Mapping Meetings: Columbia's plans

Dan Ellis

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

Outline

1 Columbia participants

- **Dan Ellis**
  Laboratory for Recognition and Organization of Speech and Audio (LabROSA)
  - signal processing, abstraction and indexing
  - speech recognition
  - computational auditory scene analysis
- **Kathy McKeown**
  Nautral Language Processing (NLP) group
  - text analysis, information extraction, generation
- **“Mapping meetings” support**
  - 2 students, 3 academic years

LabROSA

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

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

Natural Language Processing Group

- **4 faculty & senior researchers**
  13 Ph.D. students + MS ...
- **Summarization**
  - Sports News task: max info-per-word
- **Shallow parsing**
  - phrase-based (not sentences)
  - statistical
- **Intersection of similar documents**
  - common phrases → summary
  - identify contradictions etc.

Outline

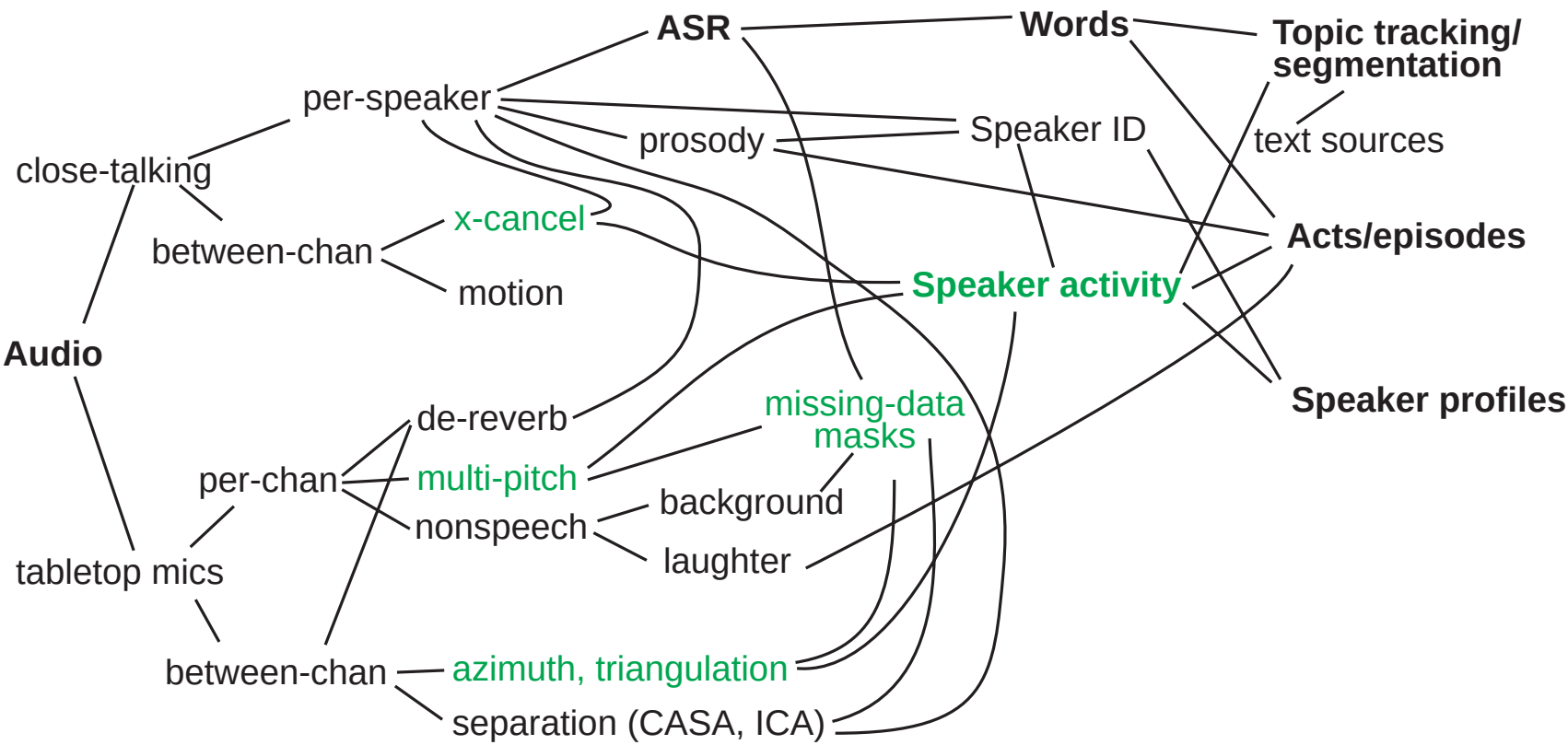
1. Columbia Participants
2. **Mapping meetings: Perspectives**
   - project themes
   - broad plans
3. Techniques
4. Summary

