SPEECH RECOGNITION AS A COMPONENT IN COMPUTATIONAL AUDITORY SCENE ANALYSIS

Dan Ellis <dpwe@icsi.berkeley.edu>

International Computer Science Institute

1947 Center Street Suite 600, Berkeley CA 94704, USA

ABSTRACT

Current speech recognition systems can perform well with high-quality speech input, but nonspeech intrusions present a serious challenge. Computational auditory scene analysis systems attempt to model listeners' ability to separate different sounds in a mixture, but efforts to use them as enhancement preprocessors for speech recognition have been disappointing. We propose integration of the speech recognizer into the scene analyzer's search for objects, thereby leveraging the recognizer's speech knowledge to help separate the sources. A conventional speech recognizer adapted to this role demonstrates the idea; a more successful implementation will require some profound changes to the recognizer, which are discussed.

1. INTRODUCTION

Robustness to extraneous, non-speech energy in the acoustic signal is one of the most pressing problems facing current speech recognition systems. Decades of research on the 'reduced problem' of recognizing clean, close-mic'd speech has allowed the community to converge on feature spaces, such as cepstral coefficients, that provide the required distinction between speech sounds in the minimum number of dimensions. Unfortunately, if we then attempt to broaden our domain to consider sounds that do not consist entirely of speech, our neatly reduced feature space is unable to make the simple speech/nonspeech distinctions that are trivially obvious to human listeners. Alternative processing, specifically oriented towards separating sources in an acoustic mixture, holds greater promise in this situation than the classification schemes constructed for isolated speech: This paper proposes an integration of a conventional speech recognizer within a system for organizing sound mixtures according to their inferred sources. Research into such systems is known as "Computational Auditory Scene Analysis" (CASA) following the psychoacoustic description of auditory organization typified by Bregman [1].

The problem of recognizing noisy speech has received considerable attention, reflecting a variety of assumptions about the character of the nonspeech noise. If the noise is continuous and has static properties that are known in advance, and if it has a fixed level relative to the speech, the simplest approach is to train the recognizer's models with speech embedded in the known noise. Alternatively, static or slowly-changing noise of unknown character can be estimated on-line and compensated for by several techniques such as spectral subtraction [2] or parameter mapping [11].

A more principled approach to recognizing combinations of sources which may have dynamic properties is to have separate models for each, then find the maximum-likelihood 'state labellings' for every source, based on a mathematically-sound combination of the signal distributions in each model. In the context of hidden Markov models, this is the HMM decomposition originally proposed by Moore [12], and used in various guises [10].

The main limitations to this approach are firstly that each noise source must have its own specific model (albeit with any number of states) – possibly a poor match to its character – and secondly that combining and testing the signal models is combinatorically expensive since, in the simplest formulation, it must be done for every possible combination of states. Again, the modeled combination implies a specific relative intensity for each simultaneous source, which must therefore be tracked, perhaps via the state.

Here, we propose an alternative approach based on the techniques of Computational Auditory Scene Analysis. In particular, we draw upon the prediction-driven CASA approach proposed in [7] to model nonspeech noises with simple parametric 'elements', possibly subject to a hierarchy of increasingly-specific signal constraints. Section 2 gives a brief introduction to CASA and then explains how speech signals could be included within such a framework as one of the possible explanations available to account for the observed mixture. Section 3 describes a specific implementation of this idea, and describes the additions to a conventional speech recognizer used in this role; the results of this preliminary system are presented. Finally, in section 4 we discuss future improvements to this approach.

2. SCENE ANALYSIS AND SPEECH

The goal of CASA systems is to construct an organization of the energy in a complex sound mixture that corresponds to the different sources that would be perceived by a human listener. This can be focussed on a specific problem such as separating voices [14], or address a more general problem of events perceived in ambient sound scenes [7]. Growing interest in this field has produced a wide variety of approaches and techniques.

Based on the 'grouping rules' elucidated by psychologists [1], early work in CASA consisted of unidirectional signal abstraction systems, in which spectral features were used to break the signal into locally-coherent patches, which were then grouped across frequency and time according to principles, such as common-period amplitude-modulation and common onset time, to assemble the complete sources identified in the mixture ([3] typifies this approach). Such algorithms which can only include or exclude elements from the initial time-frequency analysis face problems when portions of the energy corresponding to one source are distorted or masked by other parts of the signal. By contrast, numerous 'perceptual restoration' phenomena (e.g. [15]) demonstrate that, in appropriate circumstances, listeners will deduce masked signal information on the basis of context and expectations, leading to the illusion that the inferred signal was directly perceived.

