Submitted to Speech Communications, Special issue on Computational Auditory Scene Analysis
Martin Cooke and Hiroshi Okuno, guest editors.

Using knowledge to organize sound: The prediction-driven approach to computational auditory scene analysis, and its application to speech/nonspeech mixtures

Dan Ellis <dpwe@icsi.berkeley.edu>
International Computer Science Institute
1947 Center St, Berkeley CA 94704

Abstract

Computational auditory scene analysis – modeling the human ability to organize sound mixtures according to their sources – has experienced a rapid evolution as the simple principles suggested by psychological experiments have turned out to be less than the whole story. Phenomena such as the continuity illusion and phonemic restoration show that the brain is able to use a wide range of knowledge-based contextual constraints when interpreting obscured or complex mixtures: To model such processing, we need architectures that operate by confirming hypotheses about the observations rather than relying on directly-extracted descriptions. One such architecture, the 'prediction-driven' approach, is presented along with results from its initial implementation. This architecture can be extended to take advantage of the high-level knowledge implicit in today's speech recognizers by modifying a recognizer to act as one of the 'component models' which provide the explanations of the signal mixture. Although this adaptation raises a number of issues, a preliminary investigation supports the argument that successful scene analysis must exploit such abstract knowledge at every level.

1. Introduction

The work described in this paper fits into a kind of evolutionary tale of approaches to sound organization: In the beginning, there was the 'simplistic' or "blank background" view that sound objects somehow defined themselves, and that identifying a single perceptual object was as simple as picking out a figure in a child's coloring book. The experimental stimuli on which so much of our understanding of auditory organization is based – the sinusoids and bandlimited noise bursts of Bregman [1990] and others – echo this approach, since, as presented in soundproof listening booths, they would actually be amenable to such an approach.

The second stage of evolution, which we might call the 'optimistic' or "uniform background" view, emerged from the initial efforts to apply the insights of experimental results in auditory organization (especially those in [Bregman 1990]) and apply them to real sounds. Unlike sinusoids against a silent background, real sounds contain all kinds of noise and distractions to defeat simple extraction routines, and therefore demand a more sophisticated approach. However, the signal processing community is long used to dealing with noise and offers various approaches for making the best possible decisions under some simple, but useful, assumptions. These amount to a kind of template matching, such that if the form of the target and the interference can be exactly specified, the parameters of the target can be recovered in the mathematically best-possible fashion. The essence of this approach is that we can produce simple definitions of what we are looking for – sinusoids of unknown frequency, or narrowband noise energy – and we can then go through a given signal identifying and extracting just the parts that interest us, and ignoring the rest – in analogy to the way a human listener is able to 'screen out' interfering sounds that are not of interest. I consider the early models of Cooke and Brown [Cooke 1991/3, Brown 1992, Brown & Cooke 1994] to fall into this category.

We are now at the third evolutionary stage, and in this paper I will describe one view of its defining characteristics. Based on efforts to overcome the limitations of the 'optimistic' view, we might call this 'realistic' or "structured, obstructive background" approach; the key insight is that it will not always be possible to extract a signal from interference in a unique or optimal way, but rather it is necessary to bring to bear a wide range of contextual constraints and prior biases in a heuristic search for an account of the signal that is at least reasonably satisfactory. This approach represents a considerably reduced ambition for automatic sound organization, forced upon us both by the limitations of systems based on the 'optimistic' view, and by a more careful look at some experimental results that reveal how the brain operates in more taxing perceptual situations.

Perceptual illusions come to have a central importance in this approach: Throughout the history of perceptual research, illusions – i.e. percepts which somehow failed to track the objective truth of the external world – have offered rich insights into perceptual mechanisms and processing, as well as providing the basis for many crucial technologies, such as the flickering cinematic image that is perceived as smooth motion. The model of auditory organization to be developed in this paper not only draws upon specific illusions for inspiration, but sees the *existence* of illusions as a central defining feature of perceptual systems – one to be emulated, not avoided, in our computational models. In a very important sense, the starting point for the approach to be described is: What kind of architecture can we build that would be capable of 'experiencing' human-style auditory illusions?

Restoration phenomena

The specific class of illusions to which we will attach so much importance are the so-called restoration or 'auditory induction' illusions [Warren 1996]. A simple instance is the continuity illusion (to use the term of [Bregman 1990], also studied as the 'pulsation threshold' [Houtgast 1972]) in which a sine-tone is alternated with a narrow-band noise signal centered on the same frequency. If there is no gap between the two signals, and provided the noise burst is of sufficient amplitude, the sine tone can be heard as continuing during the noise bursts. This is an illusion because the signal was not, in fact, constructed by adding noise bursts to a continuous sinusoid, although it is hardly a perceptual error given that even if the tone had been continuous, the required level of the noise ensures that the two conditions (sine tone continuous or interrupted) would be difficult to distinguish even by optimal means. The key point, however, is that the listener perceives a sine tone as present during the noise bursts by inferring that the tones before and after the burst are likely to have continued during it i.e. it is *induced* from the signal context rather than directly calculated from the signal in 'real time'. The operating assumption, that it is more likely that the noise burst is superimposed on a steady sine tone than the more complex arrangement of a noise burst that starts just at the instant that the sine tone ceases, was termed 'old-plus-new' by Bregman [1990]. It is clearly a reasonable and useful perceptual principle, but it is important to remember that listeners do not apply it consciously; rather, they perceive the sine tone as affirmatively present during the noise: At the level of conscious introspection, there is no distinction between the percepts that are based directly on the observed signal, and those that have been induced from context and other biases.

