Unsupervised Audio Scene Analysis

Chris Stauffer & W. Eric L. Grimson

Artificial Intelligence Laboratory
Massachusetts Institue Of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: This work addresses the problem of robustly monitoring a scene using a microphone or microphone array. Rather than concentrating on human speech, we are attempting to model the general sounds that occur in a particular scene and to look for long-term contexts of those sounds and to look for anomalous sounds.

Motivation: While many company's are expending great effort in the field of automatic speech recognition(ASR), little attention is being paid to general audio and long-term modeling of audio in general. Even an ASR system which could give a complete transcription of the words heard in an environment would lack vital information. e.g., who was talking, when they were talking, what was the tone of the conversation, did someone slam the door, did someone use a harsh expletive, when do these conversations tend to happen, etc. In fact, this information is potentially of great value even without the transcription of what was said.

Previous Work: Dan Ellis has done significant work in decomposing audio stream into separate audio elements resulting from independent sources[1]. He has produced a summary of the work done in this area[2].

Approach: Rather than concentrating on the low-level segmentation of separate sources, which generally runs about three orders of magnitude slower than real-time, we are going to rely on our microphones being in a sparse environment. This will allow us to use a rather simple segmentation procedure and concentrate on the high-level issues.

The high-level issues we are going to approach are clustering the segmented audio into different audio sources, learning patterns of audio segments, finding audio segments which are out-of-context(e.g., breaking glass, etc.), and learning a context cycle for the types of audio that are heard in the environment (daily work cycles, traffic light patterns, etc.).

Difficulty: The goal of this project is create a system that will work in any sparse audio environment and the evaluation metric we will use is how many interesting models can be developed for completely different audio environments. The difficulty here is that we are not able to take advantage of domain specific knowledge without limiting the domains where our system can be used.

Impact: If such a system were placed on a persons wristwatch, he could ask some interesting questions... How much time did I spend in conversations this week? How much time typing? Did I get enough sleep? Tell me when I went jogging. How often did I yell at my students or children? etc. In a parking garage, such a system could be used to fire a warning if a person screams, a gunshot is heard, or glass breaks.

There are significant issues regarding who collects the audio and how that information will be used. We are also looking at the outward impact of such a technology and ways of controlling its uses and abuses.

Future Work: Future work will concentrate on integrating this system with the current surveillance and monitoring effort[3], which can already recognize general classes of objects and the activities they undergo in a particular environment.

Research Support: This work has been funded by DARPA under contract number N00014-00-1-0907 administered by the Office of Naval Research.

References:

[1] Ellis, D.P.W. Prediction-driven computational auditory scene analysis for dense sound mixtures. In *Proc. ESCA Workshop on the Auditory Basis of Speech Perception*, Keele, July 1996 (6pp).

[2] Ellis, D.P.W. Divisive issues in Computational Audiotry Scene Analysis. *http://www.icsi.berkeley.edu/ dpwe/casa/oview*, talk given March 7, 1997.

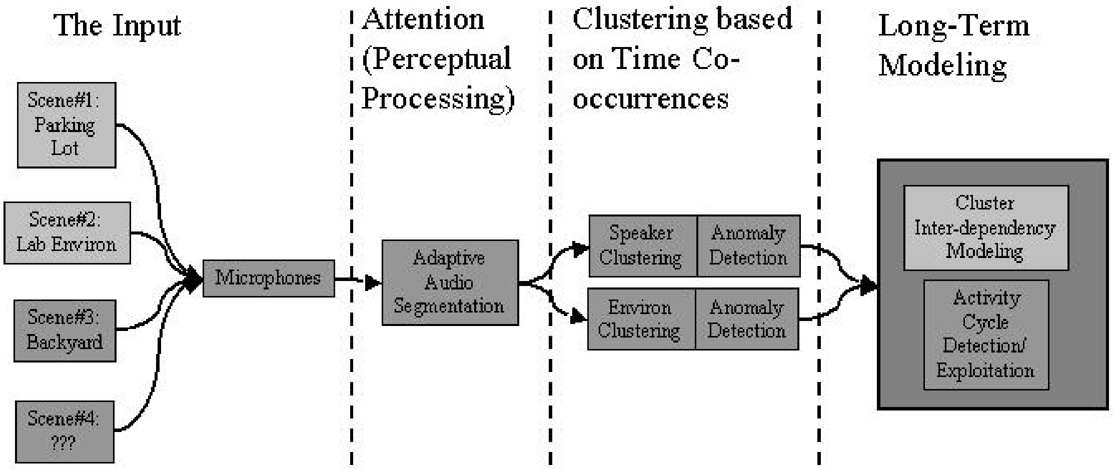

Figure 1: *This figures shows the audio processing sequence involved in modeling the general audio in a scene. First, general audio from microphones is segmented using a variation of a background subtraction algorithm outlined in [3]. Second, the audio snippets that are extracted are clustered into similar groups and anomalous snippets are tagged. Finally, the labeled audio is used to find long-term patterns and relationships between different sounds classes (possibly in different microphones).*

[3] C. Stauffer and W.E.L. Grimson. Adaptive Background Mixture Models for Real-Time Tracking. *Proc. Computer Vision and Pattern Recognition*, 246-252, 1999.