Presented at IEEE workshop on Apps. of Sig. Proc. to Aud. & Acous., Mohonk, 1997

COMPUTATIONAL AUDITORY SCENE ANALYSIS EXPLOITING SPEECH-RECOGNITION KNOWLEDGE

Dan Ellis

International Computer Science Institute
Berkeley CA 94704
<dpwe@icsi.berkeley.edu>

ABSTRACT

The field of computational auditory scene analysis (CASA) strives to build computer models of the human ability to interpret sound mixtures as the combination of distinct sources. A major obstacle to this enterprise is defining and incorporating the kind of high level knowledge of real-world signal structure exploited by listeners. Speech recognition, while typically ignoring the problem of nonspeech inclusions, has been very successful at deriving powerful statistical models of speech structure from training data. In this paper, we describe a scene analysis system that includes both speech and nonspeech components, addressing the problem of working backwards from speech recognizer output to estimate the speech component of a mixture. Ultimately, such hybrid approaches will require more radical adaptation of current speech recognition approaches.

1. INTRODUCTION

Listeners are able to interpret complex sound mixtures through the strong constraints provided by their knowledge of ‘actual sounds’. A major obstacle to researchers in computational auditory scene analysis, building computer models of this ability, is the question of collecting, representing and deploying such constraints. If the ability to understand sound mixtures is intimately bound to knowledge of real-world sound characteristics, it will be difficult to make progress in modeling one without a reasonable grasp on the other.

Fortunately, there exists a domain in which considerable achievements have been made in capturing the typical features of a class of sound: automatic speech recognition (ASR). This paper looks at integrating the approaches and domains of computational auditory scene analysis with the data-derived knowledge and ambiguity-resolution techniques of automatic speech recognition. While much previous work in CASA has been oriented towards helping the problem of speech recognition [1, 2, 3], this has almost always been formulated in terms of a decoupled preprocessor [4]; given that speech recognition is currently the further advanced of the two domains, we consider the converse possibility of using speech recognizers to help scene analysis systems, and integrating the two processes to benefit them both. (This was indeed the approach which Weintraub lamented he could not take [1]).

A current theme in CASA work is *iterative explanation*, in which an account of a scene is constructed by attending to the successive residuals left after explaining more prominent pieces [5, 3, 6]. The approach adopted in this paper is to analyze mixtures of speech and environmental sounds by hypothesizing the presence of objects of both types, then iteratively refining each component. Exploiting *general source knowledge* represented as the state of hypothesized models is the explicit goal of Ellis’s ‘Prediction-driven’ CASA [6]; it is also implicit in Moore’s decomposition of a signal as the combination of hidden Markov models [7]. However, the majority of work in CASA has concentrated on identifying the number and extent of the different sources present, while limiting them to simple models such as smoothly-varying periodic sounds [2].

A major barrier to the mutual integration of scene analysis and speech recognition systems is their incompatible representations. ASR systems assume that their input is a single voice, hence they employ representations (such as normalized cepstral coefficients) that encode phonetically-relevant signal variation in a low-dimension space, excluding detail such as voicing periodicity. By contrast, periodicity is the most popular cue in scene analysis systems which must therefore use richer representations. Even assuming speech, a CASA system would strive to separate each of several overlapped voices present in an acoustic scene; ASR systems fail miserably when they encounter such unanticipated complications. This project addresses integrating the two fields by finding a translation between their representations.

After describing an overall design for a ‘hybrid’ speech/nonspeech scene analysis system, we examine in particular the adaptations required of a conventional speech recognizer for this purpose. After showing some preliminary results, we conclude by discussing how this approach compares to previous CASA and ASR systems, and make some observations on the kinds of components required for an improved analyzer of speech-bearing acoustic scenes.

Practical motivations for this work are diverse: Robustness to nonspeech interference is a major issue in ASR, and new approaches to recognition of speech signals corrupted by additions are urgently required. Other applications could include multimedia indexing interested more in extracting the nonspeech sound effects as content indicators, but which must handle speech appropriately [8]. Another scenario is a portable aid for the deaf, providing a textual description of the sound environment in near-real time [9]. Ultimately this same information will be required by humanlike robots of the future.