Lab ROSA

Mapping meetings: Project themes

- **Discourse analysis**
  - interaction patterns: predicted, emergent
  - individuals .. pairs .. groups
  - words, speech style, timing, ... (speaker-relative)
- **Summarization & abstraction**
  - topic content
  - “meeting acts”
  - participant roles (...)
- **Browsing & visualization**
  - access methods / dimensions
  - scale (time, detail); pov orientation
  - data classes & types; rendering

Information flow



Lab ROSA

Signal organization plans

- **Speaker turns from channel energy envelopes**
  - mixing matrix inversion
- **Extracting sources from mixed signals**
  - source ID from spatial cues, pitch
  - feature/mask extraction $\rightarrow$ missing-data recog.
- **Distant signal recognition**
  - tandem modeling
  - channel compensation
  - multi-source recognition

NLP plans

(Kathy McKeown)

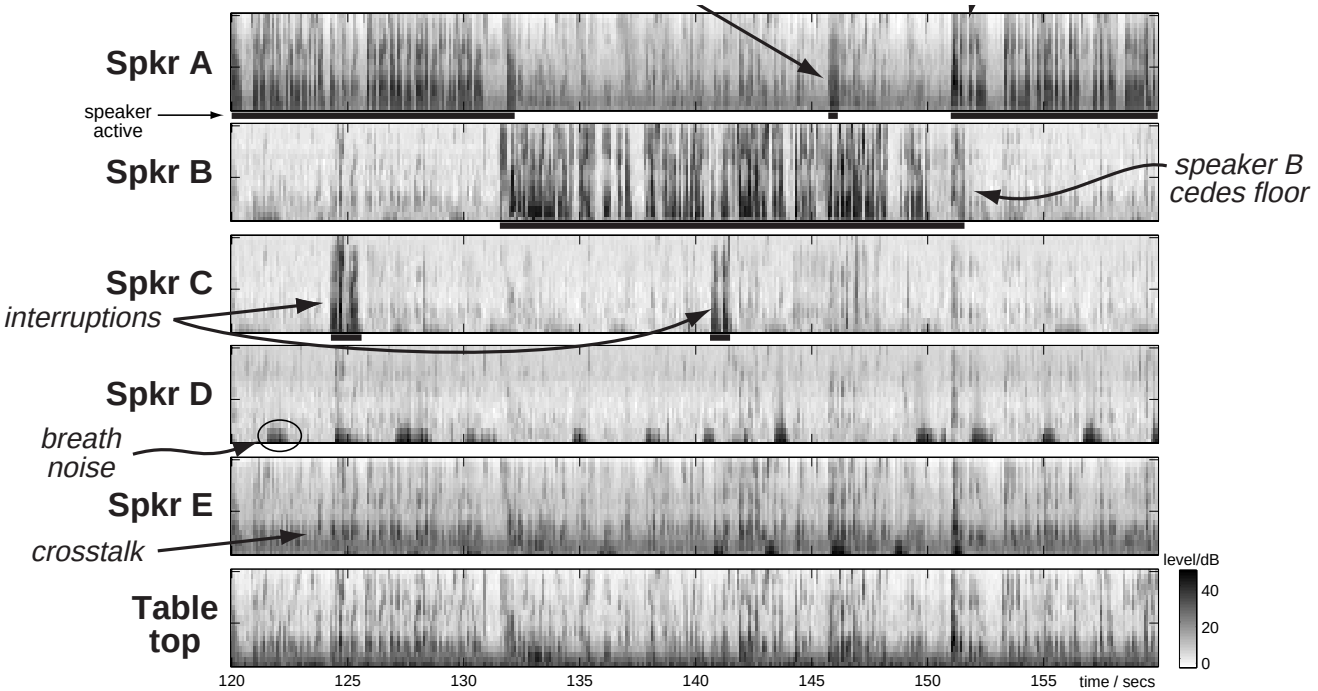
- **Argumentative Summaries**
  - special case of summarization
  - same theme, different perspectives
- **Outcomes**
  - identify points of disagreement
  - identify resulting consensus
- **Depends on**
  - information extraction: words, discourse, prosody
- **Manpower**
  - one student starting in January

