# Unstructured Audio Classification for Environment Recognition

**Selina Chu**

Signal and Image Processing Institute and Viterbi School of Engineering
Department of Computer Science, University of Southern California
Los Angeles, CA 90089, USA
selinach@sipi.usc.edu

**Abstract**

My thesis aims to contribute towards building autonomous agents that are able to understand their surrounding environment through the use of both audio and visual information. To capture a more complete description of a scene, the fusion of audio and visual information can be advantageous in enhancing the system's context awareness. The goal of this work is on the characterization of unstructured environmental sounds for understanding and predicting the context surrounding of an agent. Most research on audio recognition has focused primarily on speech and music. Less attention has been paid to the challenges and opportunities for using audio to characterize unstructured environments. Unlike speech and music, which have formantic structures and harmonic structures, environmental sounds are considered unstructured since they are variably composed from different sound sources. My research will investigate challenging issues in characterizing environmental sounds such as the development of appropriate features extraction algorithm and learning techniques for modeling the dynamics of the environment. A final aspect of my research will consider the decision making of an autonomous agent based on the fusion of visual and audio information.

## Acoustic Environment Recognition

We consider the task of recognizing environment sounds for the understanding of a scene (or context) surrounding an audio sensor. By auditory scenes, we refer to a location with different acoustic characteristics such as a coffee shop, park or quiet hallway. Consider, for example, applications in robotic navigation and obstacle detection, assistive robots, surveillance, and other mobile device-based services. Many of these systems are dominantly vision-based. When being employed to understand unstructured environments, their robustness or utility will be lost if visual information is compromised or totally absent. Audio data could be easily acquired, in spite of challenging external conditions such as poor lighting or visual obstruction, and is relatively cheap to store and compute than visual signals. To enhance the system's context awareness, we need to incorporate and adequately utilize such audio information.

Research in general audio environment recognition has received some interest in the last few years (Ellis, 1996; Huang, J. 2002; Malkin et al., 2005; Eronen et al., 2006), but the activity is much less as compared to that for speech or music. Other applications include those in the domain of wearables and context-aware applications (Waibel et al., 2004; Ellis et al., 2004). Unstructured environment characterization is still in its infancy. Most research in environmental sounds has centered mostly on recognition of specific events or sounds (Cai et al., 2006). To date, only a few systems have been proposed to model raw environment audio without pre-extracting specific events or sounds (Eronen et al., 2006; Malkin et al., 2005). Similarly, our focus is not in analyzing and recognition of discrete sound events, but rather on characterizing the general acoustic environment types as a whole.



## Time- and Frequency- Domain Feature Extraction

The first step in building a recognition system for auditory environment was to investigate on techniques for developing a scene classification system using audio features. We performed the study by first collecting real world audio with a robot and then building a classifier to discriminate different environments, which allows us to explore and investigate on suitable features and the feasibility of designing an automatic environment recognition system using audio information (Chu et al, 2006). We showed that we can predict with fairly accurate results the environment in which the robot is positioned (92% accuracy for 5 types of environments).

Many previous efforts utilize a high dimension feature vector to represent audio signals (Eronen et al., 2006). We showed in (Chu et al., 2006) that a high dimension feature set for classification does not always produce good performance. This in turn leads to the issue of selecting an optimal subset of features from a larger set of possible features to yield the most effective subset. In the same work, we utilized a simple forward feature selection algorithm to obtain a smaller feature set to reduce the computational cost and running time and achieve a higher classification rate. Although the results showed improvements, the features found after the feature selection process were more specific to each classifier and environment type. It is with these findings that motivated us to look for a more effective approach for representing environmental sounds. Toward this goal, we investigated

in ways of extracting features and introduce a novel idea of using matching pursuit as a way to extract features for unstructured sounds (Chu et al, 2008).

As with most pattern recognition systems, selecting proper features is the key to effective performances. Audio signals have been traditionally characterized by Mel-frequency cepstral coefficients (MFCCs) or some other time-frequency representations such as the short-time Fourier transform and the wavelet transform, etc. (Rabiner et al., 1993). MFCCs have been shown to work well for structured sounds such as speech and music sounds, but their performance degrades in the presence of noise. Environmental sounds, for example, contain a large variety of sounds, which may include components with strong temporal domain signatures, such as chirpings of insects and sounds of rain. These sounds are in fact noise-like with a broad flat spectrum and are not effectively modeled by MFCCs.

