Recognition & Organization of Speech and Audio

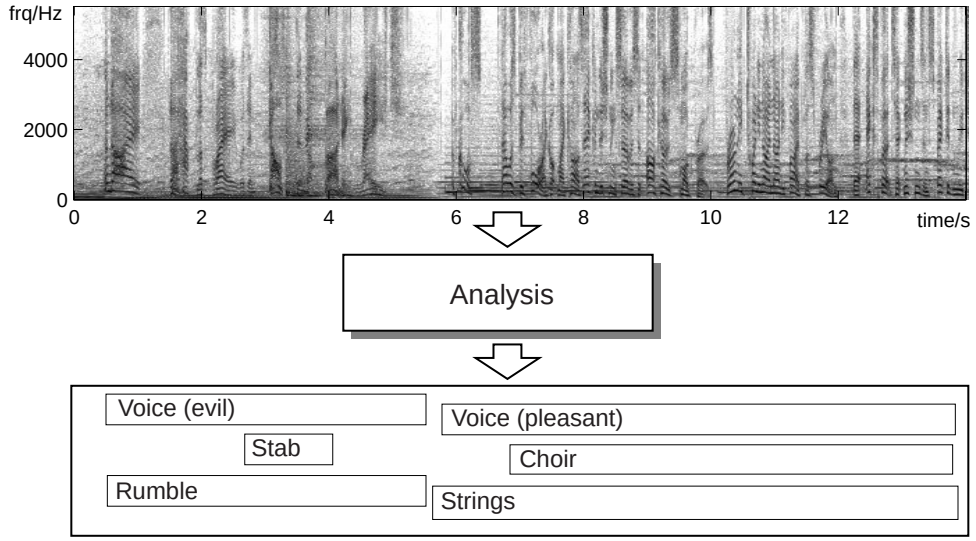
Dan Ellis
Electrical Engineering, Columbia University
<dpwe@ee.columbia.edu>

Outline

1. **Sound 'organization'**
2. **Background & related work**
3. **Existing projects**
4. **Future projects**
5. **Summary & conclusions**

Lab ROSA

1 Organization of sound mixtures

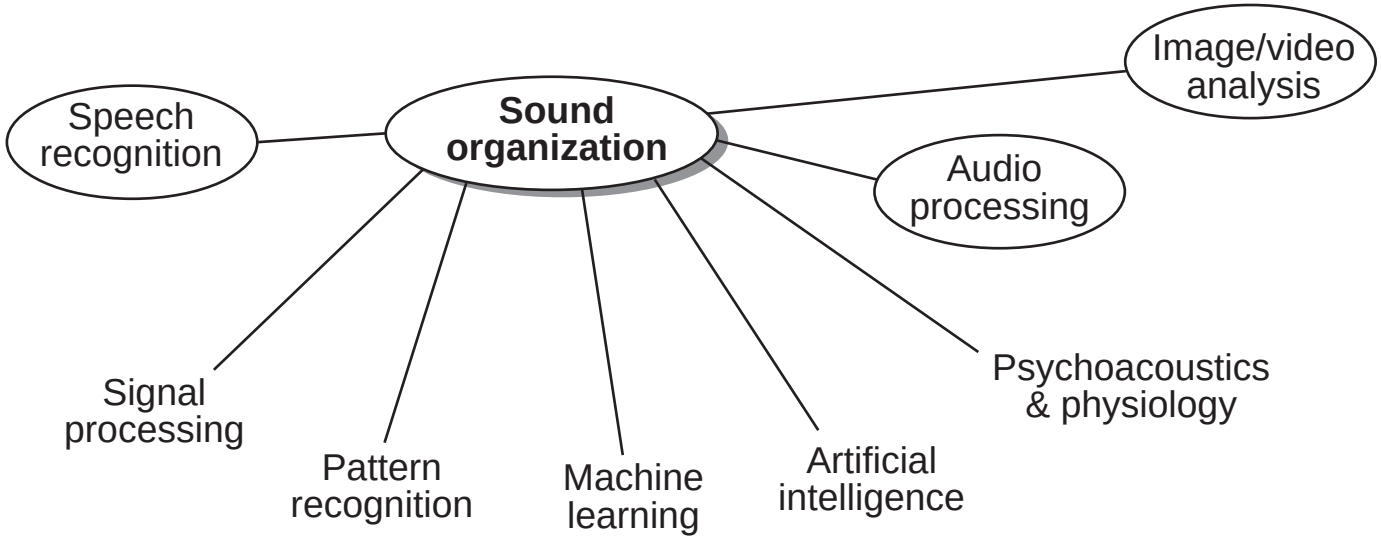


- **Core operation:**

  Converting continuous, scalar signal into discrete, symbolic representation

Lab ROSA

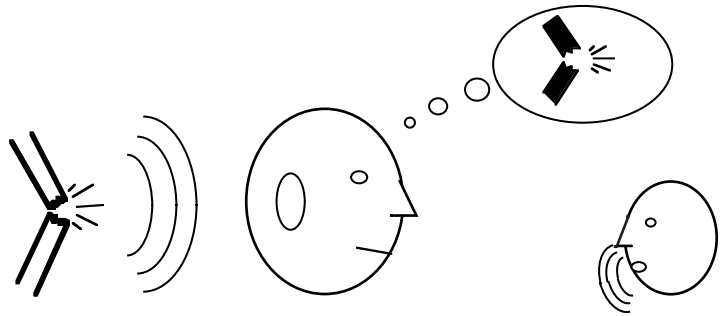
Positioning sound organization



- **Draws on many techniques**
- **Abuts/overlaps various areas**

Lab ROSA

About auditory perception



- **Received waveform is a mixture**
  - two sensors, N signals ...
  - need knowledge-based constraints
- **Psychoacoustics:**
  the study of human sound organization
  - 'auditory scene analysis' (Bregman'90)
- **Auditory perception is ecologically grounded**
  - scene analysis is preconscious ($\rightarrow$ illusions)
  - perceived organization:
    real-world objects + events (transient)
  - subjective *not* canonical (ambiguity)

