Center for Computer Research in Music and Acoustics May 1986

Department of Music
Report No. STAN-M-36

INTELLIGENT ANALYSIS OF COMPOSITE ACOUSTIC SIGNALS

by

John Chowning and Bernard Mont-Reynaud

This report contains the contents of a proposal to the National Science Foundation to continue work on the machine perception of complex sound signals. The report summarizes previous research in this area and outlines the proposed research to continue the project. The proposal has been funded with a start date of May 1, 1987 for three years.

This research was supported (in part) by the National Science Foundation under Contract NSF MCS 80-12476 and MCS 82-14350 and System Development Foundation under Grant SDF #345. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of Stanford University, any agency of the U. S. Government, or of sponsoring foundations.

TABLE OF CONTENTS

1. PROJECT SUMMARY

This research is concerned with computer analysis of complex acoustic signals, particularly with the development of a robust method to identify and track simultaneous acoustic sources in a monaural signal.

In the system, perception results from the interaction of data-driven and expectation-driven agents. Strategies for allocating resources to system agents and controlling feedback loops during the analysis of a time-varying signal are based on simulated *real-time problem-solving.*

The use of *multirate signal processing* in conjunction with focus-switching heuristics yields high resolution simultaneously in the time and frequency domains, an improvement upon traditional bandwidth-time tradeoffs.

Source coherence criteria derived from psychoacoustic observations (including correlated AM and FM modulation among partials) permit source separation when simpler methods fail.

The system's *learning co-processor* allows parameter adjustement and various forms of pattern recognition, using traditional numerical techniques as well as syntactic pattern matching, concept learning, and new hybrid methods combining parametric and structural views.

In summary, the research addresses key areas of acoustic analysis as well as important issues in AI, perception and learning. The implemented system is tested by using it as a front end for various acoustic recognition tasks, including (but not limited to) the transcription of polyphonic sound.

2. RESULTS FROM PRIOR NSF SUPPORT

The research discussed below was carried out under NSF Contract No. DCR-8214350, entitled "An Intelligent System for the Analysis of Digitized Acoustic Signals".

2.1. Summary of Completed Work

Combining signal processing techniques and semi-numerical methods with layers of domain knowledge, a computer program analyzes sound recordings of performed music and produces musical transcriptions. Many single-voice examples have been so transcribed. The results obtained with polyphonic data are quite encouraging.

The research contributes to several technical areas. In the area of signal processing, a family of algorithms for time-varying spectral analysis has been devised. The *Bounded-Q* transform allies the speed characteristic of the FFT with a non-linear frequency resolution which, like the ear's, is almost constant-Q.

The acoustic analysis subsystem mediates between signal processing and higher-level reasoning. Primarily concerned with event detection and identification, it includes novel or improved methods for segmentation and polyphonic pitch detection.

Semi-numerical algorithms for rational approximation, for hierarchical clustering, and for the recognition of subsequence patterns, have been developed within the higher-level analysis subsystem, which also includes methods specific to the musical domain.

The central theme of the work has been to develop strategies for reliable sound perception. The system acts primarily in data-driven regime, but expectation-driven feedback loops drastically improve its reliability. For example, dips in the auto-correlation of rhythmic sequences point to possible segmentation errors. This yields powerful error recovery based on acquired high-level context.

2.2. Publications

The publications listed were supported by NSF Contract No. DCR-8214350. Copies of these papers (but not of the Ph.D. thesis) are included in Appendix 2.

Chafe, C. and D. Jaffe, "Source Separation and Note Identification in Polyphonic Music." *Proc. ICASSP,* Tokyo, 1986.

Mont-Reynaud, B. "Problem-solving Strategies in a Music Transcription System." *Proc. IJCAI,* Los Angeles, 1985.

Chafe, C., D. Jaffe, K. Kashima, B. Mont-Reynaud and J. Smith, "Techniques for Note Identification in Polyphonic Music." *Proc. ICMC,* Vancouver, 1985.

Mont-Reynaud, B. and M. Goldstein, "On finding rhythmic patterns." *Proc. ICMC,* Vancouver, 1985.

Schloss, W. Andrew. *On the Automatic Transcription of Percussive Music — From Acoustic Signal to High-Level Analysis.* Ph.D. Thesis, Department of Hearing and Speech, Stanford University, Stanford California, May 1985. Department of Music Technical Report STAN-M-27.

2.3. Overview

Earlier work on this project (supported by NSF Contract MCS-8012476) was only aimed at the transcription of single-voice musical examples. For this purpose, it used a two-stage process. A sound recording was first reduced to an event list by a front end responsible for signal processing, pitch estimation and event detection [22,23]. Then the event list was submitted to musical analysis [10,37].

The current system consists of three main subsystems: signal processing, acoustic analysis and musical analysis. Each subsystem has an independent user interface and may be used in isolation, for investigations of limited scope. Interactive access facilitates debugging and experimentation and allows the user to step the system through examples that exceed the capabilities of fully automatic operation.

When the subsystems are teamed up for distributed problem-solving, acoustic analysis mediates between signal processing and higher-level processing. Reliability of the overall analysis has been enhanced by increasing the level of integration of the subsystems. This is particularly important for the analysis of polyphonic signals.

The system reliably transcribes recordings of simple melodies played on the piano or other percussive instruments. Examples that are more difficult due to the acoustic properties of the instrument, the complexity of the music or the expressiveness of the performance, are often transcribed with little or no assistance from the user.

The major limitations of the system arise in the treatment of polyphonic data. The high-level subsystem does not currently handle multiple voices. In the area of acoustic analysis of polyphonic sound, the results obtained so far have already improved the state of the art, but they also point to further research on the fundamental processes of source segregation in hearing.

The contributions of the project to various technical areas are described below.

