to oscillate between high and low tones.

6 Score Generation

Once the pitch sequence has been determined and the note boundaries established, it seems an easy task to place those notes on a staff and be finished. Many commercial software programs exist to translate a note sequence, usually a MIDI file, into a musical score, but a significant problem which is still not completely solved is determining the key signature, time signature, measure boundaries, accidentals and dynamics that make a musical score complete.

6.1 MIDI

MIDI was first made commercially available in 1983 and since then has become a standard for transcribing music. The MIDI protocol was developed in response to the large number of independent interfaces that keyboard and electrical instrument manufacturers were coming up with. As the saying goes, "The nice thing about standards is that there are so many to choose from." In order to reduce the number of interfaces in the industry, MIDI, yet another standard, was introduced. It did, however, become widely accepted and while most keyboards still use their own internal interface protocol, they often have MIDI as an external interface option as well.

MIDI stands for Musical Instrument Digital Interface, and is both an information transfer protocol and a hardware specification. The communications protocol of the MIDI system represents each musical note transition as a message. Messages for note beginnings and note endings are used, and other messages include instrument changes, voice changes and other administrative messages. Messages are passed from a controlling sequencer to MIDI instruments or sound modules over serial asynchronous cables. Polyphonic music is represented by a number of overlapping monophonic tracks, each with its own voice.

Many developments have been added to MIDI 1.0 since 1983, and the reader is referred to [Rums94] for a more complete treatment.

6.1.1 MIDI and Transposition

Many researchers have realized the importance of a standard note representation. If a system can be developed that will translate from sound to MIDI, then anything MIDI-esque can then be done. The MIDI code can be played through any MIDI capable keyboard, it can be transposed, edited and displayed. The fact that MIDI is based on note onsets and offsets suggests to some researchers that transcription research should concentrate on the beginnings and endings of notes. Between the note boundaries, within the notes themselves, very little interesting is happening, and nothing is happening that a transcription system is trying to preserve, unless dynamics are of concern.

What MIDI doesn't store is information about the key signature, the time signature, measure placement and other information that is on a musical score. This information is easily inferred by an educated human listener, but computers still have problems. James Moorer's initial two-part transcription system assumed a C major key signature, and placed accidentals wherever notes were off of the C major scale. Most score generation systems today do just that. They assume a user-defined key and time signature, and place notes on the score according to these definitions.

The importance of correctly representing the key and the time signatures is shown in [Long94], where two transcriptions of the same piece are presented side by side. An experienced musician has no problem reading the correct score, but has difficulty recognizing the incorrect one, because the melody is disguised by misrepresenting its rhythm and tonality.

6.2 Key Signature

The key signature, appearing at the beginning of a piece of music, indicates which notes will be flat or sharp throughout the piece. Once the notes are identified, one can identify which notes are constantly sharp or flat, and then assign these to the key signature. Key changes in the middle of the piece are difficult for a computer to judge because most algorithms look at the piece as a whole. Localized statistics could solve this problem, but current systems are still not completely accurate.

6.3 Time Signature

Deciding where bar lines go in a piece and how many beats are in each bar is a much more difficult problem. There are an infinite number of ways of representing the rhythmical structure of a single piece of music. An example given in [Long94] suggests a sequence of 6 evenly spaced notes could be interpreted as a single bar of 6/8 time, three pairs, two triplets, a full bar followed by a half bar of 4/4 time, or even between beats. Longuet-Higgins suggests the use of musical grammars to solve this problem, which will be described in Section 8 on page 18.

Part II
Related Topics

7 Time-frequency Analysis

Most of the pitch detection and pitch tracking techniques discussed in Section 4 rely on methods of frequency analysis that have been around for a long time. Fourier techniques, pitch detectors and cepstrum analysis, for example, all look at frequency as one scale, separate from time. A frequency spectrum is valid for the full time-frame being considered, and if the windowing is not done well, spectral information "leaks" into the neighboring frames. The only way to get completely accurate spectral information is to take the Fourier transform (or your favorite spectral method) of the entire signal, and then all local information about the time signal is lost. Similarly, when looking at the time waveform, one is aware of exactly what is happening at each instant, but no information is available about the frequency components.

