M.I.T. Media Laboratory Perceptual Computing Section Technical Report #492, April 1999

submitted to the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk, NY

TOWARDS MUSIC UNDERSTANDING WITHOUT SEPARATION: SEGMENTING MUSIC WITH CORRELOGRAM COMODULATION

Eric D. Scheirer

Machine Listening Group
Media Laboratory, Massachusetts Institute of Technology
E15-401D, Cambridge MA 02139-4307 USA
eds@media.mit.edu

ABSTRACT

The application of a new technique for sound-scene analysis to the segmentation of complex musical signals is presented. This technique operates by discovering common modulation behavior among groups of frequency subbands in the autocorrelogram domain. The algorithm can be demonstrated to locate perceptual events in time and frequency when it is executed on ecological music examples taken directly from compact disc recordings. It operates within a strict probabilistic framework, which makes it convenient to incorporate into a larger signal-understanding testbed. Only within-channel dynamic signal behavior is used to locate events; therefore, the model stands as a theoretical alternative to methods that use pitch as their primary grouping cue. This segmentation algorithm is one processing element to be included in the construction of music perception systems that understand sound without attempting to separate it into components.

1. INTRODUCTION

When human listeners are presented with musical signals, they automatically and naturally begin to hear them as collections of auditory objects. The primitive features of each object, and simple relationships amongst the objects, determine the overall *surface features* of the music. From the surface features of the music, the listener is able to make immediate judgments, such as determining the tempo, genre, composer, performer, style, complexity, and degree of polyphony in the music. There is preliminary evidence that surface information is also adequate for decoding the emotive intent of the performer or composer [1].

While there are some similarities between the auditory segmentation process and attempts to build systems for polyphonic pitch-tracking or *automatic transcription*, these processes are not identical. In particular, while transcription systems typically founder on the task of segregating "notes," especially when the notes bear a harmonic relationship to each other, there is no evidence that the human auditory system actually performs segregation to such a fine degree [2]. Rather, musical segmentation in the general case is performed only coarsely, and many times "notes" are left grouped together in perception. We often perceive chords holistically rather than analytically. Bregman [3, pp. 459-460] terms the percept of many-grouped-notes a *chimerical* auditory object.

The construction of systems that can model the surface-analysis process of music is an interesting problem in two domains. Such an effort may be treated as a scientific inquiry; there is little known about the perception of complex sound scenes such as those found in music, and efforts to build better models will further our understanding of the hearing process in general. It may also be taken as an engineering inquiry; it has been argued [4] that building models of musical hearing is the best way to approach the construction of music-analysis and music-retrieval systems.

The present paper discusses a new technique for analyzing the autocorrelogram sound-periodicity representation, and the application of this technique to the analysis of musical signals. By calculating the cross-channel comodulation behavior of the autocorrelogram, a complex musical signal may be partitioned into perceptual segments suitable for feature analysis. The comodulation technique operates at a primitive, prefeature signal level, and is thus a theoretical alternative to models that use pitch as a cue for perceptual grouping. This approach may be considered as a step towards the construction of music-understanding systems.

2. APPROACH

It is not the goal of this research to perform "signal separation" in the sense of producing multiple, cleanly synthesized output signals from a given musical scene. Rather, the goal is modeled after the ability of the human listener: to perform *understanding without separation* in the musical domain. The difference between the goal represented here and the goal represented by most previous research into computational auditory-scene analysis (CASA) systems is represented schematically in Figure 1.

In a traditional CASA system, the goal of sound-processing is to extract multiple "component" sounds from a mixture. The output sounds can then be analyzed independently to compute their features. The sounds that are the output should be the same in some perceptually important way as the sounds that acted as components of the mixture. A primary motivating factor for this approach is its potential application to automated speech recognition (ASR). ASR systems today perform passably well on clean speech without interference; this makes it attractive to imagine "cleaning up" signals so that they can be used as input to unmodified ASR systems.

