Music Content Analysis through Models of Audition

Keith D. Martin, Eric D. Scheirer, and Barry L. Vercoe
MIT Media Laboratory Machine Listening Group
Cambridge MA USA
{kdm,eds,bv}@media.mit.edu

ABSTRACT

The direct application of ideas from music theory and music signal processing has not yet led to successful musical multimedia systems. We present a research framework that addresses the limitations of conventional approaches by questioning their (often tacit) underlying principles. We discuss several case studies from our own research on the extraction of musical rhythm, timbre, harmony, and structure from complex audio signals; these projects have demonstrated the power of an approach based on a realistic view of human listening abilities. Continuing research in this direction is necessary for the construction of robust systems for music content analysis.

INTRODUCTION

Most attempts to build music-analysis systems have tried hard to respect the conventional wisdom about the structure of music. However, this model of music—based on notes grouped into rhythms, chords and harmonic progressions—is really only applicable to a restricted class of listeners; there is strong evidence that non-musicians do not hear music in these terms. As a result, attempts to directly apply ideas from music theory and statistical signal-processing have not yet led to successful musical multimedia systems. Today's computer systems are not capable of understanding music at the level of an average five-year-old; they cannot recognize a melody in a polyphonic recording or understand a song on a children's television program. We believe that to build robust and broadly useful musical systems, we must discard entrenched ideas about what it means to listen to music and start again.

The goal of this paper is to present an unconventional view of what human listeners are able and, especially, *unable* to do when listening to music. We provide a conceptual framework that acknowledges these perceptual limitations and even exploits them for the purpose of building artificial listening systems. We hope to convince other researchers to think deeply about the limitations of conventional approaches and to consider alternatives to direct application of research results from structuralist music psychology to the construction of music-analysis systems. Many of these ideas are rooted in traditional music theory and have questionable relevance to practical issues in building real computational models. In contrast, we will present evidence from our own modeling work, which advocates a more directly psychoacoustic perspective.

The paper has three main sections. First, we present a collection of broad research goals in musical content analysis. Second, we describe a research framework that attempts to address the limitations of conventional approaches. Third, we present several case studies from our own research, demonstrating the power of an approach based on a realistic view of the abilities of human listeners.

BROAD RESEARCH GOALS

Current research in the Machine Listening Group at the MIT Media Lab addresses two broad goals simultaneously. The first is the *scientific* goal of building computer models in order to understand the properties of human perceptual processes. The second is the *practical* goal of engineering computer systems that can understand musical sound and building useful applications around them. Although the framework discussed here is wholly compatible with the broader project of general sound understanding, this paper addresses only music, and only from an application-centered perspective. For the case studies we give as examples here, there are direct parallels in our research into non-musical sound and the scientific study of auditory perception in general.

Presented at ACM Multimedia '98 Workshop on Content Processing of Music for Multimedia Applications, Bristol UK, 12 Sept 1998

Computer systems with human-like music-listening skills would enable many useful applications:

- Interaction with large databases of musical multimedia could be made simpler by annotating audio data with information that is useful for search and retrieval, by labeling salient events and by identifying sound sources in the recordings (Wold *et al.* 1996; Foote in press). The computer might understand naturally expressed queries by a user and use automatic segmentation to return only those parts of the database that are of interest to the user.
- The first generation of partly-automated structured-audio coding tools could be built. As structured audio techniques (Vercoe, Gardner and Scheirer 1998) for very-low-bitrate transmission of interactive soundtracks become widely available, it will be desirable to develop content for them rapidly.
- Robust automated musical collaboration systems could be constructed. Such systems could act as teachers and accompanists, and also as "mediators" in real-time human collaboration over large distances where communication delay would otherwise make musical interaction impossible. The more powerful and humanistic the machine listener becomes, the more sensitive and musical we can make the synthetic performer (Vercoe 1984).
- More intelligent tools could be constructed for composers, whether experienced or novice. By allowing rapid access to in-progress compositions, and control over databases of timbre, samples, and other material, a "composers' workbench" incorporating musically intelligent systems would become a powerful tool (Chafe, Mont-Reynaud and Rush 1982).

In order to build effective musical multimedia systems, it is important to understand how it is that people are capable of communicating about music and understanding the musical behavior of others. To do so requires understanding the human representations of musical sound and building computer systems with compatible representations. The goal should be to make the interface natural, allowing human users to interact in the same way they do with other humans, perhaps by humming a few bars, articulating similarity judgments, or by performing a "beat-box" interpretation of a song. The best way to understand what a human can hear in music—and thus what musical "information" humans have access to—is to study the listening process itself.

