Synthetic Listeners and Synthetic Performers

Barry Vercoe
Media Laboratory
M.I.T.

We describe a new area of computer-modeled intelligence, aimed at capturing two of today's most common human activities—music perception and rhythmic response. The two are most evident in an ensemble of performing musicians, where each member is both an expert listener and an expert executor of some response—often prescribed by a score, and generally inviting the addition of expressive content.

The approach we take is to insert a computer into this mix. The goal is simple: to understand the dynamics of live ensemble performance well enough to replace any member of the group by a *synthetic performer* (computer model) so that the other live performers cannot tell the difference.

We have built several realtime models that appear to survive in such musical situations. Each is endowed with a listening skill and strategies for response that permit it to substitute for a human performer. We try to avoid systems built from reductive shorthands, such as common music notation. We want to get below the level of symbols or MIDI representations to a model which encounters music the way humans do—as an acoustic surface, vaguely but not necessarily familiar. This typically entails scanning the acoustic space for pitch and event-onset data, matching the events heard against recently-developed expectations, and (in parallel) generating a response to synchronize with immediate expectations.

The approach requires a mix of three things: *prior knowledge, good sensory perception,* and *music-interpretive instincts.* The first originates either as data from past experience or as something acquired during a performance. The others are defined by the perceptual, cognitive, and expressive acuities usually found in the better musicians (difficult), but are also surprisingly evident when untrained people are paying little conscious attention to what they are hearing.

Systems with Prior Knowledge

Our first experiments in realtime music parsing presumed considerable knowledge. In a system developed in Paris in 1982-84, and demonstrated live to the International

Computer Music Conference in late 1984, a computer tracked a live flutist via optical key sensors and acoustic microphone, and matched his performance of a *given* score with a predefined accompaniment [Vercoe, 1984]. This was notable for its extreme sensitivity to timing, due largely to two things: it had a well-defined flute score (not phrase-structured but with barlines and a meter), and it knew the relative likelihood of time displacements within a typical measure (e.g., elongation of the first sixteenth-note of a beat was likely just an expressive aberration, not necessarily implying a tempo change). The system was robust in the face of limited expression and drastic tempo changes, and was sufficiently stable to withstand dropped notes or added Baroque ornaments. However, it had no "memory" of past performances, and could not accumulate its experience—it was essentially *sight reading* on the concert stage every time.

A second implementation sought a representation of music *rehearsal and learning* [Vercoe & Puckette, 1985]. Music of the nineteenth century typifies the problem by inviting extreme expressiveness, rubato and stylistic affect—none of which is explicit in the score, but whose addition follows much the same plan at each performance. The system was accordingly given a learning capability. In accompanying a violin performance of a Kreisler work the system would initially be disturbed by the unexpected timing data, but could learn enough about the soloist's interpretation in four or five rehearsals to achieve musically acceptable synchronization.

Prior knowledge was critical to both systems. Simultaneous time shifts due to tempo and expression are hard to decode, and even harder to learn about. The experienced human listener is probably testing several solutions in parallel against some internally represented norm. For the model it was critical to know the rhythmic sequence of the performer's score, which acts as a kind of *carrier signal* upon which tempo and expressive content are separable *modulations.* Without knowledge of the carrier, both decoding and learning are difficult. In a recent implementation, the score can be entered via a MIDI performance. Although this permits *expressive* data to be entered, barline and metric information are not included, so there is no good way for the machine to learn from rehearsals.

Levels of Perception, and Instincts of Music Survival

Every music performance has its emergencies, and during these times a performer's attention to incoming auditory data can drop to near zero. A musician would not survive these times without good instincts of *rhythm.* Such background processing is expected of

the youngest player—who even as an infant could clap to a beat. Where does this ability originate, and how much processing power does it take?

The reduced level of resource allocation is not unlike building a rhythm detector using only very few parts and algorithms that must complete *in real time*. In today's technology this means simulating thousands of parallel neuronal groups on a single serial processor or small collection of processors. The success of limited per-task processing of rhythmic problems suggests that rhythmic stability has its origins in gross inexperience and inattentiveness.

Most of what humans know about rhythm is derived from polyphonic music. Because of the low conscious effort required to do multivoice audio tracking, most human musical cultures have developed complex polyphonic or polyrhythmic traditions. They have developed instruments (like the piano) whose apprehension depends on skillful signal separation, and both composers and orchestrators punctuate their melodies with rhythmic chords and percussive effects. We still know little about how the auditory system does multi-source signal separation, despite some success using a 16,384-processor Connection Machine [Vercoe, 1988]. However, rhythm does not presume perfect signal separation. Not only does the onset precede the pseudo-steady-state pitch chronologically, but it has a distinct function in such basic evolutionary needs as sound recognition and auditory localization.

Music embraces this priority. If one builds a digital pipe organ with no "chiff" attack, it fails to support the literature. Rhythmically important events in music begin with a spectral "splat," and the soloist or conductor will bring years of experience controlling these to communicate the musical intent. The basic component of a synthetic rhythm detector is a device that can sense the *relative human auditory importance* of these onsets.

Building a Synthetic Listener

We have built two systems with the capacity to analyze rapidly changing orchestral spectra. The first is an expanded Macintosh II computer, for which we have developed a Real-Time Audio Processor (RTAP) card that uses two Motorola MC56000 digital signal processing chips [Boynton & Cumming, 1988; Peterson, 1990]. Each card has 2 channels of CD quality audio in and out (AES/EBU digital audio format at 44.1 KHz), and 27 million multiply-add instructions per second (27 mips). A fully loaded MacII with 4 boards has 8 channels of audio and 108 mips of audio processing. This is host to a

library of 160 audio processing modules, including realtime auditory localization and the pitch tracker used in the Kreisler violin piece described above.

