to be presented at the ESCA workshop on the Auditory Basis of Speech Perception, Keele UK, July 1996

PREDICTION-DRIVEN COMPUTATIONAL AUDITORY SCENE ANALYSIS FOR DENSE SOUND MIXTURES

Daniel P. W. Ellis
email: dpwe@icsi.berkeley.edu
International Computer Science Institute
Berkeley CA 94704 U.S.A.

ABSTRACT

We interpret the sound reaching our ears as the combined effect of independent, sound-producing entities in the external world; hearing would have limited usefulness if were defeated by overlapping sounds. Computer systems that are to interpret real-world sounds – for speech recognition or for multimedia indexing – must similarly interpret complex mixtures. However, existing functional models of audition employ only *data-driven* processing incapable of making context-dependent inferences in the face of interference. We propose a *prediction-driven* approach to this problem, raising numerous issues including the need to represent any kind of sound, and to handle multiple competing hypotheses. Results from an implementation of this approach illustrate its ability to analyze complex, ambient sound scenes that would confound previous systems.

1. INTRODUCTION

The majority of real-world listening situations involve multiple sources of sound that overlap in time; one of the most puzzling questions in the study of hearing is how the brain manages to disentangle such mixtures into the perception of distinct external events. We would like to be able to build computer systems capable of interpreting these kinds of sound environments. As human listeners, we are so successful at organizing composite sound fields that we are tempted to overlook the significance and sophistication of the separation process, focusing on the possibly more tractable problem of how the sound of a single, isolated sound is recognized or interpreted. However, real-world applications will require in addition a solution to the problem of sound organization.

Computer modeling of the process by which people convert continuous sound into distinct, interpreted abstractions (such as the words of a particular speaker, or some inference about a remote, sound-producing event) is termed 'computational auditory scene analysis' (CASA). This title acknowledges that the work is founded on experimental and theoretical results from psychoacoustics, such as are described in Bregman's book *Auditory Scene Analysis* [1]. Several important projects in this area have focused on the problem of separating speech from interfering noise (either unwanted speech [2] or more general interference [3,4]). These approaches may be characterized as *data-driven*, that is, relying exclusively on locally-defined features present in the input data to create an output through a sequence of transformations. An equivalent term for this kind of system is *bottom-up*, meaning that information flows monotonically from low-level, concrete representations to successively sophisticated abstractions. A typical example is illustrated in figure 1, a block diagram description of the system presented in [4].

The assumption underlying this kind of processing is that the auditory system isolates individual sounds by applying sophisticated signal processing, resulting in the *emergence* of integrated sound events, grouped by one or more intrinsic cues such as common onset or periodicity. While this may be a good characterization of the processes at work in simplified demonstrations of phenomena such as fusion, there exists a wide class of perceptual phenomena which cannot be explained by bottom-up processing alone: auditory 'illusions', in which the perceived content of the sound is in some sense incorrect or different from what was actually presented. The implication is that in such situations the result of perceptual organization has been additionally influenced by high-level biases based on the wider context of the stimulus or other information; this kind of behavior is often termed *top-down* processing, and to the extent that it is central to real audition we must strive to include it in our computational models. (The case for top-down auditory models is powerfully presented in [5]).

One well-known instance of such an effect is the *continuity illusion* discussed in [1], and also used as the motivation for a specific refinement of Brown's CASA system in [6]. A simple version of this illusion is illustrated in figure 2, the spectrogram of an example from a set of auditory demonstrations [7]. The sound