Lab ROSA

Key themes for LabROSA

- **Sound organization**
  - recovering/constructing abstraction hierarchy
    - at an instant (sources)
    - along time (segmentation)
- **Scene analysis**
  - need to find attributes according to objects
  - use attributes to form objects
  - ... plus constraints of knowledge
- **Exploiting large data sets (the ASR lesson)**
  - supervised/labelled: pattern recognition
  - unsupervised: structure discovery, clustering
- **Special-purpose cases:**
  - speech recognition
  - source-specific recognizers
- **... within a 'complete explanation'**

Applications for sound organization

What do people do with their ears?

- **Robots**
  - intelligence requires awareness
  - Sony's AIBO: dog-hearing
- **Human-computer interface**
  - .. includes knowing when (& why) you've failed
- **Archive indexing & retrieval**
  - pure audio archives
  - true multimedia content analysis
- **Content 'understanding'**
  - intelligent classification & summarization
- **Autonomous monitoring**
- **Broader 'structure discovery' algorithms**

Outline

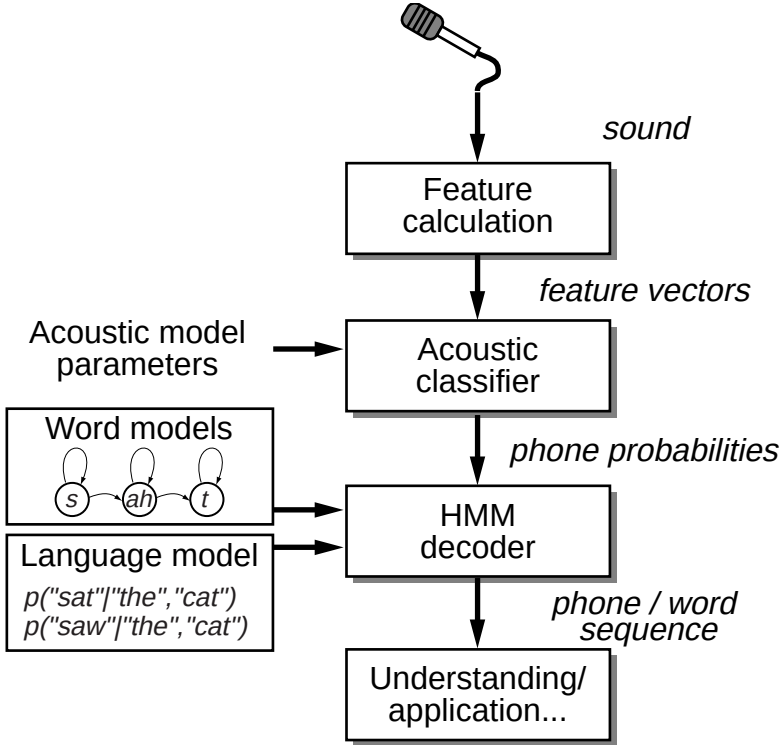
1. Sound ‘organization’
2. **Background & related work**
   - Audio coding & compression
   - Automatic Speech Recognition
   - Computational Auditory Scene Analysis
   - Multimedia information retrieval
3. Existing projects
4. Future projects
5. Summary & conclusions

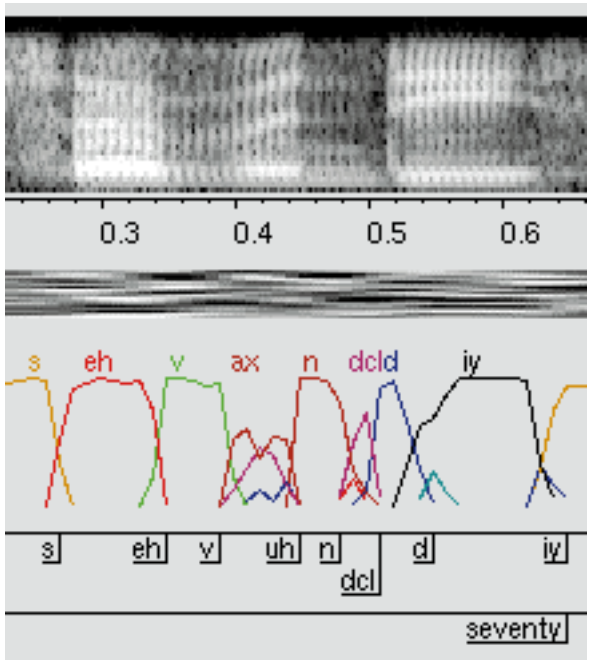
Audio coding & compression

- **Goal is reconstruction, not abstraction**
- **But criteria are 'subjective':**
  want same *percept*, not same waveform
- **MPEG-Audio:**
  - filterbanks
  - information-theoretic coding
  - psychoacoustic masking of quantization noise
- **MPEG-4 'Structured Audio'**
  - computer music synthesis model
  - instrument definition + control stream
  - automatic analysis?

