Source separation based on binaural cues and source model constraints

Ron J. Weiss, Michael I. Mandel and Daniel P. W. Ellis

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

{ronw,mim,dpwe}@ee.columbia.edu

1. Summary

- Task: separate an arbitrary number of sound sources from a stereo recording.
- Construct a probabilistic model of the mixed signal utilizing localization cues and prior model of source statistics.
- EM algorithm to learn model parameters for each source.
- Separate sources by applying probabilistic time-frequency masks to the mixture.
- Extension of the Model-based EM Source Separation and Localization (MESSL) algorithm (Mandel and Ellis, 2007).

2. Signal Model

- Observations are related to each source signal by the gain and delay that characterize the direct path and early reflections.

$$\ell(t) = \sum_i s_i(t - \tau_i^\ell) * h_i^\ell(t) \qquad r(t) = \sum_i s_i(t - \tau_i^r) * h_i^r(t)$$

- Model interaural spectrogram *and* binaural observations as independent mixtures of Gaussians.
- Assume each time-frequency cell is dominated by a single source.

1. Interaural Phase Difference (IPD):

$$\phi(\omega, t) = \angle \frac{L(\omega, t)}{R(\omega, t)} \approx \omega(\tau_i^\ell - \tau_i^r) \sim \sum_\tau \psi_{i\tau} \mathcal{N}(\omega\tau, \sigma_i)$$

2. Interaural Level Difference (ILD):

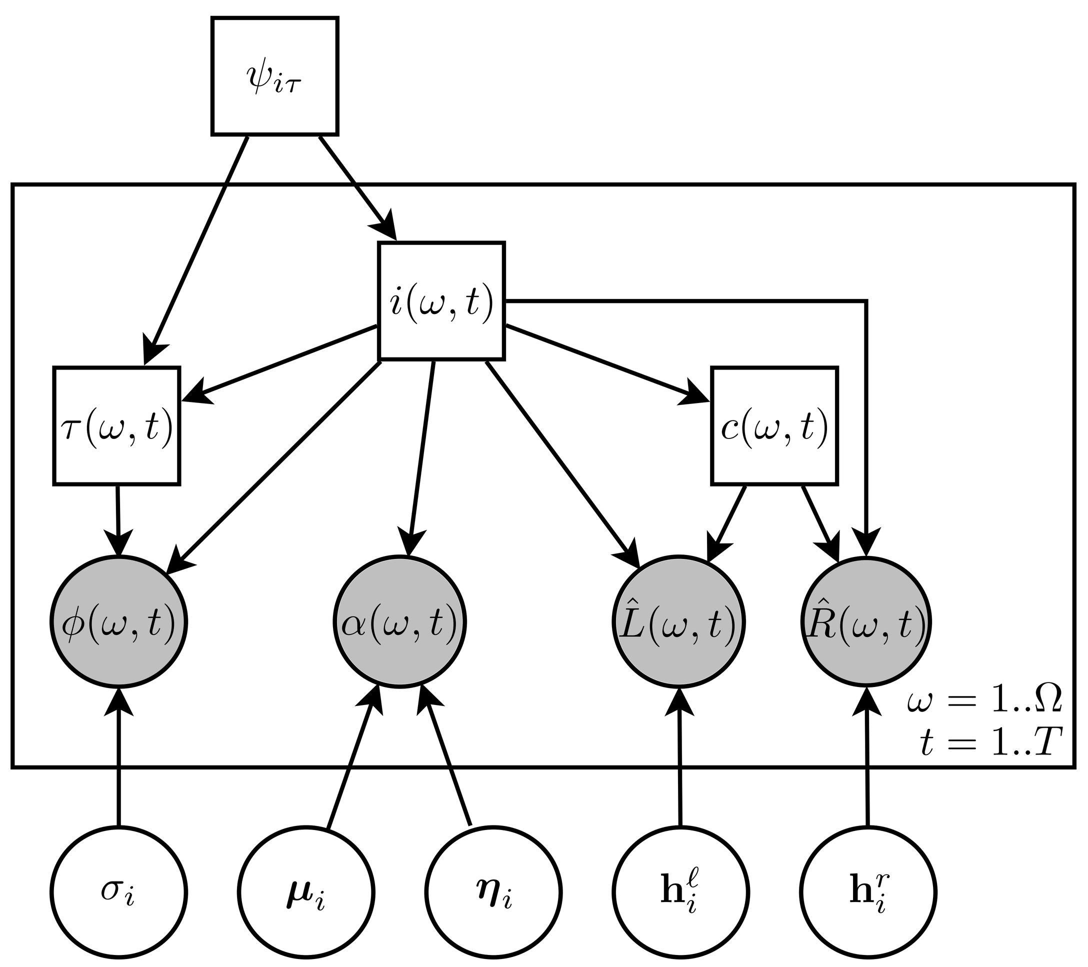
$$\alpha(\omega, t) = 20 \log_{10} \left| \frac{L(\omega, t)}{R(\omega, t)} \right| \approx \hat{H}_i^\ell(\omega, t) - \hat{H}_i^r(\omega, t) \sim \mathcal{N}(\mu_i, \eta_i)$$

