MIT Media Lab Perceptual Computing - Technical Report #224 (rev. 2)
Submitted to the 12th International Conference on Pattern Recognition, Jerusalem, October 1994

A Computer Implementation of Psychoacoustic Grouping Rules

Daniel PW Ellis
Perceptual Computing, MIT Media Lab E15-368C, Cambridge MA 02139
dpwe@media.mit.edu

We are building a functional computer model of sound organization and understanding in human listeners. In particular, we would like to be able to detect and locate events in the acoustic signal that will be perceived as separate objects. Our model aims to duplicate this function of the auditory system by the same means, although the level of correspondence is abstract and necessarily speculative given our current state of knowledge. None the less we believe that recent theories of auditory organization are sufficiently detailed to make possible computer modeling, and that such models can help develop and refine these theories.

Psychoacoustic experiments have demonstrated the importance of various 'cues' in the perception of sound as single or multiple objects [Breg90]. We describe an implementation of *grouping rules* corresponding to the cues of harmonicity, common onset, continuity and proximity. We implement these rules in a very direct fashion based upon our constant-Q sinewave representation of the sound [Elli92]. However, the rules alone are simple and prone to error; we greatly increase the robustness of the results by adding a *second* layer of grouping that looks for corroboration between groupings based on different rules. We believe that such a system of repeated hierarchic grouping is critical for the successful modeling of auditory functions.

Our preliminary results from applying the system to real sounds (isolated speech and a musical signal corrupted by impulsive noise) show an encouraging ability to identify and isolate the perceptually significant events. We hope that by extending the set of rules, and refining both the primary rules and the secondary grouping procedures, we will obtain a useful simulation of the organization and segregation performed by the human auditory system.

1. INTRODUCTION

This paper describes a computer model of part of the human auditory system, specifically the process by which disparate acoustic energy incident upon the ears is 'organized' into a small number of separately-identified real-world sound sources. This problem of *source separation* is analogous to the segmentation problem in vision: the raw information available to the perceptual system is the combination of several essentially independent sources; the most effective way to process this information (avoiding combinatorial explosion) is to isolate the contribution of each source and deal with it separately; therefore, the first stage of processing must be concerned with identifying and grouping sets of information by source.

When energy in disjoint frequency bands is perceived as arising from a single source, the separate bands are said to be *fused.* Consider for example the individual lower harmonics of a complex periodic tone: The listener usually hears only a single tone with pitch related to the common period of the harmonics, and 'quality' depending on the harmonics' amplitudes. The conditions required for such fusion have been extensively studied in the psychoacoustical laboratory, resulting in a set of empirical rules, reflecting real-world regularities, that predict how combinations of simple sine stimuli will be organized by listeners [Breg90]. These rules are of a form appropriate to further codification in a computer program. Signal processing of sound has typically been limited in its ability to cope with interfering mixtures because of the lack of a 'segmenting' front-end. This is particularly evident in speech recognizers that can only function when given clean, isolated input. Lately, several researchers have been investigating approaches based on auditory organization: [Cook91], [Brow92] and [Okun94] have focused on the problem of enhancing speech amid interference, whereas [Mell91], [Baum92] and [Kash93] have considered the related problem of separating the different melodic lines in polyphonic music.

The current project is motivated by a belief that one critical aspect of any successful model of human auditory grouping is the simultaneous use of a *range* of *different* cues. Moreover, the process of combining these cues is itself a central and unique facet of such systems, perhaps more important than the details of the individual cues themselves, making the auditory perception system a 'society' in the sense of [Mins86]. Therefore we set out to build a selection of simple grouping schemes in order to be able to experiment with the problem of combining their outputs.

While the system we describe has an essentially ad-hoc and problem-specific structure, it is worth noting its similarity,

both in outline and to some extent in detail, to the sound-understanding blackboard systems of [Carv92] and [Nawa92], an axis of abstraction we hope to investigate and develop in later work.

1.1 The motivation for modeling

Although it would certainly be useful to endow machines with the abilities to organize and interpret sound that are possessed by people, our primary motivation has been to understand the human prototype rather than solve particular problems in automatic processing of signals. We chose to pursue this by building a functional model. Independent of useful processing that may be performed by models, the implementation of theories of perception as computer programs has several benefits:-

- The theory can be verified to result in the behavior intended.
- Sometimes proposed perceptual processing structures can be very hard to analyze except by building a model.
- Models permit experimentation to resolve quantitative uncertainties.
- Building a model is a good way to check for logical gaps in a theory and helps keep the theorist 'honest'.
- The process of building the model provides a somewhat different perspective on both the task and the proposed mechanism, and can furnish insights resulting in new and/or improved theories.