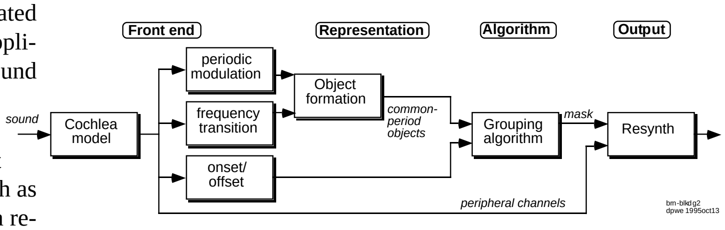

Figure 1: The data-driven computational auditory scene analysis system of [4]. Cues in the input sound are detected by the front end and used to create a representation in terms of 'objects' – patches of time-frequency with consistent characteristics. Some of these are grouped into a mask indicating the energy of the target (voiced speech) based on their common underlying periodicity. The target is resynthesized by filtering the original mixture according to this mask.

consists of a short sequences of 130 ms of sine tone alternating with 130 ms of noise centered on the same frequency. When the noise energy is low relative to the tone, the sequence is heard as an alternation between the two. But if the sine energy is decreased, the perception changes to a steady, continuous sine tone to which short noise bursts have been *added*, rather than hearing the noise burst as *replacing* the sine tone. The excitation due to the noise in the portions of the inner ear responding to the sine tone has become sufficiently high that it is impossible for the brain to say whether or not the tone continues during the noise burst. But instead of reporting that the sine tone is absent because no direct evidence of its energy can be observed, the auditory system effectively takes account of contextual factors; in the situation illustrated, the fact that the sine tone is visible on both sides of the noise burst leads to the inference that it was probably present all the time, even though its presence could not be confirmed directly during the noise. (The precise conditions under which the continuity illusion will occur have been studied extensively and are described in [1]). An interesting point to note is that the listener truly 'hears' the tone, i.e. by the time the perception reaches levels of conscious introspection, there is no distinction between a percept based on 'direct' acoustic evidence and one merely inferred from higher-level context.

Of course, it need not be a strictly incorrect conclusion for the ear to assume the tone has continued through the noise. Although the stimulus may have been constructed by alternating stretches of sinusoid with the output of a noise generator, the resulting signal is increasingly ambiguous as the noise-to-tone energy ratio increases; the unpredictable component of the noise means that there becomes no conclusive way to establish if the tone was on or off during the noise, only increasingly uncertain relative likelihoods. Thus, an alternative perspective on the continuity illusion is that it isn't an illusion at all, but rather a systematic bias towards a particular kind of interpretation in a genuinely ambiguous situation.

An even more dramatic instance of inference in hearing is *phonemic restoration* first noted in [8]. In this phenomenon, a short stretch of a speech recording is excised and replaced by a noisy masking signal such as a loud cough. Not only do listeners perceive the speech as complete (i.e. as if the cough had simply been *added*), they also typically have difficulty locating precisely where in the speech the cough occurred – and hence the phoneme that their auditory systems have inferred. Most dramatic is the fact that the content of the speech that is 'restored' by the auditory system depends on what would 'make sense' in the sentence. Of course, if the deletion is left as silence rather than being covered up with a noise, no restoration occurs; the auditory system has direct evidence that there was *nothing* in the gap, which would be inconsistent with any restoration.

2. THE PREDICTION-DRIVEN APPROACH

In order to construct computer sound analysis systems that can make these kinds of indirect inferences of the presence of particular elements, we cannot rely on data-driven architectures since they register an event only when it is directly signaled as a result of low-level processing – if low-level cues have been obscured, there is no bottom-up path by which the higher abstractions can arise. Our response is to propose the *prediction-driven* approach, whose principles are as follow:

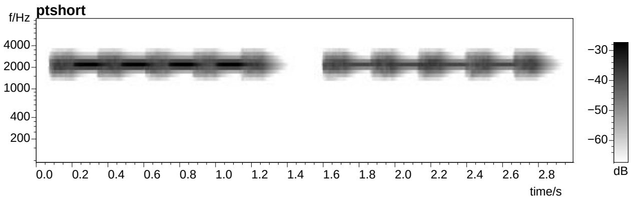

Figure 2: Spectrogram of an illustration of the continuity illusion. Repeated listening to the alternating sine-tone/noise-burst illustrated on the left gives the impression of a tone *interrupted* by noise, but when the noise energy is greater relative to the tone (as on the right), the perception is of a *continuous* tone to which noise bursts have been added. (Note the logarithmic frequency axis; this spectrogram is derived from the cochlea-model filterbank used in our system).

