M.I.T. Media Laboratory Perceptual Computing Section Technical Report No. 491, April 1999.

Sound Scene Segmentation by Dynamic Detection of Correlogram Comodulation

Eric D. Scheirer
Machine Listening Group, MIT Media Laboratory
E15-401D Cambridge, MA 02139-4307 USA
eds@media.mit.edu

Abstract: A new technique for sound-scene analysis is presented. This technique operates by discovering common modulation behavior among groups of frequency subbands in the autocorrelogram domain. The analysis is conducted by first analyzing the autocorrelogram to estimate the amplitude modulation and period modulation of each channel of data at each time step, and then using dynamic clustering techniques to group together channels with similar modulation behavior. Implementation details of the analysis technique are presented, and its performance is demonstrated on a test sound.

1. Introduction

The autocorrelogram and similar representations of subband periodicity are now established as the preferred computational representation of early sound processing in the auditory system. This is primarily due to the accuracy with which these models explain the available experimental data on pitch perception. Although there is still debate regarding the strengths and weaknesses of different periodicity representations (Irino and Patterson 1996; Slaney 1997; de Cheveigné 1998; Kaernbach and Demany 1998), it is now relatively uncontested that some sort of temporal periodicity detection is applied to the output of the cochlear filterbank in the human listening process.

A remarkable correspondence between visual motion in the autocorrelogram and the perception of auditory scene analysis was reported some time ago (Duda, Lyon and Slaney 1990). However, as yet there has been relatively little attempt to operationalize this discovery in a computational auditory-scene-analysis (CASA) system. The present paper describes initial experiments in building a CASA system around the principle of detecting subband comodulation in the autocorrelogram domain. The literature on correlogram-based scene analysis is reviewed, and those few approaches to using the correlogram for purposes other than pitch analysis highlighted. Then, the correlation-comodulation algorithm is described. Finally, results on test signals are presented and the directions of future research discussed.

2. Background

Most of the early attempts to construct computational source-grouping systems were based on sinusoidal analysis (Quatieri and McAulay 1998). Details on these sorts of systems are now widely available in the literature (for example, Brown and Cooke 1994; Ellis 1994; Rosenthal and Okuno 1998) and will not be discussed further.

The concept of subband periodicity detection was first suggested as a model for the pitch of a sound by Licklider (1951). Licklider’s model was based on a network of delay lines and coincidence detectors oriented in a two-dimensional representation. Since Licklider's formulation, this technique has been rediscovered several times, first by van Noorden (1983), who cast it in terms of the calculation of histograms of neural interspike intervals in the cochlear nerve. In the last decade, the model was reintroduced by Slaney and Lyon (1990), Meddis and Hewitt (1991), and others; it has since come to be called the *autocorrelogram* method of pitch analysis and is today the preferred model. The autocorrelogram is the 3-D volumetric function mapping a cochlear channel, temporal time delay (or *lag*), and time to the amount of periodic energy in that band at that lag and time.

Weintraub (1985) used a dynamic programming framework around Licklider's autocorrelation model to separate the voices of two talkers whose voices interfere in a single-channel recording. In his model, he attempted to track the pitch of the two voices, and allocate the different channels of the cochleagram to each of the two voices. Summerfield, Lea, and Marshall (1990) presented a convolution-based strategy for separating multiple static vowels in the correlogram. By convolving a two-dimensional wavelet kernel possessing the approximate shape of the “spine and arches” of the pitch structure in a correlogram frame with an image representing the whole frame, they showed that multiple pitches with differing F_0 could be recognized. The stimuli used were simple synthesized vowels with F_0 not harmonically related. Mellinger's (1991) thesis on music analysis contains a brief exploration of motion-based separation in the correlogram, but the techniques he developed for autocorrelogram analysis were never integrated into the main thrust of his system. He was interested in applying the results of image-processing research to sound-analysis systems; in particular, his thesis presented a number of edge detection and flow detection wavelet kernels that were applied to cochleagrams or correlograms.

The approaches of Summerfield *et al.* and Mellinger were notable for attempting to make use of the entire autocorrelogram volumetric function. In contrast, the primary approach has more recently been based on inspection of the *summary autocorrelation* (the (time × periodicity) function given by summing the autocorrelogram across frequencies). The autocorrelogram model of pitch has been extended to allow the separation of simultaneous sounds (Meddis and Hewitt 1992; de Cheveigné 1993). In its simplest form, such a system operates with a residual-driven approach: determine the strongest pitch in the signal by inspecting the summary autocorrelation; remove the filter channels whose behavior can be explained as resulting from this pitch; determine the strongest pitch in the residual; and so on.

Ellis' dissertation (1996) described a system that could analyze sound and segregate perceptual components from noisy sound mixtures, such as city-street ambience. As part of this approach, he developed a novel sound representation termed the *weft* (Ellis and Rosenthal 1998). The weft allows simultaneous representation of pitch and spectral shape for multiple harmonic sounds in a complex sound scene. He provided algorithms for extracting wefts from autocorrelograms in a residue-driven approach, as well as details on their use in sound-scene analysis. Building on the work of Ellis, Martin (1996a; 1996b) developed a music-analysis system for the separation of sound that used the correlogram as the front end. His system used the *blackboard model* to manage multiple hypotheses and data analysis in a prediction-driven framework. Martin demonstrated the performance of his system by analyzing and correctly transcribing four-voice piano music.

In addition to these scene-analysis systems, there have been a few notable attempts to extract features other than pitch from the autocorrelogram. Leman (1995) found cues for musical tonality, Scheirer (1997; 1998) demonstrated the use of the slowed-down autocorrelogram for tempo analysis, and Martin (1999) has recently demonstrated the extraction of features suitable for robust identification of musical instruments.