2. SYSTEM OVERVIEW

Figure 1 shows a block diagram of the complete system. The sound mixture is fed to the front-end, which consists of a bank of band-pass filters approximating the ‘critical bands’ of the human auditory system, followed by temporal envelope extraction. This gives a smooth

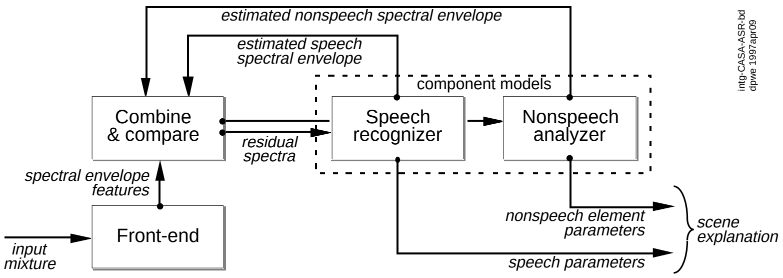
estimated nonspeech spectral envelope
estimated speech spectral envelope
component models
Combine & compare
Speech recognizer
Nonspeech analyzer
residual spectra
spectral envelope features
Front-end
input mixture
nonspeech element parameters
speech parameters
scene explanation
intg-CASA-ASR-bd dpwe 1997apr09

Figure 1: Overview of the CASA-ASR hybrid system.

representation of the signal's energy as a function of time and frequency. These spectral envelope features form the input to the comparator, which subtracts combinations of the estimates from the 'component models' (currently the speech recognizer and nonspeech analyzer, although in principle more could be added) to form 'partial residuals', which are returned to the models for re-estimation.

Each component model attempts to explain the partial spectrum it has been given according to its constraints: The speech recognizer searches for a matching phoneme sequence, and the nonspeech analyzer (which is a simplified version of the system described in [6]) tries to match its input with simple noise elements. In each case, a model will generate two outputs: abstract model parameters (e.g. the phoneme sequence or the noise profiles), and the spectral surface implied by these parameters. This estimate of the contribution of the model to the overall signal is fed back to the comparator ready for the next round of iterative estimation; this process repeats until the complementary estimates stabilize.

The sophisticated sequential constraints of the speech recognizer require a certain temporal context to identify the preferable label assignments. This 'interpretation lag' requires the entire system to work at a large temporal granularity of hundreds of milliseconds.

Two questions arise immediately with such an iterative system: how do we obtain starting estimates, and will the iteration converge? We do not have general answers to these points, but as long as the signal is reasonably close to speech, the high-level constraints of the speech recognizer will push towards a single local interpretation, leaving the less-constrained nonspeech models to mop up the remainder. We note a resemblance to the Expectation-Maximization (EM) algorithm: the system makes an allocation of the signal energy to the different component models based on the spectral estimates of the last iteration; these are then fed to the models, which searches for the parameters that maximize their fit to the allocated spectrum.

3. THE SPEECH RECOGNIZER

In the system overview, the speech recognizer constitutes a single component model, taking a partial residual spectrum as input, and generating an abstract explanation (as a sequence of phonetic labels) and a corresponding estimate of the speech component's spectrum in the mixture. This function is broken down in figure 2. The left-hand side of the figure constitutes a conventional speech recognizer: the input spectral envelope is first normalized by RASTA filtering [10], which applies a band-pass filter in the log-domain to the envelope of each frequency channel to remove long-term transfer characteristics. This normalized spectrum is smoothed with the so-called Perceptual Linear Prediction, then projected into a condensed and decorrelated feature space with a truncated cepstral transform. This gives a 13 element feature vector for every 12.5 ms frame, typical of the low-dimensional feature spaces used in speech recognition.

The next step estimates the probability that a given frame represents each of 56 phonetic labels with a neural-net classifier, trained to match a hand-labeled corpus. The network actually looks at the 13 features plus their derivatives and double-derivatives over a 9-frame context window giving 39x9 = 351 input units feeding a hidden layer of 500 units. Estimates of label probabilities are fed to a Markov decoder, which searches for the most likely label sequence in conjunction with its knowledge of word and language structure. This gives an interpretation of the input signal as a sequence of phoneme labels (and hence words). At this stage, conventional speech recognition is complete.