An uncertainty principle is at work here. The more one knows about the frequency of a signal, the less that frequency can be localized in time. The options so far have been complete frequency or complete time, using the entire signal or some small window of the signal. Is it possible to look at frequency and time together? Investigating frequency components at a more localized time without the need for windowing would increase the accuracy of the spectral methods and allow more specific processing.

7.1 Wavelets

The Fourier representation of a signal, and in fact any spectrum-type representation uses sinusoids to break down the signal. This is why spectral representations are limited to the frequency domain and cannot be localized in time: the sinusoids used to break down the signal are valid across the entire time spectrum. If the base functions were localized in time, the resulting decomposition would contain both time information and frequency information.

The wavelet is a signal that is localized in both time and frequency. Because of the uncertainty-type relation that holds between time and frequency, the localization cannot be absolute, but in both the time domain and the frequency domain, a wavelet decays to zero above or below the center time/frequency. For a mathematical treatment of wavelets and the wavelet transform, the reader is referred to [Daub90] and [Daub92].

The wavelet transform consists of decomposing the signal into a sum of wavelets of different scales. It has three dimensions: location in time of the wavelet, scale of the wavelet (location in frequency) and amplitude. The wavelet transform allows a time-frequency representation of the signal being decomposed, which means that information about the time location is available without windowing. Another way to look at it is that windowing is built in to the algorithm.

Researchers have speculated that wavelets could be designed to resemble musical notes. They have a specific frequency and a specific location in time as well as an amplitude envelope that characterizes the wavelet. If a system could be developed to model musical notes into wavelets, then a wavelet transform would be a transcription of the musical piece. A musical score is a time-frequency representation of the music. Time is represented by the forward progression through the score from left to right, and frequency is represented by the location of the note on the score.

Malden Wickerhauser contributed an article about audio signal compression using wavelets [Wick92] in a wavelet application book. This work does not deal directly with music applications, however it does have a treatment of the mathematics involved. Transcription of music can be considered lossy compression, in that the musical score representation can be used to construct an audio signal that is a recognizable approximation of the original audio file (i.e. without interpretation or errors generated during performance). The wavelet transform has also been applied as a pre-processor for sound systems, to clean up the sound and remove noise from the signal [SoWK95].

7.2 Pielemeier and Wakefield

William Pielemeier and Greg Wakefield presented a work in 1996 [PiWa96] discussing a high-resolution time-frequency representations. They argue that windowed Fourier transforms, while producing reliable estimates of frequency, are often less than what is required for musical analysis. Calculation of the attack of a note requires very accurate and short-time information about the waveform, and this information is lost when a windowed Fourier transform produces averaged information for each window. They present a system called the Modal distribution, which they show to decrease time averaging caused by windowing. For a full treatment, please see [PiWa96].

8 Musical Grammars

It has been theorized that music is a natural language like any other, and the set of rules that describe it fits somewhere in the Chomsky hierarchy of grammar. The questions are where in the hierarchy dies it fit, and what does the grammar look like? Is a grammar for 12-semitone, octaval western music different from a grammar for pentatonic Oriental music, or decametric East-Indian music? Within western music, are there different grammars for classical and modern music? Top 40 and Western? Can an opera be translated to a ballad as easily (or with as much difficulty) as German can be translated to French?

8.1 Lerdahl and Jackendoff

In 1983, Fred Lerdahl, a composer, and Ray Jackendoff, a linguist, published a book that was the result of work aimed at a challenge issued in 1973. The book is called “A Generative Theory of Tonal Music”, and the challenge was one presented by Leonard Bernstein. He advocated the search for “musical grammar” after being inspired by Chomskian-type grammars for natural language. Several other authors responded to the challenge, including Irving Singer and David Epstein, who formed a faculty seminar on Music, Linguistics and Ethics at MIT in 1974.

