Bregman's Chimerae: Music Perception as Auditory Scene Analysis

Eric D. Scheirer
Machine Listening Group, MIT Media Laboratory

Abstract
Research into the perception and cognition of music listening often contains implicit assumptions about the nature of the underlying mental representations, and about the relationship between "auditory processing" and "music perception". We attempt to highlight and problemitize some of these assumptions and to provide a more cognitively appropriate model for music perception and cognition, based on models from the field of Computational Auditory Scene Analysis. We provide initial evidence for the appropriateness of this model by describing several existing and novel auditory/musical phenomena. Finally, we describe some of the new questions such a model raises, and propose certain experiments which might be used to answer them.

I. Introduction

The goal of the field of music perception and cognition is to explain the human ability to map incoming acoustic data into emotional, music-theoretical, or other high-level cognitive representations, and to provide evidence from psychological experimentation for these explanations. It is crucial that the representations we develop must be perceptually and cognitively *appropriate*; that is, when taken as a whole, they must be able to explain the entire pathway of the acoustic input as it proceeds through auditory, perceptual, and cognitive processing.

Music-listening is fundamentally a perceptual process, in that the beginning of, or "input to" the mapping is a stream of real-world, acoustic data. It is essential that our models refer to acoustic data rather than only to abstractions from music theory, musicology, computer music, or other non-perceptually-based disciplines. Such abstractions may be convenient for analysis, but we often lack psychological evidence of their appropriateness.

In this paper, we discuss (in Section 2) a commonly-used metaphor for the lowest levels of the music perception process, that of **transcription**, and highlight its pervasiveness throughout many domains of music psychology research. We will argue, using evidence and examples from the study of auditory scene analysis, that such a metaphor is inappropriate, and can be insufficient to explain certain phenomena of music listening.

Section 3 describes an alternative model for the music-listening process strongly informed by similar representations from the auditory scene analysis and "computational auditory scene analysis" community, and proposes the study of the taxonomy of low-level mental representations appropriate to this model. In particular, we consider a more in-depth look at the low-level auditory and musical object described by Bregman and called by him an "auditory chimera".

In section 4, we describe some psychoacoustic and psychomusicological questions raised by the process model in section 3, and consider the types of experiments which might be conducted to try to answer them.

II. Bottom-Up Processing and Transcription

Figure 1 shows a typical view of the "auditory pathway" as implied or stated by many lines of research into the perception and cognition of music-listening. Acoustic events enter the ear as waves of varying sound-pressure level and are processed by the cochlea into streams of band-passed power levels at various frequencies. The harmonically-related peaks in the time-frequency spectrum specified by the channels of filterbank output are grouped into "notes" or "complex tones" using auditory grouping rules such as continuation, harmonicity, and common onset time. Properties of these notes such as timbre, pitch, loudness, and perhaps their rhythmic relationships over time, are determined by a low-level "music perception" facility.

Once the properties of the component notes are known, the relationships they bear to each other and to the ongoing flow of time can be analyzed, and higher-level structures such as melodies, chords, and key centers can be constructed. These high-level descriptions give rise to the final "emotive" content of the listening experience as well as other forms of high-level understanding and modeling, such as recognition, affective response, and the capacity for theoretical analysis.

In some ways local to individual components of the model, this description of music listening is quite adequate and well-documented by evidence from relevant experiments. For example, the way that multiple harmonic partials are grouped together into the percept of a complex tone is well understood [DeWitt and Crowder 1987, Huron 1991, McAdams 1984]; and, similarly, the factors which govern the formation of local harmonic or "key center" representations from a stream of melodic notes have been extensively researched [Krumhansl 1991].

However, there are two assumptions embodied in the model which should be examined more closely. The first is the explicitly mono-directional flow of data from "low-level" processes to "high-level" processes; that is, that the implication that higher-level cognitive models have little or no impact on the stages of lower-level processing.

We know from existing experimental data that this upward data-flow model is untrue in particular cases. For example, frequency contours in melodies can lead to a percept of accent structure [Thomassen 1982], which in turn leads to the belief that the accented notes are louder than the unaccented. Thus, the high-level process of melodic understanding impacts the "lower-level" process of determining the loudnesses of notes. Similarly, auditory streaming of harmonically-related simple tones, as extensively reported by Bregman, can affect their "assignment" as component partials of complex tones (e.g., [Bregman 90], p.216-219).

