UNDERCONSTRAINED STOCHASTIC REPRESENTATIONS FOR TOP-DOWN COMPUTATIONAL AUDITORY SCENE ANALYSIS

Daniel P. W. Ellis

Machine Listening Group
MIT Media Lab Perceptual Computing
20 Ames St, Cambridge MA 02139
dpwe@media.mit.edu

ABSTRACT

I propose a structure for the first stage of a computer system capable of performing complex auditory scene analysis similar to that accomplished by human listeners. This structure contains the following innovations over previous approaches: (1) Sound is represented as discrete elements drawn from an overcomplete vocabulary encompassing both tonal and less structured sounds, designed to highlight the interdependence in the acoustic energy. (2) Through the redundancy of the basis this analysis permits and indeed requires the imposition of additional constraints, which provides for the incorporation of top-down or context-sensitive factors. (3) A modular architecture operates on an analysis-by-synthesis principle, where processes are invoked until the representation adequately accounts for the observed sound. A common goodness-of-fit criterion allows for future expansion of the system with new explanation rules, new representational elements and more abstract levels of analysis. Some initial results of applying these ideas to scenes consisting of noise bursts and dense environmental sound are presented.

1. INTRODUCTION

1.1 Computational Auditory Scene Analysis

When people listen to sound, they are able to convert a dense combination of pressure waveforms generated by different mechanisms into an abstract conception of events in the external world. This is a remarkable accomplishment, since at first sight many acoustic situations – such as the proverbial cocktail-party – would seem to offer sound waveforms that were hopelessly confusing. We would like to build computer systems to reproduce this ability, not only because of the applications in prostheses and automatic annotation, but also because to discover how the perceptual system is able to extract useful information from such daunting evidence would probably furnish techniques of value in many other domains.

The next section considers existing approaches to this problem of computational auditory scene analysis (CASA), and highlights some of their shortcomings. Section 2 proposes the architecture of expectation-driven analysis to an overcomplete vocabulary. Section 3 describes some implementation results, and section 4 concludes with a discussion of outstanding issues and future directions for this work.

1.2 Shortcomings of CASA systems

Since Bregman published his unifying account of psychological results in auditory organization, *Auditory Scene Analysis* [1], there has been a series computational models of these principles. The dominant approach, as embodied in the dissertations of Cooke [2], Mellinger [3] and Brown [4], and elsewhere [5], may be characterized as follows: First the sound is processed by a conventional signal-processing module intended to simulate early auditory processing. This transformation is then converted into discrete elements, and through a search process these elements are organized into different groups corresponding to different 'sources' believed to exist in the sound. Within this outline, the models vary considerably in their level of physiological faithfulness, the kinds of sounds they are intended to process, the principles and strategies used for grouping, and the form of final output.

However, none of these systems is entirely satisfactory as a model of auditory analysis. Bregman has recently written a critique of such systems [6] which emphasizes the following shortcomings of current models in addressing crucial aspects of the auditory system:

- **Context insensitivity:** The data-driven, bottom-up structure of these systems gives them little ability to adapt their processing based on context beyond a very local scale. Yet ample psychoacoustic evidence shows that context is central to auditory analysis. Slaney has also made a thorough argument on the importance of this point [7].
- **Closed architecture:** Bregman bemoans the brittle reliance on a narrow range of cues. He advocates systems easily expanded to include new cues, and processing strategies that select the most applicable cues in each situation.
- **No room for ambiguity:** Considering the well-known 'duplex-perception' phenomena, where a single acoustic element can contribute to the perception of more than one auditory object, he questions the wisdom of building systems that produce a single, black-and-white interpretation of a scene.

The approach in this paper is motivated by these problems.

2. AN ARCHITECTURE FOR CASA

2.1 Discrete vocabulary elements

In common with the systems mentioned above, my starting point is the notion that the auditory scene analysis problem is best addressed by an initial re-representation of the continuous acoustic energy as a collection of discrete 'units'. The value of a representation lies in its ability to make explicit those aspects of the data relevant to the problem at hand, in this case the independent acoustic sources. Thus the best representation will consist of distinct elements, preferably containing *all* the energy from a single source, or the largest subset as can be practically identified without the risk of including energy from other sources. This suggests a representational vocabulary determined by the important kinds of structure in real sounds, e.g.:

- Pseudo-periodic signals, whose common origin is best expressed if energy across the whole spectrum that reflects the same periodicity is integrated into a single element. We have proposed a form for these elements called *wefts* [8] that are derived from correlograms [9,10].
- Transients – energy with a rapid onset and short duration – are a very significant part of our world of hard surfaces and other discontinuities, and are particularly prominent in speech.
- Noise 'clouds,' regions of disordered energy without clear boundaries. These may be adequately represented by a smooth time-frequency envelope without recording the fine structure.