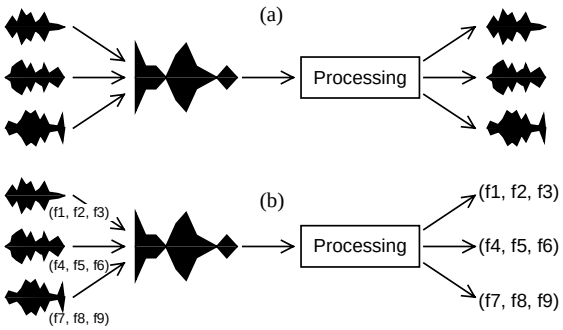

Figure 1: Different models for computational auditory scene analysis. In (a), a *sound separation system* analyzes a sound mixture to discover the sounds that comprise it. In (b), a *sound understanding system* analyzes a sound mixture to discover the *features* of the sounds that comprise it.

In contrast, the approach embodied by the present research is to robustly extract features from complex scenes that are the same as the features of the component sounds. It is apparent that this task is easier, since less time must be spent on achieving high-quality synthesis of output sounds, and that it is more similar to the human hearing process, since human listeners do not maintain multiple independent time-domain signals as an intermediate representation of complex signals.

The advantage of the understanding-without-separation approach is most apparent in the case when one component signal destroys information in another through masking or cancellation. In a sound-separation system, it is very difficult to deal with this situation properly, since the obliterated sound must be invented wholesale from models or *a priori* assumptions. In an separationless approach, the required action is one of making feature judgments from partial evidence, a problem that is treated frequently in the pattern recognition and artificial intelligence literature. Rather than having to invent a answer, the system can delay decision-making, work probabilistically, or otherwise avoid the problematic situation until a solution presents itself.

The major difficulty of this approach is evaluating the behavior of systems that embody it. When the goal of a system is to extract clean-sounding independent components, it is easy to listen to the outputs to see if the system is doing the right thing. When the goal is to extract perceptual features, for which there may or may not be any ground truth to be measured from the signal, it is necessary to continually compare the behavior of the system with that of human listeners. Although results of human listening experiments will not be presented here, comparison with human judgments is an essential part of evaluating the performance of any purportedly perceptual computing system.

3. PROCESSING

A sound-analysis system is being developed to explore new techniques of musical signal processing and to refine the understanding-without-separation paradigm. This section describes the operation of the system; due to space restrictions, the description is necessarily very concise.

The core representation in this system is the *log-lag autocorrelogram* [5]. The autocorrelogram is the volumetric function mapping time, cochlear channel, and lag to the amount of periodic energy in a signal at that point in time, frequency, and periodicity. The autocorrelogram and similar models of subband periodicity [6-8] are similar to the Licklider [9] "duplex" model of pitch perception. This is now the preferred model of early auditory processing due to the accuracy with which it explains the available experimental data on pitch perception. Ellis [5] suggested logarithmic scaling of the lag axis on the basis of maintaining similarity to pitch perception, this variant also presents additional advantages that will become clear below.

Several techniques have been proposed for the analysis of simultaneous sounds in periodicity representations. Many of them [5, 8, 10] use pitch as a cue for grouping, typically in a residual-driven approach: the dominant pitch of the mixture is calculated, a signal with this pitch is subtracted, the dominant pitch of the remaining material is calculated, and so forth.

3.1. Amplitude and period modulation

Rather than follow a pitch-driven approach, the present system follows a qualitative observation regarding the correlogram that was first reported by Duda *et al.* [11]. When the correlogram is viewed as a movie, showing one "frame" of lag × frequency data after another, cochlear channels that correspond to the same auditory object can be seen to undergo coherent visual motion. The coherent motion appears either as *amplitude modulation*, in which several channels all get louder and softer together, or as *period modulation*, in which the autocorrelation functions of several channels all are stretched and squashed at the same rate.

Period modulation is not the same as frequency modulation, since the former is a within-channel feature and the latter is an across-channel feature. Frequency modulations in signals give rise to period modulations in the correlogram; as the frequency dominating a particular cochlear channel changes, the periodic rate of modulation of the channel output changes correspondingly. This leads to a fairly strong hypothesis, which could be tested empirically, about the perception of frequency modulation: frequency modulation is only detected and incorporated into perceptual processing to the extent that it has within-band period and amplitude modulation effects.

