Speaker Turns from Between-Channel Differences

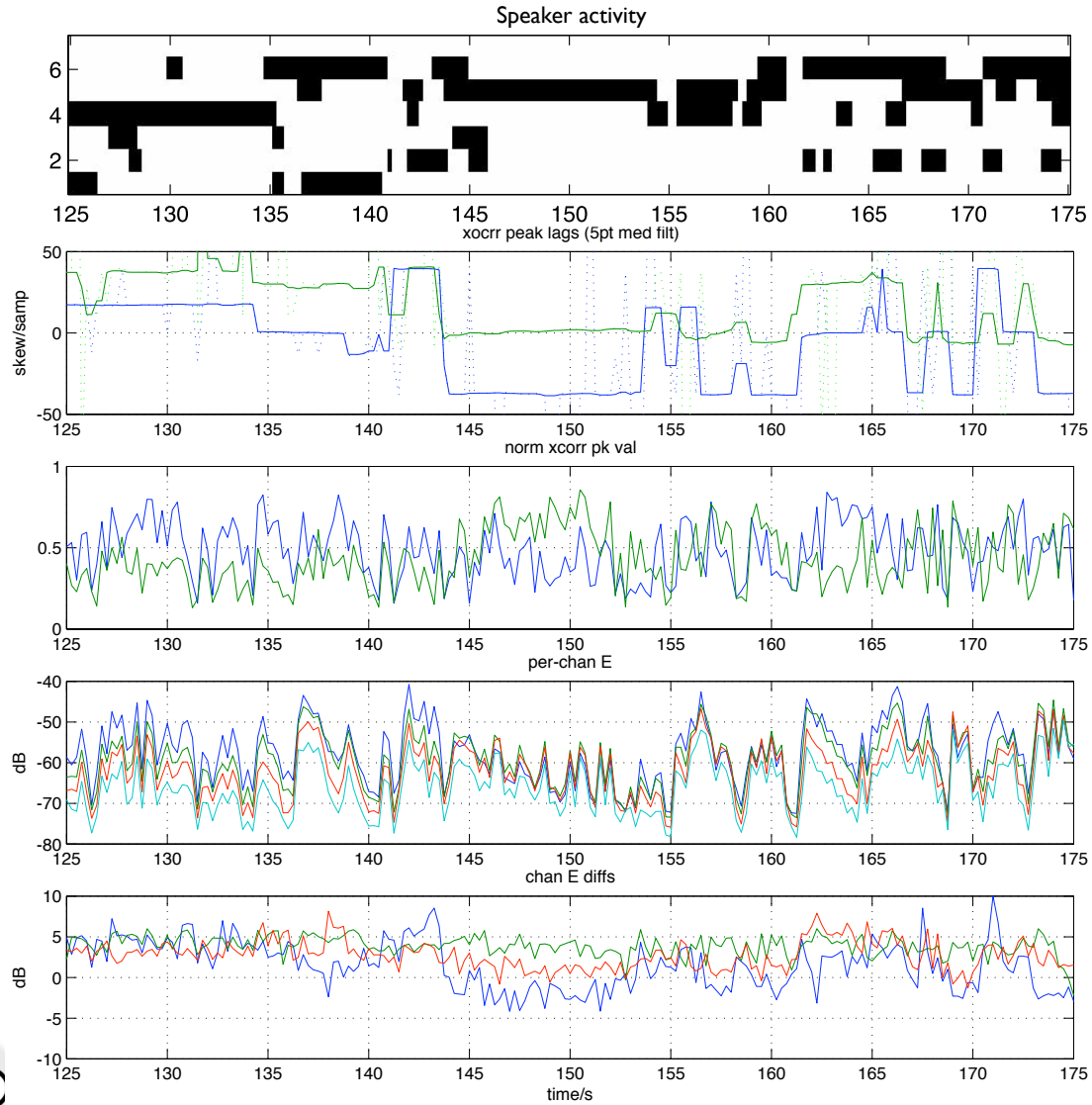
Dan Ellis & Jerry Liu
LabROSA (Columbia) & ICSI

- Between-channel features
- Timing-difference-based system
- Evaluation and results

Meeting Turn Information

- Multiple mic recordings carry information on speaker turns
  - every voice reaches every mic... (?)
  - ... but with differing coupling filters (delays, gains)
- Find turns with minimal assumptions
  - e.g. ad-hoc sensor setups (multiple PDAs)
  - differences to remove effect of source signal
    - no spectral models, < 1xRT

Between-channel cues: Timing (ITD) & Level



Speaker ground-truth

Timing diffs (ITD) (2 mic pairs, 250ms win)