- **Analysis by prediction and reconciliation:** The central operating principle is that the analysis proceeds by making a prediction of the observed cues expected in the next time slice based on the current state. This is then compared to the actual information arriving from the front-end; these two are reconciled by modifying the internal state, and the process continues.

- **Internal world-model:** Specifically, the 'internal state' that gives rise to the predictions is a world-model – an abstract representation of the sound-producing events in the external world in terms of attributes of importance to the perceiver. Prediction must work downwards through the abstraction hierarchy, resulting eventually in a representation comparable to the actual features generated by the front-end.

- **Complete explanation:** Rather than distinguishing between a 'target' component in a sound mixture and the unwanted remainder, the system must make a prediction for the entire sound scene – regardless of which pieces will be attended to subsequently – in order to create a prediction that will reconcile with the observations.

- **Sound elements & abstractions:** The bottom level of the internal world model is in terms of a mid-level representation of generic sound elements [9]. In view of the requirement for complete explanation, these elements must encompass any sound that the system could encounter, while at the same time imposing the source-oriented structure that is the target of the analysis. Higher levels of the world model are defined in terms of these elements to correspond to more specific (constrained) patterns known to or acquired by the system.

- **Competing hypotheses:** At a given point in the unfolding analysis of an actual sound there will be ambiguity concerning its precise identity or the best way to interpret with the available abstractions. Since the analysis is *incremental* (maintaining at all times a best-guess analysis rather than deferring a decision until a large chunk of sound has been observed), alternative explanations must be developed in parallel, until such time as the correct choice becomes obvious.

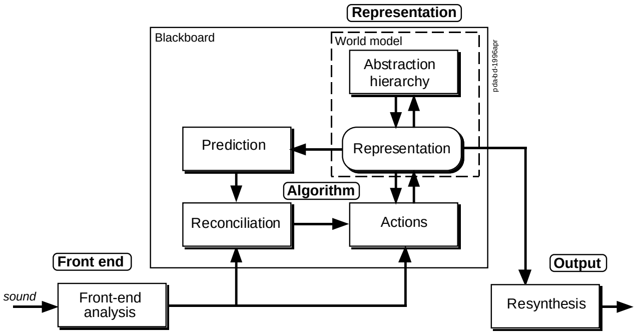

Figure 3: Block diagram of a prediction-driven sound analysis system.

The block diagram of such an analysis system is shown in figure 3. In contrast to the purely left-to-right information flow illustrated in the data-driven system of figure 1, the center of this structure is a loop, with sound elements making predictions which are reconciled to the input features, triggering modifications to the elements and abstractions that explain them.

How does such a structure address our motivation, the problem of 'perceiving' sound sources whose direct evidence may be largely obscured? The intention is that the context from which the corrupted information may be inferred has been captured in the higher-level abstractions. These will in turn generate predictions for potentially hidden elements. By the time these predictions reach the stage of reconciliation with the observed input, they will have been combined with the representation of the additional source responsible for masking them, and thus their net influence on the system's overall prediction will be slight. Assuming the internal model of the dominant interfering sound is accurate, the reconciliation will succeed without recording a serious prediction error, implicitly validating the prediction of the obscured element. Thus, the absence of direct evidence for a particular component causes no problems; its presence is tacitly admitted for as long as predictions based on its assumed presence remain consistent with observed features. This is therefore an analysis system *capable of experiencing illusions*, something we believe is very important in a successful model of human hearing.

Of course, so far this is only a rough sketch of an architecture rather than a narrow specification. In the next section we present some details of our initial implementation.

