

# FODAVA-Partner: Visualizing Audio for Anomaly Detection

## 1 Problem Statement

“Most people who handle money a lot (i.e. cashiers) can identify a lower-quality fake bill instantly just by touching it” [10]. Data analysts are like cashiers: a trained data analyst can detect anomalies “at a glance” when data is appropriately transformed. This is the goal of data visualization.

This proposal addresses the type of audio anomalies that human data analysts hear instantly: angry shouting, trucks at midnight on a residential street, gunshots. The human ear detects anomalies of this type rapidly and with high accuracy. For example, rifle magazine insertion clicks are detected with 100% accuracy at 0 dB SNR in white noise, babble, or jungle noise [1]. Unfortunately, a data analyst can listen to only one sound at a time. Visualization shows the analyst many sounds at once, possibly allowing him or her to detect an anomaly several orders of magnitude faster than “real time.” This proposal aims to render large audio data sets, comprising thousands of microphones or thousands of minutes, in the form of interactive graphics that reveal important anomalies at a glance.

Precedents for such graphical rendering are familiar to audio professionals. A simple amplitude-versus-time graph reveals silences for a speech transcriber to skip past; a spectrogram reveals details of birdsong to an ornithologist.

But many audio anomalies are not so easy to display: angry shouts versus enthusiastically shouted greetings; the clatter of overturned tables rather than mere dishwashing; spoken Thai in Berlin, or German in Bangkok. All these cases can submit to automatic anomaly detection, using probabilistic models of long- and short-term spectral features of normal activity. Unfortunately, the state of the art in automatic audio event detection is not very accurate [112]. We propose to represent audio anomaly to the analyst using a type of overcomplete lossless encoding: automatic anomaly salience scores will be displayed together with raw audio features, allowing a human analyst to drill down at any point in the data in order to resolve discrepancies in the visible display.

The goal of this proposal is to present to analysts a coherent visual summary of both probabilistic and raw spectral information. The measurable outcome of this research will be the speed with which analysts find audio anomalies that have been planted, by the experimenter, in a very large dataset. A successful research outcome will be a visual summary that lets the analyst detect most anomalies immediately (about 10,000× faster than real time), and all anomalies after brief interactive exploration (about 1,000× faster than real time). In short, the goal of this proposal is to transform, model, and reduce data for efficient effective visualization and analytic reasoning:

- A simple time series is *transformed* into audio features and probabilistic model-based features, vastly reducing the quantity of data presented at one time to the analyst (fewer observations).
- Multiple techniques of dimensionality *reduction* condense the breadth of the data presented to the analyst (fewer variables).
- Computationally inexpensive multiscale caching of all layers, from summary variables down to the raw source data, supports *interactive* investigation of hypotheses at different spatial and temporal scales.
- The interactive visualizations are *efficient*: caching increases the analyst’s decision rate.
- The visualizations are *effective*: the decisions have a measurably low error rate.

## 2 Background

### 2.1 Audio Visualization

Audio professionals use sophisticated interactive tools to review and edit recorded or synthesized audio, while producing audio CDs or soundtracks for video. After the recording equipment has

been purchased, the cost per minute of recording is negligible. Thus, vast amounts of recorded data are assimilated by these professionals. Although they aim to cull, combine, and polish raw audio into a final soundtrack, their minute-to-minute work is not that different from an analyst looking for anomalies. Much of their expertise goes into making an “edit decision list” defining how the raw audio is to be combined, and much of that work is searching and comparing, not modifying and polishing. It is thus warranted to learn from the stable tools which have found favor in this industry, e.g. Pro Tools by Digidesign, Audition by Adobe (formerly Cool Edit Pro), and Soundtrack Pro (part of Apple’s video suite Final Cut Studio).

Academic audio researchers use visualization tools of a slightly different type, typified by the audio timeline editors Praat [11] and Audacity [77]. Besides the conventional amplitude-vs-time display, these can render spectrograms, cochleograms (frequency vs. time, linear and psychoacoustic respectively), pitch tracks, and tracks of spectral peak frequencies (formants) as a function of time. Scientific audio visualization tools usually provide more computable features than professional editing tools (e.g., pitch tracks or formant frequencies), but far fewer simultaneously viewable audio channels (typically 2 to 8) and less total data (typically one hour per edited waveform). Other scientific visualization tools, though not directly focused on rapid interactive exploration, still have components which can be exploited in our context. Primary among these are the Visualization Toolkit (“vtk”) with its front end ParaView, Lawrence Livermore National Laboratory’s VisIt, Wolfram’s Mathematica, and the Mathworks products including Matlab.

## 2.2 Audio Event Detection

This proposal draws on previous research in the fields of audio content analysis [22, 75, 94] and acoustic event detection [6, 18, 82, 91]. Research in audio content analysis has typically addressed the problem of segmenting an audio stream into a small number of acoustically compact categories, using methods comparable to those of speaker diarization [94]. Research in acoustic event detection seeks to detect specified acoustic events such as gunshots [18], explosions [82, 22], speech/music transitions [91], cough events [107], or audience cheering at a sports event [6].

This proposal is based, in part, on our contribution to the audio acoustic event detection evaluations sponsored by the project “Classification of Events, Activities and Relationships (CLEAR)” [113, 112]. The CLEAR evaluation used both isolated sound databases and a continuous audio database recorded in seminars [115]. Acoustic Event Detection (AED), as a task of the CLEAR Evaluation 2006 [114], was carried out by three participant partners from the CHIL project [17]. The project included two sub-goals: (1) Classification sub-goal sought to correctly classify discrete isolated events into one of 12 labeled categories (door snak, paper shuffling, footsteps, knocking, hair moving, phone ringing, spooncup jingle, key jingle, keyboard, applause, cough, and laughter), (2) Detection sub-goal sought to correctly detect and label the same 12 event categories in a business meeting recorded by multiple tabletop, wall-mounted, and headset microphones. All tested systems performed well in the Classification sub-task (typically 90% accuracy), with the best performance achieved by a Support Vector Machine (SVM) observing log frequency filter bank parameters and four kinds of perceptual features [104]. By contrast, all systems performed quite poorly in the Detection task; the best performance (38% accuracy) was achieved by our HMM recognizer with AdaBoost feature selection [127].

## 3 Objectives: Anomaly Detection

Visual analytics has been defined as “the science of analytical reasoning facilitated by interactive visual interfaces. People use visual analytics to synthesize information and derive insight from massive, dynamic, ambiguous and often conflicting data; detect the expected and discover the unexpected; provide timely, defensible, and understandable assessments; and communicate assess-

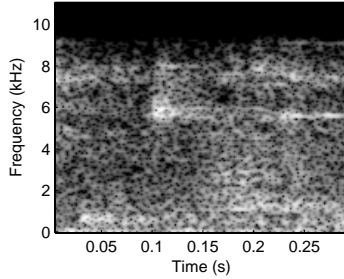


Figure 1: Single Acoustic Event (Keys Jingling)

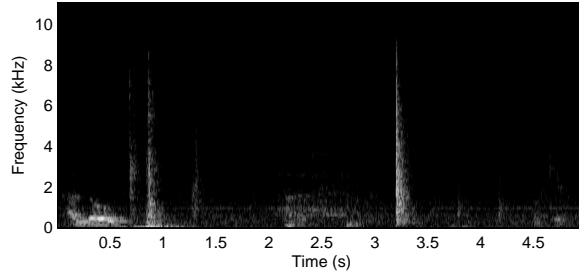


Figure 2: Multiple Acoustic Events

ment effectively for action” [116]. This proposal focuses on the sub-goal of visual analytics that we consider to be most in need of fundamental research: helping intelligence analysts to “discover the unexpected.” Specifically, we focus on the task of detecting anomalous events: events that deviate “from the common rule, type, arrangement, or form” [67].

Anomalies occupy a critical role in data analysis, because an event that is anomalous according to one hypothesis about the world may be the key piece of evidence in support of a better hypothesis. Heuer defined a method of data analysis called “alternative consideration of hypotheses (ACH),” according to which all competing hypotheses are simultaneously considered. He warned that “if analysts focus mainly on trying to confirm one hypothesis they think is probably true, they can easily be led astray by the fact that there is so much evidence to support their point of view. They fail to recognize that most of this evidence is also consistent with other explanations or conclusions, and that these other alternatives have not been refuted” [96].

The goal of our proposed research is a two-stream approach to data visualization, organized around the measurable outcome of “anomaly detection.” The first stream presents all available information about a large audio database to the human analyst in a form that can be rapidly and interactively examined. The second stream presents automatic assessments of the degree to which each second of audio is consistent with various hypotheses known to the software. The hypotheses known to the software will include the null hypothesis (“the events happening during this second of audio are common events in this database”), as well as a range of specific labeled hypotheses (non-threatening hypotheses such as “ordinary speech” and “ordinary traffic,” as well as potentially threatening hypotheses such as “angry speech” and “gunshots”). We propose that the goal of data transformation should be to present the evidence for each of these hypotheses side by side to the analyst, allowing him or her to rapidly estimate the level of support available for all competing hypotheses. “Anomaly detection” is a measurable, quantifiable outcome that will serve as a proxy for the interpretive tasks of data analysis: the proposed research is successful if it allows human subjects to rapidly and accurately detect the anomalous events in a large dataset.

## 4 Proposed Methods: Data Transformations

### 4.1 Multiscale Spectrograms and Wavelets

When a time series such as audio is rendered visually, simply drawing the values as a function of time (the “waveform”) rarely reveals interesting structure. Spectrograms are more useful.

