13 Computational auditory pathways to music understanding

Barry Vercoe
Media Laboratory, MIT, Cambridge, Massachusetts, USA

When we listen to the radio we can easily distinguish the music from the talking. But could a machine? Music and speech are both structured audio, and our telling them apart stems from interpreting the audio signals at some stage of representation. If we could fully describe the interpreted representation to some other human, we could possibly describe it to a machine. Conversely, if we could instruct a machine to tell music and speech apart, and did so using only those elementary processes we believe operate in humans, we would be close to accounting for how humans apparently do it.

Telling music apart from speech is not a very high-level goal, and we might like to imbue a machine with more sophisticated musical power. We might want it to "name that tune", or identify some rhythmic style given only the acoustic signal as input. We could go further by insisting that it do this in realtime, that it take only microphone input (for ears) and need no internal audio storage (like us). If it can pitch-track and follow a score, and its response is itself a musical signal (it sings or plays), then we would have a surrogate performer, able to participate in chamber ensembles with some degree of musicianship. It might even be able to improve its performance by learning from rehearsals.

Lest the above ideas sound like fantasy, we should point out that each has already been demonstrated to varying degrees in various contexts (Large & Kolen, 1994; Richard, 1994; Todd & Lee, 1994; Vercoe & Puckette, 1985). The computational methods that made them work have taught us something about how human music cognition appears to operate, but they are all based on different abstractions of what music is (some are scores, some are just points in

time). The purpose of this chapter is to suggest how to bring many different methodologies into a single environment, to start from an original acoustic signal and proceed to the many elements of music processing in a sequence of stages. Some stages concern auditory-peripheral information processing, where the acoustic signal finds its first multiple representations and its first pre-attentive interpretations. Others concern mental representations and associated processes that are often highly attentive and prone to being driven by emotion, and affect.

The approach we will use is two-fold. We will first look at the problem from the standpoint of the Auditory Experience, surveying what we know about neural representation of music data and associated information processing, with occasional forays into computational representations of both. We will then examine Computational Representations in practice using a realtime software environment for modelling acoustic and music data processing, one that will enable us to implement our existing knowledge in realtime and then use it as a substrate for constructing and testing new theories about how music cognition appears to work.

THE AUDITORY EXPERIENCE

It takes only a few moments contemplating any music to realise that there are so many shapes and gestures in simultaneous motion that they must keep a large set of sensors and interpretors in continuous parallel work. We can see physiological evidence of this in the cochlea, in the mass of auditory nerve fibres going from it to one way-station after another. A first reaction might be one of disbelief: surely the brain cannot be coping with all that data! What we first need to understand is the nature of the parallelism, the degree to which the parallel information is sifted and simplified, and how the complex acoustic surface is reduced to a few parallel strands of semi-interpreted information. We will begin by taking a closer look at the what the cochlea apparently does.

SPECTRAL SENSORS

The human ear probes the external world with some 2,000 hair-cell sensors, each sending different firing patterns along its 20 or so attached nerve fibres. The distribution of hair cells is roughly logarithmic with frequency, save for clustering in the middle-high registers that leads to increased frequency sensitivity. It was once believed that each hair cell acted independently, sending its own report for the brain to arbitrate and sort out as best it could. Recent research has shown that much of the simplification happens right in the cochlea, and that the information passed on is already reduced to a few elements.

An example is seen in some data analysis reported by Secker-Walker and Searle (1990). When a simple speech sound was sent to the ear of a cat, and about 200 of its auditory nerve fibres were monitored for their response, a time-domain

analysis revealed the patterns shown in Fig. 13.1. (The case for humans would be similar.) In the diagram the nerve fibres are aligned by their characteristic frequency (CF, shown left in KHz), which is the frequency to which each fibre is individually most responsive. Time proceeds from left to right, and is measured in milliseconds. In response to the stimulus, the neural firing rate for each fibre (summed over dozens of repetitions) is seen to exhibit periodic bursts of activity. The period for the low fibres (CF around 250Hz) is about 8.3 milliseconds, corresponding to a speech fundamental of 120Hz. Medium-low fibres (around 700Hz) peak collectively about every 2 milliseconds (500Hz), and medium fibres (around 1600Hz) peak collectively about every 0.7 milliseconds (1400Hz). There is also collective peaking every 0.4 milliseconds (2500Hz) as well as every 8.3 milliseconds (120 Hz) in the high CF fibres.

We can interpret this data for some facts relevant to our musical needs. First, it is apparent that auditory nerve fibres do not restrict their concerns to just their CF ratings. Most of them are willing to "vote as a block" if it concerns something in their vicinity, and some blocks collectively send two or more reports when these will not be confused. The latter is interesting: the 120Hz period sent by the high CF fibres is identical to the fundamental period reported by the low CF fibres. The high CF report is due to the "beating" that occurs when two or more harmonics of the fundamental fit within a single critical band (about 1/3 octave for most filters in this region), and this will be the case for all harmonics above the sixth. What Fig. 13.1 tells us is that if we are listening to an instrument with almost no energy at the fundamental (such as the bassoon), that will be just fine since, even though the low CF fibres will have nothing to report, the high CF fibres will send a loud and clear pitch message anyway.

Of most relevance, and perhaps most surprising, is the "block voting" effect. It appears that the natural harmonics of the glottal speech stimulus have almost no representation in the cochlea reports. Instead of 120Hz being supported by 240, 360, 480, etc. we see only 500, 1400 and 2500. What are these? They are the frequencies of the speech formants (resonances), here being encoded temporally for transmission and later processing. Not only are they not coincident with the simple harmonics, but on closer inspection the three formants are slowly changing their periods in two different directions while the fundamental remains stationary. Far from sending a mass of Fourier transformed audio data for the brain to worry about, the 40,000 fibres are sending just 4 pieces of information: the perceptual fundamental (whether or not present), and the three resonant frequencies that characterise a particular vowel quality (this one being the short a).

