MIT Media Lab Perceptual Computing Group Technical Report #219
submitted to the 1993 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics

HIERARCHIC MODELS OF HEARING FOR SOUND SEPARATION AND RECONSTRUCTION

Daniel P W Ellis

Perceptual Computing Section
MIT Media Lab E15-368C
Cambridge MA 02139
dpwe@media.mit.edu

ABSTRACT

In building a machine to detect and segregate individual components in sound mixtures, the best example to copy is the human auditory system. Several models of auditory organization implement various rules of psychoacoustic *grouping* [Breg90]; we propose in addition to model auditory *inference* as exhibited in the well-known 'phonemic restoration illusion' of [Warr70]. A hierarchy of abstracted features and source hypotheses similar to [Nawa92] allows reconstruction of obliterated detail which can then be used to recreate an 'idealized' sound without corruption. A preliminary example of fitting a harmonic model to a noisy recording of a clarinet gives a very convincing resynthesis with the interference totally removed. However, there are many issues including the design of the representation and the control architecture still to be addressed in building a more general system.

1. INTRODUCTION

1.1 The source separation problem

We have been investigating the problem of automatically 'separating' a recording of a mixture of sounds to produce reconstructed sound comprising just one of the sources. Our prototype has been the human auditory system; to be useful in the real world of coincident events, our sense of hearing has evolved highly sophisticated techniques for identifying and characterizing the contributions of each component in a sound. To the extent that these techniques are known to us through psychoacoustical experiment and auditory physiology, we are building a computer model of this processing. This model must duplicate listeners' judgements of the number and duration of sources, and also construct output sounds that listeners can identify as the components of the mixture.

Any such system must include the known principles of auditory *grouping*, by which acoustic energy is 'fused' into single objects [Breg90]. A number of researchers have produced computer models that exploit important cues such as synchronized energy onset and harmonic frequency relations, with good results in the organization and separation of real sounds (notably [Cook91], [Mell91], [Brow92]). However, these systems lack a model of *inference* in auditory processing - using higher-level patterns to 'fill in' details not specifically detected by the periphery. As a result, any reconstructions will have characteristic energy 'holes' in time-frequency where cancellations between the original sounds prevented information extraction. Previous systems addressing this loss have been limited to specialized domains such as the method for separating voiced speech described in [Quat90] which reconstructed whole time frames by interpolating between neighbors.

In section two we explain our understanding of auditory inference with reference to the 'phonemic restoration illusion', and outline a corresponding computer model. A hierarchy of abstract representations, similar to the Sound Understanding Testbed of [Nawa92] both explains the illusion and provides a restoratory mechanism.

Resynthesis is a taxing ultimate goal for the system; while abstraction generally strips detail and categorizes many instances into the same model, we cannot discard too much information and still be able to regenerate something recognizable as the original. In section three we describe an example of such reconstruction, where the steady, harmonic sound of a clarinet is successfully extracted from a recording corrupted by the impulsive sound of a can hitting a hard surface, even though the two spectra have significant overlap.

Section four touches upon some other important issues not yet addressed, and details our immediate objectives in generalizing the system to process a range of possible sounds.

2. HIERARCHIC MODELS

2.1 Auditory illusion

An extreme example of the way human listeners 'fill-in' missing information is the 'auditory restoration' of deleted phonemes masked by noise bursts [Warr70]. This is preconscious processing, since the listener genuinely 'hears' the missing sound, and, indeed, has difficulty judging the timing of the noise burst within the speech. Further experiments indicate that the process is simple illusion rather than detailed subtraction in the masked region [Repp91].

The perceptual mechanism suggested by such phenomena is one of model-fitting: There is a certain finite set of sounds, represented as models, that the listener expects to hear; the stimuli (and the context) provide evidence upon which to select one model above the others, and its parameters are then estimated. The stimulus represented as the model plus its parameters lacks details such as precisely which evidence was employed in its selection, hence the 'illusion' of restoration.

A plausible mechanism by which the known rich diversity of perceivable sounds can be achieved by the finite model-fitting we propose is through a hierarchy of abstraction: rather than going directly from auditory nerve excitation patterns to a complex 'saxophone' model, there are successive layers of perceptual organization, recognizing features of incrementally increasing complexity and abstraction. Thus the similar but distinct 'saxophone' and 'clarinet' models can be relatively small, both making use of the outputs from feature-detectors such as 'clearly-pitched', 'steady-amplitude' and 'steady-pitch' from lower layers.