3. AN IMPLEMENTATION

We have recently completed an implementation of a CASA system based on the prediction-driven approach [10]. The system follows the block diagram of figure 3, except that the world model is shallow, lacking any significant abstractions above the level of the basic sound elements. We will now focus on three aspects of this implementation: the front end, the generic sound elements, and the prediction-reconciliation engine.

3.1 The front end

Building a model of auditory function requires a commitment to a set of assumptions about the information being exploited in real listeners, and these assumptions are largely defined by the structure and capabilities of the signal-processing front-end. For the prediction-reconciliation architecture, we need to define the *indispensable* features comprising the aspects of the input sound that the internal model must either predict or be modified to match. Foremost among these is signal energy detected over a time-frequency grid; at the crudest level, the purpose of the sound processing system is to account for the acoustic energy that appears at different times in different channels of the peripheral frequency decomposition accomplished by the cochlea. Therefore, one output of the front-end is a smoothed time-frequency intensity envelope derived from a simple, linear-filter model of the cochlea [11].

The smoothing involved in producing this envelope, however, removes all the fine structure present in the individual frequency channels, fine structure that is demonstrably significant in many aspects of auditory perception. As a partial compensation, some of this detail is encoded in a summary short-time autocorrelation function or *periodogram*, to form the basis of the detection of periodic (pitched) sounds.

Figure 4 shows the block diagram of the front end implementation. The two indispensable features, intensity envelope and periodogram, are illustrated with example gray-scale representations at the top-right and bottom right corners. In addition, there are two further outputs from the front end, consulted in the formation of elements but not subject to prediction or reconciliation. These are the *onset map* – essentially a rectified first-order difference of the log-intensity envelope – and the *correlogram*, a detailed three-dimensional representation of the short-time autocorrelation in every frequency channel (introduced in [12,13]). The periodogram is a distillation of the information in the correlogram, but the construction of elements to represent wideband periodicity in the input requires reference back to the full detail in the time vs. frequency vs. lag volume.

3.2 Generic sound elements

Previous models of auditory source separation, such as the ones mentioned in the introduction, as well as [14] and our own previous work [15], have tended to concentrate on periodic sounds (such as pitched musical instruments or voiced speech) as a very significant subset of the sonic world with a promising basis for separation – the common period of modulation. However, the prediction-driven approach's goal of an explicit representation of all components of a sound scene – regardless of their eventual casting as 'target' or 'interference' – necessitated the widening of the intermediate representation beyond simple Fourier components to be able to accommodate non-tonal sounds as well. Our implementation employed three distinct kinds of basic sound elements:

- **Noise clouds** are the staple element for representing energy in the absence of detected periodicity. Based on the observation that human listeners are largely insensitive to the fine-structure of non-periodic sounds, this element models patches of the time-frequency intensity envelope as the output of a static noise process to which a slowly-varying envelope has been applied. The underlying model is of separable temporal and spectral contours (whose outer product forms the shaping envelope), although recursive

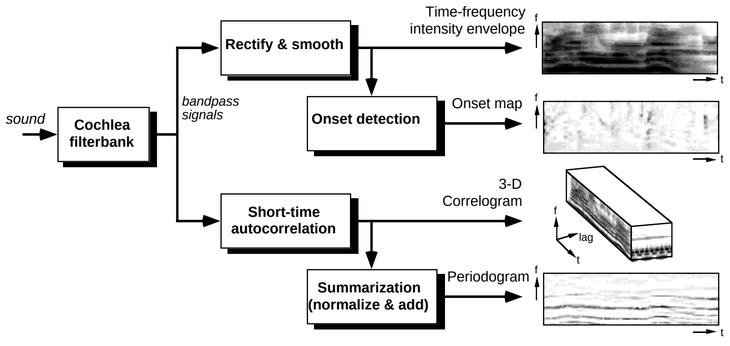

Figure 4: Block diagram of the implementation's front end.

estimation of the spectral envelope allows it vary somewhat with time. Noise clouds may also have abrupt onsets and terminations as they are created and destroyed by the reconciliation engine.