A more complex example of the same principle at work is the well-known 'phonemic restoration' phenomenon. In the classic experiment, reported in [Warren 1970], listeners were unaware when a particular phoneme in a recording of running speech was replaced with silence – provided a loud, broad-band masking signal (a cough in that example) was added to cover up the gap. Not only did listeners effortlessly restore the deleted phoneme (inferred from the lexical context), but they were unable to say exactly when the cough had occurred relative to the speech, making a typical error of 5 phonemes. Again, at the conscious level, the speech sound which was actually absent in the original signal, and which the masking noise had permitted to be restored, had no subjective distinction from the percepts arising from speech that had been heard without any obscuring signal. Later experiments went on to show that manipulating the context words could make a single deletion-plus-masker be 'restored' in several different ways – even when the disambiguating word occurred several words *later* than the deletion (e.g. "the *eel was on the axle" versus "the *eel was on the orange", where "*" indicates the noise-masked deleted phoneme) [Warren & Warren 1970].

There are two things to note here: Firstly, the required masking noise (of energy sufficient to have masked the inferred sound) shows that this not simply carelessness in the perceptual system: If a signal contains silence where there should have been a phoneme, listeners are acutely aware, and intelligibility of the modified speech is considerably disrupted. Restoration occurs only when the masking sound is present – which makes the purpose of deleting the phoneme mainly a check for the experimenter that the masker was adequate, since the listener is unlikely to be able to distinguish between masker+phoneme and masker alone. Thus, restoration reflects a 'well-adapted' response (in the evolutionary sense): Given the limitations of the auditory periphery, and the unfortunate conjunction of a loud interference with a longer speech signal, the most robust course of action is to use the surrounding context to infer, where possible, the obscured speech sound – operating at any level available, including the lexical/semantic information.

The second point is that while we can state with certainty that the deleted phoneme has been restored based on context alone – since we reduced it to silence in the original signal – the fact that its perception had no immediate conscious distinction from the surrounding unobscured speech casts doubt on what we would otherwise have assumed, namely that our perceptions of speech sounds in normal conditions are based on a direct analysis of the sound. The fact that we so readily and transparently replace obscured information with inferences based on all kinds of complex expectations raises the possibility that even in less difficult conditions a significant part of our auditory perception may be based on inference, if for some reason this is easier or more reliable that a detailed direct analysis of the actual signal. This observation could relate to a number of speech-perception phenomena such as categorical perception boundaries or our abilities to communicate despite wide differences in accent and pronunciations; such matters are beyond the scope of this paper.

'Top-down' and 'bottom-up'

As discussed in many places [Ellis 1993, Slaney 1995, Bregman 1995, Cooke 1996], restoration phenomena show the action of 'top-down' processes in perception:

On a continuum which places the physical realization of a sound at the bottom, and its most abstract conceptualization (for example, “a piece of music”) as the top level, processes that gather information at a lower level to construct descriptions higher up are termed ‘bottom-up’, and thus perception is, overall, a bottom-up process. The term ‘top-down’ refers to activity in the reverse direction, where the existence or particular state of a more abstract representation influences the fate of more concrete information at a lower level. Thus, in phonemic restoration experiments, the abstract entity of the perceived (partial) word sequence has a top-down influence on the portion of the signal during the masking noise, causing it to be interpreted as both a noise and the deleted or obscured phoneme. In practice, of course, there may not be any specific location where information is flowing ‘backwards’, but regardless of the actual implementation, such context-dependence may be termed ‘top-down’.

The concept of top-down processing is important because it is not immediately obvious as necessary. In a purely bottom-up system, fixed mechanisms process a limited region of observed evidence and generate an abstracted representation for it independent of other parts of the processing - such as generating a time/frequency/magnitude sample in a sinusoidal representation from a local maximum in a spectral slice. Such processing is largely adequate for the conceptions of perceptual organization embodied in the first two ‘generations’ of modeling described at the beginning of this paper. It is only when more complex restoration-style effects are considered that top-down processing becomes an irresistible necessity. Since, in a top-down system, the exact treatment of the local maximum in the spectral slice can depend on very much more than the immediately local features, such systems are far more complex to build and characterize, and we naturally would not consider them unless they were strongly implicated by the evidence.

Note, in contrast, that you cannot have a *purely* top-down perceptual system: at some point, the influence of the context-sensitive mechanisms must ‘make contact’ (in the words of [Darwin 1984]) with the concrete, observed information to permit bottom-up processing, even if that processing merely confirms the chain of abstractions hypothesized by the top-down processing. Thus any top-down perceptual model is necessarily a combination of top-down and bottom-up mechanisms, and the interactions and domains of these two styles of information flow define a major part of the character of any processing system. As we shall see, in the particular example of speech sounds, the regular, periodic structure of voiced speech often allows it to be extracted with little ambiguity by bottom-up mechanisms alone, whereas the lesser structure of unvoiced speech such as sibilants makes them more dependent on top-down processing for their effective characterization.

Modeling restoration

What can we say about models for the kind of processing occurring in restoration experiments? If the essence of these phenomena is that an intermediate abstraction (the perception of the ‘sound’ of the words) can be influenced both by the stimulus signal and by the abstracted context, then we need separate pieces for all three of these levels - signal, intermediate representation and abstract knowledge - and, at a minimum, mechanisms by which both extremes can influence the middle level. Also, if we compare phonemic restoration with the simpler phenomenon of the continuity illusion, we see that the kind of high-level knowledge that can exert a top-down effect may occur over a wide range of scales, from simple continuity tendencies for low-level representations to the vastly more complex processing implicit in choosing a phoneme to complete a sentence subject to semantic demands. Thus the structures providing the top-down influence might be expected to extend over several levels of abstraction themselves, quite probably containing their own mix of upwards and downwards information flow.

Although there are doubtless many possible interpretations, one processing metaphor able to provide a satisfying explanation of top-down phenomena is ‘constructive perception’ (also known as ‘analysis-by-synthesis’ [Stevens & Halle 1967]). As presented here, this conceives of perception as a process of constructing an internal model of the external world such that incoming information from the perceptual periphery can be compared to the expected behavior of the world-model as it stands, which is then confirmed or revised as appropriate. Often, it is possible closely to keep track between the external world and internal model, but circumstances such as a loud masking signal (or, more generally, some separate obscuring object) may make a temporary disconnection between the model and the observations for which it is providing an explanation. In this case, an idea of what is happening in the external world can be maintained – despite the lack of information – by allowing the model to continue evolving in the ‘most likely’ fashion (which can be defined with a wide range of sophistication), and, depending to the extent that past history forms part of the internal model, it may be appropriate to revise the model when more detailed observations again become available, to maximize agreement with all the available information. Such an account seems a plausible starting point for a computational description of restoration phenomena.