We note that any such hierarchy is not likely to be neatly structured or layered. In particular, feedback from later stages can be important for the efficiency of early processing [Nawa92].

2.2 An analogous computational model

Our proposed computational approach to sound analysis is shown as a block diagram in figure one. The primary time-varying spectrum of the sound is subjected to data-driven self-organization to label explicitly basic psychoacoustically-inspired features such as:

- stable frequency components i.e. well-defined slowly-moving spectral peaks, suggesting a periodic generating mechanism
- regions of locally uncorrelated 'noise' energy with some steady ensemble properties, such as average energy
- isolated energy bursts, limited in time and frequency.

These elements are then organized into primitive intrinsic structures of known perceptual salience, including:

- trains of bursts of similar frequency , directly comparable to the stable components above
- broad-band energy with synchronized onsets
- simultaneous frequency components spaced at multiples of a common fundamental frequency i.e. harmonic patterns.

The higher level stages of analysis are less clearly defined at this point. Essentially, they are hypotheses as to the origins of the sound, which will involve several further layers of more abstract features and top-level categorizations. The complexity of such a network will overwhelm bottom-up processes and require hypothesis-directed search as well. We have indicated with the thin and dotted arrows the influence of such top-down control on all the previous stages, for instance, to relax threshold constraints on track formation in the light of strong evidence from other harmonics.

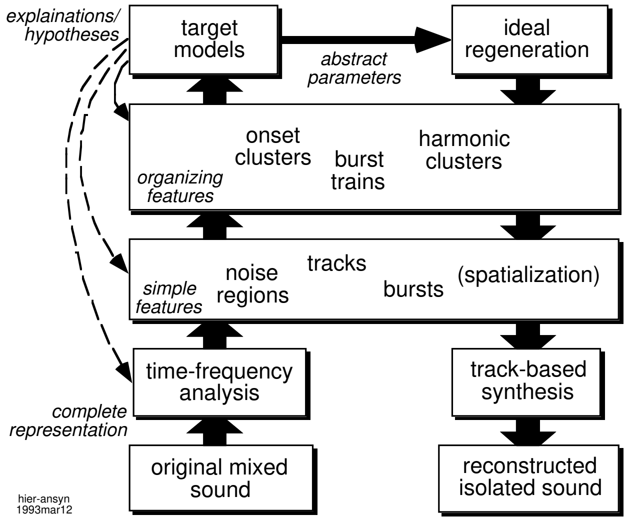

Figure 1: A hierarchic analysis and reconstruction of sound.

2.3 Reconstruction from a hierarchic analysis

The right-hand side of figure one deals with the converse problem of reconstructing the sound from the abstract analysis. Once a sound, or part of a sound, has been classified as an instance of a particular high-level model, reconstruction is a matter of estimating the parameters of that model based on the available evidence, then regenerating a representation back down the abstraction right through to a reconstructed sound.

Since abstraction (and perception) involves the projection of many different configurations onto fewer models, the inverse process will be underconstrained and require the 'invention' of extra parameters based on some ideal example. The difficult issue here is constructing a hierarchy of abstractions and parameterizations that discards only information not important to the perceived nature of the sound, so that the free choices in reconstruction do not alter that nature.

3. IMPLEMENTATION EXAMPLE

In this section we describe our implementation of the ideas outlined above, currently at an early stage of development. We include an example of the reconstruction of which we are currently capable, in which the system is able to completely remove interfering impulsive noise from a recording of clarinet while retaining the natural quality of the original instrument.

3.1 Front-end model

This system is built upon our constant-Q sine-wave model [Elli91], [Elli92], itself based upon the Sinusoid Transform system [McAu86]. The sine-wave model represents sound as the energy-maxima contours in the output of a constant-Q filterbank. Resynthesis by using each magnitude/frequency contour pair as control inputs to a sine-wave oscillator has very good perceptual identity with the original sound. This supports the contention that the 'track set' has a reasonable correspondence to an internal representation employed at some level in the auditory system.