The book begins by presenting a detailed introduction to the concept of musical grammar, from the point of view of linguistics and artistic interpretation of music. Rhythmic grouping is discussed in the first few chapters, and tonal grammar is discussed in the last few chapters. The intent is not to present a complete grammar of all western music, but to suggest a thorough starting point for further investigations. The differences between a linguistic grammar and a musical grammar are presented in detail, and an interesting point is made that a musical grammar can have grammatical rules and *preferential* rules, where a number of grammatically correct structures are ranked in preference. The difference between a masterpiece and an uninteresting étude is the adherence to preferential rules. Both pieces are “grammatically” correct, but one somehow sounds better.

8.1.1 Rhythm

In terms of rhythmic structure, Lerdahl and Jackendoff begin by discussing the concept of a grouping hierarchy. It seems that music is grouped into motives, themes, phrases, periods and the like, each being bigger than and encompassing one or more of the previous group. So a period can consist of a number of complete phrases, each being composed of a number of complete themes and so on. This kind of grouping is more psychologically correct than sorting the piece by repetition and similarity. While one can identify similar passages in a piece quite easily, the natural grouping is hierarchical.

Where accents fall in a piece is another important observation which aids in the perception of the musical intent. Accents tend to oscillate in a self-similar strong-weak-strong-weak pattern. There is also a definite connection between the accents and the groups - Larger groups encompassing many subgroups begin with very strong accents, and smaller groups being with smaller accents.

The full set of well-formedness rules and preferential rules for the rhythm of a musical passage is presented in Appendix A. There are two sets of well-formedness and preferential rules for the rhythm of a piece, these are grouping rules and metrical structure rules.

8.1.2 Reductions

The rules quoted in Appendix A provide structure for the rhythm of a musical passage, presented here to provide the flavor of the grammar developed by Lerdahl and Jackendoff. To parse the tonality of a passage, the concept of *reduction* is needed. Much discussion and motivation stems from the reduction hypothesis, presented in [LeJa83] as:

> The listener attempts to organize all the pitch-events into a single coherent structure, such that they are heard in a hierarchy of relative importance.

A set of rules for well-formedness and preference are presented for various aspects of reduction, including time-span reduction and prolongational reduction, but in addition to the rules, a tree structure is used to investigate the reduction of a piece. Analogies can be drawn to the tree structures used in analysis of grammatical sentences, however they are not the same. The intermediate forms at different levels of the trees, if translated, would form grammatically correct musical pieces. The aim of reduction is to bit by bit strip away the flourishes and transitions that make a piece interesting until a single pitch-event remains. This single fundamental pitch-event is usually the first or last event in the group, but this not necessary for the reduction to be valid. As with linguistic reductions, the goal is first to ensure that a sentence (or passage) is grammatically correct, but more importantly, to discover the linguistic (or musical) properties associated with each word (or pitch-event). Who is the subject and who is the object of "Talk"? Is this particular C chord being used as a suspension or a resolution? It is these questions that a grammar of music tries to answer, rather than "Is this piece of music grammatically correct in our system?"

8.2 Longuet-Higgins

In a paper discussing Artificial Intelligence[Long94], Christopher Longuet-Higgins presented a generative grammar for metrical rhythms. His work convinces us that there is a close link between musical grammars and the transcription of music. If one has a grammar of music and one knows that the piece being transcribed is within the musical genre that this grammar describes, then rules can be used to resolve ambiguities in the transcription, just as grammar rules are used to resolve ambiguities in speech recognition. He calls his grammar rules "realization rules" and are reproduced here for a 4/4-type rhythm.

