Using Learned Source Models to Organize Sound Mixtures

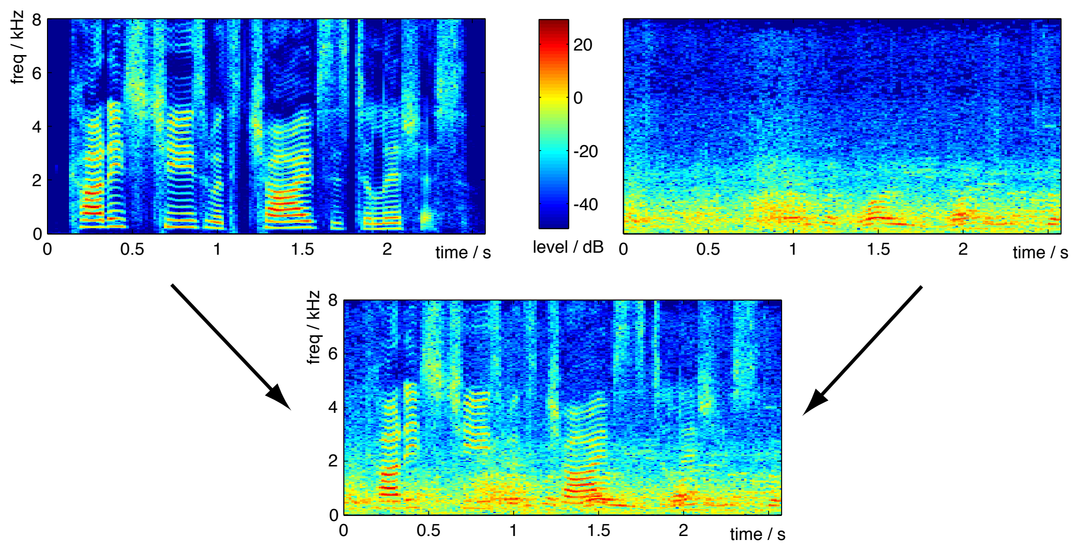
Dan Ellis
Laboratory for Recognition and Organization of Speech and Audio
Dept. Electrical Eng., Columbia Univ., NY USA

dpwe@ee.columbia.edu http://labrosa.ee.columbia.edu/

1. Source Models as Constraints
2. Examples of Model-Based Systems
3. Acquiring and Using Models
4. Biological Relevance?

The Problem of Scene Analysis

- How do we achieve 'perceptual constancy' of sources in mixtures?



  - no obvious segmentation of objects
  - underconstrained: infinitely many decompositions
  - time-frequency overlaps cause obliteration