Outline

1. Columbia Participants
2. Mapping Meetings: Perspectives
3. **Techniques**
   - speaker turn detection
   - per-speaker signal extraction
   - speech recognition
4. Summary

Crosstalk cancellation

- **Baseline speaker activity detection is hard:**



- **Noisy crosstalk model:** $\mathbf{m} = \mathbf{C} \cdot \mathbf{s} + \mathbf{n}$
- **Estimate column C_{xA} from A's peak energy**
  - ... including pure delay (10 ms frames)
  - ... then linear inversion

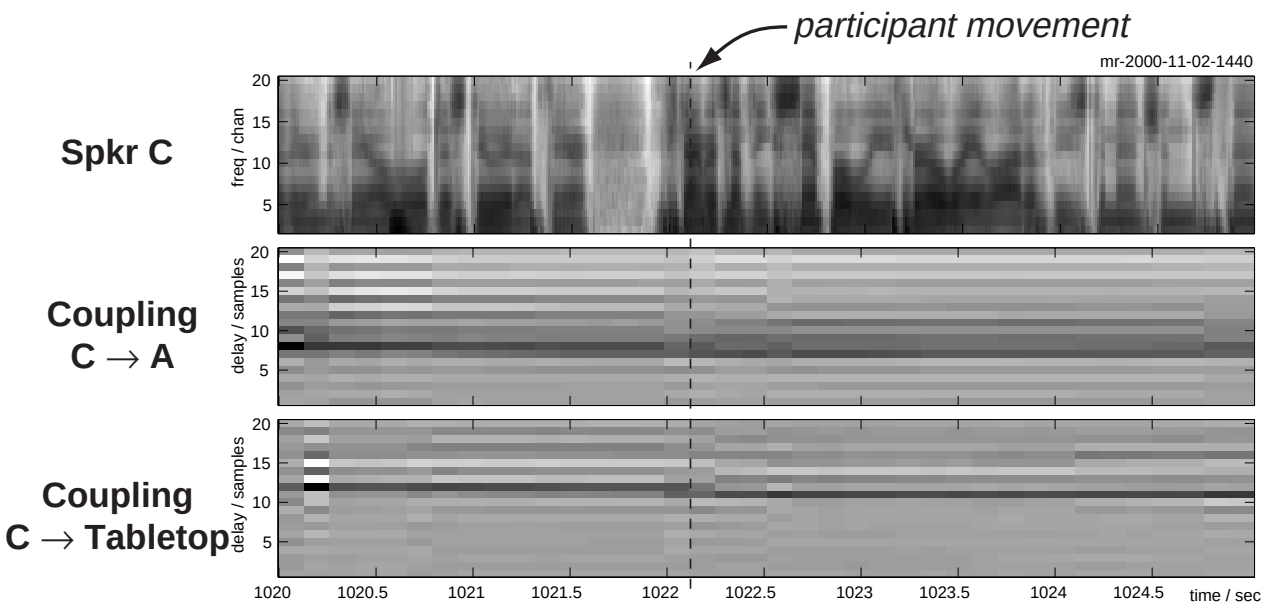