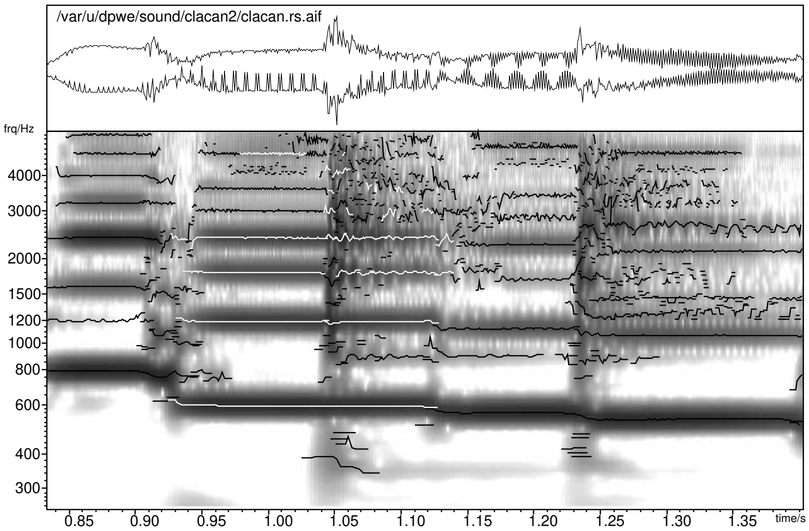

Figure 2: Waveform, scaleogram and track representation of corrupted clarinet. Can intrusions occur at t = 1.04s and 1.24s.

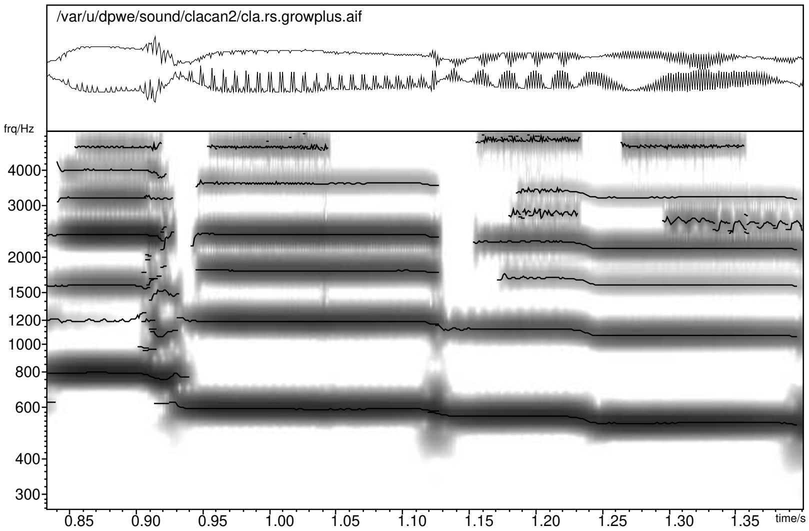

Figure 3: Waveform, scaleogram and generating tracks of reconstructed clarinet. The can has been completely removed.

3.2 Analysis

Figure two illustrates the analysis of a sound by this front-end. The sound is approximately half a second of clarinet corrupted with two bounces of a metal can hitting a hard surface. The can hits are visible as sharp bursts in the upper panel, which shows the time waveform going left to right.

The lower panel has frequency as its (exponential) vertical axis against the same timebase. The time-frequency distribution of energy (i.e. the filterbank output) is shown as the gray intensity; the distinct lower harmonics of the clarinet are very clear, as are the broad vertical bands of noise representing the can hits.

The lines drawn over this scaleogram are the tracks and bursts identified by the front-end model. The harmonics of the clarinet are well represented in the absence of interference, but ringing from the can obscures and distorts them.

The lines drawn in white are an example of one object found by our current harmonic-grouping algorithm. This works by generating a hypothesis of a harmonic sound based on each track above some magnitude threshold, then searching for support for each hypothesis among the other tracks. While it has successfully identified several components of the clarinet, it has omitted others, presumably because it employs only average frequency at present, and thus can be confounded by modulation.

3.3 Synthesis

Figure three shows the reconstruction of the clarinet. While the spectrum is clearly much reduced in detail, the resynthesized signal sounds remarkably natural, particularly in the context of more complete reconstruction of the uncorrupted notes before and after. Any hint of the can has been completely removed. The black lines drawn over the spectrum in this case are the reconstructed tracks used to *generate* the sound.

These tracks were produced in the following manner. The output of the bottom-up analysis was fed to a single high-level model searching for sets of harmonics with parallel amplitude and frequency modulation. This is our ‘clarinet’ model (although it works only for a portion of the spectrum of a rather small class of clarinet sounds!) This model parameterizes a clarinet note by a single frequency and amplitude contour pair, and a set of constant ratios describing the relationship of several harmonics to the characteristic contour. For the corrupted sound, the entire contour can be extracted from the lowest harmonic (the fundamental) since this escaped collision with the can sound. The ratios defining four other harmonics can be estimated by averaging over the time during which these tracks are clearly extracted. When these tracks disappear due to the interference, the model ‘finishes off’ the sound based on the contour and ratios. This is perhaps the most simple example of abstraction and resynthesis we could have concocted, but the results are surprisingly convincing and promising for broader extensions of this approach.

