*To be presented at the NATO Computational Hearing Advanced Study Institute, Il Ciocco, Italy, July 1-12, 1998*

# TOWARD AUTOMATIC SOUND SOURCE RECOGNITION: IDENTIFYING MUSICAL INSTRUMENTS

*Keith D. Martin*

kdm@media.mit.edu
MIT Media Lab Machine Listening Group
Room E15-401
20 Ames Street
Cambridge, MA 02139

## ABSTRACT

One of the broad goals of research in computational auditory scene analysis (CASA) is to create computer systems that can learn to recognize sound sources in a complex auditory environment. In this paper, a set of acoustic features is proposed that relate to the physical properties of sound-producing objects. In particular, a set of orchestral musical instrument sounds is presented as representative of the class of sounds produced by quasi-periodic excitation of resonant structures, acoustic properties of this class are considered, and the log-lag correlogram is presented as a signal representation that codes many of the proposed features. Specific examples are given of features extracted from violin, trumpet, and flute tones. Extensions to Ellis's prediction-driven CASA framework are proposed in the form of a hierarchy of sound-source models represented by frames. It is suggested that the goal of building an artificial system for sound source recognition in complex mixtures may be well served by such an approach.

## 1. INTRODUCTION

Recognizing objects in the environment from the sounds they produce is arguably the primary function of the auditory system. An organism that can sense a threat at a distance has a competitive advantage (in the evolutionary sense) over one that cannot. Recognition is possible, in part, because acoustic features of sounds often betray physical properties of their sources. As a simple example, large objects tend to produce sound energy at frequencies lower than those produced by small objects. If an organism's goal is to recognize sounds as arising from particular source classes, recognition should be based—if possible—on those acoustic features that are invariant across the sounds within each class yet distinguish between the sounds of different classes. For many classes of sound sources, acoustic characteristics that correlate with physical or behavioral properties are examples of such highly discriminatory features.

One of the broad goals of computational auditory scene analysis research is to create computer systems that can learn to recognize the sound sources in a complex auditory environment. In this paper, the class of sounds generated by quasi-periodic excitation of resonance structures will be considered. This class includes many animal vocalizations, but the discussion will be limited to sounds produced by a set of orchestral musical instruments, including members of the string, brass, and woodwind families. Although recognizing musical instruments is clearly not a task of evolutionary significance, humans can become skilled at identifying the types of musical instruments (e.g., clarinet, violin, etc.) independent of a particular performer and, to a large degree, of the acoustic environment. To date, no artificial system has been built that can demonstrate the same competence, but enough is known about the acoustic features that allow listeners to distinguish among the instrument classes that we might hope to be able to build a system that can do so.

The remainder of this paper comprises three sections. Section 2 describes relevant research in computational auditory scene analysis. In Section 3, a set of acoustic features is proposed—related to physical properties of sound-producing objects—that can be extracted from a simple auditory model. In Section 4, a method is described for constructing an artificial system—using the principles of auditory scene analysis—that employs these features to recognize musical instruments. In particular, extensions to Ellis's prediction-driven computational auditory scene analysis framework [1] are proposed that will enable hierarchical sound source classification and automatic acquisition of new models as novel sound sources are encountered.

The difficult problems associated with learning new features for discrimination will not be considered. Instead, a set of previously learned or hard-wired features—one sufficient to distinguish among the sound classes of interest—will be presupposed. This is not to say that every feature is relevant for every sound; each particular feature may be relevant to the recognition of only a subset of the sound source classes of interest.

The system described here is still in the early stages of implementation. One of the goals of this paper is to solicit feedback from members of the computational hearing community about the proposed approach.

## 2. BACKGROUND: CASA

*Auditory scene analysis* is the process of explaining sound energy arriving at the ears in terms of coherent acoustic sources. Bregman describes it as a complex interaction of grouping heuristics and learned schemata [2]. His grouping heuristics organize sound energy by harmonicity, common onset, common modulation, and common spatial location. Sequential integration is mediated by principles such as similarity of pitch, loudness, and timbre. These heuristics may operate in either a bottom-up (data-driven) or top-down (schema- or prediction-driven) fashion, depending both on the particular heuristic and the context. Several attempts have been made to build computational auditory scene analysis (CASA) systems based on these principles. Early efforts were limited by inadequate cues (including limited implementations), inextensible algorithms, rigid evidence integration, and inability to handle obscured data (as discussed in [1]).