Lab ROSA

Passive motion detection

- **Cross-correlation recovers impulse response**

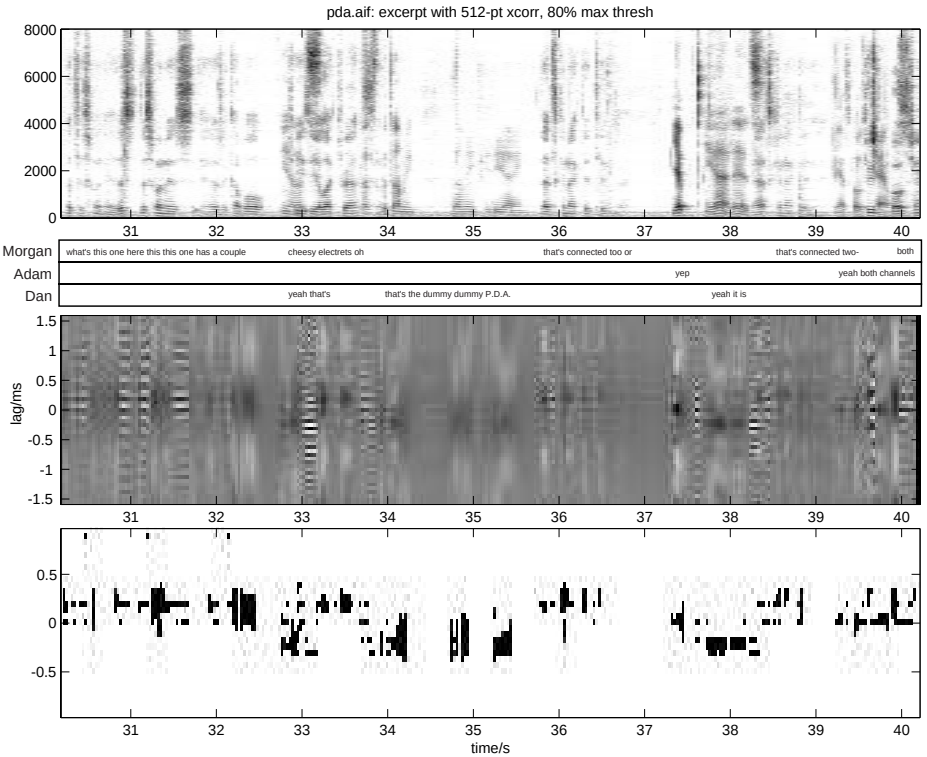
$$S_{xy} = H_{xy} \cdot S_{xx}$$

- **Coupling to each mic gives distances; can infer which are moving**



PDA-based speaker change detection

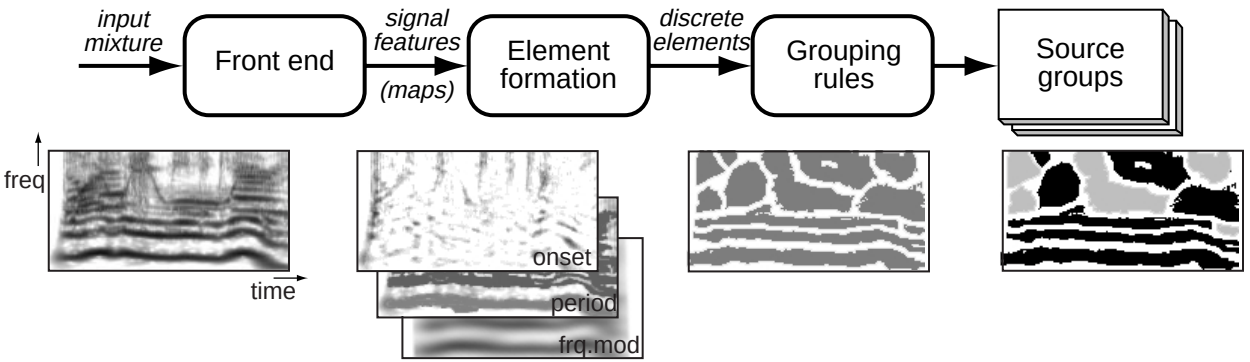
- **Goal: small conference-tabletop device**
- **Speaker turns from PDA mock-up signals?**



- **SCD algo on spectral + interaural features**
  - average spectral + per-channel ITD, $\Delta\phi$