Peak correlation coefficient *r*

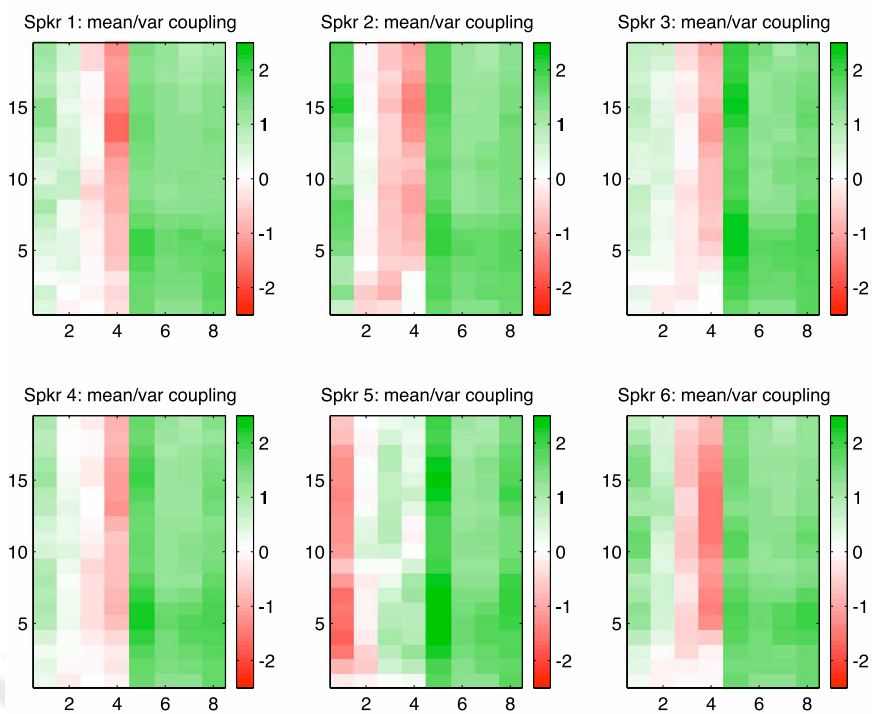
Per-channel energy

Between-channel energy differences

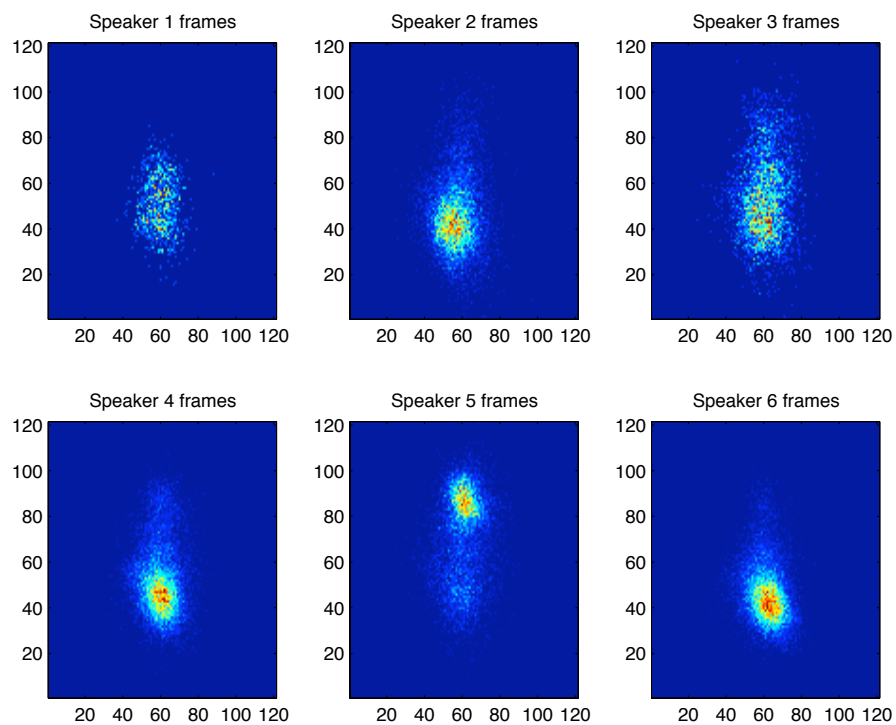
Level Cues

- Level at each mic relative to average of all
- Project *nmic* x *nfreq* arrays onto PCA
- Compare to ground truth ... poor separation

