Building a Binaural Source Separator

MICHAEL I MANDEL, DANIEL P W ELLIS, AND TONY JEBARA
mim@ee.columbia.edu, dpwe@ee.columbia.edu, jebara@cs.columbia.edu · Columbia University, New York, NY

1 The problem

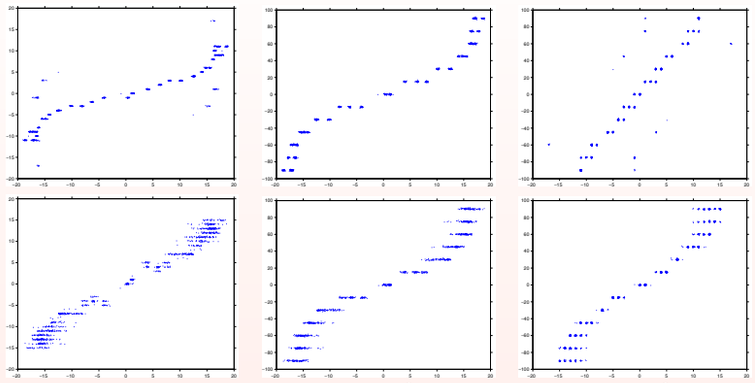
- Locate and separate sound sources using stereo recording
- Localization for acoustic scene description, pointing
- Separation for source recognition, classification, description

2 Overview of current system

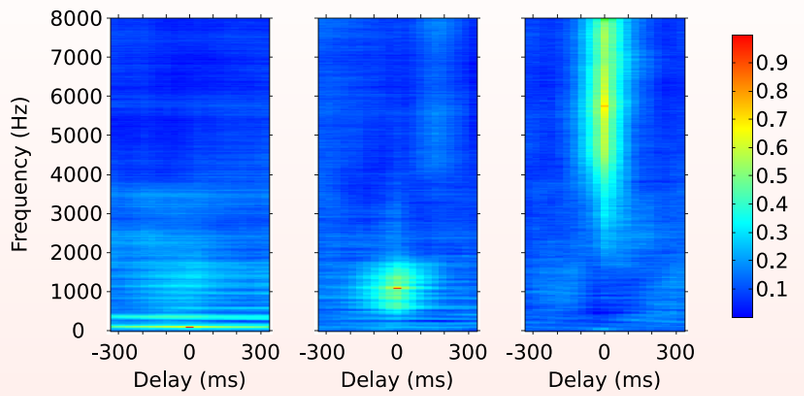
- Parametric probability model of interaural phase difference (IPD)
- Described in Mandel et al. (2006)
- EM algorithm (repeat):
  - Calculate probability of spectrogram points belonging to sources and delays given current parameter estimates
  - Re-estimate parameters for each source and delay to maximize the total likelihood given those memberships

3 Shortcomings

- There are frequencies at which the IPD alone cannot distinguish between two sources
- Nearby points in the spectrogram tend to come from the same source, i.e. are not independent
- Localization information available in the interaural level difference (ILD) is not being used
- Segmentation information available in each of the monaural inputs alone is not being used



Relationships between ILD(a), ITD (τ), and direction of arrival (θ) for high frequencies (top) and low frequencies (bottom).



The correlation between points at three frequencies and their **neighbors** in ground truth masks.