- **Transient clicks** are intended to model brief bursts of energy which are perceived as essentially instantaneous. Although this could be argued as a special case of the definition of noise clouds, the particular prominence and significance of this kind of sound-event in our world of colliding hard objects, as well as their perceptual distinction from, say, a continuous background rumble, led to their definition as a separate class of sound element.

- **Wefts** are the elements used to represent wideband periodic energy in this implementation [9] ('weft' is the Anglo-Saxon word for the parallel fibers running the length of a woven cloth – i.e. the part that is not the warp). Previous systems typically started with representations amounting to separate harmonics or other regions of spectral dominance, then invested considerable effort in assembling these pieces into representations of fused periodic percepts. Our intuition was that this common-period fusion occurred at a deeper level than other kinds of auditory grouping (such as good continuation), and we therefore defined a basic sound element that already gathered energy from across the whole spectrum that reflected a particular periodicity. As described in more detail in [9] and [10], the extraction of weft elements starts from common-period features detected in the periodogram, but then refers back to the correlogram to estimate the energy contribution of a given periodic process in every frequency channel.

It should be noted that, taken together, these elements form a representational space that is considerably overcomplete – at a given level of accuracy, there may be several radically different ways to represent a particular sound example as a combination of these elements, depending on whether the sound is considered transient, or how clearly its periodic nature is displayed, quite apart from the choice between modeling as one large element or several overlapping smaller ones. For a data-driven system seeking to convert a single piece of data into a single abstract output, such redundancy would cause problems. The top-down, context-sensitive nature of the prediction driven approach is largely impervious to these ambiguities; the current state will typically make a single alternative the obvious choice, although that choice will depend on the circumstances (i.e. the context) – much as it does in real audition.

3.3 The reconciliation engine

The final major component of the implementation is the engine that creates and modifies the sound elements in response to the comparison between predictions generated by the existing representation and the features from the front-end. In the implementation, this was accomplished by a blackboard system, inspired by and indeed based upon the IPUS system [16,17]. Blackboard systems are well suited to problems of constructing abductive inferences for the causes of observed results: they support competing hierarchies of explanatory hypotheses, and, in situations where the ideal algorithm is unknown or unpredictable, can find a processing sequence based upon the match between the data and the 'knowledge sources' or processing primitives provided.

Referring again to the block diagram of figure 3, the starting point for the reconciliation loop is the gathering of predictions for the complete set of sound elements attached to a particular hypothesis. In the implementation, simple models of temporal evolution were provided in each of the basic element types, although in a more complete implementation predictions guided by higher-level explanations for those elements would provide better anticipation of the input. Predictions in the domains of intensity envelope and periodogram are combined according to the appropriate signal-processing theory, then compared to the equivalent features from the front end. Summary statistics for this comparison measure the total shortfall or excess of prediction in each domain; when these measures exceed simple thresholds, a 'source-of-uncertainty' flag is raised. Through the blackboard's scheduling mechanism, this triggers a search for an appropriate action, which will seek to remove the discrepancy by an appropriate modification of the elemental representation. This might be the addition or removal of a whole element, or simply the modification of a given element's parameters. Situations that result in several plausible reconciliation actions can result in a 'forking' of the hypothesis to create competing alternatives, on the assumption that the correct choice will become obvious at some later time-step.

Changes to the sound-elements may trigger changes further up the abstraction hierarchy through the triggering of additional blackboard rules. The blackboard paradigm was initially developed to handle problems with very large solution spaces in which anything approaching complete search is impractical. Each hypothesis on the blackboard has an associated *rating*, reflecting its likelihood of leading to a useful solution. Computational resource is conserved by pursuing only the most highly rated partial solutions. In the implementation, ratings were calculated based upon a minimum-description-length (MDL) parameter which estimated the total length of a code required to represent the input signal using the model implicit in the hypothesis [18]. Formally equivalent to Bayesian analysis, MDL permits the integration of model complexity, model parameterization complexity, and goodness-of-fit into a single number and comprised a consistent theoretical basis for ratings assigned to otherwise disparate hypotheses.