Many other examples of similar interaction among "levels" can be found within the relevant literature; as a result, we should conclude that this processing model is incorrect, and there is much more interaction among the stages of processing than is represented here.

The second assumption is represented in Figure 1 as the "notes" label leading from the "Auditory" to "Musical" stage of processing. In computer-music research, the process of turning a digital-audio signal into a symbolic representation of the same musical content is termed the **transcription** problem, and has received much study [Moorer 1975, Maher 1989, Scheirer 1995, Martin 1996]. By assuming that "notes" are the fundamental mental representations of all musical perception and cognition, we require that there be a transcription facility in the brain to produce them. This assumption, and especially the implicated requirement, are largely unsupported by experimental evidence.

The transcription, or note-based, model of music-listening is extremely pervasive. Nearly all research on musical timbre attempts to examine the timbre of notes [Gray 1977], studies of "tonal fusion" discover the circumstances under which notes blend together into chords [Sandell 1995], and models of rhythmic perception [Handel 1990] assume that the incoming auditory stream has already been parsed into "events", i.e., notes.

Higher-level research on cognitive music models similarly assumes an accurate parsing of the acoustic data into notes; in fact, many theories of melodic and harmonic implication [Narmour 1990, Lerdahl and Jackendoff 1983] argue from examples of musical notation alone, with little or no reference to the actualization of the music into sound and its subsequent re-interpretation by the brain. It is as if the composer's intent -- the "dots on lines" representation of the music -- could be communicated directly to the listener, without the complicated performance/listening transduction process "interfering".

An intuitive consideration of the process of listening to texturally complex music also discredits this model. We have no percept of most of the individual notes which comprise the chords and rhythms in the densely-scored inner sections of a Beethoven symphonic development. While highly-trained individuals may be able to "hear out" some of the specific pitches and timbres through a difficult process of listening and deduction, this is surely not the way in which the general experience of hearing music unfolds. Rather, as we discuss in the next section, we group many sounds together, and they are perceived and processed together.

III. Predictive Models and Auditory Events

Figure 2 shows an alternative model of the music-listening process. Rather than extracting "notes" from the auditory stream, we only highlight psychoacoustic cues in the data, such as "tracks", noisy regions, or onsets in the time-frequency spectrum. These cues are grouped into percepts as part of, not preliminary to, a circular process of predictive event formation.

In this model, predictions based on the current musical context (which is dependent on what has been previously heard, and what is known about the musical domain from innate constraints and learned acculturation) are compared against the incoming psychoacoustic cues. The agreements and/or disagreements between prediction and realization are reconciled and reflected in a new representation of the musical situation.

Note that within this model, the types of representations actually present in a mental taxonomy of musical context are as yet unspecified. One element of the internal representation of music which has been somewhat underexamined is called an **auditory chimera** by Bregman:

> [Music often wants] the listener to accept the simultaneous roll of the drum, clash of the cymbal, and brief pulse of noise from the woodwinds as *a single coherent event* with its own striking properties. The sound is chimeric in the sense that it does not belong to any single environmental object. [Bregman 1990 p. 460, emphasis added]

Again arguing from intuition, it seems likely the majority of the inner-part content of a Beethoven symphony is perceived in exactly this manner. That is, multiple non-melodic voices are grouped together into a single virtual "orchestral" sound object which has certain properties analogous to "timbre" and "harmonic implication", and which is, crucially, *irreducible* into perceptually smaller units. It is the combined and continuing experience of these "chimeric" objects which gives the music its particular quality in the large -- that is, what the music "sounds like" on a global level. In fact, it seems likely that a good portion of the harmonic and textural impact of a piece of complex music is carried by such objects.

An auditory example, which to our knowledge has not been presented before, shows how at least one particular type of harmonic understanding depends integrally on very low-level principles of auditory scene analysis, and thus provides evidence that harmonic understanding can be a "low-level" process in some ways.. Consider a stream of alternating eighth notes, [... C - C# - C - C# - C ...]. The percept is of an undulating melodic line. At some point during this stream, we play, in addition to the stream, the A below and E above the line for an eighth note. Thus, depending on whether this dyad is synchronized with the C or the C#, we present either an A major or an A minor triad for a moment.