Assertions of the correspondence between the value of a variable in a model and some *measurable* parameter in the real world can be confirmed by experiments. But while we can describe in general terms the intended function of our system to identify events and sources 'like a person would', the only concrete definitions we have for person-like behavior involve psychoacoustic detection tests. These involve a large amount of human machinery (such as the ability to understand and carry out instructions) that are not relevant to the limited scope of our model.

So although we would very much like to work within a more rigorous and defensible paradigm, we are at present limited to qualitative arguments for the validity of the models, and consequently weaker conclusions and implications. It is hoped that despite this, the insights and results of modeling will be found sufficiently interesting to warrant the attention they are given.

1.2 Structure of the paper

The next section gives an overview of the current project, and sections 3, 4 and 5 describe in more detail the representation, primary grouping and secondary grouping respectively. Section 6 introduces some preliminary results of applying the system to real sounds, and section 7 concludes with a brief discussion of our planned developments.

2. OVERVIEW OF THE SYSTEM

Figure 1 shows how the current work fits into the context of a complete machine audition system. At the bottom of the diagram, real-world events generate sound, which is analyzed into a time-frequency representation by a filterbank approximating the function of the cochlea. The output of the filterbank is represented as time-frequency contours (or *tracks*) lying along energy maxima, as described in section 3. This representation is then subject to grouping into objects: The two rounded boxes labeled "primary grouping" and "secondary grouping" correspond to the work described in this paper.

The first box, primary grouping, is made up of rules for identifying the basic psychoacoustic cues such as harmonicity, common onset and proximity, and is described in more detail in section 4. These rules each produce multiple groups of time-frequency tracks.

These groupings are called *part-objects* in the diagram to underline their preliminary nature. They are fed to the secondary grouping stage which performs somewhat different operations to produce more robust track groupings by integrating the results of the primary rules. These operations include looking for duplicate and (near) subsets among the groups and reducing these down to a smaller set of canonical groupings.

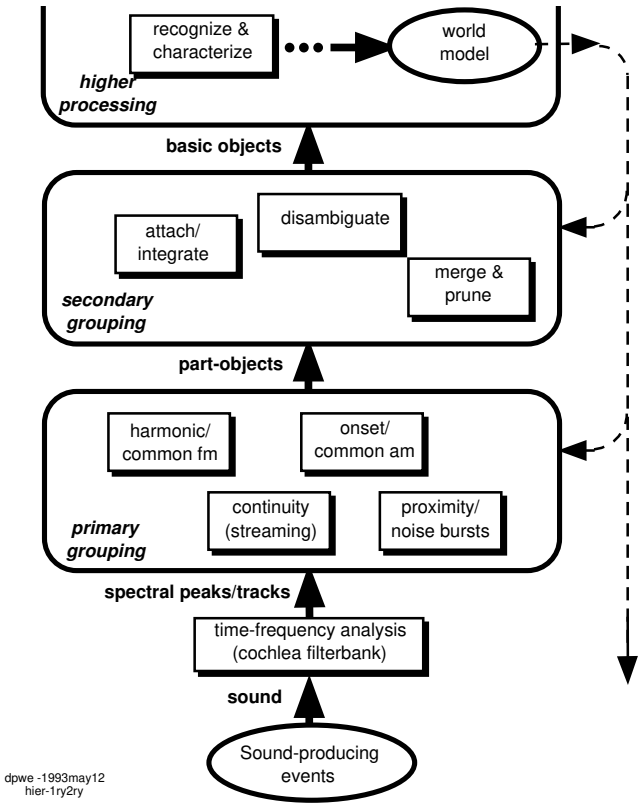

Figure 1: Block diagram of a 'complete' sound processing system, including the primary and secondary grouping stages described in this paper.

The pruned and non-redundant track groups produced by this second stage are the *basic objects* of the diagram, since they are very likely to be perceived as individual entities, yet they are still relatively simple in structure. In particular, they have been created only by consideration of general principles of the structure of real-world sounds, but without any knowledge or memories of sounds they might resemble or evoke. This kind of acquired or knowledge-based processing would be accomplished by the next layer, labeled "higher processing" in the diagram, and deliberately drawn as incomplete since we do not have strong ideas about how far upwards this block diagram should really extend. The dotted arrows pointing down from the "world model" indicate the possibility of 'goal-driven' processing that would guide searches to reveal such details based on expectations.