UNDERLYING PRINCIPLES

Every research agenda is based on a set of principles regarding the topic under study. Our principles differ greatly from those typically employed in musical content analysis, and we believe that our approach will yield more useful computer-music systems in both the near and far term. The following summary is not exhaustive, but it is intended to convey a sense of how our approach differs from the usual one.

The abilities of non-musicians are of primary interest.

Much of music psychology relies on experimental results from "expert" listeners. Indeed, advanced music students at the university level are often used as experimental subjects. This creates confounds in the experiments, as they are often intended to examine exactly those aspects of music (key sense, harmonic perception, pitch equivalence) that the subjects have been taught for many years. A review of psychological research using *non-musician* listeners (Smith 1997) indicates that many of the musical abilities assumed, and, unsurprisingly, detected in musicians are not present in non-musicians. Non-musicians cannot recognize intervals, do not make categorical decisions about pitch, do not understand the functional properties which theoreticians impute to chords and tones, do not recognize common musical structures, and might not even maintain octave similarity.

This is not to dismiss non-expert listeners as musically worthless; rather, it is to say that unless we want music systems to only have relevance to the skills and abilities of listeners who are graduate-level musicians, we must be cautious about the assumptions we follow. Naïve listeners can extract a great deal of information from musical sounds, and they are likely to be the primary users of many of the applications we would like to build. For example, non-musicians are generally able to identify the beat in music performed with arbitrary timbres, find and recognize the melody in a recording of a complex arrangement, determine the identity of the sound sources in a recording, divide a piece of music into sections (e.g., verse, chorus, bridge), articulate whether a piece of music is "complex" or "simple," identify the genre of a piece of music, and so forth. At present, computers cannot perform any of these tasks as well as even the least musically-sophisticated human listener.

We believe that much of the process of listening to and understanding music is simply the application of general hearing processes to a musical signal. In that regard, there is no reason to consider musicians as special; they only have stronger cognitive schemata that allow them to make different use of the same perceptual data. We assume that the abilities of a non-musician are prerequisite to, and to some extent separable from, the abilities of a musician.

"Notes" are not important.

To a non-expert listener, the acoustic structure or "surface" of a piece of music is more important than its written form. Although many music-psychological models require *transcription* (whether explicitly or implicitly) into a note-level representation prior to analysis, this transcriptive metaphor is simply an incorrect description of human perception (Scheirer 1996). For most kinds of music, there are no notes in the brain. We hear groups of notes—chords—as single objects in many circumstances, and it can be quite difficult to "hear out" whether or not a particular pitch has been heard in a chord. In the next section, we will show that, as one example, beat tracking of arbitrarily complex music does not require transcription (or even explicit "onset events").

Analysis of music into "notes" is also unnecessary for classification of music by genre, identification of musical instruments by their timbre, or segmentation of music into sectional divisions. Transcription should be viewed as a separate engineering problem, possibly of interest for practical reasons, rather than as a prerequisite for music understanding.

We must understand the constraints of the problem.

In a discussion about understanding the processes of visual perception, David Marr writes:

> "[T]rying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers: It just cannot be done. In order to understand bird flight, we have to understand aerodynamics; only then do the structure of feathers and the different shapes of birds' wings make sense." (Marr 1982, p. 27)

To take an example from audition, acoustic resonances are ubiquitous in sound-producing systems; the physical process of energy build-up within a resonance provides constraints that determine the most important features of musical instrument timbre. Similarly, the sound of the human voice (either speaking or singing) is best understood by considering the constraints of sound production from the viewpoint of an acoustic source (the airflow through the glottis) injected into a resonant system (the vocal tract).

In addition to constraints on the acoustic production process, the structure of music depends on many constraints arising in the *listener* as well. The signal-processing limitations of the physiological auditory periphery and the high-level limits of attention and memory have acted historically to shape the development of musical styles. Thus, by understanding the auditory process, we gain insight into the structure of music as well.

Listening is an activity.