3. Binaural observations:

$$\hat{L}(\omega, t) \approx \hat{S}_i(\omega, t) + \hat{H}_i^\ell(\omega, t) \sim \sum_c \pi_c \mathcal{N}(\mu_c + h_i^\ell, \Sigma_c)$$

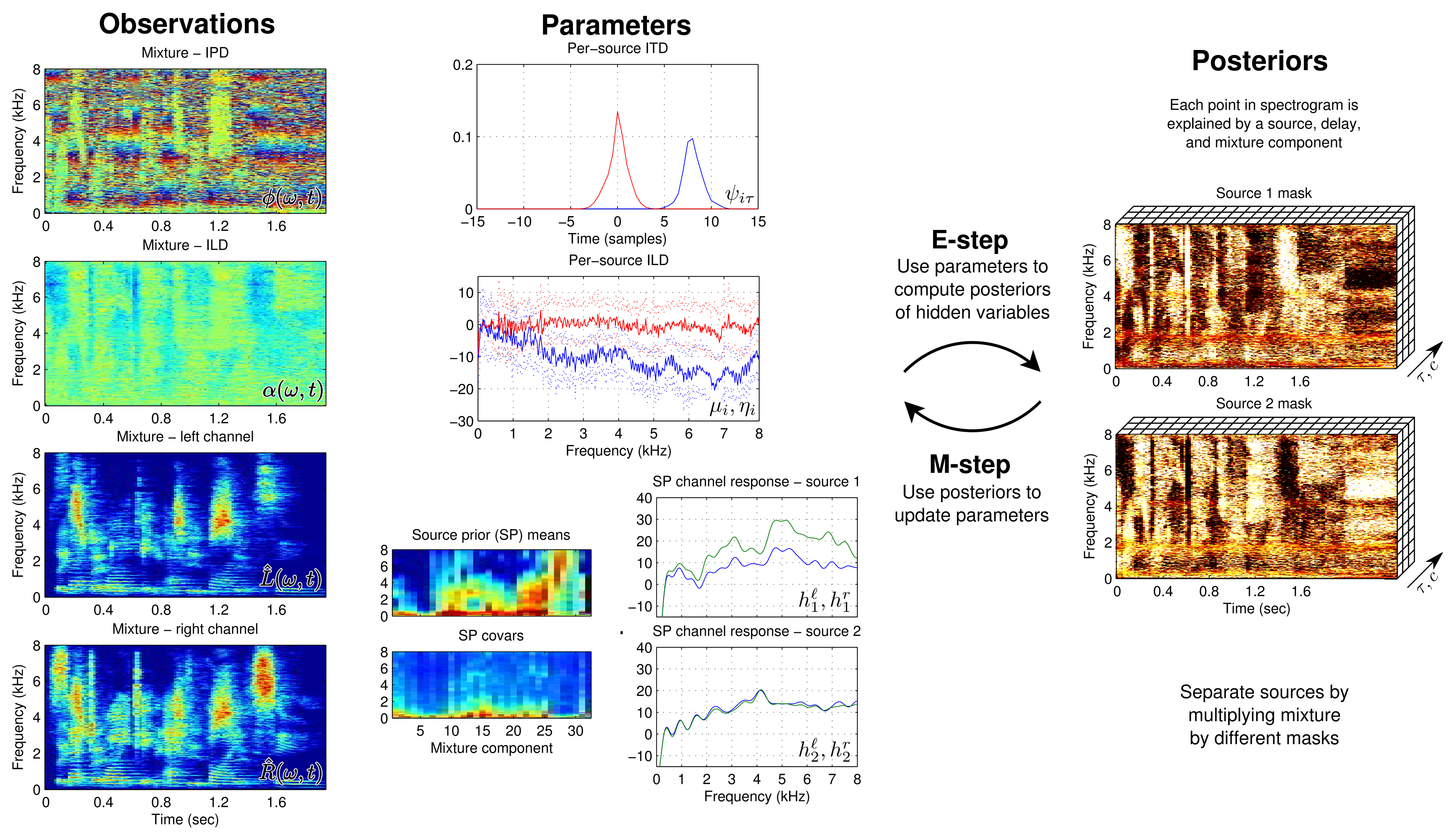
$$\hat{R}(\omega, t) \approx \hat{S}_i(\omega, t) + \hat{H}_i^r(\omega, t) \sim \sum_c \pi_c \mathcal{N}(\mu_c + h_i^r, \Sigma_c)$$

- Each point in spectrogram is explained by a given source, time delay, and source model component.