Scene Analysis as Inference

Ellis'96

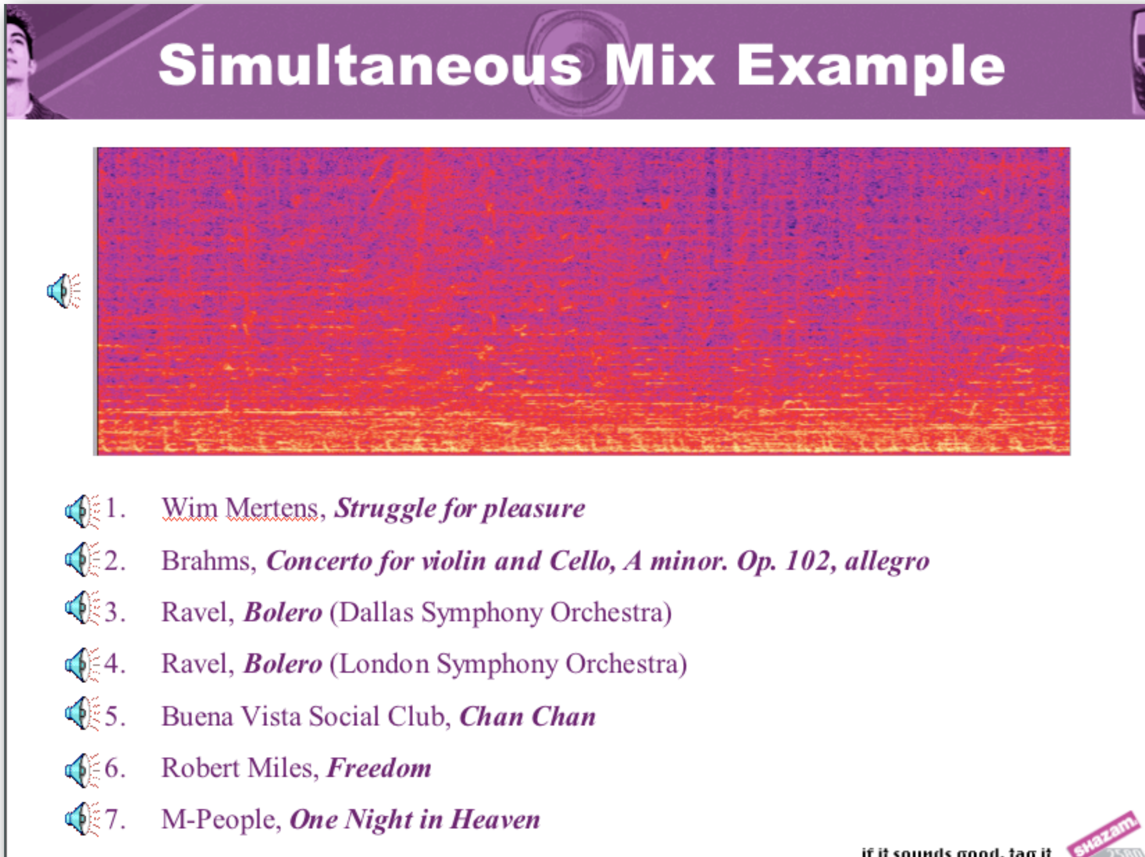
- Ideal separation is rarely possible
  - i.e. no projection can completely remove overlaps
- Overlaps ⇒ Ambiguity
  - scene analysis = find "most reasonable" explanation
- Ambiguity can be expressed probabilistically
  - i.e. posteriors of sources $\{S_i\}$ given observations X:

$$P(\{S_i\}|\,X) \propto \underbrace{P(X\,|\{S_i\})}_{\text{combination physics}}\ \underbrace{P(\{S_i\})}_{\text{source models}}$$

- Better source models → better inference
  - .. learn from examples?

An Example: Fingerprinting

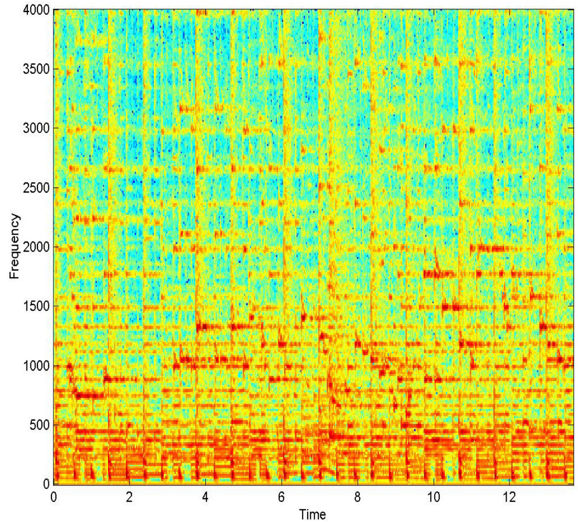
- "Impossible" separation task (Avery Wang)