$\left(\frac{1}{2}\right)$-unit → $\left(\frac{1}{2}\right)$-note or $\left(\frac{1}{2}\right)$-rest
or 2 × $\left(\frac{1}{4}\right)$-units

$\left(\frac{1}{4}\right)$-unit → $\left(\frac{1}{4}\right)$-note or $\left(\frac{1}{4}\right)$-rest
or 2 × $\left(\frac{1}{8}\right)$-units

$\left(\frac{1}{8}\right)$-unit → $\left(\frac{1}{8}\right)$-note or $\left(\frac{1}{8}\right)$-rest
or 2 × $\left(\frac{1}{16}\right)$-units

$\left(\frac{1}{16}\right)$-unit → $\left(\frac{1}{16}\right)$-note or $\left(\frac{1}{16}\right)$-rest

Different rules would be needed for 3/4-type rhythms, for example, and to allow for dotted notes and other anomalies. The general idea, however, of repeatedly breaking down the rhythm into smaller segments until an individual note or rest is encountered is insightful and simple.

He also discussed tonality, but does not present a generative theory of tonality to replace or augment Lerdahl and Jackendoff's. Discussions are made instead about resolution of musical ambiguity even when the notes are known. He compares the resolution of a chord sequence to the resolution of a Necker cube, which is a two dimensional image that looks three-dimensional, as seen in Figure 2. It is difficult for an observer to be sure which side of the cube is facing out, just as it is difficult to be sure of the nature of a chord without a tonal context. He insists that a tonal grammar is essential for resolving ambiguities.

8.3 The Well-Tempered Computer

In 1994, at the same conference where [Long94] was presented, Mark Steedman presented a paper with insights into the psychological method by which people listen to and understand music, and these insights move toward a computational model of human music perception [Stee94]. A "musical pitch space" is presented, adapted from an earlier work by Longuet-Higgins.

Later in the paper Steedman presents a section entitled "Towards a grammar of melodic tonality", where he draws on the work of Lerdahl and Jackendoff. Improvements are made that simplify the general theory. In some cases, claims Steedman, repeated notes can be considered as a single note of the cumulative duration, with the same psychological effect and the same "grammatical" rules holding. In a similar case, scale progressions can be treated as single note jumps. The phrase he uses is a rather non-committal "seems more or less equivalent to". A good example of the difficulties involved in this concept is the fourth movement of Beethoven's Choral symphony, Number 9 [Beet25]. In this piece, there is a passage which contains what a listener would expect to be two quarter notes, and in fact this is often heard, but the score has a single half note. Similar examples can be made to chromatic runs in the same movement. Part of the reason that the two quarter notes are expected and often heard is that the theme is presented elsewhere in the piece, with two quarter notes instead of the half note.

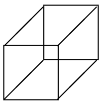
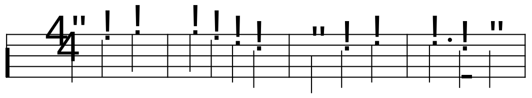
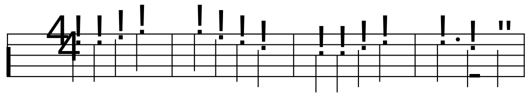


Figure 2: A Necker Cube.

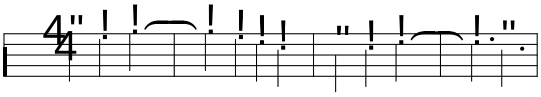
This is the way that the passage is written,



but when played, it can sound like this,



or like this,



depending on the interpretation of the listener. Since different versions of the score can be heard in the same passage, it is tempting to say that the two versions are grammatically equivalent, and that the brain cannot tell one from the other. However, it more accurate to say that the brain is being fooled by the instrumentalist. We are not confused, saying "I don't know if it is one way or the other", we are sure that it is one way and not the other, but we are not in agreement with our colleagues as to which way it is.

The substitutions suggested by Steedman can be made in some cases, but not in all cases, and the psychological similarity does not seem to be universal. It is important to discover how reliable these substitutions are before inserting them into a general theory of musical tonality.