Automatic Speech Recognition (ASR)

- **Standard speech recognition structure:**





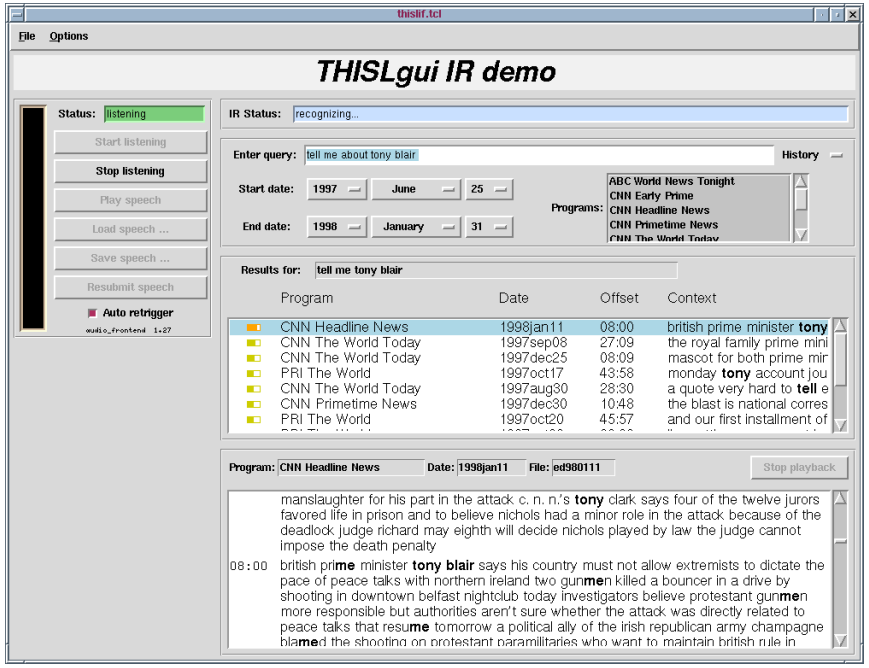
DATA

- **'State of the art' word-error rates (WERs):**
  - 2% (dictation) - 30% (telephone conversations)
- **Segmentation of speech & nonspeech**
  - ... recognizer wouldn't notice!

Lab ROSA

Spoken document retrieval

- **Text-based IR on ASR transcripts**
  - e.g. news broadcasts (CMU's Informedia, Thisl)

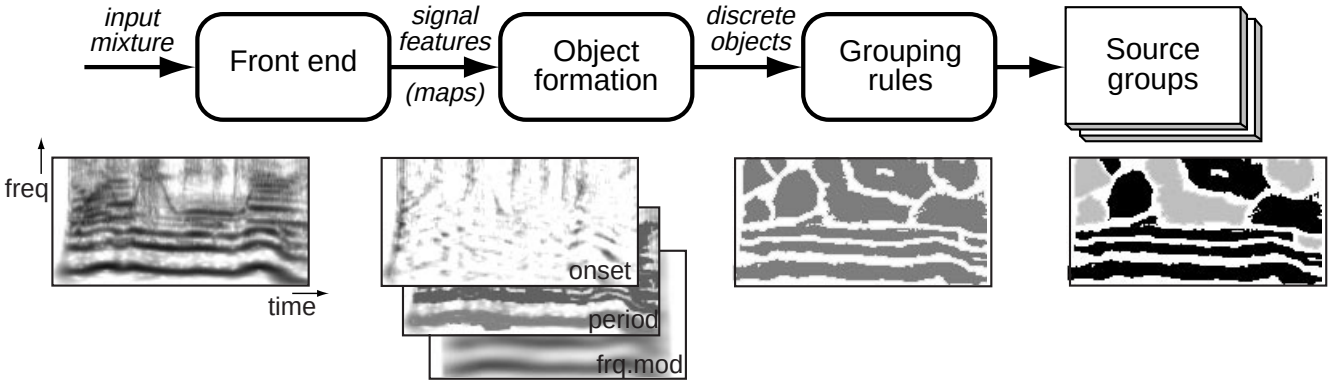


- **Recognition errors are not the limiting factor**
  - TREC-98 results: average precision 0.5→0.4
- **Weak at word level, but OK over paragraphs**
  - replay the audio, don't show the text!

Lab ROSA

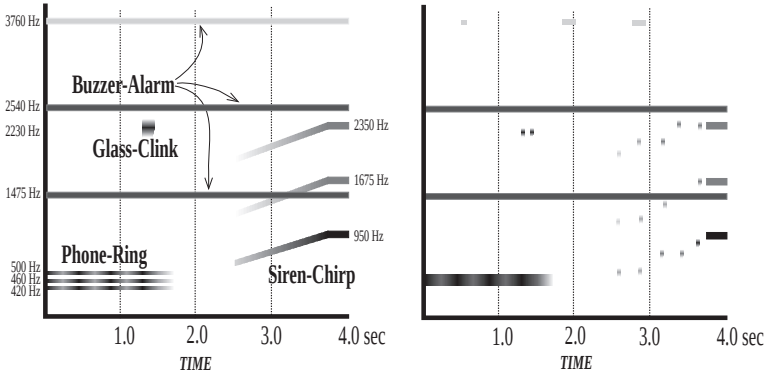
Computational Auditory Scene Analysis (CASA)

- **Implement psychoacoustic theory? (Brown'92)**



  - what are the features? how are they used?