Although the approach outlined below has not yet been validated with data from psychoacoustic experiments, this step would be necessary in order to further the case that it is a useful model. Darwin and Carlyon (1995) review experiments that suggest the role that frequency and amplitude modulation play in source grouping, although there is not a great deal of literature on this topic.

2.1. Some properties of the autocorrelogram

The grouping approach described in Section 3 is based on modulation properties of the autocorrelogram that are briefly described in this section. This discussion is applicable to periodicity-analysis methods other than autocorrelation; the modulation behavior described here holds for any analysis technique that begins with an array of bandpass filters and then detects periodicity in their output.

As a sound signal evolves over time, the response pattern of each channel of the autocorrelogram modulates in two ways. First, it undergoes *amplitude modulation* in response to changing power in the frequency subband of the signal associated with that channel. As the power increases, the energy output of the cochlear filter increases. Second, it undergoes *period modulation* in response to changes in the frequency of stimulation dominating that cochlear channel.

The bandpass filters comprising the cochlear filterbank can only produce output that is similar to a modulated sinusoid, where the frequency of the sinusoidal carrier is near the center frequency of the filter and the modulation function is very low-passy relative to the carrier. In the frequency region where phase-locking to the waveform occurs, the half-wave rectification and smoothing processes change the shape of the wave function, but not the underlying periodicity. Thus, for any input signal, the autocorrelation function in the low-mid subbands must be roughly periodic with period near that corresponding to center frequency of the subband.

As a time-varying signal evolves, the particular frequency dominating the filter channel – that is, acting as the carrier signal – changes; this change is reflected in corresponding changes in the periodicity of the autocorrelation function. The changes in the autocorrelation function are a sort of modulation, termed *period modulation* for the purposes of the present paper. A simple example is presented in Figure 1.

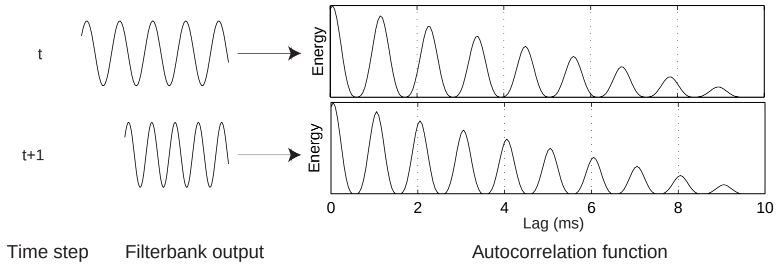

Figure 1: Autocorrelation of a windowed, modulated sinusoid. As the frequency of output from a particular filter channel changes, the period of the autocorrelation function of that output changes. The change in dominant frequency becomes a *period modulation* in the autocorrelation.

Period modulation of the autocorrelation signal in a subband does not only occur in response to frequency modulation of the input signal. It occurs any time that the frequency that is dominating the channel changes. This can happen due to amplitude modulation of the spectral components of the input as well as its frequency modulation. Amplitude modulation and period modulation are not independent; as the frequency dominating a channel changes, it moves closer or farther from the center frequency of the channel, and thus leads to changing output response from the filter.

For simple signals such as those presented below, the changes in periodicity are smooth and easy to analyze. More complex signals present more complex changes in dominant periodicity and are left for future work.

3. Approach

The goal of the system described in the present paper is to allocate channels from a cochlear model into a partitioned representation suitable for further analysis. As such, it is only one component of a complete CASA system. There are no innovations presented regarding the cochlear model or the periodicity analysis; therefore, these components of the approach will only be discussed briefly. Further, although incorporation of top-down information (Ellis 1996; Slaney 1998) is necessary in order to build robust CASA systems, it is not explicitly treated here.

The goal of this work is not to separate sound in the sense of creating multiple output sounds that can be summed to reconstruct a scene. Rather, it is to *partition* the sound data in the correlogram domain so that feature analysis can be undertaken. Many researchers today consider the sound-separation problem to be more of an engineering problem than a scientific one (Martin, Scheirer and Vercoe 1998). The human auditory system does not separate signals in the sense of actually extracting multiple, distinct time-domain sound objects from an acoustic stimulus. Rather, human listeners have the remarkable ability to understand sounds in spite of the noise and overlapping sounds present in the world. The focus of the present research, like the ability of the human listener, is *sound understanding without separation.*

The principle underlying the operation of the system is that articulated by Duda *et al.* (1990): parts of the sound scene that belong together can be seen to undergo coherent modulation when the correlogram is visualized as a moving picture. Slaney and Lyon (1991) produced an excellent “hearing demo reel” videotape that effectively illustrates this principle for many sorts of sounds, including multiple talkers, speech-in-noise, and symphonic music. These demonstrations suggest that the source-partitioning problem might be solved by estimating modulation in the subbands of the autocorrelogram and using that information to group the cochlear channels.

The correlogram-comodulation analysis system is divided into five rough stages that will be described more fully in the subsequent sections. They are: frequency analysis, rectification, and smoothing of sound through models of the cochlea and inner hair cells; subband periodicity analysis with the autocorrelogram; subband modulation detection; clustering of the modulation data to discover comodulation patterns; and feature analysis of the partitioned sound objects. The third and fourth steps make up the new approach, therefore the description is most detailed there.

3.1. Filterbank and periodicity detection

The front-end system used for this research is very similar to others reported in the literature. The particular implementation was programmed by Martin (1999), following the work of Slaney (Slaney 1994) and Ellis (1996). The cochlear filterbank is modeled as a set of 54 eighth-order *gammatone* filters; this model for the cochlea was introduced by Patterson (Patterson *et al.* 1992). The phase-locking behavior of the inner hair cells in the cochlea is modeled with half-wave rectification and smoothing. The output of the cochlear filterbank and rectification processing for a synthetic sound signal, the “McAdams oboe” (McAdams 1984) is shown in Figure 2.