3. A REPRESENTATION FOR SOUND: THE CONSTANT-Q SINEWAVE MODEL

Figure 2 briefly introduces the underlying representation used for the processing, the Constant-Q Sinewave Model based upon the Sinusoid Transform of [McAu86] and described in more detail in [Elli92]. We review the process very swiftly here. The continuous variation in air pressure as a function of time is converted into a set of discrete contours or 'tracks' via three stages: first, the acoustic energy is calculated as a function of time and frequency by passing the signal through a filterbank. In our implementation this is an array of Gaussian constant-Q filters. At each instant (i.e. for each vertical slice through the time-frequency plane) the magnitude and frequency of the local maxima of the spectrum are recorded; these are matched between successive time frames to form tracks that extend across time. These tracks comprise a set of discrete pairs of frequency and magnitude functions that represent the original sound. Although this representation is not perfectly invertible, a reconstructed signal with high perceptual similarity to the original sound may be produced by a simple resynthesis process of using each contour pair as control inputs to a sinewave oscillator. This process is also effective for limited subsets of the tracks, which is useful for segregation. The key properties of the representation for the current application are that, excluding spectral collisions, each track represents energy of only a single source, and yet has meaningful attributes such as frequency modulation rate and magnitude variance.

Figure 3(a) shows the graphical format we use to display this bottom-level analysis, in this case for the spoken word 'spoil'. Time goes from left to right; this example is about 0.8 seconds long as indicated by the scale at the bottom. The top panel shows the envelope of the magnitude waveform. The lower panel has logarithmic frequency as its vertical axis and shows the energy output of the filterbank as shades of gray. We see the initial sibilant, stop and release, followed by a period of voicing. During the vowel, we see the first five or six harmonics separately resolved as in a narrowband spectrogram, but above this we begin to see individual pitch pulses as vertical stripes as in a wideband spectrogram.

The lines drawn over the filterbank output, particularly visible along the centers of the resolved harmonics, are the frequency contours of the tracks formed by the analysis described above. The noisy sibilant is represented as a number of short-duration, haphazardly scattered sinusoids with uncorrelated modulation.

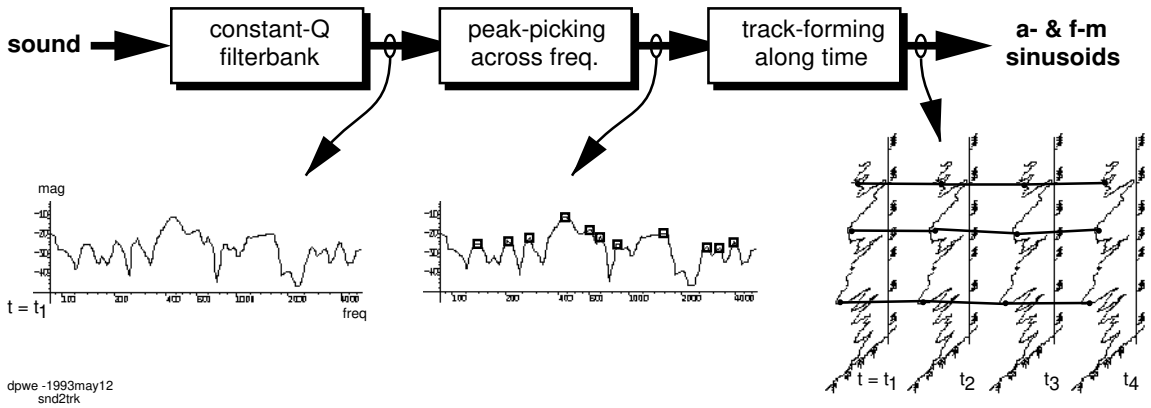

Figure 2: Sound is converted to a representation consisting of energy contours in time-frequency (or 'tracks') via three stages as illustrated.