8.4 Discussions

There seems to be such a similarity between music understanding and language understanding that solutions in one field can be used analogically to solve problems in the other field. What needs to be investigated is exactly how close these two fields really are. The similarities are numerous. The human mind receives information in the form of air pressure on the eardrums, and converts that information into something intelligible, be it a linguistic sentence or a musical phrase. Humans are capable of taking a written representation (text or score) and converting it into the appropriate sound waves. There are rules that language follows, and there seem to be rules that music follows - It is clear that music can be unintelligible to us, the best example being a random jumble of notes.

Identification of music that sounds good and music that doesn't is learned through example, just as language is. A human brought up in the western tradition of music is likely not to understand why a particular East Indian piece is especially heart-wrenching. On the other hand, music does not convey semantics. There is no rational meaning presented with music as there is with language. There are no nouns or verbs, no subject or object in music. There is, however, emotional meaning that is conveyed. There are specific rules that music follows, and there is an internal mental representation that a listener compares any new piece of music to.

9 Conclusions

Since Piszczalski's landmark work in 1986, many aspects of Computer Music Analysis have changed considerably, while little progress has been made in other areas. Transcription of monophonic music was solved before 1986, and since then improvements to algorithms have been made, but no really revolutionary leaps. Jimmy Kapadia's M.Sc. Thesis defended in 1995 had little more than Piszczalski's ideas from 1986 in it [Kapa95]. Hidden Markov Models have been applied to pitch tracking, new methods in pitch perception have been implemented, and cognition and perception have been applied to the task of score generation. Polyphonic music recognition, however, remains a daunting task to researchers. Small subproblems have been solved, and insight gained from computer vision and auditory scene analysis, but the problem remains open.

Much work remains to be done in the field of key signature and time signature recognition, and a connection needs to be drawn between the independent research in musical grammars and music transcription, before a complete working system that even begins to model the human system is created.

The research needed to solve the music understanding problem seems to be distributed throughout many other areas. Linguistics, psychoacoustics and image processing have much to teach us about music. Perhaps there are more areas of research that are also worth investigating.

10 References

[Beet25] Beethoven, Ludwig Van. *Symphony No. 9 D Minor, Op. 125, Edition Eulenburg.* London: Ernst Eulenburg Ltd., 1925. First Performed: Vienna, 1824.

[BrCo94] Brown, Guy J. and Cooke, Martin. "Perceptual Grouping of Musical Sounds:

A Computational Model." J. New Music Research, Vol. 23, 1994, pp 107-132.

[BrCo95] Brown, Guy J. and Cooke, Martin. "Temporal Synchronization in a Neural Oscillator Model of Primitive Auditory Scene Analysis." IJCAI Workshop on Computational Auditory Scene Analysis, Montréal, Québec, August 1995, pp 40-47.

[BrPu92] Brown, Judith C. and Puckette, Miller S. "Fundamental Frequency Tracking in the Log Frequency Domain Based on Pattern Recognition." J. Acoust. Soc. Am., Vol. 92, No. 4, October 1992, p 2428.

[Breg90] Bregman, Albert S. *Auditory Scene Analysis.* Cambridge: MIT Press, 1990.

[ChSM93] Chazan, Dan, Stettiner, Yoram and Malah, David. "Optimal Multi-Pitch Estimation Using the EM Algorithm for Co-Channel Speech Separation." IEEE-ICASSP 1993, Vol. II, pp 728-731.

[Daub90] Daubechies, Ingrid. "The wavelet transform, Time-Frequency Localization and Signal Analysis." IEEE Trans. Information Theory, Vol. 36, No. 5, 1990, pp 961-1005.

[Daub92] Daubechies, Ingrid. *Ten Lectures on Wavelets.* Society for Industrial and Applied Mathematics, 1992.

[DeGR93] Depalle, Ph., García, G. and Rodet, X. "Tracking of Partials for Additive Sound Synthesis Using Hidden Markov Models." IEEE-ICASSP 1993, Vol. I, pp 225-228.