Mean/StdDev per-spkr coupling matrices



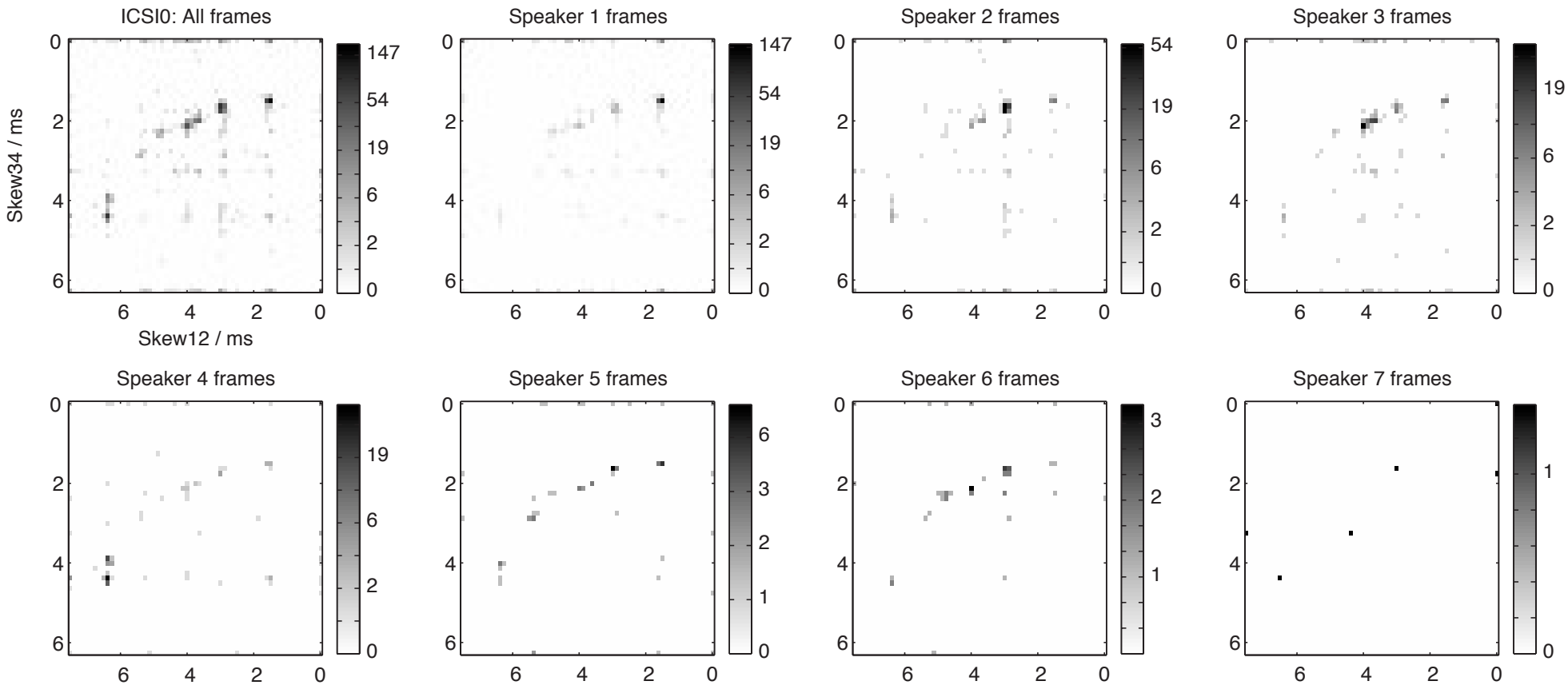
Per-spkr PCA 1,2 projections



ROSA
Laboratory for the Recognition and Organization of Speech and Audio
Columbia University
ICSI
INTERNATIONAL COMPUTER SCIENCE INSTITUTE

Timing Cues (ITD)