2. The prediction-driven architecture

Prediction-driven computational auditory scene analysis [Ellis 1996] is proposed as an architecture that follows the structural requirements presented so far. Full details are given in that document, but here we review the key principles of operation, focussing on how they can

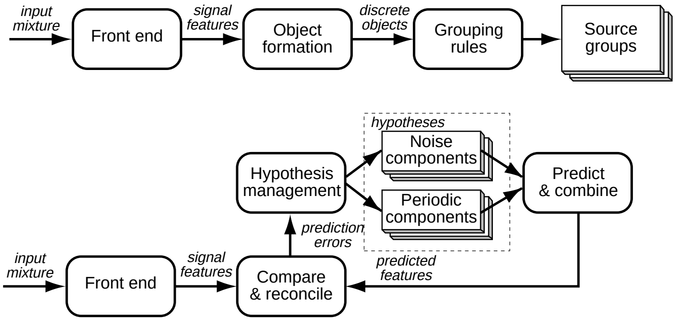

Figure 1: The principle modules of data-driven sound organization (upper panel) such as [Brown 1992] contrasted with the prediction-driven approach (lower panel) described in this paper.

encompass restoration-style effects. As illustrated in figure 1, in the prediction-driven approach, the straight, left-to-right processing chain of a data-driven sound organization system is folded back into a loop, taking the predictions derived from the internal 'world model' representation to be compared against the input as the impetus for changes to the system's state. In this respect, it is reminiscent of the 'residue-driven architecture' proposed by Okuno et al [1996] - except here the comparison is made in perceptually-motivated domains, rather than directly on the time-domain signal, and the outcome of comparison reflects a probabilistic assessment that the observation matches the predictions, rather than a simple remainder after subtraction to be dealt with independently of the terms being differenced.

Starting with the incoming signal, the first stage is the front-end processing which sets the most concrete level of the analysis. This consists of a time-frequency energy envelope, formed by smoothing the energy output by a set of gammatone band-pass filters approximating the peripheral filtering of the cochlea [Patterson & Holdsworth 1990]. Experiments show that a great deal of information is carried by even a rather coarse approximation to the signal at this level [Drullman et al. 1994], thus the envelopes contain no detail below the scale of a few milliseconds. However, certain aspects of the fine-structure of sound are highly perceptually significant, and there is thus a second representation in terms of a correlogram [Slaney & Lyon 1992], which adds an autocorrelation-lag dimension to make a three-dimensional representation for sound which reveals the pitch-range periodic modulations that may be dominant in each peripheral frequency channel. This representation is collapsed across channels to give a 'periodogram' (i.e. the summary autocorrelation as a function of time), which is the foundation of the formation of 'wefts', the representations of wideband periodic signals used in the system [Ellis 1997a], further described below.

This dual representation of the observed acoustic signal feeds one side of the comparison block, whose other input comes from the predictions based on the world-model. This latter internal representation is conceived of as a hierarchy of increasingly specific source descriptions whose bottom level is expressed in terms of a few types of 'generic sound element', each one subject to different constraints on its parameters depending on the particular hierarchy it is supporting. In [Ellis 1996], there were three kinds of sound element: wefts, representing pitched sound via a pitch-track and a broadband energy envelope, noise clouds, which had the same kind of energy envelope but no periodicity information to represent sound without a clear pitch, and click transients, which were characterized by an onset time and an exponential decay rate in each channel, recognizing the particular ubiquity and importance of such collision-style sounds in our world.

These three classes recall the division into periodic and stochastic components for musical tones in [Serra 1989], and the streams and noise beds of [Klassner 1996], as well as many other extended sinusoidal models. However, it is worth explaining that using a broad spectrum of periodically-excited energy as the lowest level of representation rather than tracking individual sinusoids - even though they might be resolved - was a deliberate effort to get away from the laborious and unsatisfying stage of harmonic grouping necessitated in sinusoid models. As a philosophical point, the ideal mid-level representation should introduce as much structure as possible that can be unambiguously observed, where doing so will not obstruct the later processing [Ellis & Rosenthal 1995].

Given a world-model in terms of these sound elements, as well as constraints or most-likely estimates for their future evolution derived from the higher-level constructs of which they form a part, it is straightforward to generate predictions for the input observations to feed to the comparator. A key idea here is that by combining the predictions from different elements which may overlap in both time and frequency, and by requiring only inexact agreement to the observations (expressed as a 'confidence bound' on the predictions), phenomena such as the 'induction' of the sine tone or phoneme may be accommodated: As long as the predictions of the model of the noise burst and the model of the continuing narrow-band pitched tone combine into an overall prediction that satisfactorily matches the noisy observation, the comparator will report agreement between model and reality, and no changes will be required. Of course, depending on the preceding context, there are very many possible combinations of elements that might have predicted equally acceptable matches to the observation, which is to say that there is a consistent bottom-level resolution of the top-down variability inherent to this approach.