An alternative system has been developed which appears to have a greater chance of long-term maturity. The recent advent of Reduced Instruction Set Computers (RISC processors), and the rush of vendors to put these into affordable workstations, means that desk-top audio processing with sufficient power to implement the models we have discussed above will soon be commonplace. We have recently developed a realtime software audio-processing system, Csound, whose audio spectral analysis power already approaches that of the RTAP [Vercoe, 1990].

While frequency resolution is important in judging the quality of a richly orchestrated chord, overall amplitude provides the best account of metric stress. It is sufficient to model this with a bank of constant-Q, logarithmically-spaced filters, modified by Fletcher-Munson data. An implementation with 70 to 80 discrete Fourier transform (DFT) filters, each tuned to the equal-tempered notes of the piano and with a frequency/bandwidth Q of 33 has been found adequate. We currently calculate this transform every 10 or 20 milliseconds.

Inspection of auditory-nerve firing rates due to tone-bursts at different energy levels reveals a two-stage encoding that affirms the importance of onsets. We are thus led to a scheme that isolates spectral onset motion for special treatment: each filter is monitored for positive-going change, and that difference is carried in a *Positive Difference Spectrum* (PDS), in parallel with the original. The PDS is likewise shaped by Fletcher-Munson data.

Sensations of Rhythm and Meter

Researchers studying auditory masking have shown that auditory energy is absorbed over time. When the absorbing of one note is interrupted by the arrival of another in the same frequency band, the first is incomplete and *appears* to have less emphasis. This would account for the sensation of rhythmic stress in music. In fact, the most common rhythmic patterns in music, East or West, have a short-long, short-long grouping with a perceptual emphasis on the long. Reversing the perception requires special energy (at least 4 decibels) on the shorter note.

We can simulate the *integration of energy* in the total spectrum by convolving the PDS cells with a persistence function and summing across the frequency bins. The sequence

of energy estimates is then examined over a short time-interval (loosely related to short-term memory) by a network that can sense the presence of regular patterns.

The network is a narrowed form of auto-correlation [Brown & Puckette, 1989]. Autocorrelation had lost favor as a sensory analysis method, because it did not account for the high resolutions and small JND's evident in practice. Narrowed auto-correlation, which has multiple collection points, is not so limited; but the function is unfortunately zero-phase. We have developed a Phase-Preserving Narrowed Autocorrelation (PPNAC) that provides not only a sharpened account of past rhythmic activity but also a phase-correct sense of *expected musical events.* Implementation of this as a chain of energy-sensing neurons in the musically developing child is not hard to imagine.

A musically interesting feature occurs as a simple property of this analysis: combining multiple terms will also suggest rhythmic patterns that are the *harmonics* of the actual stimuli. That is to say, a regular pulse of quarter-notes will induce a weak prediction of eighth-note activity, and a weaker expectation of triplet-eighths, sixteenths, and so on. The listener perceiving a quarter-note phrase is thus already prepared for its natural subdivisions. How music moves naturally between different levels of rhythmic pattern had previously seemed mysterious.

A Basis for Response

We have described a *synthetic performer* that can participate in ensemble music it knows a lot about (has the full score or can learn from rehearsals). We then advocated theories of rhythmic perception that would permit response to rhythmically organized music not previously heard. More advanced skills—as when a jazz performer joins in a tune (chord sequence) he has never encountered—raises issues about music knowledge acquisition we can only begin to contemplate. It is likely that the multi-level *agents* described by colleague Marvin Minsky in his *Society of Mind* [Minsky, 1987] will provide a framework for future understanding of musical experience.

What we have built here is no experienced listener—nor one at times paying much attention. But the grid for rhythmic hierarchy, on which the complex web of musical organization is built, is already there. We have a foot-tapper. That our most complex art-form is based on such primitive responses would explain how the sensory-overworked musician can survive the complex society of ensembles, their rehearsals, and performances.

References

Boynton, L. & D. Cumming, (1988). "A Real-Time Acoustic Processing Card for the MacII," International Computer Music Conference, *Proceedings.* 349-356.

Brown, J. & Puckette, M. (1989). "Calculation of a Narrowed Autocorrelation Function," *J.Acoust.Soc.Am.* 85(4), 1595-1601.

Minsky, M. (1987). *The Society of Mind,* Simon and Schuster, New York.

Peterson, K. (1990). *Pseudo-Static Scheduling of Music Processing Algorithms for a Small-Scale Multiprocessor,* M.S. Thesis, M.I.T.Media Lab & Dept of Electrical Engineering.

Vercoe, B. (1984). "The Synthetic Performer in the Context of Live Performers," *Int. Computer Music Conf., Proceedings,* 199-200.

Vercoe, B. & Puckette, M. (1985) "Synthetic Rehearsal: Training the Synthetic Performer," *Int. Computer Music Conf., Proceedings,* 275-278.

Vercoe, B. (1988). "Hearing Polyphonic Music with the Connection Machine," *Proceedings,* First Workshop on A.I. and Music, AAA-88, St. Paul, MN. pp. 183-194.

Vercoe, B. (1990). "Realtime Csound: Software Synthesis with Sensing and Control," Int. Computer Music Conf., *Proceedings.*