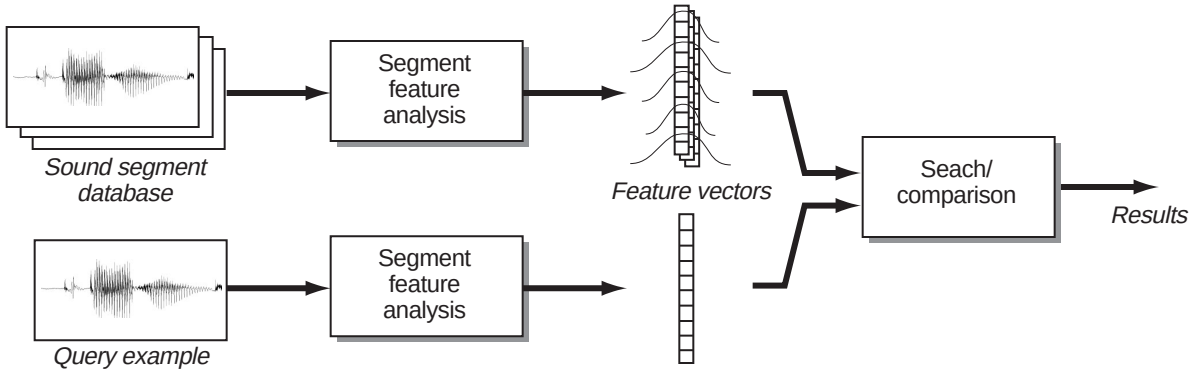
- **Top down constraints are needed (Klassner'96)**



Lab ROSA

Audio Information Retrieval

- **Searching in a database of audio**
  - speech .. use ASR
  - text annotations .. search them
  - sound effects library?
- **e.g. Muscle Fish "SoundFisher" browser**
  - define multiple 'perceptual' feature dimensions
  - search by proximity in (weighted) feature space



  - features are 'global' for each soundfile, no attempt to separate mixtures
  - segmentation...

Lab ROSA

Music analysis

- **Automatic transcription (score recovery)**
  - classic ‘hard problem’: can people do it even?
  - recent success in reduced forms
    e.g. melody, drum track (Goto’00)
- **Instrument identification**
  - ideas from speaker identification (basic PR)
    + instrument family hierarchies (Martin’99)
- **Fingerprinting**
  - spot recordings despite noise, distortion
  - relies on *perceptual* invariants
- **Music clustering**
  - e.g. music recommendation based on signal
  - correlate objective features with user ratings?

Multimedia description

- **MPEG-7 'Metadata'**
  - MPEG is known for audio/video *compression* standards;
  - also develop standards for *search and indexing*
- **MPEG-7 is a standard format for *metadata*: Well-defined categories for content description**
- **Focus is on framework & infrastructure**
- **Audio descriptor categories:**
  - from ASR
  - from computer music community
  - uses still to emerge

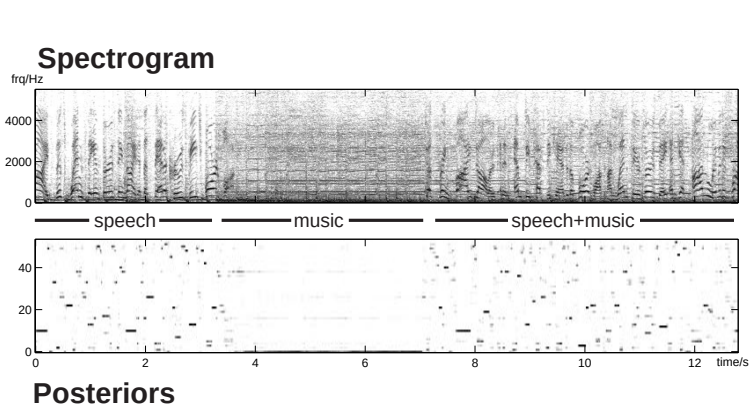
Outline

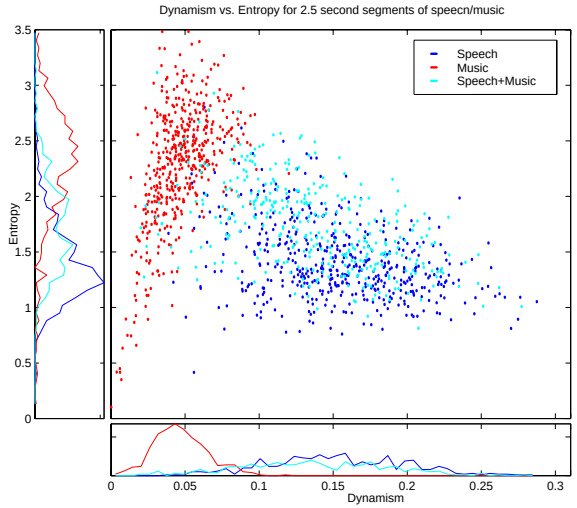
1. Sound ‘organization’
2. Background & related work
3. **Existing projects**
   - Acoustic change detection
   - Robust speech recognition
   - Nonspeech event detection
   - Prediction-driven CASA
4. Future projects
5. Summary & conclusions

Acoustic change detection

(with Williams/Sheffield, Ferreiros/UPMadrid)

- **Approaches:**
  - ‘metric’: find instants of maximal change
  - ‘model-based’: best alignment of model set
  - ‘bayesian’: generate models when warranted