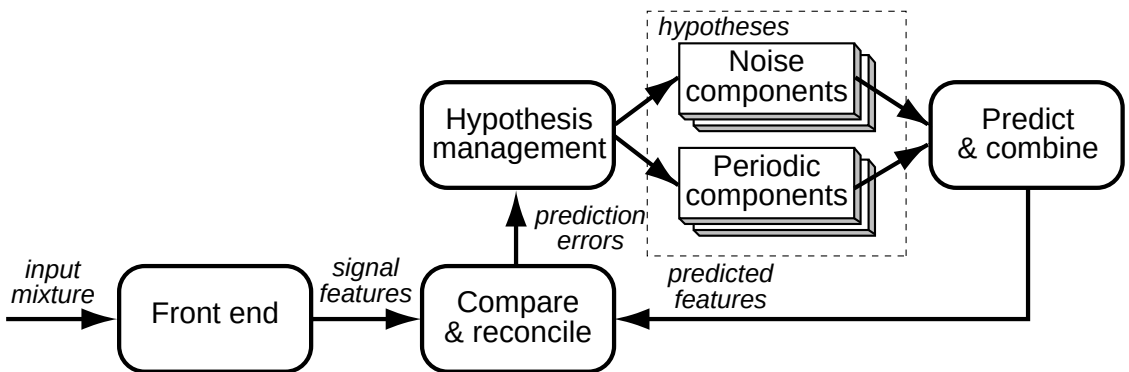
Computational Auditory Scene Analysis (CASA)

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



  - what are the features? how are they used?

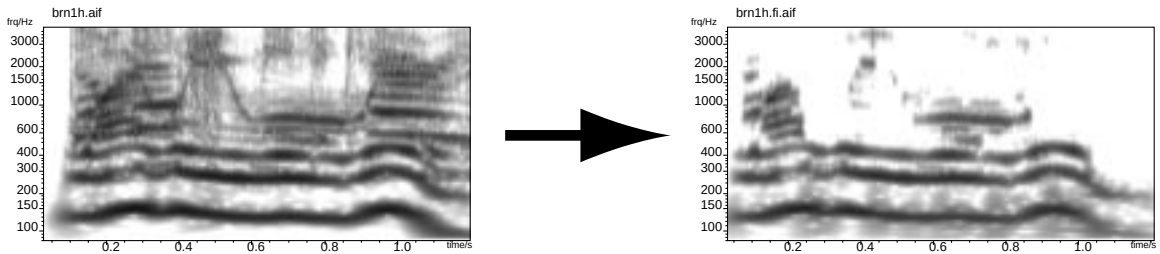
- **Need top-down constraints:**



Lab ROSA

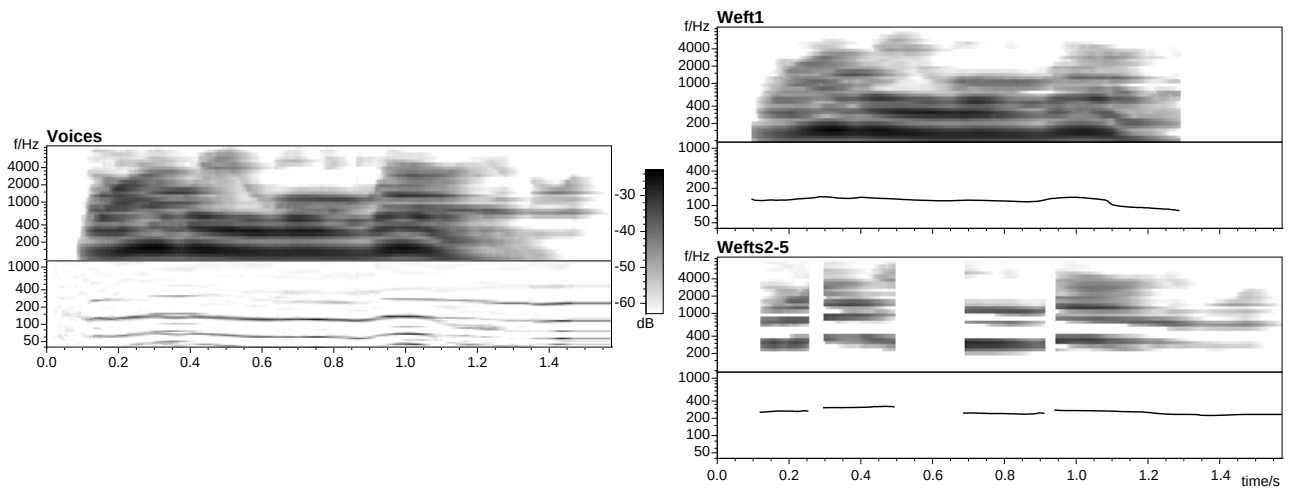
Pitch-based monaural separation

- **Bottom-up**



  - time-frequency cells tagged with fundamental
  - delete cells not dominated by target f0

