The Listening Machine:
Sound Source Organization for Multimedia Understanding

Daniel P.W. Ellis, Electrical Engineering, Columbia University

Identifying the individual sources present in a real-world sound recording is difficult: Almost without exception, sounds of interest are embedded in a context of competing sounds, and it is rare to be given an unobstructed view of an ideal, isolated target. Human listeners, in common with other auditorily-equipped animals, are adept at handling such mixed signals, but our best computational audition systems — for instance automatic speech recognizers — are highly vulnerable to added interference, even at levels that listeners barely notice.

This program is about developing algorithms and systems for the analysis of sound mixtures in the context of automatic multimedia scene analysis. In comparison with video and image analysis, there has been little work on the general problem of organizing everyday sounds into the objects and events perceived by listeners. Unlike noise-robust speech recognition, which seeks simply to minimize the impact of the nonspeech components on the derived signal features, sound organization involves identifying and separately characterizing each significant contribution in a sound.

Central to the proposed approach is the idea of *sound fragment recognition*: Although a sound mixture may not afford unobstructed views of an entire sound source (voice, telephone ring, musical instrument), there will often be limited 'glimpses' in time and frequency when part of the signal can be observed relatively undistorted. By identifying and grouping such regions, and employing recognition algorithms modified to make correct classifications under incomplete observations, the combined sound mixture can be successfully interpreted as the combination of different, separately-modeled sounds.

Technical advances in pattern recognition and machine learning, increasingly illuminating results from neurophysiology and experimental psychology, and enormous advances in available computation power, make this a particularly profitable time to advance automatic sound organization, while the rapid growth in the volume of available multimedia content makes it urgent for better automatic content-based analysis tools to become available, e.g., to be able to find a particular event in a recording of the previous month's day-to-day activities.

The impact of this project includes the education of the graduate students who will conduct the research and other students who will pursue projects at the Laboratory for Recognition and Organization of Speech and Audio (LabROSA). This new lab will offer a unique agenda of analysis and abstraction of sound in all its forms from music to meetings. Results and illustrations will be transferred into courses at both undergraduate and graduate level, and used as demonstrations at Open House events aimed at the diverse local high school population in New York City.

The ability to analyze and describe complex sounds in terms of the different events they contain will facilitate a wide range of novel applications in content-based multimedia indexing, machine perception systems for robots interacting in human environments, and prosthetic devices for the perceptually impaired. These eventual outcomes will have an enormous impact on a broad section of the general public.

The Listening Machine:
Sound Source Organization for Multimedia Understanding

Daniel P.W. Ellis, Electrical Engineering, Columbia University

1 Introduction: The problem of machine listening

> *Imagine that you are on the edge of a lake [...] Your friend digs two narrow channels up from the side of the lake [...] Halfway up each one, your friend stretches a handkerchief [...] As waves reach the side of the lake, they travel up the channels and cause the two handkerchiefs to go into motion. You are allowed to look only at the handkerchiefs and from their motion to answer a series of questions [such as] How many boats are there on the lake and where are they? [...] Solving this problem seems impossible, but it is a strict analogy to the problem faced by our auditory systems.*
> Albert Bregman, *Auditory Scene Analysis* (Bregman 1990), pages 5–6.

The sense of hearing exists because it provides information that helps an organism to adapt and thrive, by identifying opportunities and threats in the environment. In order to analyze acoustic information, the information specific to each independent active sound source must first be separated, yet the overwhelming majority of research in sound recognition — typified by work in speech transcription — has assumed that the target sound is isolated, i.e. it is the only significant component in the received acoustic field. This may be a reasonable assumption when the signal comes from a head-worn microphone, but it is almost never true in a natural environment, where a distant source of interest will be competing with other simultaneous distractor sources.

> *To survive in such a densely populated world, the organism must be capable of segregation, [which] allows a cat to hear the faint sounds made by a mouse in the rustling grass, and might be of use to the mouse in that same situation [...] The hypothesis explored here is that the auditory system is in large part designed for that task.*
> Alain de Cheveigné, "The Auditory System as a Separation Machine" (de Cheveigné 2000).

This proposal consists of a program of research and learning to establish a radically new approach to extracting information from sound — one which, like the auditory systems of human and other animals, treats the organization of the received sound into features attributable to different sources as a central and indispensable aspect of the sound processing problem. As discussed below, this results in a problem that is significantly more complicated than the traditional speech transcription or acoustic event classification paradigm, yet it is inescapable: a sense of hearing that ceases to work when more than one sound source is active has little or no practical use.

1.1 Recognizing sound mixtures

> *Machine recognizers [...] cannot currently be used in normal environments without a close-talking or directional microphone or without using a push-to-talk switch because*

> *desired speech inputs cannot be separated from other environmental sounds. Common transient and intermittent environmental sounds may be interpreted by many modern high-performance recognizers as well-formed sentences.*
> Richard Lippmann, "Speech recognition by machines and humans" (Lippmann 1997).

The current state of sound recognition, in which it is assumed that signal segmentation is unnecessary because only a single source is present, has arisen because even given this simplification, the problems of sound recognition, such as classifying the myriad realizations of a given word as representing the same item, have proven to be very challenging (Gold and Morgan 2000). Moreover, preliminary investigations have revealed the problem of separating signals, such as reducing background noise in a speech signal, to be very hard (Cooke and Ellis 2001), even though people perform it effortlessly.

Recent advances in the basic speech recognition mechanisms mean that for many applications the interference of competing noise is the most important challenge. Unlike the studio-quality read-speech of the early 1990s, speech recognition tasks devised in the past few years have involved high levels of real-world background noise as a major feature (Pearce 1998; Singh, Seltzer, Raj, and Stern 2001), and have exposed the vulnerability of the 'all one source' assumption of the current recognition paradigm.

At the same time, computational models of the kinds of source-separation tasks performed by listeners (known as Computational Auditory Scene Analysis or CASA) have been increasing in sophistication. For example, recordings of music involving two or three instruments can now be transcribed into score-like representations with usable levels of accuracy (Goto 2001; Klapuri 2001). An increased understanding of the psychological basis of source segregation in listeners, coupled with an improved appreciation of how these processes can be simulated by and related to algorithms, means that the time is ripe for a concerted effort to find practical, general-purpose approaches to making sense of complicated sound mixtures.