2.1 Prediction-driven CASA

The prediction-driven approach was proposed in [7] as an architecture for auditory scene analysis that could incorporate listener-like knowledge-based signal inferences in the face of masking and corruption. The process, illustrated in figure 1, centers around a

comparison between observed signal features (such as subband energy envelopes) and predictions derived from the objects currently hypothesized to comprise the scene. Prediction errors trigger rules that modify the hypotheses either by changing the parameters of the constituent elements, or by adding and/or removing elements. Such an analysis-by-synthesis system can experience restoration-style 'illusions' as long as the predictions derived from the inferred signal elements are not inconsistent with the actual signal (in contrast to the requirement of earlier systems that each element be directly observable). In [7], hypotheses were constructed from three 'generic sound elements' corresponding to periodic complexes (such as voiced speech), transient 'clicks' and regions of aperiodic noise energy. The intention was to build, within a blackboard framework [4], a hierarchic abstraction of the objects present in the scene based upon these all-purpose elements, adding co-dependencies and constraints implied by the identification of, for instance, a musical instrument or a closing door. [7] presents results in which that system broadly matched the responses of listeners on a variety of 'ambient' sound examples such as the sound of a construction site or a city street.

Figure 1: Block diagram of the prediction-driven computational auditory scene analysis architecture. Predictions based on hypothesized elements are compared to observed features, and any errors drive modifications of the hypotheses. The "speech elements" are the innovation over [7] presented in this paper.

2.2   CASA and speech recognition

A major motivation for computational auditory scene analysis has been the possibility of improving the recognition of speech corrupted with other sources. The earliest system that sought to model auditory organization of utterance-sized examples [16] used recognition improvement as its evaluation metric. However, using CASA as a 'speech-enhancement preprocessor' has been disappointing owing to a mismatch between the signals expected by recognizers and those that CASA systems can provide [5] (recent innovations have ameliorated this e.g. [14], [6]). In particular, local signal cues such as periodicity that form the essence of conventional CASA systems cannot reconstruct the much more complex structure of speech signals. We presume that listeners use their acquired knowledge of speech structure when perceiving speech in mixtures; our task, then, is to find a way to incorporate this kind of knowledge into a CASA system.

The prediction-driven architecture suggests an approach to this end: Although [7] proposed introducing high-level knowledge to improve the separability of hypotheses, no suitable body of knowledge describing 'real-world' sounds was identified. However, this is exactly the information contained in speech recognition systems, which are trained to embody the characteristic spectro-temporal regularities corresponding to real speech. Integrating this within the prediction-driven architecture can be achieved by introducing a new kind of element into the vocabulary available to the scene hypotheses that will account specifi-
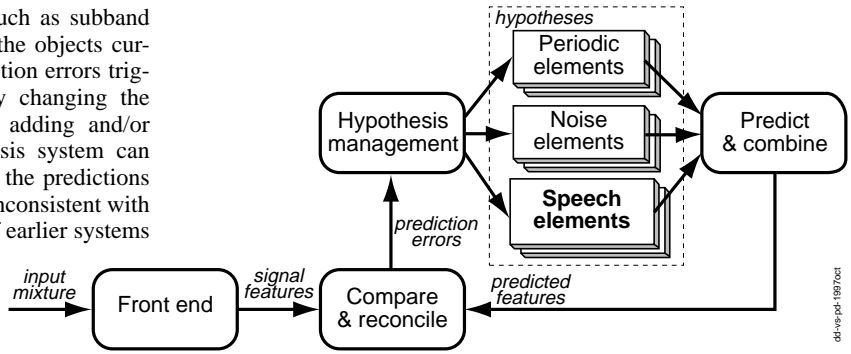
cally for speech signals. By basing this element on a speech recognition system, it will be able to incorporate the spectral and sequential constraints of real speech into the prediction-reconciliation process at the heart of the signal analysis, rather than applying them only after the initial analysis is complete.

The role of this element in its ideal form is to find a set of parameters based upon an input signal (or residual prediction error) that can control a model of speech to approximate that signal. This internal hypothesis of the speech signal, expressed in terms of abstract speech units (such as phonemes or even words) and characteristics specific to that particular utterance, is then used to generate a prediction of the speech energy in the mixture which can be combined with other hypothesized elements to test the adequacy of the entire explanation. Thus, the speech element effects a kind of projection from the input signal residual to the space of 'speech sounds', thereby introducing constraints able to make the distinction between speech and nonspeech in the mixture.