- **Model-based (wefts)**

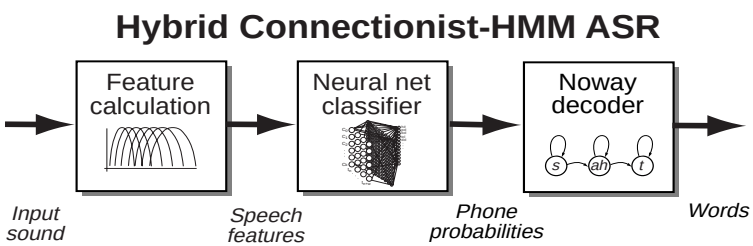


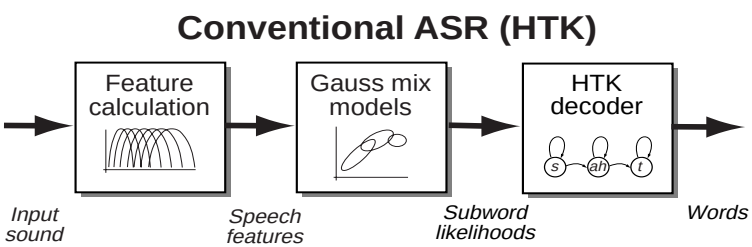
  - estimate multiple periodicities per cell

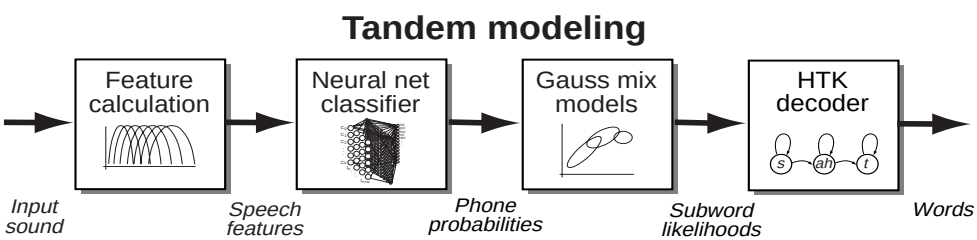
Tandem speech recognition

(with Manuel Reyes)