When a spectrogram’s time axis covers a long duration, it is often the case that the number of pixels in the image display is insufficient to display all computed spectra. In such situations, the software must somehow adjust the temporal resolution of the spectrogram. The most common solution is to simply drop the spectra that fall between pixels. Unfortunately, the resulting spectrogram misses important signal details: for example, compare Fig. 1 to 2, or Fig. 3 to 4.

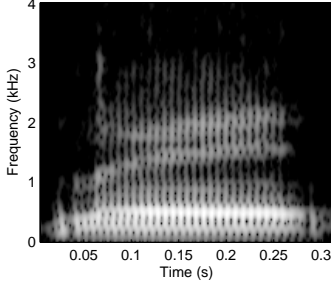


Figure 3: Speech fragment (word “right”)

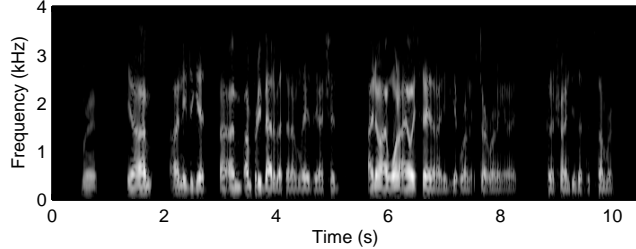


Figure 4: Speech (“right yeah I’m I need to get I’m I’m pretty bad about that I’m lazy I should I know ...”)

We propose to investigate at least four other, more useful display options. First, the length of the FFT can be adjusted to match the time span covered by one pixel in the display. Increasing the length of the FFT results in increased frequency resolution (fewer Hertz per pixel). High frequency resolution may be useful (if events of interest are coded in spectral fine structure), or it may hinder analysis (if events of interest are coded in the coarse spectral shape, as is, for example, the phonetic quality of a vowel). Because increased frequency resolution is not always useful, the scientific visualization community generally insists that an audio visualization tool must allow separate adjustment of the temporal resolution and frequency resolution of the display [11].

Second, the spectrogram can be matched to display resolution by downsampling the power spectral estimator. The Bartlett-Welch estimator computes a smoothed power spectral density estimate  $|Y(\omega, t)|^2$  as follows:

$$|Y(\omega, t)|^2 = \sum_{\tau=-\infty}^{\infty} h[\tau] |X(\omega, t + \tau)|^2 \quad (1)$$

where  $X(\omega, t + \tau)$  is the short-time Fourier transform of the original audio,  $|Y(\omega, t)|^2$  is the smoothed power spectral density, and  $h[\tau]$  is the impulse response of a low-pass filter. The Blackman-Tukey estimate performs a similar computation by averaging the autocorrelation (inverse transform of  $|X(\omega, t)|^2$ ). Both of these estimators emphasize high-amplitude frames; low-amplitude frames may be masked. The human ear computes a similar sort of temporal masking, but only over intervals of a few hundreds of milliseconds [81].

Over longer time intervals, humans appear to integrate perceptions of “loudness.” Loudness has been defined to be the cube root of power [36], so a third estimate of the smoothed spectrogram may be computed as

$$|Y(\omega, t)|^{2/3} = \sum_{\tau=-\infty}^{\infty} h[\tau] |X(\omega, t + \tau)|^{2/3} \quad (2)$$

The practical effect of Eq. 2 is quite different from that of Eq. 1: downsampled spectrograms computed using Eq. 2 retain information about the low-amplitude frames.

The fourth approach to variable time-scaling that we will test is a wavelet representation: a representation that uses different time scales in different frequency bands [97]. It has been demonstrated that, in the study of non-speech audio, wavelets can represent certain types of events better than a spectrogram [19, 111, 122, 123].

Because our proposed testbed applications view different time scales flexibly, and because computing Fourier and wavelet transforms at these massive scales is slow even with fast shortcuts [38], we propose to precompute and cache spectrograms and wavelets at multiple time and frequency resolutions, using all four of the proposed temporal scaling methods. By pre-computing and caching spectrograms, we will make it possible for the analyst to smoothly zoom through many time scales

without noticeable latency. Pre-computing and caching of spectrograms at multiple time and frequency scales can be performed in the data recorder, in real time, while the data are being acquired: the cost of an  $N$ -sample FFT is  $\log_2 N$  operations per sample, so computing many such transforms is not computationally constrained. Indeed, the key constraint will be disk space: the number of cached temporal and frequency resolutions will be optimized to match available disk storage.

## 4.2 Harmonicity, Spectral Center of Gravity, and Source Separation

We propose also to compute and test a number of audio features inspired by auditory scene analysis, including features that summarize the harmonicity and spectral dynamics of the signal (features based on the correlogram and rate-scale representations), features that estimate the degree to which a signal can be separated into multiple sources (features based on independent components analysis and non-negative matrix factorization), and compact summary features including the zero-crossing rate and spectral center of gravity.

Computational Auditory Scene Analysis (CASA) segregates from a mixture of sounds the isolated sounds that a human listener would identify [14, 35, 118]. Recently CASA has focused on segregating speech from interfering sounds (speech segregation) [64, 84, 85, 89, 99, 109, 119] or recognizing speech in the presence of other sources [7, 20]. It first decomposes the acoustic mixture into sensory elements or segments, and then groups segments that likely originate from the same source. Commonly, CASA addresses a known number of sound sources with known characteristics, e.g. stationary sources [69, 103, 4, 78], transient or tonal sources [14], and harmonic structure on “wefts” [35, 31]. Some attempts have been made to segregate time-varying sounds [14, 121]. CASA biomimetically solves blind source separation [117], much like the statistical technique of independent components analysis (ICA) [8, 27], but more effective with narrowband sources [117].

CASA widely uses the audio correlogram to summarize signal harmonicity in multiple bands. The correlogram represents a signal in 3D by taking the short-time autocorrelation of frequency bands in a 2D spectrogram. Since the correlogram reveals the full complexity of an audio signal’s periodicity, it helps human analysts investigate such signals [105, 49]. The correlogram can be inverted in order to reconstruct sounds in real time [105, 35].

Recent research has proposed a four-dimensional representation, sometimes called the rate-scale features, based on experimental measurements of the signals that activate nerve cells in the auditory cortex [120, 16]. It has been demonstrated that the over-complete scale-space representation provides discrimination among speech events [79] that is robust against most types of noise and reverberation [56]. Widespread adoption of the rate-scale features has been slow, because there are so many features available in the complete four-dimensional image. To our knowledge, the only methods that have successfully applied rate-scale features for either automatic speech or audio processing are methods of severe downsampling (e.g., the rate-scale features may be downsampled to estimate the MFCC [56]) or methods that can handle high feature dimension, such as SVMs [51].

Another approach inspired by the CASA literature is the method of non-negative spectral factorization [106]. Non-negative factorization determines an overcomplete set of basis functions that optimally represent any given database, under the constraint that all of the expansion coefficients are non-negative. Coefficients are usually selected to minimize a metric that encourages sparse representations (ones whose coefficients are mostly zero) [62]. Because they are sparse, the coefficients in a non-negative matrix factorization have statistics similar to the statistics of neurons. The basis sets learned by non-negative factorization approximate those learned by neurons in the visual cortex [86] and auditory cortex [66]. In audio, sparse non-negative factorization exploits the non-overlapping character of nonzero spectral magnitudes [99]: by explicitly computing the basis sets corresponding to each audio source, both sources can be resynthesized into their “unmixed” states [106].

Other informative but difficult-to-use features include knowledge-based features such as explicit periodicities, frequency transitions, and onsets/offsets in auditory nerve firing patterns based on the known topographical organization of the higher auditory pathways [14]. Bitar and Espy-Wilson proposed using a large subset of such features as the input to SVMs for automatic speech recognition [9]; Hasegawa-Johnson et al. demonstrated that a hybrid SVM-dynamic Bayesian network [51] or SVM-HMM [12] is able to use such high-dimensional feature vectors in order to detect acoustic-phonetic landmarks.

### 4.3 Data-Driven Feature Extraction

We have proposed a new framework for selecting and analyzing features for acoustic event detection (AED) [127]. This framework characterizes features by quantifying their relative discriminant capabilities using the Kullback-Leibler divergence (KLD) [63]. An algorithm based on Adaboost [93, 37] selects the most discriminating feature set from a large pool of features.

In automatic speech recognition (ASR), features are chosen based on how speech is produced and perceived. Phonemes are discriminated better with the spectral envelope than with spectral fine structure. ASR methods—feature sets for speech—such as MFCC [23] and perceptual LPC (PLP) [54] commonly smooth out fine structure with bandpass filters. Moreover, to match the ear’s non-uniform frequency resolution they adopt non-uniform critical bands, to increase low-frequency resolution. These optimizations are inappropriate for non-speech sounds, where neither human vocal apparatus nor human auditory apparatus apply.

To analyze events’ spectral structure and design suitable features for AED, we analyze features’ discriminatory capability with KLD. Intuitively, a discriminative feature component should separate an acoustic event from other events (and speech). KLD measures each feature component’s discriminatory capability for each audio event,  $d_{ij} = D(p_{ij}||q_i)$ , where  $p_{ij}$  and  $q_i$  are the  $i^{th}$  feature component’s respective distributions, given the  $j^{th}$  audio event and given a mixture of all audio events. KLD defines the difference between two probability distributions,  $p_{ij}$  and  $q_i$ , to be their cross entropy minus  $p_{ij}$ ’s self-entropy:  $d_{ij}(p_{ij}||q_i) = \int p_{ij}(x) \log(p_{ij}(x)/q_i(x))$ .