1.2 The need for sound organization

Although speech recognition provides the most visible example of the sound mixture problem, the basic ability to organize sound scenes is necessary for any kind of device that aims to approximate an intelligent, human-like response to an unconstrained natural environment. This includes future portable and autonomous devices that may need to distinguish between spoken commands and other possible sounds, or to adapt to their changing contexts based on their senses, just as animals do. The envisaged sound organization technology would enable a 'semantic hearing aid' that could provide a textual or otherwise re-represented version of the ongoing acoustic environment for hearing impaired individuals, alerting them to situations they might not notice (Goldhor 1992).

The same process of converting a complex sound into a high-level abstract description can substitute for the human annotators (such as documentary filmmakers) who, at present, must review audio-visual material in order to provide a searchable index. Even a relatively crude automatic solution to this problem would make large archives of media content — whose limited interest makes the economics of human annotation infeasible — useful and available for scholars and other interested parties.

2 Research

As a concrete target on the path to a full simulation of a human listener, our goal in this project is a system that can recognize a wide range of sound-events mixed in with noise, voices, music etc. The sound events to be recognized will, at least initially, be based on sets of training examples, but the recognition must succeed despite enormous differences between the *context* of training and test examples. Our principle target application will be the identification of salient events in large audio archives for description (summarization) and indexing (retrieval). We will work with recordings of 'personal space', continuously collected by a portable audio memory aid, and with the soundtrack of video content such as the corpus used in the TREC Audio-Video spoke (NIST/TREC 2001). Hence, the most significant technical problem is the question of finding an efficient way to apply the existing tools of pattern recognition when the combinatorics of multiple sound sources make enumeration intractable, but when the overlap of the different sound signals preclude direct recognition of each one.

2.1 Background

Work on automatic recognition for real-world nonspeech sounds has taken a global-features approach rather than trying to isolate the properties of individual sources. Muscle Fish (Wold, Blum, Keislar, and Wheaton 1996) developed an early content-similarity-based browser for sound effects, and examples of similar work include applications to similar general sound databases, (Li 2000; Zhang and Kuo 2001), vehicle noise (Couvreur and Bresler 1998), and machine sounds (Atlas, Ostendorf, and Bernard 2000).

Soundtrack segmentation into a few classes (such as speech/music) has been important for speech recognition applications. Various features and classification schemes have been proposed in (Saunders 1996; Scheirer and Slaney 1997; Siegler, Jain, Raj, and Stern 1997; Chen and Gopalakrishnan 1998) among others; we have also worked in this area, tightly coupling classification to the speech acoustic model (Williams and Ellis 1999).

Work on separating simultaneous sound sources has been pursued from the perspective of modeling the perceptual phenomena described in psychology (Bregman 1990), which is known as Computational Auditory Scene Analysis (CASA) (Cooke and Ellis 2001). Most often, signal cues of harmonicity (e.g. for voiced speech) and common onset (across frequency channels) are used to group together time-frequency cells apparently relating to the same source. Our work has investigated the use of top-down 'prediction-driven' constraints (Ellis 1996; Ellis 1999).

CASA is often contrasted with blind source separation through Independent Component Analysis (ICA) in which a parameterized separation algorithm is adjusted to maximize the statistical independence of the 'unmixed' outputs (Bell and Sejnowski 1995; Hyvärinen and Oja 2000). ICA's elegant simplicity is also a weakness, in that more esoteric, arbitrary constraints (such as prior source models) and the need for multiple alternative solutions cannot easily be incorporated.

2.2 Technical Approach

Our approach starts with the idea of a set of models of the individual sounds that can occur in our mixtures, similar to the models used in speech recognition. The classical application of statistical pattern recognition is to find a maximum a-posteriori probability fit across a range of class models M_i to a set of signal features $\mathbf{X}$ i.e. the features are interpreted as an instance of the model M^*, where:

$$M^* = \underset{M_i}{\operatorname{argmax}} P(M_i|\mathbf{X}) \tag{1}$$

Rearranging via Bayes' rule gives:

$$M^* = \underset{M_i}{\operatorname{argmax}} \frac{P(\mathbf{X}|M_i)P(M_i)}{P(\mathbf{X})} \tag{2}$$

where the prior term $P(\mathbf{X})$ does not vary across the models, so can be dropped. Individual classes M_i are thus represented by distribution models $P(\mathbf{X}|M_i)$, which are a convenient way to represent prior class knowledge: The feature values observed in a set of training instances are generalized, typically as Gaussian mixture models (GMMs). The implicit assumption is that observations at classification time will be fully and directly comparable to the training examples on which the distribution models are based.

In the case of sound mixtures, however, any single target sound may appear in an infinite variety of acoustic contexts, formed by different combinations of different background sounds. We may describe this mathematically by defining a new variable, $\mathbf{Y}$, as the actual feature observations of the total, compound mixture, and our classification problem becomes:

$$M^* = \underset{M_i}{\operatorname{argmax}} P(M_i|\mathbf{Y}) = \underset{M_i}{\operatorname{argmax}} P(\mathbf{Y}|M_i)P(M_i) \tag{3}$$

One approach to recognizing the target sound buried in a mixture is to directly train distribution models, $P(\mathbf{Y}|M_i)$, to include the 'typical' effects of background sounds, either by training on noisy tokens (the 'multicondition training' paradigm used, for instance, in the Aurora task (Pearce 1998)), or by synthetically combining model representations of clean targets with models of isolated interference sounds to predict the appearance of various possible forms of corruption (known variously as 'HMM decomposition' (Varga and Moore 1990) or 'parallel model combination' (Gales and Young 1993)). It is, however, difficult or impossible to construct a training corpus with any kind of generality: not only are an uncountable number of possible background sound objects, but when combining models the absolute signal level can no longer be simply normalized away: instead, any pair of sounds must be modeled at an enumerated range of relative levels. This combinatoric explosion results in models that are either overly broad (because a single model is being made to stand for a broad range of noise or levels), or prohibitively expensive to create and to use (because a very large number of individual models must be tested).