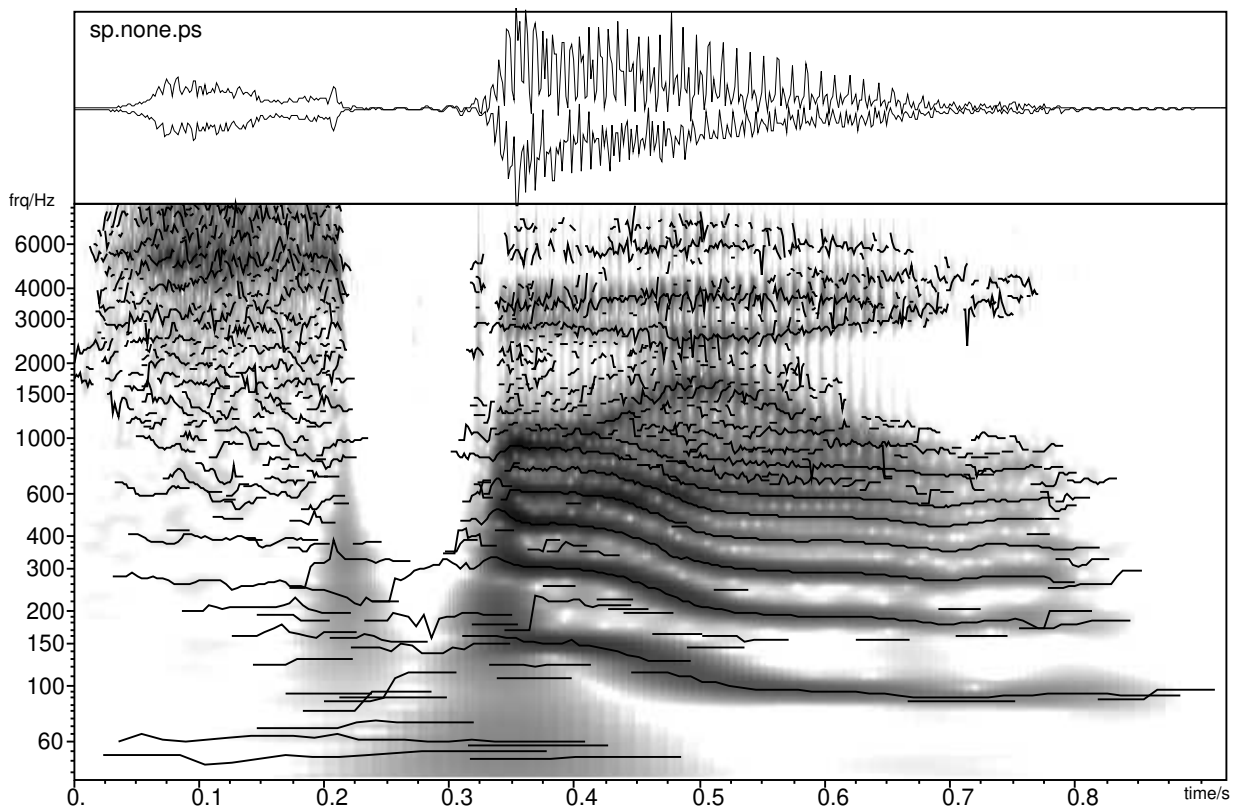

Figure 3(a): Representation of the word 'spoil'. The top panel shows the amplitude waveform, with time going from left to right. The lower panel shows the energy output of the filterbank as gray density. The frequency contours of the tracks analyzed from the sound are drawn over the energy display.

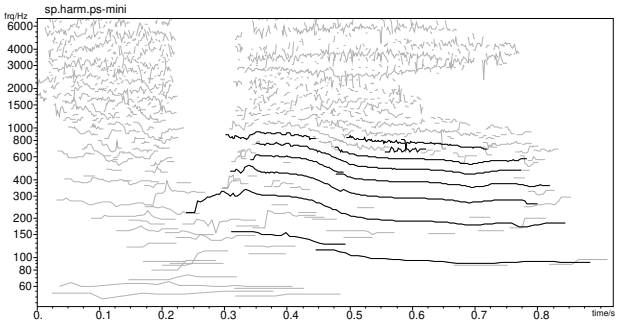

Figure 3(b): The tracks from the lower panel of 3(a). A group of tracks formed by the harmonicity rule is emphasized.

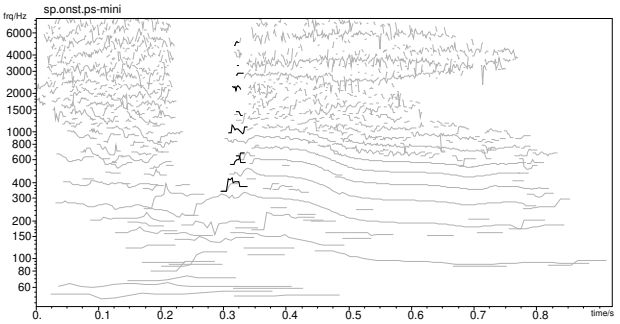

Figure 3(c): As 3(b), except the emphasized tracks have been grouped by the common onset rule.

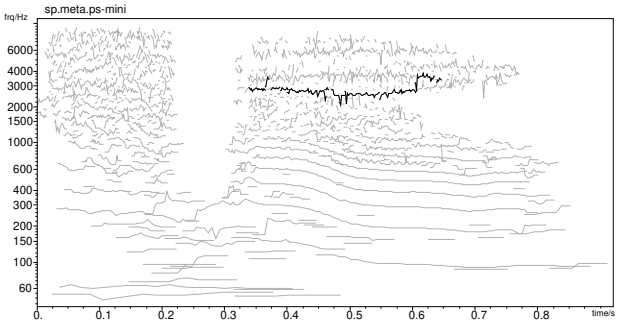

Figure 3(d): As 3(b), except the emphasized tracks have been grouped by the continuity rule.