At slow tempi of presentation, this is exactly the percept -- a stream of eighth notes with a triad played in the middle. However, as the tempo increases, the chord quality becomes indistinct, and the percept of the interruption becomes simply that of an open fifth. That is, the *harmonic* implication of the musical structure is influenced by the auditory streaming of the component notes.

It is hard to see how such an experience could be explained with a purely bottom-up model of harmonic perception, since, at the particular moment the chord is played, the acoustic input is the same as that of the chord presented in isolation. Similarly, there is no explanation in the music notation, or in a music-theoretical analysis, to explain this phenomenon. It is only the *perceptual* impact of what has preceded (and, perhaps, what has followed) that changes the harmonic perception.

Models such as these are being extensively examined in the so-called Computation Auditory Scene Analysis community [Ellis 1996, Slaney 1995, Oppenheim and Nawab 1992], in an attempt to build real-time computer systems capable of understanding their acoustic surroundings, separate speech of multiple simultaneous speakers, index sound databases, and so forth. It seems likely that similar computational systems could be built for real-time analysis of digital music input.

IV. Questions and Experiments

To examine the validity of the above-claimed processing model, and to confirm the importance of the chimeric sound object, there are some naturally-implied questions that should be answered through empirical experimentation:

- Are there good musical analogies to the linguistic phenomenon of "phonemic restoration"?
- When a number of harmonic partials are presented together, how are they grouped into auditory objects? How does "spectral fusion" of partials into notes relate to "blend" of notes into chord objects?

- Under what circumstances can single notes and monophonic lines be perceived or "followed" in a piece of music; under what circumstances is this difficult or impossible?

- How hard is it for humans to hear out pitches from chords?   What is the percept of a "chord"? (Is it just the sum of the percepts of the component notes?)

- What are the fundamental units of harmonic perception? How does the percept of "key" as derived from melodic understanding resemble and differ from one derived from chord presentation?

It is our hope that careful examination of these and other questions in this area will lead to a better understanding of music perception and cognition, and a fruitful dialogue on the nature of the music-listening process.

References

Bregman, Albert, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

DeWitt, Lucinda, and Robert Crowder, "Tonal fusion of consonant musical intervals". *Perception and Psychophysics* 41 (1), 1987.

Ellis, Dan, *Prediction-Driven Computational Auditory Scene Analysis.* Ph.D. dissertation, MIT Media Lab, 1996.

Gray, J.M, "Multidimensional perceptual scaling of musical timbres". *J. Acoust. Soc. America* 61 (5), 1977.

Handel, Stephen, *Listening*. Cambridge, MA: MIT Press, 1989.

Huron, David, "Tonal consonance versus tonal fusion in polyphonic sonorities". *Music Perception* 9 (2), 1991.

Krumhansl, Carol, *Cognitive Foundations of Musical Pitch.* Oxford: Oxford University Press, 1991.

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

McAdams, Stephen, *Spectral Fusion, Spectral Parsing, and the Formation of Auditory Images*. Ph.D. Dissertation, CCRMA Stanford University, 1984.

Maher, Robert, *An Approach for the Separation of Voices in Composite Musical Signals.* Ph.D. dissertation, University of Illinois, 1989.

Martin, Keith, "A blackboard system for automatic transcription of simple polyphonic music". Perceptual Computing Technical Report #385, MIT Media Lab, July 1996.

Moorer, James, *On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer.* Ph.D. dissertation, CCRMA Stanford University, 1975.

Narmour, Eugene, *The Analysis and Cognition of Basic Melodic Structures*. Chicago: Chicago University Press, 1990.

Oppenheim, Albert, and S. Hamid Nawab, *Symbolic and Knowledge-Based Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1992.

Plompt, R. and W. J. M. Levelt, "Tonal consonance and critical bandwidth". *J. Acoust. Soc. America* 38, 1966.

Sandell, Greg, "Roles for spectral centroid and other factors in determining "blended" instrument pairings in orchestration". *Music Perception* 13(2), 1995.