The alternative approach is retain $P(\mathbf{X}|M_i)$, the clean feature distribution model, as the basic representation of each source, but to further model the relationship between the clean features $\mathbf{X}$ and the compound observations $\mathbf{Y}$. In general, we can integrate the $P(M_i|\mathbf{Y})$ term in equation 3

over the unknown values of $\mathbf{X}$:

$$P(M_i|\mathbf{Y}) = \int P(M_i, \mathbf{X}|\mathbf{Y}) d\mathbf{X} \quad (4)$$

$$= \int P(M_i|\mathbf{X}, \mathbf{Y}) P(\mathbf{X}|\mathbf{Y}) d\mathbf{X} \quad (5)$$

The first term in the integral reduces to $P(M_i|\mathbf{X})$, since the value of total observation $\mathbf{Y}$ is immaterial given the target features $\mathbf{X}$. To express this in terms of our original distribution model, $P(\mathbf{X}|M_i)$, we can apply Bayes' rule to this first term to give:

$$P(M_i|\mathbf{Y}) = P(M_i) \int P(\mathbf{X}|M_i) \frac{P(\mathbf{X}|\mathbf{Y})}{P(\mathbf{X})} d\mathbf{X} \quad (6)$$

(Note in this case that $P(\mathbf{X})$ is not a constant, and cannot be dropped.) In this form, the relationship between target source features $\mathbf{X}$ and composite observations $\mathbf{Y}$ is defined by $P(\mathbf{X}|\mathbf{Y})/P(\mathbf{X})$, the *change* in the likelihood of a particular value of $\mathbf{X}$ given knowledge of $\mathbf{Y}$.

When the multidimensional distributions $P(\mathbf{X}|M_i)$ are represented as mixtures of diagonal-covariance Gaussians (GMMs), the likelihood of each mixture component can be calculated as the product of the likelihoods of the individual feature dimensions, e.g., for a mixture of Q Gaussians indexed by k, over P dimensions indexed by j, we have:

$$P(\mathbf{X}|M) = \sum_{k=1}^{Q} P(k|M) \prod_{j=1}^{P} P(x_j|k, M) \quad (7)$$

where x_j is a scalar element of the feature vector. Assuming a similar decomposition of the prior $P(\mathbf{X})$, we can use this to decompose equation 6 to give:

$$P(M_i|Y) = P(M_i) \sum_{k=1}^{Q} P(k|M_i) \prod_{j=1}^{P} \int P(x_j|k, M_i) \frac{P(x_j|\mathbf{Y})}{P(x_j)} dx_j \quad (8)$$

where each $P(x_j|k, M_i)$ is a simple unidimensional Gaussian. The relationship between observed and target features has thus been decomposed to a likelihood change of the individual target feature elements due to the observations, $P(x_j|\mathbf{Y})/P(x_j)$

Even with this factorization, evaluating the full integral over every dimension of $\mathbf{X}$ will be tractable only under certain special conditions. In the 'missing data' approach to speech recognition (Cooke, Green, Josifovski, and Vizinho 2001), it is assumed that some elements of the observation feature vector are likely to be dominated by the target sound, thereby enabling at least part of the clean-signal model to be used unmodified. This is a good match to the situation if our features are spectral energies: Many sounds concentrate their energy at any moment into a few frequency bands (such as the formants in speech), and these bands can 'poke through' the energy of background sounds to permit largely unobstructed observations of those parts of the spectrum. This is in contrast to the more commonly-used cepstral features, where a change in any single frequency band will, in general, change *every* cepstral coefficient.

Given a way to decide which observation elements reliably reflect the underlying target features, and which ones have been corrupted, we have several choices for evaluating the per-dimension integral in equation 8:

- For dimensions considered reliable, $P(x_j|\mathbf{Y})$ is a Dirac delta at the assumed value $\hat{x}$, so the integral reduces to

$$P(x_j|k, M_i)/P(x_j)|_{x_j=\hat{x}} \tag{9}$$

- If, for a particular element, the masking due to energy from other sound sources meant that nothing could be inferred about the underlying x_j, $P(x_j|\mathbf{Y})/P(x_j)$ would be unity, and the integral over the complete pdf $P(x_j|k, M_i)$ would also reduce to unity.

- Even if the target signal is masked at a particular frequency, we know the observed spectral energy at that frequency, and we can infer that the actual target energy is not more that this value. Thus a more accurate treatment of such dimensions is given by the 'bounded marginalization' approach (Cooke, Green, Josifovski, and Vizinho 2001), where $P(x_j|\mathbf{Y})$ is zero for x_j greater than some ceiling x_{max}. While the bounded distribution of x_j may be difficult to express, the ratio $P(x_j|\mathbf{Y})/P(x_j)$ can be given a simpler form: zero for $x_j > x_{max}$ and a constant value F for $x_j \leq x_{max}$, where F is the normalization constant that preserves $P(x_j|\mathbf{Y})$ as a true pdf, i.e.:

$$F = \frac{1}{\int_{-\infty}^{x_{max}} P(x_j) dx_j} \tag{10}$$

  which is simply a lookup of a value of the error function erf when $P(x_j)$ is a Gaussian.

- More complex assumptions about the relationship between $\mathbf{Y}$ and x_j can be accommodated through other relationships. For instance, in 'soft missing data' (Barker, Green, and Cooke 2001), the true value of x_j is taken to be close to the masking ceiling in regions adjacent to unmasked energy.

Thus, we see that using spectral feature models, a simple masking assumption, and some mechanism for distinguishing between masked and unmasked elements, the model likelihoods in equation 8 can be evaluated in most cases with high computational efficiency.

The outstanding question is how to identify which frequency channels should be considered as reliable, and which to treat as corrupt. Here, too, there are several alternatives:

- We could infer a simple model of the interference, for instance by estimating a fixed 'noise floor' in each frequency band. Energy that exceeds the floor is taken as belonging to the target sound. This is the approach taken in the basic 'missing data' approach to recognizing speech in noise (Cooke, Green, Josifovski, and Vizinho 2001).

- We could treat the reliable/unreliable segregation as part of the inference problem, i.e. solve for the largest posterior value $P(M_i, S|\mathbf{Y})$, where S represents a particular segregation hypothesis, indicating the elements of source model $\mathbf{X}$ believed to have been reliably observed. This approach is taken in the 'speech fragment decoder' described in (Barker, Cooke, and Ellis 2002).

- We could use information from the observed signal, perhaps beyond that being modeled by $P(\mathbf{X}|M_i)$, to indicate which parts of the signal belong to different sources. This is one way to introduce the organization cues of Computational Auditory Scene Analysis (CASA) into a probabilistic sound-understanding framework (Barker, Cooke, and Ellis 2001).