An in-depth report of a processing model that can measure the dynamic behavior of the autocorrelogram has been recently presented [12]. In brief, the amplitude modulation for each channel is calculated by comparing the output power in that channel over one time interval to the output power over the next. The period modulation for each channel is calculated by finding the peak cross-correlation between the autocorrelation function at one point in time and the autocorrelation function at the next. The cross-correlation technique for estimating period modulation works because of the use of the log-lag autocorrelogram. When the lag axis is calculated with logarithmic spacing, the stretch-squash behavior of period modulation is represented as simple shifts of the channel to the left or right.

The complete report of this method shows its detailed operation on a sound that is perceptually segregated due to common-frequency-modulation cues (the "McAdams oboe"). The period-modulation-estimation method for analyzing the correlogram is similar in some ways to the method presented by Mellinger [13], for grouping "partials" based on their frequency-modulation behavior. It is different in important aspects, however; notably, the Mellinger technique is a cross-channel integration technique, while the period-modulation analysis is a within-channel technique. Also, the system presented here is not based on "partials" or other primitive objects.

3.2. Dynamic clustering

The two primitive "prefeatures" (amplitude and period modulation) are presented to an untrained dynamic clustering framework, which groups together channels to form *object masks* that may be applied to the cochleagram or autocorrelogram. The dynamic clustering operates in two stages: an instantaneous estimation of cluster density, followed by a Viterbi procedure that analyzes the dynamic changes in group membership of the cochlear channels.

The cluster density process operates on a frame-by-frame basis, by using the EM algorithm [14] to estimate the parameters of a Gaussian mixture model [15]. In each frame, the two prefeatures span a two-dimensional feature space within which each cochlear channel is a point (since an ordered pair—the two prefeatures—is calculated for each channel). The Gaussian mixture model determines a probability density function around centers of common modulation in this feature space and the *a posteriori* likelihood that each channel is a member of each cluster. Currently, the number of clusters is set intuitively, but this could be extended to include a more principled approach in the future.

The Viterbi procedure computes maximum-likelihood paths for each channel, using the posterior grouping probabilities calculated in the clustering step and *ad hoc* prior probabilities for the movement of channels from object to object. This procedure itself is divided into two stages. In the first stage, the association of clusters with objects is computed. This stage is necessary since the EM procedure does not produce any correspondence from frame to frame concerning which cluster is labeled "Cluster #1." The Viterbi algorithm [15] is used to compute the maximum-likelihood path of cluster-to-object associations under the assumption that every channel stays in the same group from time-step to time-step. The result of this looks something like "at time 1, object A is cluster 1 and object B is cluster 2, while at time 2, object A is cluster 2 and object B is cluster 1". In the second stage, given the associations computed in the first stage, the Viterbi algorithm is used again to compute the maximum-likelihood membership of each cochlear channel at each time.

The final result of this processing is a membership function $F(n,t)$ that maps from a cochlear channel n and a time step t to the identity of the object to which that channel belongs to at that time. This is an exclusive allocation (to use the term of Bregman [3]) model in which each cochlear channel belongs only to one object at a time. The set of time/frequency points M_k for object k

$$M_k = \{ (n,t) : F(n,t) = k \}$$

may be taken as a *mask* that can be used to select a particular region of the time-frequency space for analysis.

Each object, according to the corresponding mask, may be analyzed for features directly from the masked data. No attempt is made to "clean up" the masks for resynthesis; any cleanup necessary for perceptually modeling should be applied in the feature-estimation process. The masks, as shown in the next section, are not intended to *separate* the sounds such that there is no leakage from one object into another. If they must be given an acoustic interpretation at all, they could be considered as enhancing one part of the signal relative to another. Within the present approach, it is preferred simply to consider them as the "places to look" in the signal to estimate features.