Two recent CASA approaches are sufficiently novel to merit special mention here. Ellis attempted to address the limitations of previous systems by building a system that maintains a world-model consisting of low-level sound objects (noise clouds, transients, and quasi-periodic tonal elements). His system uses short-term prediction to infer masked or obscured information and is remarkably successful at grouping low-level time-frequency energy into perceptually salient objects—for example, car horns and slamming doors in a complex, noisy street scene.

Although it was not explicitly modeled after human auditory scene analysis, the IPUS Sound Understanding Testbed (SUT) [3] is unique among existing CASA systems in its use of explicit sound source models. SUT contains 40 sound source models divided into five categories (chirp, harmonic, impulsive, repetitive, and "transients"). Like Ellis's prediction-driven system (which also used the IPUS blackboard framework as its architectural structure), SUT integrates top-down (in the form of model-based inference) and bottom-up (data-driven) processing. Although it is likely that model-based inference plays an important part in auditory scene analysis, the SUT implementation has a number of severe limitations. For example, its models are based on single instances of particular sounds—used both for training and testing of the system—rather than generalizations from a set of training examples. Also, SUT does not include any mechanisms for dealing with a larger set of models, so combinatorial explosion in search will limit the degree to which the system can be expanded.

CASA systems are slowly expanding in complexity, but nothing approaching robust real-world performance has been demonstrated in any listening tasks to date. One might speculate that a more sophisticated world-model is needed; to that end, a combination of short-term prediction with a hierarchy of source models is proposed, within a framework that supports an interplay of top-down and bottom-up processing. Better source models—and flexible methods of reasoning about them—may be the key to building computer systems that can perform sound source identification in natural environments. This type of approach is a candidate for difficult tasks such as identifying musical instruments within a large ensemble performance, where nearly all sound sources are partially masked by others.

## 3. FEATURES FOR RECOGNITION

As stated in the introduction, acoustic characteristics that correlate with physical or behavioral properties of sound sources are examples of highly discriminatory features for source recognition. The literature on musical instrument acoustics (e.g., [4]) suggests a set of features that serve as a useful starting point. Features that appear to be important for musical instrument recognition include (but are not limited to): resonance characteristics (e.g., the frequencies and bandwidths of formants), amplitude envelope (attack, decay, and tremolo characteristics), inharmonicity, spectral centroid (which is known to correlate with perceived "brightness" [8]), onset asynchrony (the relative attack times of low- and high-frequency partials), pitch, and frequency modulation (e.g., vibrato, jitter). In sounds produced by natural sources, these features will strongly covary; for example, a source with a narrow resonance (indicating loose coupling between excitation and resonant body) will exhibit a slower attack than one with a broad resonance.

In an artificial recognition system, it is desirable that the signal representation capture as many of these features as clearly as possible and that it contain a degree of information similar to that contained in the auditory system. For the features identified above, the *log-lag correlogram* appears to be a good choice of signal representation.

The correlogram representation adopted here is based on the one underlying Ellis's prediction-driven CASA system [1]. Processing occurs in three stages. In the first, the raw acoustic signal is passed through a gammatone filterbank [5], which models the frequency resolution of the cochlea and retains a great deal of information in the output time signals of each channel. The filter outputs are half-wave rectified and lightly smoothed as a rough model of inner hair cell transduction. In high-frequency channels, these operations remove fine timing structure while preserving the envelopes of the signals. This process does not model adaptation or dynamic range compression, but does retain the desirable feature of coding signal intensity transparently.

In the third stage, the output of each channel is subjected to short-time autocorrelation, implemented by a simple delay/multiply/smooth architecture with a smoothing constant of approximately 20 ms. Autocorrelation output is computed as a function of time for lags spaced evenly on a logarithmic scale. In addition, the zero-lag autocorrelation gives a measure of the short-time energy in each channel.

These computations are expensive but easily adapted to parallel processing architectures. By computing each lag separately rather than using a window-based convolution method, it is possible to compute outputs for arbitrarily large or small lags—independent of the averaging-window length.