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



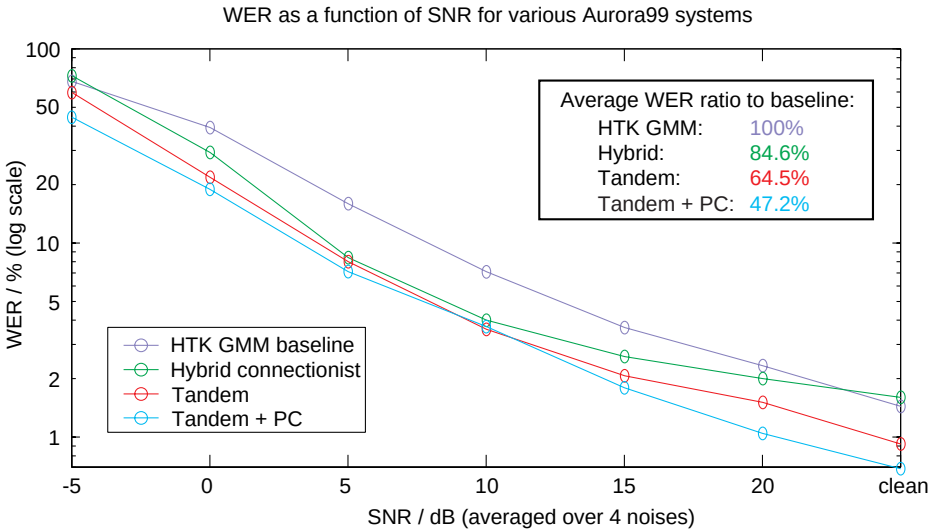


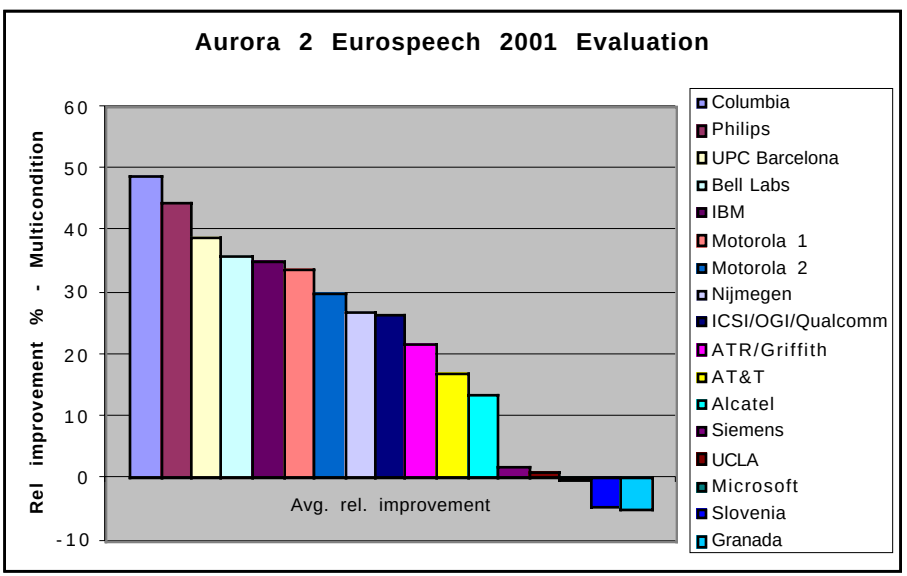


- **Train net, then train GMM on net output**
  - GMM is ignorant of net output ‘meaning’

Lab ROSA

Tandem system results: Aurora digits





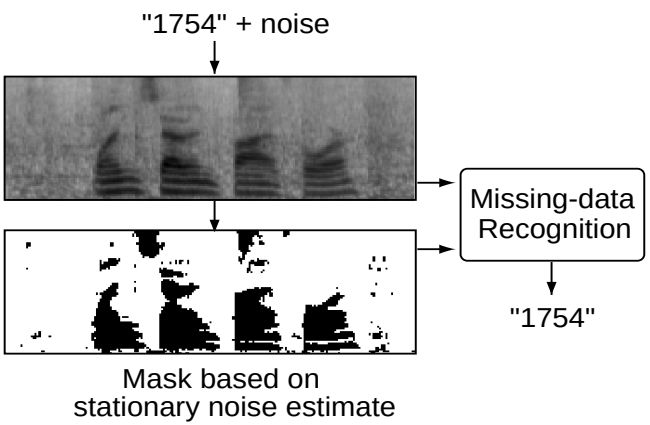
Missing data recognition

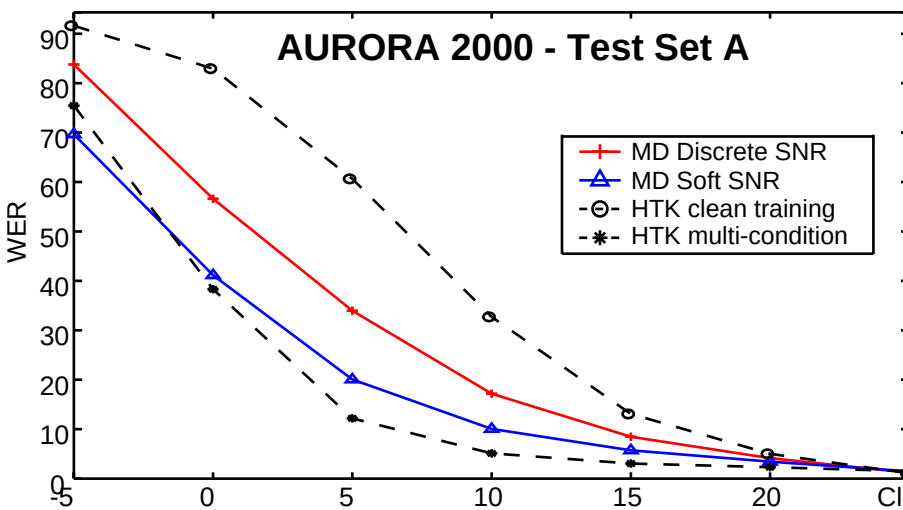
(with Cooke, Green, Barker @ Sheffield)