The  $i^{th}$  feature component’s *global discriminant capability* is defined by  $d_i = \sum_j P_j d_{ij}$ , where  $P_j$  is the  $j^{th}$  acoustic event’s prior probability. A large  $d_i$  means that the  $i^{th}$  feature’s distribution has large differences among acoustic events, and thus discriminates well.

Our system, proven in CLEAR competitions [114, 112], uses AdaBoost to select the feature set, but not to linearly combine several classifiers. Along with silence and speech, each audio event in the CLEAR development set is segmented into several acoustic event instances. These labeled event instances serve as labeled examples for AdaBoost. AdaBoost’s weak classifiers are of just one type: an example is correctly classified iff the correctly labeled feature’s generalized Markov model (GMM) log likelihood exceeds that computed by the global GMM. We concatenate the features thus selected into an event-independent feature vector. A hidden Markov model (HMM) then segments and finally classifies events.

Since audio events have varying spectral structure, we propose to use this feature selection framework to customize feature sets for each audio event. These features will include not only conditional speech features such as MFCC, PLP and filter bank parameters, but also image-based features such as HOG [70] and SIFT [74], which describe the local structure of the spectrogram and correlogram. We will also explore biomimetic and psychomimetic features [14].

### 4.4 Log Likelihood Features

Audio features such as the spectrogram, correlogram, pitch, and non-negative factorization are useful for describing an audio scene to a human analyst. Most audio analysts are accustomed to reading spectrograms and pitch tracks; if the temporal resolution is sufficiently fine, most analysts can get useful information from such a display. If the temporal resolution is too coarse, however, a

spectrogram is useless, and a pitch track is nearly useless. We propose to augment audio features using a variety of *model-based features*, described in this section and the next section. Model-based features measure the degree to which the audio on any given microphone, at any given time, fits a known audio model. An anomaly may be operationally defined as “something that has never been seen before:” in probabilistic terms, something “never before seen” is something that fails to fit a model of the previously observed audio. We define model-based features  $L(x_t)$  to be real-valued measures of the degree to which the audio, at time  $t$ , “fails to fit.” Model-based features are intended to *augment* audio features such as the spectrogram and pitch track, not to replace them: for example, model-based features might be used to color a spectrogram (e.g., “green” signals anomaly, “red” signals events of a known but dangerous event class), or alternatively, model-based features might be displayed on a set of axes parallel to the audio features.

The problem of automatically detecting anomalies in an input audio stream may be expressed using the notation of statistical hypothesis detection. Suppose that we are able to extract an audio feature vector  $\vec{x}_t$  at regular intervals. Under the null hypothesis  $H_0$  (no anomaly present), the distribution of the observed  $\vec{x}_t$  is modeled by a function  $p_0(\vec{x}_t)$ . An anomaly can be detected ( $H_0$  is rejected) if  $L_0(\vec{x}_t)$  is small, where

$$L_0(\vec{x}_t) = \ln p_0(\vec{x}_t) \quad (3)$$

A standard anomaly detector would “fire” (declare a detected anomaly) if  $L_0(\vec{x}_t)$  were less than some threshold. Instead of thresholding  $L_0(\vec{x}_t)$ , we propose to present the real-valued log likelihood (or its monotonic nonlinear transform) to the human analyst, possibly in the form of a color code, e.g., hue specifies the form of the PDF model  $p_0(\vec{x}_t)$  that has generated the lowest log likelihood, and color saturation signals the degree of confidence of the anomaly detector. Several different estimates of the PDF  $p_0(\vec{x}_t)$  will be considered. Unless the data lead us to a different choice, all of the choices will be related, in one way or other, to the mixture Gaussian PDF:

$$p_0(\vec{x}_t) = \sum_{k=1}^K c_k |R_k|^{1/2} e^{-(\vec{x}_t - \vec{\mu}_k)^T R_k (\vec{x}_t - \vec{\mu}_k)} \quad (4)$$

where  $c_k$  are weights,  $\mu_k$  are mean vectors, and  $R_k = 0.5\Sigma_k^{-1}$  are precision matrices. We will test systems that train Eq. 4 using both parametric and non-parametric methods, as follows. Parametric methods will follow the usual Expectation-Maximization techniques for training a mixture Gaussian model, i.e., we will first fix the number of Gaussians, then adjust the parameters  $c_k$ ,  $\vec{\mu}_k$ , and  $R_k$  to optimally represent a large number of training vectors [57]. Non-parametric methods will create a new Gaussian for each training vector: the mean vectors  $\vec{\mu}_k$  will be set equal to the training vectors, and the precision matrices will be set to  $R_k = \gamma I$  for some constant kernel precision  $\gamma$ . We propose to test two different non-parametric training methods, both of which have been carefully studied and therefore have known error bounds. First, we will test Parzen window PDF estimates with Gaussian windows; a Parzen window estimate is a PDF of the form shown in Eq. 4, with each mixture weight set to  $c_k \propto \frac{1}{K}$  [25]. Second, we propose to test one-class SVMs with Gaussian kernels. A one-class SVM is a measure of the “typicalness” of a feature vector  $x_t$  [15]; the negative of the one-class SVM score is a measure of anomalousness. Although the one-class SVM is not a log likelihood measure, it nevertheless has exactly the form shown in Eqs. 3 and Eq. 4; the key difference is that the mixture weights  $c_k$  are chosen not in order to optimally estimate the within-class probability, but in order to minimize an upper bound on the anomaly detector’s expected error probability [15].

#### 4.5 Log Likelihood Ratios

The mixture Gaussian PDF introduced in Eq. 4 is a universal approximator. As  $K \rightarrow \infty$ , Eq. 4 approximates any non-negative function of  $\vec{x}$  with vanishingly small error [108]. For finite  $K$ , however,

Eq. 4 inadequately models many important categories of random processes. Word frequencies, for example, are distributed with a Zipf distribution,  $p(x) \propto x^{-0.5}$  [129], which cannot be modeled by Eq. 4 using finite  $K$ . A finite- $K$  mixture Gaussian estimate of the Zipf distribution underestimates the probability of outliers, because the Gaussian PDF ( $e^{-x^2}$ ) is more compact than the Zipf ( $x^{-0.5}$ ). An anomaly detector based on Eq. 3, therefore, necessarily misclassifies *null-class outliers* (unusual feature vectors  $\vec{x}_t$  that fit within the null class, e.g., unusually loud bursts of wind) as anomalies (feature vectors not drawn from the null class). This misclassification may in fact be desirable: analysts may call null-class outliers anomalous. To determine its desirability, we intend to perform perceptual experiments comparing sets of data tagged using  $L_0(\vec{x}_t)$  (Eq. 3) with other sets of data tagged using a metric less susceptible to null-class outliers, the log likelihood ratio  $\text{LLR}_j(\vec{x}_t)$ .

Log likelihood ratios are extensively used in automatic speaker verification and language identification [83, 95, 110, 126], thus our research draws on the results of many prior experiments (our own and those of others). Suppose that, in addition to a large quantity of unlabeled data, we have a small amount of labeled data from several audio event classes: some of the classes are non-threatening (code white: normal traffic, dish clatter), while some are known to be threatening (code red: gunshots, explosions, angry shouting). The  $j$ th event  $H_j$  is detected if  $\text{LLR}_j(\vec{x}_t)$  is large, where

$$\text{LLR}_j(\vec{x}_t) = \ln \left( \frac{p_j(\vec{x}_t)}{p_0(\vec{x}_t)} \right) \quad (5)$$

and where  $p_j(\vec{x}_t)$  is a model of the PDF of feature  $\vec{x}_t$  given hypothesis  $H_j$ , and  $p_0(\vec{x}_t)$  is a comparable model conditioned on the null hypothesis  $H_0$  (no event present). The probability models  $p_j(\vec{x}_t)$  and  $p_0(\vec{x}_t)$  make similar assumptions; for example, both distributions might be mixture Gaussians (Eq. 4) with the same number of kernels ( $K$ ), but with slightly different mean vectors ( $\vec{\mu}_k$ ). Similarity between the numerator and denominator of Eq. 5 compensates for modeling errors: systematic under-estimation of the likelihood of null-class outliers,  $p_0(\vec{x}_t)$ , tends to be compensated by a matching under-estimation of the event likelihood  $p_j(\vec{x}_t)$ . Compensation for other types of variability can be imposed by designing the feature vector to reject known sources of error [125], by selecting features from a large pool of candidate features according to a maximum-mutual-information [87, 88] or minimum-error criterion [127, 128], or by adjusting the precision matrices in Eq. 4 to emphasize distinctions that can be confidently measured [126].

Unlike Eq. 3, Eq. 5 requires knowledge of the type of anomaly being detected. We propose to characterize anomalies with event classes,  $1 \leq j \leq J$ , chosen to include specific salient classes (gunshots, explosions, sirens) as well as classes covering every possible audio measurement vector (e.g., one class will use a mixture Gaussian model with mean vectors uniformly distributed between the smallest and largest possible measurement in each feature dimension). The proposed class PDF models' training will not include test anomalies, i.e., signals deliberately inserted into data during human perceptual experiments. Only some test anomalies will be drawn from classes for which PDFs were trained (e.g., gunshots, explosions); others will lie outside such classes (e.g., gorilla roar). Since some event models will of necessity be trained using few training tokens, to avoid over-fitting the training data we will use regularized training methods, including MAP adaptation of a universal background model [95] and two-class training of a radial basis function SVM approximating Eq. 5.