[DoNa94] Dorken, Erkan and Nawab, S. Hamid. "Improved Musical Pitch Tracking Using Principal Decomposition Analysis." IEEE-ICASSP 1994, Vol. II, pp 217-220.

[DoRo91] Doval, Boris and Rodet, Xavier. "Estimation of Fundamental Frequency of Musical Sound Signals." IEEE-ICASSP 1991, pp 3657-3660.

[DoRo93] Doval, Boris and Rodet, Xavier. "Fundamental Frequency Estimation and Tracking Using Maximum Likelihood Harmonic Matching and HMMs." IEEE-ICASSP 1993, Vol. I, pp 221-224.

[GoWo92] Gonzales, R. and Woods, R. *Digital Image Processing.* Don Mills: Addison Wesley, 1992.

[GrBl95] Grabke, Jörn and Blauert, Jens. "Cocktail-Party Processors Based on Binaural Models." IJCAI Workshop on Computational Auditory Scene Analysis, Montréal, Québec, August 1995, pp 105-110.

[KaHe95] Kapadia, Jimmy H. and Hemdal, John F. "Automatic Recognition of Musical Notes." J. Acoust. Soc. Am., Vol. 98, No. 5, November 1995, p 2957.

[Kapa95] Kapadia, Jimmy H. *Automatic Recognition of Musical Notes.* M.Sc. Thesis, University of Toledo, August 1995.

[Kata96] Katayose, Haruhiro. "Automatic Music Transcription." Denshi Joho Tsushin Gakkai Shi, Vol. 79, No. 3, 1996, pp 287-289.

[Kuhn90] Kuhn, William B. "A Real-Time Pitch Recognition Algorithm for Music Applications." Computer Music Journal, Vol. 14, No. 3, Fall 1990, pp 60-71.

[Lane90] Lane, John E. "Pitch Detection Using a Tunable IIR Filter." Computer Music Journal, Vol. 14, No. 3, Fall 1990, pp 46-57.

[LeJa83] Lerdahl, Fred and Jackendoff, Ray. *A Generative Theory of Tonal Music.* Cambridge: MIT Press, 1983.

[**Long94**] Longuet-Higgins, H. Christopher. "Artificial Intelligence and Music Cognition." Phil. Trans. R. Soc. Lond. A, Vol. 343, 1994, pp 103-113.

[**Mcge89**] McGee, W. F. "Real-Time Acoustic Analysis of Polyphonic Music." Proceedings ICMC 1989, pp 199-202.

[**MeCh91**] Meillier, Jean-Louis and Chaígne, Antoine. "AR Modeling of Musical Transients." IEEE-ICASSP 1991, pp 3649-3652.

[**MeOM95**] Meddis, Ray and O'Mard, Lowel. "Psychophysically Faithful Methods for Extracting Pitch." IJCAI Workshop on Computational Auditory Scene Analysis, Montréal, Québec, August 1995, pp 19-25.

[**Moor77**] Moorer, James A. "On the Transcription of Musical Sound by Computer." Computer Music Journal, November 1977, pp 32-38.

[**Moor84**] Moorer, James A. "Algorithm Design for Real-Time Audio Signal Processing." IEEE-ICASSP 1984, pp 12.B.3.1-12.B.3.4.

[**NGIO95**] Nakatani, T., Goto, M., Ito, T. and Okuno, H.. "Multi-Agent Based Binaural Sound Stream Segregation." IJCAI Workshop on Computational Auditory Scene Analysis, Montréal, Québec, August 1995, pp 84-90.

[**Oven88**] Ovans, Russell. *An Object-Oriented Constraint Satisfaction System Applied to Music Composition.* M.Sc. Thesis, Simon Fraser University, 1988.

[**PiGa79**] Piszczalski, Martin and Galler, Bernard. "Predicting Musical Pitch from Component Frequency Ratios." J. Acoust. Soc. Am., Vol. 66, No. 3, September 1979, pp 710-720.