The correlogram representation is three-dimensional. The first dimension (cochlear position) yields critical-band frequency resolution, which is capable of resolving the first five or six harmonics of a periodic signal. The second dimension (autocorrelation lag) is a logarithmic representation of periodicity, corresponding to the nearly logarithmic pitch resolution exhibited by humans. The third dimension is time. The main panels of Figures 1–3 display snapshots of the correlogram output for violin, trumpet, and flute tones respectively.

Many of the features claimed here to be useful for recognizing quasi-periodic sounds are captured vividly in the correlogram representation:

**Pitch** – Pitched sounds will exhibit vertical organization within the correlogram: energy at a fixed lag (corresponding to the pitch period) will be present over a range of cochlear positions (frequencies). By integrating over the cochlear dimension—forming the so-called *summary autocorrelation*—likely pitch candidates can be identified by finding local peaks. Examples of the summary autocorrelation are shown in the bottom panels of Figures 1–3. Comparisons between the pitch of a musical tone and the normal playing ranges of known musical instruments may be useful for identifying potential instrument models—and ruling out others—during the recognition process.

**Frequency modulation** – It has been suggested that the presence and character of frequency modulations, including *jitter* (random modulations) and *vibrato* (periodic modulations), are characteristic of some sound sources [6]. For example, Figure 4 displays pitch tracks of tones produced by the violin, trumpet, and flute. The violin and flute tones exhibit pronounced vibrato. Brass instruments often exhibit a characteristic pitch modulation at the onset of a tone, due to loose coupling between the vibration of the player's lips and the instrument [7] (this can be seen in Figure 4 up to approximately 500 ms). It may be useful to identify a set of sub-features related to frequency modulation, including the presence/absence/degree of periodic modulation, "scoop" at onset, and random variations.

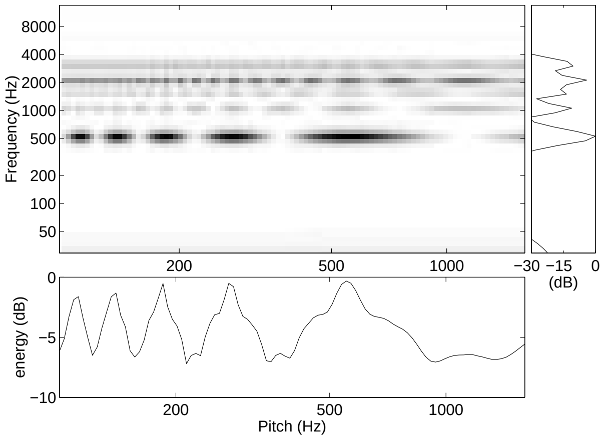

**Figure 1.** Correlogram snapshot of a violin tone. The horizontal axis, labeled "pitch," is the inverse of autocorrelation lag. The vertical axis, labeled "frequency," corresponds to cochlear position. The lower panel displays the summary autocorrelation (the correlogram integrated over the cochlear dimension). The right-hand panel displays the zero-lag energy, which for isolated periodic sources is equal to the spectral envelope.