## 5 Proposed Methods: Interactive Testbeds

Two graduate research assistants (GRAs) will each develop a test application. The first application visualizes multi-day audio on a timeline, while the second, geographically distributed audio.

Each GRA will develop a large number of LLR and log-likelihood models, of the forms given in Eqs. 5 and 3, using varieties of feature vectors, definitions of anomalous event classes, types of feature vector normalization, and types of model training. Each GRA will estimate the utility of

the trained anomaly detectors using development test data (not including the test anomalies). On the basis of development test data, the GRAs will choose a set of measures  $LLR_j(\vec{x}_t)$  and  $L_0(\vec{x}_t)$  that will be mixed to define the hue, saturation, and value (HSV) of pixels in their test applications.

### 5.1 Multi-day Timeline Audio

This interactive application is a multi-parameter zoomable timeline, in the spirit of “non-linear editing” video editing suites. Its source data is a single audio recording several days long. It displays many parameters derived from the source data as HSV, over a horizontal timeline (Fig. 5).

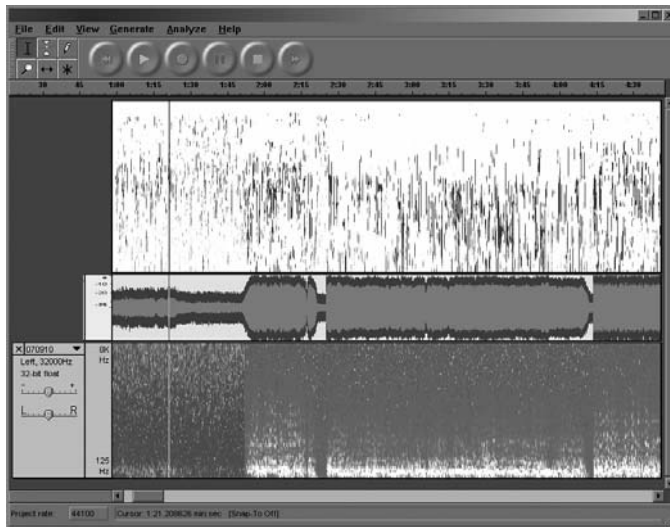


Figure 5: Mock-up of timeline application.

The analyst initially views the whole recording (during which hours was anything happening), then zooms in to smaller interesting time periods (multi-scale). Cached data values at multiple time scales let the analyst zoom not merely in discrete steps, but continuously. This works well for investigating large data sets like the human genome [3]; its ancestor is arguably the film “Powers Of Ten” [26]. Such zooming may be simplest with a mouse scroll-wheel (like Google Maps), but we will also experiment with 3Dconnexion’s Space Navigator 6-degree-of-freedom puck in the non-mouse hand.

Besides spectrograms  $\log |Y(\omega, t)|$  at different temporal resolutions, other parameters rendered above the time axis will include:

- color saturation, indicating degree of anomalousness according to a probabilistic model;
- hue indicating anomaly class, e.g., red=sudden, blue=speechlike, green=rumble.

Anomalousness of an interval is defined as the most anomalous part thereof, so that, even when the display is zoomed out to a resolution of one horizontal pixel per hour, that pixel’s colors will represent the most anomalous sounds during that hour.

We can easily prototype the interactive part of this application by extending the open-source project Audacity [77], using for inspiration its plugin Hi-Jacker, its connections to the familiar GUI package wxWidgets, and others’ extensions [68]. True interactivity will require some additional coding (for which we have adequate expertise [102]). A key part thereof will be caching parameter values, precomputed inexpensively at many zoom levels. Rudimentary caching has already been prototyped in Audacity [68].

This application will demonstrate the algorithms’ suitability to first-responder handheld computers. Current trends suggest that for this application, a 2010 handheld has about the same power as a 2008 desktop: near-terabyte solid state disk (SSD), 1 GHz CPU, 1 GB of RAM, and gigabit

networking. Battery life and screen size are the only significant distinction between the two. We propose to begin development on desktops at the start of this 3-year project, then migrate to a handheld during the third year.

## 5.2 The Milliphone

At the other limit of computational display hardware, this interactive application demonstrates the algorithms' suitability to large emergency-management control rooms. It is intended to run in the Beckman Cube, a fully immersive high-end virtual reality theater [130, 60, 100, 102, 101] (Fig. 7) whose software and technology have been robustly deployed in public museums [59, 58, 76]. Its hardware and software are a mature outgrowth of the original CAVE, which has been moved next door to the Cube and remains in heavy use [21].

This application renders the sound recorded by a thousand mobile microphones (hence "milliphone") that have been deployed in public places throughout a city. Geographical location is displayed horizontally, time vertically. A thousand  $(x, y, t)$  threads hang above the map. Each thread's  $(x, y)$  position at height  $t$  represents the position of the corresponding microphone at time  $t$ ; its color represents the recorded sound's anomalousness (Fig. 6).

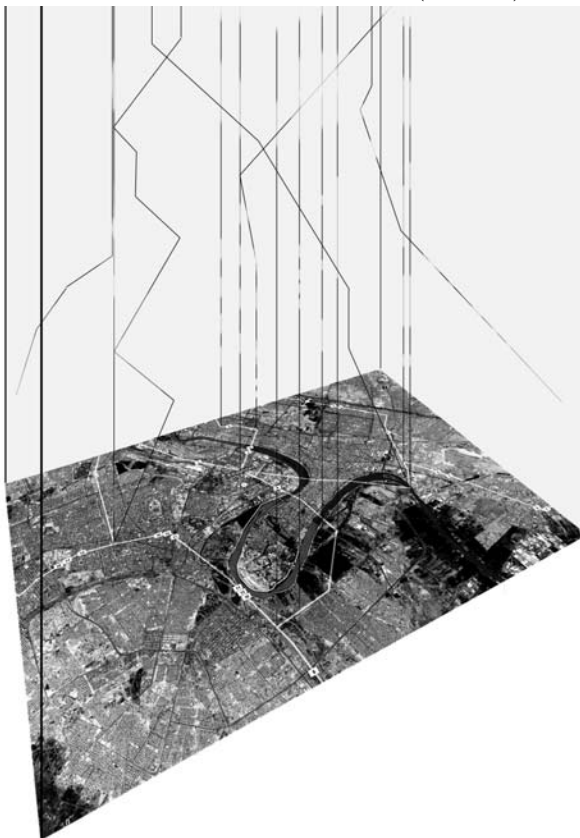


Figure 6: Mock-up of Milliphone application.

The analyst walks around and through the threads for global understanding. As with the handheld timeline application, the analyst can zoom continuously through the data, "drilling down" in space as well as in time. Traditional stepwise zooming is possible as well, with the familiar metaphor of click-dragging to select a rectangular  $(x, y, t)$ -subset of the data.

This display is inherently and effectively three-dimensional. Since the threads are hardly wide enough to obscure each other, the analyst can instantaneously grasp the spatial distribution of

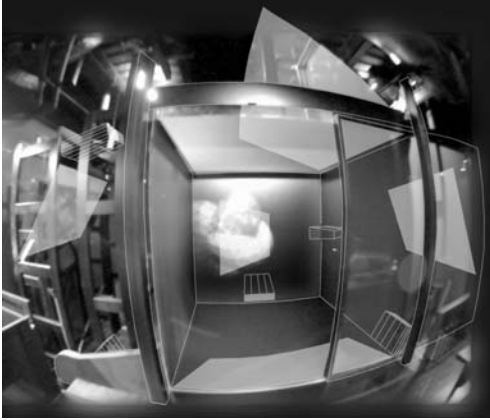


Figure 7: The Beckman Cube. Projectors and mirrors are highlighted for clarity.

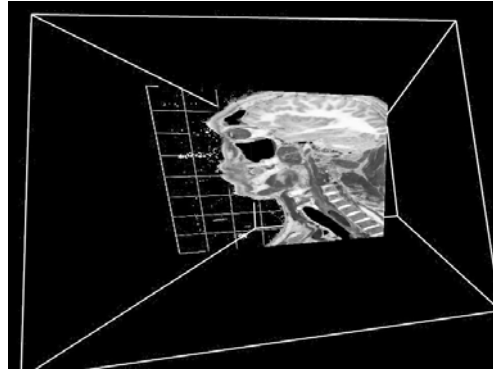


Figure 8: The Visible Human interactive explorer, running on a desktop simulator of the Beckman Cube.

recorded anomalous events at each time. Alpha-channel (semitransparency) may further prevent nonanomalously colored thread segments from occluding more anomalous segments.

A simple way to prototype the interactive part of this application is by extending our mature interactive renderer [58] of the Visible Human Project [2] (Fig. 8), designed particularly for the Beckman Cube.

One Co-PI has worked for over a decade in both virtual reality [60, 58, 100, 102] and the even more real-time-intensive field of software-based musical instruments [46, 48, 47]. This expertise is relevant to guiding the development of both of these testbeds.