For our task of mixture interpretation, however, we also need to reconstruct an estimate of the 'inferred' spectrum for the speech component. This is the role of the right-hand half of the figure, which converts the phone labels back into the spectral domain. Firstly, the labels are converted into the recognizer feature space – the converse of the neural-net classifier. This stage is problematic, since it is notoriously difficult to work a neural net 'backwards' to identify the actual regions of feature space that correspond to a particular output. For the moment, we substitute the mean vector calculated over all correspondingly-labeled training frames. This reconstituted feature vector is transformed to the normalized spectral domain with an inverse cepstral transform.

The final step is to reverse the normalization of the initial RASTA step, which removed the slowly-varying portion of the temporal envelope in each band; denormalization is a matter of restoring this low-frequency portion from the input signal. Converting this summation back into the linear domain gives the desired estimated spectral contribution, which carries short-term variation determined by the modeled phonetic label sequence superimposed on the slowly-varying component of the original input spectrum.

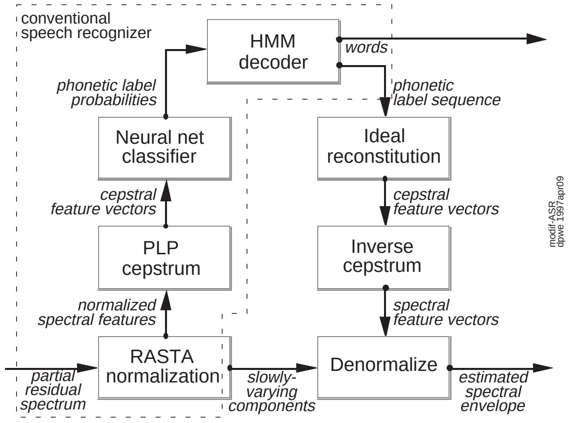

Figure 2: Components of the modified speech-recognition module.

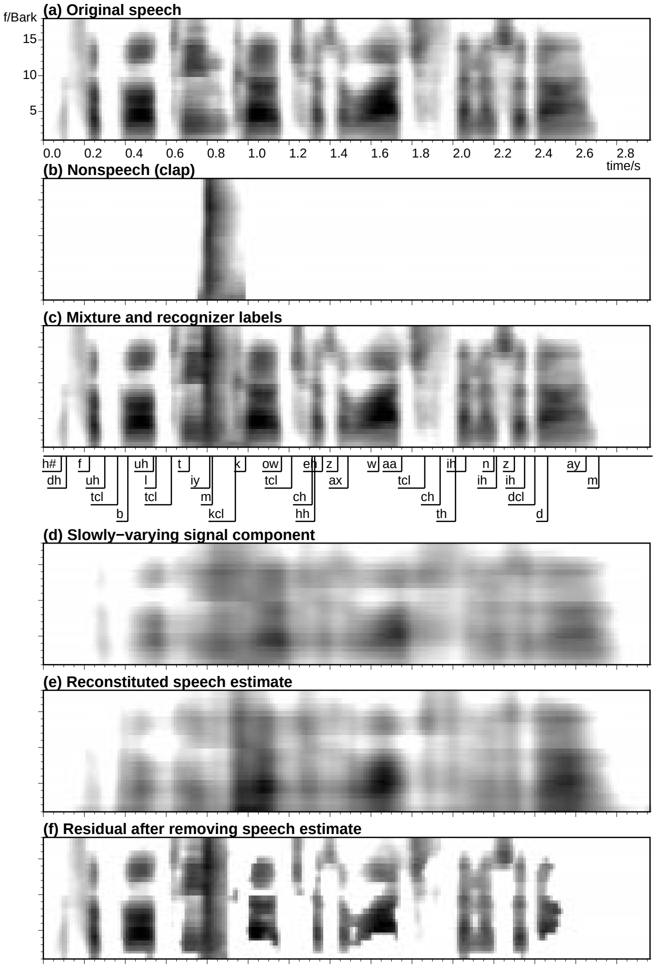

Figure 3: Analysis of speech/nonspeech mixture.

