Auditory Context Awareness via Wearable Computing

Brian Clarkson, Nitin Sawhney and Alex Pentland
Perceptual Computing Group and Speech Interface Group
MIT Media Laboratory
20 Ames St., Cambridge, MA 02139
{clarkson, nitin, sandy}@media.mit.edu

Abstract

We describe a system for obtaining environmental context through audio for applications and user interfaces. We are interested in identifying specific auditory events such as speakers, cars, and shutting doors, and auditory scenes such as the office, supermarket, or busy street. Our goal is to construct a system that is real-time and robust to a variety of real-world environments. The current system detects and classifies events and scenes using a HMM framework. We design the system around an adaptive structure that relies on unsupervised training for segmentation of sound scenes.

1. Introduction

In this paper we consider techniques that utilize auditory events in the environment to provide contextual cues for user interfaces and applications. Such cues enable a context aware application to provide relevant or timely information to the user, change its operating characteristics or take appropriate actions in its physical or virtual environment.

We will discuss a system for auditory scene analysis (ASA) and discuss the techniques used for detecting distinct auditory events, classifying a temporal sequence of auditory events, and associating such classes with semantic aspects of the environment. Our ASA system divides the auditory scene into two layers, sound objects (the basic sounds: speech, telephone ring, passing car, etc.) and sound scenes (busy street, supermarket, office, etc.). Models for sound objects are obtained using typical supervised training techniques. We show that it is possible to obtain a meaningful segmentation of the auditory environment into scenes using only unsupervised training.

This framework has been used as part of a wearable computing application, *Nomadic Radio*. Future work will address issues related to adaptive learning for different environments and utilizing the hierarchical structure for dynamically adding new classes to the system.

2. Related Work

Several approaches have been used in the past to distinguish auditory characteristics in speech, music and audio data using multiple features and a variety of classification techniques.

Scheirer [5] used a multi-dimensional classification framework on speech/music data (recorded from radio stations) by examining 13 features to measure distinct properties of speech and music signals. They have concluded that not all features are necessary to perform accurate classification. This suggests using a set of features that automatically adapt to the classification task. Foote [12] used MFCC's with decision trees and histograms to separate various audio clips. He claimed these techniques could be used to train classifiers for perceptual qualities (e.g. brightness, harmonicity, etc.). Other researches have used cluster [4] and neural net-based [3] approaches for similar level of sound classification.

Saint-Arnaud [6] presents a framework for sound texture classification that models both short-term and long-term auditory events. We also use this concept to organize sound classes into scenes and their constituent objects.

Ellis [2] starts with a few descriptive elements (noise clouds, clicks, and wefts) and attempts to describe sound scenes in terms of these. Brown and Cooke [11] construct a symbolic description of the auditory scene by segregating sounds based on common F0 contours, onset times and offset times. Unlike others, these two systems were designed for real-world complex sound scenes. Many of the design decisions for our system have been motivated by their work.

All of these systems represent solid progress towards auditory scene analysis, but none are real-time and few are appropriate for real-world input.

3. The Current System

We now describe our ASA system for exploring the ideas presented above. The system begins with a coarse segmentation of the audio stream into events. These events are then divided into sound objects which are modeled with Hidden Markov Models (HMMs). Scene change detection is implemented by clustering the space formed by the likelihoods of these sound object HMMs.

3.1 Auditory Input

The system is designed for real-time real world I/O. No special considerations were made in the selection of the microphone except that it be stable, small and unobtrusive. Currently we use a wireless lavalier microphone (Lectro-sonics M 185) because it can be easily integrated into personal mobile computing platforms. We believe that wearable systems will have the most to gain from an ASA system because the auditory environment of a mobile user is dynamic and structured. The user's day-to-day routine contains recurring auditory events that can be correlated amongst each other and the user's tasks.

3.1 Event Detection

The first stage in the system's pipeline is the coarse segmentation of auditory events. The purpose of this segmentation is to identify segments of audio which are likely to contain valuable information. We chose this route because it makes the statistical modeling much easier and faster. Instead of integrating over all possible segmentations, we have built-in the segmentation as prior knowledge.

The most desirable event detector should have negligible computational cost and low false rejection rate. The hypothesized events can then be handed to any number of analysis modules, each specialized for their classification task (e.g. speech recognizer, speaker identification, location classifier, language identification, prosody, etc.).

We used a simple and efficient event detector, constructed by thresholding total energy and incorporating constraints on event length and surrounding pauses. These constraints were encoded with a finite-state machine.