### 5.3 Data Sets

For the timeline application, we will generate our own test data. Downloadable subsets of the data will be published on the web, and the entire dataset will be shared with other FODAVA institutions. At minimal cost and using available expertise, we will build a device to record a continuous week of studio-quality in-the-field audio, far from power sockets. Our rural location uniquely lets us easily record mile-distant interstate highway traffic, or even countryside with zero audible human activity. One Co-PI has industry experience with long single audio recordings (Wagner’s four-opera Ring Cycle on a 1997 CD-ROM) [80].

Into such long “boring” data sets we will insert anomalous sounds: gunshots, explosions, Switchboard conversation snippets, orators, symphonies, sci-fi sound effects, glasses clinking, bar fights, whales, gorillas. These sounds will come from large sound effects libraries. 99% of the inserted sounds will come from 2 or 3 “background classes” in the effects library’s ontology (e.g., perhaps, the classes of “vehicles” and “speech”); the remaining 1%—the key anomalies for subjects to discover—will come from other classes (e.g., animal sounds, shouting, singing, explosions). Giving the event times a Poisson distribution with mean 3/hour should be realistic. With appropriate fadein/fadeout, we can ensure that only the desired sound is available as a cue to listeners—no signal discontinuities or changes in “room tone” will be learnable by anomaly detection algorithms. Inserting these sounds ourselves at known time offsets means there is no data gathering effort—or human error—involved in noting where they are.

We will create tasks of increasing difficulty to test analysts’ ability to find subtle anomalies with the multi-day audio interface. First, they will detect anomalous speech events (non-English languages, shouting, singing) added to a long speech database such as Switchboard [42]. Second, for the most difficult task in the multi-day audio experiment, we will record audio under an approach path of our municipal airport, and ask subjects to find flight times of all airplanes taking off or landing.

Data for the Milliphone application will be synthesized from the ambience tracks of a large sound effects database, from original recordings, from data acquired by other FODAVA institutions, and from LDC corpora (Linguistic Data Consortium: Illinois has been a member most years since 1996). Simulated microphones will be placed in and around a prototype city at locations including cafés, commercial and residential street corners, highways, construction sites, and airports. A map of the same city will be drawn at the base of the Milliphone display (Fig. 6).

#### 5.4 Human Subject Protocols

Human subjects will be recruited to play the part of data analyst in each testbed. Each of the two GRAs will recruit five subjects per year, to test new developments in his or her testbed application. Each subject will participate in four two-hour sessions using the interface. In each session, the subject will be asked to explore two different datasets, for a total of eight datasets per year.

The first two-hour session will consist of two different training exercises, designed to give the subject some expertise in the task prior to testing. In the first training exercise (one hour), the subject will be told where to look in the data for each anomaly. Subjects will be led through the features of the visualization tool in tutorial fashion. In the second training exercise, subjects will be asked to find anomalies, and will be given immediate feedback (anomalous vs. not anomalous) for each event they claim to have detected.

Remaining sessions will be test sessions, measuring how quickly a subject finds the anomalies in six consecutive datasets. Datasets will differ in up to four types of independent variables, selected in different combinations in each year of the proposed research, generating a total of six different datasets per year: (1) the set of audio features provided, (2) the set of model-based features provided, (3) the type of anomalous events, and (4) the type of background (non-anomalous) audio. Dependent variables will include the elapsed time in which subjects find 50%, 75%, 90%, and 95% of the target anomalies. Statistical analysis (ANOVA) will test for significant relationships between the dependent and independent variables, singly and in interaction.

All human subject protocols will be subject to the prior approval of the University of Illinois Institutional Review Board.

## 6 Summary

This document has proposed a two-stream mathematical approach to the visualization of audio. In the first stream, we propose to take advantage of the expertise of a human analyst by displaying, to him or her, a set of audio features that will either be already familiar to any trained audio analyst (spectrogram, pitch tracks, energy) or that can be rapidly learned (correlogram, non-negative factorization). In the second stream, we propose to add signaling information (model-based features) for the purpose of drawing the analyst’s attention to events that the model cannot explain. Our two-stream paradigm is premised upon analyst expertise: we assume that an analyst listening to the temporal segment of audio containing an anomaly will detect that anomaly *immediately*. The goal of the proposed algorithms is to guide the analyst, as rapidly as possible, to the anomalous segments, so that he or she can find all anomalies in a large dataset without tediously listening to every second of the data. We want to reduce the analyst’s workload, but we don’t presume to read the analyst’s mind.

An analyst looking for the unexpected must be ready to spot events that are unexpected in any way. We have proposed, in this document, model-based features that will detect many types of “unexpected events,” but not all; therefore we propose that the model-based features should be used by the analyst to augment his or her view of the data, but not to obscure it. An image analogy captures our goal: our goal is not to label the images (“red car in front of a blue house”), but to enhance them (so that the analyst can more rapidly read the license plate).

## 6.1 Milestones

This proposal includes theoretical and experimental aspects. Theoretical aspects will develop new audio- and model-based features to help analysts rapidly detect anomalous events in an audio segment. Experiments will test these features' utility using two testbed applications.

Each of the two GRAs will (1) develop and optimize audio and model-based features, (2) implement a testbed, and (3) test the testbed with human subjects. Experimental tests will measure each feature's contribution to the analyst's efficiency. A key metric is error rate versus time compression: a human subject taking ten seconds to scan a week of audio may miss a few anomalies, while ten minutes of interactive investigation might reduce the error rate to zero. The GRAs' milestones are:

### GRA #1: Multi-day Timeline Audio

- **Year 1:** Develop multi-scale spectrogram-like features, pitch tracks, energy. Compile the Audacity toolkit. Implement continuous zooming through a multiscale spectrogram. Build field-recording computer, acquire multi-week audio recording of a remote site, add anomalies to the recording. Test the user interface for browsing of the complete database.
- **Year 2:** Develop a model-based feature set (log likelihood or log likelihood ratio) for emphasis of anomalies. Experiment with non-traditional spectral cues including correlogram, wavelet expansion, and/or non-negative matrix factorization. Implement model-based features in the Audacity interface; experimentally test the impact on human subjects.
- **Year 3:** Implement at least one novel audio feature set in the Audacity interface, coupled to the spectrogram and model-based features. Experimentally test the benefit of new feature set for human subjects. Begin porting the interface to a handheld computer.

### GRA #2: The Milliphone

- **Year 1:** Develop and test audio feature sets capable of summarizing activity of a microphone using color of a single point; such feature sets might include energy, harmonicity, and spectral center of gravity. Construct a simulated urban soundscape by merging ambience recordings from sound effect CDs and other sources. Render the colored threads corresponding to all sources. Allow user to sample each audio track by touching its thread with a mouse. Test the ability of user to browse the database in this way.
- **Year 2:** Develop model-based features for the detection of anomaly. Develop a method to integrate model-based features into the soundscape display. Test the utility of model-based features by asking subjects to find anomalies. Implement spatiotemporal zooming.
- **Year 3:** Develop information-theoretic criteria for the selection and/or weighted combination of features displayed in the milliphone threads. Implement caching adequate to inspect the full Milliphone data set without slowing down the application. Test with users.

The PI, at least one graduate student, and at least one co-PI will attend the FODAVA kickoff meeting and every annual program review. Databases and software acquired and/or created during this research will be shared iteratively with the FODAVA-Lead institution and other FODAVA institutions using secure ftp or grid-ftp (multithreaded sftp developed by the Teragrid project) [13], and reasonable-sized subsets will be published on the web; similar measures will be used to acquire, from the FODAVA-Lead institution, datasets collected at other FODAVA-funded institutions.

## 6.2 Intellectual Merit

This proposal contributes to the intellectual discourse in three fields of scientific inquiry: audio processing, anomaly detection, and visualization.

To the field of *audio processing*, we intend to contribute experimental measures of the correlation between audio features and analyst efficiency. Audio features for speech coding, synthesis, and

recognition have been exhaustively studied for decades, yielding a set of features that have been iteratively optimized for carefully prescribed tasks [5, 23, 55]. No type of audio has been analyzed as thoroughly as speech, but considerable research energy has been expended in two other areas: music information retrieval [24, 32, 34, 30, 44, 61], and CASA [14, 28, 29, 33, 98] including the sub-tasks of frequency segmentation [64, 84, 99] and temporal segmentation [124]. Beyond the task of segmentation, little research has addressed the general study of non-speech audio. Detection methods have been proposed for specific types of non-speech audio [107], but only recently has detecting and classifying general audio events been addressed [113]. It is clear from our research that optimal features for event detection differ greatly between speech and non-speech [127], and that much research remains to be done.

To *anomaly detection* we offer a new application. “Anomaly detection,” historically a sub-discipline of theoretical multivariate statistics, finds use in computer networks [90], hyper-spectral imaging [45, 41], SONAR [43], and surveillance video [40]. A current search for “audio” and “anomaly” on a standard engineering abstracts database (Inspec) yields no relevant hits (the only paper containing both words in its abstract describes a mere voice activity detector [92]). We propose to address the task of audio anomaly detection with standard statistical models from the anomaly detection community (log likelihood ratios and one-class SVMs), noting that considerable experimentation may be necessary to optimize the model inputs (spectral and hyperspectral transformations of the audio signal) for this completely unexplored task.

Finally, to *visualization* we offer a new theory of audio visualization. Audio visualization tools abound, but almost always for displaying and editing audio with known content, which needs no visualization beyond mere text labels. We propose to create two challenging audio visualization testbeds, to enrich them with multiple versions of two test corpora, and to turn subjects loose.