2.4. Problem-solving framework

The approach combines mathematically-based methods with domain reasoning and significant use of context. The knowledge expressed in data representation choices, algorithms and heuristics covers selected aspects of instrumental acoustics, psychoacoustics, music theory, notation conventions, and performance practice.

The system uses a layered representation, in which frame-like objects are stored in a blackboard. Data descriptors span many layers of abstraction, from the surface signal to various approximations of its meaning. Useful descriptive elements include frequency-domain signal transforms, spectral peaks and spectral lines, scale tones, performed durations, rhythmic values of notes, accent markers, global meter and key markers, and patterns of varying nature and scope.

The predominant direction for information flow in the system is bottom-up. In addition, feedback links from the higher-level context to the lower levels of analysis allow powerful error recovery, as will be discussed in the sequel. The system generally avoids backtracking. This is done by carrying along small sets of hypotheses in parallel. The degree of ambiguity is controlled at each level of description with pruning techniques which take advantage of a multiplicity of evaluation criteria [37, 38].

2.5. Signal processing

The most critical signal processing step is the transformation of the sampled input signal to a frequency-domain representation. The Fast Fourier Transform is attractive for its efficiency, but the linear spacing of the frequency bins yields poor discrimination in the lower frequency range, compared with that of the ear. In

order to improve this situation while retaining the FFT's speed, we have devised a method of time-varying spectral analysis, called *Bounded-Q Frequency Transform* (or BQFT) which uses FFTs within a recursive scheme of octave decimation. The algorithm is described and illustrated in Appendix 2a. The BQFT offers better than semitone resolution, which is sufficient for the effective processing of polyphonic textures by the further levels of analysis.

2.6. Acoustic Analysis

- Segmentation.

A sudden rise in signal amplitude is the most obvious clue for onset detection. Reliable attack detection for the piano and for percussive instruments in general is obtained by thresholding instantaneous slopes obtained by linear regression from a suitably computed amplitude envelope. Applying the same method to a high-pass filtered signal is even better [50].

Other kinds of musical articulations, such as slurred bowed string attacks, motivate the use of frequency domain onset detectors. In addition to these data-driven event detection methods, messages from higher level routines may suggest candidate events for tentative identification.

- Identification.

Within spectrally stable segments after each detected event, the BQFT spectrum is searched for candidate sources. The presence of a source is suggested by a set of peaks in the instantaneous spectrum that can be grouped as partials of a given fundamental. Periodicity estimation (polyphonic pitch detection) uses a variant of Amuedo's algorithm [2]. Chords are separated into source hypotheses and the groups of partials are tracked in time.

Each source hypothesis is compared to a model describing the distinct features of the instrument's acoustics. For the piano these features include inharmonicity, exponential decay and amplitude modulation from beating among multiple strings. Hypotheses are weighed accordingly and "good" notes are added to the note list, labelled with timing, pitch and dynamic information. More details are given in appendices 2c and 2e.

2.7. High-Level Processing

Whenever possible, this research favors the development of general algorithms over the addition of specialized rules, expressing a bias in the *"power vs. knowledge"* controversy. As a result, many essential algorithms and strategies developed within the musical analysis subsystem are applicable outside the musical domain.

- Semi-numerical Algorithms.

Methods for hierarchical clustering and rational approximation have been developed or adapted to deal with timing fluctuations. These domain-independent algorithms assist in the conversion of performed durations to a metrical grid [37,38] but also have applications in the frequency domain.

- Detection of Structural Patterns.

Pattern detection methods explore auto-correlations in discrete sequences of features ([39] and Appendix 2d). One algorithm arranges repeated subsequences into a hierarchical data structure, and "subsumed" patterns are weeded out.

- Musical Analysis.

Earlier methods [10,37] for determining musical accents, tempo changes, rhythmic values and other aspects of musical context have been strengthened, notably by the use of clustering and pattern detection algorithms, in order to expand their power and applicability. For example, important aspects of metrical organization are revealed in observing regularities in the placement of instances of patterns.

- Feedback to Acoustic Analysis.

Some of the high-level processing creates context suitable for feedback to lower levels of analysis. In the system, this often takes the form of *peer pressure,* used as a strategy for error detection and recovery ([38], also Appendix 2b). A striking application of this idea uses pattern detection techniques, followed by statistics over pattern instances, to reveal spurious events and identification errors. A closely related technique uses *near-miss analysis* to suggest not only spurious events, but also missing events. It specifies the time at which an event was expected to be found (see [39] and Appendix 2d).

3. PROJECT DESCRIPTION

3.1. Objectives and significance

The proposal is part of a continuing effort towards computer understanding of complex acoustic signals. Earlier research discussed in Section 2 (and in Appendix 2) has led to the development of a framework of representations, techniques and strategies for sound perception, and to a better understanding of some of the key issues in acoustic analysis. The proposed research, while building upon the previous work, moves away from specific aspects of the musical domain. It focuses on fundamental processes of acoustic analysis that form the basis of domain-independent approach to the perception of composite sounds.

The central goal of the proposed work is to develop a system for the perception of monaural sound, capable of reliably separating simultaneous sound sources.

In the next three sections, we define the various aspects of the problem more precisely, outline the approach, and discuss suitable domains for the research. The significance of the work is discussed throughout but is also the object of a separate final section.

The problem

The goal of the proposed system is to perform an automatic acoustic analysis of signals resulting from simultaneous sound sources. Reliable performance in the tracking and identification of individual sound sources is the overriding consideration.

The research is not aimed at modeling the specific mechanisms of human hearing. Naturally, comparisons with human performance helps gaining a sense of direction and a measure of success. Also, psychoacoustical findings provide important insights. Research on human hearing has shown that listeners use a substantial number of distinct clues when hypothesizing sources, as well as complex conflict resolution strategies. Source tracking, which is one of the most fundamental principles at work, exploits both short-term and long-term continuity in the manifestations of each source. Short-term continuity is straighforward, once a proper metric is

chosen, but long-term continuity involves intelligent memory processes, with the ability to extract and recognize source behavior patterns.