This method's flaw is the possibility of arbitrarily long events. An example is walking into a noisy subway where the level of sound always exceeds the threshold. A simple solution is to adapt the threshold or equivalently scale the energy. The system keeps a running estimate of the energy statistics and continually normalizes the energy to zero mean and unit variance (similar to Brown's onset detector [11]). The effect is that after a period of silence the system is hypersensitive and after a period of loud sound the system grows desensitized. Figure 1 shows the effects of adaptation for a simple tone (actual energy is on top and adapted energy is on the bottom). Notice that after 500 frames (8 secs), the system is ready to detect other sounds despite the continuing tone.

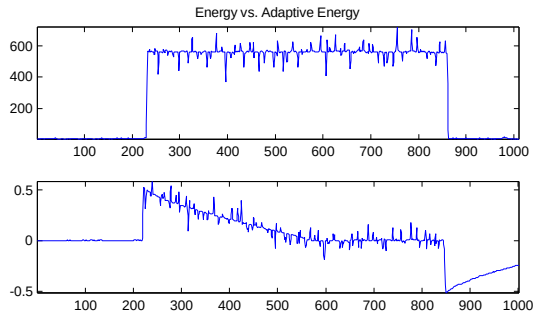

Figure 1: The event detector uses a normalized version (bottom) of raw energy (top) to gradually ignore long-lasting sounds.

3.2 Feature Extraction

Currently our system uses mel-scaled filter-bank coefficients (MFCs) and pitch estimates to discriminate, reasonably well, a variety of speech and non-speech sounds. We have experimented with other features such as linear predictive coefficients (LPC), cepstral coefficients, power spectral density (PSD) [7], energy, and RASTA coefficients. RASTA is appropriate for modeling speech

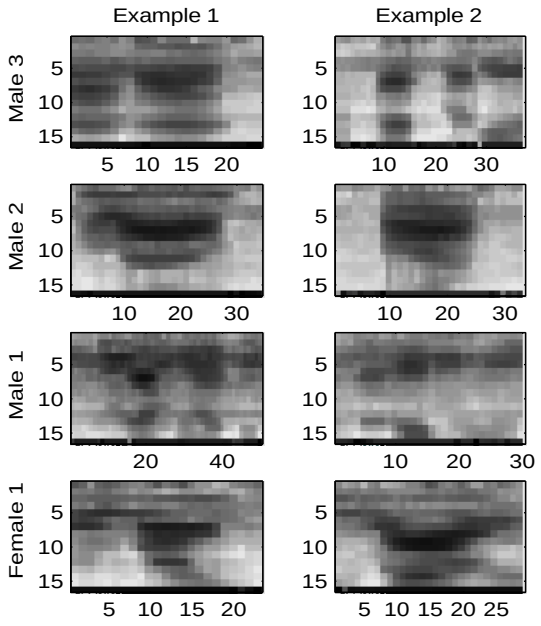

Figure 2: Comparison of speakers using 15 mel-scaled filter-banks. Notice that the gross spectral content is distinctive for each speaker. (Frequency is vertical and time is horizontal.)

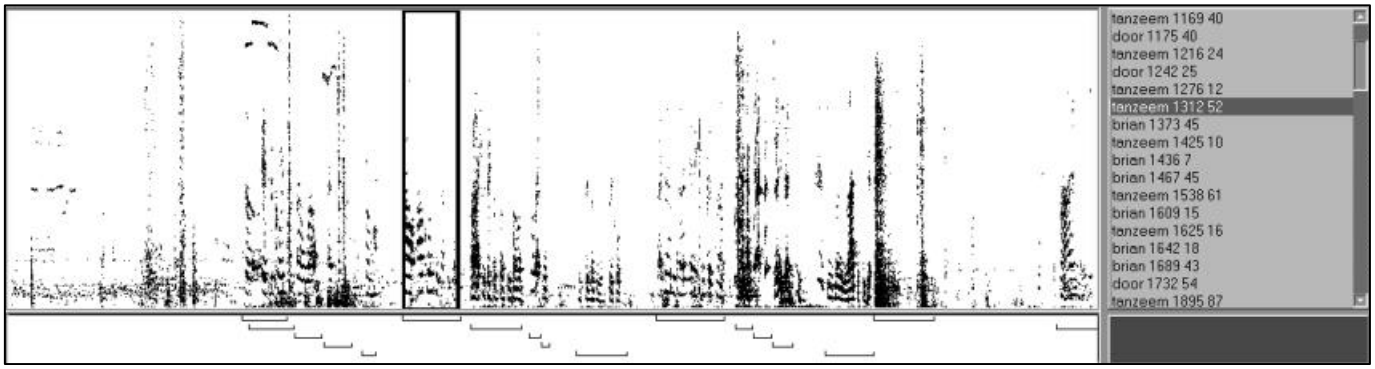

Figure 3: Transcription of detected events. Spectrogram (upper left), event boundaries (lower left), labels (upper right).