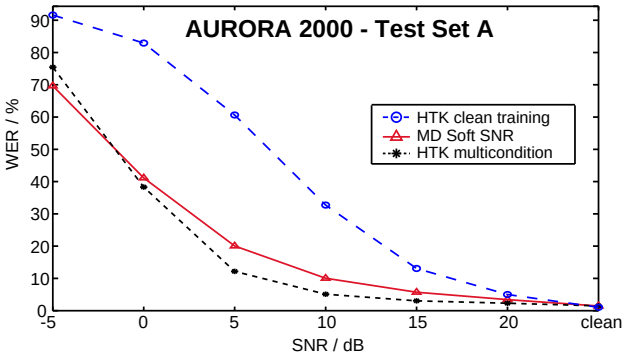

Figure 1: Word error rate vs. signal-to-noise ratio for several approaches. "HTK clean training" uses conventional modeling and recognition, trained on clean data only. "HTK multicondition" uses conventional models, but where the training examples have been mixed with noise similar to that used in the test conditions, at similar levels. "MD Soft SNR" uses models trained on clean data only, but using a 'soft' variant of the missing-data recognition to do the classification.

2.3 Preliminary results

2.3.1 Missing data speech recognition

Figure 1 shows example results from the missing-data approach used for the standard Aurora noisy digits task (Pearce 1998). In this task, fluent digit strings (e.g. "eight one seven three oh") are artificially mixed with various real-world noise backgrounds (restaurant, car, airport etc.) at a range of signal-to-noise ratios (SNRs). There are two alternative training sets: the "clean" set consists only of the digits, to test how well systems can deal with completely unanticipated noise; the "multicondition" training set includes training examples mixed with noise at a range of SNRs from clean to 5 dB. The test set consists of distinct digit strings mixed with four noise types at seven SNRs (clean to -5 dB) for a total of 28 conditions (separate test sets include noise less similar to the noises in the multicondition set, and channel coloration). The task also specifies a 'baseline recognizer' built from the well-known HTK toolkit, using a standard (but somewhat optimized) set of features and parameters.

The missing-data system used spectral features (instead of the Mel cepstra of the baseline) and estimated a static background noise level from the first 100 ms of each sound file in the test set; time-frequency cells significantly above the noise floor were taken as reliable, those below were subjected to bounded marginalization, and cells whose energy was within a few dB of the estimated noise floor made a 'soft' contribution to overall likelihood, calculated as a linear mix of reliable and masked estimates (Barker, Green, and Cooke 2001).

Figure 1 shows that using the same clean-data models as the baseline recognizer, missing data recognition achieves a substantial reduction in the word error rate for higher signal-to-noise ratios, bringing performance close to that achieved by multicondition training — but, unlike the multi-condition system, without any prior knowledge of the corrupting noise styles, making it far more robust to variation in test conditions.

In this system, the connection between observed features ($\mathbf{Y}$ in the exposition above) and target feature ($\mathbf{X}$) is made via the bounded missing data assumption, and the fixed noise floor model. In

subsequent work, we have investigated additionally comparing different segregation choices to find the most likely one, with some promising results (Barker, Cooke, and Ellis 2002).

2.3.2 Alarm sound detection

A second preliminary investigation into the fragment-recognition approach focused on the alarm sounds generated by many man-made devices such as telephones, sirens etc. (Ellis 2001). Alarm sound recognition would be useful in a portable warning device for hearing impaired users, but only if it is able to operate with very unfavorable signal-to-noise ratios. Alarm sounds are designed to be easily heard and readily recognized, making the task less daunting.

With no prior work, we had to build our own database and our own baseline system for comparison. A collection of 50 example alarms was assembled from CDs and the web, including car horns, fire alarms, doorbells and telephones. For the test set, these examples were mixed with a range of real background sounds at 0 dB SNR (equal power in alarm and noise during the 'active' alarm segments). Since alarms often have sparse, sustained spectra, they were usually quite easily to hear against the noise.

The baseline system used a global-feature approach, consisting of a multi-layer perceptron (MLP) classifier trained to discriminate between time windows with and without alarms. The training set was a collection of alarms in noise, similar to the test examples (although the noises and alarms were different).

The second, fragment recognition system operated as follows: First, the sound was subject to a time-frequency analysis emphasizing the narrow, sustained harmonics typical of alarm sounds. These concentrations were then represented with sinusoid models, and tracks starting at about the same time and with similar shapes were grouped together into composite objects. In this way, the preprocessing approximated the psychoacoustic grouping principles of common onset and common modulation.

Each grouped object was summarized by parameters such as average frequency variation, average magnitude decay, amplitude modulation depth etc. This representation relies implicitly on partial observations of the target alarms — the discrete frequency peaks — and is largely independent of the background noise level, until the alarm sound is entirely buried. These parameters are then passed to a classifier based on properties extracted from the training set to decide if the object is an alarm.

The outputs of the two alarm detection systems are illustrated in Figure 2. The output of the MLP classifier shows a rapid variation in the classification of individual, short frames; smoothing this result and median filtering yields detected alarms, shown by the thick boxes. Because the neural network can only generalize across the noise backgrounds used in training, it is vulnerable to the many false alarms seen here. The fragment-recognition system exhibits many fewer false alarms, but has also missed one of the true alarms.

In a complete test presenting the 25 test alarms each in four different noise backgrounds, the overall error rate of both systems was large — 192% for the MLP system, and 197% for the sound-object based system. The breakdown by error type and noise conditions revealed interesting differences: the MLP system made half as many false-rejects as the object-based system (22% vs 50%), which made more than 90% of its false-alarms (insertion errors) against a pop-music

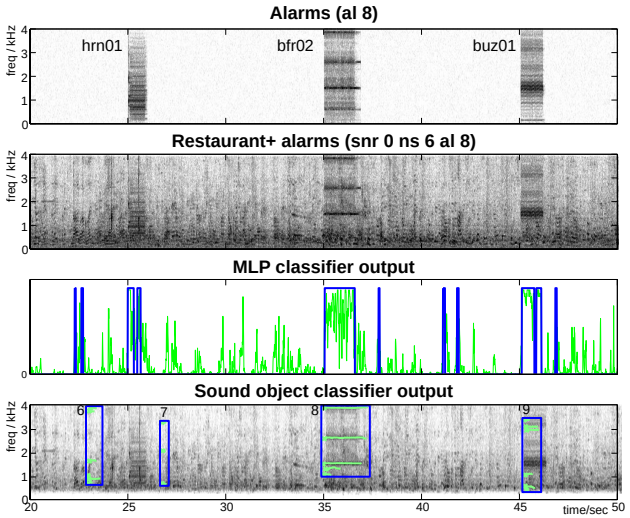

Figure 2: Example of the alarm detection systems. Top panel shows spectrogram of three example alarms from the test set; the second panel has 'restaurant' noise added at 0 dB SNR. The third panel shows the output of the MLP frame-level classifier, which detects the alarms but also experiences numerous false alarms due to differences between the restaurant noise and the noises used in its training set. The bottom panel shows the sound objects classified as alarms by the preliminary fragment-recognition system. There are many fewer false alarms, although one of the true alarms has been missed. Blue boxes outline detected alarm events in bottom two panels.

background, which contained many sustained pitches to trigger the alarm object detector (Ellis 2001).