The technical facets of acoustic analysis include segmentation (or event detection), source segregation, source identification (including pitch detection) and source tracking. These subproblems are not independent, however, from one another or from the larger context of signal interpretation. The somewhat tangled situation that results is typical of non-trivial perception tasks, and dictates important aspects of the problem-solving architecture required for the task.

As part of the initial assumptions underlying this work, we postulate the need for a layered data representation. We also believe that perceptual choices result from the convergence between observation (sensation, bottom-up processes) and expectation (prediction, top-down processes). Such a formulation leads to a wealth of technical issues, which we propose to address by experimenting with a program for "intelligent listening":

What causes sources to be formed against a background of noise and other sources? How are sources characterized? Can this process be modeled by the propagation of information in a descriptive hierarchy, and if so, how? What constitutes a proper balance of data-driven and expectation-driven reasoning, and how can it be achieved? How do adaptation and learning processes affect perception?

The approach

The proposed approach inherits its general character from prior research, which has been rather successful with single-source acoustic recognition problems, and, up to a point, with multi-source problems. The proposed approach, however, includes substantial new developments. It is described below using four key points.

- Computational Framework.

The primary global data structure (the "memory") is a two dimensional structure organized by time and level of abstraction. Information is collected and propagated by system agents, which are loosely coupled processes communicating via this memory. A key assumption shared with the earlier work is that perception occurs as the result of a careful balance of sensation and expectation. Accordingly, the implemented system relies on the interaction of two types of agents, data-driven

agents which propagate information from more detailed levels to more abstract levels, and context-driven agents which propagate information in the opposite direction. Feedback loops involving both types of agents are important for the robustness of the system. But since these loops require delicate regulatory mechanisms, their use in prior research had been limited to a small number of special cases where regulatory strategies could be devised. Loop control will now be addressed by viewing computation as unfolding in (simulated) real-time. This approach offers a uniform way to regulate feedback, but it also leads, as the speed of hardware increases, to true real time processing.

- Multirate Signal Processing.

Closely related to, but distinct from, abstraction/time tradeoffs are frequency/time resolution tradeoffs. The use of *multirate signal processing* together with suitable focus-switching strategies should allow one to obtain high resolution simultaneously in time and frequency. This type of process provides an efficient method by which the system can get around usual forms of the bandwidth-time tradeoff, and remain close to theoretical limits in the greatest part of the useful range.

- Expanded Psychacoustic Knowledge.

Recent work on fusion and streaming provides important insights into the processing of multi-source textures by the human ear [6, 14, 33, 34]. Criteria for the grouping of acoustic features suggested by these experiments aim at forming one or more coherent *auditory images*. In addition to familiar time-domain and frequency-domain clues, these criteria include more subtle effects such as correlated AM and/or FM modulation across partials. These clues may be expected to succeed in some cases where all else fails, but the computational cost and benefits remain to be investigated.

- Adaptation and Learning.

Adaptive techniques and unsupervised learning have important contributions to make to machine perception. Knowledge-intensive systems tend to focus on too narrow a domain. Training is known to be an attractive (if challenging) alternative to building specialized knowledge into the system. In some cases, a learning phase can be carried out "off-line" using a combination of trial-and-error, optimization algorithms, and other methods. But we are most concerned here with "on-line" learning

and adaptation. Our prior research offers examples of the use of both parameter adaptation and learning of stuctural descriptions, in order to yield context-gathering and error-recovery mechanisms. The intention is to generalize this type of strategy by postulating a "learning co-processor" as part of the memorization process.

Interpretation domains

For the purpose of experimenting with the tracking and segregation of simultaneous sources, polyphonic data remains quite attractive. Musical data *do* present eminently desirable characteristics for the effective pursuit of the research issues, including the investigation of learning techniques in the context of perception, but they may not be unique in this respect.

The "domain" for this work is that of composite acoustic signals. The choice of specific interpretation domains is important, yet subordinate to other research concerns. This contrasts with prior research, which was specifically focused on the music transcription task. Since the research has progressed towards more fundamental issues in sound perception, it seems useful to open the investigation to other acoustic domains and tasks, if only to reveal domain dependencies.

The basic intention is to provide a tool for acoustic analysis that is as general as possible, and can be shown to perform well *at least* in the context of polyphonic music. This is further discussed as part of the "Research Plans".

Significance

- Signal Processing and Acoustic Analysis.

In the signal processing area, the implications of simultaneous multirate signal processing could be quite deep. The idea is in principle applicable whenever the use of a fixed time-frequency resolution tradeoff is inadequate, and a finer control of bandwidth-time tradeoffs is desired. The implementation raises non-trivial issues, and it remains to be seen if the theoretical beauty of the scheme carries over to practical situations.

The most obvious contributions from this research are in key areas of acoustic analysis: acoustic segmentation, source identification (notably polyphonic pitch

detection) and, particularly, source segregation.

The problem of "hearing" multiple acoustic sources is a well-known challenge for machine perception research – one which is significant for both theoretical and practical reasons. The experience acquired in past work, combined with the new directions outlined in this proposal, offer the promise of a vigorous attack on the problem. The unique character of the proposed approach is reflected in the simultaneous use of multiple sampling rates, in the implementation of a psychoacoustically mature notion of acoustic source coherence, and in the AI concepts in use.

The applications of multiple-source acoustic analysis are numerous. As Amuedo [2] has justly observed, single-source problems in the presence of a significant amount of noise are best approached as multiple-source problems. [In practice, noise and untracked sources are effectively indistinguishable – one is reminded of the "cocktail-party effect"]. The possible application domains include music recognition, speech recognition (in multi-speaker *or* in noisy situations), underwater signal analysis, ecological studies from sound recordings, and many others.

