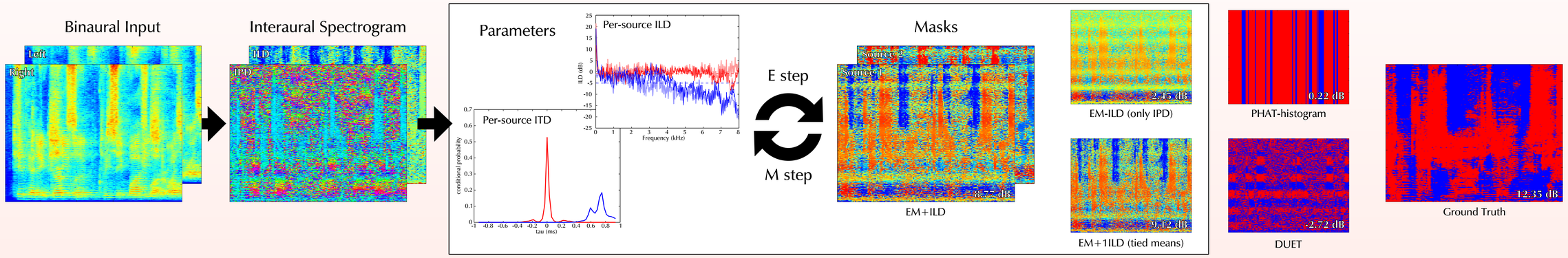
EM localization and separation using interaural level and phase cues

Michael I Mandel and Daniel P W Ellis

LabROSA · Dept of Electrical Engineering · Columbia University, New York

{mim,dpwe}@ee.columbia.edu



Algorithm: A binaural recording is transformed into **observations** of interaural phase and level differences, IPD and ILD, from which model **parameters** are estimated to improve the **total log-likelihood**, from which each point's membership in sources and delays is calculated.

Example analysis: Speakers are located at 0° and 90° in a reverberant room. Per-source ILD shows ground truth (thick lines), estimates from mixture (thin lines), and frequency independent estimates (dashed lines).

EM masks: These masks were created by our system using different ILD models. All masks are annotated with the SNR improvement, in dB, that they produce for this example.

Comparison masks: PHAT-histogram mask derived from Aarabi (2002), the DUET mask from Yilmaz and Rickard (2004), and the ground truth 0 dB mask.

1. Summary

- The goal of our system is to locate and separate an arbitrary number of sound sources from a stereo recording
- We use a probabilistic model of the interaural spectrogram of a single source that treats every time-frequency point independently (Mandel and Ellis, 2006)
- A probabilistic mixture of those models reduces the multi-source problem to the single source problem
- Our system estimates the maximum likelihood parameters of the mixture model with an EM algorithm
- The only assumption made is that each time-frequency point is dominated by one source

2. Advantages of our approach

- Creates true probabilistic masks
- Makes no assumptions about source statistics (e.g. Gaussianity, stationarity)
- Can separate more sources than observations
- Estimates parameters from the whole spectrogram
- Requires minimal training
- Works in reverberation

3. Algorithm outline

- Initialize source delays from PHAT-histogram, initialize source level differences to 0 dB
- Repeat 5–15 times or until convergence:
  - E step: Evaluate the probability of each spectrogram point under each source model, probabilistic masking
  - M step: Estimate the parameters of each source as the average of sufficient statistics from the interaural spectrogram weighted by that source's mask

4. Probabilistic model

- Model of binaural hearing of a single source with reverberant noise:

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

- Predicted interaural spectrogram with model **parameters** for frequency-dependent level and frequency-independent time differences, $a(\omega)$ and τ:

$$\frac{L(\omega,t)}{R(\omega,t)} = 10^{a(\omega)/20} e^{-j\omega\tau} N(\omega,t)$$

- **Observations** of the interaural level and phase differences, $\alpha(\omega,t)$ and $\phi(\omega,t)$:

$$\frac{L(\omega,t)}{R(\omega,t)} \equiv 10^{\alpha(\omega,t)/20} e^{j\phi(\omega,t)}$$

- Phase residual, the deviation of the observed phase difference from a delay of τ:

$$\hat{\phi}(\omega,t;\tau) = \arg\left(e^{\phi(\omega,t)} e^{-j\omega\tau}\right)$$