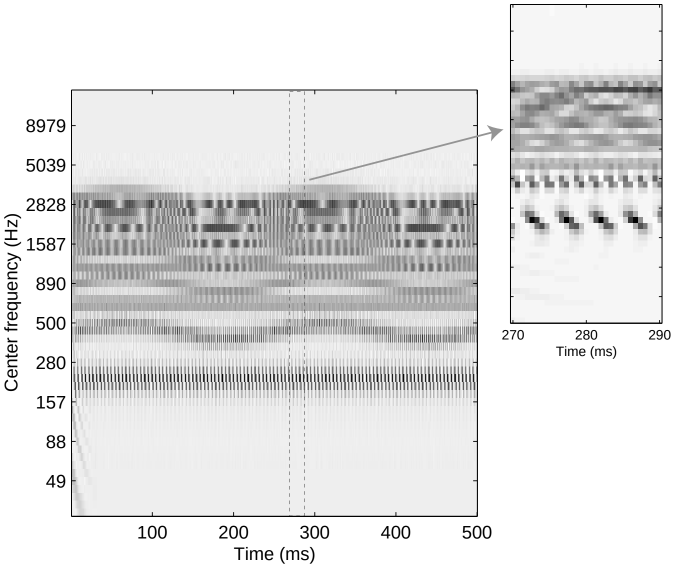

Figure 2: The cochleagram of the “McAdams oboe” sound (McAdams 1984). This sound is comprised of the first ten harmonics of a 220 Hz fundamental with coherent vibrato (10% depth, 4 Hz) applied to the even harmonics only. The percept is that of a clarinet-like sound with pitch at 220 Hz and a soprano-like sound with pitch at 440 Hz. The main panel shows the broad frequency resolution of the filterbank; the vibrato in the second and fourth harmonics can be easily seen. The small panel presents a closer view of the time range around 280 ms; phase-locking in the middle frequencies and lack of phase-locking in the high frequencies is observed.

Periodicity detection is performed using the running log-lag autocorrelogram. In earlier implementations of the autocorrelogram (Slaney 1994), it was calculated on a frame-by-frame basis, by windowing the half-wave rectified output signals of the cochlear filterbank and using the FFT to compute the autocorrelation. More recently, Ellis (1996) and Martin (1999) have suggested using a running autocorrelation rather than a windowing operation. This eliminates edge effects (the decay shown in Figure 1) caused by windowing the signal before calculation, and it also makes it easier to sample the lag axis nonlinearly.

Ellis (1996) observed that as human pitch perception is roughly logarithmic with frequency, it makes more sense to sample the lag axis with logarithmic spacing. He termed this the *log-lag autocorrelogram*. In Martin’s implementation of the log-lag autocorrelogram, delay lines are used to calculate the continuous running autocorrelation without an analysis window. The delay line outputs are computed using fractional delay filters, and after multiplication with the undelayed signal, each lag signal is smoothed with a lowpass filter. This model is much more computationally intensive than Slaney's FFT-based model, but has key modulation properties that will become clear in the next section. Three frames of the autocorrelogram of the synthetic test sound are shown in Figure 3. For the analysis presented here, the autocorrelogram is sampled at a 100 Hz frame rate.

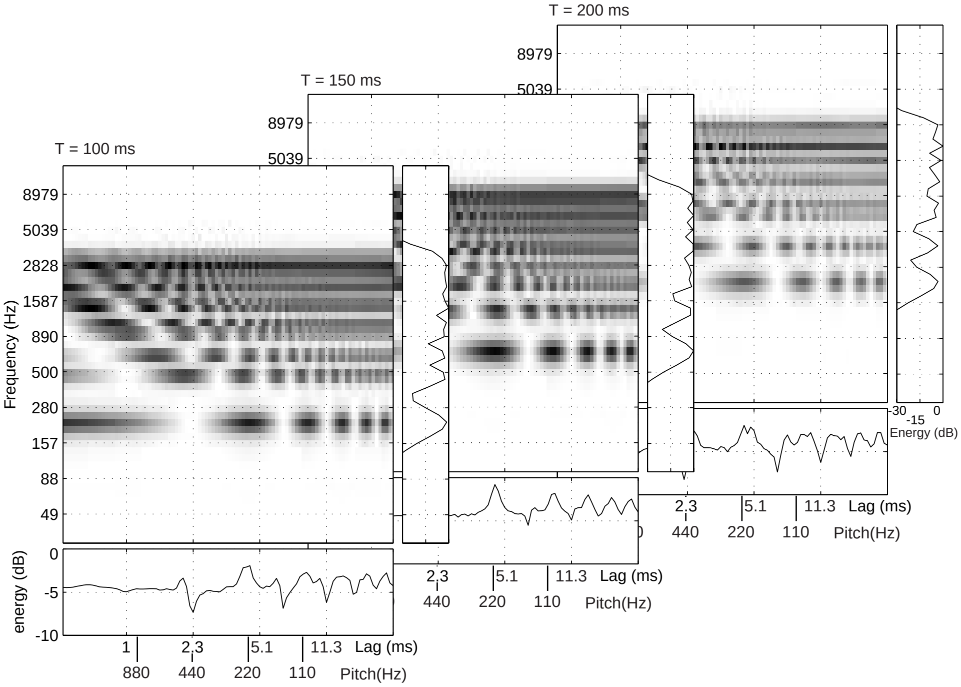

Figure 3: Three frames of the log-lag autocorrelogram of the McAdams oboe. For each frame, the main panel shows a constant-time slice through the autocorrelogram volume, calculated as described in the main text; the bottom panel shows the *summary autocorrelogram*, which is the sum across frequencies of the periodic energy at each lag (the sum is calculated in the linear-energy domain and then converted to dB scale for presentation); and the right panel shows the *energy spectrum*, which is the zero-lag energy in each channel. The three frames highlight different portions of the vibrato phase for the even harmonics; the first shows a point in time at which the even harmonics are sharp relative to the odd harmonics, the second at which the even harmonics are in tune with the odd harmonics, and the last when the even harmonics are flat. Readers familiar with pitch-based source separation techniques will observe the difficulty in distinguishing the in-tune from detuned partials using only the information in the summary autocorrelation.