The lesson learned here is that the essence of both pitch and musical timbre is already determined in this early processing, and we would be safe in basing our computer pitch detectors and timbral recognisers on processing of this form. The key lies in the method used to process the data, for the information sent on by the cochlea does not take the form of Fig. 13.2a, but that of Fig. 13.2b. Figure 13.2a

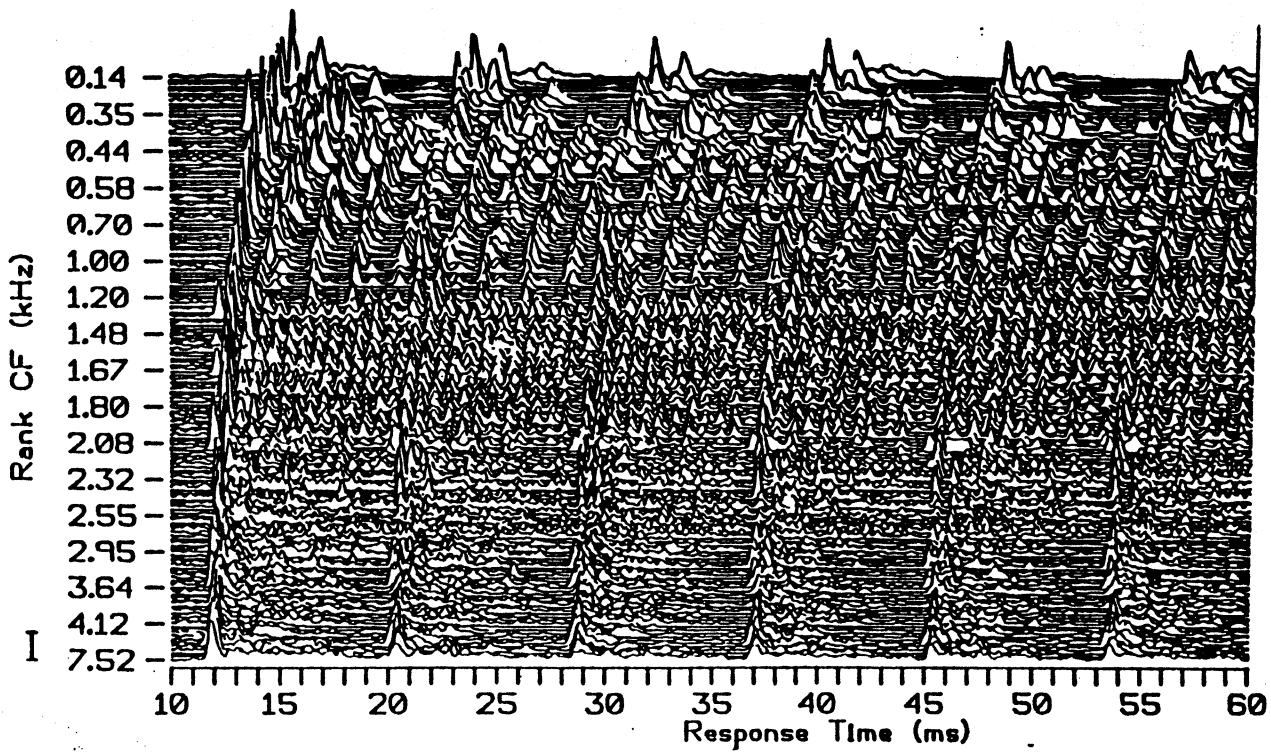

FIG. 13.1. Auditory nerve fibre response to a stopped vowel, showing broad-band synchrony across groups of fibres. Secker-Walker and Searle (1990). Used by permission.

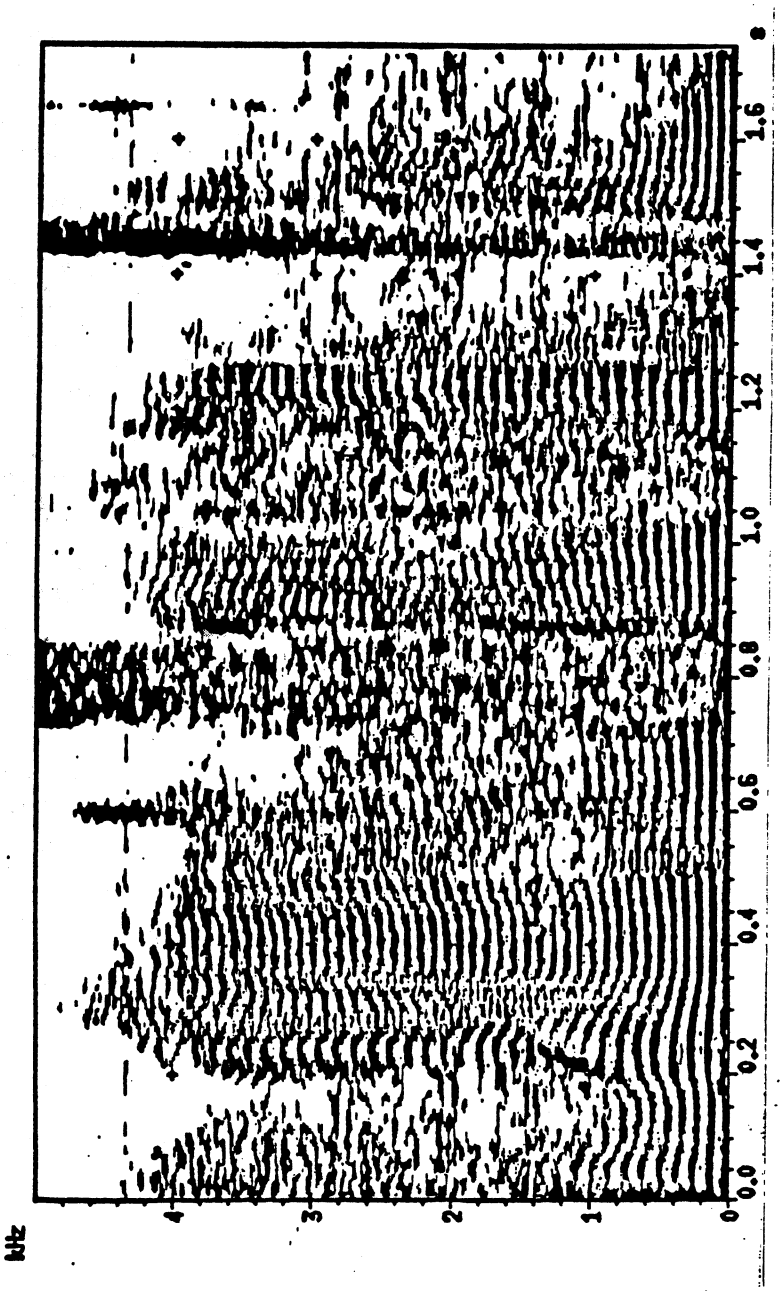

FIG. 13.2a. FFT linear-frequency spectrogram of the phrase "away in Southampton".