In this example, noisy observations are related to clean target models through the sound object extraction system, which rejects at an early stage any acoustic energy not broadly 'alarm-like' (narrow spectral peak, sustained frequency and magnitude). The system can in theory recognize even partial extraction of these already fragmentary descriptions, if the remaining, extracted harmonics still form a pattern whose composite characteristics fit the learned 'alarms' class.

2.4 Research plan

The research part of this program will focus on extending our preliminary results in fragment-based recognition to extract a higher degree of information from a wider range of sound sources and conditions. So far, we have shown results comparable to standard approaches for two special cases, but improved implementations, model representations, and particularly segregation cues (as discussed above) should reveal the dramatic improvements obtainable when the mixed nature of sound is directly addressed. Developing the core recognition engine will involve many sub-projects; our specific plans include:

- **Recognizing multiple sources at once:** In both preliminary tasks, the goal was to identify a single, specific target, and ignore the remainder. The analysis, however, is completely symmetric over the labeling of foreground and background; any number of simultaneous sources

can be simultaneously recognized, and segregating their features in a single operation should be more efficient and more accurate. (This draws on ideas developed in (Barker, Cooke, and Ellis 2002); see the attached commitment letter from Sheffield University.)

- **Better target model descriptions:** Our current models of the clean target examples (parameter distribution GMMs) are simplistic: they make little accommodation for large but highly-structured deviations from the training set that can be anticipated, for instance, the problem of absolute level variation. This becomes much more important when dealing with mixtures that cannot be simply normalized to a fixed energy level (since the target may not be the dominant signal, and thus normalization would adjust it to an unexpected level). Source attributes such as overall level and fixed or slowly-varying spectral coloration can be modeled as 'hyper-parameters', using Bayesian network/graphical model techniques of the kind demonstrated in (Jojic and Frey 2001; Beal, Jojic, and Attias 2002).

- **Online model acquisition:** Rather than predefining the model objects, a system that can define a set of objects to explain its composite objects would have a much more robust, human-like ability to organize sound. One approach is to identify sounds that happen to be high-SNR examples, and to cluster them into incrementally-refined 'concepts'.

As mentioned at the start of this section, the application area will be everyday sound environments, following on from the alarm sound detection work. We are collecting a database of real-world sound by carrying a portable recorder around in normal day-to-day circumstances. Summarization and browsing of this kind of data — already easy to collect, but currently next to useless — will be a major application example for this work.

A second application domain will be the soundtrack of videos, leveraging our close ties to the Digital Video / MultiMedia group in the department, which has a well-established reputation in image and video analysis and searching (Chang, Chen, Meng, Sundaram, and Zhong 1997; Zhong and Chang 1999). This collaboration will further involve industrial partner IBM, and will permit our involvement in the Audio-Video TREC spoke, a new evaluation for multimedia content search (NIST/TREC 2001).

Because the area of sound organization and understanding is so novel, there are no established sound-only evaluation metrics beyond speech recognition word error rate. Such standards are very helpful in interpreting one's own results, as well as in promoting a field, so we will devote a substantial portion of our effort to developing and promoting well-defined, general-purpose evaluation standards for sound organization tasks. Initially this will consist of finding known sound events in a test corpus, but as the usefulness of the technology is made clear through prototype auditory memory agents, evaluation will be based on simulations of practical tasks, like location ambiguously-described events.

3 Educational aspects

This project will support the development of a recently-established lab within Columbia's Electrical Engineering Department, the Laboratory for Recognition and Organization of Speech and Audio. **LabROSA** is currently taking shape, but needs secure support for the future. The lab is

founded on a unique vision of transferring the advanced statistical and machine-learning tools employed in speech recognition to the recognition of the full range of real-world sound, and of taking ideas from auditory modeling and perceptual sound organization to break through the logjams currently facing speech recognition. The lab is unique in this perspective, and also in its situation at Columbia, where it is both intellectually and physically adjacent to the ADVENT Digital Video / MultiMedia group of Prof. Shih-Fu Chang, widely recognized as a leading source of ideas and solutions in video content processing (Chang, Chen, Meng, Sundaram, and Zhong 1997), and also the Natural Language Processing group of Prof. Kathy McKeown, at the forefront of high-level analysis and summarization of language (McKeown, Klavens, Hatzivassiloglou, Barzilay, and Eskin 1999). The three groups have already been involved in several joint proposals for multimodal content analysis and access (Ellis, Chang, and McKeown 2000; Chang, Ellis, and McKeown 2001).

LabROSA has a commitment to supporting the research community by providing data and software. PI Ellis is the maintainer of a large neural network speech recognition package, the ICSI SprachCore, which has been used in numerous publications. Our website http://labrosa.ee.columbia.edu provides links to a variety of resources, including student- compiled literature reviews and a collection of Matlab implementations of various sound processing algorithms, developed in the lab. We consider the dissemination of working code examples to be of comparable importance to our published papers.

The immediacy to all people of sound in general, and music in particular, makes LabROSA a particularly effective conduit for introducing the department to a wider audience. Thus, we have been one of a small number of labs featured in three Engineering Open House days over the past year, where demonstrations of music deconstruction and similarity-based music recommendation truly engage the local high school students at whom these events are aimed (and who already associate music with computers thanks to MP3s). We hope to exploit the accessibility of our work to further promote the discipline of electrical engineering by motivating young students and the public at large.

As an example, we are planning to release a software 'plugin' for the popular music-playing program WinAMP to help people organize their music collections using the content-analysis techniques developed in our lab. In addition to its intrinsic value to the user, the plugin will, given the user's permission, report back anonymized statistics to help us evaluate the success of our algorithms, at the same time giving the individual a sense of involvement with and contribution to the academic research enterprise (Ellis, Whitman, Berezweig, and Lawrence 2002).

This proposal will provide support for **research mentorship** through one graduate research assistantship for the duration of the project. If all goes according to plan, this would be taken by one of the current lab members who is currently being supported through teaching assistantships and startup funds. Since he is already a year into his graduate studies, the goal would be for him to graduate in the third year of the project, and to bring in a new graduate student for the remainder.