4. DISCUSSION

The general model introduced in section two raises a number of important issues not covered in the example above. In this section we briefly address some of the most significant remaining questions.

4.1 Designing the hierarchy

The nature of the results obtained with such a hierarchic analysis will depend critically on the components of the hierarchy itself - both the kinds of objects from which representations are assembled at a given level, and the rules by which information is abstracted between levels. So far all of these choices have been quite empirical: the representational elements are based on a speculative model of audition, and the rules of analysis are very heuristic. Although this is an uncomfortable situation, the alternatives are few at present. Ideally, we would like the system to acquire or ‘learn’ models, but this is a distant goal.

4.2 Resynthesis of obscured ‘typical’ features

When an identified sound contains, in its abstract representation, a feature such as a broadband noise burst that has otherwise been completely obscured, the problem of reconstruction can become very difficult. While the absence of acoustic evidence might mean that to a listener any ‘plausible’ example will do (since we assume the listener too will have had the feature obscured), the knowledge necessary to construct such a plausible example from scratch can be considerable. In the clarinet example, if *all* the harmonics had been lost for some period, the resynthesis might have attempted to reconstruct the underlying contour by reconnecting two parts; to do this convincingly would require at the very least some model of the short-term fluctuations (jitter) of the contour to avoid an artificially pure signal. At present our system has no such detailed knowledge, which must be obtained by measuring real sounds.

4.3 Knowledge-based signal processing and control flow

While a deep, parameterized hierarchy provides for the representation of a wide range of signals, it simultaneously presents practical difficulties in deciding among this large space of states. This point is underlined in the description of the Sound Understanding Testbed [Nawa92], which employs an advanced ‘blackboard’ architecture [Carv92] to process all layers of a hierarchy in a uniform and efficient manner. This is extended to the control of the numerical parts of the system: expensive calculations (such as long Fourier transforms) are only performed when ‘demanded’ by some hypothesis generated by other data (a hallmark of ‘knowledge-based’ signal processing systems). Clearly such considerations are crucial for systems able to deal with anything like the diversity encountered in the real world.

4.4 Conclusions and future work

We have presented our overview of some important aspects of human hearing, and what we believe is the cognitive machinery behind them. Our preliminary experiments with a computer simulation help to clarify the ideas, and have given surprisingly successful results for the separation of a musical instrument corrupted by impulsive noise. This demonstrates the feasibility of sound analysis, organization and reconstruction based on a hierarchy of parameterized abstractions. We have also mentioned briefly some of the important issues remaining to be addressed, such as designing the abstractions and the flow of control in realistically-sized systems.

Our immediate goal is increase the range of features understood by our system to enrichen the range of sounds it can represent. We also plan to integrate all the stages into a unified architecture to facilitate flexibility in operation and experimentation.

ACKNOWLEDGEMENTS

This work was carried out in the Music Cognition group of the Perceptual Computing Section of the MIT Media Lab. The support of this group by the Television of Tomorrow Consortium is gratefully acknowledged. Thanks also to Professor Barry Vercoe and the other group members for their continuing support.

REFERENCES

[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) “Modelling auditory processing and organisation,” PhD thesis, CS dept, Univ, of Sheffield

[Elli91] DPW Ellis, BL Vercoe, TF Quatieri (1991) “A perceptual representation of audio for co-channel source separation,” IEEE Workshop on Apps. of Sig. Proc. to Audio and Acous.

[Elli92] DPW Ellis (1992) “A perceptual representation of audio,” MS thesis, EECS dept, MIT

[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.

[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

[Quat90] TF Quatieri, RG Danisewicz (1990) “An approach to co-channel talker interference suppression using a sinusoidal model for speech,” IEEE Tr. ASSP 38(1)

[Repp91] BH Repp (1991) “Perceptual restoration of a ‘missing’ speech sound: Auditory induction or illusion?” Haskins Lab. Status Rpt. on Sp. Rsrch. SR-107/108

[Warr70] RM Warren (1970) “Perceptual restoration of missing speech sounds,” Science 167