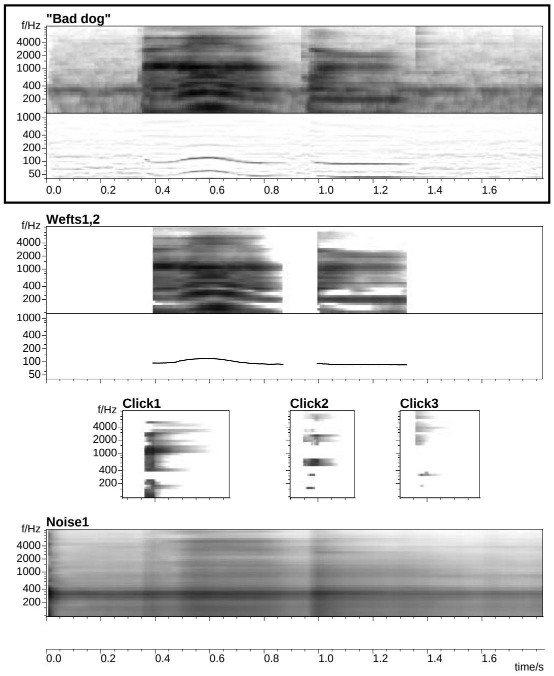

Figure 5: The system's analysis of a male voice speaking "bad dog".

4. RESULTS

Although the format of this paper precludes a more thorough description of operation of our system, some example results will provide a flavor. Figure 5 illustrates the system's analysis of some speech from a single male speaker against a background of office noise. The top panel, outlined with a thick line, shows the front-end's analysis of the sound, both as a time-frequency intensity envelope (upper) and a periodogram (lower). (Although the vertical axis of the periodogram is actually lag (autocorrelation period), the log-time lag axis used in the system has been flipped vertically and labeled with the equivalent frequency to be more comparable to the other displays). The remaining panels display the six elements generated to account for this sound in the final answer hypothesis. At the bottom of the figure is the noise-cloud element initiated at the start of the sound accounting for the more-or-less static background noise. The voice (which is the phrase "bad dog") starts around t=0.35 s, generating a transient click element (Click1), leading straight into the first weft element to follow the voiced portion of the word. (All elements are displayed by their extracted time-frequency intensity envelopes; the weft displays additionally show their period-tracks on the same axes as the periodogram). Click elements are generated for the three consonant-stop releases in the phrase, with two wefts accounting for the voiced portions of the syllables.

One goal of the elemental representation was to embody sufficient information to permit perceptually-satisfying resynthesis – confirmation that all the perceptually-relevant information has been recorded. Resynthesis of the particular elements we used was relatively straightforward; however, the results were not completely satisfactory. In an example of this kind, listening to a resynthesis of just the voice-related elements (Click1-Click3 and Wefts1,2) reveals that the clicks and the periodic wefts do not 'fuse' together as in natural speech. Presumably, some of the cues used to integrate the different sounds of speech have been unacceptably distorted in the analysis-resynthesis chain. (All the sound examples may be experienced over the World-Wide Web at `http://sound.media.mit.edu/~dpwe/pdcasa/`).