Slaney, Malcolm, "A critique of pure audition" In *Readings in Computational Auditory Scene Analysis,* Hiroshi Okuno and David Rosenthal, eds. New York: Erlbaum Publishing, Inc., 1996 (forthcoming).

Scheirer, Eric, "Extracting expressive performance information from recorded music". MS Thesis, MIT Media Lab, 1995.

Thomasson, J.M. "Melodic accent: Experiments and a tentative model". *Journal of the Acoustical Society of America* 71 (6), 1982.

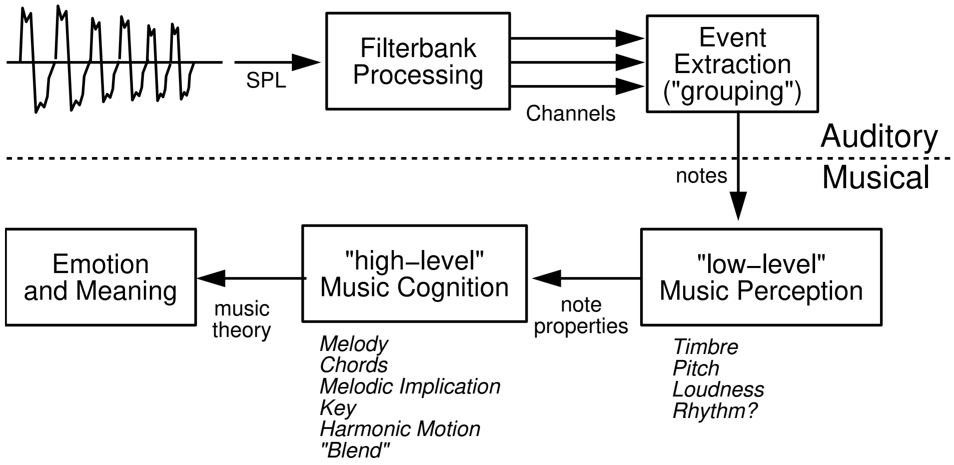

Figure 1. A bottom-up model of musical perception and cognition. Boxes contain "facilities" or processes which operate on streams of input and produce streams of output. Arrows denote these streams and are labeled with a rough indication of the types of information they might contain. Italicized labels beneath the "music perception" and "music cognition" boxes indicate into which of these categories various musical properties might fall.

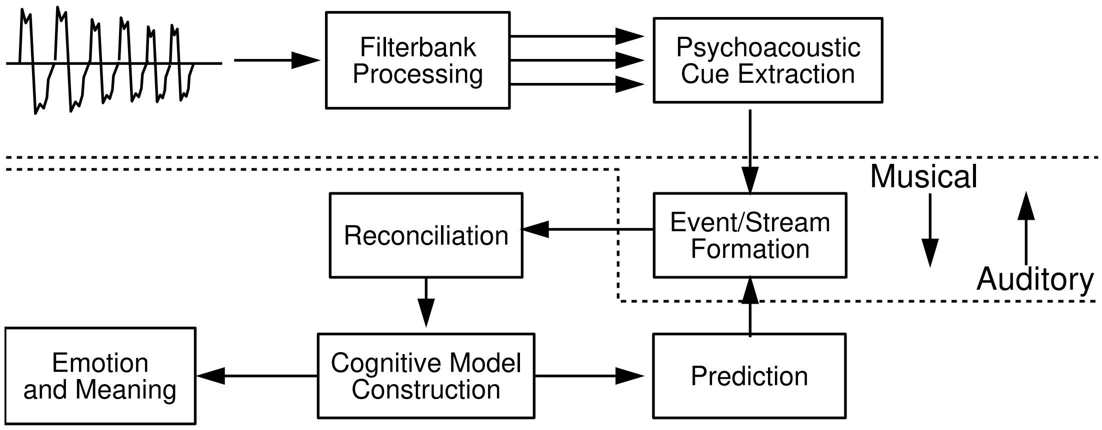

Figure 2: A "top-down" or "prediction-driven" model of music perception and cognition (adapted from a similar model in [Ellis 1996]). Boxes again represent processing facilities; arrows are unlabeled to indicate less knowledge about the exact types of information being passed from box to box.