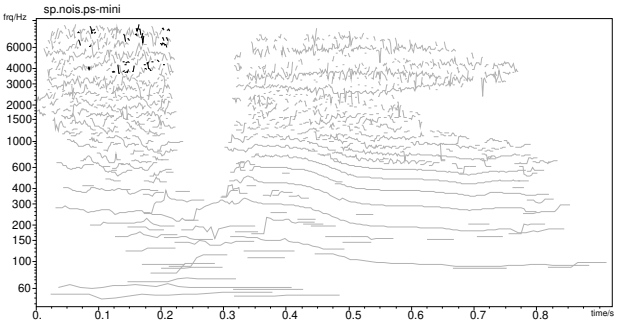

Figure 3(e): As 3(b), except the emphasized tracks have been grouped by the proximity (noise) rule.

4. PRIMARY GROUPING RULES

The first stage of grouping consists of more-or-less direct implementations of the cues known to us from psychoacoustics. The cues we consider are:-

- Harmonicity, i.e. separately-resolved harmonics sharing a common period;
- Common onset, i.e. bands of energy that start at the 'same' instant;
- Common amplitude modulation, i.e. different frequency bands where the energy increases or decreases at the same time;
- Continuity, the association between sequential energy bursts in a narrow frequency band; and
- Proximity, the association between energy bursts of similar general characteristics over a broad frequency band.

Each of these will be discussed in more detail below. For more or less arbitrary reasons, each rule 'method' takes a single track element as a 'seed'; it returns zero or more *groups* of track elements which it 'grew' from the seed.

Each rule is written to generate several answers when grouping is ambiguous, and if a particular group contains several seeds, that group is likely to be generated several times. Thus the overall strategy is to generate a large number of track groups which will, with high confidence, contain the desired groups among many others, and then prune poor-quality and duplicate groups at a later stage.

4.1 Harmonicity rule

The grouping of acoustic energy by harmonicity derives from the real-world observation that sound is often generated by near-periodic mechanisms — the oscillation of the human vocal folds being a particularly important example. A Fourier analysis, of the kind performed by the filterbank in our model or the mammalian cochlea, will represent a periodic signal by a series of distinct harmonics, at least for the region of the spectrum over which the analysis bandwidth is less than the spacing between the harmonics. These harmonics, which occur at integer multiples of the repetition frequency or *fundamental,* correspond to the Fourier series expansion of the periodic signal. The harmonicity rule finds regions of energy that show this pattern.

It is clear that the track representation is particularly suitable for this purpose, since each resolved harmonic will typically result in a single, isolated track. The harmonicity rule considers all tracks that overlap in time with the seed track, and that are close enough in frequency to possibly be one of the harmonics resolvable by our filter bank (the constant Q nature limits this; harmonics above the sixth or seventh will not be resolved).

For each candidate member of the harmonic group, the frequency ratio to the seed track is calculated for each sample point during their time overlap. In order to be added to the group, the ratio has to be close to integer, and close to constant over the track duration.

Figure 3(b) shows the result of applying the harmonic rule to one of the harmonics of the voiced speech of figure 3(a). In this case, the track corresponding to the fundamental is broken during a rapid frequency slide, but the second harmonic is complete and was the seed. The implementation is able to handle seeding by tracks above the fundamental because it repeats the search with the seed cast successively in the role of first through fourth harmonics, with or without a fundamental track.

4.2 Common onset rule

The 'ecological' basis for the common onset rule (i.e. the regularity of the real world that makes it useful) is that if a particular physical process generates energy in various frequency bands, it is likely that energy will start in each of those bands at the same moment. This would seem to be simple to detect among the tracks, however:

- There has to be some tolerance of asynchrony both to accommodate measurement errors, and because there may be slight frequency-dependent delays introduced by the acoustic environment.
- As the bandwidth of our constant-Q filters varies with frequency, so does their time-uncertainty; their fine time resolution at high frequencies gets progressively worse lower down. Consequently, our timing tolerance must increase as we descend in frequency. Also, two events which are clearly separate in time at high frequency may merge into a single feature at lower frequencies, giving rise to ambiguous groupings.
- The wide time support of our filters at low frequency makes the very beginning of a track a poor estimate of its actual onset, which would be better located some little way into the track. This is illustrated in figure 4, which shows magnitude (in dB) as a function of time (in milliseconds) for an example track.

