Toward Automatic Sound Source Recognition: Identifying Musical Instruments

Keith D. Martin (with Youngmoo E. Kim) - MIT Media Lab Machine Listening Group - kdm@media.mit.edu

Goal: sound source recognition

The goal of this research is to build computer systems that can learn to recognize sound sources in complex auditory environments. This is one of the broad goals of research in computational auditory scene analysis (CASA).

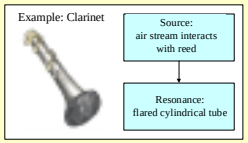
Recognition is possible because the acoustic features of a sound betray physical properties of its source. The most perceptually-salient acoustic features relate to the physical properties of source excitation and resonance.

Previous CASA systems have had only limited *world models*. Better source models may be the key to building computer systems that can perform sound source identification in natural environments. The system described here extends Ellis's prediction-driven CASA framework with a taxonomy (inheritance hierarchy) of sound source models.

How are musical instruments recognized?

Many of the acoustic features that are important to humans for sound source identification are related to the excitation and resonance properties of the source. For musical instrument sounds, these features include:

- pitch (vibrato/jitter/range)
- spectral envelope (formants)
- amplitude envelope (tremolo/onset/decay)
- onset asynchrony (excitation, resonances)
- inharmonicity

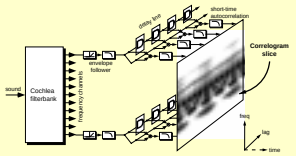


Mid-level representation: the log-lag correlogram

The log-lag correlogram is a three-dimensional representation that captures a great deal of information about quasi-periodic sounds (including musical instrument tones).

The three dimensions are:

1. Cochlear position - with critical-band frequency resolution
2. Autocorrelation lag - a logarithmic representation of periodicity (the inverse of pitch)
3. Time

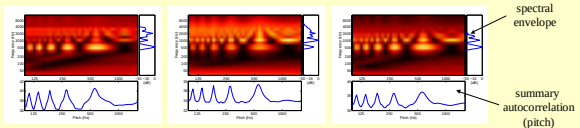


Here, the correlogram has been implemented with a delay/multiply/smooth architecture, after Ellis (1996). The correlogram is a representation; it is *not a hearing model.* It may or may not correspond to the neural representation in humans.

Feature support

The log-lag correlogram vividly encodes many of the acoustic features that are important for musical sound source identification, as can be seen in the following comparison of violin, trumpet, and flute tones.

Correlogram snapshots

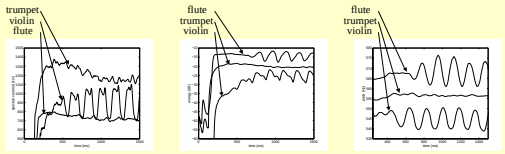


violin trumpet flute

After identifying the time-varying pitch of a quasi-periodic signal, it is possible to measure the energy in the pitched signal—as a function of cochlear position, or frequency—yielding an estimate of the source's (possibly time-varying) spectral envelope.

It is then possible to estimate aspects of the sound source's resonant properties, including its formant structure and spectral centroid ("brightness"). These features may vary over time in a manner that is characteristic of the source excitation and/or resonance.

Time-varying features for the three example tones



spectral centroid amplitude envelope pitch

Isolated tone data for pilot study

In an effort to quantify the utility of various acoustic features for recognition, a set of 1023 instrument tones was recorded from a sampler CD. These tones cover the full playing ranges (and a small range of articulation styles) for 14 orchestral instruments, including members of the brass, woodwind, and string families.

For each tone, the log-lag correlogram was computed. The pitch and spectral envelope were extracted as functions of time, and a number of auxiliary features (e.g, spectral centroid, vibrato frequency and strength) were computed.