3.2. Modulation Analysis

In this section, new techniques for analyzing the modulation behavior of channels of the autocorrelogram are presented. The purpose of modulation analysis is to convert the dynamic motion of the autocorrelogram into static features that are suitable for inclusion in a pattern-analysis system.

On a linear lag axis, a simple frequency modulation such as a vibrato corresponds to a period modulation that can be described as *stretching* and *squashing*. That is, when the frequency of the signal stimulating a particular filter channel increases, the output of the filter also increases in frequency. The autocorrelation function thus is squashed, with peaks closer together, as in Figure 1. As the signal frequency decreases, the output of the filter decreases in frequency and the peaks of the autocorrelation function are stretched.

The utility of the log-lag autocorrelogram for detecting period modulation now becomes evident. When the lag axis is scaled logarithmically, the stretch-squash effect of period modulation becomes a simple shift to the left or right, which is easily to analyze. Slices of the log-lag autocorrelogram for two cochlear channels (one steady, and one undergoing period modulation) for the synthetic sound used in the previous figures are shown in Figure 4.

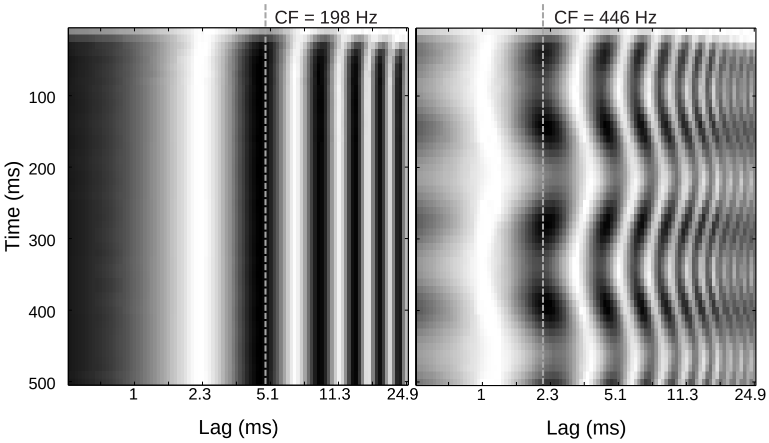

Figure 4: Two slices through the autocorrelogram of the McAdams oboe sound. Each panel shows the autocorrelation response of a single filter over time. The left panel corresponds to a cochlear filter with center frequency 198 Hz; thus, this channel is dominated by the steady partial at 220 Hz in the sound. The right panel shows the response of a cochlear filter with center frequency 446 Hz; thus, this channel is dominated by the partial that frequency-modulates about 440 Hz. Since the lag axis is logarithmically scaled, the period modulation is reflected as linear shifting behavior over time, not stretching and squashing—all of the curves in the right panel are parallel.

In the log-lag domain, cross-correlation can be used to detect period modulation. At each time step, the autocorrelation function in each channel is cross-correlated with the autocorrelation function in the same channel from a previous time step. If the channel is period-modulating, the peak of this cross-correlation function is off-center. Peak-picking suffices to determine the period modulation, at least for the simple examples tested so far.

The finite length of the autocorrelation vector provides an implicit windowing function in the cross-correlation. That is, the running autocorrelation is only calculated over a finite set of lags, and so it can be viewed as the application of a rectangular (boxcar) window function to the true, infinitely-long, autocorrelation function. This windowing function biases the peak estimate in the cross-correlation towards the center, since it has a triangular autocorrelation function. In order to provide accurate estimates, the cross-correlation is unbiased by multiplication with the inverse triangle window before peak-picking.

The domain of period-modulation values is *lag scale per second*. At each time step, the autocorrelation function is scaled by some ratio, which may be detected by looking for shifts in the log-lag autocorrelogram. The primary region of interest in this domain is -2 % to +2% lag scale per 10 ms frame; modulations more extreme than these tend to be hidden by the center-bias of the windowed cross-correlation. Very small modulations, between –0.2 % and +0.2 % per frame, are difficult to detect due to the lack of high-frequency information in the autocorrelation function.

The output of the period-modulation detection is shown in Figure 5. This figure, a *period modulogram*, shows the two-dimensional function mapping cochlear channel and time into the period-modulation estimation in that channel at that time.

