Auditory Representation and Sound Separation

by
Malcolm Slaney and Richard F. Lyon
Apple Computer, Inc.
Mail Stop 76-2H
20525 Mariani Avenue
Cupertino, CA 95014
(408) 974-4535, (408) 974-6111
malcolm@apple.com, lyon@apple.com

There are many ways to represent and process sound. This paper will describe perceptual models of the human auditory system. We believe the solution to the sound separation problem is to use the proper representation and we have been studying how the brain solves this problem.

We live in a rich acoustic environment. We are able to understand speech, even with interfering speakers and unrelated background noise. The entire process is effortless. We can easily separate the sounds that we hear into separate sources and form what some researchers have called an auditory image.

Conventional representations of speech like LPC or Fourier analysis assume a single source-resonator model of production or a linear model of perception. But these assumptions aren't always valid. We want to be able to understand sounds coming from multiple sources as well as the human ear and the human ear is anything but linear.

Our work during the last several years has been to build better models of the human auditory system. This includes more accurate cochlear models and and now models of higher level auditory functions.

The brain uses several clues for sound separation. Binaural localization, onsets and common modulation are some of the more important ones. To this end, we have been exploring the use of a perceptually-motivated three-dimensional representation of sound called the correlogram. The correlogram represents sound as a moving image of cochlear place (or frequency) and short-time autocorrelation versus time. The result is a compelling visualization of sound, which encodes many of the perceptually important clues in a form where these clues are easy to detect.

Our model of human sound separation includes a cochlear model, the correlogram, basic perceptual object detectors and some sort of grouping mechanism. We are currently working on object detectors based on motion in frequency or pitch and have started to think about onset and binaural detectors. The purpose of these detectors is to summarize key information about each component of the sound and to provide input to the grouping mechanism which tracks these objects over time and decides which components are coming from a single source.

This paper will describe our work with the correlogram, possible neurophysioligical and silicon implementations, and our work with motion detectors in the auditory system. In addition, our initial success detecting motion in the correlogram using optic flow will be described. The talk will be illustrated with audio and video examples.