

**Schlumberger workshop**  
**on**  
**Mathematical Models of Sound Analysis**

14-15 juin 2012

RÉSUMÉS

**Francis Bach**, *INRIA et ENS*

**Machine Learning for Audio Source Separation**

Blind source separation has attracted a lot of attention in signal processing and machine learning over the last two decades. However, unsupervised separation of several tracks from a single signal remains a largely open problem. In this talk, I will present recent work based on machine learning aimed at providing new solutions, one based on structured sparsity-inducing norms, and one based on semi-supervised learning (joint work with Augustin Lefèvre and Cédric Févotte).

**Douglas Eck**, *Google*

**Mixing Audio Signals and User Models for Large Scale Music Recommendation**

I'll start by giving a short guided tour of music recommendation in Google Play Music and discuss some of the challenges faced in building such a product. Audio feature extraction and machine learning plays a large role in what we do. I'll discuss some aspects of our current audio/ML pipeline, highlighting areas where more research is needed, including audio feature extraction, vector quantization, multi-class ranking and rapid retrieval and scoring at scale.

I'll also discuss strategies for integrating user feedback into such a model. Though I'll describe many technical aspects of our work, the main goal of the talk is to foster a better understanding of the real-world challenges we face in providing large-scale audio-driven music recommendations. I will make sure to leave plenty of time for audience feedback and discussion.

**Rémi Gribonval**, *INRIA Rennes*

**Sparse Audio Models for Inverse Audio Problems**

Inverse problems are ubiquitous in audio processing, from the restoration of saturated single channel audio to the localization of audio sources using arrays of two or more microphones. In the last decade, a solid mathematical and algorithmic framework has emerged to address such problems, based on the notion of sparse atomic decompositions in signal dictionaries. In practice, choosing a dictionary often requires prior expert knowledge. I will discuss some recent alternatives to pre-chosen dictionaries, including designs exploiting the underlying physics, as well as data-driven approaches where the model is learnt from a training corpus.

**Hynek Hermansky, Johns Hopkins University**  
**Dealing with Unknown Unknowns in Speech**

Stochastic machine learning approaches build a model of the world by optimizing their performance on some training data. This assumes that the world will not change in the future. Problems occur when this assumption is violated. The world is full of "unknown unknowns", among these belong signal distortions not seen in the training data, or lexical items not present in a vocabulary of a machine, which all create problems for the current ASR. The problem is inherent to machine learning and will not go away unless alternatives to extensive reliance on false beliefs of unchanging world are found. We discuss a unified way for dealing with such unexpected harmful variability (noise) and with unexpected lexical items (out-of-vocabulary words) in speech. In both cases, we are motivated by earlier observations of human performance on such problems, which indicate existence of multiple parallel processing streams in human speech processing cognitive system.

**Aren Jansen, Johns Hopkins University**  
**Automatically Learning the Structure of Spoken Language Without Supervision**

The dominant paradigm in the speech recognition community for the past four decades has been to train automatic systems with as much transcribed data we can get our hands on. This strategy has led to the development of highly accurate systems that have finally found a place in our daily lives in the form of popular applications such as Apple iPhone's Siri. An unfortunate consequence of this trajectory, however, is that state-of-the-art recognition performance can only be achieved on languages and domains for which vast transcribed training resources either exist or can be easily obtained. With public internet resources like YouTube and PodCasts, untranscribed speech audio is easy to obtain and contains a wealth of hidden information regarding the acoustic-phonetic, lexical, and grammatical structure of the language being spoken. The trick is uncovering this structure automatically, an endeavor that will require new machine learning techniques, algorithms scalable to massive problem sizes, and a lot of patience. I will provide an overview of my efforts in these directions and describe some useful language- and domain-independent technologies that have been produced along the way.

**Stéphane Mallat, IHÉS et École Polytechnique**  
**Invariants for Sound Classification**

Physiological models of the cochlea as well as many classification algorithms are based on the calculation of wavelet coefficients. The modulation spectrum of these coefficients appeared to play an important role for classification but also in physiological models. Why wavelets? Why and how to compute such modulation spectrum? Is it specific to sounds or is it a particular instance of more general principles of data organization? This lecture will address these issues through the construction of invariant representations. Invariants play a crucial role to eliminate irrelevant variability for classification. It will be shown that stable invariants preserving sufficient information for classification are deeply related to wavelet transforms and iterated modulation spectrum. Invariants to time shift and frequency transpositions will be studied, with applications to sounds and image classification.