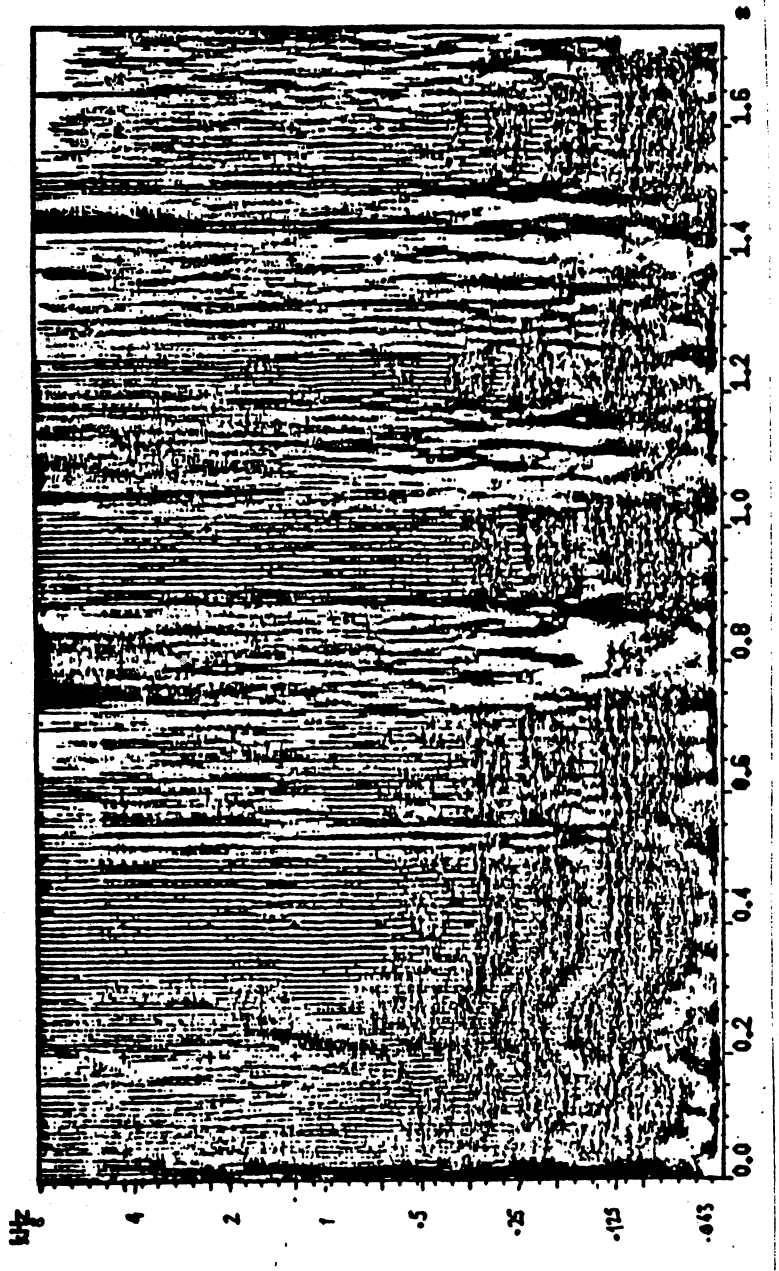

FIG. 13.2b. Constant Q log-frequency spectrogram of the same phrase.

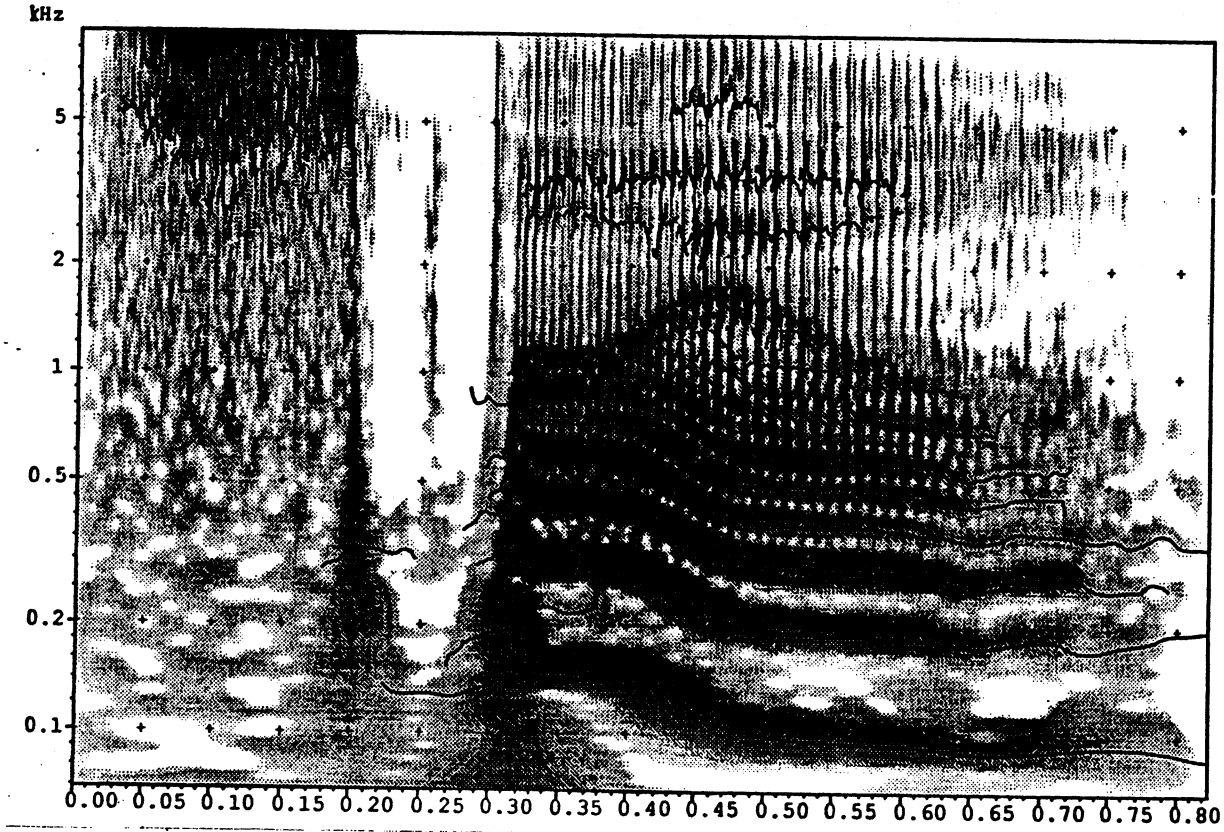

FIG. 13.2c. Constant Q log-frequency spectrogram of the word "spoil".

is a mathematically convenient view of the utterance "away in Southampton". It is a computer representation, obtained via a Fast Fourier Transform (FFT) which has computational elegance and gets the job done in record time. This kind of "spectrogram" is often still seen in texts and in legal proceedings, yet it does not represent what the cochlea reports: it has linearly spaced filter bins (y-axis) with the same bandwidth at all frequencies, while the cochlea has near logarithmic spacing of hair cell frequencies with roughly proportional bandwidths (constant ratio to the CF). The FFT gives poor frequency resolution in the lower octaves, and too much in the upper, and since its bandwidths are constant it entirely misses the "beating" that can make up for a missing fundamental.