### 6.3 Broader Impact

The proposed research will contribute data, testbeds, and theory to the broader scientific community. All software and datasets created under this proposal will be released on the web, under open-source/open-data licenses similar to those governing our Syzygy virtual reality operating system [102] and AVICAR database [65], and will be made available immediately (and with frequent updates) to other FODAVA-funded institutions.

Theoretical developments will be published in professional conferences and journals. The best-performing algorithms will be distributed with the testbeds’ software.

The algorithms developed and tested with audio data will generalize to many other time series: seismic sensors (low-frequency audio), meteorology sensors, weather satellite data, temperature (HVAC), subway detectors of gases like hydrogen sulfide, radiation detectors, and a host of other kinds of sensor data. All these cases are similar to audio in some ways: (they are amenable to spectral and wavelet analysis, albeit typically with lower bandwidth and coarser temporal resolution than audio), and different in some ways (statistical models of the signal and noise that are appropriate for audio may be inappropriate in these other cases). All of these other signals are also unified in another way: all of them are topics of active current research at the Beckman Institute.

The Beckman Institute for Advanced Science and Technology is a 313,000 square foot building dedicated to inter-disciplinary research. Faculty and graduate students from more than 60 academic departments at the University of Illinois have offices in the Beckman Institute. Research at the Beckman Institute is entirely investigator-driven, but is nevertheless loosely organized around three Major Research Themes: Biological Intelligence, Electronic and Molecular Nanostructures, and Human-Computer Intelligent Interaction. The Integrated Systems Laboratory (ISL: home of co-PIs Kaczmarek and Goudeseune, and of the Beckman CUBE virtual reality theater) is funded in part by Beckman Foundation funds, for the purpose of providing virtual reality infrastructure (ISL’s core scientific expertise) to other groups in the Beckman Institute. ISL is also funded by the Beckman

Institute for the purpose of demonstrating virtual reality and current Institute research to groups touring the Beckman Institute, and at the annual Beckman Institute Open House (every March). There are typically 2-10 tour groups weekly, including frequent K-12 field trips, science camps, and other educational outreach programs. Through the University's Office of Public Engagement, the ISL frequently recruits potential students currently in their junior and senior years of high school, focusing on underrepresented groups from major metropolitan areas. First-year undergraduates at the University of Illinois have been known to cite ISL's virtual reality demonstrations (first seen when they were in elementary or high school) as a reason for their interest in science.

The proposed research has potential impact beyond the scientific community. Computer vision has recently become useful in a wide range of intelligent monitoring: traffic safety, factory and industrial safety, safety of child care and eldercare facilities, military applications, home security, and building security. In many modern computer vision applications, colocating an inexpensive microphone with a camera would be routine if only it were useful. Unfortunately, audio is not currently useful for intelligent monitoring applications, because automatic systems cannot detect audio anomalies accurately enough, and because human analysts don't have the time to listen to every sound recorded by every microphone. We propose to let computers do what computers can do, so that humans can do what humans can do: the computers will compute summary features over multiple views of the same data, so that human analysts can more rapidly drill down to find interesting anomalous events.

## 7 Results from Prior NSF Support

Audiovisual Phonologic-Feature-Based Recognition of Dysarthric Speech Mark Hasegawa-Johnson, Jon Gunderson, Thomas Huang, and Adrienne Perlman; November 9, 2005 to November 8, 2008; \$668,575; IIS 0534106; A53831.

Although automatic dictation software with high word recognition accuracy is now widely available, people with gross motor impairment, including some with cerebral palsy and closed head injuries, have not enjoyed these advances. This is because their motor impairment includes dysarthria, reduced speech intelligibility caused by neuromotor disability.

In this research, we have recorded audiovisual speech dictated by twelve talkers with spastic dysarthria, using an array of microphones and cameras originally developed for the AVICAR database [65]. We have shown that even talkers with low intelligibility (less than 50% correct transcription by naïve listeners) can dictate using a 26-word vocabulary such as the international radio alphabet with automatic word recognition accuracy exceeding 88% [52]. During the period funded by the grant, students and faculty funded by this research published conference papers and technical reports on audiovisual speech recognition [39, 50, 53, 72, 73], speaker identification [125], and speaker adaptation [71].

The AVICAR database (100 talkers recorded in a moving automobile [65]) is available via secure ftp download from the University of Illinois. Over twenty laboratories worldwide have downloaded and tested it. Data recorded by dysarthric talkers are currently being prepared for similar distribution. All dysarthric talkers were asked to specifically approve or disapprove three possible uses of the recordings: (1) research at the University of Illinois, (2) presentation at professional conferences, and (3) distribution to researchers at other institutions. All subjects except M06 voluntarily approved all three uses of their data.

## References

- [1] Kim S. Abouchacra, Tomasz Letowski, and Timothy Mermagen. Detection and localization of magazine insertion clicks in various environmental noises. *Military Psychology*, 19(3):197–216, 2007.
- [2] M.J. Ackerman. The visible human project. In *Medicine Meets Virtual Reality II: Interactive Technology & Healthcare: Visionary Applications for Simulation Visualization Robotics*, pages 5–7. Aligned Management Associates, 1994.
- [3] R.M. Adams, B. Stancampiano, M. McKenna, and D. Small. Case study: a virtual environment for genomic data visualization. In *Proc. IEEE Conf. on Visualization*, pages 513–516, 2002.
- [4] P.F. Assmann and Q. Summerfield. Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies. *Journal of the Acoustical Society of America*, 88:680–697, 1990.
- [5] B. S. Atal and Suzanne L. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 50(2):637–655, 1971.
- [6] M. Baillie and J.M. Jose. Audio-based event detection for sports video. *Lecture Notes in Computer Science*, 2728:61–65, 2003.
- [7] J.P. Barker, M.P. Cooke, and D.P.W. Ellis. Decoding speech in the presence of other sources. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 4, pages 270–3, 2000.
- [8] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.
- [9] Nabil Bitar and Carol Espy-Wilson. A knowledge-based signal representation for speech recognition. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pages 29–32, Atlanta, 1996.
- [10] BlackX, Twerty, Tom Viren, Brandywine, and Sondra C. How to detect counterfeit US money. Downloaded November 11, 2007 from [www.wikihow.com](http://www.wikihow.com).
- [11] Paul Boersma and David Weenink. Praat: doing phonetics by computer. Technical report, Institute of Phonetic Sciences, University of Amsterdam, 2003.
- [12] Sarah Borys and Mark Hasegawa-Johnson. Distinctive feature based SVM discriminant features for improvements to phone recognition on telephone band speech. In *Proc. Interspeech*, pages 697–700, Lisbon, Portugal, 2005.
- [13] John Bresnahan, Michael Link and dGaurav Khanna, Zulficar Imani, Rajkumar Kettimuthu, and Ian Foster. Globus gridftp: What’s new in 2007 (invited paper). In *Teragrid Conference*, 2007.
- [14] Guy J. Brown and Martin Cooke. Computational auditory scene analysis. *Computer Speech and Language*, 8:297–336, 1994.
- [15] C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Knowledge Discovery and Data Mining*, 2(2), 1998.

- [16] Robert P. Carlyon and Shihab Shamma. An account of monaural phase sensitivity. *J. Acoust. Soc. Am.*, 114(1):333–348, 2003.
- [17] CHIL. Computers in the human interaction loop. <http://chil.server.de/>, 2006.
- [18] C. Clavel, T. Ehrette, and G. Richard. Events detection for an audio-based surveillance system. In *IEEE International Conference on Multimedia & Expo*, pages 1306–1309, 2005.
- [19] Ronald R. Coifman and Victor Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Trans. on Information Theory*, 38(2):713–718, 1992.
- [20] Martin Cooke and Daniel P.W. Ellis. The auditory organization of speech and other sources in listeners and computational models. *Speech Communication*, 35:141–177, 2001.
- [21] Carolina Cruz-Neira, Dan Sandin, Tom Defanti, R.V. Kenyon, and J.C. Hart. The cave: Audio-visual experience automatic virtual environment. In *Communications of the ACM*, volume 35, pages 65–72, 1992.
- [22] Rui Cui, Lie Lu, Hong-Jiang Zhung, and Liun-Hong Cai. Highlight sound effects detection in audio stream. In *ICME03*, pages III: 37–40, 2003.
- [23] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-28(4):357–366, August 1980.
- [24] J. Stephen Downie. Music information retrieval. *Annual Review of Information Science and Technology*, 37, 2003.
- [25] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley & Sons, New York, 2001.
- [26] Charles Eames and Ray Eames. *Powers of Ten (film, 9 minutes)*. IBM, 1977.
- [27] F. Ehlers and H. G. Schuster. Blind separation of convolutive mixtures and an application in automatic speech recognition in a noisy environment. *IEEE Trans. Signal Processing*, 45:2068–2612, 1997.
- [28] Dan Ellis. Prediction-driven computational auditory scene analysis for dense sound mixtures. In *Proc. ESCA Workshop on Auditory Basis of Speech Perception*, Keele UK, Jul. 1996.
- [29] Dan Ellis. Using knowledge to organize sound: The prediction-driven approach to computational auditory scene analysis, and its application to speech/nonspeech mixtures. *Speech Communication*, 33, 1999.
- [30] Dan Ellis and Graham E. Poliner. Classification-based melody transcription. *Machine Learning Journal*, 2006.
- [31] Dan Ellis and David F. Rosenthal. Midlevel representations for computational auditory scene analysis: The weft element. In D.F. Rosenthal and H.G. Okuno, editors, *Computational Auditory Scene Analysis*, pages 257–272. Lawrence Erlbaum, Mahwah, NJ, 1998.
- [32] Daniel P.W. Ellis. Extracting information from music and audio. *Communications of the ACM*, 49(8):32–36, Aug 2006.