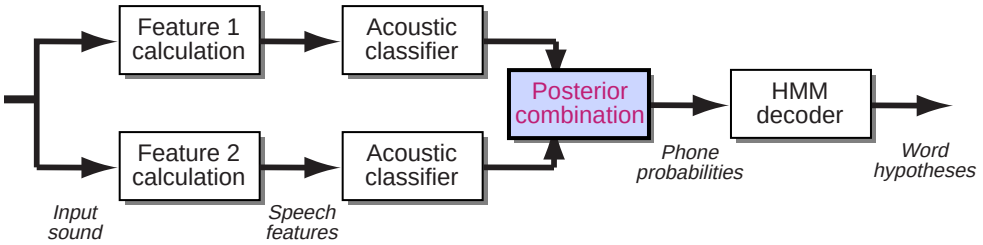
- **Typically agnostic about underlying problem**
  - use any features, find any changes
- **Good for ASR adaptation, otherwise...**

Lab ROSA

Speech feature combination

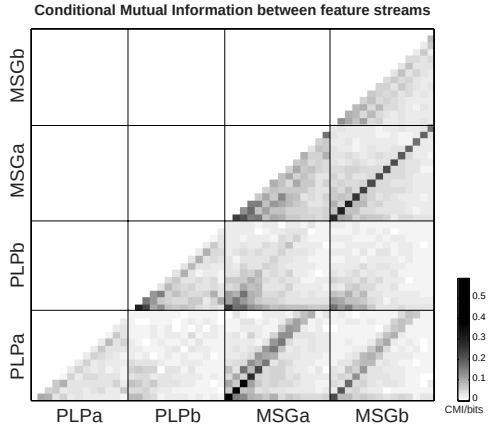
(with Bilmes/UW, Hermansky/OGI, ICSI)

- **'Multistream' approaches**



  - streams can correct each other → big gains

- **Which feature streams to combine?**
  - *low* mutual information between *classifiers* indicates complementary streams

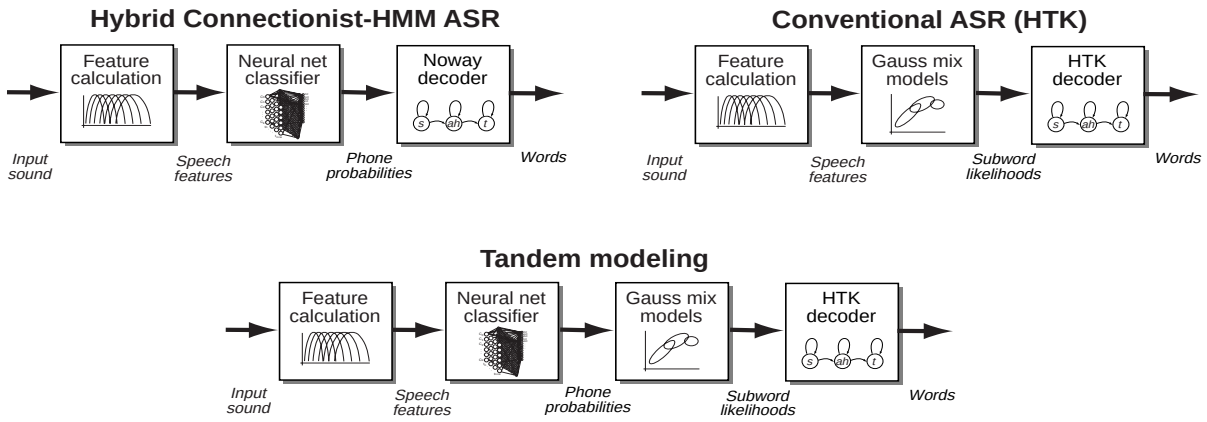


Lab ROSA

Tandem speech recognition

(with Hermansky, Sharma & Sivadas/OGI, Singh/CMU)

- **Neural net estimates phone posteriors; but Gaussian mixtures model finer detail**
- **Combine them!**

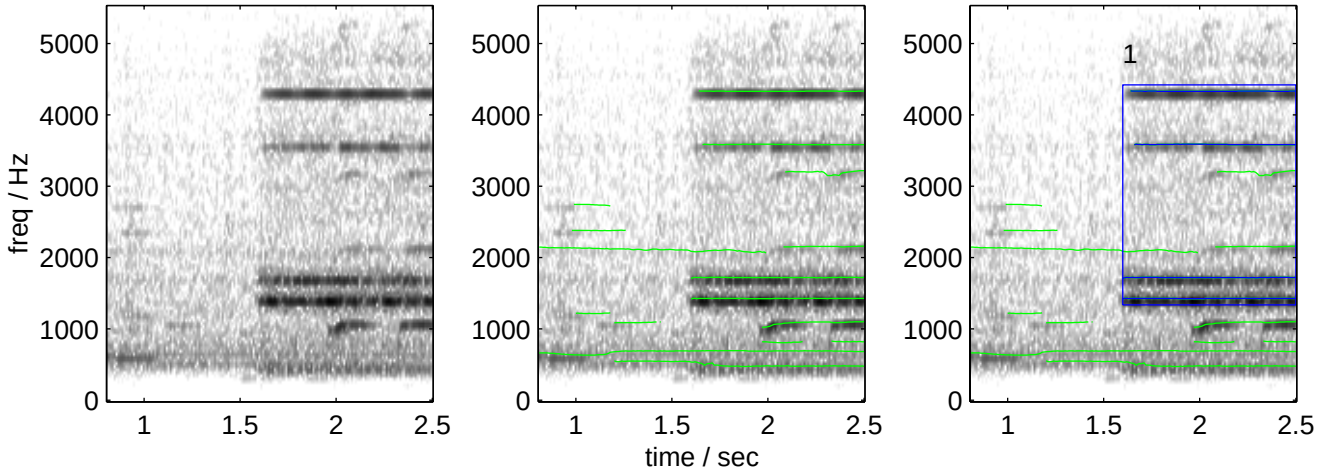


- **50% relative improvement over GMMs alone**
  - different statistical modeling schemes get different info from same training data