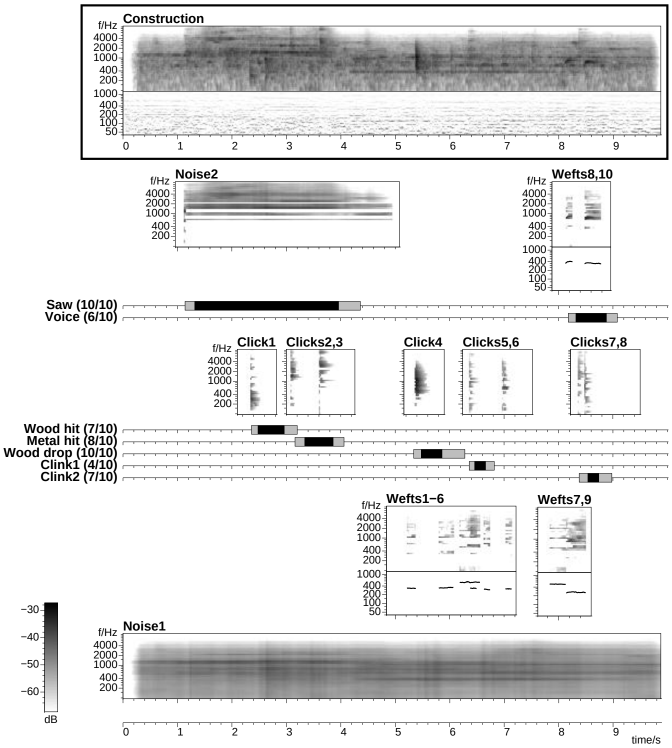

Figure 6: The 'Construction' sound example: front-end analysis (top), system-generated elements and subjective event labels. All components are displayed on a single left-to-right time axis labeled at the bottom.

The second example is of the kind of dense, ambient sound scene that originally motivated the system. Figure 6 illustrates the system's analysis of the 'Construction' sound example, a 10-second extract of "construction-site ambience" from a sound-effects CD-ROM [19]. The time-frequency intensity envelope of the original sound gives some idea of the complexity of this example, with mixtures of background noises, clicks and bangs. The system's analysis was composed of a total of 20 elements as illustrated.

Assessing the analysis is problematic, since it is far from obvious what sound events *ought* to be registered – a sound of this kind gives the subjective impression of an undifferentiated background ambience from which more identifiable events occasionally emerge. To address this uncertainty, subjective listening tests were performed in which listeners were asked to indicate the times of the different events they could distinguish, as well as giving a label to each one [10]. The summaries of these responses are displayed as the horizontal bars on figure 6; the solid bars connect average onset and offset times, with the gray extensions indicating the maximum extents. The bars are labeled with a title summarizing the subjects' consensus as to the event's identity, as well as how many of the 10 subjects actually recorded that event. In this example, there is a remarkable correspondence between the events recorded by the subjects and the elements generated by the analysis. Note however that many of the short wefts identified by the system did not correspond to subjectively-perceived events, presumably merging into the background 'babble'. 'Noise1' corresponds to the background ambience which the subjects generally did not record, although we may assume they were all aware of it.

5. CONCLUSIONS

The prediction-driven approach to computational auditory scene analysis that we have presented offers many advantages over previous approaches to this problem. Its global view of the sound scene obliges it to handle the complete range of real-world sounds as opposed to the restricted domains previously considered, and our implementation has reflected this. The orientation towards 'ambient' sound scenes, in which many different sounds overlap to a high degree and the computer attempts to make sense of as many of them as possible is in contrast to more carefully-constructed voice-on-voice examples of some previous system; while this more challenging domain may limit the achievable fidelity of extraction and reconstruction, it is perhaps closer to the real-world scenarios we would ultimately wish our sound-understanding machines to interpret.

Another important aspect of the architecture is its extensibility. A data-driven system typically has its fixed processing algorithm deeply embedded into its structure. Since a blackboard system determines its execution sequence on-the-fly to accommodate the local analysis conditions, it is a far simpler matter to add new kinds of knowledge and processing rules; as long as they can be described in terms of the known states of the blackboard's problem-solving model, they will automatically be invoked as appropriate once they have been added. A particular example of this is the possibility of adding higher levels of abstraction in the world-model hierarchy. Although the world-model in the implementation consisted almost entirely of the bottom-level sound elements, the most exciting part of the approach is the potential to benefit from higher-level knowledge in the form of sound abstractions. This even opens the possibility of a sound-understanding system that can *acquire* abstractions through experience – addressing the important question of *learning* in auditory organization.