Perception is causal and takes place in real time. Therefore, a listener's understanding of a piece of music at a given time cannot depend on portions that have not yet occurred; models of "music perception" which act on whole pieces or whole sections cannot be psychologically justified. Listening relies on an interaction of bottom-up (signal-driven) and top-down (prediction-driven) processes. The expectations of the listener as a piece of music unfolds in time greatly affect what is perceived. Processes of attention are important for determining what a listener is able to hear, and although we do not currently understand the mechanisms underlying attention, the study of *auditory streams* within the field of auditory scene analysis seems to be a good first step toward addressing such problems (McAdams and Bregman 1979; Bregman 1990).

Systems must be built for and tested on real music.

Too often, psychoacoustic experiments employ simple stimuli that have little relation to sounds that occur in the environment. As a result, most music analysis systems are tested only with simple stimuli, such as additive synthesis tones, Shepard tones, perfect-modulation vibratos, and so forth. If these systems are tested on ecological music signals—signals which actually occur in the real world of music—it is quickly discovered that the system cannot handle the additional complexity, noise, temporal characteristics, etc. Overly simplified stimuli are not a "building block" to the construction of functional, robust systems, but a distraction from efforts to understand real music.

There is no ground truth for music.

In speech communication, there is often a goal of communicating a particular piece of information. Speech researchers can compare the output of a speech-recognition system to a human transcription of the text to determine the effectiveness of processing. In building music systems, researchers sometimes take a similar goal to reconstruct the notated score of the music or to otherwise re-create the intuitions of theoreticians. Such models of validation must be carefully justified through experiment before they are used. The only true validation of music-listening systems is through comparison to human listening behavior and judgments. The distractions provided by music theory—for example, several systems for finding the roots of chords have been "validated" through participation in a theoretical debate regarding *one* chord in *one* piece of music—must be avoided if we care about building systems that are useful for non-music-theorist listeners.

Psychoacoustics and auditory scene analysis is the right starting point.

Since we've put most music theory and music psychology aside, one might ask what is left. We believe the perceptually-motivated study of auditory scene analysis, which deals with mixtures of sounds, is a good place to start. Psychoacoustic studies performed using real musical stimuli also have great potential. The fields of acoustics and signal processing often highlight the constraints that perceptual systems rely upon; from that viewpoint they are useful places to look for features to be used in music-understanding systems.

CASE STUDIES

In this section, we describe several case studies of building systems based on the alternative principles articulated in the foregoing. Two of these projects are complete, one is ongoing, and one is highly speculative. Each of them has more complete references elsewhere in the literature, as noted in the respective descriptions.

Speech/music discrimination

While it is not a music-analysis system *per se*, our results in building a system which could distinguish speech from music in real-time (Scheirer and Slaney 1997) illustrate the emergence of a perceptual focus in a system constructed only with engineering goals in mind.

The goal of this project was to build a system that could listen to a radio tuner and continuously classify the signal as "speech" or "music," assuming that there were no regions of overlap. To accomplish this, 13 features that were thought to be useful discriminators were selected, and they were combined using four multidimensional classification frameworks. The performance of the system was exhaustively tested using a database of speech and music excerpts collected directly off a radio tuner for this purpose. The best classifier performed with 5.8% error counting on a frame-by-frame basis, and with 1.4% error when integrating long frames (2.4 seconds) of sound.

While the overall results are not important in reflecting the perceptual principles discussed above, some post-hoc analysis is illustrative. One of the tests we conducted examining the performance of the system examined the subsets of features from which we could build good classifiers. That is, under the conditions tested, we could obtain equivalent performance from a classifier working on the full 13-dimensional feature space and a smaller one using only a subset of the features and ignoring the rest. This indicates that not all features are useful; some are poor discriminators, and some are sufficiently correlated that only one need be used.

In this analysis, we found that the best performance could be achieved with one of several three-dimensional classifiers; various three-dimensional classifiers performed statistically equivalently. Among these was the *perceptual feature set*, consisting of the three features for which there is at least anecdotal evidence of perceptual significance:

1. The *spectral centroid variance*. The spectral centroid is the "balancing point" of the spectral power distribution. It is known to be a strong correlate of perceptual "brightness" of sounds (Grey and Gordon 1978) and an important contributing factor in perception of musical "blend" (Sandell 1995). The variance of this measure calculates how stable the spectral centroid is over time; signals for which the spectral centroid is stable are likely to be music.