Lab ROSA

Alarm sound detection

- **Deconstructing sound mixtures**



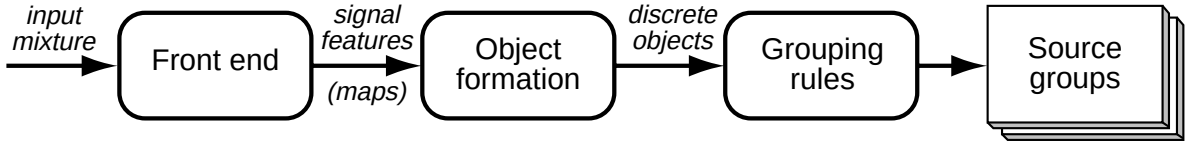
  - representation of energy in time-frequency
  - formation of atomic elements
  - grouping by common properties (onset &c.)

- **Alarm sounds have particular structure**
  - people ‘know them when they hear them’
  - build a generic detector?

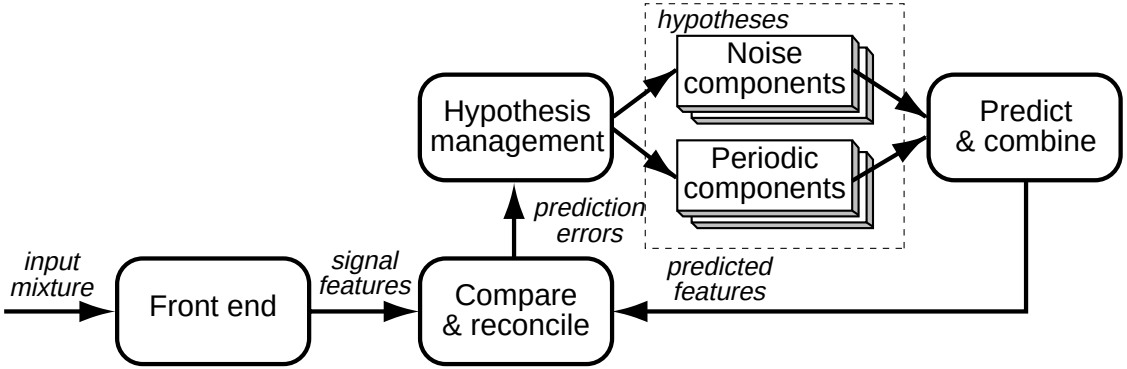
Lab ROSA

Prediction-driven CASA

- **Data-driven (bottom-up) fails for noisy, ambiguous sounds (most mixtures!)**



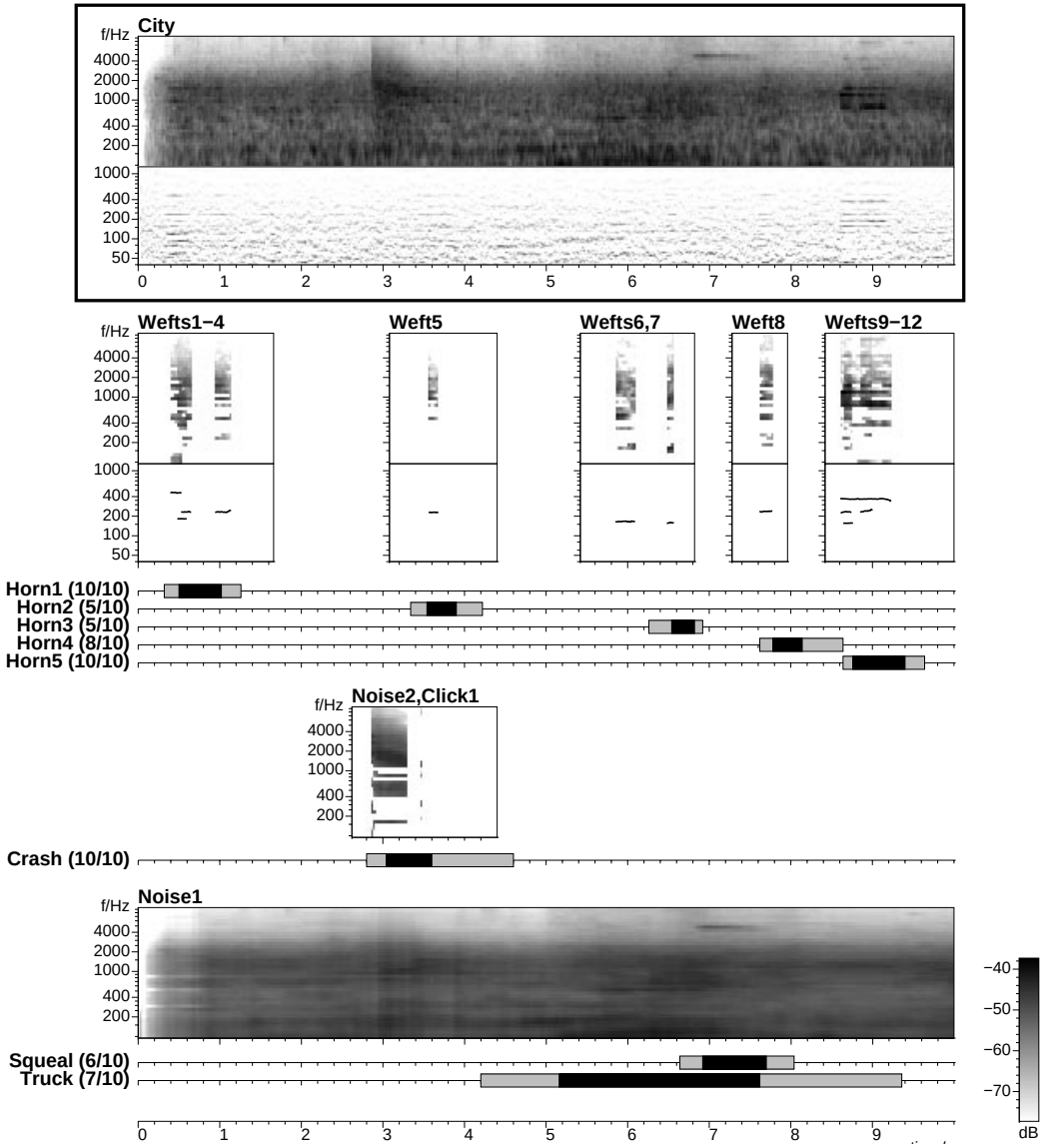
- **Need top-down constraints:**



  - fit vocabulary of generic elements to sound ... bottom of a hierarchy?
  - account for entire scene
  - driven by prediction failures
  - pursue alternative hypotheses

Lab ROSA

PDCASA example



Lab ROSA

Outline

1. Sound ‘organization’
2. Background & related work
3. Existing projects
4. **Future projects**
   - *somewhat concrete*
     - Meeting Recorder
     - Missing-data recognition & CASA for ASR
     - Structure from audio-video features
   - *provisional*
     - Speech & speaker recognition
     - Music organization
     - Audio archive structure discovery