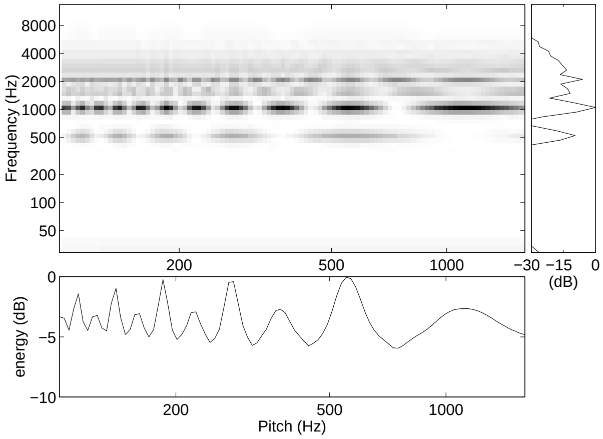

**Figure 2.** Correlogram snapshot of a trumpet tone. See the caption to Figure 1 for a description of the panels.

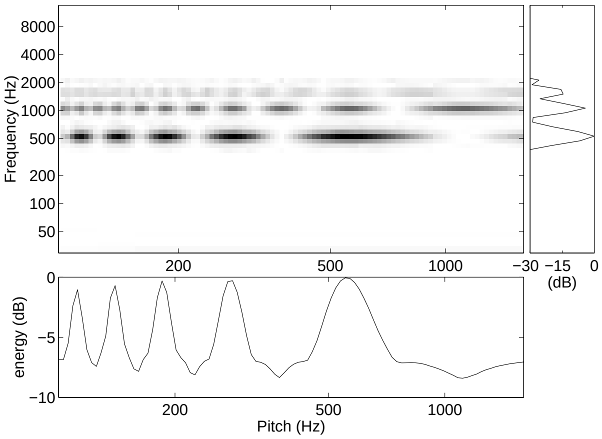

**Figure 3.** Correlogram snapshot of a flute tone. See the caption to Figure 1 for a description of the panels.

**Spectral envelope** – Following Ellis's *weft* representation [1], after the pitch-period of a sound is estimated, the correlogram can be examined to measure periodic energy at the corresponding lag as a function of cochlear position. This corresponds to a spectral envelope calculated with critical-band resolution, which may be used in conjunction with the pitch-track to recover the resonance structure of the sound source (e.g., see [6]). Approximations to the spectral envelopes of sample violin, trumpet, and flute tones are shown in the right-hand panels of Figures 1–3 respectively (for an isolated quasi-periodic signal, the zero-lag energy in each frequency band is approximately equal to the energy recovered by the technique described above).

**Spectral centroid** – After the spectral envelope has been estimated, it is a simple matter to calculate its centroid. Research has demonstrated that the spectral centroid correlates strongly with the subjective qualities of "brightness" or "sharpness" (e.g., [8]). The variation of spectral centroid over time for the violin, trumpet, and flute tones are shown in Figure 5.

**Intensity** – The autocorrelation output at zero lag corresponds to a running estimate of the intensity in each cochlear channel. The sum of these intensities is a simple correlate of perceived loudness. Beauchamp [9] has suggested that the ratio of spectral centroid to intensity is an important characteristic of musical instrument sounds; as tones grow louder, they become "brighter" (i.e., the spectral centroid shifts to a higher frequency) in a relationship that might aid instrument recognition.

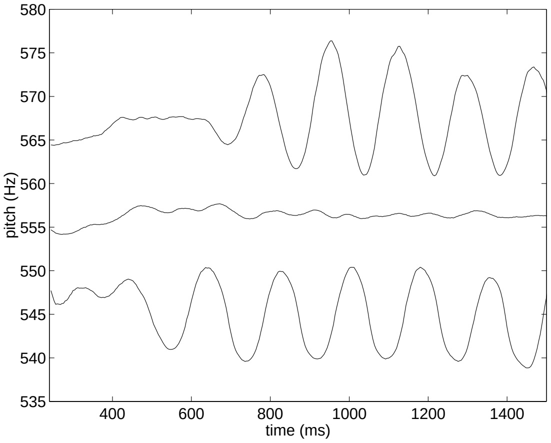

**Figure 4.** Frequency modulation in tones produced by [bottom to top] violin, trumpet, and flute. (The three tones were performed at the same pitch; for display purposes, the violin's pitch-track has been offset by –5 Hz—the flute's by +5 Hz.) The violin and flute tones exhibit periodic frequency modulations consistent with musical *vibrato*. The trumpet tone exhibits random frequency modulations consistent with *jitter*.

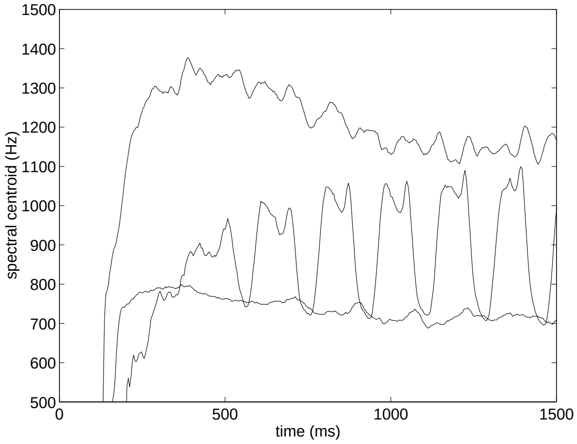

**Figure 5.** Spectral centroid for tones produced by [bottom to top] flute, violin, and trumpet. The trumpet tone is "brighter" than the violin and flute tones. Also note the large degree of variation of the violin tone's spectral centroid during vibrato as compared to the flute's.