The reconciliation engine, which updates the world-model in response to errors reported between the observed and predicted signals, is the most critical part

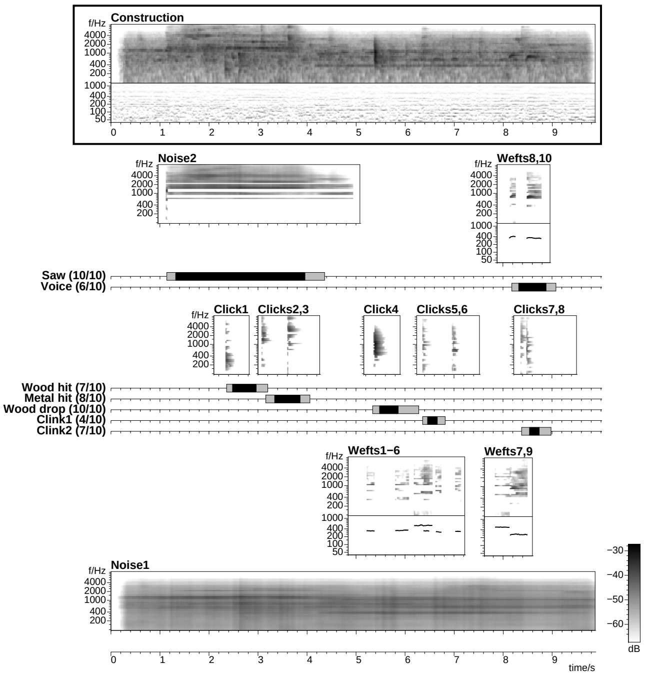

Figure 2: Analysis of the "construction-site" ambience by the initial implementation of the prediction-driven approach (from [Ellis 1996]). The top panel shows the original signal, represented by its time-frequency energy envelope and periodogram. Subsequent panels depict the various elements constructed by the system to explain the sound. Horizontal bars show the distinct components identified by subjects, along with the consensus as to their identity. This sound example can be heard at the web site: http://sound.media.mit.edu/~dpwe/pdcasa/.

of the system. It is implemented as a rule-base (within a blackboard framework [Carver & Lesser 1992]) which classifies the different kinds of prediction inadequacies depending on the elements involved. The basic operation is to adjust the parameters of existing elements, provided the deviations are not too great; when the discrepancies are larger it will add new elements or terminate existing ones as appropriate. Since very often the best course of action or the most appropriate kind of new element will be ambiguous, the reconciliation engine is also able to split models into multiple alternative hypotheses; these are pursued in parallel until the associated goodness-of-fit metrics indicate a particular winner.

Reconciliation is pursued in both domains of the front-end representation, namely energy envelope and periodogram. While any kind of element contributes to the energy envelope, only the periodically-excited wefts can account for pitch-indicating features in the periodogram, and thus the flow of causality from periodic modulation in the input to formation of a weft element in the representation is essentially still data-driven - in contrast with noise energy, which is far more context-dependent; this distinction gives a hint for the reason that pitch is so valuable in auditory organization.

As we have discussed, restoration effects often rely on very abstract knowledge or constraints to influence the interpretation of the signal: How does this appear in the prediction-driven approach? Although barely developed in [Ellis 1996], the idea was that the addition of generic sound elements in the world model would trigger the construction, within the same blackboard architecture, of higher-level abstractions, which would then invoke top-down mechanisms to reflect their particular character in the predictions of the elements supporting them. This general idea of incorporating constraints through the components of the world model forms the basis of integration with speech recognition described in part 3 of this paper.

Practical results

To demonstrate the initial implementation, the system of [Ellis 1996] was applied to a small number of 'ambient sound scenes' (such as "city street" or "construction site"), and the results compared to those obtained from human listeners. The rationale for this testing domain was that (a) sound ambiences represent the kind of dense acoustic scene that human listeners face for the majority of the time, where every sound is overlapping with several others essentially all the time, and (b) by sticking mainly to environmental sounds (bangs, squeals etc.) rather than more highly-structured human sounds (speech and music, the domains of virtually all previous work in the area), it would be possible to examine the 'fundamental' problem of sound organization – as might be being solved by cats and rats – while stripping away the complex adaptations operating in human listeners, which exploit the very complex knowledge that the brain can bring to bear, but with which the model was not endowed. An example result of this analysis is given in figure 2 (taken from [Ellis 1996], which shows the machine analysis of the "construction site" ambience example. The first panel shows the bottom-level representation of the sound, in terms of its time-frequency energy envelope and the corresponding periodogram. The boxes below each illustrate separate generic elements used by the system to account for the scene: in this implementation, the world model consisted of just this bottom layer rather than adding any more complex explanatory abstractions; this is tolerable because of the limited structure of the particular sound sources in the example. The response of listeners in the subjective tests is displayed as horizontal bars indicating the average times when listeners pressed a key 'in time' with a particular one of the sources they heard in the mixture: This allows us to confirm that the objects extracted by the system do indeed correspond to generally-perceivable auditory entities.

Issues with the prediction-driven approach

The development of the prediction-driven architecture was motivated by the theoretical concerns presented in the introduction, obviously without complete foreknowledge of the implementational issues that would arise. In practical terms, the most significant questions and areas that emerged in the system were as follow:

- As mentioned above, the **rule-base** that maps from the prediction errors to the appropriate modifications of the world-model elements constitutes the real heart of the system, almost completely determining the overall system behavior, and requiring, in practice, a great deal of empirical tuning to obtain satisfactory behavior. The rules basically consisted of sets of heuristics and thresholds for when to hypothesize elements of different kinds based on unaccounted-for features in the observations, or which existing element to terminate if the predictions exceeded what the observations could support.