3. Separation algorithm

- Initialize source delays from PHAT-histogram (Aarabi, 2002), initialize all other parameters to 0.
- Repeat 5–15 times or until convergence:



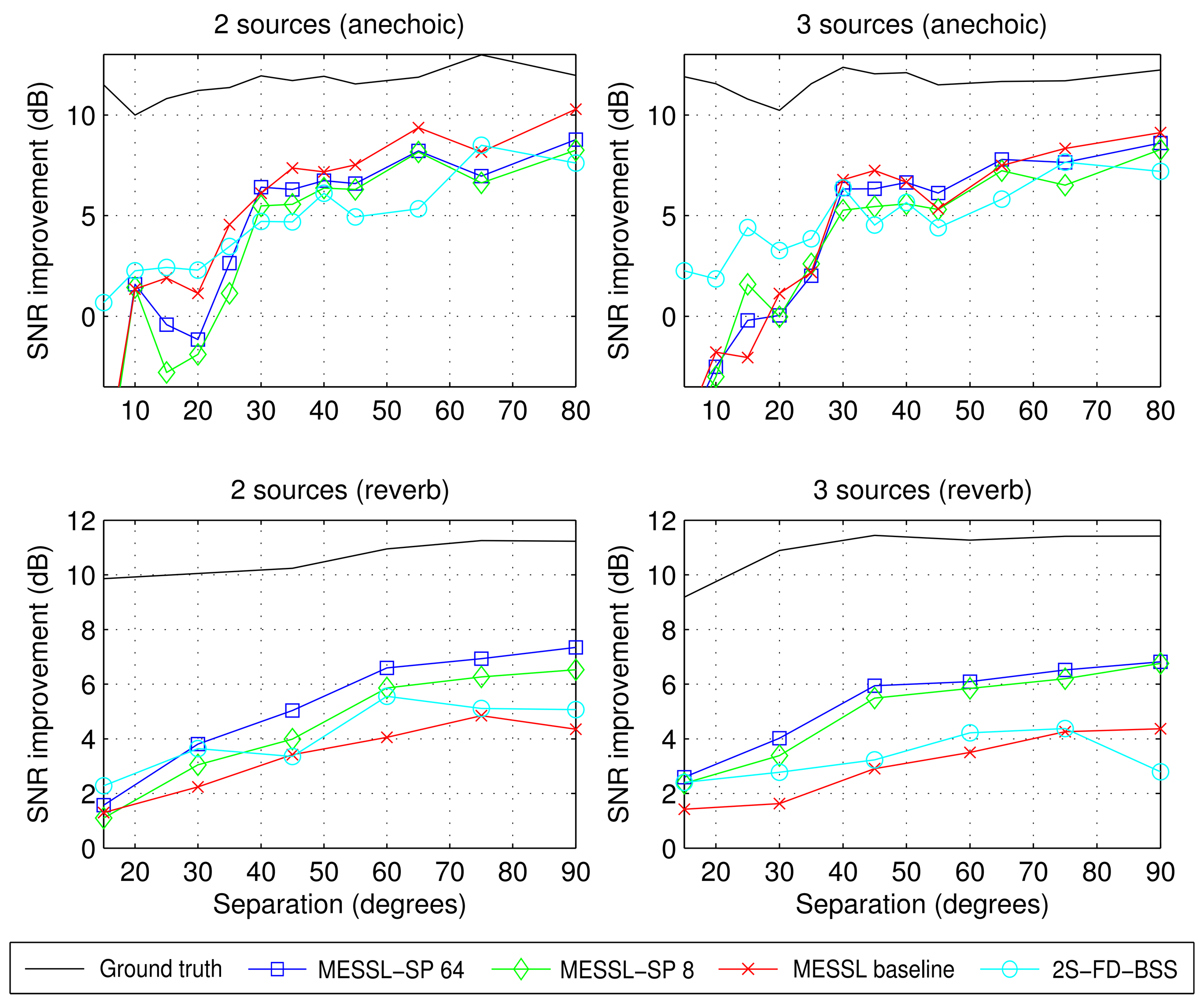
4. Evaluation

- Speech signals from GRID dataset (Cooke and Lee, 2006)
- Speaker independent GMM trained over all speakers
- Evaluated algorithms in 4 conditions
  - Anechoic (A) and reverberant (R) simulations using binaural impulse responses from KEMAR dummy head
  - 2 and 3 simultaneous sources selected from 15 GRID utterances
- Compared SNR improvement of separation:

$$20 \log_{10} \frac{||M_i S_i||}{||S_i - M_i \sum_j S_j||} - 20 \log_{10} \frac{||S_i||}{||\sum_{j \neq i} S_j||}$$