events such as speech recognition. The more direct spectral measurements, such as cepstral, LPC, and MFC, give better discrimination of general sounds (such as speaker and environment classification). Although for the purposes of this paper we restrict ourselves to a single set of features, we strongly believe that our system should include mechanisms for generating new features candidates as needed, and automatically selecting the appropriate features for the task.

To get a sense of what information this particular feature set extracts, we can compare the voices of different speakers. Figure 2 below shows the MFC features for 8 speech events (extracted with the event detection algorithm). There are 2 examples for each speaker to show some possible variations. These diagrams use 15 mel-scaled filter-banks (ranging from 0 to 2000Hz log-scale, each about 0.5 secs long) to show the rough spectral peaks for these speakers. Discrimination of speaker id for 4 speakers is quite simple as indicated by our 100% recognition accuracy (on a test set) using HMMs on these features. As more speakers are registered with the system (using only 15 mel-scaled filter-banks), the accuracy drops drastically. Adding pitch as an additional feature increases accuracy. An actual system for auditory scene analysis would need to be able to add features like this automatically. More complex methods would allow the discrimination of more speakers, but usually physical context (such as being in the office vs. at home) can restrict the number and identity of expected speakers.

3.3 Sound Object Classification

Methods

The features extracted from an event form a time series in a high dimensional space. Many examples of the same type of sound form a distribution of time series which our system models with a HMM. Hidden Markov Models capture the temporal characteristics as well as the spectral content of an event. Systems like Schreirer's [5], Saint-Arnaud [4], and Foote [12] ignore this temporal knowledge.

Since our ASA system is event-driven, the process of compiling these training examples is made easier. The event detection produces a sequence of events such as in Figure 3. Only those events that contain the sound object to be recognized need to be labeled. Also, it is not necessary to specify the extent of the sound object (in time) because the event detector provides a rough segmentation. Since the event might span more time than its sound object (as it usually does when many sound objects overlap), it implicitly identifies context for each sound object.

Once HMMs have been estimated for the needed sound objects, new sounds can be compared against each HMM. Similar sound objects will give high likelihoods and dissimilar objects low likelihoods. At this point the system can either classify the sound as the nearest sound object (highest HMM likelihood) or describe the sound in terms of the nearest N sound objects. The last option is necessary for detecting the case where more than one sound object may be present in the event.

Application

Nomadic Radio is a wearable computing platform that provides a unified audio interface to a number of remote information services [8]. Messages such as email, voice mail, hourly news, and calendar events are automatically downloaded to the device throughout the day, and the user must be notified at an appropriate time. A key issue is that of handling interruptions to the listener in a manner that reduces disruption, while providing timely notifications for relevant messages. This approach is similar to prior work by [9] on using perceptual costs and focus of attention for a probabilistic model of scaleable graphics rendering.

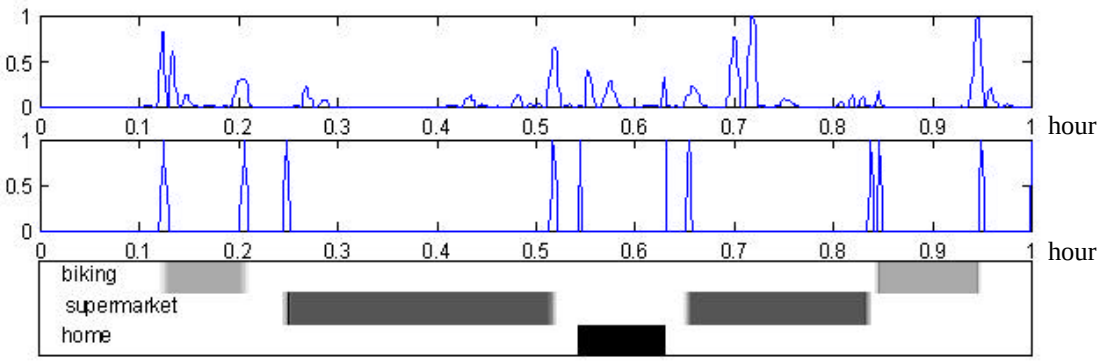

Figure 4: Scene change detection via clustering (top), hand-labeled scene changes (middle), scene labels (bottom).