  Although standard procedure for traditional Artificial Intelligence techniques such as blackboards, it is worth noting that because models of this kind require the researcher to define explicitly all of the system's knowledge and behaviors, they lack the appeal of more statistical techniques such as those used in speech recognition where aspects of the behavior can be derived automatically from training examples.
- When the rule-base did not change the set of explanatory elements, or even when it did, there remained the **error allocation** problem of splitting-up the deviations (e.g. in energy) between prediction and observation between the several elements that might overlap in a particular time-frequency cell. This too was accomplished through a complex heuristic that favored accommodating changes within elements whose parameters had the least confidence or stability. Thus if a particular noise-cloud had been observed for some considerable time and had exhibited stable properties, it would absorb much less of the difference than a newly-created transient click at the same location whose peak energy was still to be determined. (The form of the allocation did, however, ensure that they were both allocated at least a little).
- We have noted that the world-model could split into alternative hypotheses to maintain a step-by-step analysis of ambiguous situations. This raised the issue of **hypothesis ratings**, to allow the system to compare the success of the alternates in explaining the subsequent evolution of the signal, and to decide which path ultimately to pursue. The rating scheme attempted to calculate the probability of the observations given each particular model (i.e. a posterior criterion), while also accounting for the uncertainty in model parameters; a near-miss to a model with very tightly-defined parameters may or may not be better than a slightly worse fit to a more uncertain model.
- One aspect specifically emphasized in [Ellis 1996] but leaving clear room for improvement was the question of **resyntheses**. By arguing that a model of perceptual organization ought to represent every aspect of the signal that is perceptually salient, we can expect that the world-model representation should contain sufficient detail for satisfactory resyntheses of the separate components that have been identified; contriving a suitable resynthesis algorithm, however, is a different question. Although the generic sound elements were structurally simple to synthesize, subjects in listening tests gave rather modest similarity scores for extracted components when compared to the real mixtures, indicating weaknesses in analysis, synthesis or both.
- The final limitation of [Ellis 1996] is one we have already mentioned: Although a major attraction of the prediction-driven approach was its capacity to incorporate higher-level knowledge through an **explanation hierarchy**, this was left almost completely unexplored in the original implementation. While the appropriate structure and representations for this kind of knowledge are still very difficult questions, the work on integration with speech recognition to be described in the remainder of this paper constitutes one approach to injecting abstract knowledge of the structure of particular sound classes into the prediction-driven framework.

3. CASA and speech recognition

Although speech recognition systems have, in the past 20 years, developed from awkward laboratory demonstrations to successful consumer products, they have always been highly susceptible to disruption by non-speech signals, reflecting the widely-adopted restricted formulation of the problem as recognizing spoken words given that the input signal is speech. Consequently, the idea of auditory organization models that could separate mixtures and thereby free recognizers of this interference has long been attractive, and has been the more or less explicit motivation behind many of the important projects in the field [Weintraub 1985, Cooke 1991/3, Brown 1992, Okuno et al 1996]. These projects adopted an 'enhancement preprocessor' architecture, seeing the CASA system as an independent first stage to supply a conventional speech recognizer with a version of the unobstructed speech; the results as measured by recognition enhancement were disappointing – principally because the particular distortions introduced by the separation schemes were hugely disruptive to speech recognition systems, far beyond their perceptual impact. This enhance-then-recognize approach has been roundly critiqued in [Cooke 1996].

Viewed from the prediction-driven perspective, this independence of operation between organization and recognition is exactly wrong: The recognizer must include a great deal of high-level information about the

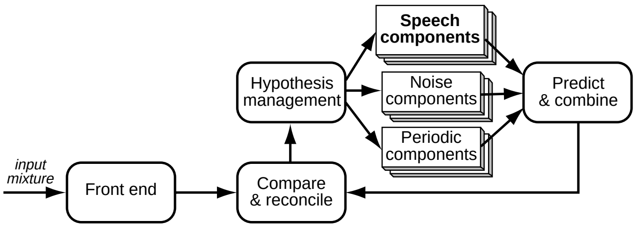

Figure 3: Extending the prediction-driven architecture to incorporate a 'speech component model', which is a conventional speech recognizer adapted to generate predictions.

structure and constraints of the speech signal, and this is exactly the kind of information that should be helpful in forming the best explanation of a mixture in terms of sound components that may include speech. As Weintraub [1985] observed, human-level performance will require a close interaction between organization and recognition, with each able to influence the other.

Using the concepts introduced so far, the ideal way to integrate speech recognition within the prediction-driven approach would be to construct a hierarchy of 'speech explanations' that worked in the domain of the three 'generic elements' (periodic wefts, transient clicks and noise clouds) to identify and refine hypotheses of speech signals within a sound mixture. Unfortunately (and this has surely been an influence on other researchers integrating speech recognition with auditory organization), existing speech recognition systems are very complex and highly specific to the particular problem formulation they have adopted; there is no simple path to adapt them to work within a different control structure or with a radically altered signal representation. Similarly, the prospect of building a recognition system from scratch within a new problem formulation that would be able to rival the complexities (and hence performance) of current state-of-the-art systems is extremely daunting. The most promising path, in the short-to-medium term, is to find ways in which existing speech recognition systems can be used with minimal alterations within novel processing schemes.

The compromise solution we have developed is to introduce, alongside the wefts, clicks and clouds, a new kind of bottom-level signal explanation element specifically matched to speech sounds, as shown in figure 3. These elements are the result of a specific 'speech component model', which is a conventional speech recognizer extended to be able to generate the predictions required by the architecture. While we have achieved only preliminary implementations of this approach at the present time, it represents a genuine integration of the signal knowledge embedded in the recognizer into mixture-organization process, and as such overcomes the severe limitations of the enhance-then-recognize approach.

Analyzing speech/nonspeech mixtures

The system operates as follows: When a new mixture is presented to the system, the speech module attempts to interpret it as speech, in the process applying its intrinsic constraints concerning the acceptable forms of a 'speech signal'. The output of the module is an estimate of the speech component in the mixture, reflecting these constraints, which is then used in the reconciliation module to focus the nonspeech explanatory elements on the remaining signal features. The predictions from these elements similarly form a 'projection' of the residual which was not accounted for by the speech module into the (less constrained) space of signals conforming to the simple nonspeech models (possibly augmented by higher explanatory levels). This estimate of all the nonspeech elements can then be used iteratively to re-estimate the speech component taking account of the nonspeech signals hypothesized to be in the mixture. Assuming that the initial estimates have some validity, cycling through this iteration should lead to a stable and satisfying explanation of the original mixture.