3. AN EXAMPLE IMPLEMENTATION

Prediction-driven CASA is an incremental algorithm, with 'best' hypotheses maintained throughout the evolution of the signal. However, for our initial implementation we chose not to modify the core of speech recognizer, but to investigate a two-pass iterative system: In the first pass, the sound mixture is passed to the speech hypothesis module, which performs conventional recognition on the signal and, based on the labels assigned, returns an estimate of the speech component within the mixture over the entire utterance. In the second pass, this pre-calculated 'prediction' is used within the normal incremental prediction-reconciliation analysis to construct hypotheses of addition nonspeech
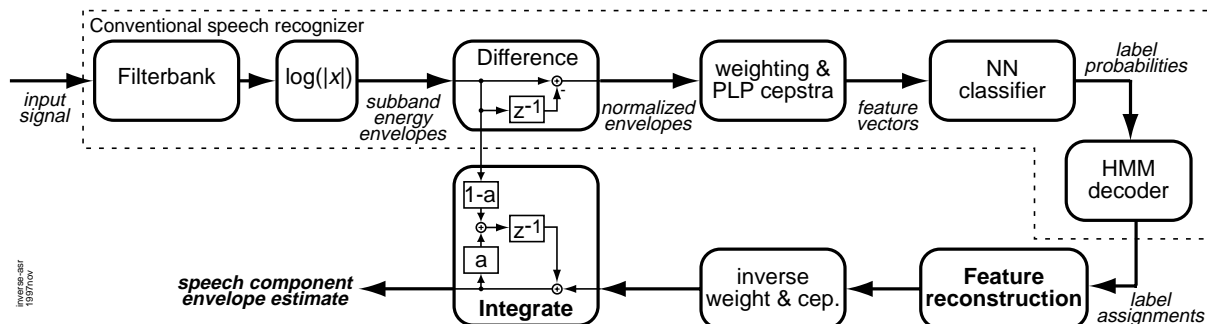

Figure 2: Block diagram of the speech element module, which consists of a conventional recognizer extended to reconstruct an estimate of the time-frequency envelope corresponding to the recognized speech.
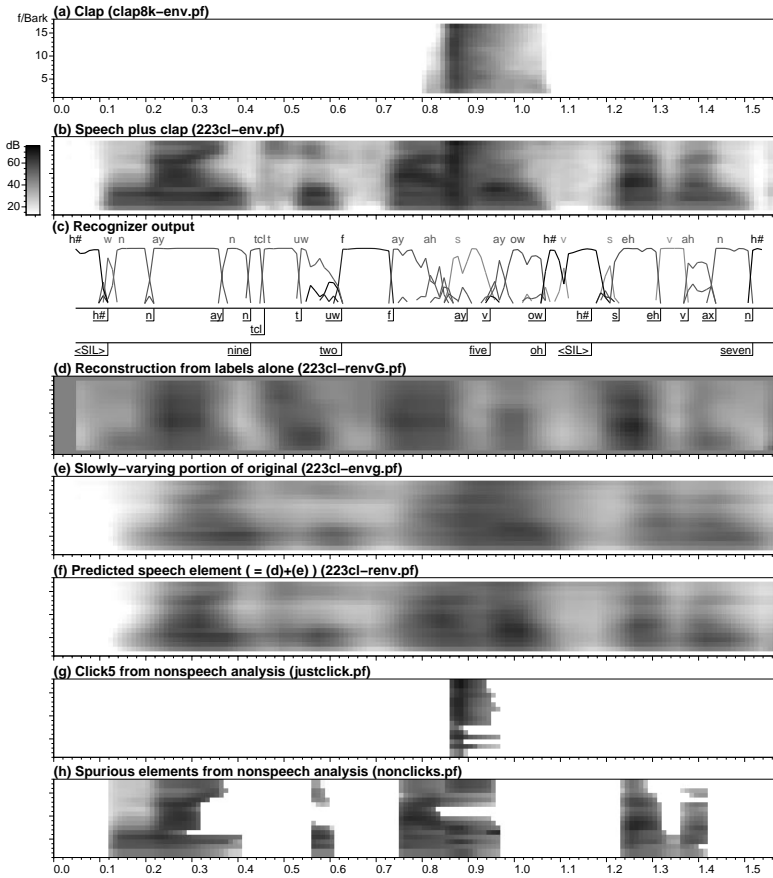
Figure 3: Analysis of speech/nonspeech mixture. See text for description.

label. Reconstruction then amounted to assigning the features at each frame to the means for the corresponding label. This drastic approximation was improved somewhat by the overlap between successive frames; because our classifier looks at a temporal context of nine frames, each reconstructed feature frame included overlap from surrounding labels. Moreover, contributions from each label were weighted by the inverse of the variance for that model dimension, thereby attaching greater weight to the more consistent values. Without this overlap, the reconstructed features would have shown highly unnatural abrupt transitions. (Other possible improvements for this step are discussed in section 4).