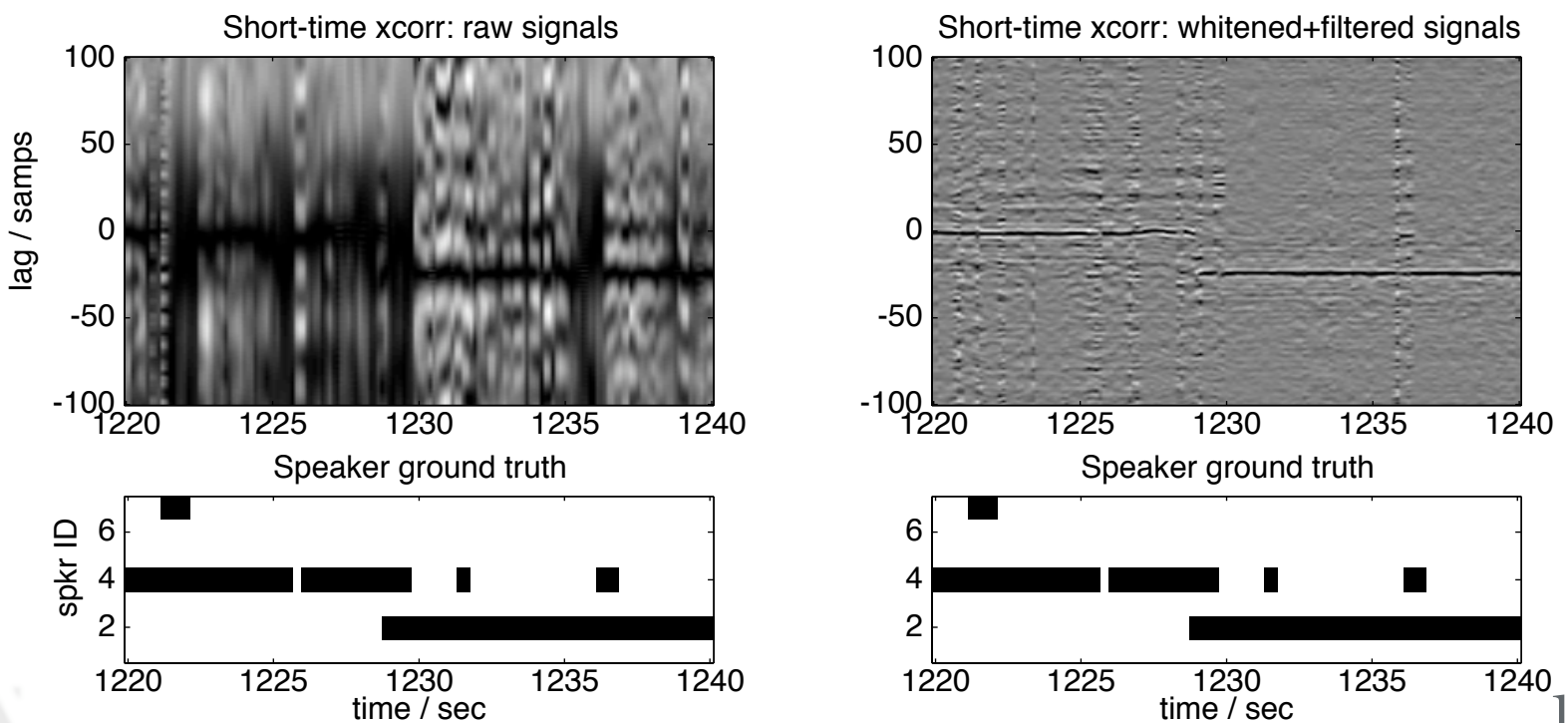
- Cross-correlation between two 2-mic pairs
- Compare to ground-truth...



- Promising, but still ambiguous (overlaps)

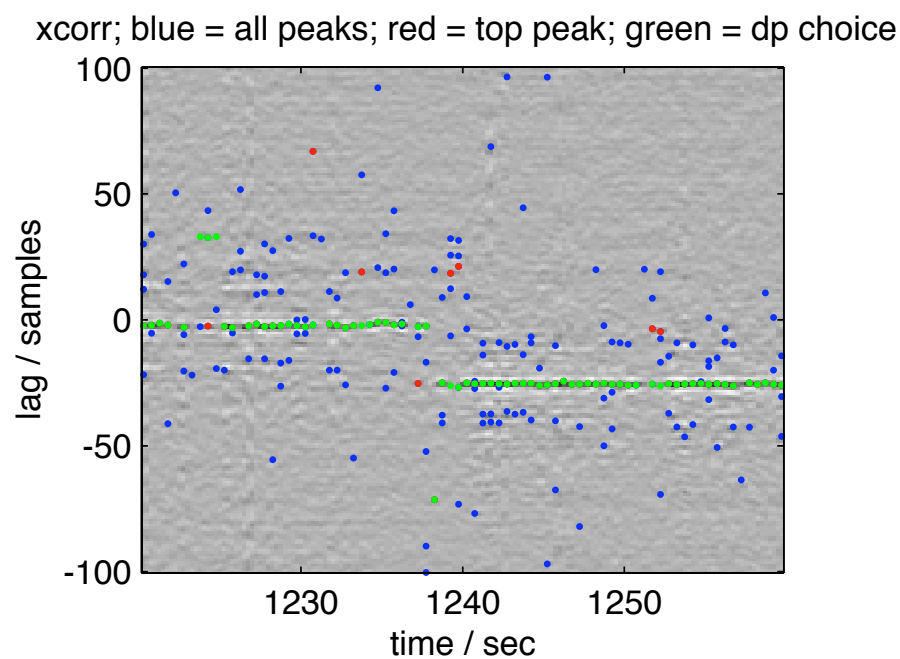
Pre-whitening for ITD

- Inverse-filter by 12-pole LPC models (32 ms windows) to remove local resonances
- Filter out noise < 500 Hz, > 6 kHz
- Then cross-correlate...