The issues arising from this process are:

- The speech recognition system needs to generate, in addition to the abstract description of current recognizers (i.e. the word sequence), an estimate of the actual form of the speech, less any interference, in the domain of the basic signal representation.
- The speech module should be able to take advantage of estimates for nonspeech components in the mixture, available from the iterative reprocessing.
- To get started on a new signal, there needs to be a way to get a 'bootstrap' estimate of the speech and/or nonspeech elements.

Each of these points is now discussed in more detail.

The extension of a conventional speech recognizer to reconstruct an estimate of the speech signal is described in [Ellis 1997b] and [Ellis 1998]. Although this is very similar to the task faced in speech synthesis, it cannot easily use results from that field since the problem is not to generate just any rendering of speech uttering the extracted words, but rather the precise one present in the original signal – including all the peculiarities of timing and pronunciation. We approach this problem as a question of inverting each of the stages in speech recognition – i.e. signal normalization and classification. The phonetic-state label sequence (Viterbi path) assigned by the hidden Markov model decoder is the starting point, since it contains all the information about how the signal interacted with the high-level constraints of phoneme structure, phonotactics, lexicon and grammar.

Inverting these symbolic labels to a continuous-valued feature space is the stage at which the most information has to be re-introduced, since it is effectively reversing the many-to-one projection accomplished by the classifier (Gaussian-mixture or neural-network) in the recognition path. We have experimented with weighted-overlap of simple mean templates, but more complex neural-net estimates that can track particular output tra-

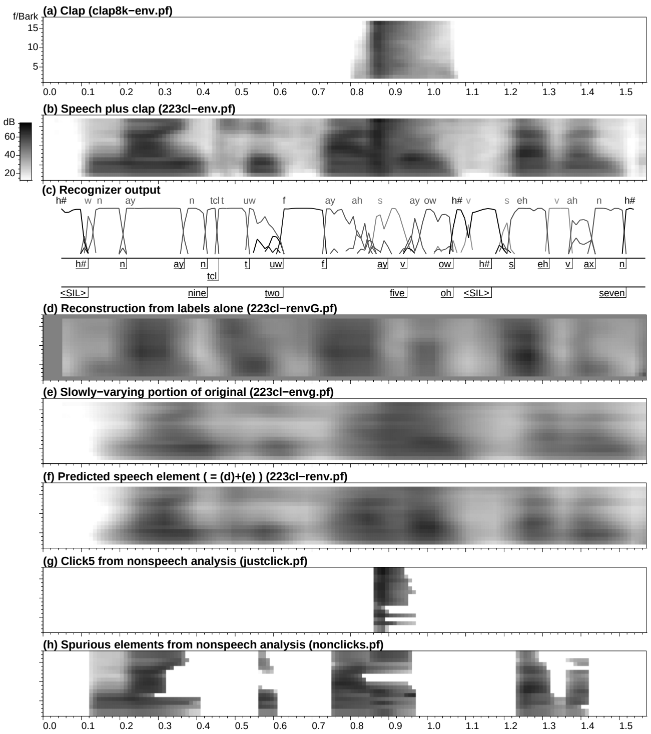

Figure 4: Preliminary results of the speech-component model. Panel (a) shows the cochlea-model spectrogram of a nonspeech intrusion (a clap), which is added to speech in panel (b). Panel (c) shows the speech recognizer's analysis both as the posterior label estimates from the classifier, and the phoneme and word labels from the HMM decoder. Panels (d), (e) and (f) show the stages of reconstructing the speech-only estimate, and panels (g) and (h) show the nonspeech elements additionally constructed to complete the signal explanation. Many of these are spurious owing to deficiencies in the reconstructed speech estimate (from [Ellis 1998]).

jectories, starting from a finer classification symbol set (accommodating phonetic context and perhaps more allophonic pronunciations), should improve this stage. The key objective, however, is that the system be able to produce good estimates of the speech features in time-frequency regions where nonspeech components have obscured direct observation; thus, a further possibility is to train a system to detect these situations explicitly (perhaps by looking at something like the entropy of the recognizer's classification outputs) and adapt the reconstruction process accordingly.

Undoing the stages of normalization, such as the equalization of average dynamic range achieved by RASTA processing [Hermansky & Morgan 1994], can be defined more precisely: Careful design and analysis of the normalization algorithms permit the generation of additional information which can be used in synthesis to reintroduce these phonetically-irrelevant aspects of variation. When speech recognition is the only goal, there is a trade-off between normalization effort and classifier complexity: A very powerful classifier (such as a large neural-network or a multiple-component mixture) can alleviate the need to normalize-away certain forms of variation. When resynthesis is also desired, however, normalization – which can often be exactly inverted – is preferable to the information loss associated with classification, leading to different priorities in constructing the recognizer portion of the speech module.

Figure 4 illustrates the operation of the extended speech module, showing how the inverted label sequence and the reconstructed normalization information combine to approximate the speech component of a mixture. At this stage of development, distortions introduced in the reconstruction lead to the generation of a number of spurious nonspeech elements; the enhancements proposed in this paper will alleviate this.

Recognizing speech given nonspeech estimates

The second issue mentioned as arising from the prediction-driven approach is the question of using the nonspeech estimates to help guide the speech recognizer. Since current recognizers are, generally, built to recognize signals presumed to consist of a single voice, they have no conception of nonspeech energy in the signal, let alone any way to make use of that information. If the central expression of the acoustic model in a speech recognizer is the emission probability distribution of the hidden Markov model, i.e. the probability

$$p(Y|q_s)(1)$$

where Y is a signal feature-vector frame and q_s is the phonetic label assignment for that frame under a given hypothesis, then a system that exploits information about additional, nonspeech elements needs to calculate

$$p(Y|q_s,Z) \qquad (2)$$

where Z is the information describing the combined nonspeech elements, expressed, for instance, in the same feature space as Y. This is reminiscent of the 'HMM decomposition' approach [Moore 1986, Varga & Moore 1990] which treats a mixture as the combined output of several processes, each described by a hidden Markov model. In that case, the observation probability is

$$p(Y|q_s,q_n) \qquad (3)$$

where the probability is additionally conditioned on the state of the second or subsequent signal models, q_n etc. The attraction of HMM decomposition is that every signal is treated the same way, and the existing speech-related tools can be used to build models of other structured interference such as helicopter noise and machine-gun fire. By contrast, eqn. (2), needed for the approach described here, depends on signal models that can generate resynthesis estimates but otherwise can be of any form. This offers several benefits: it carries no penalty