- AI Architectures, Perception and Learning.

Other contributions concern machine perception and artificial intelligence in terms that are not specific to the acoustic domain. In particular, the systems aims at achieving an on-line synergy of perception and learning agents in an intelligent and highly adaptive analysis system, and the principles involved are readily generalizable.

Specific aspects of the proposed AI framework that may become significant in broader AI contexts include resource allocation and pruning strategies, and the use of "real-time" as the basis of a methodology for controlling expectation-based feedback paths.

Contributions may also expected in the area of machine learning, where researchers have recently turned to ***hybrid*** methods combining parametric and symbolic points of view [27, 20], and many recognize the importance of addressing learning issues in the realistic context of perceptual tasks [45, 46].

3.2. Research Plans

The proposal focuses on the development of an acoustic analysis front end suitable

for use in a partially unspecified intelligent system. The acoustic analysis system must be exercised and tested as part of larger systems devoted to various intelligent acoustic recognition tasks. The first subsection below discusses specific tasks and domains that seem suitable for this purpose. The next four subsections respectively expand the four major points on which the approach is based.

Experimental domains

As indicated earlier, a significant difference between the prior work and the proposed research is that the latter focuses on more fundamental problems of acoustic analysis in a largely domain-independent manner. Working with several acoustic analysis tasks appears to be a good way to properly factor domain dependencies. In any case, the choice of domain is subordinate to the pursuit of two key research issues, acoustic source formation and pattern acquisition.

While automatic transcription of polyphonic sound is no longer the primary goal for the research, the use of musical data continues to be eminently desirable. In no other domain is the simultaneity of acoustic sources exploited as systematically as in polyphonic music. Also, data easily available for the research present great opportunities for controlling independently the various dimensions of difficulty inherent in the problem. In particular, under the enormous variety of sound manifestations, musical passages usually present a number of levels of organization, some of which are good targets for feature abstraction followed by pattern detection, leading to unsupervised learning. Rather than being constructed according to a pre-specified grammar, a musical composition typically establishes some simple patterns and then uses them as building blocks, providing a good test for learning methods.

Other tasks that point to similar technical issues should be useful, in order to broaden the experimental basis of the work without diluting the effort. A task that presents good characteristics in this respect is the automatic recognition and classification of bird songs. Bird song patterns appear to be good targets for learning methods. As usual, the complexity of the problem depends on whether a vocabulary of features is given by experts. [Note that most birds emit a sound formed by a single sine wave, subject to (possibly rapid) modulation.]

The "multi-bird recognition" problem, which appears to meet all our research-

driven criteria, also turns out to have practical applications. An important indicator of the ecological evolution of specific habitats is the variety of bird species that inhabit it. Since it is expensive and inconvenient to send experts to listen directly to the birds at specific times and places, sound recordings are made locally and sent to the expert, who then identifies species and attempts to count individuals, or at least species variety. For some regions of the world, very few experts are capable of making the necessary judgements from these field recordings. Computer analysis of the recordings may be just what is needed in some cases.

While it is not clear how far to proceed with bird songs, within the context of the proposed investigation, it is obvious that taking at least some steps in that direction will help insure that the acoustic analysis system being constructed present a fair degree of domain independence.

Problem-solving framework

The data representation or *memory* is primarily organized in terms of a vertical dimension (the level of abstraction) and a horizontal dimension (time). The structure underlying the "vertical axis" is a partial order over the set of feature classes. This partial order, whose representation is a part of the self-description of the system, indicates what feature types may offer evidence for what other feature types.

The object instances are frame-like structures stored in a blackboard. They are organized in a network which uses evidence links as well as temporal succession links, pattern/instance links, etc.

Pure data-driven hypothesis formation stores values determined from lower-level slot values into higher-level slots and sets up the corresponding evidence links (H1 "provides direct evidence for" H2)

Ideally, a slot holds a single value. But it may have none or several, resulting in incomplete or ambiguous descriptions which are still very useful. Often the fact that a slot contains missing or ambiguous information ends up being irrelevant, because some other piece of evidence takes over along a different path.

Sets of alternative values are pruned on the basis of partial orders over multiple evaluation criteria. This technique helps maintain a level of ambiguity appropriate

to the context. It helps avoids backtracking by making it practical to carry several choices in parallel until the addition of further context allows a decision to be made. This is further discussed in [37, 38] and in Appendix 2b.

System agents are loosely-coupled processes communicating via the blackboard. The gathering of possible representations of the signal proceeds by the parallel activity of all enabled agents. The major flow of activity is bottom up and results in upward evidence relations. The use of predictive agents (that postulate lower-level features on the basis of higher-level context – these are rarely used) or context-sensitive explanatory agents (bottom-up agents that also use context parameters from higher up) causes the addition of downward (context) links in the network.

The interactions between agents may also be understood in terms of the two dimensions in the data representation. The bulk of the computation is carried out by straighforward bottom-up agents responsible for incremental steps in abstraction, i.e. in the *vertical* structure. In order to visualize the structure of the larger system, one begins with a data flow graph linking features in a strictly upward (and thus loop-free) manner. Then one may add *optional* context parameters that may affect the behavior of many bottom-up agents. Since context parameters are also computed in the data flow graph, we obtain feedback loops, which can be quite useful in achieving various forms of adaptation and focusing.

Naturally, feedback loops also present a potential for instability (over-reaction to change) or stubbornness (inability to adjust to change) and for an explosion in computational complexity. It seems difficult to explicitly limit the number of iterations of any given context feedback loop, because such loops only exist as a by-product of evidence and context propagation paths, and there is no centralized mechanism to keep track of all possible loops. In some cases it is possible to trust that convergence (a recognizable situation of negligible state change which stops the transitive propagation of consequences) will obtain quickly. This happens with context-gathering mechanisms that have a built-in "central tendency" – this is the key idea of the peer pressure strategy described in [38] (see Appendix 2b).