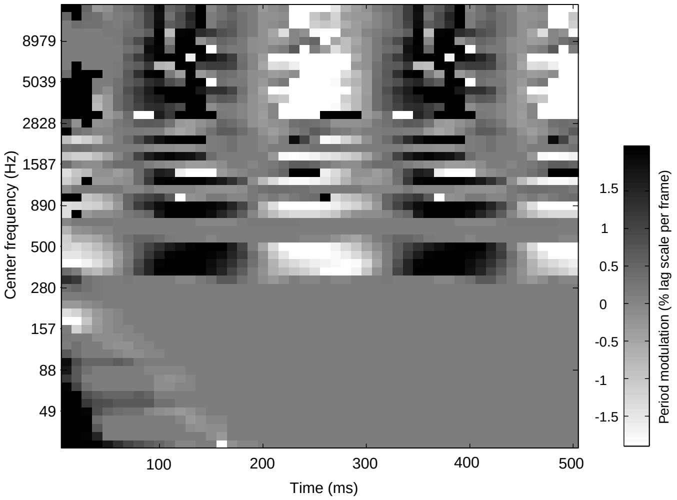

Figure 5: The period modulogram, showing the period modulation of each channel at each time step. The period modulation is measured with cross-correlation as described in the text; the legend at right shows the correspondence between the gray level in the main panel and the degree of period modulation. Period modulation is measured in lag scale per time; a value of 1.5 % means that the lag axis in that channel at that time must be stretched by a factor of 1.5 % (that is, a multiplicative scaling of 1.015) in order to approximate the lag axis in the same channel at the next time step. Similarly, a value of -1.5 % corresponds to a multiplicative squashing factor of 0.985. Compared to Figure 3, the channels responding to the vibrato are clearly visible.

The amplitude modulation in each channel is also measured, by dividing the zero-delay autocorrelation—the energy—in each channel by the zero-delay autocorrelation from a previous frame. The *energy scale per frame* is greater than 0 dB when the channel is increasing in power, and less than 0 dB when the channel is decreasing in power. The output of the amplitude-modulation detection process—the *amplitude modulogram*—is shown in Figure 6. There is no explicit amplitude modulation in this signal after the onset, so all of the amplitude modulation arises from coupling to frequency modulation.

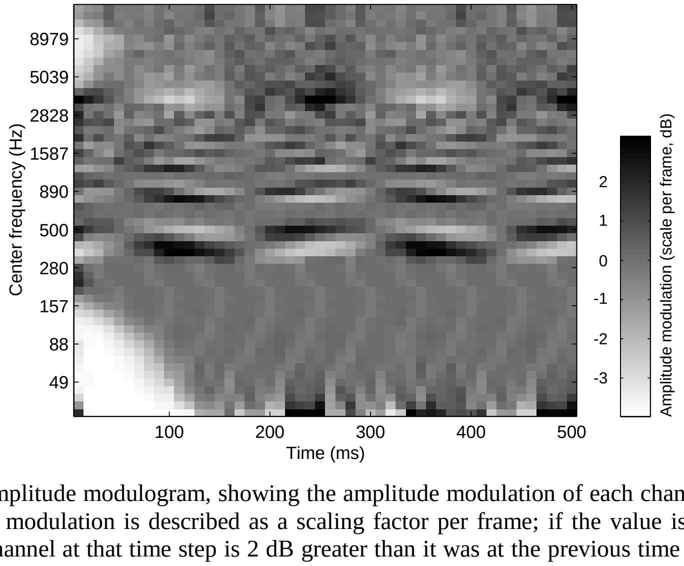

Figure 6: The amplitude modulogram, showing the amplitude modulation of each channel at each time step. Amplitude modulation is described as a scaling factor per frame; if the value is 2 dB, then the energy in that channel at that time step is 2 dB greater than it was at the previous time step. When the figure is compared to Figure 5, the correlation between period modulation and amplitude modulation is observed. The structure in the very low and very high channels is spurious; there is very little energy in these channels as can be seen in the energy profiles plotted in Figure 3, and so the amplitude modulation measurement there is not meaningful.

There is potentially more information to be retrieved from the modulation patterns by using more than one previous frame; that is, by cross-correlating the autocorrelation function at time *t* in a channel with that of the same channel at 1 ms, 10 ms, and 100 ms previous. With appropriate smoothing, this sort of *multiscale* processing could give information on coherent motion at many time resolutions, from glottal jitter to syllabic or note-to-note transitions. The modulation plots shown here were computed using a 40 ms (four frame) time step; this delay was chosen empirically to give the best visual results.

3.3. Dynamic clustering

The modulation-detection principles described in the previous section convert the dynamic motion of the correlogram into static features, and the cross-channel concept of frequency modulation into the within-channel concept of period modulation. The next step in comodulation analysis is to determine which channels are modulating in the same way. There is a substantial literature on clustering, grouping, and modeling the distribution of data scattered in a multidimensional feature space. In the case here, the feature space has only two dimensions; each channel at each time is mapped to an ordered pair (*p*,*a*), where *p* is the current period modulation and *a* the current amplitude modulation. Within each frame, only those channels containing significant acoustic power (at least –15 dB compared to the channel with maximum power, see Figure 3) are considered. Thus, at each time step there are at most 54 ordered pairs, one for each cochlear channel, but usually less, since not every channel has energy at every time. Figure 7 shows a scatterplot of the feature space at four points in time.

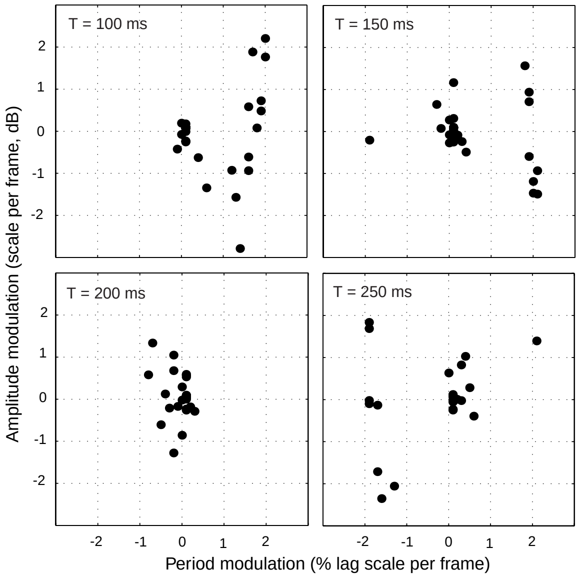

Figure 7: Scatterplots of the modulation data shown in Figure 5 and Figure 6 at four different time steps. Each point corresponds to the behavior of one cochlear channel at one point in time. The latest three frames may be compared to the three frames in Figure 3 (the time markings differ due to the delay in computing modulation). In these frames, the even harmonics are becoming respectively sharper, not changing, and flatter with respect to the odd harmonics.