Therefore, we proposed a novel feature extraction method that utilizes the matching pursuit (MP) algorithm to select a small set of time-domain features (Chu et al, 2008), which we called MP-features. MP-features have shown to classify sounds where the frequency domain features (e.g., MFCCs) fail and can be advantageous when combining with MFCCs to improve the overall performance. Extensive experiments were conducted to demonstrate the advantages of MP-features as well as joint MFCC and MP-features in environmental sound classification. To the best of our knowledge, we were the first to propose using MP for feature extraction for environmental sounds. This method has shown to perform well in classifying fourteen different audio environments, achieving 83% classification accuracy. This result is very promising, considering that, due to the high variance and other difficulties in working with environmental sounds, recognition rates have thus far been limited as the number of targeted classes increases, i.e. approximately 60% for 13 or more classes (Eronen et al., 2006).

## Adaptive Audio Background Modeling

The next goal of this thesis is to develop a way to perform background modeling that can capture the dynamic nature of the environment. Such models should have the following characteristics: 1) require none or little assumptions on prior knowledge, 2) able to adapt to audio changes over time, and 3) able to handle multiple dynamic audio sources. Most methods on background modeling for audio have been adaptation of works for video (Cristani et al., 2004). The adaptation used for audio has been using specific fixed parameters for each type of context or environment. A starting point would be to extend the work of existing background modeling technique to make the system more flexible, by coming up with ways to integrate some learning/statistical model into the process.

## Decision Making in Dynamic Environments

The final goal of this thesis is to perform planning and decision making in dynamic environments using a fusion of audio and visual information. We will investigate on methods and techniques for dealing with decision making under uncertainty that utilize basic methods based on probability, decision and utility theories. We start by building a framework for decision making that incorporates different modalities. The first step is to develop a framework for information fusion from the two different modalities, audio and visual sensor information, to enhance the characterization of the environment or scene context. A major issue in this area is the synchronization or coupling of events from different data streams. An initial approach would be to build a system where the decisions for determining certain scenes or environment are made in each individual information stream, based on low-level information processing. Then from further processing, such as correlation analysis on the training data, we can extract linkage information between features of the different data streams to combine the two decisions for making predictions for decision making and planning.

## References


Cai, R., Lu, L., Hanjalic, A., Zhang, H., and Cai, L.-H. 2006. A flexible framework for key audio effects detection and auditory context inference. In *IEEE Trans on Audio, Speech and Language Processing,* 14(3):1026–1039.

Cristani, M., Bicego, M., Murino, V. 2004. On-Line Adaptive Background Modelling for Audio Surveillance. In *proc. Of ICPR.*

Chu, S., Narayanan, S., and Kuo, J. C.-C. 2006. Content Analysis for Acoustic Environment Classification in Mobile Robots. In *Proc. of AAAI Fall Symposium.*

Chu, S., Narayanan, S., and Kuo, J. C.-C. 2008. Environmental Sound Recognition using MP-based Features. In *proc. of IEEE ICASSP.*

Ellis, D. P. W. 1996. Prediction-driven computational auditory scene analysis, Ph.D. thesis, MIT Dept. of EE and CS.

Ellis, D. P. W. and Lee, K. 2004. Minimal-impact audio-based personal archives. In *proc. of CARPE.*

Eronen, A., Peltonen, V., Tuomi, J., Klapuri, A., Fagerlund, S., Sorsa, T. Lorho, G., Huopaniemi, J. 2006. Audio-based context recognition. In *IEEE Trans On Speech and Audio Processing.*

Huang, J. 2002. Spatial auditory processing for a hearing robot. *In proc. of ICME.*

Malkin, R., Waibel, A. 2005. Classifying User Environment for Mobile Applications using Linear Autoencoding of Ambient Audio. In *Proc. of IEEE ICASSP.*

Rabiner, L. and Juang, B.-H. 1993. *Fundamentals of Speech Recognition. Prentice-Hall.*

Waibel, A., Steusloff, H., Stiefelhagen, R., and the CHIL Project Consortium 2004. Chil - computers in the human interaction loop. In *proc. of WIAMIS.*