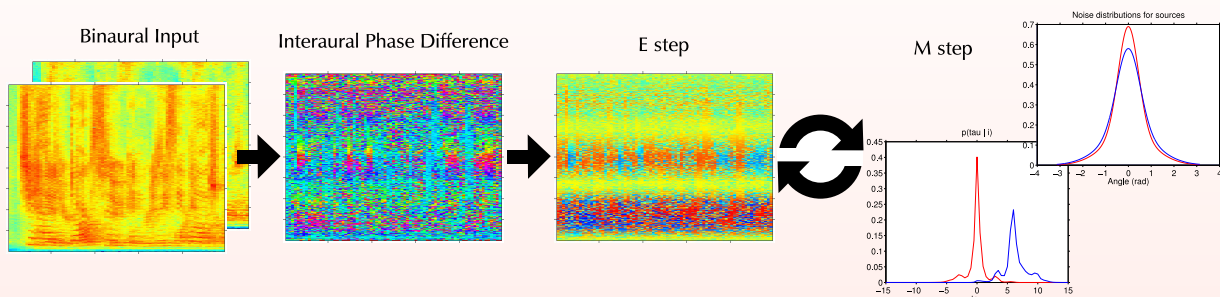


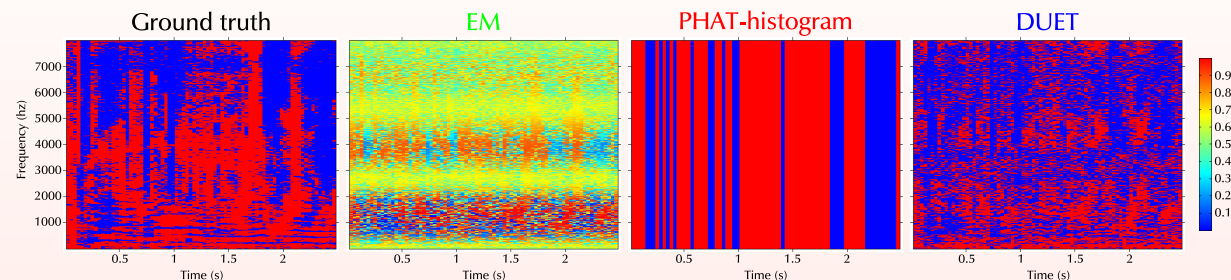
# An EM algorithm for localizing multiple sound sources in reverberant environments

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



Algorithm: A binaural recording is processed to extract the phase of the **interaural spectrogram**, from which each point's membership in sources and delays is calculated, the **parameters** are then updated to improve the **total log-likelihood** given those assignments.



Example masks: ground truth 0 dB mask, probabilistic mask from our EM algorithm, PHAT-histogram mask derived from Aarabi (2002), and DUET mask from Yilmaz and Rickard (2004). The example has two speakers in reverberation at 0° and 45°.

## 1 The problem

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

## 2 Prior approaches

- Brandstein and Silverman (1997): Microphone arrays for source localization
- Aarabi (2002): Generalized cross correlation with results pooled over time to create a pseudo probability distribution, "PHAT-histogram"
- Rennie (2005): EM algorithm for localizing entire spectrogram frames
- Yilmaz and Rickard (2004): DUET algorithm for localizing sources and assigning regions of spectrograms to sources

## 3 Our approach

- Parametric probability model of interaural phase difference
- 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

## 4 Algorithm

- Hearing model:

$$\ell(t) = a_\ell s(t - \tau_\ell) * n_\ell(t) \quad r(t) = a_r s(t - \tau_r) * n_r(t) \quad (1)$$

- **Interaural spectrogram**:

$$\frac{L(\omega, t)}{R(\omega, t)} = \alpha(\omega, t) e^{\phi(\omega, t)} = e^{a-j\omega\tau} N(\omega, t), \quad (2)$$

- Define:

$$\hat{\phi}(\omega, t; \tau) = \arg \left( \frac{L(\omega, t)}{R(\omega, t)} e^{j\omega\tau} \right) \quad (3)$$

- Model **parameters**:

$$\theta \equiv \{p(i, j, \tau), \sigma_{ij} \quad \forall i, j, \tau\} \quad (4)$$

- **Total log-likelihood**:

$$\log p(\phi(\omega, t) | \theta) = \sum_{\omega t} \log \sum_{ij\tau} \psi_{ij\tau} \mathcal{N}(\hat{\phi}(\omega, t; \tau) | 0, \sigma_{ij}^2) \quad (5)$$

## 5 Advantages of our approach

- Estimates probabilities over directions and sources for arbitrary regions of spectrogram, unlike Aarabi (2002); Rennie (2005)
- Probabilistic setting makes it easy to incorporate other cues and techniques, unlike Yilmaz and Rickard (2004)
- Can localize more sources than microphones, even in reverberation, unlike microphone arrays
- Makes no assumptions about source statistics, good for speech and music, unlike ICA

## 6 Experiments

- Compared to DUET, PHAT-histogram, and 2 controls
- Total of 300 mixtures used in experiments
- Evaluated algorithms on 4 conditions and 4 metrics
- **Conditions**:
  - Anechoic and reverberant simulations using binaural impulse responses from KEMAR
  - 2 and 3 simultaneous sources selected from 15 TIMIT utterances
- **Metrics**:
  - Localization mean-square error
  - Mutual information between estimated mask and ground truth mask
  - Signal to noise ratio of energy passed through mask
  - W-disjoint orthogonality: SNR times penalty for eliminating signal energy

## 7 Conclusions

- Our algorithm performs quite well at localization
- Outperforms others in anechoic conditions
- Performs as well as PHAT-hist in reverberation

### 7.1 Future work

- Use interaural level difference as well
- Combine with monaural source separators
- Exploit correlations between points in spectrogram

## References

- Parham Aarabi. Self-localizing dynamic microphone arrays. *IEEE transactions on systems, man, and cybernetics*, 32(4), November 2002.
- M. Brandstein and H. Silverman. A practical methodology for speech source localization with microphone arrays. *Computer, Speech, and Language*, 11(2):91-126, April 1997.
- Steven J. Rennie. Robust probabilistic TDOA estimation in reverberant environments. Technical Report PS1-TR-2005-011, University of Toronto, February 2005.
- Ozgun Yilmaz and Scott Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on signal processing*, 52(7):1830-1847, July 2004.



The results of our experiments for four different conditions (↓) and four different metrics (↔)