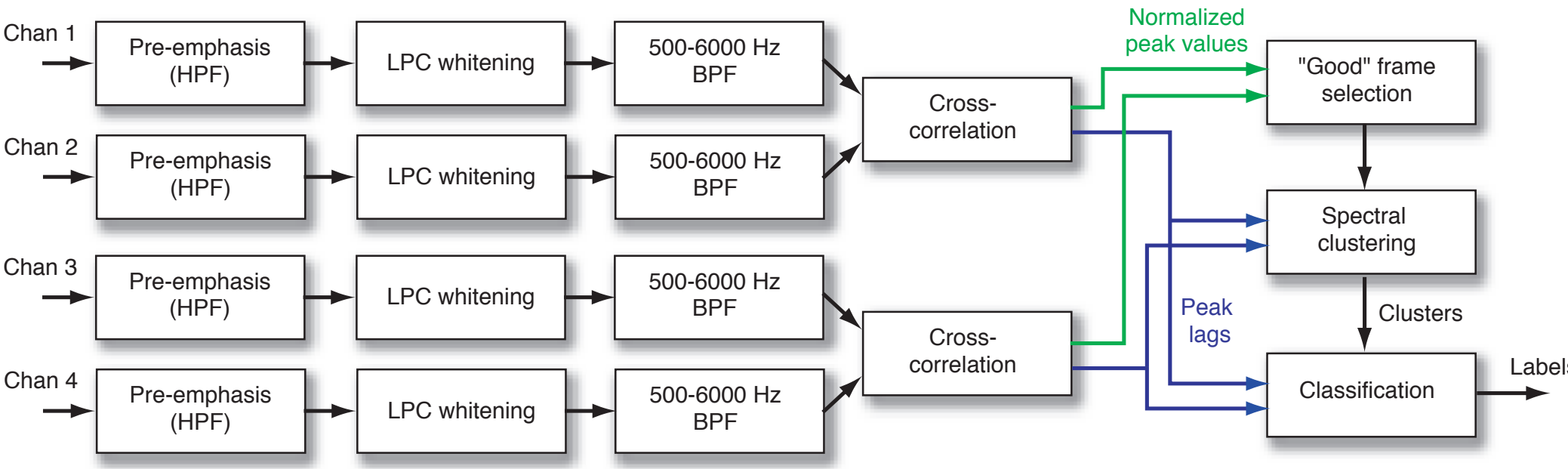
Dynamic Programming for peak continuity

- How can we exclude ITD outliers?
- Consider top N cross-corr peaks, optimize combined peak height + step cost



- Helps and hurts...

System overview



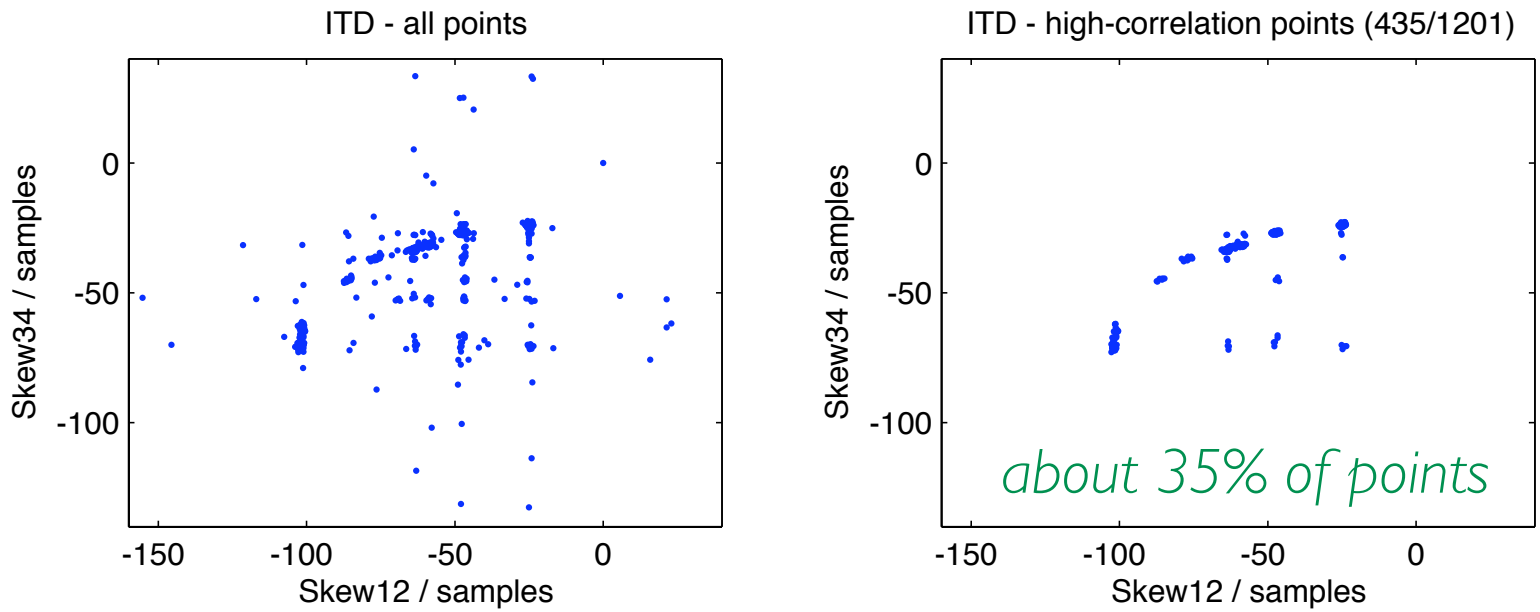
- Many noisy points remain in ITD...
- Models based on "good" (high-*r*) frames
- Spectral clustering to identify speakers
- Then classify all frame by (single) best match