[**Pisz77**] Piszczalski, Martin. "Automatic Music Transcription." Computer Music Journal, November 1977, pp 24-31.

[**Pisz86**] Piszczalski, Martin. *A Computational Model for Music Transcription.* Ph.D Thesis, University of Stanford, 1986.

[**PiWa96**] Pielemeier, William J. and Wakefield, Gregory H. "A High-Resolution Time-Frequency Representation for Musical Instrument Signals." J. Acoust. Soc. Am., Vol. 99, No. 4, April 1996, pp 2383-2396.

[**QuEn94**] Quirós, Francisco J. and Enríquez, Pablo F-C. "Real-Time, Loose-Harmonic Matching Fundamental Frequency Estimation for Musical Signals." IEEE-ICASSP 1994, Vol. II, pp 221-224.

[**Rich90**] Richard, Dominique M. "Gödel Tune: Formal Models in Music Recognition Systems." Proceedings ICMC 1990, pp 338-340.

[**Road85**] Roads, Curtis. "Research in Music and Artificial Intelligence." ACM Computing Surveys, Vol. 17, No. 5, June 1985.

[**Rowe93**] Rowe, Robert. *Interactive Music Systems.* Cambridge: MIT Press, 1993.

[**Rums94**] Rumsey, Francis. *MIDI Systems and Control.* Toronto: Focal Press, 1994.

[**SaJe89**] Sano, Hajime and Jenkins, B. Keith. "A Neural Network Model for Pitch Perception." Computer Music Journal, Vol. 13, No. 3, Fall 1989, pp 41-48.

[**Sche95**] Scheirer, Eric. "Using Musical Knowledge to Extract Expressive Performance Information from Audio Recordings." IJCAI Workshop on Computational Auditory Scene Analysis, Montréal, Québec, August 1995, pp 153-160.

[Smit94] Smith, Leslie S. "Sound Segmentation Using Onsets and Offsets." J. New Music Research, Vol. 23, 1994, pp 11-23.

[SoWK95] Solbach, L., Wöhrmann, R. and Kliewer, J. "The Complex-Valued Wavelet Transform as a Pre-processor for Auditory Scene Analysis." IJCAI Workshop on Computational Auditory Scene Analysis, Montréal, Québec, August 1995, pp 118-124.

[Stee94] Steedman, Mark. "The Well-Tempered Computer." Phil. Trans. R. Soc. Lond. A, Vol. 343, 1994, pp 115-131.

[Tang88] Tanguiane, Andranick. "An Algorithm For Recognition of Chords." Proceedings ICMC 1988, pp 199-210.

[Tang93] Tanguiane, Andranick. *Artificial Perception and Music Recognition.* Lecture notes in Artificial Intelligence 746, Germany: Springer-Verlag, 1993.

[Tang95] Tanguiane, Andranick. "Towards Axiomatization of Music Perception." J. New Music Research, Vol. 24, 1995, pp 247-281.

[Wang95] Wang, DeLiang. "Stream Segregation Based on Oscillatory Correlation." IJCAI Workshop on Computational Auditory Scene Analysis, Montréal, Québec, August 1995, pp 32-39.

[Wick92] Wickerhauser, Malden V. "Acoustic Signal Compression with Wavelet Packets." in *Wavelets - A Tutorial in Theory and Applications.* C. K. Chui (ed.), Academic Press, 1992, pp 679-700.

Part III
Appendices

A Musical Grammar Rules from [LeJa83]

Grouping Well-Formedness Rules:

GWFR1: Any contiguous sequence of pitch-events, drum beats, or the like can constitute a group, and only contiguous sequences can constitute a group.

GWFR2: A piece constitutes a group.

GWFR3: A group may contain smaller groups.

GWFR4: If a group $G1$ contains part of a group $G2$, it must contain all of $G2$.

GWFR5: If a group $G1$ contains a smaller group $G2$, then $G1$ must be exhaustively partitioned into smaller groups.