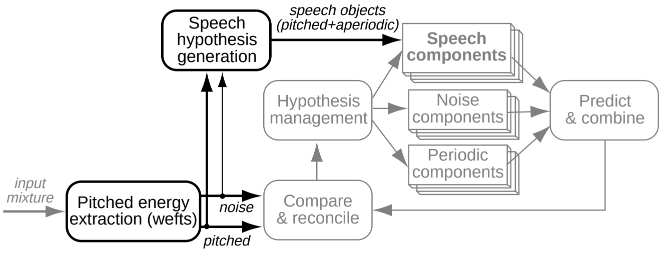

Figure 5: Proposed structure for an improved speech/nonspeech signal organization architecture that uses bottom-up estimates of pitched signals to bootstrap estimates of the speech components within the mixture.

for signals which, unlike speech, are not appropriately modeled as a sequence of discrete states; it does not grow in complexity as the number of additional signals increases, since any number of interfering signals can be combined into a single masking *Z*; and finally, assuming a representation such as spectral energy is being used, the calculation of eqn. (2) can be defined algorithmically (by backing off to priors for channels that are dominated by interference), rather than the explosively large number of enumerated cases required to cover every possible combination of states in eqn. (3). (Both approaches require tracking and accounting for the absolute level of each source, which was easily normalized out of eqn. (1), but this is a very natural thing to do in the prediction-driven domain).

Boostrapping the mixture analyzer

The final issue identified in the speech/nonspeech analysis approach considers the initial starting state of the system. In the algorithm described, the speech component model first attempts to recognize the entire mixture as speech, but in order for this to be a useful approach, the mixture should be capable of a reasonable analysis by the (conventional) speech-recognizer part of the model. When this is not the case – which, experience dictates, constitutes the majority of speech/nonspeech mixtures, as well as signals that do not contain speech – this first step will be inappropriate. What is needed is a more sophisticated mechanism for deciding when the speech model should be invoked, and for generating an initial estimate of the speech in the mixture.

The key here is to use the pitch cue: The system we have described actually uses only the time-frequency energy-envelope feature, which does not reflect any pitch information. The pitch-related features in the correlogram and periodogram have not been used, yet they offer a basis for (a) providing a somewhat robust separation between the speech (or at least its periodic, voiced parts) and other signals in the mixture, and (b) acting as a specific trigger and guide for invoking the speech-component module. The ‘pitch cue’ of coherent modulation periodicity across frequency channels was used as the sole basis for speech separation in previous CASA-for-speech systems [Brown & Cooke 1994]; like those previous systems, since the periodic features detected in the periodogram and explained by the weft elements depend far less on the state of the world-model, they can to a large part be derived directly from the signal in a data-driven fashion. Current speech recognizers do not use pitch information or make distinctions between voiced (pitched) and unvoiced sounds because this has not offered any advantages in the isolated-voice domains for which they have been developed. However, there is no obstacle in principle to building a speech recognizer that *does* operate on features divided by this criterion, for instance a recognizer trained on pairs of spectral slices, one reflecting the pitched energy (extracted perhaps by weft-style processing [Ellis 1997a]), and a second representing the unvoiced energy. (The pitch information itself would not be used (although it could be), but merely the distinction between periodic and aperiodic portions of the spectrum.) Such a recognizer could then estimate the speech content of a signal represented only by its pitched energy – estimated from a mixture without any additional knowledge – by treating the unvoiced energy (which is incapable of purely-bottom up separation) as ‘missing data’ dimensions [Cooke et al 1997].

The vision, then, is to further modify a conventional speech recognizer to treat periodic and aperiodic energy as separate feature dimensions, and then incorporate them into a prediction-driven scene analysis system as illustrated in figure 5, where bottom-up periodicity-based element extraction leads to an initial estimate of any speech components. These estimates will then result in predictions, now including the unvoiced energy that would be expected for the estimated utterance, thereby facilitating a top-down, knowledge-based organization of the remainder of the signal. Such a system constitutes a sophisticated combination of bottom-up and top-down processing for the organization of mixtures containing both speech and nonspeech, and appears to be a fair model of the sophisticated processing performed by human listeners in accomplishing this task.

4. Summary and conclusions

A consideration of some specific ‘illusory’ perceptual phenomena leads to the conclusion that auditory perception relies on the application of knowledge-based constraints at a range of levels to be able to make reasonable and useful analyses of mixtures in which direct feature observation is not possible. Building computer models of this kind of processing requires an approach different from the unidirectional data-driven calculations of conventional signal processing and more like the construction hypothesis-confirmation mechanisms that have been investigated for speech recognition. The ‘prediction-driven’ architecture embodies this alternative approach, and holds the promise of incorporating higher-level knowledge via the abstraction hierarchy in the internal world-model components.

Speech constitutes a particularly important class of highly-structured signals, and research into automatic speech recognition has resulted in systems that embody a great deal of knowledge about that structure, knowledge which ought to be helpful in separating speech from other sounds in a mixture; hitherto, however, this problem has been largely ignored. By adapting a conventional speech recognizer into a 'speech component model' this knowledge can be exploited for scene analysis within the prediction-driven framework, although further study is needed of the consequent issues of regenerating speech-component 'predictions' that reflect the abstract analysis. One major problem, that of making an initial estimate of the speech component in a mixture, could be addressed by building a speech recognizer that works separately on periodic and aperiodic spectral energy, then using bottom-up estimates of the former to seed the initial speech hypothesis.