Choosing "Good" Frames

- Correlation coef. r ~ channel similarity:

$$r_{ij}[\ell] = \frac{\sum_n m_i[n] \cdot m_j[n+\ell]}{\sqrt{\sum m_i^2 \sum m_j^2}}$$

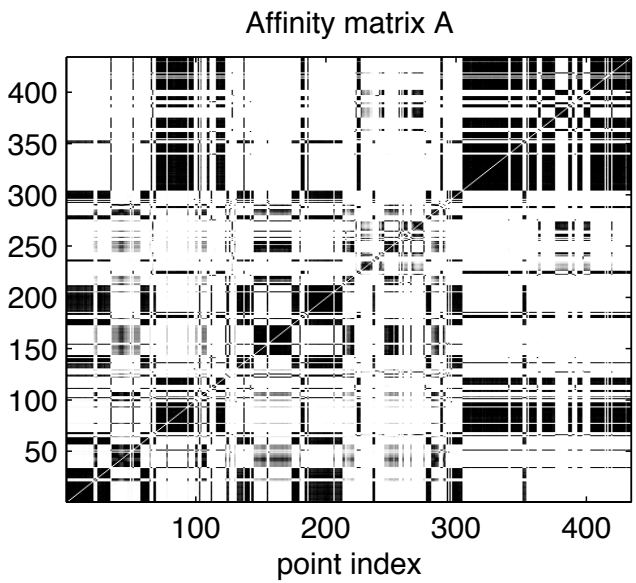
- Select frames with r in top 50% in both pairs

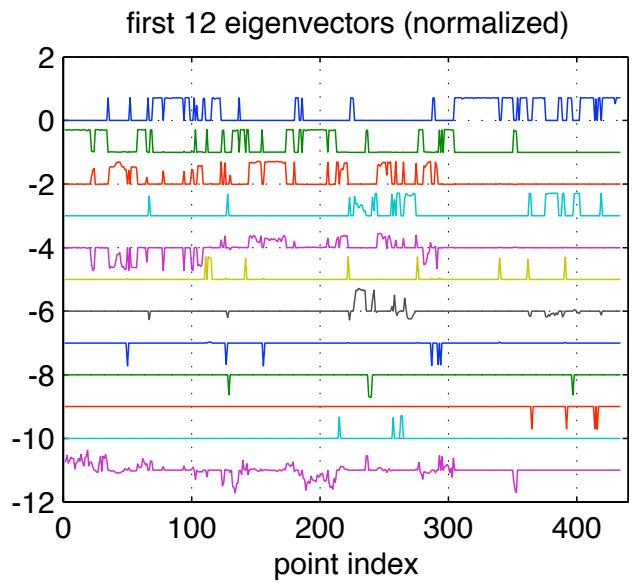


- Cleaner basis for models

Spectral clustering

- Eigenvectors of "affinity matrix" **A** to pick out similar points:



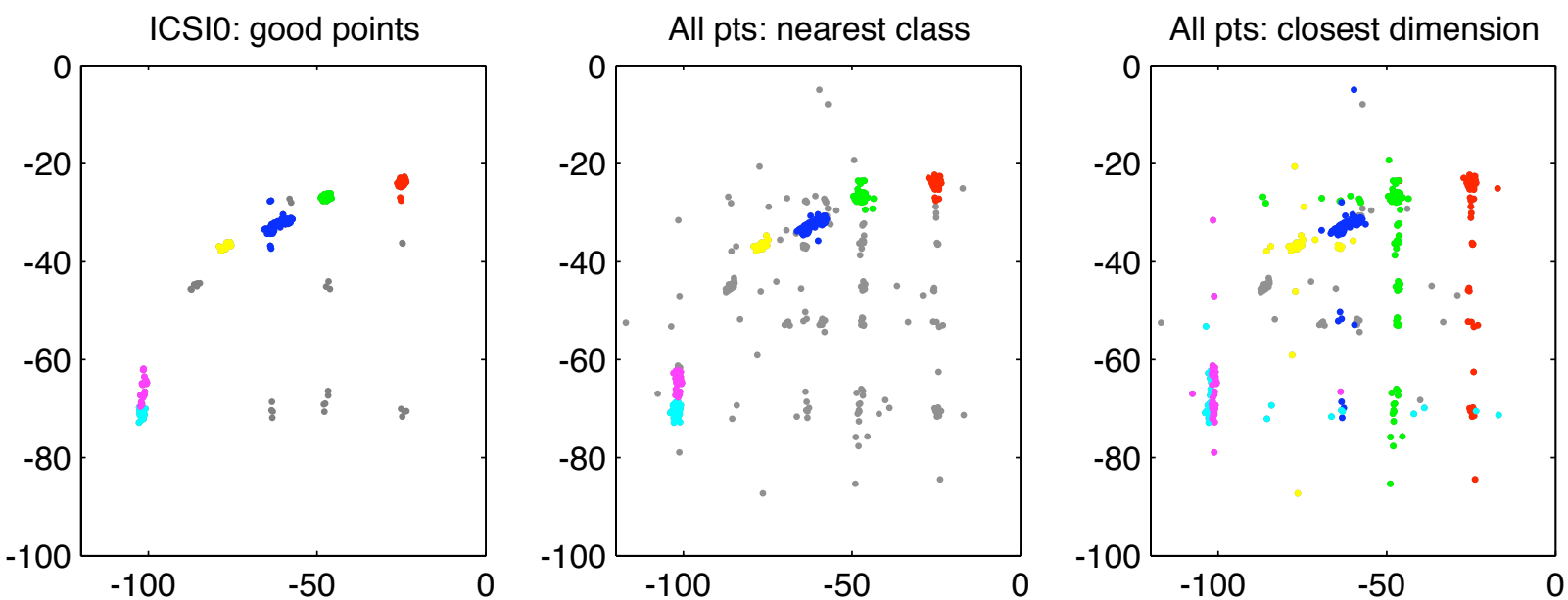


$$a_{mn} = exp\{-\|\mathbf{x}[m] - \mathbf{x}[n]\|^2 / 2\sigma^2\}$$

- Ad-hoc mapping to clusters
  - Number of clusters K from eigenvalues ≈ points

Speaker Models & Classification

- Actual clusters depend on σ and K heuristic
- Fit Gaussians to each cluster, assign that class to all frames within radius
  - or: consider dimension independently, choose best