We overcome this last problem by calculating a special 'onset time' from the slope of the magnitude contour of each track. For each point in track, the magnitude slope (in dB/second) is compared with the maximum slope so far along the track; The onset is located at the point where the slope falls below some threshold. This point is shown by the cross-hair in figure 4.

Onset groups are iteratively extended from a low frequency seed, so a single seed may give rise to a number of plausible onset groups. Figure 3(c) shows an example of an onset group.

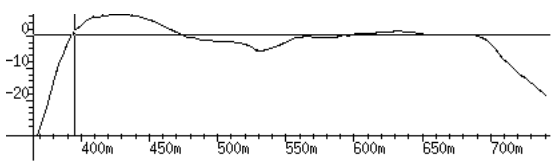

Figure 4: The magnitude (in dB) versus time (in milliseconds) for a particular track. The crosshairs show where the onset was placed, according to the magnitude slope.

4.3 Continuity rule

The rationale behind this rule is that acoustic energy in a certain frequency band resulting from a single source may undergo brief interruptions or periods of undetectability, after which it will reappear at close to the same frequency. When the 'gaps' are short in duration compared to the energy bursts on either side, it is desirable to make some kind of connection across the gap to indicate the similarity and closeness between the two parts.

This grouping of tracks across short time gaps may be necessary due to energy modulation in the source, or can be viewed as a pragmatic measure to 'clean up' the output of the track-formation stage, by compensating for occasions when that stage 'lost track' of particular energy regions; there can be an arbitrarily small difference between a signal that is analyzed as a single track with very deep amplitude modulation, and one that results in a large number of short tracks. To reflect the essential similarity of these two signals, we can organize the latter case into a continuity group (or *metatrack*) that in many ways behaves like the single, deeply modulated track.

The actual implementation is very straightforward; a time-frequency region just ahead of the end of the seed is searched for onsets of possible continuations. These are scored for time-frequency proximity, and any that pass some threshold are added to the growing group; if there are multiple candidate continuations, several alternative groups are spawned and grown. The continuation is terminated when energy of the tracks being added falls to some threshold below the average magnitude of the group. Figure 3(d) shows a group formed by the continuity rule, containing the energy from a vowel's formant.

4.4 Proximity (noise) rule

While harmonic complexes and resonant (formant) bursts are represented very successfully as sums of modulated sinusoids, such a transformation is less obviously appropriate for wideband, sustained 'noise' such as speech sibilants. These analyze into a large number of densely packed, uncorrelated tracks. We would like some method to group such tracks into the single perceptual element to which they correspond. A smoothed spectrum is obtained by convolving the spectral magnitude function at a particular instant with a Gaussian kernel; a candidate noise band is then defined as a range in this smoothed spectrum that lies within a magnitude threshold of its average value for at least a certain bandwidth. This search is then repeated at successive time instants; if highly-overlapped frequency bands are found at several adjacent times, a complete time and frequency range for the noise region is defined. Any tracks lying in this region of appropriate energy and duration are recruited to the corresponding noise group.

Figure 3(e) shows an example of a group formed by this rule.

4.5 Amplitude-modulation groups

The last major monaural psychoacoustic grouping cue is common amplitude modulation, the association of energy in different frequency bands that exhibits synchronized fluctuations. Such a cue would help in collecting the separate formant trains of a single voice (since they will all be amplitude-modulated by the pitch pulses), and, with some modification, could help attach the formants to the resolved harmonics. We have yet to implement such a rule. This should be natural and straightforward - in a manner similar to the harmonic grouping, pairs of tracks (or metatracks) could be scored for amplitude modulation correlation during their common time support.

5. SECONDARY GROUPING

We have now described in detail our implementation of the primary layer of group-formation, whereby the well-known psychoacoustic cues are used to generate a large number of groups of tracks showing various degrees of association according to the cues. The groups themselves are not yet useful for interpreting or explaining the sound; there are too many of them, there are many redundant groups, and some of them have only a low confidence score. To produce a small number of less redundant, high confidence groups, we apply a second stage of grouping which aims to prune and integrate the results of the primary rules.

5.1 Pruning

All the groups generated by a particular primary rule are sorted into one list, according to total energy. Then each group is compared to all the groups below it in the list; if a lower group is a proper subset of the larger group, or if the energy in their difference is less than some threshold, the smaller group is deemed a (near) subset which is adequately represented by the larger, and it is removed from further consideration. This typically effects a 10:1 or greater reduction in the number of track groups.

5.2 Correlation

Pruning is only applied within a particular rule since the existence of highly-overlapped groups *between* different rules is very important evidence of genuine coherence for that group. The next stage of processing searches for this coherence by looking for highly-overlapped pairs. If such a pair has energy in their common tracks above some threshold of their total energy, a new 'supergroup' is formed by merging their component tracks.