Computational Auditory Scene Analysis systems able to approach the ability of human hearing to separate sounds in extremely demanding circumstances will require both the appropriate processing architecture and the incorporation of suitably-represented knowledge. While our current models represent only the beginnings of such a system, the results presented offer an encouraging avenue of investigation.

Acknowledgments

The work described in this paper was conducted at the MIT Media Lab and at the International Computer Science Institute, both of whom are acknowledged. Work at ICSI was funded by the European Union under project SPRACH (2077). My thanks to the ICSI Realization group for their patience in introducing me to speech recognition.

References

A.S. Bregman (1990), Auditory Scene Analysis (MIT Press, Cambridge MA).

A.S. Bregman (1995), "Psychological data and computational ASA," working notes for workshop on Comp. Aud. Scene Analysis, *Int. Joint Conf. on Artif. Intell.* (Montréal).

G.J. Brown (1992), Computational Auditory Scene Analysis: A representational approach (Ph.D. thesis CS-92-22, Sheffield University).

G.J. Brown & M.P. Cooke (1994), "Computational auditory scene analysis," Comp. Speech & Lang. 8.

N. Carver & V. Lesser (1992), "Blackboard systems for knowledge-based signal understanding," in: A. Oppenheim & S. Nawab, eds., *Symbolic and knowledge-based signal processing* (Prentice Hall, New York).

M.P. Cooke (1991), Modelling auditory processing and organisation (Ph.D. thesis, Sheffield University, published by Cambridge University Press, 1993).

M.P. Cooke (1996), "Auditory organisation and speech perception: Arguments for an integrated computational theory," Proc. Int. Workshop on the Aud. Basis of Speech Percep. (Keele).

M.P. Cooke, A.C. Morris & P.D. Green (1997), "Missing data techniques for robust speech recognition," Proc. IEEE Int. Conf. on Acous., Speech & Sig. Proc. (Munich).

C.J. Darwin (1984), "Perceiving vowels in the presence of another sound: Constraints on formant perception," Journal of the Acoustical Society of America 76, pp. 1636-1647.

R. Drullman, J.M. Festen & R. Plomp (1994), "Effect of temporal envelope smearing on speech reception," Journal of the Acoustical Society of America 95, pp. 1053-1064.

D.P.W. Ellis (1993), "Hierarchic models of hearing for sound separation and reconstruction," Proc. IEEE Workshop on Apps. of Sig. Proc. to Audio and Acoust. (Mohonk, NY).

D.P.W. Ellis (1996), Prediction-driven computational auditory scene analysis (Ph.D. thesis, Dept. of EECS, MIT).

D.P.W. Ellis (1997a), "The Weft: A representation for periodic sounds," Proc. IEEE Int. Conf. on Acous., Speech & Sig. Proc. (Munich).

D.P.W. Ellis (1997b), "Computational Auditory Scene Analysis exploiting Speech Recognizer knowledge," Proc. IEEE Workshop on Apps. of Sig. Proc. to Audio and Acoust. (Mohonk, NY).

D.P.W. Ellis (1998), "Speech recognition as a component in computational auditory scene analysis," submitted to: Proc. IEEE Int. Conf. on Acous., Speech & Sig. Proc. (Seattle).

D.P.W. Ellis & D.F. Rosenthal (1995), "Mid-level representations for computational auditory scene analysis," working notes for workshop on Comp. Aud. Scene Analysis, *Int. Joint Conf. on Artif. Intell.* (Montréal).

H. Hermansky & N. Morgan (1994), "RASTA processing of speech," IEEE Trans. Spech & Audio Proc. 2(4).

T. Houtgast (1972), "Psychophysical evidence for lateral inhibition in hearing," Journal of the Acoustical Society of America 51, pp. 1885-1894.

F.I. Klassner, III (1996), Data reprocessing in signal understanding systems (Ph.D. dissertation, Computer Science dept., Univ. of Massachusetts, Amherst).

R.K. Moore (1986), "Signal decomposition using Markov modeling techniques," tech. memo 3931, *Royal Sig. Res. Estab.*(Malvern UK).

H.G. Okuno, T. Nakatani & T. Kawabata (1996), "A new speech enhancement: Speech stream segregation," Proc.

Int. Conf. on Spoken Lang. Proc. (Philadelphia).

R.D. Patterson & J. Holdsworth (1990), “A functional model of neural activity patterns and auditory images,” in: W.A. Ainsworth, ed., *Advances in speech, hearing and language processing vol. 3* (JAI Press, London).

X. Serra (1989), A system for sound analysis/transformation/synthesis based on a deterministic-plus-stochastic decomposition (Ph.D. thesis, Stanford University).

M. Slaney (1995), “A critique of pure audition,” working notes for workshop on Comp. Aud. Scene Analysis, *Int. Joint Conf. on Artif. Intell.* (Montréal).

M. Slaney, R.F. Lyon (1992), “On the importance of time – a temporal representation of sound,” in: M. Cooke, S. Beet & M. Crawford, eds., *Visual Representations of Speech Signals* (John Wiley).

K. Stevens & M. Halle (1967), “Remarks on analysis by synthesis and distinctive features,” in: Models for the perception of speech and visual form, proceedings of a symposium (MIT Press, Cambridge MA) pp. 88-102.

A.P. Varga & R.K. Moore (1990), “Hidden Markov model decomposition of speech and noise,” Proc. IEEE Int. Conf. on Acous., Speech & Sig. Proc. (Albuquerque).

R.M. Warren (1970), “Perceptual restoration of missing speech sounds,” Science 167, pp. 392-393.

R.M. Warren (1996), “Auditory illusions and perceptual processing of speech,” in: N. Lass, ed., *Principles of experimental phonetics* (Mosby, St. Louis), pp. 435-466.

R.M. Warren & R.P. Warren (1970), “Auditory illusions and confusions,” Scientific American 223(12), pp. 30-36.

M. Weintraub (1985), A theory and computational model of monaural sound separation (Ph.D. thesis, Stanford Univ.).