The ISODATA technique (Therrien 1989) is used to compute the dynamic clustering of the data. ISODATA is a heuristic procedure which iteratively assigns data points into groups based on the distance between the cluster means. It is controlled by parameters that specify the approximate number of groups and how large (in terms of spread in the feature space) they are allowed to be. Figure 8 shows the same time steps as Figure 7, with the cluster assignments determined by ISODATA. In this case, the control parameters were fixed and empirically chosen. The data have been successfully partitioned into clusters based on comodulation patterns.

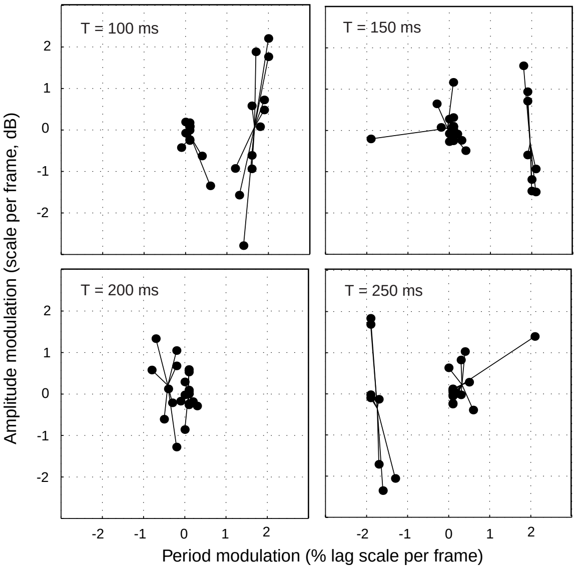

Figure 8: Assignment of the modulation data into clusters. The clustering is performed automatically by a slightly-modified version of the ISODATA procedure (see text). Each cluster corresponds to a group of cochlear channels that are doing the same thing at a particular time.

The standard ISODATA algorithm can only be used on individual frames of data, since it has no notion of *time-dependent* cluster. Thus, for use in this system, the ISODATA procedure is modified slightly. At each time step, the cluster distribution at the previous time step is given as an *a priori* cluster distribution. Doing this encourages ISODATA to maintain continuity within clusters. Also, at each time step, the ISODATA control parameters could be modified depending on how successful the clustering was at the previous time step. If the clustering was an excellent model of the data, then the control parameters could be adjusted to make it more difficult to change the clustering in the subsequent time step. This dynamic adjustment proved unnecessary for the example presented here.

These modifications were undertaken somewhat empirically and have not been thoroughly tested on sound scenes with different properties. It would probably be desirable to move away from ISODATA to more sophisticated dynamic clustering models. There are many such models in the literature. In particular, many operate on a more formalized maximum-likelihood Bayesian model. They are thus more easily modified to use constraints such as the desired inertia of clusters over time in a principled way, by expressing of the constraints as probabilistic priors.

3.4. Feature analysis

Based on the dynamic clusters formed in the previous stage of analysis, the sound scene can be segmented into a number of acoustic objects. Each object is coherent in dynamic correlogram motion; that is, at each time step, each channel within an object is modulating in the same way in period and amplitude.

Figure 9 shows the object masks that result from the cluster analysis in the previous section. The partitioning technique follows Bregman’s principle of *exclusive allocation* (Bregman 1990, p. 12); each cochlear channel at each time is assigned only to one object. As can be seen in this figure, the energy in the scene has been assigned to two objects, one corresponding to the odd harmonics, which are static, and one corresponding to the even harmonics, which are modulated. Although this scene is composed only of two objects, there is no apparent reason that the technique should not work for more than two objects, as long as each has independent and coherent modulation properties.

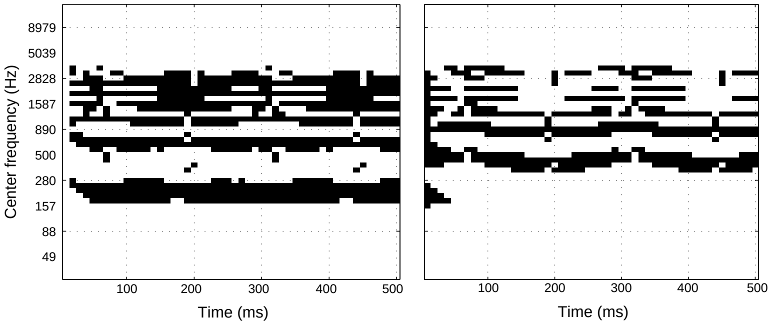

Figure 9: Object masks showing the segmentation of the scene into objects. In each panel, the dark regions of the time-frequency plane correspond to energy that undergoes coherent modulation in the correlogram. The left panel corresponds to the static (odd) partials of the McAdams oboe sound; the right panel to the modulating (even) partials. The high frequencies, low frequencies, and occasional middle frequencies are not assigned to any object because they have relatively little energy.

It may be observed in Figure 9 that the high midrange (around 1500-3000 Hz) is assigned inconsistently, first to one object, then the other. In this frequency region, the frequency modulation causes the modulating partials to interfere with the static partials, in different cochlear channels.. Thus, the sum of the two objects manifest as regular patterns of within-channel beating. Techniques for analyzing this behavior have not yet been determined; currently, the present technique seems to work robustly only for partials that are resolved by the filterbank.