**Josh McDermott, *New York University***  
**Auditory Texture Representation**

Humans infer many important things about the world from the sound pressure waveforms that enter the ears. In doing so we solve a number of difficult and intriguing computational problems. We recognize sound sources despite large variability in the waveforms they produce, extract behaviorally relevant attributes that are not explicit in the input to the ear, and do so even when sound sources are embedded in dense mixtures with other sounds. This talk will describe my recent work using auditory texture representation as a window into how we accomplish these feats. Sound textures are produced by superpositions of large numbers of similar acoustic features (as in rain, swarms of insects, or galloping horses). They are noteworthy for being stationary, which raises the possibility that time-averaged statistics might capture their structure. I will describe work testing this idea. The work stems from two premises: first, that understanding perception requires understanding real-world sensory stimuli and their representation in the brain, and second, that a theory of the perception of some property should enable the synthesis of signals that appear to have that property. I will show how the synthesis of textures from statistics of biological auditory models provides evidence for statistical texture representations. I will also discuss how synthetic textures can be used to reveal new aspects of sound segregation.

**Nima Mesgarani, *University of California, San Francisco***  
**Representation of Spectrotemporal Features of Speech in Human Auditory Cortex**

Humans possess a remarkable ability to attend to a single speaker's voice in a multi-talker background. How the auditory system manages to extract intelligible speech under such acoustically complex and adverse listening conditions is not known, and, indeed, it is not clear how attended speech is internally represented. Here, using multi-electrode surface recordings from the cortex of subjects engaged in a listening task with two simultaneous speakers, we demonstrate that population responses in non-primary human auditory cortex encode critical features of attended speech: speech spectrograms reconstructed based on cortical responses to the mixture of speakers reveal the salient spectral and temporal features of the attended speaker, as if subjects were listening to that speaker alone. We find that task performance is well predicted by a rapid increase in attention-modulated neural selectivity across both single- electrode and population-level cortical responses. These findings demonstrate that the cortical representation of speech does not merely reflect the external acoustic environment, but instead gives rise to the perceptual aspects relevant for the listener's intended goal. A model of cortical processing of sound is discussed with various applications for speech processing technologies.

Israel Nelken, *Hebrew University*

### **The representation of Surprise in Auditory Cortex**

Neurons in auditory cortex are sensitive to the history of the incoming sound streams at multiple time scales. They respond differently to the same sound when it appears with different probabilities, but they also respond differently to the same sound, with the same probability, when it appears more or less regularly. While most of our work has been done with pure tones, I will also show that neurons are also sensitive to the probability of appearance of complex spectro-temporal patterns. Finally, I will present a normative model for the detection of surprise, where surprise is conceptualized here as a prediction error that depends on an internal representation of sound probabilities. Internal representations may have different complexity, and result in different quality of prediction; complexity and prediction error can be traded, resulting in a set of optimal representations that have maximal quality of prediction for a given complexity. I will show that prediction errors calculated from these optimal representations account for a substantial amount of variance of the neuronal responses on a trial-by-trial basis, suggesting that the responses of neurons in auditory cortex explicitly represent surprise.

Shihab Shamma, *University of Maryland*

### **Mathematical Models of Auditory Cortical Processing**

Adaptive multiscale analyses of complex sounds

A computational perspective on the anatomy and physiology of the auditory pathway is presented that relates underlying biological mechanisms to modern audio processing strategies and algorithms. Two overarching principles will be highlighted: I. Multiresolution spectro-temporal representation of sound; II. Temporal coherence in the perceptual organization of complex sounds and its relation to deep auto-encoder networks. These principles have provided versatile insights and guided research in central auditory processing and the design of systems in audio applications.

Simon Thorpe, *CNRS Toulouse*

### **Brain Mechanisms for Learning to Recognize Sounds**

Humans are able to recognize briefly presented sounds on the basis of remarkably little information. This is particularly true for sounds that we are very familiar with, such as our favorite pieces of music, or the themes to well-known television and radio programs. How does the brain store this sort of sensory pattern? I will argue that some relatively simple neural mechanisms involving STDP (Spike-Time Dependent Plasticity) may be a key to understanding this ability. Both psychophysical and modeling studies suggest that a few tens of repetitions may often be enough to store a memory trace for a particular sound pattern. Furthermore, these mechanisms may allow such memory traces to be maintained intact for decades.