4. EXAMPLES

This section graphically presents results produced by the segmentation algorithm. The sound examples used, along with a variety of similar images, can be found on the author's WWW page at <http://sound.media.mit.edu/~eds>. An appropriate evaluation of these results has not yet been undertaken, since it involves a fair amount of work with human subjects to determine if the grouping results presented here are in accord with the perceptions of listeners. These results only serve to demonstrate the sort of scene partitioning that is currently achieved by the system. The plots should not be interpreted as an attempt to convince the reader that the system "works."

Figure 2 shows three different musical excerpts automatically converted into object masks. Each of the excerpts is an "ecological" music signal sampled directly from radio tuner input at 22 kHz sampling rate. Simply by inspecting the features of the object masks, we can correctly interpret many aspects of the musical signals: the first is the most complex, the second has sporadic broadband energy (snare drum hits), and so forth. However, it is also apparent that the sequential grouping of the algorithm is rather poor. This is highlighted more clearly in Figure 3.

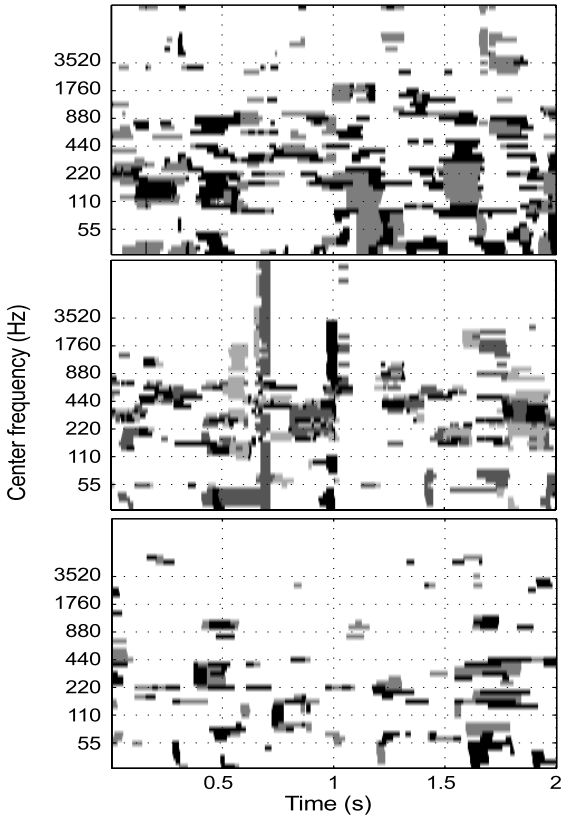

Figure 2: Object masks for three different musical excerpts. Top, a rock example, partitioned into three objects; middle, a jazz piano trio, partitioned into four objects, and bottom, a Mozart symphony, partitioned into three objects. The "background" is one of the objects in each case. Each color corresponds to one object mask – that is, the algorithm asserts that all the black-colored time-frequency cells belong together in one object, all the dark gray cells in another, and so on. When these masks are inspected in comparison to the sound of the acoustic signals, it is apparent that many of the perceptual objects have been located.

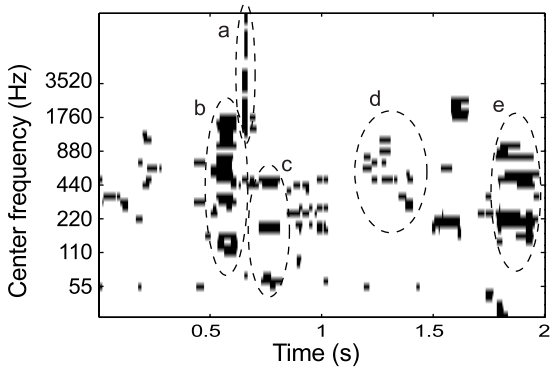

Figure 3: Object #2 from the middle (jazz) example above. It is seen (through hand-analysis) that many of the individual time segments correspond to individual perceptual objects; however, objects from different sources have been grouped together. (a) is a cymbal sound; (b) and (e) are piano chords; (c) is a bass note; and (d) is a run of piano notes.

It is unsurprising that the sequential grouping behavior is not accurate, since the method presented has essentially no way to make these judgments. Inclusion of the sort of feature-extraction capabilities necessary to correctly perform sequential integration is one of the tasks for future research.