These students, along with the other students based in LabROSA, will gain an introduction to the academic research community, including both technical skills of structuring and conducting experiments (and how to develop and exploit the tools required), and the socialization aspects of how best to contribute as part of the scientific community through the development and sharing of resources, participation in and organization of meetings, reviewing, and of course writing publications and making presentations. All LabROSA students are closely involved in developing the

direction of their own research projects, rather than simply following directions or acting purely as 'assistants'.

Other LabROSA members include visitors: I am currently hosting a graduate student on a one-year secondment from University College, Dublin, as well as a shorter visit from the student of a colleague at the Oregon Graduate Institute. A separate pending proposal seeks funds to support a postdoc, and I have an international candidate in mind, a contact made through my past involvement in a series of international collaborations sponsored by the European Union (Robinson, Cook, Ellis, Fosler-Lussier, Renals, and Williams 2002).

Since starting at Columbia, I have been responsible for two **courses**: the undergraduate/masters' level introductory Digital Signal Processing (DSP) course, and a new graduate course of my own design, Speech and Audio Processing and Recognition. Both these courses have been developed according to 'open courseware' principles, with all course materials (lecture notes, assignments, exams, solutions etc.) being distributed via web sites that are universally and permanently available from http://www.ee.columbia.edu/~dpwe/. While I have not actively publicized these materials, I have happily granted several requests from far-flung individuals who have found them via web search engines, and who wish to re-use or adapt some of the material.

In both courses, I have pursued the educational philosophy of balancing and interrelating theoretical and practical aspects. Both courses involve weekly computer-based assignments, relying on the excellent interactive data manipulation package Matlab to enable students to experiment with the techniques discussed in class when applied to real data. True facility with the mathematical and algorithmic tools of signal processing and recognition requires both a clear understanding of the theoretical basis *and* sufficient practical experience to develop a proper intuition. It can take years to fully develop such an intuition, but by giving the students the ability to go away and experiment on their own it is possible to start them on this path. Both courses also involve a final project, for which I require an actual, practical component, preferably in Matlab.

For the undergraduate **Digital Signal Processing** class, sound forms an excellent demonstration medium, because it is one dimensional (and therefore simpler to conceptualize than images or video), and because the process of 'listening' to the effects of, say, bandpass filtering or quantization makes a deep impression and gives a visceral, alternative way to understand a mathematical concept. My computer-generated sound examples have attracted particularly positive comments from students.

The current program will allow me to further enhance props of these kind by developing more sophisticated special-purpose demonstrations. As an example, I currently use a crude animation to illustrate the process of cross-correlation to detect a particular pulse shape buried in noise; by building up the cross-correlation function one point at a time, and showing (via computer projection) the alignment of the correlation kernel with the noisy signal resulting in each point in the correlation, insight into the process is obtained. The Matlab commands used to generate this display are distributed via the website; however, my goal is to build a more polished demonstration, making the interaction and experimentation more inviting and more flexible, while still retaining the transparency of a Matlab-based implementation.

The graduate level **Speech and Audio Processing and Recognition** class is directly related to the technical content of the research in this proposal. As the only advanced course on audio processing currently offered by the department, the syllabus ambitiously includes everything from

music compression to speech database querying. My goal has been to cover a different broad topic (e.g., psychoacoustics, music synthesis and analysis) in each week's lecture, including a single detailed investigation of one area (for instance, simultaneous masking or sinusoid modeling) within each lecture. Each class is complemented by a reading and a Matlab practical tailored to the current topic. Support under this program would facilitate the development of more custom practicals for this course, often involving direct transfer of the research tools and results.

An example will illustrate the two-way flow between research and coursework: A current research project (see section 7) involves recording of real, multi-participant meetings with several tabletop microphones. I included some of this data among the set of 'project data' offered to students of the DSP class. A team of two talented students worked with this data to infer speaker locations in a 2-D plane from channel cross-correlations — a project I had been planning to pursue with a graduate student. I continued this work with one of the students as independent research for another semester (one of 8 such projects I have supervised in the past two years), to extend the localization to 3 dimensions; the work will be submitted to ICASSP this year. The results were so interesting that I converted them into a very successful practical for my Speech and Audio class.

Developing a good practical, like preparing a good lecture, is intricate and time consuming. I have been fortunate to be able to borrow some courseware developed by colleagues: The Matlab Auditory Demonstrations from Sheffield (Cooke, Parker, Brown, and Wrigley 1999) formed the core of a practical on auditory perception, and my European project colleagues from IDIAP in Switzerland graciously let me use their practical introducing hidden Markov models. In turn, I have developed a practical on sinusoid modeling into a tutorial, available from my web site, which several colleagues have told me they found useful. Sharing practicals benefits both recipient and donor, not least in the form of contributed enhancements.

Columbia's School of Engineering, and the Electrical Engineering department in particular, is actively engaged in developing new mechanisms to **evaluate the effectiveness of educational activities**. Course assessments are web-based, affording consistency, anonymity, and automatic statistical analysis; by emphasizing the importance of these assessments, I have achieved 80-100% participation among students in my classes. By tracking quantitative ratings in categories such as quality of lectures, relevance of material, and appropriateness of workload, an objective measure of the effect of incremental changes introduced into a course can be gained. The department also conducts formal, standardized surveys of its graduating students to gauge both their overall experience in the program, and any comments specific to individual aspects or classes.

While clearly at home within Electrical Engineering, LabROSA has many **interdepartmental links**, most obviously to Computer Science, but also to other disciplines including Psychology and Music. A major advantage of Columbia's wide range of schools and departments is to facilitate connections and interactions of this kind; several longer-term curricular developments to be pursued under this program involve co-ordinated developments with faculty in other departments, and Columbia will be unique in offering this combination of opportunities.

Computer Science has a large Natural Language Processing group (led by Kathy McKeown), and we are already collaborating on the Meeting Recorder project described below. This fall, Julia Hirschberg is joining CS as senior faculty to start a group researching spoken language processing and dialog systems, and we have begun discussions of an integrated program in speech and language, which could also involve psychology faculty such as Robert Remez. Columbia also has a

highly-regarded, and engineering-oriented Computer Music Center, with whom I am already collaborating to the extent of trading guest lectures. The goal is to divide my currently overstuffed Speech and Audio class into two **new classes**: one containing the speech material, and with less emphasis on signal processing, to suit the joint EE/CS program in speech and language. (To make sure this course is fully accessible to CS students, we may develop a brief signal processing 'crash course' for students without an EE background.)