The final step is to invert the normalization applied to the feature vectors. In RASTA processing, variations in channel characteristics are removed by differencing the subband energy envelopes in the log domain, thereby removing any slowly-varying constant offset, equivalent to a constant gain in the linear domain. We cannot invert this differencing directly by integrating the reconstructed features; as such a non-decaying integrator would accumulate all the distortion introduced by the intermediate processing. Instead, we accumulate a weighted average between the integrated signal and the original envelope which it is approximating, as shown in figure 2. The weighting parameter a defines a low-frequency breakpoint (within each subband envelope) between the signal that is recovered from the reconstructed features and the slowly-varying portion, normalized away by RASTA, copied directly from the original input.

The net result of this processing is an estimate of the speech spectrum in the original mixture formed of the general trend of input signal upon which has been superimposed a more rapidly-varying structure derived from the label assignment made by the recognizer (illustrated in figure 3 as described below). Rapid variation which cannot be interpreted as speech is thus excluded and left for the nonspeech elements to explain.

3.2 Preliminary results

Figure 3 gives an illustration of the current system analyzing a mixture of telephone speech (taken from the OGI "Numbers" database) corrupted with a hand-clap. The top panel shows the clap alone, and the second panel the speech/nonspeech mixture. (All displays are energy envelopes in the Bark-scaled frequency axis used in RASTA-PLP processing). The speech recognizer gives the probability estimates and label assignments shown in panel (c); note that although the clap sound disrupted the local probability estimates, the constraints of the lexicon and grammar result in a final labelling that is essentially correct despite the corruption. Panel (f) shows the reconstructed speech envelope, which is the sum of the slowly-varying portion of the original signal (panel (e)) and the features reconstructed from the decoder labels, represented in panel (d).

Using this reconstructed speech envelope within the complete scene analysis system results in a number of hypothesized nonspeech elements to account for residual prediction errors. Panel (g) shows the transient element from this analysis that has located the clap in the mixture. However, as shown in panel (h), six other objects were constructed, arsing from shortfalls between the speech envelope and the reconstruction. Sound examples can be found at http://www.icsi.berkeley.edu/~dpwe/research/icassp98.

elements required to provide a full explanation for the signal. These preliminary estimates of the nonspeech components may then be used to help relax the constraints on the speech recognizer in later iterations, although this reprocessing has yet to be applied in practice. The net result of the analysis is a set of hypothesized objects to explain the signal, consisting of the recognized speech and the nonspeech components.

3.1 A modified speech recognizer

To implement this system, we need to regenerate an estimate of the speech signal from the recognizer output i.e. to perform the 'projection' into the space of speech signals. Our approach, illustrated in figure 2 (based on our earlier work [9]), was to invert each stage of the recognizer back to the smoothed subband energy envelopes used as the principal domain of the nonspeech analysis. To generate the label probability estimates for the hidden Markov model decoder, a speech recognizer must normalize the signal into feature vectors that exhibit some consistency across different versions of the same token, then classify these vectors into the different label classes. It is these two stages, normalization and classification, which must be inverted in the modified recognizer.

Our recognizer consisted of RASTA-PLP features [11] followed by a neural-network classifier [13]. Given the best-path (Viterbi) label assignments from the recognizer, inverting classification implies the reconstruction of feature-vector sequences. Classification is implicitly many-to-one, so this problem is ill-formed, and our opaque neural-net classifier hinders still further. For the initial implementation, we used our labelled training speech database to construct a single diagonal-covariance Gaussian model for each

In theory, the system would use these initial nonspeech estimates to go back and revise the speech object. This is not currently implemented, because a simple subtraction of the nonspeech estimates leaves holes in the original signal; better reconstruction and a more intelligent way to use the information are both required.