Figure 13.2b is a logarithmic analysis of the same speech phrase. The filters are log spaced in frequency (see the KHz on the left) and constant Q (bandwidths have constant ratio to the CF). This means that medium- to high-frequency filters are broad-band, which makes them quick to respond to changes; they will also exhibit "beating" above the sixth harmonic, which accounts for the vertical striations visible in the figure. A closer view is seen in Fig. 13.2c, which shows a similar analysis of the word "spoil". The constant Q filters give separate resolution to the lower six harmonics, above which the "oi" formant motion is being pulse-modulated by the vertical striations of the broadband higher filters. (Comparison with Fig. 13.1, though upside down in frequency, is informative here.) Mammals have an abundance of broad-band filters; the fast response is ideal for sensing data where precision timing is critical. In mid-range this helps a binaural system sense the direction of a sound. At higher frequencies the pulse-time perturbations would help collect all the harmonics emanating from a soft, life-theatening footstep into a single percept. In music we have learned to capitalise on fast acting filters. An effective time-marking percussion instrument is one with a predominance of high frequencies. And the violin is successful because its broad spectral nuances are gathered aurally into a single meaningful note-event. A computer with broad-band high-frequency filters would have full access to this timing and integrating information.

SPECTRAL SEPARATORS

In addition to sensing the musical structure of a single line, humans also experience the extra dimension of polyphonic music. Because of the low conscious effort seemingly 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 composers and orchestrators have felt free to punctuate their melodic lines with rhythmic chords and percussive effects. Yet we know very little about how the auditory system does multi-source signal separation. From where amongst the mass of mixed partials does it find the grouping cues? One clue stems from the work of McAdams

(1984), who showed that if the harmonic partials of three synthetic vowel sounds are summed with no vibrato or amplitude motion, it is impossible for the human ear to separate and recognise them; yet if those of any group are given some coherent frequency motion, they will fuse into a separate recognisable vowel sound. The effect is especially strong if the motion also causes amplitude change, as when partials are being independently modified by steeply shaped resonances.

It seems that sound from a natural vibrating system conveys a signature in the form of micro-perturbations—amplitude and frequency modulations that distinguish its partials from those of other sources. We observe with characteristic hindsight that when instrument makers of the mid-1500's experimented on the early viol with motion-inducing enhancements (removal of frets to permit vibrato; an ornate body for steeper resonances) they "invented" the modern violin (aurally more separable, the future vehicle of the concerto). We also note that when two strings are driven by the same bow, the coupling is still loose enough to impart two distinct perturbation patterns to the mix.

Signal separation is a complex problem, and the challenge of emulating what the human ear does so well (both in and out of the concert hall) has been squarely confronted by researchers in the emerging field of Computational Auditory Scene Analysis. The most effective techniques first divide the signal into energy strands (using roughly constant Q filters), find the local energy peaks amongst neighbouring frequencies (block voting), then group the energy tracks into clusters which likely stem from a single source (exhibit common onset time, integer frequency ratios, and co-modulation of frequency, amplitude and phase) (Ellis, 1994). These techniques are showing much progress, and will eventually allow computers to do things like polyphonic pitch tracking and multi-timbral identification with the confidence of skilled musicians.

In the meantime, we cannot simply ignore the polyphonic problem. Although music in some cultures is primarily melodic (e.g. Native-American chant with slow drum accompaniment), that of most cultures incorporates multi-source strands in which the interpretation of one strand (either rhythmic or harmonic) is informed by its relation to events in another. Such powerful influences need some kind of representation and this remains one of our goals, even though machine separation of a polyphonic web is still inferior to that of the average music listener. We can see this most easily in the rhythmic domain.

EVENTS AND EVENT PATTERNS

What is a musical event, how is it encoded, and what induces the hierarchy of events that we describe as musical rhythm? We can look to neurophysiology for some initial assistance. While investigating auditory-nerve firing rates due to simple tone-bursts at different energy levels, Delgutte (1980) found a two-stage encoding process that appeared to treat event onsets and their steady-state

continuation in quite different ways. During event onsets the neural firing rate at first exhibits a very rapid increase, then a rapid adaptation, after which it eventually settles on a steady state acknowledgement of the event's continued presence. Significantly, the energy level (note intensity) was encoded almost entirely in the onset flash (lasting about 10 milliseconds); a second adaptation (50 milliseconds) and the steady state continuation gave only mild recognition to the intensity of the event. In subsequent research (Smith, Brachman & Goodman, 1983) it was found that an event with a slow onset receives a compromise encoding; the peak firing rate depends on the slope of the envelope and its target intensity.

What do these two neurophysiological results suggest for computational music cognition? The first can be taken almost literally. We have always known that note attacks are important: if one builds an electronic 'pipe' organ with no simulated acoustic chiff at each onset it simply fails to support the literature; and we see that composers annotate the important with fortepiano (fp) markings and percussive doublings. Consequently, a simple event detector and rhythm interpreter might encode Delgutte-like onset-only pulses (the louder the stronger), then look for patterns. But the second result above (slow onsets) suggests that something else is going on. To investigate that we turn to perceptual psychology.

Any perceived event tends to have a persistence in the perceiver. Persistence (impulse response) can model phenomena as distinct as two-tone forward masking and the perception of accents in equitone sequences. This latter was studied by Povel and Okkerman (1981), who presented subjects with equitone sequences (same frequency, intensity, timbre) with slightly different inter-onset intervals. When the inter-onset intervals (alternating long and short) differed from each other by less than 8%, the first of each adjacent note-pair seemed louder. Yet when the difference was increased beyond 8%, the accent seemed to move from the first of the pair to the second (as in a 6/8 lilt). In no case was any physical accent actually present.

So what is happening here? Presumably at least two things. The first effect can be attributed to incomplete recovery from adaptation, making the second note of each group appear softer. But what of the second effect? Given its shorter inter-onset interval, the effect is apparently due to the time we initialy need to estimate the energy in a tonal stimulus. The energy of a single pulse is integrated over some 200–300 milliseconds, and when the integration is interrupted by the arrival of another pulse the tail of the first integration is lost. The amount of loss depends on the inter-onset interval, and its significance here apparently exceeds that of incomplete recovery going into the second note. In related research by Zwislocki and Sokolich (1974), when the tones are of different frequency but in the same critical band, the tail of the first actually enhances the integration of the second, suggesting that integrators are perhaps "warmed up" by a nearby preceding tone. But either way the effect is the same: our perception of a lilting

sequence is that the first of the adjacent pair is at a lower intensity than the second.

This seemingly simple effect is one of the prime generators of human-perceived rhythm and meter. For within the window of our preferred beat size (about 600 milliseconds), there is a perceptual mechanism that artificially weights the sub-events so that the longer ones seem louder than the shorter ones. Of course, performers know about this: if asked to impart a dotted rhythm so the audience will perceive the beat on the short note, they know they must reverse this effect by giving the short note extra stress (about 4dB according to Povel and Okkerman). The energy integration curve is not a simple one. To a first approximation it is roughly exponential. A multiple time-constant has been advocated (Todd, 1994), although this possibly involves perceptual/motor constructs. Moreover, in preliminary experiments conducted by the present author the time constant differs widely amongst individuals; it is also frequency dependent, but has not yet been adequately mapped over this domain. About all we can safely say about energy integration is that it is widespread, and has a very large influence on how we hear musical patterns. And it does not occur naturally in machines.