- [33] Daniel P.W. Ellis. Model-based scene analysis. In D. Wang and G. Brown, editors, *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Wiley/IEEE Press, 2006.
- [34] Daniel P.W. Ellis and John Arroyo. Eigenrhythms: Drum pattern basis sets for classification and generation. In *Proc. Intern. Symp. Music Information and Retrieval ISMIR-04*, pages 101–106, Barcelona, Oct 2004.
- [35] D.P.W. Ellis. *Prediction-driven computational auditory scene analysis*. Ph.D. thesis, MIT, 1996.
- [36] Harvey Fletcher and W.A. Munson. Loudness, its definition, measurement and calculation. *Journal of the Acoustical Society of America*, 5(2):82–108, 1933.
- [37] Yoav Freund and Robert E. Schapire. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780, 1999.
- [38] Matteo Frigo and Steven G. Johnson. Fftw: Fastest fourier transform in the west, 2007. Open-source multiplatform FFT implementation, version 3.1.2 released 2007 July 5.
- [39] Yun Fu, Xi Zhou, Ming Liu, Mark Hasegawa-Johnson, and Thomas Huang. Lipreading by locality discriminant graph. In *Proc. Internat. Conf. Image Proc. (ICIP)*, 2007.
- [40] Zhouyu Fu, Weiming Hu, and Tieniu Tan. Similarity based vehicle trajectory clustering and anomaly detection. In *International Conference on Image Processing*, pages 602–5, 2005.
- [41] J.-M. Gaucel, M. Guillaume, and S. Bourennane. Whitening spatial correlation filtering for hyperspectral anomaly detection. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pages 333–6, 2005.
- [42] J.J. Godfrey, E.C. Holliman, and J. McDaniel. SWITCHBOARD: telephone speech corpus for research and development. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pages 517–520, San Francisco, 1992.
- [43] A. Goldman and I. Cohen. Anomaly detection based on an iterative local statistics approach. In *IEEE Convention of Electrical and Electronics Engineers in Israel*, pages 440–3, 2004.
- [44] Masataka Goto and Yoichi Muraoka. Musical understanding at the beat level: Real-time beat tracking for audio signals. In D.F. Rosenthal and H.G. Okuno, editors, *Computational Auditory Scene Analysis*, pages 293–308. Lawrence Erlbaum Associates, 1998.
- [45] F. Goudail, N. Roux, I. Baarstad, T. Loke, P. Kaspersen, M. Alouini, and X. Normandin. Some practical issues in anomaly detection and exploitation of regions of interest in hyperspectral images. *Applied Optics*, 45(21):5223–36, 2006.
- [46] Camille Goudeseune. *Composing with parameters for synthetic instruments*. Doctoral thesis, University of Illinois at Urbana-Champaign, 2001.
- [47] Camille Goudeseune. Interpolated mappings for musical instruments. *Organised Sound*, 7(2):85–96, 2002.
- [48] Camille Goudeseune and Hank Kaczmarski. Composing outdoor augmented-reality sound environments. In *Proc. International Computer Music Conference*, 2001.

- [49] Svante Granqvist and Britta Hammarberg. The correlogram: A visual display of periodicity. *JASA*, 114(5):2934–2945, 2003.
- [50] Mark Hasegawa-Johnson. A multi-stream approach to audiovisual automatic speech recognition. In *IEEE Workshop on Multimodal and Multimedia Signal Processing*, Chania, Greece, 2007.
- [51] Mark Hasegawa-Johnson, James Baker, Sarah Borys, Ken Chen, Emily Coogan, Steven Greenberg, Amit Juneja, Katrin Kirchhoff, Karen Livescu, Srividya Mohan, Jennifer Muller, Kemal Sonmez, and Tianyu Wang. Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2005.
- [52] Mark Hasegawa-Johnson, Jonathan Gunderson, Adrienne Perlman, and Thomas Huang. HMM-based and SVM-based recognition of the speech of talkers with spastic dysarthria. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2006.
- [53] Mark Hasegawa-Johnson, Karen Livescu, Partha Lal, and Kate Saenko. Audiovisual speech recognition with articulator positions as hidden variables. In *International Congress on the Phonetic Sciences*, Saarbrücken, 2007.
- [54] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *the Acoustical Society of America*, 87:1738–1752, 1990.
- [55] Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.
- [56] Woojay Jeon and Beeing-Hwang Juang. Speech analysis in a model of the central auditory system. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(6):1802–1817, 2007.
- [57] Bin H. Juang, Stephen E. Levinson, and Man Mohan Sondhi. Maximum likelihood estimation for multivariate mixture observations of Markov chains. *IEEE Transactions on Information Theory*, 32(2):307–309, 1986.
- [58] H. Kaczmarski, C. Goudeseune, B. Schaeffer, L. Chong, R. Marshack, L. Hendrickson, and J. Crowell. Application framework for canvas, the virtual reality environment for museums. In *Electronic Imaging, the Visual Arts and Beyond (EVA London)*, 2005.
- [59] H. Kaczmarski and K. Harleman. Canvas, a virtual reality environment for museums. In *Electronic Imaging, the Visual Arts and Beyond (EVA Florence)*, 2005.
- [60] H. Kaczmarski, B. Schaeffer, C. Goudeseune, M. Zuffo, L. Soares, M. Cabral, B. Raffin, and J. Allard. Commodity clusters for immersive projection environments. In *Course taught at ACM SIGGRAPH*, 2003.
- [61] Kunio Kashino, Kazuhiro Nakadai, Tomoyoshi Kinoshita, and Hidehiko Tanaka. Application of the Bayesian probability network to music scene analysis. In D.F. Rosenthal and H.G. Okuno, editors, *Computational Auditory Scene Analysis*, pages 293–308. Lawrence Erlbaum Associates, 1998.
- [62] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T. Lee, and T. J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural Computation*, 15:349–396, 2003.

- [63] V. Krishnamurthy and J. Moore. On-line estimation of hidden markov model parameters based on the kullback-leibler information measure. *IEEE Trans. on Signal Processing*, 41(8):2557–2573, 1993.
- [64] Trausti Kristjansson, Hagai Attias, and John Hershey. Single microphone source separation using high resolution signal reconstruction. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2004.
- [65] Bowon Lee, Mark Hasegawa-Johnson, Camille Goudeseune, Suketu Kamdar, Sarah Borys, Ming Liu, and Thomas Huang. AVICAR: Audio-visual speech corpus in a car environment. In *Interspeech*, 2004.
- [66] M. S. Lewicki. Efficient coding of natural sounds. *Nature Neuroscience*, 5(4):356–363, 2002.
- [67] LLC Lexico Publishing Group. anomaly. In *Dictionary.com Unabridged (v 1.1)*. Random House, Inc., 2007.
- [68] Beinan Li, John A. Burgoyne, and Ichiro Fujinaga. Extending audacity for audio annotation. In *Proc. International Conference on Music Information Retrieval (ISMIR)*, pages 379–380, 2006.
- [69] J. S. Lim and A. V. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proc. IEEE*, 12:1586–1604, 1979.
- [70] Ming Liu, Huazhong Ning, Xi Zhou, Mark Hasegawa-Johnson, and Thomas S. Huang. On frequency domain correspondence: when vision techniques meet with speech problems. In *INTERSPEECH07*, 2007.
- [71] Ming Liu, Xi Zhou, Huazhong Ning, Mark Hasegawa-Johnson, and Thomas S. Huang. Speaker normalization via frequency domain correspondence: A data driven approach. In *Proc. Interspeech*, 2007.
- [72] Karen Livescu, Özgür Çetin, Mark Hasegawa-Johnson, Simon King, Chris Bartels, Nash Borges, Arthur Kantor, Partha Lal, Lisa Yung, Ari Bezman, Stephen Dawson-Hagerty, Bronwyn Woods, Joe Frankel, Mathew Magimai-Doss, and Kate Saenko. Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU summer workshop. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.
- [73] Karen Livescu, Özgür Çetin, Mark Hasegawa-Johnson, Simon King, Chris Bartels, Nash Borges, Arthur Kantor, Partha Lal, Lisa Yung, Ari Bezman, Stephen Dawson-Hagerty, Bronwyn Woods, Joe Frankel, Mathew Magimai-Doss, and Kate Saenko. Articulatory feature-based methods for acoustic and audio-visual speech recognition: 2006 JHU summer workshop final report. Technical Report WS06, Johns Hopkins University Center for Language and Speech Processing, 2007.
- [74] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *Journal of Comput. Vision*, 60:91–110, 2004.
- [75] L. Lu. Content analysis for audio classification and segmentation. *IEEE Trans. Speech and Audio Processing*, 10:504–516, 2002.