5.3 Closure

In order to form robust groups that are not harmonically based we need some way of combining the results of continuity and proximity (noise) rules. Currently we merge together all such rules that intersect with principal onset groups. Then all non-harmonic groups are ranked by total energy and pruned as described above; those with the highest energy are carried forward as higher-level objects.

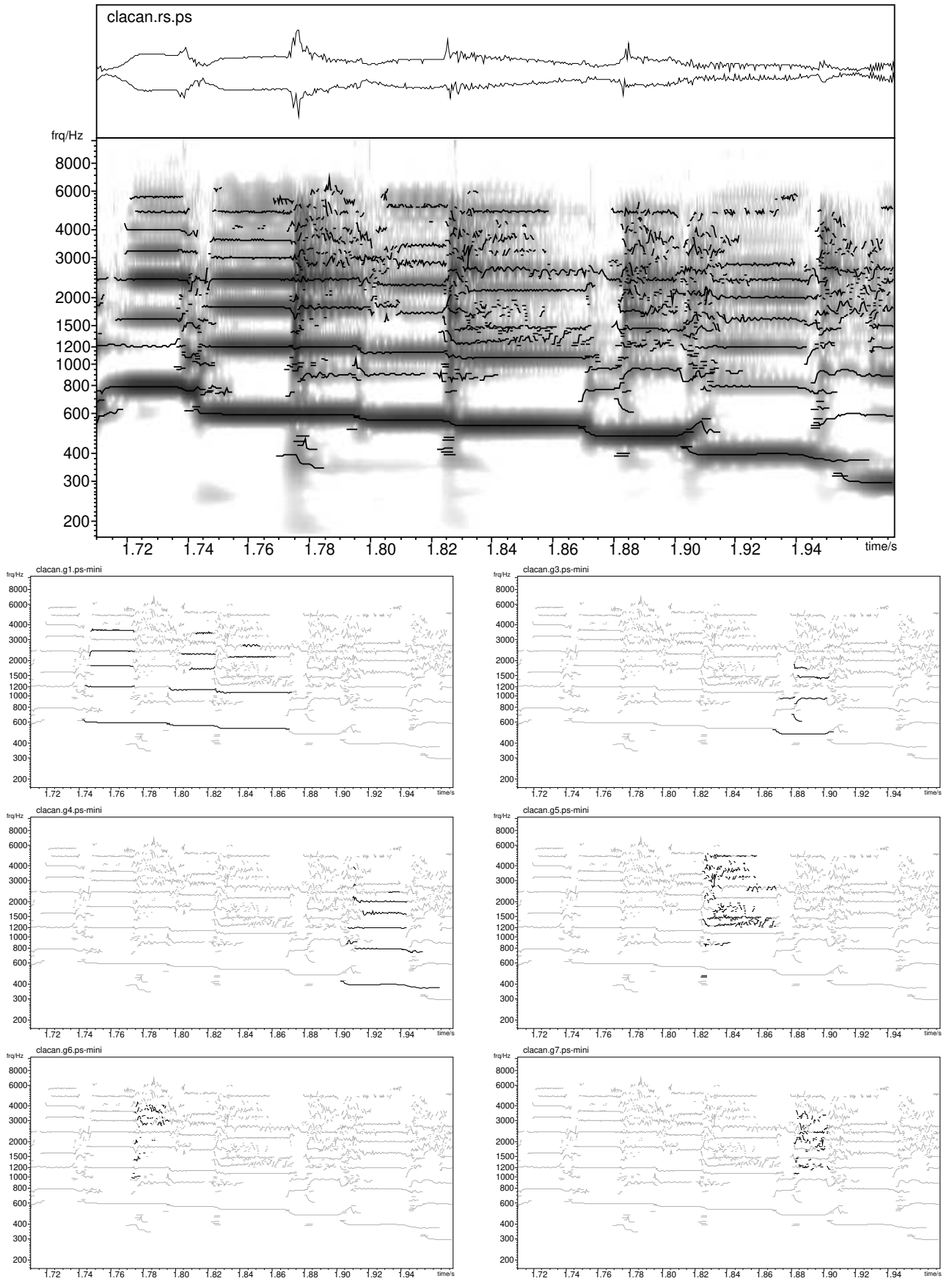

Figure 5: The top image shows the analysis of the original mixture, a clarinet melody with repeated intrusions of a can hitting a hard surface. The lower six images show the most important groups formed by the system.