However, there is another approach to the control of feedback loops, which may be used in conjunction with peer pressure or separately. The idea is to assume that all processing happens in "real time," i.e. simultaneously with reading input samples. The notion of time used here is "simulated real-time" but it is pleasant to

know that if sufficient computing bandwidth is present in the processing network, the analysis model actually turns into a true real-time system.

The basic principle is that system agents take finite time to do their work, even though they work in parallel. Every analysis step is *causal* and it delivers at time t+D the results of observing the situation at time t. Since positive delays add up in every computational path, what appears at first to be a "vertical loop" (a recursive system of equations in the instantaneous system variables) actually has a horizontal drift. In other words, feedback loops cannot be "pure repetitions" operating with the same lower-level data over and over again.

Further, as long as we use simulation, the delay incurred in any one step need not reflect the computational complexity of this step. By choosing delays appropriately one effectively assigns a "time constant" to every feedback mechanism. Intuition (aided, when available, by psychophysical data) will guide the choice of processes and delays to be used.

The idea of simulating "real-time" ought to offer a promising approach to the control of attention and the stabilization of feedback loops between data-driving and context-driving. In the same spirit, we will provide each level of data description with only finite buffering, and introduce *forgetting* mechanisms that enforce the space constraint gracefully. Thus, the *memory* has a tendency to decay as time passes, unless features are remembered as part of patterns that remain active.

In the memory structure there is also a vertical gradient of quality as well as quantity. Features found higher up tend to describe larger time spans, and are also computed with greater delays. Feedback from high-level context may eventually affect the lowest levels of computation. Naturally, the higher the feedback context, the greater the delay. This is as it ought to be. When dealing with a given "chunk" (phoneme, word, note, sentence, musical phrase) the urgency of identifying the chunk typically rises for some time after the chunk is finished, and the acceptable delay is essentially proportional to the duration of the chunk.

In addition to this orderly bottom-up progression, context or redundancies create expectations which, for example, allow one to understand a sentence before it is finished, or to predict the next word to come. This phenomenon is essential to perception, although it is easy to go overboard with the use of predictive models.

While prediction will help recognize an input word if it matches one of the expected words, the ability of recognizing unexpected chunks must remain strong. Coupling mechanisms that foster a correct balance between expectation and sensation are actively sought in this research. Prior work that emphasized the *peer pressure* strategy will be further developed, notably by adapting these ideas to the use of "real-time" notions.

Multirate Signal Processing

The Bounded-Q Frequency Transform (BQFT) is an algorithm for time-varying spectral analysis which has proved useful for obtaining high-resolution spectral data. The algorithm is described in Appendix 2a. In brief, one computes successive identically-sized windowed FFTs of the original signal and of the derived signals obtained by repeatedly down-sampling by factors of 2. The surprising efficiency of the method, both in terms of compute time and in terms of the size of the data, comes from the fact that (1 + 1/2 + 1/4 + 1/8 + ...) does not exceed 2.

The BQFT provides a redundant spectral representation which yields a range of frequency/time tradeoffs simultaneously. The reader is invited to examine Figures 1 through 7 at this point (Appendix 2a). The display uses 4 by 4 pixels with 17 gray levels from 0 to 16. Although the displays are not fully calibrated and show some artifacts, one gets the intention clearly by comparing figures. Consider for example the first event in Figure 7. Its frequency is defined within a half of a semitone, which is the resolution of the visual display. Note that the timing of this event is quite fuzzy. But if one locates this event in Fig 3 or Fig 2, the timing becomes clearer, resolved down to a single 4 by 4 pixel representing a 20 msec interval.

It is important to realize that neither an FFT analysis nor a constant-Q analysis can deliver simultaneously such accuracy in both time and frequency. The way we "cheat" is by claiming that the first event visible in Fig 7 is the same as that in Fig 3. In the presence of sufficient noise or polyphonic complexity, such claims may become hazardous. But among the many strong partials of a tone, chances are some are relatively noise-free and can be used for accurate frequency measurement.

In the course of acoustic analysis, it frequently happens that subtle articulations are detectable only at one level of resolution and barely visible at others. This

suggests a concurrent approach in which event detection and identification rely on running detectors in parallel and heuristically merging their results.

Informal experiments to assess the power of the BQFT to resolve subtle phenomena have been encouraging. For example, in Fig 5, the first five events may be labeled with the (transposed) musical pitches B, C, D, C and B. Immediately after this appears a more confusing situation (near the center of the figure). On the way down to G, there is first a weak C, and before that, an even weaker and shorter almost D-sharp tone. By ear it was possible to tell that "something" was happening. Well-trained musicians actually recognized the passing C without much difficulty. But they did not hear the D-sharp until the digital recording was played at one quarter the speed. It is comforting to see that visual scanning of the BQFT data was correctly pointing to subtle acoustic features.

At the very least, in the lower octaves, the method surpasses the FFT in frequency resolution, and surpasses constant-Q in time resolution. (Of course, one can do repeated constant-Q analyses with successive factors of 2 in BT tradeoffs, but by then the method is almost equivalent to the BQFT and less efficient. Another remark: when the BQFT itself becomes fuzzy in the lower frequency range (tens of Hz) it is time to abandon the "place" theory of pitch and switch to time-domain pitch detection.

The BQFT eliminates the difficulty that Amuedo saw in his own periodicity estimation algorithm, which computes a *virtual pitch* by detecting clusters of subharmonics [2]. Because the algorithm relies on the original FFT only (without decimation), it has trouble with pitches below middle C, due to the lack of resolution in the lower part of the useful range.