So how can a machine possibly hear musical rhythms the way we hear them? Somehow we need to build the above effect into the collecting mechanism of our musical machine. We could elect to develop a "rule system" that would recognise the condition that causes loudness weighting, then apply some algorithmic modification of the physical signal measurement to simulate the human auditory bias. Imagine the overload, however, when the rule set is applied to a large mass of polyphonic music: would the computer have time to find the beat? A less belaboured way would be to build this auditory bias into the way spectral analysis filters encode intensity. We have examined two effects above, one in which sudden intensity is encoded in a short-lived firing-rate flash (Delgutte), the other in which slow-rising intensity and successive pulses demonstrate persistence in the ensuing central auditory system (Povel & Okkerman and Smith et al.). In all cases the important thing seems to be positive change; note releases and decays do not disturb this system in any significant way. Moreover, short-term adaptation and response to an increasing stimulus each appear to be additive, wherein the increased response is independent of the state of adaptation (Smith & Zwislocki, 1975), and several computational models have been assessed for adherence to this empirical data and for suitability as front-ends to speech recognisers (Hewitt & Meddis, 1991).

However, these can get to be computationally expensive. In order that we might have enough computational power remaining for higher-level realtime polyphonic processing on affordable machines, we might prefer to identify only the most salient causes of the above effects (although one must be vigilant of approximation errors that are amplified by later stages). We could look for a simple representation that combined positive change with impulse response

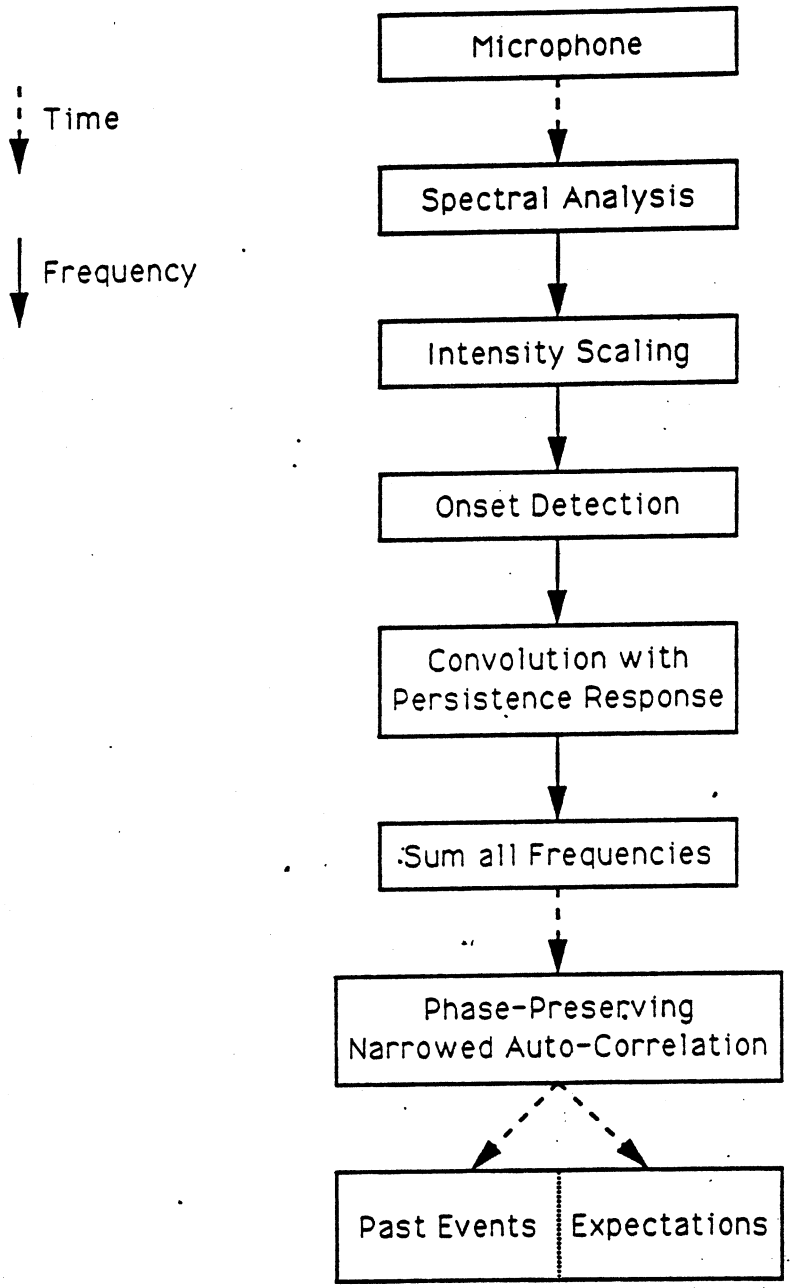

FIG. 13.3. Sequence of audio operations needed to perform auditory-based analysis of rhythmic acoustic input.

```
          sr = 32000
          kr =  1000
          ksmps = 32

          instr   1
asig      in
wsig1     spectrum asig, .01, 8, 12, 8, 0, 1, 0
wsig2     specscal wsig1, 3, 4
wsig3     specdiff wsig2
wsig4     specfilt wsig3, 5
          specdisp wsig1, .04, 0
          specdisp wsig3, .04, 0
          specdisp wsig4, .04, 0
ksum4     specsum  wsig4, 1
ktempo    tempest  ksum4,.01,.1,3,1,30,.005,.5,90,2,.04,1
koct,ka   specptrk wsig1, 1.3, 7.4, 8.9, 8.1, 25, 3, .5
koct      =        (koct > 7.4 ?  koct-7.4 : 0)
          display  koct, .04, 72
          out      asig
          endin
```

FIG. 13.4. Csound program to perform auditory-based analyis of rhythmic acoustic input.

(persistence) in each frequency channel to capture the essence of these auditory effects in a single step. We will use the notion of a Positive Difference Spectrum convolved with an Impulse Response in the representation developed below.

Meanwhile, there has been effective research into computer methods of assessing rhythmic pattern, meter, tempo and place, some using acoustic input and others using MIDI keyboard data. In an early example of realtime acoustic sensing (Vercoe, 1984) a flute was tracked using a combination of optical key sensors and realtime acoustic analysis, and the resulting event sequence compared to a predefined score. The comparison was then used to direct a realtime accompaniment to remain in sync with the soloist, who could speed up or slow down the duo performance at will. The beat tracking used both pitch and timing data, including phase locking onto the metre at two hierarchical levels, and the results were sufficiently stable to be used in public concert performances. Auditory input is also the stimulus for a beat induction model using sensory-motor filters (Todd & Lee, 1994). Two separate filters are tuned to the most natural periods for beats (600ms) and body sway (about 5 seconds), within which events are grouped by temporal proximity, and relative accent is derived from the relative and absolute distance between events.