The second course will consist of the remaining topics in audio signal processing, and will be made attractive to Computer Music students; one way to accommodate students with different backgrounds is to recognize the different kinds of projects that will be completed in each case. However, in my experience, a range of perspectives in the classroom leads to positive interations and student involvement.

Also in Computer Science, Tony Jebara, who joined Columbia this year, conducts research in machine vision as a basis for behavior and action; machine listening forms an obvious complement to this work. I am very interested in using his advanced machine learning techniques, so we are currently discussing collaboration both in research and in classes — for instance, to introduce audio processing into the practical aspects of his class on machine perception.

4 Diversity

I have a deep idealogical commitment to social responsibility in engineering, and one aspect of this is an effort to promote and support diversity in LabROSA and in the department and school in general. One of my five graduate students is female, which is too low, but better than the department average of under 10% female, something we are moving to improve. My first graduate student, Manuel Reyes, is Hispanic, and I am very pleased to be supporting him through his Ph.D. I work hard to come across as approachable in class, particularly to encourage the less confident students; of the seven students for whom I have written letters of recommendation, two were women. I am currently working with an African American female student who is retaking my undergraduate DSP course this summer to complete her degree requirements. I have made LabROSA a fixture in the Engineering Open House events for local high-school students, and I make a point of sitting with students least like our current student body at the lunches associated with these events, to encourage their emerging interest in engineering.

5 Timeline

The anticipated sequence of activities in this program is broadly as follows:

Year 1: Construct baseline fragment-based recognition systems for detecting alarm sounds and other events in everyday sound, demonstrating the fundamental advantage over global-feature approaches. Develop and enhance in-class demonstrations and practicals, and actively pursue sharing them with colleagues at parallel institutions.

Year 2: Collaborate to define and collect evaluation datasets and criteria for the nonspeech recognition work, with the goal of establishing widely-adopted standards. Investigate integration of information from other modalities for use in Audio-Video TREC (NIST/TREC 2001) in

collaboration with IBM (see attached commitment letter). Adapt the Speech and Audio class to co-ordinate with Computer Science courses in speech and language processing, including a new 'signal processing crash course' for CS students.

Year 3: Graduation of student working on fragment recognition; recruit a follow-on. Organize a special session at a major conference to promote these new results and to promote interaction with related researchers. Further development of the graduate Speech class to co-ordinate with CS machine perception classes through shared projects etc.

Year 4: Consider problem of acquisition and generalization of nonspeech sound models from unsupervised data. Transfer results from research into classroom courses in the form of demonstrations, datasets, and basic concepts. Develop a complementary Digital Audio Processing class for EE and Computer Music students, containing material excluded from the Speech class.

Year 5: Revise evaluation tasks to accommodate new aspects of model acquisition (i.e. much more data, but much less annotation). Review progress of integrated CS/EE speech and language courses to see how best to divide the material between courses and faculty.

6 Conclusion

This program addresses the hitherto neglected problem of recognizing the individual components in an everyday sound ambiance. By adapting and developing missing-data techniques already shown to be effective in recognizing speech corrupted by noise, we will develop systems able to describe multimedia content in terms that match users' impressions.

The immediate beneficiaries of this work will be students: those directly involved in the research project, other students doing projects at LabROSA who will share the ideas and resources emerging from the project, and students participating in the various classes, existing and newly-developed, mentioned above. Successful solutions to the problem of general sound understanding will have a much broader impact in the form of new applications in multimedia content description, intelligent interactive machines, and perceptual prostheses that, in addition to enhancing lives, will also promote the status and allure of engineering throughout society.

7 Results from prior support

PI Ellis is a co-PI on one current NSF project: NSF IIS-0121396 ($1,402,851), Title: ITR/PE+SY: Mapping Meetings: Language Technology to make Sense of Human Interaction, Award period: 2001-09-01 to 2005-08-31, PI: Nelson Morgan, International Computer Science Institute. This project is concerned with the application of speech recognition and other automatic signal analysis techniques to extracting information from recordings of natural, unconstrained meetings between human participants. The one graduate student in LabROSA supported on this grant is currently looking at unsupervised clustering of the extensive data so far collected as a way to define and locate "interesting" events in the recordings, and there will likely be transfers of these techniques into the current project. Publications arising from this work have yet to be completed, although the initial data collection and general goals are described in (Morgan, Baron, Edwards, Ellis, Gelbart, Janin, Pfau, Shriberg, and Stolcke 2001).

References

Atlas, L., M. Ostendorf, and G. D. Bernard (2000). Hidden markov models for monitoring machining tool-wear. In *Proc. ICASSP-2000*.

Barker, J., M. Cooke, and D. P. Ellis (2001). Combining bottom-up and top-down constraints for robust asr: The multisource decoder. In *Proceedings of the CRAC-2001 workshop*. http://www.ee.columbia.edu/crac/papers/barker.pdf.

Barker, J., M. Cooke, and D. P. Ellis (2002). *Speech Communication*. Submitted.

Barker, J., P. Green, and M. Cooke (2001). Linking auditory scene analysis and robust asr by missing data techniques. In *Proc. Workshop on Innovation in Speech Processing WISP-2001*. http://www.dcs.shef.ac.uk/research/groups/spandh/projects/respite/publications/barker_wisp_01.pdf.

Beal, M., N. Jojic, and H. Attias (2002). A self-calibrating algorithm for speaker tracking based on audio-visual statistical models. In *Proc. ICASSP-02*.

Bell, A. J. and T. J. Sejnowski (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation 7*(6), 1129–1159. ftp://ftp.cnl.salk.edu/pub/tony/bell.blind.ps.Z.

Bregman, A. S. (1990). *Auditory Scene Analysis*. Bradford Books, MIT Press.

Chang, S.-F., W. Chen, H. Meng, H. Sundaram, and D. Zhong (1997). Videoq - an automatic content-based video search system using visual cues. In *ACM Multimedia Conference*.

Chang, S.-F., D. Ellis, and K. McKeown (2001). Empowering video storytellers: Concept discovery and annotation for large audio-video-text archives. Proposal to the 2001 NSF-ITR program. http://www.ee.columbia.edu/~dpwe/proposals/ITR01-AVT.pdf.

Chen, S. and P. Gopalakrishnan (1998). Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*. http://www.nist.gov/speech/publications/darpa98/pdf/bn20.pdf.