In *Nomadic Radio* the primary contextual cues used in the notification model include: *message priority level* from email filtering, *usage level* based on time since last user action, and the *conversation level* estimated from real-time analysis of sound events in the mobile environment. If the system detects the occurrence of more than several speakers over a period of time (10-30 seconds), that is a clear indication of a conversational situation, then an interruption may be less desirable. The likelihood of speech detected in the environment is computed for each event within (10-30 second) window of time. In addition, the probabilities are weighted, such that most recent time periods in the window are considered more relevant in computing the overall speech level. A weighted average for all three contextual cues provides an overall notification level. The speech level has an inverse proportional relationship with notification i.e. a lower notification must be provided during high conversation.

The notification level is translated into discrete notification states within which to present the message (i.e. as an ambient or auditory cue, spoken summary or preview and spatial background or foreground modes). In addition, a latency interval is computed to wait before playing the message to the user. Actions of the user, i.e. playing, ignoring or deactivating the message adjust the notification model to reinforce or degrade the notification weights for any future messages during that time period.

We are currently refining the classifier's performance, and evaluating the effectiveness of the notification model. We are considering approaches for global reinforcement learning of notification parameters. We also plan to use additional environmental audio classes for establishing context. Hence, the system would find it less desirable to interrupt the user when he is speaking rather than during normal conversational activity nearby.

3.4 Sound Scene Segmentation

Methods

A sound scene is composed of sound objects. The sound objects within a scene can be randomly dispersed (e.g. cars and horns on the street) or have a strict time-ordered relation (e.g. the process of entering your office building). Thus, to recognize a sound scene it is necessary to recognize both the existence of constituent sound objects and their relationships in time.

We want to avoid manually labeling sound scenes in order to build their models. Thus, the approach we take is to build a scene segmentor using only unsupervised training. Such a segmentor does not need to perform with high accuracy. However, a low miss rate is required because the boundaries produced will be used to hypothesize contiguous sound scenes. Actual models of sound scenes can then be built with standard MMI (maximum mutual information) techniques.

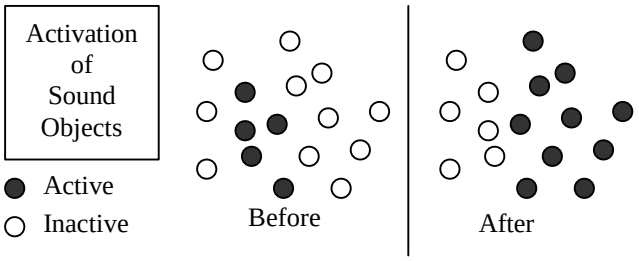

Figure 5: The basis of scene change detection is the shift in sound object composition.

Since most sound scenes are identifiable by the presence of a particular group of sound objects it may be possible to segment sound scenes before knowing their exact ingredients. An unsupervised training scheme, such as clustering, can segment data according to a given distance metric. The appropriate distance metric for segmenting sound scenes (i.e. detecting scene changes) would measure

the difference in sound object composition (as in Figure 5). We conducted an experiment to test this hypothesis as follows.

Experiment

We recorded audio of someone making a few trips to the supermarket. These data sets were collected with a lavalier microphone mounted on the shoulder and pointed forwards:

Supermarket Data Set: (approx. 1 hour)

1. Start at the lab. (late night)
2. Bike to the nearest supermarket.
3. Shop for dinner.
4. Bring groceries home.
5. (turn off for the night)
6. Start at home. (morning)
7. Walk to the nearest supermarket.
8. Shop for breakfast.
9. Bike to work.

Figure 6: A rough transcript of the data set used to test the scene segmentor. (left) The microphone setup used for the data collection. (right)

The Supermarket data was processed for events. For the purpose of model initialization, these events were clustered into 20 different sound objects using K-Means which uses only spectral content and ignores temporal relationships. HMMs were then trained for each of the 20 sound objects. These HMMs represent the canonical sounds appearing throughout the data and use the temporal and spectral information in each event for discrimination. Each of these HMMs was scored on each event extracted previously and a time-trajectory in a 20-dimensional space (of canonical sounds) is the result.