Grouping Preference Rules:

GPR1: Avoid analyses with very small groups - the smaller, the less preferable.

GPR2 (Proximity): Consider a sequence of four notes $n_1n_2n_3n_4$. All else being equal, the transition $n_2 - n_3$ may be heard as a group boundary if

a. (Slur/Rest) the interval of time from the end of n_2 to the beginning of n_3 is greater than that from the end of n_1 to the beginning of n_2 and that from the end of n_3 to the beginning of n_4, or if

b. (Attack-Point) the interval of time between the attack points of n_2 and n_3 is greater than that between the attack points of n_1 and n_2 and that between the attack points of n_3 and n_4.

GPR3 (Change): Consider a sequence of four notes $n_1 n_2 n_3 n_4$. All else being equal, the transition $n_2 - n_3$ may be heard as a group boundary if

a. (Register): the transition $n_2 - n_3$ involves a greater intervallic distance than both $n_1 - n_2$ and $n_3 - n_4$, or if

b. (Dynamics): the transition $n_2 - n_3$ involves a change in dynamics and $n_1 - n_2$ and $n_3 - n_4$ do not, or if

c. (Articulation): the transition $n_2 - n_3$ involves a change in articulation and $n_1 - n_2$ and $n_3 - n_4$ do not, or if

d. (Length): n_2 and n_3 are of different lengths and both pairs n_1, n_2 and n_3, n_4 do not differ in length.

(One might add further cases to deal with such things and change in timbre or instrumentation)

GPR4 (Intensification): Where the effects picked out by GPRs 2 and 3 are relatively more pronounced, a larger-level boundary may be placed.

PR5 (Symmetry): Prefer grouping analyses that most closely approach the ideal subdivision of groups into two parts of equal length.

GPR6 (Parallelism): Where two or more segments of the music can be construed as parallel, they preferably form parallel parts of groups.

GPR7 (Time Span and Prolongational Stability): Prefer a grouping structure that results in more stable time-span and/or prolongational reductions.

Metrical Well-Formedness Rules:

MWFR1: Every attack point must be associated with a beat at the smallest metrical level at that point in the piece.

MWFR2: Every beat at a given level must also be a beat at all smaller levels present at that point in the piece.

MWFR3: At each metrical level, strong beats are spaced either two or three beats apart.

MWFR4: The tactus and immediately larger metrical levels must consist of beats equally spaced throughout the piece. At subtactus metrical levels, weak beats must be equally spaced between the surrounding strong beats.

Metrical Preference Rules:

MPR1 (Parallelism): Where two or more groups can be construed as parallel, they preferably receive parallel metrical structure.

MPR2 (Strong Beat Early): Weakly prefer a metrical structure in which the strongest beat in a group appears early in the group.

MPR3 (Event): Prefer a metrical structure in which beats of level L_i that coincide with the inception of pitch-events are strong beats of L_i.

MPR4 (Stress): Prefer a metrical structure in which beats of level L_i that are stressed are strong beats of L_i.

MPR5 (Length): Prefer a metrical structure in which a relatively strong beat occurs at the inception of either

a. a relatively long pitch-event,

b. a relatively long duration of a dynamic,

c. a relatively long slur,

d. a relatively long pattern of articulation

e. a relatively long duration of a pitch in the relevant levels of the time-span reduction, or

f. a relatively long duration of a harmony in the relevant levels of the time-span reduction (harmonic rhythm).

MPR6 (Bass): Prefer a metrically stable bass.

MPR7 (Cadence): Strongly prefer a metrical structure in which cadences are metrically stable; that is, strongly avoid violations of local preference rules within cadences.

MPR8 (Suspension): Strongly prefer a metrical structure in which a suspension is on a stronger beat than its resolution.

MPR9 (Time-Span Interaction): Prefer a metrical analysis that minimizes conflict in the time-span reduction.

MPR10 (Binary Regularity): Prefer metrical structures in which at each level every other beat is strong.