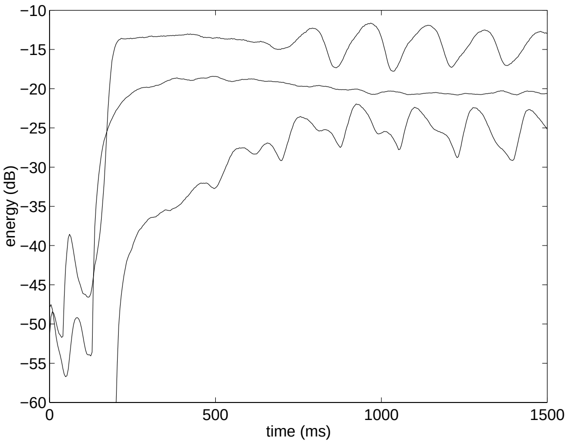

**Figure 6.** Amplitude envelope for tones produced by [bottom to top] violin, trumpet, and flute. (For display purposes, the violin pitch has been offset by –5 dB—the flute by +5 dB.) Both the violin and flute tones have clear amplitude modulation. The violin takes much longer (nearly 500 ms!) to reach its steady state energy level than the other two instruments. The flute's onset is nearly instantaneous.

**Amplitude envelope** – The amplitude envelope is simply the intensity measured as a function of time. It carries information about the source excitation and its coupling to the resonant body. For example, impulsive sounds such as plucked or struck strings decay exponentially with a rate that varies inversely with the tightness of coupling between the vibrating material (the string) and the resonant body [7]. Some instruments have faster rise times than others, indicating tight coupling between source and resonant structure. In particular, bowed string instruments have very slow attacks, as can be seen in Figure 6, which shows the amplitude envelopes for the onsets of tones performed on violin, trumpet, and flute.

**Amplitude modulation** – As with frequency modulation, small variations of the amplitude envelope of a sound can be important characteristics of natural sound sources. Articulated brass instrument tones, for example, often have low-amplitude, inharmonic "blips" at onset [4]. For some instruments, such as the flute, large periodic amplitude modulations (tremolo) are found in conjunction with vibrato, and may be an important identifying characteristic. It may be useful to identify a set of sub-features related to amplitude modulation, including the presence/absence/degree of periodic and random variations, as well as the rise time and rate during onset (as mentioned in the description of the amplitude envelope).

**Onset asynchrony** – By observing the spectral envelope over time, it is possible to track the rise of periodic energy in the various cochlear channels. In some instruments, high-frequency components rise more slowly than low-frequency components [4], and the ratio of the rise times may be a useful feature for recognition.

**Inharmonicity** – Because of mechanical stiffness, freely vibrating strings produce inharmonic partials. In such string tones, the upper partials exhibit frequencies higher than integer multiples of the first partial [7]. Inharmonicity may be observed in the correlogram as deviations from strict vertical organization in the vicinity of the estimated pitch period.

This list of features is not exhaustive, but it is representative of the physical characteristics that may be used to distinguish among quasi-periodic sound sources during auditory scene analysis. This point has not been stressed in the presentation so far, but it is important to note that in everyday listening situations instruments are not usually recognized by isolated tones. Indeed, a short piece of a musical phrase leads to far better recognition than isolated tones [10]. The timbre literature has unfortunately concentrated on the characteristics of isolated tones, placing undue emphasis on note onsets; in natural sounds, small variations of the "steady state" convey more robust information for identification. In light of this observation, it may be necessary to extend the proposed feature set to include more characteristics of note transitions and of the steady state.

## 4. PROPOSED MODEL

To build an artificial recognition system, I propose to extend Ellis's prediction-driven CASA approach with a hierarchical taxonomy of sound source models that supports inheritance of feature properties. Where Ellis's system attempts to explain the acoustic energy of an auditory scene in terms of noise clouds, transients, and quasi-periodic tonal elements, the proposed system will seek to explain only the quasi-periodic energy. The system will calculate the summary autocorrelation over time and will try to identify sources that account for the peaks by explaining features in the correlogram like those proposed in Section 3.