Now, during a sound scene the trajectory clusters in a particular section of the space. During a sound scene *change* the trajectory should shift to a new region. To test this, we clustered the trajectory (these clusters theoretically represent static scenes) and measured the rate at which the trajectory hops between clusters. The result is the top panel of Figure 4. In order to evaluate these results we also hand-labeled some scene transitions (middle and last panel).

It turns out that we have with very typical unsupervised clustering been able to detect 9/10 of the scene transitions (to within a minute). The graph also shows a few transitions detected where the hand-labels have none. The cost of these false positives is low since we are only trying to generate hypothetical scene boundaries. The cost of missing a possible scene boundary is however high.

Interestingly, the most prominent false detection was caused by some construction occurring in the supermarket that day. In terms of the scenes we were interested in (biking, home, supermarket) there was no scene change (the construction was a part of the supermarket scene). However, this shows how difficult it is to manually label scenes.

4. The Future System

New sounds and environments will always occur which the original designers would not have provided for. Therefore, adaptation and continuous learning are essential requirements for a natural UI based on audio. The modular (HMMs) and tree-like (object vs. scene) structure of our system make it amenable to extension. New sound objects and scenes can be added without having to re-train the rest of the system. Also, the supervised training techniques outlined in this paper, will assist in directing the process of adding new scene models as they are needed.

In a fixed feature space, as the number of classes or models grow it will become more difficult to distinguish among them. Eventually, it will be necessary to add new features. However, this solution will soon lead to prohibitively high dimensionality. We plan to use a hierarchical feature space where for each class the features are sorted by ability to discriminate, as in a decision tree. The scoring of each class would involve only the N most discriminating features.

5. Conclusion

We have given preliminary evidence of classifying various sound objects, such as speakers. We have also begun the development of incremental learning with the development of automatic scene change detection. The classification and segmentation have been implemented on Pentium PC and runs in real-time. There are still many problems to overcome. A major one is adapting models trained in one environment (such as a voice at the office) for use in other environments (the same voice in a car). This is a problem of overlapping sounds, which work by Bregman [1] and Brown & Cooke [11] should provide insight. However, our preliminary ASA system has allowed us to experiment with the use of auditory context in actual wearable/mobile applications. In the future, a variety of software agents can utilize the environmental context our system provides to aid users in their tasks.

References

[1] Bregman, Albert S. Auditory Scene Analysis: The Perceptual Organization of Sound. MIT Press, 1990.

[2] Ellis, Daniel P. "Prediction-driven Computational Scene Analysis", Ph.D. Thesis in Media Arts and Sciences, MIT. June 1996.

[3] Feiten, B. and S. Gunzel, "Automatic Indexing of a Sound Database using Self-Organizing Neural Nets". Computer Music Journal, 18:3, pp. 53-65, Fall 1994.

[4] Nicolas Saint-Arnaud, "Classification of Sound Textures". M.S. Thesis in Media Arts and Sciences, MIT. September 1995.

[5] Schreirer, Eric and Malcolm Slaney, "Construction and Evaluation of a Robust Multi-feature Speech/Music Discriminator", *Proc. ICASSP-97*, Apr 21-24, Munich, Germany, 1997.

[6] Wold, E., T. Blum, D. Keislar, and J. Wheaton. "Content-based Classification Search and Retrieval of Audio". IEEE Multimedia Magazine, Fall 1996.

[7] Sawhney, Nitin. Situational Awareness from Environmental Sounds. Project Report for Pattie Maes, MIT Media Lab, June 1997. *http://www.media.mit.edu/~nitin/papers/Env_Snds/EnvSnds.html*

[8] Sawhney, Nitin "Contextual Awareness, Messaging and Communication in Nomadic Audio Environments", MS Thesis, Media Arts and Sciences, MIT, June 1998. *http://www.media.mit.edu/~nitin/msthesis/*

[9] Horvitz, Eric and Jed Lengyel. Perception, Attention, and Resources: A Decision-Theoretic Approach to Graphics Rendering. *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI'97)*, Providence, RI, Aug. 1-3, 1997, pp. 238-249.

[10] Pfeiffer, Silvia and Fischer, Stephan and Effelsberg, Wolfgang. Automatic Audio Content Analysis. University of Mannheim, 1997.

[11] Brown, G. J. and Cooke, M. Computational Auditory Scene Analysis: A representational approach. University of Sheffield, 1992.

[12] Foote, Jonathan. A Similarity Measure for Automatic Audio Classification. Institute of Systems Science, 1997.