Cooke, M. and D. Ellis (2001). The auditory organization of speech and other sources in listeners and computational models. *Speech Communication 35*(3–4), 141–177. http://www.ee.columbia.edu/~dpwe/pubs/tcfkas.pdf.

Cooke, M., P. Green, L. Josifovski, and A. Vizinho (2001). Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication 34*(3), 267–285. http://www.dcs.shef.ac.uk/research/groups/spandh/projects/respite/publications/cooke_speechcom_00.pdf.gz.

Cooke, M. P., H. Parker, G. Brown, and S. Wrigley (1999). The interactive auditory demonstrations project. In *Proc. Eurospeech-99*. http://www.dcs.shef.ac.uk/~martin/MAD/docs/mad.htm.

Couvreur, C. and Y. Bresler (1998, Jul.-Aug.). Automatic classification of environmental noise sources by statistical methods. *International Journal of Noise Control Engineering 46*(4), 1–16.

de Cheveigné, A. (2000). The auditory system as a separation machine. In *Proc. Intl. Symposium on Hearing*. http://www.ircam.fr/pcm/cheveign/sh/ps/ATReats98.pdf.

Ellis, D. (1999). Using knowledge to organize sound: The prediction-driven approach to computational auditory scene analysis, and its application to speech/nonspeech mixtures. *Speech Communications 27*(3–4), 281–298. http://www.icsi.berkeley.edu/~dpwe/research/spcomcasa98/spcomcasa98.pdf.

Ellis, D. (2001). Detecting alarm sounds. In *Proc. Workshop on Consistent and Reliable Acoustic Cues CRAC-2000*.

Ellis, D., S.-F. Chang, and K. McKeown (2000). The multimedia lexicon: Automatic object and structure discovery in audio-video-text content. Proposal to the 2000 NSF-ITR program. http://www.ee.columbia.edu/~dpwe/proposals/ITR00-mmlex.pdf.

Ellis, D., B. Whitman, A. Berezweig, and S. Lawrence (2002). The quest for ground truth in musical artist similarity. In *Proc. International Symposium on Music Information Retrieval ISMIR-2002*.

Ellis, D. P. (1996). *Prediction-driven Computational Auditory Scene Analysis*. Ph. D. thesis, MIT Dept. of Electrical Engineering and Computer Science. http://web.media.mit.edu/~dpwe/pdcasa/pdcasa.pdf.

Gales, M. and S. Young (1993). Hmm recognition in noise using parallel model combination. In *Proc. Eurospeech-93*, Volume 2, pp. 837–840.

Gold, B. and N. Morgan (2000). *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. John Wiley.

Goldhor, R. (1992). Environmental sound recognition. proposal to the National Institutes of Health by Audiofile, Inc.

Goto, M. (2001). A predominant-f0 estimation method for cd recordings: Map estimation using em algorithm for adaptive tone models. In *Proc. ICASSP-2001*.

Hyvärinen, A. and E. Oja (2000). Independent component analysis: Algorithms and applications. *Neural Networks 13*(4–5), 411–430. http://www.cis.hut.fi/aapo/papers/IJCNN99_tutorialweb/.

Jojic, N. and B. J. Frey (2001). Learning flexible sprites in video layers. In *Proc CVPR-01*. http://www.psi.toronto.edu/layers.html.

Klapuri, A. (2001). Multipitch estimation and sound separation by the spectral smoothness principle. In *Proc. ICASSP-2001*. http://www.cs.tut.fi/sgn/arg/music/icassp2001_klap.pdf.

Li, S. Z. (2000). Content-based audio classification and retrieval using the nearest feature line method. *IEEE Tr. Speech and Audio Processing 8*(5), 619–625.

Lippmann, R. (1997). Speech recognition by machines and humans. *Speech Communication 1*(22), 1–15.

McKeown, K., J. Klavens, V. Hatzivassiloglou, R. Barzilay, and E. Eskin (1999). Towards multidocument summarization by reformulation: Progress and prospects. In *Proc. 17th Nat. Conf. on Artif. Intel. AAAI-99*, pp. 453–460.

Morgan, N., D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke (2001). The meeting project at icsi. In *Proc. Human Lang. Techn. Conf.* ftp://ftp.icsi.berkeley.edu/pub/speech/papers/hlt01-mr.pdf.

NIST/TREC (2001). Guidelines for the trec-2001 video track. Technical report, National Institute of Standards and Technology. http://www-nlpir.nist.gov/projects/trecvid/revised.html.

Pearce, D. (1998). Aurora project: Experimental framework for the performance evaluation of distributed speech recognition front-ends. Technical report, European Telecommunications Standards Institute.

Robinson, A., G. Cook, D. Ellis, E. Fosler-Lussier, S. Renals, and D. Williams (2002). Connectionist speech recognition of broadcast news. *Speech Communication 37*(1–2), 27–45.

Saunders, J. (1996). Real-time discrimination of broadcast speech/music. In *Proc. ICASSP-96*.

Scheirer, E. and M. Slaney (1997). Construction and evaluation of a robust multifeature speech/music discriminator. In *Proc. ICASSP-97*.

Siegler, M. A., U. Jain, B. Raj, and R. M. Stern (1997). Automatic segmentation, classification and clustering of broadcast news audio. In *Proc. DARPA 1997 Broadcast News workshop*. http://www.nist.gov/speech/proc/darpa97/index.htm.

Singh, R., M. Seltzer, B. Raj, and R. Stern (2001). Speech in noisy environments: robust automatic segmentation, feature extraction, and hypothesis combination. In *Proc. ICASSP-2001*.

Varga, A. and R. Moore (1990). Hidden markov model decomposition of speech and noise. In *Proceedings of the 1990 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 845–848.

Williams, G. and D. Ellis (1999). Speech/music discrimination based on posterior probability features. In *Proc. Eurospeech-99*. ftp://ftp.icsi.berkeley.edu/pub/speech/papers/euro99-mussp.pdf.

Wold, E., T. Blum, D. Keislar, and J. Wheaton (1996). Content-based classification, search, and retrieval of audio. *IEEE Multimedia 3*, 27–36.

Zhang, T. and C.-C. J. Kuo (2001). Audio content analysis for online audiovisual data segmentation and classification. *IEEE Tr. Speech and Audio Processing 9*(4), 441–457.

Zhong, D. and S.-F. Chang (1999). An integrated system for content-based video object segmentation and retrieval. *IEEE Tr. Circuits and Systems for Video Technology 9*(8), 1259–1268.