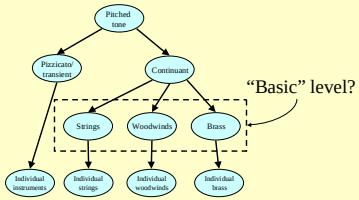
	Instruments		Features
strings	violin (bowed, muted, pizzicato)		average pitch
	viola (bowed, muted, pizzicato)		average spectral centroid
	cello (bowed, muted, pizzicato)		avg. normalized spect. cent.
	double bass (bowed, muted, pizz.)		maximum onset slope
brass	trumpet (C, Bach, C with Harmon mute)		onset duration
	horn (normal, muted)		vibrato frequency
	tenor trombone (normal, muted)		vibrato strength
	tuba		tremolo frequency
woodwind	flute	air reed	tremolo strength
	piccolo		slope of onset harmonic skew
	oboe	double reed	variance of ons. harm. skew
	english horn		amplitude decay
	bassoon		odd/even harmonic ratio
	Bb clarinet	single reed	

Pilot study: hierarchical classification

Psychological studies indicate that objects are often recognized first at a "basic" level within an appropriate taxonomy. For example, an animal is recognized as a "dog" before being identified as a "golden retriever". The *basic level* is that at which the most information can be gained with the least effort (Rosch, 1976).

Sound source recognition may operate similarly. A pitched sound might be recognized as arising from a brass instrument before the source is identified as a trumpet. Knowledge of the instrument family (or more specifically, the means of source excitation) enables the prediction of many of the salient acoustic properties of the resulting musical tones. For example, a bowed string will typically have a relatively long onset, will often exhibit frequency modulation characteristic of vibrato, and will become slightly inharmonic when the bow is released.

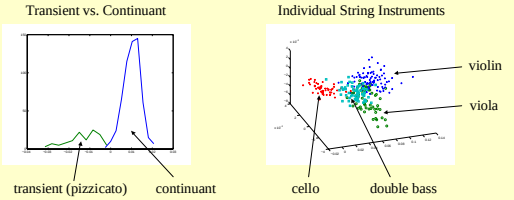
This taxonomy (based on expert knowledge rather than the acoustic signals) is one possible organization for the recognition of orchestral instrument sounds:



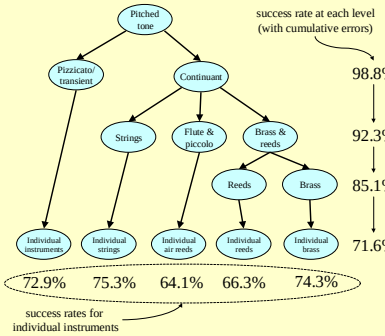
Starting with this taxonomy, a classifier tree was constructed using a Fisher multiple-discriminant analysis at each node.

The classifiers were cross-validated by using only 70% of the instrument samples for training—holding out 30% for testing.

Examples of Fisher projections



Classification results



It turned out to be necessary to modify the proposed classification hierarchy somewhat. The "woodwind" family of orchestral instruments does not appear to be acoustically unified, at least for the features examined here. The best classification results were achieved when the "air" reed instruments (flute/piccolo) formed a separate class, as shown here.

These preliminary results are extremely promising. The hierarchical-classification scheme performed better than several non-hierarchical schemes, including a *k*-nearest neighbor classifier.

Future

This research is still in its early stages. New features are continuing to be investigated, and work is beginning on a system that will analyze musical phrases rather than isolated tones.

One of the primary issues in this work, which has not yet been adequately addressed, is the topic of *generalization*. Systems that can recognize a *particular instrument* played by a *particular performer* in a *particular acoustic environment* are not interesting from a perceptual viewpoint. Much effort has been made to choose features that are likely to be performer-independent and only marginally dependent on the acoustic environment, but this aspect of the work has not yet been tested. Only when the computer can recognize a known instrument performed by a previously-unheard performer will these models be provably relevant to human perception.

Additionally, it is essential that models like this one be equipped to deal with *mixtures* of sounds. Thinking about the problem in terms of auditory scene analysis is a step in the right direction, and the correlogram seems promising as a representation for mixtures of a small number of quasi-periodic sounds, but it is clear that the recognition framework needs to be much more flexible to be able to deal with partially-masked sounds.

References

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

Rosch, E, et al., Basic objects in natural categories, Cognitive Psychology 8, 382-439, 1976.

Acknowledgements

This research is being performed in the Machine Listening Group at the MIT Media Lab and would not be possible without the support of Professor Barry Vercoe. Thanks to DAn Ellis and Eric Scheirer for helpful discussions and occasional suggestions, and to Youngmoo Kim and several students in the Media Lab's Spring '98 pattern recognition class, who worked with the pilot-study features and supplied valuable feedback.