- **Total log-likelihood**, phase residual and interaural level difference modeled as Gaussian, marginalizing over hidden source assignment and delay, i and τ:

$$\log p(\phi(\omega,t), \alpha(\omega,t) \mid \theta) = \sum_{\omega,t} \log \sum_{i,\tau} \psi_{i\tau} \mathcal{N}(\hat{\phi}(\omega,t;\tau) \mid 0, \sigma_i^2) \times \mathcal{N}(\alpha(\omega,t) \mid \mu_i(\omega), \eta_i^2)$$

- Derive E and M steps of algorithm from total log-likelihood

5. Evaluation and comparison

- Evaluated against DUET, PHAT-histogram, and 2 controls
- Total of 300 mixtures used in experiments
- Evaluated algorithms in 4 conditions
  - Anechoic and reverberant simulations using binaural impulse responses from KEMAR
  - 2 and 3 simultaneous sources selected from 15 TIMIT utterances
- Compared RMSE of localization
- Compared SNR improvement of separation:

$$10\log_{10}\frac{||M_i S_i||^2}{||S_i - M_i \sum_j S_j||^2} - 10\log_{10}\frac{||S_i||^2}{||\sum_{j\neq i} S_j||^2}$$

6. Conclusions

- Inclusion of ILD improves separation in high frequencies, resolves ambiguities in IPD (Mandel et al., 2007)
- Per-source IPD differences at low frequencies allow identification of per-source ILD differences
- We would like to add prior models connecting ILD, ITD, and position
- We predict that people cannot disambiguate the position of low frequency, narrowband sources

References

P. Aarabi. Self-localizing dynamic microphone arrays. *IEEE Trans. systems, man, and cybernetics*, 32(4), November 2002.

M. Mandel and D. Ellis. A probability model for interaural phase difference. *SAPA*, 2006.

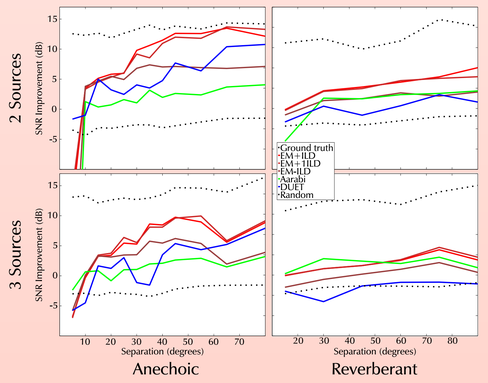
M. Mandel, D. Ellis, and T. Jebara. An EM algorithm for localizing multiple sound sources in reverberant environments. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *NIPS 19*. MIT Press, Cambridge, MA, 2007.

O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. Sig. Proc.*, 52(7):1830–1847, July 2004.

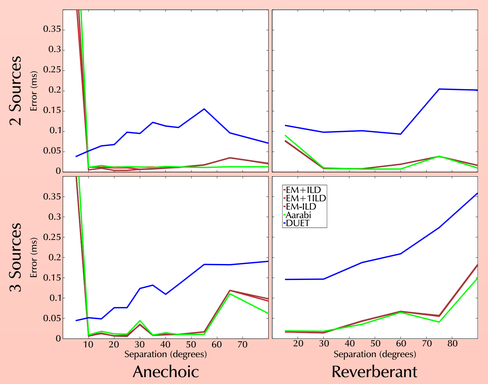
SNR improvement by condition: **A**nechoic or **R**everberant with 2 or 3 simultaneous speakers.

	2A	2R	3A	3R	Average
Gnd Truth	12.97	12.10	13.25	12.45	12.69
EM+ILD	**8.00**	**4.05**	3.74	2.08	**4.47**
EM+1ILD	7.46	3.86	**4.00**	1.94	4.31
EM−ILD	5.58	1.97	2.87	0.70	2.78
Aarabi	1.07	2.14	1.34	**2.46**	1.75
DUET	4.53	−0.39	1.83	−1.73	1.06
Random	−2.72	−2.01	−2.63	−2.04	−2.35

SNR improvement by distractor angle



Localization error by distractor angle



WASPAA 2007, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 21-24, 2007, New Paltz, NY