2. The *4 Hz modulation energy*. Speech is known to have a characteristic energy modulation peak around the 4 Hz syllabic rate. Signals with peaks at 4 Hz in the modulation spectrum (the time-varying spectrum of the signal envelope) are likely to be speech.
3. The "*pulse metric*." This was a novel feature intended to correlate with the perceptual sense of a strong rhythmic beat in the signal. The beat of a piece of music is one of the clearest features of the music to both musicians and non-musicians; if a regular beat is found in a signal, it is almost certainly a musical signal.

Thus, even in this system strictly motivated by engineering concerns, we found no solutions superior to those achieved with perceptually-motivated features.

Acoustic beat and tempo tracking

We have constructed several systems that can accurately determine tempo and locate the beat in musical signals of arbitrary polyphonic complexity and containing arbitrary timbres (Scheirer 1997; Vercoe 1997; Scheirer 1998). The analysis is performed causally, online, and in real-time, and can be used predictively to guess when beats will occur in the future. We engaged in extensive analysis and verification of the second system, demonstrating its performance on a wide variety of musical samples and comparing it to the performance of human listeners in a short validation experiment. Real music, taken directly from FM radio, was used to validate this system and compare its performance to that of human listeners.

The first of these systems (Vercoe 1997) operated by decomposing the signal using a constant-*Q* spectrogram, analyzing each channel for regions of sharply increasing energy, summing these regions across channel, and then calculating a phase-preserving narrowed autocorrelation to calculate tempo. The second (Scheirer 1998) was similar; it operated by decomposing the signal into six bands with sharply-tuned bandpass filters, and then analyzing the periodicity of each band's envelope independently using envelope detection and banks of parallel comb filters. The periodicity is detected within each channel by searching through the lag space of the multiple comb filters; the feedback delay with maximum response is selected. The estimates from the multiple subbands are combined to give an overall estimate, and then the beat phase of the signal is estimated using simple heuristics.

These algorithms perform very well on any sort of music in which the beat is conveyed clearly without a lot of syncopation. They have a difficult time analyzing signals where the beat is very slow, conveyed by instruments with slow onsets, or highly syncopated. This performance is similar to that of human listeners engaged in a "tapping task" where they tap along with the musical signals.

The method used in these systems is different from other rhythm-analysis techniques that have been presented in the literature in two ways relevant to the present discussion. Most importantly, no explicit segmentation, note analysis, or source-separation is undertaken to enable the analysis. Other systems explicitly or implicitly assume that the signal has already been, or needs to be, transcribed into sequences of note onsets before rhythmic analysis can occur. Our systems are much "closer to the signal." Rather than representing tempo analysis as the high-level juxtaposition of multiple notes output by a separation system, our tempo model views tempo itself as a fundamental low-level feature of the acoustic signal.

Additionally, there are strong connections between the analysis model developed for these systems and components of modern hearing theory (Scheirer 1997). Both our rhythm-analysis model and the "periodicity" or "rate-place" model of pitch hearing combine a front-end filterbank and rectification process with a subband periodicity analysis and cross-band integration stage. In fact, Scheirer (1997) demonstrated that existing models of pitch perception could be used directly to analyze tempo from acoustic musical signals.

The conclusion is not only that it is *possible* to incorporate a "direct perception" musical model and build a perceptually-motivated framework into beat-tracking systems, but that doing so results in a system which is *better* than other such systems reported in the literature.

Timbre classification

Musical timbre is not well understood, to the point that most recent articles discussing timbre begin by bemoaning the inadequacy of current definitions of the word (e.g., Houtsma 1997). Arguments abound about the relative importance of various features of timbre, yet no systems have been built that can recognize instruments with any meaningful generality. We are currently trying to address these issues by building a computer system that can recognize musical instruments and identify them by name. Such a system would have practical applications in automatic annotation and transcription. It might also provide insight

into the processes of sound-source recognition in humans and lead to a better understanding of what we nebulously call "timbre."

We have recently reported the results of a pilot study (Martin 1998) wherein we recorded approximately 1000 isolated instrument tones from the McGill University Master Samples collection (Opolko and Wapnick 1987), covering the entire playing ranges and a small set of articulation styles of 14 orchestral instruments. A number of perceptually-motivated features, related mostly to the resonance properties of the instruments, were extracted from the samples. Several different statistical pattern-recognition techniques were used to build classifiers and to evaluate the relative importance of the various features. Our best classifiers were able to identify the instrument family correctly for approximately 85% of the test samples, and the particular instrument for approximately 70% of the samples.