This list might be expanded or modified, but the point is to define the elements of representation according to a theory of the cues and structure employed by the low-level auditory analysis to separate acoustic objects, which in turn reflect the operative physical regularities. Given such elements, our computational auditory scene analysis problem reduces to a search in this representational space.

We immediately notice the major difference of this space compared to, say, that defined by sinusoidal elements : it is highly redundant. If a given signal may be represented as a large noise cloud, it may be equally accurately represented as the overlap of several clouds.

2.2 Reconciliation with observed signal

A second major difference is the statistical nature of this space. Since portions of the input may be represented as the result of noise processes characterized only by their general properties, the match between the observed energy in given time-frequency cells and the predictions of an element is only probabilistic. This is in contrast to sinusoidal modeling, where the mean-squared error between observed and represented energy could be very small.

In order to compare the representation with the observed signal, the appropriate goodness-of-fit metric is to derive from the noise elements the expected signal energy and variance for each time-frequency cell. This permits calculation of the likelihood that a given representation corresponds to an observed signal (i.e. the variance-weighted squared-error or Mahalanobis distance).

Elements that overlap in time-frequency can be combined to produce a composite surface according to the principles of stochastic signals. Non-noise elements simply have a smaller variance, although interference between sinusoids of unknown phase can be accommodated in this way too, as perhaps can psychoacoustic masking.

Statistical reconciliation to the input forms the unifying concept that permits the combination of diverse elements into a single architecture.

2.3 Data-driven analysis

Having defined a representational basis and a goodness-of-fit metric, we can consider how to analyze a given input signal. Since the representation is underconstrained, there will be no direct method or unique answer, as the ambiguity of the domain demands. Rather, the influence of top-down context-derived expectations can resolve the choices at the bottom level.

In the absence of specific high-level guidance, the process is as illustrated in figure 1: The input sound is analyzed into a time-frequency energy surface via an approximate cochlea filterbank model. This energy surface then feeds the comparator unit, whose other input is the expected value and variance in each time-frequency cell derived from the current representational elements. The comparison identifies time-frequency locations where the representation is inadequate to explain the observations, and these areas form the focus for modifications. This style of analysis is reminiscent of the 'residue-driven architecture' [11].

Each kind of element has its own 'generate' method, whose inputs are the input energy to be explained along with the upper bound on energy to be added in each cell. Each method will inspect these surfaces and create an object that best accounts for the energy if it can.

2.4 Top-down effects

If analysis of the sound to date has resulted in certain expectations, this information can readily be incorporated into the element-generation stage; rather than starting from scratch to create an element, the analysis can first attempt to match one or more of the predicted elements with the new observations. Extension of existing elements to incorporate new energy also falls into this category.

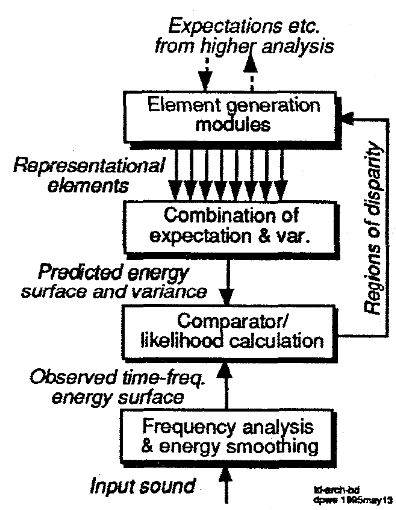

Figure 1: The analysis loop.

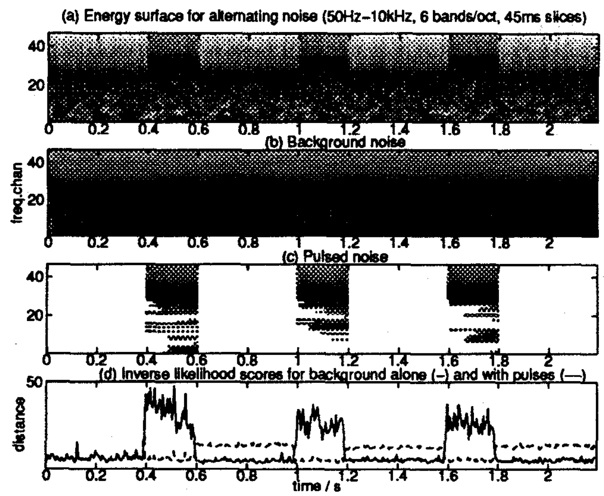

Figure 2: Analysis of alternating noise bursts.

A working system can be constructed even without this aspect of the architecture fully developed. Rather, it is critical to leave a hook of this kind unoccupied to permit the future enhancement of that system with different and more sophisticated levels of analysis.