MIDI keyboard performance data is the source for a connectionist analysis using formal models of entrainment (Large & Kolen, 1994). The event onsets serve to perturb the phase and period of a nonlinear oscillator, variably open to disturbance at certain times during its cycle. This produces a robust beat given

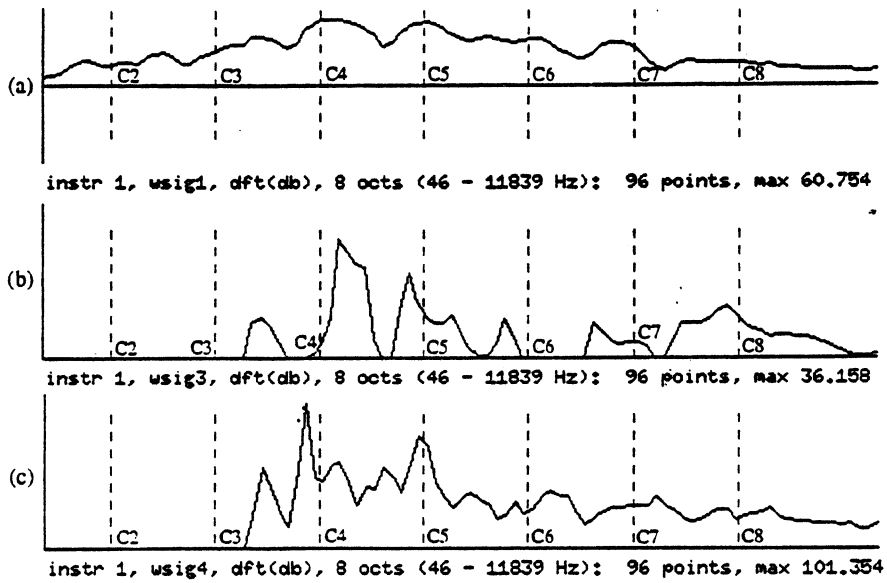

FIG. 13.5. a) Constant Q spectral snapshot of the F3/D4 dyad of the Bach excerpt (piano). b) Positive difference spectrum (PDS) reflecting the spectral changes currently occuring in a). c) Spectral output of persistence filters following the injection of all PDS data up to the moment of b).

complex metrically structured input, but the single oscillator model experiences difficulty when the input is highly expressive and contains heavy rubato.

Extracting expressive content involves a multiple task of identifying a metric grid and tempo, while also discerning the temporal warping of that grid (and oftentimes of polyphonic deviations from it) as parallel channels of communicative information. These channels are charged with expressive power, and true artists are skilled at their manipulation. In a detailed analysis of the renowned Cuban percussion ensemble Los Munequitos de Matanzas (Bilmes, 1993), it was shown that the minute deviations that constitute expressive content are themselves built from tiny temporal atoms (which the author called a Tatum), and that a computer could systematically remove and reinsert these various layers for reinterpreted performance. This line of research shows that computer analysis and representation can go beyond that for which we have either a written notation or even a fully conscious listening sense.

It will be some time before we have computer encoding of the full auditory musical experience. The point of present-day representations is two-fold: to help systematise that which we already know, and to form a basis for more exploratory research. The most demanding test of how well we are doing is de-representation (music performance based on the encoding), and some of the above techniques are happily being developed in this demanding forum. We will next examine an

encoding method that is embedded in a system designed for realtime audio signal analysis and synthesis, one in which the critical performance check is always available.

COMPUTER REPRESENTATIONS

We now describe a comprehensive environment for realtime audio processing in which the concepts introduced above can be embodied as processes operating on data. Csound is a software audio processing system with a rich array of processing modalities and the capacity to do its work in realtime (Vercoe & Ellis, 1990). It is widely distributed as freeware to research communities, is well documented, and has a large community of users who lend mutual assistance through various networks. The examples given below can be run on the standard distribution. They are not intended to prove a theory, but rather to demonstrate a unified system for modelling music perception and cognition using acoustic-only input. The reader is invited to test these and develop them further (see Appendix 7: Adding your own Cmodules to Csound, in (Vercoe, 1995)).

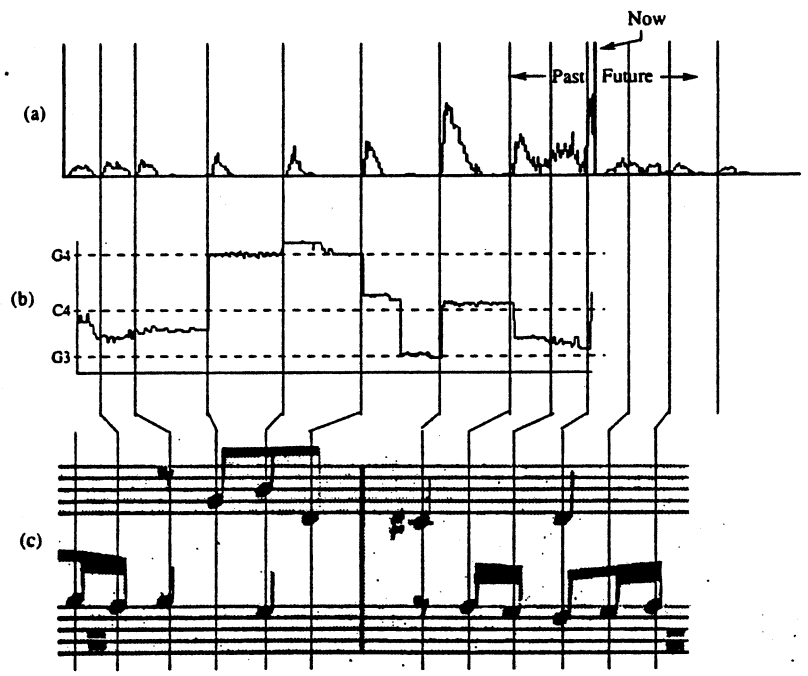

FIG. 13.6. a) Time-magnitude representation of the perceptual importance of events emerging from the summed persistence filters. The 3-second windows on either side of the central "now" represent the immediate past (left) and the expected future (right). b) Pitch analysis of the events of 13.6a. c) Score fragment of Bach Fugue used as acoustic input for the spectral, rhythmic and pitch analysis of Figures 13.5 and 13.6.