There are three aspects of this work that are of particular interest in the context of the foregoing discussion:

1. Acoustic and perceptual constraints are the keys to understanding timbre. In our system, the most valuable features for source identification are related to the speed of energy buildup—as a function of frequency—during the onset of a note. Since the buildup speed in a particular frequency region is directly related to the *Q* (ratio of center frequency to bandwidth) of the nearby resonances in the system, this "onset" feature is actually related closely to what is normally considered a "steady-state" feature of the sound. With further work, this result may eventually appease some of the debate over the relative importance of onset and steady state.

   The resonances in the *listener* are important as well. The human auditory physiology is not capable of separately analyzing harmonic partials that are closer together in frequency than approximately one third of an octave. From a practical standpoint, this means that, above roughly the sixth or seventh, partials do not have individual perceptual properties. Rather, they have *group properties* in conjunction with neighboring partials (Charbonneau (1981) has convincingly demonstrated this in a psychophysical experiment with synthetic musical-instrument tones).

2. The signal representation used for feature extraction in our system, called the log-lag correlogram, is based on a widely used scientific model of human pitch perception (Licklider 1951; Meddis and Hewitt 1991a; Meddis and Hewitt 1991b). This choice of representation was motivated by its merits as a perceptual model and by the simplicity with which it robustly encodes many salient features, including formant structure, pitch vibrato and jitter, tremolo, and onset skew. All of these features have been shown to be important for source identification by humans and for subjective judgements of timbre (see Handel 1995, for example). This feature analysis is achieved without resorting to short-time Fourier analysis, formation of sinusoidal "tracks," or assumptions about "onset" and "steady state." The *group properties* of partials are represented automatically through the use of processing that is functionally similar to that employed by the human auditory physiology.

3. There is a great deal of psychological research demonstrating that humans organize the types of objects in the world into taxonomies (Rosch *et al.* 1976; Rosch 1978). When an object is recognized, it is initially identified at a middle, or "basic," level in the taxonomy, where the most information can be gained with the least effort; for example, an animal is recognized as a "dog" before being identified as a "golden retriever." In our system, we tried several classification techniques and found that taxonomic classifiers—which start by recognizing the instrument family or articulation style and then identify the particular instrument—performed better than classifiers that identify the particular instrument directly. In addition, the taxonomy not only improved classifier performance, but also led to a less complex classifier.

We are currently at work on replicating the results of the pilot study with musical excerpts sampled directly from commercial recordings. We intend to demonstrate that the choice of perceptually-salient features and a perceptually-motivated classifier structure leads to an instrument recognition system that generalizes well to recordings by different performers in different acoustic environments—and one that requires relatively little training data.

Music perception systems

We are beginning a project in which we will use the principles described above to construct a model of the early stages of human music perception. A first approximation of the goal of this system is as follows. Consider a musically-unskilled listener turning on a radio and hearing five seconds of sound from the middle of a never-before-heard piece of music. We wish to build a system that can make some of the same judgments about this piece of music as the human listener can.

While such scenarios are not typically considered in music-psychology studies or experiments, it is clear that the listener can say many interesting things about the music that are beyond our current ability to model. The listener will be able to identify the genre of the music, discuss what other pieces or kinds of music it bears similarity to, have an emotional reaction to the music, perhaps identify the instruments in the music, perhaps sing back a certain voice in the music, verbalize a "sketch" of the music, perhaps identify the composer or performer, clap along with rhythms in the music, classify the music as "simple" or "complicated," identify social scenarios in which the music is appropriate or inappropriate, make claims about the emotive intent of the composer or performer, think of other people who might like the music, and so on.

These capabilities fall along a continuum of mental abilities, from the very simple (tapping along) to highly conceptual, cognitive, and complex (identifying appropriate social scenarios). However, it is clear that little music input, long-term stimulus structure, or musical skill is needed to use these abilities. Further, any of these skills, if robustly modeled, would be highly useful as the basis for constructing musical multimedia systems. We wish to examine those aspects of the music-listening process responsible for organizing the "surface structure" of the five-second musical excerpt into a perceptual/cognitive structure that allows for other cognitive abilities to be brought to bear.

To accomplish this, we will build a system that segregates and groups blended musical objects—the perceptual correlates of chords—from the correlogram. We will investigate the properties of these objects in perception and build statistical classifiers which can use the objects and their properties to make musical judgments.