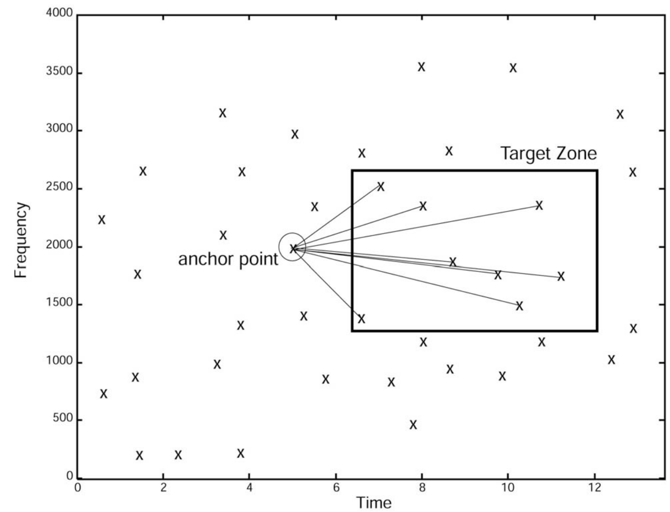


  - separation and restoration!

Fingerprinting: How it Works

- Library of songs (>1M) described by hashes





- After ~10s, song/segment identified > 98%
- Key ideas:
  - known-item database of exact waveforms
  - tiny part of signal used (... the most robust part)

Example 2: Mixed Speech Recog.

- Cooke & Lee's Speech Separation Challenge
  - short, grammatically-constrained utterances:

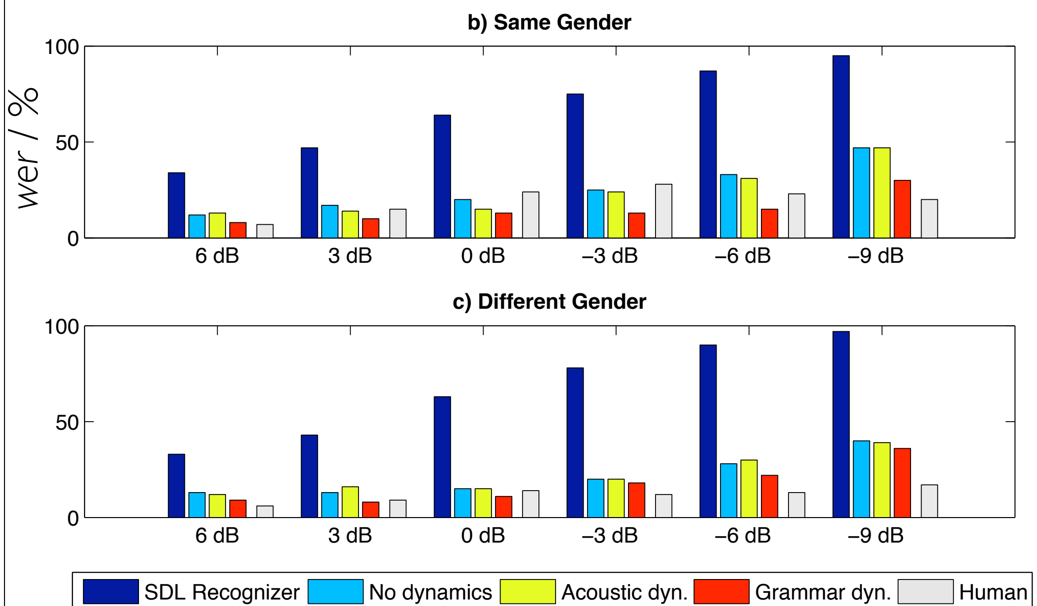
<command:4><color:4><preposition:4><letter:25><number:10><adverb:4>

e.g. "bin white at M 5 soon"

t5_bwam5s_m5_bbilzp_6p1.wav

Kristjansson et al. Interspeech'06

- IBM's "superhuman" recognizer:



  - Model individual speakers (512 mix GMM)
  - Infer speakers and gain
  - Reconstruct speech
  - Recognize as normal...
- Grammar constraints a big help

Scene Analysis as Recognition

- We don't want waveforms
  - limits to what listeners discriminate
  - .. especially over long term
- The outcome of perception is percepts
  - source identities (categories)
  - .. plus some salient parameters
- Scene analysis: recovering source + params
  - classification + parameter estimation
  - .. implies predefined set of classes = source models

What are the Models?