A final remark on sound preprocessing: For purposes of this project, we may well be content with the BQFT family of algorithms for "efficient" spectral analysis. But it is possible that greater attention needs to be given to a more detailed understanding of the human ear. (See [1] for a recent review of cochlear models). Such concerns had until recently been excluded from serious consideration in our system, partly because of the substantial computational burden of realistic cochlear models. However, an accurate model of the human ear may soon become available in hardware implementation that makes its use practical [29]. We have begun some experiments comparing the information provided by our BQFT front-end with that

given by Richard Lyon's model. Unfortunately, at the time of this writing, the results are not yet available. In any case, for the time being, the plan is to rely on the BQFT for the initial audio transform.

Acoustic Source Formation

Source segregation is an ill-named problem. When attempting to "separate" sources, the actual emphasis is on "grouping" subsets of clues in order to form coherent *acoustic images.*

Intuitions backed up by psychoacoustic research suggest that both horizontal (time-axis) grouping and vertical (frequency-axis) grouping of acoustic elements play important roles. Another important element is the recognition of previously encountered sources. It thus appears that reliable identification (including the determination of pitch, loudness, timbral identity, etc.) stems from the interaction between grouping processes and memory processes. The more refined memory processes, which support adaptation, learning and prediction, are discussed in the next section.

The approach for source segregation involves an initial segmentation. Time-domain cues are used, but abrupt spectral change also causes event detection. This is followed, first by the tracing of spectral lines, and next by grouping processes that attempt to assign spectral lines to source labels according to the coherent behavior of sub-groups of lines (partials). Criteria for coherence are expressed in the grouping rules.

To put it another way, initial processing breaks up the data into time slices and frequency bins, resulting in a large number of small cells. The task is then to regroup these into one or more coherent *acoustic images,* together with some background noise. Instantaneous energy peaks found in the original frequency domain transform must be grouped both in time and in frequency to form an image. Steps towards forming *coherent* images include horizontal grouping resulting in spectral lines and vertical grouping resulting in harmonic sets.

The first and last of the grouping rule sets operate horizontally. The others are vertical, allowing the building of sources on the basis of coherent patterns in the amplitude and frequency behavior (as functions of time) of the spectral lines. One

must decide whether a source arises as a by-product of the emergence of a pattern of coherence, or if it claims lines (partials) on the basis of mutual coherence.

The hypothesis that a new source is present may be triggered by one of two criteria. The first is a transient in the global amplitude envelope. The second is activity in any frequency channel that appears incoherent (largely uncorrelated) with its recent activity or with activity in adjacent channels. Once a source is hypothesized, it begins to make claims on existing spectral lines, starting with ones appearing in the temporal vicinity of its initiation. It makes claims based on the simultaneous grouping criteria. The strength of the claim depends on the degree to which the line in question fits with the ensemble labeled by the source. Thus each line may have several weighted claims on it as a partial of the existing hypothesized sources [7].

In some cases, modeling can reduce the amount of computation by setting up expectations about reasonable continuation of the present state of the world. In so doing, constraints are formed on possible interpretations of the present acoustic situation.

The first series of rules address the horizontal grouping of peaks into spectral lines. The next set of rules serves to group spectral lines into sources (vertical linking of partials). Another set of horizontal grouping rules serves to group source events into streams exhibiting continuity of sources. These rules are derived from work in human auditory organization [6, 32, 33, 34]. The latter set is actually wide open to the influence of other principles, as will be discussed later.

- Partial Following (Horizontal Grouping).

This procedure has decision-making on both local and global scales of time and frequency. Locally, one assumes that there is a certain amount of inertia in the behavior of a component frequency: frequencies do not generally make sudden jumps or changes in direction. Finding a coherence discrepancy in the following of a spectral line may provoke a more fine-grained acoustic analysis and propose some reasonable possibilities, thus constraining the breadth of the search necessary to resolve the problem. For example, once a component is labeled as a partial of a given source, it is most likely that its variation in frequency and amplitude will be correlated with the behavior of the ensemble of partials believed to comprise that

source. This, in effect, constitutes a constraining feedback mechanism from higher to lower levels of decision making. This level of operation is more important once a view of the world is constructed and is being monitored for change.

- Amplitude modulation behavior [6, 14, 47]

Partials of same source grow, sustain and decay in a coherent fashion; this includes the principle of synchronicity which states that partials from the same source start at approximately the same time. A good rule of thumb for instruments might be that anything occurring within a group spread of 25 msec or less can be considered as synchronous.

- Frequency modulation behavior [33, 34].

Partials of the same source change in frequency in a coherent fashion (maintaining constant ratios among themselves) in sustaining forced-vibration systems once stability is established. One needs to verify the parallelism of frequency trajectories, in the original form and also after removing vibrato, if any.

- Resonance structure behavior [48, 26, 34].

Low frequency FM-AM coupling gives an indication of resonance behavior which varies in a coherent fashion for a given source. These structures generally vary more slowly than the frequencies themselves, making them accessible to analysis. Deducing aspects of the resonance structure provides a handle on the identity of the source.

- Frequency structure characterization (Vertical Grouping).

Musical sources are partly characterized by the frequency and amplitude of partials. Specific structures are associated with struck or plucked strings, and with tuned metal and wood objects. If harmonic, these structures may modulate. Severe modulation in the amplitudes of the components will be present with inharmonic sources.

- Spatial location.

Partials of the same source come from a location that is either fixed or slowly moving, compared to sound propagation. This is only included for completeness – we are planning to use a single input channel, omitting the development of a stereo localization model; other spatial clues are too unreliable for analysis purposes.