There are many problems to address in building such a system; one of the foremost is of *indexing*. The system must be able to choose appropriate explanatory models from a potentially large library of source models. In the case of multiple simultaneous sources, where occlusion and masking is very likely to occur, it is important to be able to choose such models based on only limited information. One reasonable approach to solving this problem is to perform hierarchical classification [11].

Musical-instrument sounds form a natural hierarchy based on their acoustic properties—a hierarchy that largely corresponds with the traditional instrument-family breakdown. At the highest level, instrument tones are classified as either transient (percussive) or sustained. Sustained sounds are further classified as blown or bowed, and the blown tones may be further divided into brass and woodwind classes. This hierarchy can be extended to the level of individual instrument type (e.g., clarinet or violin) and perhaps even further, to the level of individual performers.

Each of the categories mentioned above has characteristic acoustic properties. For example, transient sounds have rapid onsets and decay exponentially. Within the class of sustained sounds, bowed strings have very long onsets (the harmonic partials take a long time—often more than 250 ms—to reach "steady state"). Within the class of wind instruments, brass instruments tend to have simple formant structures, as well as amplitude "blips" and characteristic pitch modulations at onset.

The instrument models at each node of the taxonomy must be coded in some form of representation. A reasonable approach is to represent models with *frames* [12] whose slots correspond to features like those described in Section 3. A slot may have a default value or a restriction on the acceptable range of values, and these may be inherited from abstract class prototypes (parents in the hierarchy). For example, the trumpet model might inherit a slot describing the likely presence of amplitude blips at onset from the brass family model. There may be multiple frames (a *frameset)* that describe a single instrument class, perhaps representing different pitch ranges (registers) or playing styles. For example, the high register of the clarinet exhibits more pronounced even harmonics than the low register. When a particular instrument model is hypothesized to account for a quasi-periodic sound in an auditory scene, a new instance of that instrument's frame is created, and its slot values are adjusted to match measurements made from the signal. Multiple hypotheses may be simultaneously entertained until enough evidence is accumulated to favor a single interpretation of the sound energy.

There are two particularly interesting learning problems within this framework: learning the taxonomy itself from examples, and placing a new model within the hierarchy when a novel source is encountered. With traditional pattern classification and artificial neural network techniques, the first problem requires that the entire database of sounds is available during training, and the second problem has no supported solutions. In a realistic auditory environment, the principles of auditory scene analysis may be used to group chains of notes together that are likely to have arisen from the same source (e.g., a phrase of a melodic line). If none of the known source models match the assembled data, a new, unnamed model can be placed within the taxonomy by adding a new node as a child of the most appropriate abstract parent class.

As stated in the introduction, the proposed system is still in the early stages of construction. Although I have high expectations that the described approach will be fruitful, there is, as yet, no objective proof of the claims made here.

## 5. REFERENCES

[1] D. P. W. Ellis, *Prediction-driven computational auditory scene analysis,* Ph.D. thesis, MIT, 1996.

[2] A. S. Bregman, *Auditory Scene Analysis,* Cambridge: MIT Press, 1990.

[3] F. I. Klassner, *Data reprocessing in signal understanding systems,* Ph.D. thesis, University of Massachusetts at Amherst, 1993.

[4] D. Luce, *Physical correlates of nonpercussive musical instrument tones,* Ph.D. thesis, MIT, 1963.

[5] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank," Apple Computer Technical Report #35, 1993.

[6] S. McAdams, *Spectral fusion, spectral parsing and the formation of auditory images,* Ph.D. thesis, Stanford University, 1984.

[7] A. H. Benade, *Fundamentals of Musical Acoustics,* Second edition, New York: Dover, 1990.

[8] S. Handel, "Timbre perception and auditory object identification," In Moore, B.C.J., editor, *Hearing.* New York: Academic, 1995.

[9] J. W. Beauchamp, "Synthesis by spectral amplitude and 'brightness' matching of analyzed musical instrument tones," *J. Audio Eng. Soc*., 30(6): 396–406, 1982.

[10] R. A. Kendall, "The role of acoustic signal partitions in listener categorization of musical phrases," *Music Perception*, 4(2): 185–214, 1986.

[11] S. Ullman, *High-level Vision*, Cambridge: MIT Press, 1996.

[12] M. Minsky, "A framework for representing knowledge," MIT A.I. Lab Memo #306, 1974.