Models allow world knowledge (experience) to help perception

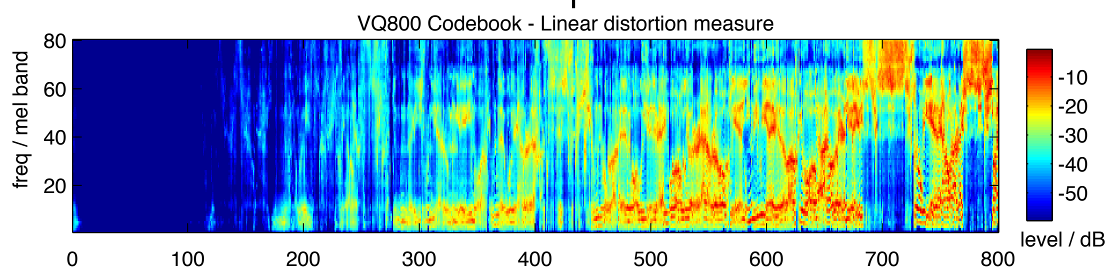
- Explicit Models (dictionaries)
  - can represent anything (“non-parametric”)
  - conceptually simple but inefficient in space/time
- Parametric Models (subspaces)
  - encapsulate broader constraints (e.g. harmonicity)
  - rely on actual regularity in the domain
  - may not be easy to apply (fit)
- Middle ground?
  - e.g. locally-learned manifolds
  - or dictionaries + parametric transformations

Learning, Representing, Applying

- Models encapsulate experience/environment
  - evolutionary scale (hardware)
    vs. lifetime scale (conventional learning)
- Tradeoff between an efficient domain and a flexible learner
  - auditory percepts already factor out e.g. channel characteristics (phase, reflections, gain)
- Learned knowledge must be easy to apply
  - e.g. representations that are easier to recall/match

Dictionaries vs. CASA

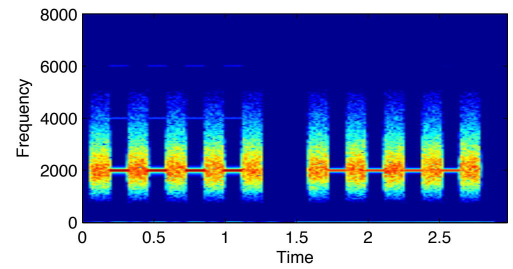
- Source models can learn harmonicity, onset
  - ... to subsume rules/representations of CASA



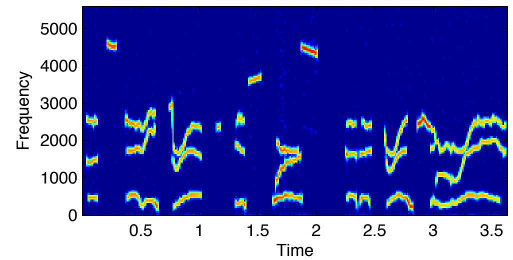
  - can capture spatial info too *[Pearlmutter & Zador'04]*
- Can also capture sequential structure
  - e.g. consonants follow vowels
  - ... like people do?
- Maybe equivalent results in the end
  - .. i.e. algorithm, not computational theory

Biological Relevance of Models

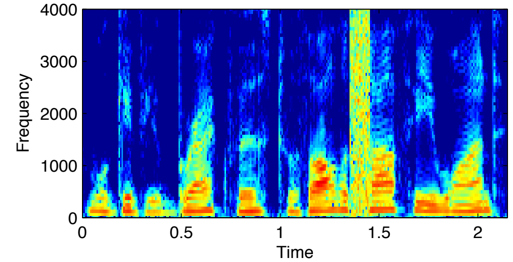
- How do we explain illusions?
    - pulsation threshold



    - sinewave speech



    - phonemic restoration



- Something is providing the missing (illusory) pieces

Summary

- Scene Analysis is possible only thanks to constraints
  - most sound combinations are unlikely
- Listeners care about individual sources
  - .. in a wide range of combinations
- Statistical source models can be learned from the environment
  - exactly how is more of a detail...