- Source Continuity [5, 8, 32]

The events emitted by a given source continue to behave in more or less the same way, i.e. the ensemble properties do not change too suddenly. This principle may be understood as the simple numeric continuity over short-term spans. This understanding may then expand to to a variety of pattern matching and learning rules (cf. next section) representing an intelligent memory-based predictive process. Finally, the latter formulation is clearly open to arbitrarily varied semantic constraints. Thus we have reached the zone at which a properly acoustic model gives way to learning and cognition.

Adaptation and Learning

It was argued earlier that "training" can be an attractive alternative to building massive knowledge into a system. When possible, one may want to use algorithms capable to "tune" quickly into the essential properties of a given situation, using a spectrum of adaptation techniques. From another front, psychoacoustic studies point to the importance of intelligent memory processes that range from following low-level parametric continuity to higher forms of predictability based on larger patterns.

In response to these felt needs, we postulate a "learning co-processor" which regroups a potentially large and varied collection of learning agents. The co-processor metaphor implies (at least partial) asynchrony with the problem-solving activity, but it does not exclude further parallelism among learning agents, each of which is loosely supervised by a critic (or "coach") which selects training goals and training data for the agent.

There are several classes of agents, and great disparities in their levels of complexity. Classical pattern-matching agents may account for the bulk of the processes that operate below the "event" level, including short-term temporal continuity and the recognition of qualities such as timbre. Other agents look for higher-level patterns in a level of representation where discrete events are already formed and at least partially labeled. Yet other agents may look both above and below the event level in order to correlate discrete qualities with more continuous, lower-level features.

Adaptive algorithms and learning methods suited for the various contexts include parametric techniques of optimization, clustering and numerical taxonomy; symbolic techniques related to grammatical inference, string matching, and concept formation; and hybrid techniques which, like conceptual clustering, combine parametric and structural views. [35, 20, 27].

It is reasonable to expect that hybrid techniques are important for perception, as well as they are bound to play an increasing role in learning research, as they represent a more realistic and generally useful model of learning situations. Recently, several authors have commented on the fact that a more genuine kind of learning problem results from operating in the context of a perception task. Partridge [45] says: "Consider the AI paradigm of rule-learning but in the empirical world rather than an abstracted concept characterized by drastically pruned descriptions. The act of describing removes most [...] attributes of each event before the learning algorithm sets to work – it is fed predigested reality... ". Phelps and Musgrove [46] express a similar opinion: "... but the most difficult part of the work has already been accomplished when the relevant features for rule formation have been found." They use 2D visual perception problem for studying realistic learning situations. Fisher and Langley, while presenting their approach to conceptual clustering, introduce it as a hybrid method of the kind just discussed, a cross between numerical taxonomy and concept learning [20].

There are now good reviews of the range of methods available for adaptation and learning (see, e.g., [35]). We plan to examine the role that such algorithms (and others to be developed) can play in perception, guided by the various needs of acoustic grouping processes leading to source formation. This may not be an easy task, but it appears that the level of maturation reached by learning and perception research should now permit a smooth junction. In fact, it would seem that the perception context in which this research operates, as well as the specific domains considered for experimentation, create an ideal environment for a thorough investigation of the synergy of perception, memory and learning.

In a situation so full of exciting ramifications, it is important to keep in mind the task goal, which is to reliably and accurately parse composite sounds into its component sources.

4. BIBLIOGRAPHY

[1] J. B. Allen, "Cochlear Modeling," *IEEE ASSP Magazine,* vol 2, no 1, 3-29, 1985.

[2] J. Amuedo, "Periodicity Estimation by Hypothesis-Directed Search," *Proc. ICASSP,* Tampa, FL, 1985.

[3] M. BenDaniel, "Automated Transcription of Music," B.Sc. Thesis, Department of Electrical Engineering and Computer Science, M.I.T., Cambridge, MA, 1983.

[4] T.O. Binford, "Survey of Model-based Image Analysis Systems," *Int. J. Robotics Research,* vol 1, no 1, 1982.

[5] A.S. Bregman, "Asking the 'What For' Question in Auditory Perception," in M. Kubovy and J. Pomerantz (eds.), *Perceptual Organization,* Erlbaum, Hillsdale, NJ, 1978.

[6] A.S. Bregman and S. Pinker, "Auditory Streaming and the Building of Timbre," *Can. J. Psych.,* vol 32, 19-31, 1978.

[7] A.S. Bregman and Y. Tougas, "Propagation of Constraints in Auditory Organizat unpublished manuscript, McGill University, Montreal, 1979.

[8] A.S. Bregman, "The Formation of Auditory Streams," in J. Requin (ed.), *Attention and Performance VII,* Erlbaum, Hillsdale, NJ, 1981.

[9] A.S. Bregman et al., "Spectral Integration Based on Common Amplitude Modulation," unpublished manuscript, McGill University, Montreal, 1983.

[10] C. Chafe, B. Mont-Reynaud and L. Rush, "Toward an Intelligent Editor of Digital Audio: Recognition of Musical Constructs," *Computer Music Journal,* vol 6, no 1, 1982.

[11] C. Chafe et al., "Techniques for Note Identification in Polyphonic Music," *Proc. Int. Conference on Computer Music, 1985,*

[12] C. Chafe and D. Jaffe, "Source Separation and Note Identification in Polyphonic Music," *Proc. ICASSP,* Tokyo, 1986

[13] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing,* Prentice-Hall Inc., Englewood Cliffs, NJ, 1983.

[14] G.L. Dannenbring and A.S. Bregman, "Stream Segregation and the Illusion of Overlap," *J. Exp. Psych./Human Perc. Perf.,* vol 2, 544-555, 1976.

[15] R. De Mori and M. Gilloux, "Inductive Learning of Phonetic Rules for Automatic Speech Recognition," *Proc. CSCSI,* London, Ontario, 1984.

[33] S. McAdams, "Spectral Fusion, Spectral Parsing and the Formation of Auditory Images," Ph.D. Thesis, Department of Speech and Hearing, Stanford University, Stanford CA, *Department of Music Technical Report STAN-M-22,* June, 1984.