Using acoustic input, we will first develop a model for encoding auditory processes that exhibits human-like response to rhythmic structure—the phenomenon resulting from the interaction of perceptual grouping and higher-level preference rules. We will then use the structure to extract a tempo which can drive other parts of this realtime system. As a convenient guide to where we are, we will also include a pitch tracker. The overall plan is outlined in Fig. 13.3, and the program that implements it is shown in Fig. 13.4. The program syntax is a simple one: the central column describes a sequence of operations on data; the input data (if any) is on its right and the results of an operation are assigned to its left. Results appearing on the left of an operation are then available as input to any subsequent operation. There are three types of results, distinguishable by name: those beginning with "a" are audio signals, names beginning with "w" are spectral data types, and names beginning with "k" are control signals. Audio signal rates are set to carry high-fidelity audio, and control signal rates are set to convey data at roughly the speed of neural communication (about 1KHz).

Given acoustic input (a segment of Keith Jarret's performance of Bach's G minor Fugue from Book I of the WTC), accessing the signal is programmed by the monaural input operator 'in', which places audio data into 'asig'. To get a spectral representation we pass 'asig' to a spectrum analyser 'spectrum', which divides the signal into 96 bands (8 octaves, 12 bands per octave) to simulate nerve fibre transmission (each "fibre" corresponds to one semitone on a piano). The filters are constant Q with a Q of 8 (CF/bandwidth), and their output in decibels is reported as a new spectrum 'wsig1' every 0.01 seconds. Since it is known that the human auditory periphery biases its loudness assessment along Fletcher-Munson curves (due to the outer-middle ear transfer function and the non-linear spread of response along the basilar membrane), we employ the 'specscal' unit to reshape our spectral data in similar fashion. We could have done this prior to spectral filtering, but prefer to do it here because the data is now in decibels and we can use a log frequency mapping across the spectral data.

The program of Fig. 13.4 is also seen to contain three spectral display requests ('specdisp'), which are projected in this case using X11 windows on an SGI monitor. The output of these are shown in Fig. 13.5a, b, and c, but to interpret them we must first understand something of the way time is handled in this realtime program. As stated above, the 'spectrum' opcode derives a new spectral cross-section every 0.01 seconds, which is thus the rate at which the 'wsig' cells are refreshed with new information. Each unit receiving a refreshed 'wsig' cell then has work to do. The 'specdisp' units, however, are asked to display only every 0.04 seconds (or 25 times a second), and the screen images are therefore downsampled snapshots of the 0.01 spectral cross-sections. The three spectra of Fig. 13.5 are thus snapshots of momentary spectra at a specific point in the acoustic signal, that point being made clear in Fig. 13.6a, b, and c, all three of which are time displays. In Fig. 13.6a the point is labelled "now", and in Fig. 13.6c the point is the onset of the final D4 of the treble fugal entry.

Returning to the spectral processing, Fig. 13.5a is a snapshot of the momentary decibel values emanating from the 96 filters. The output is visibly smooth, due to the broad-band filters used. This would have been even smoother had we used strict 1/3-octave filters (Q of 4), but our choice of 1/6 octave (Q of 8) for our sparse, symmetric filters is to acknowledge the very steep cutoff of human auditory filters (the cochlea has about 400 overlapping bands/octave in this region, and its filters are asymetric with a cutoff on one side of hundreds of db per octave). As implied above, the spectrum of Fig. 13.5a will be Fletcher-Munson scaled before being passed to the next unit.

The task of onset detection is performed by 'specdiff'. As shown earlier, neural firing rates become intensely active during event onsets, and we have theorised that representing the onset slope by a sequence of positive changes (and ignoring the negatives) might capture this sensitivity. 'Specdiff' creates a new spectrum, comprised of just the positive changes seen in each filter channel from one spectral cross-section to the next (every 0.01 seconds). The resulting positive difference spectrum (PDS) can be seen in Fig. 13.5b. Also evident in this figure is the energy just being received from the pitches of the newest notes F3/D4.

We now inject the PDS into a set of persistence filters using 'specfilt'. These are simple recursive filters, one per spectral channel, with individual rates of sustaining new input specified in half-life values (the period for which any new input is reduced to one-half the original). The values are kept in a table (no. 5), and were set by running experiments in the style of Povel and Okkerman, extended to include frequency dependence. The effect of injecting PDS impulses into the filters is to accumulate and prolong the perceptual life of events. The result of this step can be seen in Fig. 13.5c, where the latest PDS of Fig. 13.5b has just been added to the persistence mix.

The energy in the total spectrum is now estimated by summing across the frequency bins of 'wsig4' using 'specsum'. This approximates the summation that occurs when humans assess the loudness of complex tones; there has been much work done on this (Zwicker & Fastl, 1990), and we could develop more accurate methods that incorporate frequency-related suppression. However, the real goal of our summing is to compare energies across time, so we will accept the approximation. Next, the sequence of energy estimates in 'ksum' is examined over small time-intervals using 'tempest'. We use a window size of three seconds to loosely relate it to echoic memory, but the real point of "tempest" is to detect the presence of regular patterns and to estimate the tempo of their recurrence.

The display generated by 'tempest' (Fig. 13.6a) gives the best view of its operation. Unlike the above, this display is time-based, and information is seen moving across the entire window as time passes. The display is in two parts, separated by a vertical line that represents 'now'. The data to its left is the three-second echoic memory, and information that begins at the 'now' (coincident with what you hear) moves increasingly left over this period, gradually decaying

to zero as it leaves the screen. The data on the right is a prediction of the future; information beginning small at its right will gather strength as it moves left towards the "now". There is interaction between the perceptual past and the anticipated future: as each expected event hits the "now" some portion of it is rolled into what becomes "perceived". Conversely, the pattern of already perceived events contributes via a feed-forward network to the growth of expectations.

The feed-forward network within 'tempest' is a variant of the Narrowed Auto-Correlation (NAC) algorithm developed by Brown and Puckette (1989). There is growing evidence that the auditory system uses autocorrelation in post-peripheral processing (Hartman, 1989), but the once-popular method lost favour because it did not account for the high resolutions and small JND's evident in practice. Narrowed auto-correlations do not have this problem. The general function is defined by:

$$|S_N(\tau)|^2 = < \left| f(t) + f(t-\tau) + f(t-2\tau) \ldots + f(t-\overline{N-1\tau}) \right|^2$$

and the peaks have a width of 2T/N, where T is the period (2PI/w), and N the number of terms used. The NAC method has recently been used to sharpen the analysis of auditory-nerve firing patterns (de Cheveigne, 1989). Our use of it here for echoic memory pattern recognition is of course on a very different time scale, but the technique appears useful and extendable.

The problem with auto-correlations is that the resulting function is zero-phase. In order that our rhythmic analysis can preserve its placement in time, and can develop expectations about the future, we have developed a Phase-Preserving Narrowed Autocorrelation (PPNAC). The result is not only a sharpened account of past rhythmic activity, but a gradual formation of expected events (Fig. 13.6a). As the expectations move into current time, they are confirmed by the arrival of new peaks in the auditory analysis; if the acoustic source fails to inject new energy, the expectations will atrophy over the same short-term memory interval. Implementation of this as chain of cognitive processes devoted to temporal event patterns is not hard to imagine.