In conclusion, we have proposed an approach to computational auditory scene analysis that promises to be better able than its data-driven predecessors to emulate subtle but important aspects of real audition such as restoration and illusion. While the current implementation has not fully investigated all the possibilities of this approach, we hope that it has provided an interesting and convincing illustration of the concepts involved. We envisage that future CASA systems derived from this approach will result in machines capable of the kinds of sophisticated and robust sound understanding currently reserved to people.

ACKNOWLEDGMENTS

This work was conducted while the author was at the M.I.T. Media Laboratory whose support is gratefully acknowledged.

REFERENCES

[1] A. S. Bregman. *Auditory Scene Analysis.* MIT Press 1990.

[2] M. Weintraub. "A theory and computational model of auditory monaural sound separation," Ph.D. thesis, Dept. of Elec. Eng., Stanford Univ., 1985.

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

[4] G. J. Brown. "Computational auditory scene analysis: A representational approach," Ph.D. thesis CS-92-22, Dept. of Comp. Sci., Univ. of Sheffield, 1992.

[5] A. S. Bregman. "Psychological data and computational ASA," in *Readings in Computational Auditory Scene Analysis*, ed. H. Okuno & D. F. Rosenthal, Lawrence Erlbaum, 1996.

[6] M. Cooke, G. Brown. "Computational auditory scene analysis: Exploiting principles of perceived continuity," *Speech Communication* 13, 1993, 391-399.

[7] A. J. M. Houtsma, T. D. Rossing, W. M. Wagenaars. "Auditory Demonstrations," Audio CD, Philips/Acous. Soc. Am., 1987.

[8] R. M. Warren, "Perceptual restoration of missing speech sounds," *Science* 167, 1970.

[9] D. P. W. Ellis, D. F. Rosenthal. "Mid-level representations for Computational Auditory Scene Analysis," in *Readings in Computational Auditory Scene Analysis*, ed. H. Okuno & D. F. Rosenthal, Lawrence Erlbaum, 1996.

[10] D. P. W. Ellis. "Prediction-driven Computational Auditory Scene Analysis," Ph.D. thesis, Dept. of Elec. Eng. and Comp. Sci., M.I.T., 1996.

[11] M. Slaney. "An efficient implementation of the Patterson-Holdsworth auditory filterbank," Technical report #35, Apple Computer Co.

[12] R. O. Duda, R. F. Lyon, M. Slaney. "Correlograms and the separation of sounds," Proc. IEEE Asilomar Conf. on Sigs., Sys., & Comps., 1990.

[13] M. Slaney, R. F. Lyon. "On the importance of time – A temporal representation of sound," in *Visual Representations of Speech Signals*, ed. M. Cooke, S. Beet & M. Crawford, John Wiley, 1992.

[14] D. K. Mellinger. "Event formation and separation in musical sound," Ph.D. thesis, CCRMA, Stanford Univ.

[15] D. P. W. Ellis. "A computer implementation of psychoacoustic grouping rules," Proc. 12th Intl. Conf. on Pattern Recog., Jerusalem, 1994.

[16] N. Carver, V. Lesser. "Blackboard systems for knowledge-based signal understanding," in *Symbolic and Knowledge-based Signal Processing*, ed. A. Oppenheim and S. Nawab, Prentice Hall, 1992.

[17] J. M. Winograd, S. H. Nawab. "A C++ software environment for the development of embedded signal processing systems," Proc. Intl. Conf. on Acous., Speech & Sig. Proc., Detroit, 1995.

[18] J. R. Quinlan, R. L. Rivest. "Inferring decision trees using the Minimum Description Length principle," *Information and Computation* 80(30), 1989, 227-248.

[19] Aware, Inc. "Speed-of-sound Megadisk CD-ROM #1: Sound effects," Computer CD-ROM, 1993.