4. RESULTS

Figure 3 illustrates the results of the current implementation, which has focused on the speech signal reconstitution. The top image is the spectrogram of a clean speech signal, shown in the auditory filter bank domain that is the base representation of the system. The next panel shows the spectrogram for a nonspeech addition (a clap), and the third image is the spectrogram of their mixture: the noise of the clap is visible superimposed upon the speech. In the first iteration of the system, no nonspeech elements have been proposed, so the entire signal is fed to the speech recognizer; the phonetic labels assigned by the decoder are shown below the mixture. The next stage takes the low-frequency portion of the input signal (panel (d)) and superimposes more rapid fluctuations derived from the labels to reconstitute an estimate of the speech component in the mixture (panel (e)). Note the absence of a transient aligned with the nonspeech burst in the mixture, since this was not reflected in the phonetic labels.

The bottom frame shows the residual after removing the speech estimate from the original signal; this is the input to the nonspeech scene analyzer. The clap transient – which should be explained as nonspeech – is prominent as we expect. However, many other patches of energy appear where the reconstituted speech has failed to reach the peak energy in strong vowel segments. We are currently working to resolve this misalignment, as well as the further integration of the nonspeech models.

When complete, this residual will be modeled as a combination of noise bursts by the nonspeech analyzer; this nonspeech estimate will in turn be subtracted from the full mixture, giving a new residual to be relabeling by the recognizer, onward through repeated iterations until convergence is reached.

5. DISCUSSION

Comparison to other approaches

It is only relatively recently that speech recognition has been good enough to allow researchers to consider mixed signals. A common approach to recognizing speech in noise has been to base a recognizer on features that are distorted by adding noise to the training database, or by a suitable transformation of the templates [11]. This approach treats the nonspeech component as stationary; by contrast, scene analysis explicitly detects and models interference and can exploit structure in the nonspeech component.

HMM decomposition [7] is also able to exploit prior knowledge of dynamic nonspeech additions by finding combinations of hidden Markov models for both speech and interference that fit the mixture. While HMMs are excellent for speech, which is well modeled as a sequence of discrete symbols, their applicability to sounds such as footsteps or passing cars is less clear: their attributes may vary continuously, resisting the assignment of discrete states. Also, the speech-specific features used with HMMs are inadequate for many nonspeech distinctions. The biggest practical limitation of HMM decomposition models is the calculation of the conditional probabilities of state assignments given the input observations, which typically requires integrating across all possible divisions of the observation and fixed relative levels of the models. The simpler nonspeech models of the current system avoid these problems but sacrifice a rigorous probabilistic foundation.

The biggest distinction between the current system and most work in CASA is that it does not use the pitch cue. Speech recognizers similarly ignore periodicity, although perceptual experiments demonstrate pitch to be an important basis for sound organization [12]. Few CASA systems have exploited much structure in their signals beyond local data features; top-down components rely on stored templates [5].

Some problems and possible improvements

The iteration between speech and nonspeech presents a start-up problem: one or other component has to make a preliminary effort to recognize the mixture. The speech part incorporates stronger constraints, but to start with it implies that a mixture which confounds a conventional speech recognizer will also defeat this system. It is interesting to speculate how humans handle this bootstrapping problem; the paradigm employed here of ‘latching on’ to a speech signal then looking for additional explanations of whatever is left, has intuitive appeal.

The noise-cloud nonspeech model should work for many ‘environmental’ sounds, but music and other speech will require more radical adaptations of the speech recognition component to employ a distinction of periodic and aperiodic signals, and more involved integration of the nonspeech analysis module, which must include periodicity-based separation.

Adaptive recognizers, which shift their classification boundaries to match inferred speaker characteristics, could improve nonspeech discrimination by supporting more accurate estimates of the speech component.

Cooke [4] makes a number of observations concerning desirable properties of speech recognizers to be used with scene analysis systems. The current system would benefit from his idea of a recognizer that penalizes *absence* of energy more strongly than *excess* (since excess could be caused by mixture components, whereas absence cannot).