2.5 Resynthesis

The scheme described is essentially analysis-by-synthesis – arriving at model parameters through an indirect scheme, then checking and refining them by calculating the input they predict. This is easily modified to generate an actual waveform, and thus the simple and useful tool of partial resynthesis is available for this architecture, with attendant benefits for assessment and applications.

3. RESULTS

Figures 2 and 3 present the results of some preliminary implementations of these ideas. Figure 2 deals with a particular stimulus used by Bregman [6] as an example of the importance of context sensitivity – particularly the old-plus-new heuristic. It consists of a 400 ms burst of noise low-pass filtered at 1 kHz alternating with 200 ms of noise low-pass filtered at 2 kHz. Below 1 kHz, the spectra of the two components matches, and since there is no gap between them, the auditory system is able to interpret the sequence as continuous low-frequency noise with pulses of bandpass (1-2 kHz) noise every 600 ms. Panel 2(a) shows the energy surface for this stimulus as processed by the seven octave constant-Q filterbank used as a front-end.

The analysis of this signal hypothesizes a noise-cloud to account for the unstructured noise at the beginning of the signal, as shown in panel 2(b). At the same time, the likelihood that each successive time frame was indeed generated by the noise cloud is calculated. This continues successfully until the higher noise band appears at 400 ms, at which point the inverse likelihood score (the solid line in panel 2(d)) suddenly becomes very large.

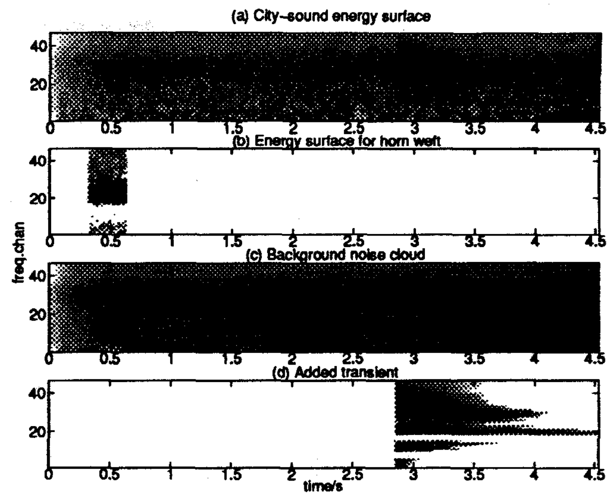

Figure 3: Analysis of the "city-sound".

Following perceptual principles, the system prefers to retain its existing explanation if possible, and accounts for the change by introducing a second noise burst composed of input energy significantly in excess of that accounted for by the noise cloud. The energy surface of this burst is shown in panel 2(c); it is mostly concentrated in the 1-2 kHz band, although it also picks up a little energy from lower frequencies. The likelihood score based on the combination of these two elements shows them to have made a successful explanation of the input, as seen from the dotted line in panel 2(d).

By comparing the likelihood scores for explaining the input with and without the burst, the system notices that at 600 ms the sound is once again best explained as background noise alone. The burst hypothesis remains, however, forming a prior expectation that such an element may occur, and allowing continued comparison of the two hypotheses in panel 2(d). When the burst re-enters at 1.0 s, the system immediately reinstates the second element. This cycle can then repeat.

Figure 3 is a rather more complex sound comprising a short extract of 'city street ambience' from a sound effects library. This sound has a rich texture from which more distinct events emerge. Panel 3(a) shows the overall energy surface which is rather featureless.

Panel 3(b) shows a weft (wide-band periodic) element extracted as described in [8], corresponding to a car horn at 0.4 s. The remainder of the input energy is incorporated into the background noise shown in panel 3(c), a recursive estimate of the energy at each frequency. This succeeds in explaining the input until a rather abrupt increase in energy in the high frequencies at t=2.85 s. The background cannot adapt to this, but the addition of a transient element can account for the excess energy, which away over the next second or so. Thus the dense sound scene is explained as a combination of periodic, transient and steady noise elements, moving towards the perceptual experience.

The sound examples corresponding to each of these components can be heard over the World-Wide Web by visiting:
`http://sound.media.mit.edu/~dpwe/waspaa95.html`.

4. DISCUSSION

There are many issues related to this architecture not addressed by the examples, including:

Blackboards: Numerous aspects of the problem – its abductive-inferential nature, the need for competing hypotheses, the desirability of a modular, extensible system – make it an obvious application for a blackboard architecture [12]. This is considerably eased by the use of an application framework toolkit for the IPUS architecture [13].

Fine structure: There are many perceptually distinct sounds that match at the level of a smoothed time-frequency energy envelope, including voice. The weft representation seeks to address this using the extra lag dimension of the correlogram. A better model of auditory processing would require a more exacting comparison between input and representation while retaining the basic architecture.