- **Noisy training seems to miss the point**
  - rather have single 'clean' models
- **Use missing feature theory...**
  - integrate over missing data dimensions x_m

$$p(q|x_o) = \int p(q|x_o, x_m) p(x_m|x_o) dx_m$$

  - trick is finding good/bad data mask
  - soft classification improves

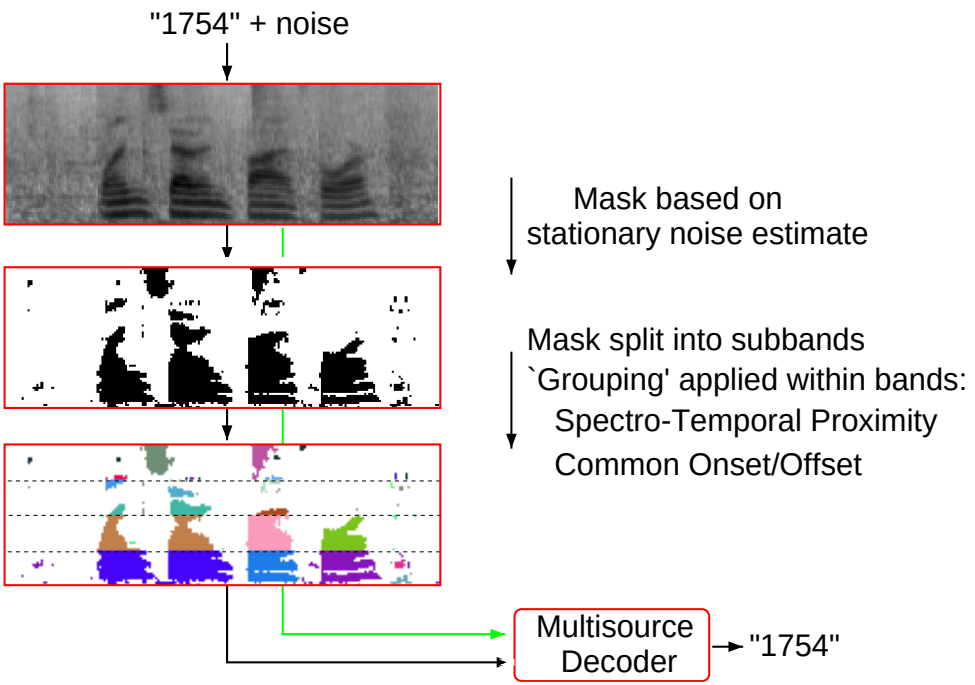




Multi-source decoding

(Jon Barker @ Sheffield)

- **Search of sound-fragment interpretations**



- **Comparing different masks**
  - evaluate $p(M,K|O) = p(M|K,O) \cdot p(K|O)$
- **CASA for masks/fragments**

Lab ROSA

Outline

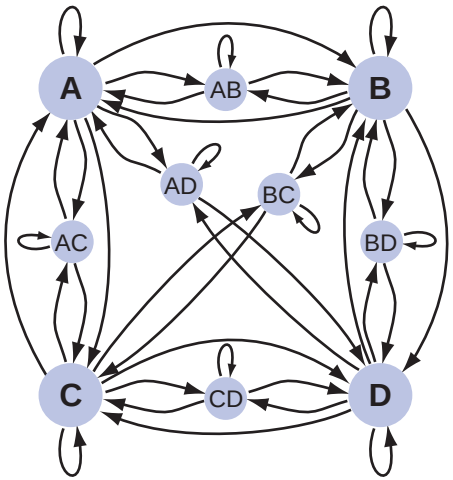
Summary

- **Columbia:**
  Audio Organization,
  Language Processing
- **Meetings:**
  Raw information sources,
  Multiple analyses
- **Audio signals (close talk / tabletop):**
  Turns
  Isolated speech
  Other information

Random ideas ...

Speaker turn (H)MM

- **Markov model for speaker changes**
  - optimal path for ambiguous cues



- **Transition matrix represents...**
  - turn durations (self-loops)
  - response patterns
- **ML choice between alternate trans. matrices**
  - detect different meeting *modes*: presentation, debate, conflict...

Marginalizing turn features

- **Each turn may have features**
  - pitch range, duration, rate, fluency
- **Features depend on speaker *and* discussion mode**
  - marginalize two ways
  - speaker-relative features indicate mode