5. Summary & conclusions

Meeting recorder

(with ICSI, UW, SRI, IBM)

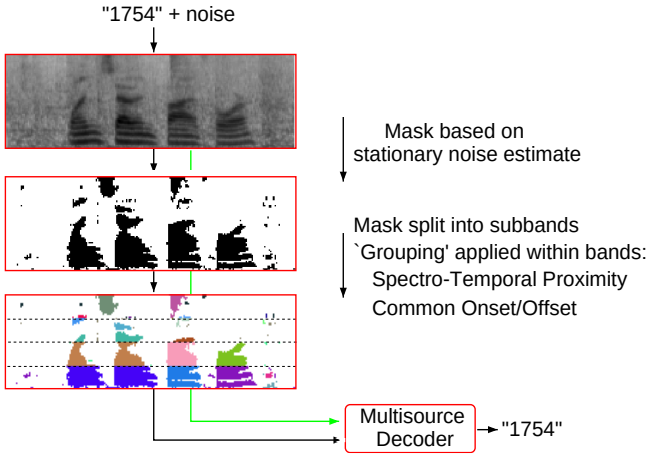
- **Microphones in conventional meetings**
  - for transcription/summarization/retrieval
  - informal, overlapped speech
- **Data collection (ICSI and ...):**

- **Research: ASR, nonspeech, organization**
  - unprecedented data, new applications

Missing data recognition & CASA

(with Barker, Cooke, Green/Sheffield)

- **Missing-data recognition**
  - integrate across ‘don’t-know’ values
  - ‘perfect’ mask → excellent performance in noise
- **Multi-source decoder**
  - Viterbi search of sound-fragment interpretations



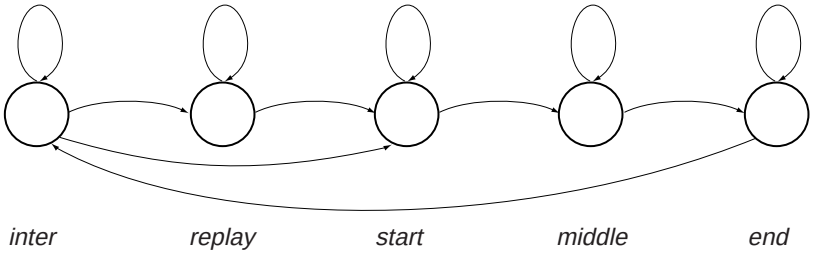
- **CASA for masks/fragments**
  - larger fragments → quicker search

Lab ROSA

Structure from audio-video features

(Peng Xu)

- **HMM modeling of sports video**



- **Distribution of camera motion labels, color features**
  - also need within-state sequential structure
- **Add features from audio**
  - could be orthogonal/complementary
- **Audio feature toolkit?**
  - simple feature vectors, boundaries, classes
  - wealth of potential applications!