Cue detection: A more direct method to create the transient element in the city-sound example would be to have onset-detectors that cause the expectation of transient elements. This kind of bottom-up cue detection can exploit the same architectural features that permit top-down adaptability in the system.

Plausibility: The basic motivation for this work is to find an appropriate way to emulate perceptual phenomena such as context-sensitivity and ambiguity: Ultimately, an ability to reproduce these effects parsimoniously would validate the approach. While it is difficult to argue the correspondence between these kinds of algorithms and circuits built of neurons, the former may still establish the basic principles.

4.1 Future Work and Conclusions

All aspects of this system need development. The vocabulary elements can be refined and expanded. The operation of the analysis system is rather unstructured at present; hopefully, more detailed principles will emerge.

A system that analyzed sound to a higher level of abstraction would be able to make more interesting predictions and a more convincing demonstration of the advantages of underconstrained choice. Evidently, the kinds of abstract analyses built on top of such a system are unlimited in scope.

In conclusion, this architecture has the potential to exhibit the same flexibility as the human prototype, the capacity to handle the full range of sounds that exist in the real world, and the expandability that must be a priority in this kind of research. Hopefully future developments to exploit that expandability will ultimately arrive at a useful system for analyzing complex sound mixtures.

ACKNOWLEDGMENTS

Thanks to Bill Gardner, Barry Vercoe and the Machine Listening Group for their support. Thanks also to Malcolm Slaney, Tom Ngo and Interval Research Corp. for being stimulating hosts during the summer of 1994. And thanks to Dick Duda for his patient interest in this work.

REFERENCES

[1] A. S. Bregman. *Auditory Scene Analysis: The perceptual organization of sound.* MIT Press, Cambridge, MA, 1990.

[2] M. P. Cooke. "Modeling auditory processing and organisation," Ph.D. thesis, CS Dept., Univ. of Sheffield, 1991.

[3] D. K. Mellinger. "Event Formation and Separation in Musical Sound," Ph.D. thesis, CCRMA, Stanford Univ., 1991.

[4] G. J. Brown. "Computational Auditory Scene Analysis: A Representational Approach," Ph.D. thesis CS-92-22, CS Dept., Univ. of Sheffield, 1992.

[5] D. P. W. Ellis. "A Computer Model of Psychoacoustic Grouping Rules," Proc. 12th Intl. Conf. on Pattern Recog., Jerusalem, October 1994. ftp://sound.media.mit.edu/pub/Papers/dpwe-icpr94.ps.gz

[6] A. S. Bregman. "Psychological Data and Computational ASA," in working notes for the workshop on Comp. Aud. Scene Analysis at the Intl. Joint Conf. on Artif. Intell., Montreal, August 1995.

[7] M. Slaney. "A Critique of Pure Audition," in working notes for the workshop on Comp. Aud. Scene Analysis at the Intl. Joint Conf. on Artif. Intell., Montreal, August 1995. ftp://ftp.interval.com/pub/papers/malcolm/PureAudition.psc.Z

[8] D. P. W. Ellis and D. F. Rosenthal. "Mid-Level Representations for Computational Auditory Scene Analysis," in working notes for the workshop on Comp. Aud. Scene Analysis at the Intl. Joint Conf. on Artif. Intell., Montreal, August 1995. ftp://sound.media.mit.edu/pub/Papers/dpwe-ijcai95.ps.gz

[9] R. O. Duda, R. F. Lyon and M. Slaney. "Correlograms and the Separation of Sounds," Proc. IEEE Conf. on Sigs., Sys., and Computers, Asilomar, 1990.

[10] M. Slaney and R. F. Lyon. "On the importance of time – a temporal representation of sound," in *Visual Representations of Speech Sounds*, ed. M. Cooke, S. Beet, M. Crawford, Wiley, 1993.

[11] T. Nakatani, T. Kawabata and H. G. Okuno. "A computational model of sound stream segregation with multi-agent paradigm," Proc. Intl. Conf. on Acous., Speech and Sig. Proc., Detroit, May 1995.

[12] S. H. Nawab and V. Lesser. "Integrated Processing and Understanding of Signals," in *Symbolic and Knowledge-Based Signal Processing*, ed. A. V. Oppenheim and S. H. Nawab, Prentice-Hall, Englewood Cliffs, NJ, 1992.

[13] J. M. Winograd and S. H. Nawab. "A C++ Software Environment for the Development of Embedded Signal Processing Systems," Proc. Intl. Conf. on Acous., Speech and Sig. Proc., Detroit, May 1995. ftp://engc.bu.edu/pub/kbsp/ICP/icp.icassp.ps.Z