Using the object masks in Figure 9, we can perform feature estimation on each object separately. For example, we can calculate the pitch of each object, using the standard summary-autocorrelation procedure, but at each time step, only considering the channels that are associated with one object or the other. The pitch-tracks of the two objects are shown in Figure 10. Similar feature analysis could be used to produce estimates of broad spectral shape, formant positions, or other desired acoustic features. Unlike most of the scene-analysis techniques put forth recently in the literature, the scene analysis and pitch estimation are not performed jointly. Rather, the scene analysis step in this model precedes the pitch estimation, thus avoiding the problematic situation (such as for the McAdams oboe example presented here) in which the overlap of two objects blurs the pitches in the summary autocorrelation. By using the dynamic motion cues in the periodicity representation, the sources can be separately analyzed without using pitch or harmonicity as a grouping heuristic.

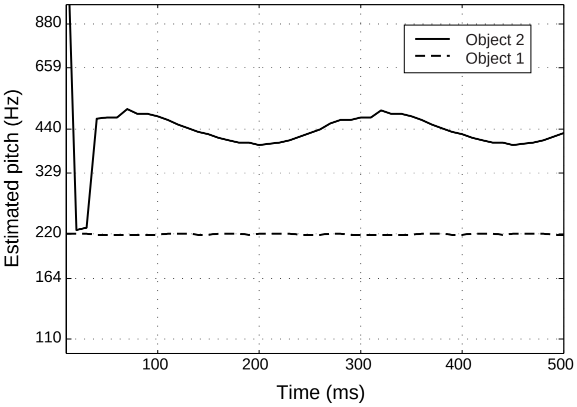

Figure 10: Pitch tracks of the two objects shown in Figure 9. After a brief startup transient, the pitches of the two perceptual objects in the scene are correctly determined.

4. Summary and Future Work

This paper has presented a new technique for segmenting a sound scene into acoustic objects. The principle of *correlogram comodulation detection* is identified as a useful heuristic for computational auditory-scene-analysis systems. By analyzing the modulation cues within the correlogram or equivalent subband periodicity representations, the dynamic behavior of the correlogram may be converted into static features suitable for cluster analysis. The scene can be partitioned into objects, and the features of the objects analyzed, based on automatic grouping of the cochlear channels by their modulation properties. The utility of the technique was demonstrated by testing it on the “McAdams oboe” sound, a sound that is difficult to analyze using frame-by-frame correlogram techniques.

The material discussed in the present paper should be considered a work in progress. The results tend to raise more questions than they answer. In particular, more evaluation of the period-modulation heuristic is needed. It is important to discover whether this technique may apply to noisy sounds and sound scenes that are more complex than the example presented here. This technique seems only applicable to the analysis of periodic and quasi-periodic sounds; thus, in a full CASA system intended to deal with a variety of sound sources, it would need to be augmented by other techniques more suited to noisy sounds and sound textures.

The question of proper time scale for modulation analysis has not yet been addressed; the examples presented here were created with an arbitrarily chosen time scale. Further expansion on this topic could include the analysis of multiscale processing in the same framework.

In the long run, if this technique proves useful for the computational analysis of different sorts of sound scenes, it will be natural to examine the psychoacoustic hypotheses that it puts forth. For example, the model cannot use periodicity modulation to group cochlear channels in which phase-locking to the signal is not present. There is little data in the psychoacoustic literature on this grouping behavior in human listeners. Also, the model strongly argues for an exclusive-allocation approach to scene analysis. If experiments were to reveal “conjoint” (in the language of Summerfield *et al.* 1990) grouping strategies in the human auditory system, the model presented here would have to be greatly revised to account for such data.

The long-term goal of the research that includes the present paper is to create computational *music perception systems* that can robustly make judgments (such as genre classification) on real examples taken directly from compact-disc recordings. By examining the perception of music as a serious problem in auditory scene analysis, the normally disparate fields of music perception and psychacoustics may be drawn together. Such a prospect requires re-evaluation of models, systems, and theories from psychology, psychoacoustics, and musical signal processing (Scheirer 1996; Martin *et al.* 1998).

Acknowledgments

The research presented here was supported by the Digital Life Consortium of the MIT Media Laboratory. Thanks to Keith Martin for ongoing discussion, collaboration, and criticism, and especially for allowing me to use his implementation of the Patterson filterbank and log-lag autocorrelogram. Thanks also to Profs. Perry Cook and Roz Picard and to Dr. Malcolm Slaney for their research supervision. As always, Prof. Barry Vercoe and the Machine Listening Group at the Media Lab were essential through their support and suggestions.

References

Bregman, A. (1990). *Auditory Scene Analysis*. Cambridge MA: MIT Press.

Brown, G. J. & Cooke, M. (1994). Perceptual grouping of musical sounds: A computational model. *Journal of New Music Research* **23**, 107-132.

Darwin, C. J. & Carlyon, R. P. (1995). Auditory grouping. In B. C. J. Moore (ed.) *Hearing* (pp. 387-424). San Diego: Academic Press.

de Cheveigné, A. (1993). Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. *Journal of the Acoustical Society of America* **93**(6), 3271-3290.

de Cheveigné, A. (1998). Cancellation model of pitch perception. *Journal of the Acoustical Society of America* **103**(3), 1261-1271.

Duda, R. O., Lyon, R. F. & Slaney, M. (1990). Correlograms and the separation of sounds. In *Proceedings of the 1990 IEEE Asilomar Workshop*. Asilomar CA.