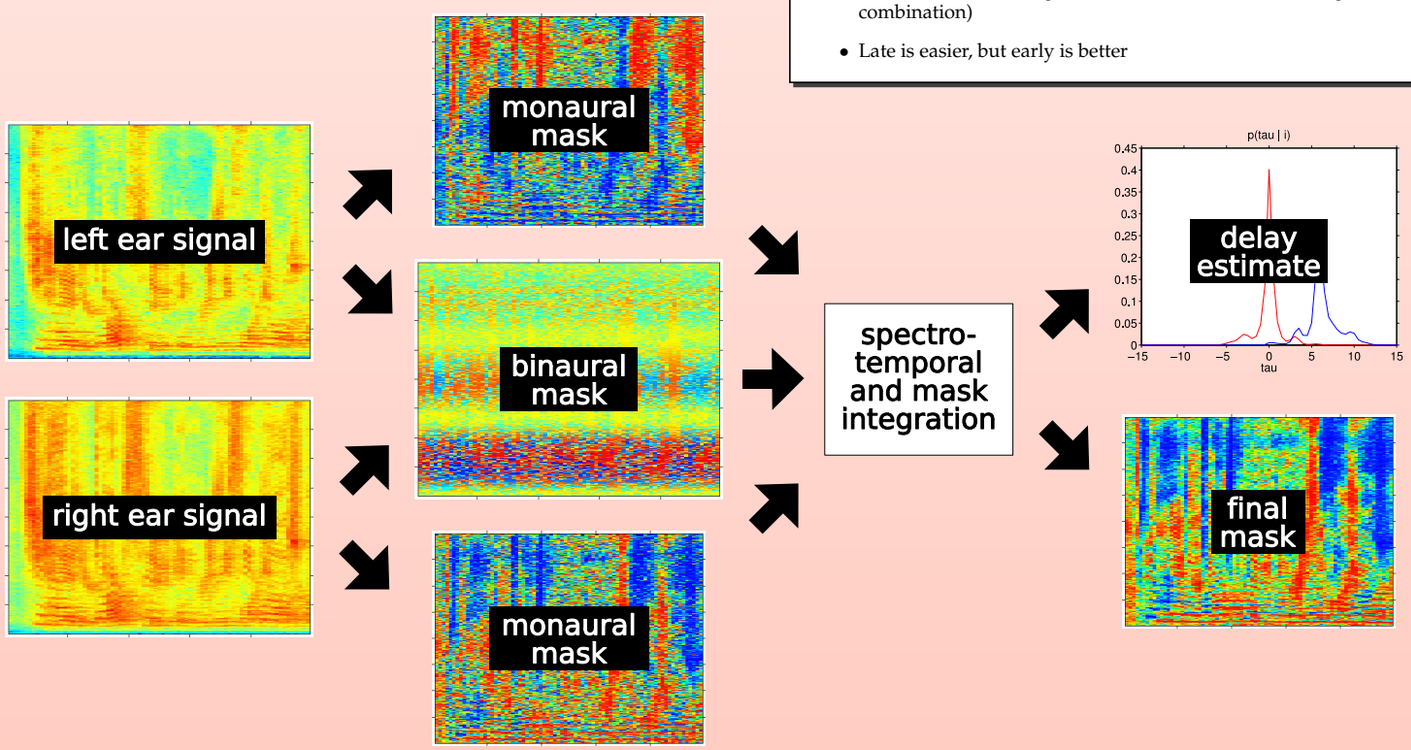
4 Information combination

4.1 Combining neighbors

- Nearby points in spectrograms usually come from the same source
- Influential neighbors vary with frequency
- Could use Markov random field to combine observations across neighbors

4.2 Combining different cues

- Could create many probabilistic masks from separate cues and then combine masks (late combination)
- Or could build one big model that uses all cues and neighbor relationships (early combination)
- Late is easier, but early is better



A sketch of the algorithm. Binaural inputs are analyzed together and separately to extract features. Information is then pooled across features and neighboring points to estimate masks and source locations.

5 Additional cues

Additional cues can add information and resolve ambiguities inherent to interaural phase difference

5.1 Interaural level difference (ILD)

- Magnitude of the interaural spectrogram
- Log-normally distributed
- Orthogonal to interaural phase difference

5.2 Monaural cues

- Separate target speaker from noise like Ellis and Weiss (2006)
- Separate two speakers from each other like Pearlmutter and Zador (2004); Kristjansson et al. (2006)
- Spectral models typically create masks
- Could apply to each channel and then combine masks with binaural cues and each other

5.3 Reliability cues

- Wilson and Darrell (2006) estimate the reliability of interaural parameters from monaural spectrograms
- Acts like the precedence effect in humans, trusting onsets most
- Could aid localization and separate direct path from reflections

References

Daniel P. W. Ellis and Ron Weiss. Model-based monaural source separation using a vector-quantized phase-vocoder representation. In *ICASSP-06*, pages V–957–960, May 2006.

Trausti Kristjansson, John Hershey, Peder Olsen, Steven Rennie, and Ramesh Gopinath. Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system. In *Proc. Int. Conf. Spoken Language Processing*, 2006.

Michael I. Mandel, Daniel P. W. Ellis, and Tony Jebara. An EM algorithm for localizing multiple sound sources in reverberant environments. In *Proc. Neural Information Processing Systems*, 2006.

Barak A. Pearlmutter and Anthony M. Zador. Monaural source separation using spectral cues. In *Proc. Fifth International Conference on Independent Component Analysis ICA-2004*, 2004.

Kevin Wilson and Trevor Darrell. Learning a precedence effect-like weighting function for the generalized cross-correlation framework. *IEEE Transactions on Speech and Audio Processing*, 2006.