A musically interesting feature falls out as a simple property of PPNAC analysis: the additive combination of multiple terms will also tender reports at frequencies 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. It is important to realise that this listener need not be experienced, nor even paying much attention. But the grid for rhythmic hierarchy is already there.

Finally, the acoustic input is scanned for identifiable pitches by the unit "specptrk", and a scaled version of its output is displayed in Fig. 13.6b. As seen

in the program, the pitch tracker uses the same spectral data as the event and tempo detectors, and its response is reasonably accurate. On close examination, the treble B-flat is at first accurately estimated then quit in favour of the G below. In fact the performer had articulated this phrase with a very short B-flat; the G is an octave error of the lower sounding G (which the tracker does get when the performer also releases the next D).

All the elements of Fig. 13.5a, b, c and Fig.13.6a, b were generated and displayed in realtime by the program of Fig. 13.4, with acoustic input direct from a CD.

CONCLUSION

We have shown that acoustic input can be the stimulus for computational models of an entire range of processes involved in the perception and cognition of music, and the prospect of how to organise this is now receiving serious thought (Leman, 1994). Parsing an acoustic signal for hierarchical rhythmic content is a case of determining the relative auditory importance of the acoustic events. A computer representation of the auditory processes involved can lead to confirmation of theories; it can also lead to systems for automatic sensing and interactive performance. It is clear that the auditory system has evolved some elegant solutions to the various problems, and a big challenge in computer representation is to match that elegance. The example of spectral formants quickly inducing wholesale synchrony across large groups of auditory fibres (Secker-Walker & Searle, 1990) is a model for that cause.

The questions we would like to pose are of the form "how do pitch and rhythmic hierarchies arise, how do they operate, how do they achieve flexibility and robustness, and how much computation do they take?" The above model is computationally efficient (it runs in realtime), but does not readily accommodate things like tempo modulation. When the metrical stimulus changes pace, both the PPNAC and expectations become time-smeared. Although this permits adequate tracking of slow tempo changes, it does not handle the extreme changes in tempo rubato that we have successfully modelled elsewhere (Vercoe & Puckette, 1985). If focussed listening begins from the model we have outlined above, it must also be paying attention to short-time transforms, from which it can construct other representations that are strongly tempo-variable. This can be investigated using other facets of the Csound processing language, but that work still remains to be done.

The purpose of representing musical processes by computer should however be clear. The approach is one that permits theories of music cognitive processes to be posed, tested, and incrementally explored. The goal is to find how complex musical data is represented and processed by the human auditory-cognitive system. Only then will we understand why the music that exploits this capacity has the structure it does.

REFERENCES

Bilmes, J. (1993). *Timing is of the Essence: Perceptual and Computational Techniques for Representing, Learning, and Reproducing Expressive Timing in Percussive Rhythm.* (MS Thesis, Media Lab, Mass. Inst. Technology).

Brown, J., & Puckette, M. (1989). Calculation of a Narrowed Autocorrelation Function. *Journal of the Acoustical Society of America, 85*(4), 1595–1601.

de Cheveigne, A. (1989). Pitch and the narrowed autocoincidence histogram. *Proceedings of the 1st International Conference on Music Perception and Cognition* (pp. 67–70).

Delgutte, B. (1980). Representation of speech-like sound in the discharge patterns of auditory nerve fibers. *Journal of the Acoustical Society of America, 68*(3), 843–857.

Ellis, D. (1994). A computer implementation of psychoacoustic grouping rules. *Proceedings of the 12th International Conference on Pattern Recognition* (C108–C112). Jerusalem, Israel.

Hartman, W. (1989). Auditory grouping and the auditory periphery. *Proceedings of the 1st International Conference on Music Perception and Cognition* (pp. 299–304).

Hewitt, M., & Meddis, R. (1991). An evaluation of eight computer models of mammalian inner hair-cell function. *Journal of the Acoustical Society of America, 90*(2), 904–917.

Large, E., & Kolen, J. (1994). Resonance and the Perception of Musical Meter. *Connection Science, 6*, 177–208.

Leman, M. (1994). Signals, images, schemata, and mental representations. In I. Deliège, (Ed.), *Proceedings of the 3rd International Conference for Music Perception and Cognition* (pp. 203–204). Liège, Belgium: European Society for the Cognitive Sciences of Music.

McAdams, S. (1984). *Spectral fusion, spectral parsing, and the formation of auditory images.* (Ph.D. dissertation, Stanford University).

Povel, D., & Okkerman, H. (1981). Accents in equitone sequences. *Perception & Psychophysics, 30*, 565–572.

Richard, D.M. (1994). Name that tune: the Turing revisited. In I. Deliège, (Ed.), *Proceedings of the 3rd International Conference for Music Perception and Cognition* (pp. 175–176). Liège, Belgium: European Society for the Cognitive Sciences of Music.

Secker-Walker, H., & Searle, C. (1990). Time-domain analysis of auditory-nerve-fiber firing rates. *Journal of the Acoustical Society of America, 88*, 1427–1436.

Smith, R., Brachman, M., & Goodman, D. (1983). Adaptation in the auditory periphery. In Parkins & Anderson, (Eds.), *Cochlear prosthesis. Annals of New York Academy of Sciences, Vol 405*, 79–93.

Smith, R., & Zwislocki, J. (1975). Short-term adaptation and incremental responses of single auditory nerve fibers. *Biological Cybernetics, 17*, 169–182.

Todd, N. (1994). The auditory primal sketch. *Journal of New Music Research, 23*(1), 25–70.

Todd, N., & Lee, C. (1994). An Auditory-Motor Model of Beat Induction. *Proceedings of the International Computer Music Conference.*

Vercoe, B. (1984). The Synthetic Performer in the Context of Live Performers. *Proceedings of the International Computer Music Conference* (pp.199–200).

Vercoe, B. (1995). *Csound: A manual for the Audio Processing System and Supporting Programs with Tutorials.* Cambridge, Mass: Media Lab, MIT.

Vercoe, B., & Ellis, D. (1990). Realtime Csound: Software Synthesis with Sensing and Control. *Proceedings of the International Computer Music Conference* (pp.209–211).

Vercoe, B., & Puckette, M. (1985). Synthetic Rehearsal: Training the Synthetic Performer. *Proceedings of the International Computer Music Conference* (pp.275–278).

Zwicker, E., & Fastl, H. (1990). *Psychoacoustics.* Berlin: Springer-Verlag.

Zwislocki, J., & Sokolich, W. (1974). On loudness enhancement of a tone burst by a preceding tone burst. *Perception and Psychophysics, 16*, 87–90.