[34] S. McAdams, "The Auditory Image: A Metaphor for Musical and Psychological Research on Auditory Organization," in R. Crozier and A. Chapman (eds.), *Cognitive Processes in the Perception of Art,* North-Holland, Amsterdam, 1984.

[35] R. Michalsky et al., *Machine Learning: An Artificial Intelligence Approach,* Tioga Press, Palo Alto, CA, 1983.

[36] M.L. Minsky, "A Framework For Representing Knowledge," in *The Psychology of Computer Vision,* McGraw-Hill, New York, 1975.

[37] B. Mont-Reynaud et al., "Intelligent Systems for the Analysis of Digitized Acoustic Signals, Final Report.," *Department of Music Technical Report STAN-M-15,* Stanford University, Stanford, CA, 1984.

[38] B. Mont-Reynaud, "Problem-Solving Strategies in a Music Transcription System," *Proc. IJCAI,* Los Angeles, 916-918, 1985.

[39] B. Mont-Reynaud and M. Goldstein, "On Finding Rhythmic Patterns," *Proc. Int. Computer Music Conference,* Vancouver BC, 1985.

[40] J.A. Moorer, "On the Segmentation and Analysis of Continuous Musical Sound," Ph.D. Thesis, Department of Computer Science, Stanford University, Stanford CA, *Department of Music Technical Report STAN-M-3,* 1975.

[41] J.A. Moorer, "On the Transcription of Musical Sounds by Computer," *Compute Music Journal,* 4:1 1977.

[42] J.A. Moorer, "Algorithm Design for Real-Time Audio Signal Processing," *Proc. ICASSP,* 12B.3.1-3.4, 1984

[43] H. Nii and E. Feigenbaum, "Rule-based Understanding of Signals," in D. Waterman and F.Hayes-Roth (eds.), *Pattern-directed Inference Systems,* Acader Press, New York, 1978.

[44] H. Nii et al., "Signal-to-Symbol Transformation: HASP/SIAM Case Study," *AI Magazine,* vol 3, no 2, 1982.

[45] D. Partridge, "Input-Expectation Discrepancy Reduction: A Ubiquitous Mechar *Proc. IJCAI,* Los Angeles, 267-273, 1985.

[46] R. Phelps and P. Musgrove, "A Prototypical Approach to Machine Learning," *Proc. IJCAI,* Los Angeles, 698-700, 1985.

[47] R. Rasch, "The Perception of Simultaneous Notes Such as in Polyphonic Music," *Acustica,* vol 40, 21-33, 1978.

[48] X. Rodet, "Time-domain Formant-Wave-Function Synthesis," in J.C. Simon (ed.), *Spoken Language Generation and Understanding,* Reidel, Dordrecht, 1980.

[49] R. Schafer and L. Rabiner, "A Digital Signal Processing Approach to Interpolatic *Proc. IEEE,* vol 61, 692-702, June, 1973.

[50] A. Schloss, "On the Automatic Transcription of Percussive Music," Ph.D. Thesis, Department of Speech and Hearing, Stanford University, Stanford CA, *Department of Music Technical Report STAN-M-27,* June 1985.

[51] G. W. Schroeder, "Models of Hearing,," *Proc. IEEE,* vol 63, 1332-1350, 1975.

[52] G. W. Schwede, "An Algorithm and Architechture for Constant-Q Spectrum Analysis.," *Proc. ICASSP,* vol 29.2, 1384-1387, Boston, April 1983.

[53] R.N. Shepard, "Psychological Relations and Psychophysical Scales: On the Status of 'Direct' Psycho-physical Measurement," *J. Mathematical Psychology,* vol 24, 21-57, 1981.

[54] R.N. Shepard, "Structural Representations of Musical Pitch," in D. Deutsch (ed.), *Psychology of Music,* Academic Press, New York, 1982.

[55] H. Simon and R. Sumner, "Pattern in Music," in B. Kleinmutz (ed.) *Formal Representations of Human Judgement,* John Wiley and Sons, New York, 1968.

[56] J. O. Smith and B. Friedlander, "High-Resolution Spectral Estimation Programs Technical Memo 5466-05, Systems Control Technology, April 1984.

[57] J. Stautner, "Analysis and Synthesis of Music Using the Auditory Transform," MS Thesis, Department of Electrical Engineering and Computer Science, M.I.T. Cambridge, Mass., May 1983.

[58] L. Steels, "Constraints as Consultants," *Proc. European Conf. on AI,* University of Kaiserlauten, Kaiserlauten WG, 1982.

[59] G. Sussman and G. Steele, "Constraints: A Language for Expressing Almost-Hierarchical Descriptions," *Artificial Intelligence,* vol 14, 1-39, 1981.

[60] W. Ulrich, "The Analysis and Synthesis of Jazz by Computer," *Proc. IJCAI,* Pittsburgh, 1977.

[61] W. Von Bekesy, *Experiments in Hearing,* McGraw-Hill, New York, 1960.

[62] W.E. Ward, "Musical Perception," in J.V. Tobias and E.D. Hubert (eds.), *Foundations of Modern Auditory Theory* , vol 1, Academic Press, New York, 1970.

[63] P. Winston, "Learning Structural Descriptions from Examples," in *The Psychol of Computer Vision,* McGraw-Hill, New York, 1975.

[64] P. Winston, "Learning and Reasoning by Analogy," *CACM,* vol 23, no 12, 1980.

[65] J. Wolcin, "Maximum A Priori Estimation of Narrow-Band Signal Parameters," *JASA,* vol 80, no 1, 174–178, 1980.

[66] V.W. Zue, "The Use of Speech Knowledge in Automatic Speech Recognition," *Proc. IEEE,* vol 73, no 11, 1985.