- [76] R. Marshack, N. Duchnowski, A. Watt, J. Jackson, and H. Kaczmarek. Creating immersive art without a programmer: The first year for canvas, a virtual reality environment for museums. In *Electronic Imaging, the Visual Arts and Beyond (EVA Florence)*, 2007.
- [77] Dominic Mazzoni and Roger Dannenberg. Audacity, 2007. Open-source multiplatform audio editor, version 1.2.6 released 2007 July 26.
- [78] R. Meddis and M. Hewitt. Modelling the identification of concurrent vowels with different fundamental frequencies. *Journal of the Acoustical Society of America*, 91:233–245, 1992.
- [79] Nima Mesgarani, Shihab A. Shamma, and Malcolm Slaney. Speech discrimination based on multiscale spectrotemporal features. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 601–4, Montreal, 2004.
- [80] S.B. Miller. Lost in the ‘ring’? click on wotan. *The New York Times*, March 23, 1997, 1973.
- [81] Brian C.J. Moore. *An Introduction to the Psychology of Hearing*. Academic Press, San Diego, CA, 1997.
- [82] M. R. Naphade, A. Garg, and T.S. Huang. Duration dependent input output markov models for audio-visual event detection. In *ICME01*, page 65, 2001.
- [83] Jiří Navrátil. Automatic language identification. In Tanja Schultz and Katrin Kirchhoff, editors, *Multilingual Speech Processing*, pages 233–272. Elsevier Academic Press, 2006.
- [84] J. Nix, M. Kleinschmidt, and V. Hohmann. Computational auditory scene analysis by using statistics of high-dimensional speech dynamics and sound source direction. In *Proc. Eurospeech*, pages 1441–1444, 2003.
- [85] H.G. Okuno, T. Nakatani, and T. Kawabata. a new speech enhancement: Speech stream segregation. In *International Conference on Spoken Language Processing 1996*, pages IV: 2356–2359, 1996.
- [86] B.A. Olshausen and D.J. Field. Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14:481–7, 2004.
- [87] M. Kamal Omar and Mark Hasegawa-Johnson. Maximum mutual information based acoustic features representation of phonological features for speech recognition. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2002.
- [88] M. Kamal Omar and Mark Hasegawa-Johnson. Maximum conditional mutual information projection for speech recognition. In *Eurospeech*, pages 505–8, 2003.
- [89] Kalle J. Palomaki, Guy J. Brown, and Jon P. Barker. Techniques for handling convolutional distortion with ‘missing data’ automatic speech recognition. *Speech Communication*, 43:123–142, 2004.
- [90] Animesh Patcha and Jung-Min Park. An overview of anomaly detection techniques: Existing solutions and technological trends. *Computer Networks*, 51(12):3448–3470, 2007.
- [91] J. Piquier. Robust speech / music classification in audio document. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages III: 2005–2008, 2002.

- [92] P.G. Raeth and D.A. Bertke. Finding events automatically in continuously sampled data streams via anomaly detection. In *Proc. IEEE National Aerospace and Electronics Conference (NAECON)*, pages 580–7, 2000.
- [93] Gunnar Ratsch, Takashi Onoda, and Klaus-Robert Muller. Soft margins for adaboost. *IEEE Trans. on Signal Processing*, 42:287–320, 2001.
- [94] D.A. Reynolds and P. Torres-Carrasquillo. Approaches and applications of audio diarization. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 953–956, 2005.
- [95] Douglas A. Reynolds and Richard C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Trans. Speech and Audio Processing*, 3(1), 1995.
- [96] Jr. Richards J. Heuer. *The Psychology of Intelligence Analysis*. Central Intelligence Agency, Washington, DC, 1999.
- [97] Olivier Rioul and Martin Vetterli. Wavelets and signal processing. *IEEE Signal Processing Magazine*, 1:14–38, Oct 1991.
- [98] David F. Rosenthal and Hiroshi G. Okuno, editors. *Computational Auditory Scene Analysis*. Lawrence Erlbaum Associates, 1998.
- [99] Sam T. Roweis. One microphone source separation. In *Neural Information Processing Systems*, volume 10, pages 626–632, 2000.
- [100] B. Schaeffer, P. Brinkmann, G.K. Francis, C. Goudeseune, and H. Kaczmarski. Myriad: scalable vr via peer-to-peer connectivity, pc clustering, and transient inconsistency. In *Computer Animation and Virtual Worlds*, volume 18, pages 1–17. John Wiley & Sons, 2007.
- [101] B. Schaeffer, H. Kaczmarski, L. Chong, M. Flider, L. Vanier, and Y. Hasegawa-Johnson. Tele-sports tele-dance: Full body network interaction. In *ACM Symposium on Virtual Reality Software and Technology*, 2003.
- [102] Benjamin Schaeffer and Camille Goudeseune. Syzygy: Native PC cluster VR. In *IEEE Virtual Reality*, 2003.
- [103] M. T. M. Scheffers. *Sifting vowels: Auditory pitch analysis and sound segregation*. Ph.D. thesis, University of Groningen, 1983.
- [104] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, US, 2002.
- [105] M. Slaney, D. Narr, and R.F. Lyon. Auditory model inversion for sound separation. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pages 77 –80, 1994.
- [106] Paris Smaragdis. Convolutional speech bases and their application to supervised speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1):1–12, 2007.
- [107] Jaclyn A. Smith, John E. Earis, and Ashley A. Woodcock. Establishing a gold standard for manual cough counting: video versus digital audio recordings. *Cough*, 2:6:1–6, 2006.
- [108] H. W. Sorenson and D. L. Alspach. Recursive bayesian estimation using gaussian sums. *Automatica*, 7:465–479, 1971.

- [109] Soundararajan Srinivasan and DeLiang Wang. Schema-based modeling of phonemic restoration. In *Proceedings of the European Speech Technology Conference (EUROSPEECH)*, pages 2053–2056, Geneva, Switzerland, 2003.
- [110] D.E. Sturim, D.A. Reynolds, R.B. Dunn, and T.F. Quatieri. Speaker verification using text-constrained gaussian mixture models. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2002.
- [111] C. Tantibundhit, J.R. Boston, C.C. Li, J.D. Durrant, S. Shaiman, K. Kovacyk, and A. El-Jaroudi. New signal decomposition method based speech enhancement. *Signal Processing*, 87:2607–2628, 2007.
- [112] Andrey Temko. CLEAR 2007 AED evaluation plan. <http://isl.ira.uka.de/clear07>, 2007.
- [113] Andrey Temko, Robert Malkin, Christian Zieger, Dusan Macho, Climent Nadeu, and Maurizio Omologo. Acoustic event detection and classification in smart-room environments: Evaluation of CHIL project systems. *Cough*, 65:5–11, 2006.
- [114] Andrey Temko, Robert Malkin, Christian Zieger, Dusan Macho, Climent Nadeu, and Maurizio Omologo. Acoustic event detection and classification in smart-room environments: Evaluation of CHIL project systems. *Cough*, 65:5–11, 2006.
- [115] Andrey Temko and Climent Nadeu. Classification of meeting-room acoustic events with support vector machines and variable-feature-set clustering. In *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pages V: 505–508, 2005.
- [116] James J. Thomas and Kristin A. Cook, editors. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center, 2004.
- [117] Andre J. W. van der Kouwe, DeLiang Wang, and Guy J. Brown. A comparison of auditory and blind separation techniques for speech segregation. *IEEE Trans. on Speech and Audio Processing*, 9(3):189–195, 2001.
- [118] D. Wang and G. Brown. *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Wiley-IEEE Press, 2006.
- [119] DeLiang Wang and Guy J. Brown. Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Trans. Neural Networks*, 10:684–697, 1999.
- [120] K. Wang and S. Shamma. Spectral shape analysis in the central auditory system. *IEEE Transactions on Speech and Audio Processing*, 3(5):382–395, 1995.
- [121] M. Weintraub. *A theory and computational model of monaural auditory sounds separation*. Ph.D. thesis, Stanford University, 1985.
- [122] G.W. Wornell. A karhunen-loeve-like expansion of  $1/f$  processes via wavelets. *IEEE Trans. on Information Theory*, IT-36, July 1990.
- [123] G.W. Wornell and A.V. Oppenheim. Wavelet based representations for a class of self-similar signals with application to fractal modulation. *IEEE Trans. on Information Theory*, IT-38, March 1992.
- [124] Tong Zhang and C.-C. Jay Kuo. Audio content analysis for online audiovisual data segmentation and classification. *IEEE Trans. Speech and Audio Processing*, 9(4):441–457, 2001.

- [125] Xi Zhou, Yun Fu, Ming Liu, Mark Hasegawa-Johnson, and Thomas S. Huang. Robust analysis and weighting on MFCC components for speech recognition and speaker identification. In *Proc. Interspeech*, 2007.
- [126] Xi Zhou, Jiří Navrátil, Jason W. Pelecanos, Ganesh N. Ramaswamy, and Thomas S. Huang. Intersession variability compensation for language detection. Submitted to *International Conference on Acoustics, Speech, and Signal Processing*, 2008.
- [127] Xi Zhou, Xiaodan Zhuang, Ming Liu, Mark Hasegawa-Johnson, and Thomas Huang. HMM-based acoustic event detection with AdaBoost feature selection. In *Classification of Events, Activities and Relationships*, Baltimore, May 2007.
- [128] Xiaodan Zhuang, Xi Zhou, Thomas S. Huang, and Mark Hasegawa-Johnson. Feature analysis and selection for acoustic event detection. Submitted to *International Conference on Acoustics, Speech, and Signal Processing*, 2008.
- [129] G.K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, 1949.
- [130] M. Zuffo, H. Kaczmarski, L. Soares, P. Bressan, B. Raffin, and P. Augerat. Commodity clusters for immersive projection environments. In *Course taught at ACM SIGGRAPH*, 2002.