4. CONCLUSIONS & FUTURE WORK

The essential idea behind this work is that constraints on what constitutes both speech and nonspeech sounds may be applied to iteratively separate one from the other. Our initial demonstration serves to illustrate the idea and indicate its viability, but also highlights the remaining barriers to a full and useful implementation.

4.1 Anticipated improvements

The most pressing problem is the inadequacy of the speech reconstruction that led to the construction of spurious nonspeech objects. While this reconstruction is a difficult problem, a number of potential improvements are indicated:

Inverting the *classification* to generate features from labels could be improved by simple measures such as using a greater repetoire of labels, making each class is more specific. In particular, defining separate classes for phoneme-centers versus transition regions could considerably sharpen the modeling. A more sophisticated approach would be to train a neural net to reconstruct features based on a time-context of labels, and perhaps additional information such as the pre-decoder probabilities.

To improve the benefits of inverting *normalization*, we could employ a recognizer that uses more sophisticated normalization techniques such as spectral warping. The more of the speech variation that can be modeled and removed by normalization rather than being left to the classifier to absorb, the more specifically the reconstruction process can mirror the actual speech signal.

A radical change would be to train a recognizer on a signal divided into periodic and non-periodic portions (separated by a mechanism such as [8]). This might get around the bootstrapping problem of finding an initial speech hypothesis when the signal is highly corrupt: a periodic component could generate a speech hypothesis which would then indicate which portions of the aperiodic component should be regarded speech, and which should be explained by other means. This could be a satisfying account of the human ability to separate speech from noise.

In order to make appropriate use of the nonspeech elements in speech re-estimation, we need a classifier that can ignore portions of the signal that have been masked by other sounds. High energy in the nonspeech hypotheses could be used to back-off the label probability estimates to their priors in the corresponding frames, but a more principled transition would be desirable.

The question of projecting a signal into 'the space of all speech sounds' could evidently benefit from the techniques of speech synthesis, but this has yet to be investigated.

4.2 Conclusions

In real-world applications, the problem of nonspeech energy appearing in speech cannot be avoided. Human listeners are extremely adept at separating such mixtures, and an effort to model this auditory organization holds much promise for improving speech recognition systems. Illusory phenomena in hearing suggest a constructive analysis system, and the prediction-driven architecture provides a suitable framework; including speech models alongside the nonspeech elements creates a single framework for recognizing the speech at the same time as modeling the nonspeech components, with each component aiding the characterization of the others. This approach operates without detailed *a priori* knowledge of the noise signals or complex precalculation of combined distributions. The current implementation has obvious limitations, but the way is clear for detailed modelling that permits more accurate reconstructions of the speech component, and hence complete scene analysis.

REFERENCES

[1] A. S. Bregman, *Auditory Scene Analysis*. Cambridge MA: MIT Press, 1990.

[2] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, 1979.

[3] G.J. Brown & M.P. Cooke, "Computational auditory scene analysis," *Comp. Speech & Lang.* 8, 1994.

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

[5] M.P. Cooke, "Auditory organisation and speech perception: Arguments for an integrated computational theory," *Proc. Int. Wkshp on the Aud. Basis of Speech Percep.*, Keele, 1996.

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

[7] D.P.W. Ellis, *Prediction-driven computational auditory scene analysis*, Ph.D. thesis, MIT, 1996.

[8] D.P.W. Ellis, "The Weft: A representation for periodic sounds," *Proc. IEEE Int. Conf. on Acous., Speech & Sig. Proc.*, Munich, 1997.

[9] D.P.W. Ellis, "Computational Auditory Scene Analysis exploiting Speech Recognizer knowledge," *Proc. IEEE Wkshp on Apps. of Sig. Proc. to Audio & Acous.*, Mohonk, 1997.

[10] M.J.F. Gales & S.J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech & Audio Proc.* 4(5), 1996.

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

[12] R.K. Moore, "Signal decomposition using Markov modeling techniques," *Royal Sig. Res. Estab.*tech. memo 3931, 1986.

[13] N. Morgan & H. Bourlard "Continuous speech recognition: An introduction to the hybrid HMM/connectionist approach", *IEEE Signal Processing* 12(3), May 1995.

[14] H.G. Okuno, T. Nakatani, T. Kawabata, "A new speech enhancement: Speech stream segregation," *Proc. Int. Conf. on Spoken Lang. Proc.*, Philadelphia, 1996.

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

[16] M. Weintraub, *A theory and computational model of monaural sound separation*, Ph.D. thesis, Stanford Univ., 1985.