CONCLUSION

If we ever hope to be able to build robust music analysis techniques capable of augmenting multimedia systems in a useful way, we must redirect the present focus of research. Less effort should be spent on building transcription systems, monophonic systems, single-timbre systems, and on validating them with simple synthesized examples; more effort should be spent attempting (initially struggling, to be sure) to build systems that operate directly on ecological music. Results from our own research with ecological sound signals indicate that there are many profitable avenues still to be explored.

REFERENCES

Bregman, A. (1990). *Auditory Scene Analysis*. Cambridge: MIT Press.

Chafe, C., B. Mont-Reynaud and L. Rush (1982). "Toward an intelligent editor of digital audio: recognition of musical constructs." *Computer Music Journal* **6**(1): 30-41.

Charbonneau, G. R. (1981). "Timbre and the Perceptual Effects of Three Types of Data Reduction." *Computer Music Journal* **5**(2): 10-19.

Foote, J. (in press). "An overview of audio information retrieval." *ACM Multimedia Systems Journal*.

Grey, J. M. and J. W. Gordon (1978). "Perceptual effects of spectral modifications on musical timbres." *Journal of the Acoustical Society of America* **63**(5): 1493-1500.

Handel, S. (1995). "Timbre perception and auditory object identification." In *Hearing*, B. C. J. Moore, ed. New York: Academic Press.

Houtsma, A. J. M. (1997). "Pitch and Timbre: Definition, Meaning and Use." *Journal of New Music Research* **26**: 104-115.

Licklider, J. C. R. (1951). "A duplex theory of pitch perception." *Experientia* **7**: 128-133.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: W.H. Freeman & Co.

Martin, K. D. (1998). "Toward automatic sound source recognition: identifying musical instruments." In *Proc. NATO Computational Hearing Advanced Study Institute*, Il Ciocco IT.

McAdams, S. and A. Bregman (1979). “Hearing Musical Streams.” *Computer Music Journal* **3**: 26-43, 60.

Meddis, R. and M. J. Hewitt (1991a). “Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification.” *J. Acoust. Soc. Am.* **89**: 2866-2882.

Meddis, R. and M. J. Hewitt (1991b). “Virtual pitch and phase sensitivity of a computer model of the auditory periphery. II: Phase sensitivity.” *J. Acoust. Soc. Am.* **89**: 2883-2894.

Opolko, F. and J. Wapnick (1987) McGill University Master Samples [Compact disc], Montreal, Quebec: McGill Univeristy.

Rosch, E. (1978). “Principles of Categorization.” In *Cognition and Categorization*, E. Rosch and B. B. Lloyd, ed. Hillsdale, NJ: Lawrence Erlbaum.

Rosch, E., C. B. Mervis, W. D. Gray, D. M. Johnson and P. Boyes-Braem (1976). “Basic objects in natural categories.” *Cognitive Psychology* **8**: 382-439.

Sandell, G. J. (1995). “Roles for spectral centroid and other factors in determining “blended” instrument pairings in orchestration.” *Music Perception* **13**(2): 209-246.

Scheirer, E. D. (1996). “Bregman’s chimerae: Music perception as auditory scene analysis.” In *Proc. International Conference on Music Perception and Cognition*, Montreal.

Scheirer, E. D. (1997). “Pulse tracking with a pitch tracker.” In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk, NY.

Scheirer, E. D. (1998). “Tempo and beat analysis of acoustic musical signals.” *Journal of the Acoustical Society of America* **103**(1): 588-601.

Scheirer, E. D. and M. Slaney (1997). “Construction and evalution of a robust multifeature speech/music discriminator.” In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich.

Smith, J. D. (1997). “The place of musical novices in music science.” *Music Perception* **14**(3): 227-262.

Vercoe, B. L. (1984). “The synthetic performer in the context of live performance.” In *Proc. ICMC*, Paris.

Vercoe, B. L. (1997). “Computational auditory pathways to music understanding.” In *Perception and Cognition of Music*, I. Deliège and J. Sloboda, ed. London: Psychology Press**:** 307-326.

Vercoe, B. L., W. G. Gardner and E. D. Scheirer (1998). “Structured audio: The creation, transmission, and rendering of parametric sound representations.” *Proceedings of the IEEE* **85**(5): 922-940.

Wold, E., T. Blum, D. Keislar and J. Wheaton (1996). “Content-based classification, search, and retrieval of audio.” *IEEE Multimedia* **3**(3): 27-36.