Ellis, D. P. W. (1994). A computer implementation of psychoacoustic grouping rules. In *Proceedings of the 1994 12th ICPR*. Jerusalem.

Ellis, D. P. W. (1996). *Prediction-Driven Computational Auditory Scene Analysis*. Ph.D. thesis, MIT Dept. of Electrical Engineering and Computer Science, Cambridge MA.

Ellis, D. P. W. & Rosenthal, D. F. (1998). Mid-level representations for computational auditory scene analysis: The weft element. In D. F. Rosenthal & H. G. Okuno (eds.), *Readings in Computational Auditory Scene Analysis* (pp. 257-272). Mahweh NJ: Lawrence Erlbaum.

Irino, T. & Patterson, R. D. (1996). Temporal asymmetry in the auditory system. *Journal of the Acoustical Society of America* **99**(4), 2316-2331.

Kaernbach, C. & Demany, L. (1998). Psychophysical evidence against the autocorrelation theory of auditory temporal processing. *Journal of the Acoustical Society of America* **104**(4), 2298-2306.

Leman, M. (1995). *Music and Schema Theory*. Berlin: Springer-Verlag.

Licklider, J. C. R. (1951). A duplex theory of pitch perception. *Experientia* **7**, 128-134.

Martin, K. D. (1996a). Automatic transcription of simple polyphonic music: Robust front-end processing. MIT Media Laboratory Perceptual Computing Technical Report #399, Cambridge MA.

Martin, K. D. (1996b). A blackboard system for automatic transcription of simple polyphonic music. MIT Media Laboratory Perceptual Computing Technical Report #385, Cambridge MA.

Martin, K. D. (1999). *Toward a Machine Listener: Recognizing Sound Sources*. Ph.D. thesis, Massachusetts Institute of Technology Department of Electrical Engineering, Cambridge, MA.

Martin, K. D., Scheirer, E. D. & Vercoe, B. L. (1998). Musical content analysis through models of audition. In *Proceedings of the 1998 ACM Multimedia Workshop on Content-Based Processing of Music*. Bristol UK.

McAdams, S. (1984). *Spectral Fusion, Spectral Parsing, and the Formation of Auditory Images*. Ph.D. thesis, Stanford University CCRMA, Dept of Music, Stanford, CA.

Meddis, R. & Hewitt, M. J. (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *Journal of the Acoustical Society of America* **89**(6), 2866-2882.

Meddis, R. & Hewitt, M. J. (1992). Modeling the identification of concurrenet vowels with different fundamental frequencies. *Journal of the Acoustical Society of America* **91**(1), 233-244.

Mellinger, D. K. (1991). *Event Formation and Separation in Musical Sound*. Ph.D. thesis, Stanford University Dept. of Computer Science, Palo Alto CA.

Patterson, R. D., Robinson, K., Holdsworth, J. *et al* (1992). Complex sounds and auditory images. In Y. Cazals, K. Horner & L. Demany (eds.), *Auditory Physiology and Perception* (pp. 429-446). Oxford: Pergamon Press.

Quatieri, T. F. & McAulay, R. J. (1998). Audio signal processing based on sinusoidal analysis/synthesis. In M. Kahrs & K. Brandenburg (eds.), *Applications of Digital Signal Processing to Audio and Acoustics* (pp. 343-411). New York: Kluwer Academic.

Rosenthal, D. F. & Okuno, H. G. (eds.) (1998). *Computational Auditory Scene Analysis*. Mahweh, NJ: Lawrence Erlbaum.

Scheirer, E. D. (1996). Bregman's chimerae: Music perception as auditory scene analysis. In *Proceedings of the 1996 International Conference on Music Perception and Cognition* (pp. 317-322). Montreal: Society for Music Perception and Cognition.

Scheirer, E. D. (1997). Pulse tracking with a pitch tracker. In *Proceedings of the 1997 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Mohonk, NY: IEEE.

Scheirer, E. D. (1998). Tempo and beat analysis of acoustic musical signals. *Journal of the Acoustical Society of America* **103**(1), 588-601.

Slaney, M. (1994). Auditory toolbox. Apple Computer, Inc. Technical Report #45, Cupertino CA.

Slaney, M. (1997). Connecting correlograms to neurophysiology and psychoacoustics. In *Proceedings of the 1997 XIth International Symposium on Hearing*. Lincolnshire UK.

Slaney, M. (1998). A critique of pure audition. In D. F. Rosenthal & H. G. Okuno (eds.), *Computational Auditory Scene Analysis* (pp. 27-42). Mahweh, NJ: Lawrence Erlbaum.

Slaney, M. & Lyon, R. F. (1990). A perceptual pitch detector. In *Proceedings of the 1990 ICASSP* (pp. 357-360). Albequerque.

Slaney, M. & Lyon, R. F. (1991). Apple Hearing Demo Reel. Apple Computer, Inc. Technical Report #25, Cupertino CA.

Summerfield, Q., Lea, A. & Marshall, D. (1990). Modelling auditory scene analysis: strategies for source segregation using autocorrelograms. *Proceedings of the Institute of Acoustics* **12**(10), 507-514.

Therrien, C. W. (1989). *Decision, Estimation and Classification: An Introduction to Pattern Recognition and Related Topics*. New York: Wiley.

van Noorden, L. (1983). Two-channel pitch perception. In M. Clynes (ed.) *Music, Mind, and Brain: The Neuropsychology of Music* (pp. 251-269). New York: Plenum Press.

Weintraub, M. (1985). *A Theory and Computational Model of Auditory Monaural Sound Separation*. Ph. D. thesis, Stanford University Dept. of Electrical Engineering, Palo Alto, CA.