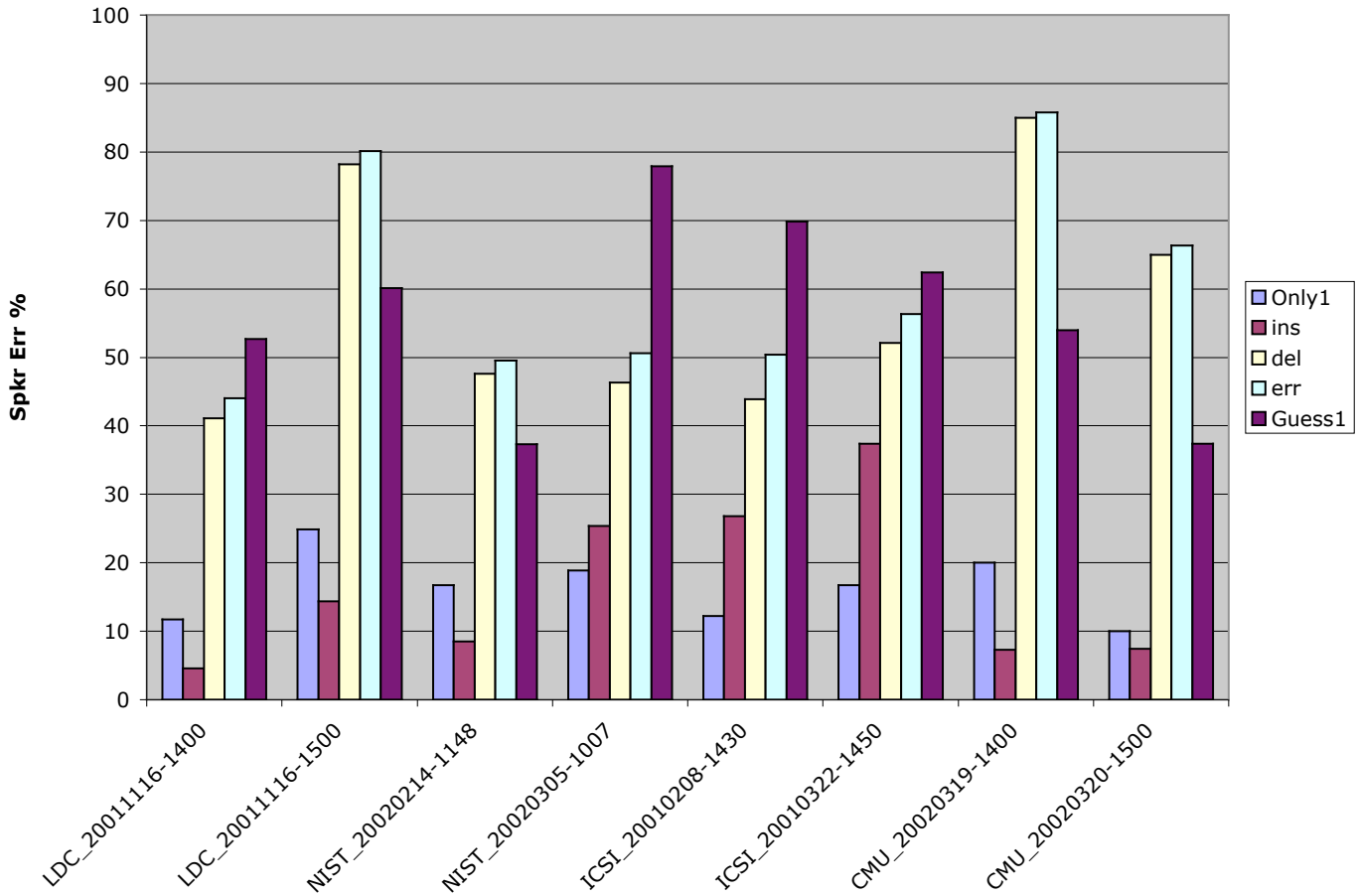


Evaluation issues

- Correspondence of results to ground truth
  - greedy selection – choose by highest overlap
- If you guess just 1 speaker is always active...
  - upper bound on error (dep. dominant speaker)
- If you report only 1 speaker/time frame
  - lower bound on error (dep. overlap)

Spkr Err%	Guess1	Only1
LDC1	52.7%	11.7%
LDC2	60.1%	24.9%
NIST1	37.3%	16.7%
NIST2	77.9%	18.9%
ICSI1	69.8%	12.2%
ICSI2	62.5%	16.7%
CMU1	54.0%	20.0%
CMU2	37.4%	10.0%

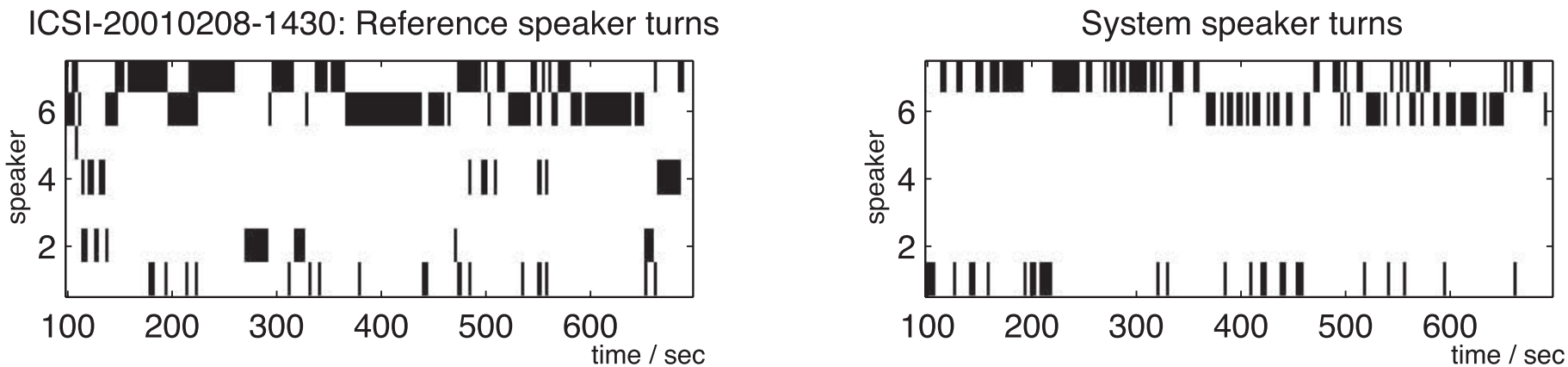
Dev-set results



- 'Best dimension' + median filter
- Worse than Guess1 in 4 out of 8 cases!
  - many deletions of outlier points

Performance Analysis

- Compare reference & system activity maps:



  - system misses quiet speakers 2,3,4 (deletions)
  - system splits speaker 6 (deletions+insertions)
  - many short gaps (deletions)

Future work

- Use more channels
  - can add mic pairs → more ITD dimensions
  - select pairs with best average correlation r
- Classify on partial data
  - classify each dimension separately & vote? (8%...)
  - throw out dimensions with lower r
- Merge clusters based on source spectrum
  - look at distribution of MFCCs in each cluster;
  - reunites speakers who move...