This ties into one of the key ideas of the prediction-driven approach of Ellis [6], that the ubiquity of masking is a problem for models based on subtraction and residuals: a distinction must be made between the absence of energy in a given channel, and the situation when an existing element has accounted for all the input energy at a level that could be masking other contributions. His solution is to use probabilistic representations of spectral level, incorporating positive and negative deviation bounds around a specified level: zero energy in the input has very tight bounds, zero residual behind a masking prediction has considerably more latitude. Expressing signal estimates as probability densities would permit a more meaningful reconstitution from phonetic labels, which relate more naturally to spectral distributions.

6. SUMMARY AND CONCLUSIONS

Any sound understanding system, even if its primary focus is speech signals, will need to handle both speech and nonspeech sounds in the real world. Human listeners are proficient at organizing sound mixtures, thanks to their general knowledge of sound. To build a computer system that can approach such human abilities we need to solve the incorporation of this kind of knowledge.

Automatic speech recognition has become a practical reality thanks to statistical approaches to collecting and exploiting knowledge of the structure of speech sounds. This work looked at employing this structural knowledge of speech in a wider sound understanding domain. We combined a speech recognition module into a computational auditory scene analysis framework that can interpret parts of the signal either as speech or with models of nonspeech sounds. This required a way to work backwards from phonetic label assignments to an estimate of the speech signal. Using this transformation, an iterative algorithm for estimating speech and nonspeech components in a mixture is made possible.

The problem of handling sound mixtures must be solved by intelligent recognition of nonspeech as well as speech. Integrated systems of the kind described have many applications beyond speech input, including content-based indexing of multimedia databases and aids for the hearing impaired. Future developments will include speech recognition components better suited to partially-obscured signals, and models for nonspeech able to characterize a wider range of the sounds we encounter. Combing the best speech recognition with the most powerful ideas from scene analysis will lead to integrated systems that perform both tasks far better than any approach that takes a less realistic view of real-world sounds.

ACKNOWLEDGMENTS

The author would like to thank the European Union for support through their basic research grant to project THISL (23495), and the entire ICSI speech group for teaching him about speech.

REFERENCES

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

[2] G. J. Brown, *Computational Auditory Scene Analysis: a representational approach*, Ph.D. thesis, CS dept., Sheffield Univ, 1992.

[3] H. G. Okuno, T. Nakatani, T. Kawabata, “Interfacing sound stream segregation to speech recognition systems – Preliminary results of listening to several things at the same time,” *Proc. AAAI-96* (2), pp. 1082-9, 1996.

[4] M. Cooke, “Auditory organisation and speech perception: Arguments for an integrated computational theory,” *Proc. ESCA workshop on the Aud. Basis of Speech Percep.*, Keele 1996.

[5] V. R. Lesser, S. H. Nawab, F. I. Klassner, “IPUS: An architecture for the integrated processing and understanding of signals,” *AI Journal* 77(1), 1995

[6] D. P. W. Ellis, *Prediction-driven Computational Auditory Scene Analysis*, Ph.D. dissertation, EECS dept., M.I.T., 1996.

[7] R. K. Moore, “Signal decomposition using Markov modeling techniques,” *Royal Sig. Res. Estab.* Tech. memo no. 3931, 1986

[8] S. Pfieffer, S. Fischer, W. Effelsberg, “Automatic audio content analysis,” *ACM Multimedia'96*, Boston, 1996.

[9] R. S. Goldhor, Audiofile Inc., private communication, 1992.

[10] H. Hermansky, N. Morgan, “RASTA processing of speech,” *IEEE Tr. Speech & Aud. Proc.*, 2(4), pp. 578-589, 1994.

[11] M. F. Gales, S. J. Young, “Robust continuous speech recognition using parallel model combination,” *IEEE Tr. Speech & Aud. Proc.*, 4(5), pp. 352-9, 1996.

[12] P. F. Assmann & Q. Summerfield, “Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies,” *J. Acous. Soc. Am.* 88(2), pp. 680-697, 1990.