6. RESULTS

The examples we have used in the explanations of the representation and rules are based upon a single spoken word. This is a curious example, since we are trying to model perceptual grouping but a single word normally constitutes a single perceptual entity, not four separate components as in our examples! While this is partially overcome by the secondary grouping, another example will better demonstrate the intended operation of the system. Figure 5 shows the track representation of a test signal composed of a solo clarinet melody to which the sound of a can being dropped on a hard surface has been added. In the picture, the straight, parallel harmonics of the clarinet can be seen in a descending staircase across the entire section; the very rapid transients from the can collisions appear as sudden vertical disturbances at t=1.775 s, 1.825 s, 1.885 s etc.

The lower panels in the figure show the most important groups found by the two-stage grouping scheme. The first few identify the harmonic clusters comprising different clarinet notes in the melody, which result in the correspondence between onset and harmonicity groups. Alongside these, however, we also see groups which have found the individual can impacts.

7. CONCLUSIONS AND FUTURE WORK

We have emphasized that this is a preliminary implementation, and there are many outstanding issues we are anxious to address.

At the lowest level, there are still questions about the adequacy of the underlying track representation. The precise nature of the filterbank response is highly idealized (compared to physiological data) and is essentially linear; addressing either of these could have far-reaching implications for subsequent processing.

We would also like eventually to provide for top-down processing, where the data provided by the early stages can be changed in response to higher level inferences and deductions. The subtraction of components in the case of an implicit spectral collision is a specific although difficult example in this area.

Considering the primary grouping rules, we have noted some shortcomings in the description above. Most pressing are the need for a common-AM grouping rule, and further refinements or innovations in a 'noise region' grouping rule.

For the secondary grouping stage, we anticipate developing several new methods in addition to those described to make the best use of the primary results. We are particularly interested in addressing the problem of ambiguity, where mutually incompatible interpretations are in competition for pieces of observed evidence. Unlike previous levels of analysis, the secondary grouping stage permits the explicit recognition and resolution of such conflicts. There are also opportunities to add processing on top of this layer, for instance to detect still larger-scale structures, perhaps with some kind of learning scheme.

In conclusion we find that the sinusoid track representation of sound provided for a very natural and straightforward implementation of grouping rules based upon the cues to source formation known from psychoacoustics. Having the results of these first-layer rules allowed us to experiment with secondary grouping methods, which led to reasonably high-level, robust structures identified in real sound examples. We are very hopeful that a refined and expanded set of rules will allow us to build a truly useful model of human auditory event formation.

ACKNOWLEDGMENTS

This work was generously supported by the MIT Media Laboratory, in particular the Television of Tomorrow consortium. The author is in the United States as a Harkness Fellow of the Commonwealth Fund of New York, whose support is gratefully acknowledged.

REFERENCES

[Baum92] U Baumann (1992) "Pitch and onset as cues for segregation of musical voices," presented to the 2nd Int'l Conf on Music Perception and Cognition, Los Angeles, Feb 1992

[Breg90] AS Bregman (1990) *Auditory Scene Analysis*, MIT Press

[Brow92] GJ Brown (1992) "Computational auditory scene analysis: A representational approach," PhD thesis CS-92-22, CS dept, Univ. of Sheffield

[Carv92] N Carver, V Lesser (1992) "Blackboard systems for knowledge-based signal understanding," in *Symbolic and knowledge-based signal processing*, ed. AV Oppenheim & SH Nawab, Prentice Hall

[Cook91] MP Cooke (1991) "Modeling auditory processing and organisation," PhD thesis, CS dept, Univ. of Sheffield

[Elli92] DPW Ellis, BL Vercoe (1992) "A perceptual representation of sound for auditory signal separation," presented to the 123rd mtg of the ASA, Salt Lake City, May 1992

[Kash93] K Kashino, H Tanaka (1993) "A sound source separation system with the ability of automatic tone modeling," Proc. Int. Comp. Music Conf, Tokyo

[McAu86] RJ McAulay, TF Quatieri (1986) "Speech analysis/synthesis based on a sinusoidal representation" IEEE Tr. ASSP 34

[Mell91] DK Mellinger (1991) "Event formation and separation in musical sound," PhD thesis, CCRMA, Stanford Univ.

[Mins86] M Minsky (1986) *The Society of Mind*, Simon and Schuster

[Nawa92] SH Nawab, V Lesser (1992) “Integrated signal processing and understanding of signals,” in *Symbolic and knowledge-based signal processing*, ed. AV Oppenheim & SH Nawab, Prentice Hall

[Okun94] HG Okuno, T Nakatani, T Kawabata (1994) “Auditory stream segregation in auditory scene analysis with multi-agent system,” submitted to AAAI-94