Although only three examples have been presented here, validating the performance of music-signal-processing systems requires continuing attention to a variety of input signals. It is insufficient to claim good performance based on a few carefully-chosen tests – a convincing argument must be produced that the technique functions for *all* signals in the domain under investigation. This is an continuing goal of the present research project.

5. FUTURE WORK

There is a great deal more work that must be undertaken in order to demonstrate a robust understanding-without-separation system. It is to be emphasized that the present paper is only one component of a large project still in progress. Fundamental questions remain to be addressed with regard to both the engineering aspects of the system, and the further evaluation of the system as a model for the hearing process.

From an engineering perspective, immediate work is focused on improving sequential-integration aspects of the system. This will take two forms. This first is the construction of an improved model for associating the clusters in the prefeature space to auditory objects. This will include a "birth/death" model of dynamic changes to the number of clusters. Second, more attention to the features of auditory objects in this framework will lead to better models for knowing when two objects in the scene should be sequentially connected together. Top-down information may also play a role in this stage.

From a scientific perspective, it is a natural step to examine the application of this grouping model to the known data on frequency modulation detection, frequency-modulation based segregation, comodulation release from masking, and related phenomena. If the model could be used to concisely explain these data and make new testable predictions, then it could be viewed as a contribution toward a better model of the hearing process.

6. ACKNOWLEGEMENTS

The author is grateful to Dr. Keith Martin, formerly of the MIT Media Lab, for the use of the log-lag autocorrelogram software implementation as well as ongoing discussion and criticism.

7. REFERENCES

[1] L.-L. Balkwill and W. F. Thompson, "A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues," *Music Perception*, in press.

[2] E. D. Scheirer, "Bregman's chimerae: Music perception as auditory scene analysis," in *Proc. International Conference on Music Perception and Cognition*, Montreal, 1996, pp. 317-322.

[3] A. Bregman, *Auditory Scene Analysis*. Cambridge MA: MIT Press, 1990.

[4] K. D. Martin, E. D. Scheirer, and B. L. Vercoe, "Musical content analysis through models of audition," in *Proc. ACM Multimedia Workshop on Content-Based Processing of Music*, Bristol UK, 1998.

[5] D. P. W. Ellis, *Prediction-Driven Computational Auditory Scene Analysis*. Ph.D. Thesis, MIT Dept. of Electrical Engineering and Computer Science. Cambridge MA, 1996.

[6] R. Meddis and M. J. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," *Journal of the Acoustical Society of America*, vol. 89, no. 6, pp. 2866-2882, 1991.

[7] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory Physiology and Perception*, Y. Cazals, K. Horner, and L. Demany, Eds. Oxford: Pergamon Press, 1992, pp. 429-446.

[8] A. de Cheveigné, "Cancellation model of pitch perception," *Journal of the Acoustical Society of America*, vol. 103, no. 3, pp. 1261-1271, 1998.

[9] J. C. R. Licklider, "A duplex theory of pitch perception," *Experientia*, vol. 7, pp. 128-134, 1951.

[10] R. Meddis and M. J. Hewitt, "Modeling the identification of concurrent vowels with different fundamental frequencies," *Journal of the Acoustical Society of America*, vol. 91, no. 1, pp. 233-244, 1992.

[11] R. O. Duda, R. F. Lyon, and M. Slaney, "Correlograms and the separation of sounds," in *Proc. IEEE Asilomar Workshop*, Asilomar CA, 1990.

[12] E. D. Scheirer, "Sound scene segmentation by dynamic detection of correlogram comodulation," MIT Media Laboratory Perceptual Computing Technical Report #491, 1999.

[13] D. K. Mellinger, *Event Formation and Separation in Musical Sound*. Ph.D. Thesis, Stanford University Dept. of Computer Science. Palo Alto CA, 1991.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society (series B)*, vol. 39, no. 1, pp. 1-38, 1977.

[15] C. M. Bishop, *Neural networks for pattern recognition*. New York: Oxford University Press, 1995.