Speech & speaker recognition

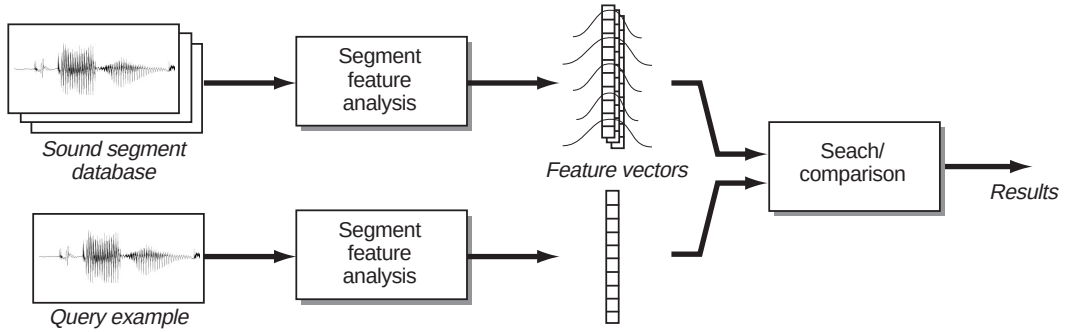
- **Words are not enough;**
  Confidence-tagged alternate word hypotheses
- **Other useful information:**
  - speaker change detection
  - speaker characterization
  - phrasing & timing
  - prosodic cues to dialog state
  - laughter, pauses, etc.
- **Integration with other analyses**
  - segmentation for adaptation
  - nonspeech events to ignore
  - video-derived information...

Music organization

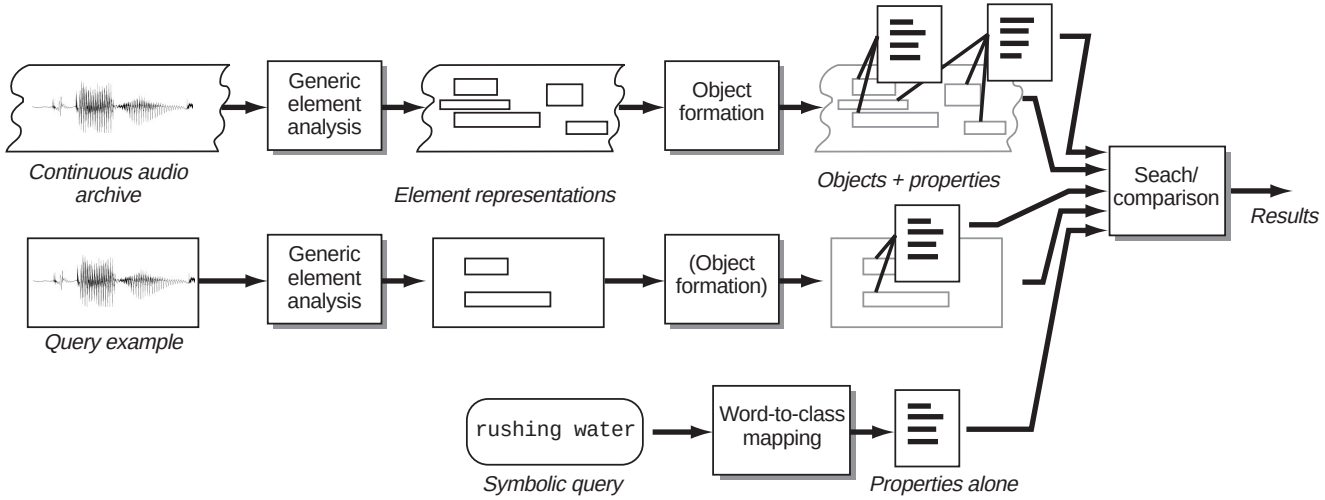
- **Music is a special case**
  - lots of structure
  - highly significant
- **Trick is to find meaningful, tractable questions**
  - boundary between speech and music?
- **New (counter-intuitive) approaches?**
  - perceive as whole, not by voice (Scheirer'00)
    $\rightarrow$ global features for chord structures
  - generic 'event' cues + local feature classification
  - more provisional notion of instruments/voices

CASA for audio retrieval

- **Muscle Fish system uses global features:**



- **Mixtures → need elements & objects:**



- **features calculated on grouped subsets**

Lab ROSA

Audio archive structure discovery

- **What can you do with a large unlabeled training set (e.g. multimedia clips from the web)?**
  - bootstrap learning: look for common patterns
  - have to learn generalizations in parallel: e.g. self-organizing maps, EM HMMs
  - post-filtering by humans may find ‘meaning’ in clusters
- **Associated text annotations provide a very small amount of labeling**
  - .. but for a very large number of examples – sufficient to obtain purchase?
  - maximize label utility through NLP-type operations (expansion, disambiguation etc.)
  - goal is automatic term-to-feature mapping for term-based content queries

Outline

1. Sound ‘organization’
2. Background & related work
3. Existing projects
4. Future projects
5. **Summary & conclusions**

Summary

DOMAINS

- Broadcast
- Movies
- Lectures
- Meetings
- Personal recordings
- Location monitoring

ROSA

- Object-based structure discovery & learning
- Speech recognition
- Speech characterization
- Nonspeech recognition
- Scene analysis
- Audio-visual integration
- Music analysis

APPLICATIONS

- Structuring
- Search
- Summarization
- Awareness
- Understanding

Conclusions

- **Sound is more than just speech!**
  - speech is a special case
  - most auditory perceivers don't understand speech
- **Object-based analysis is critical**
  - it's what people do
  - the world presents acoustic mixtures
- **Whole-scene representation is the way**
  - it's what people do
  - provides mutual constraints of overlap
- **Broad range of approaches for a broad range of phenomena**