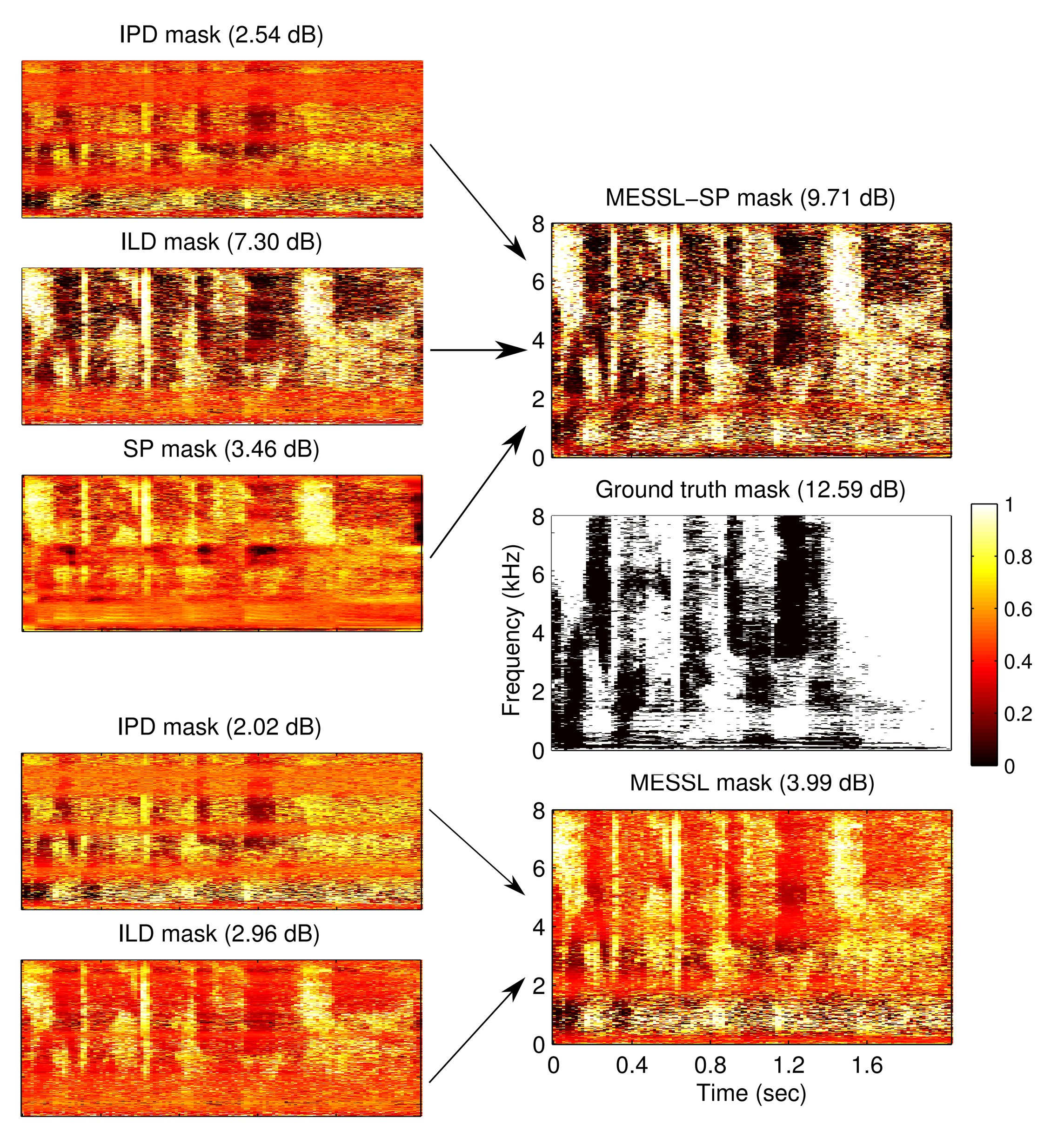
- Compare with ground truth mask, baseline MESSL, ICA-based method 2S-FD-BSS (Sawada et al., 2007)

System	2A	3A	2R	3R	Avg
Ground Truth	11.57	11.62	10.60	10.93	11.18
MESSL-SP 64	3.65	3.66	**5.21**	**5.33**	**4.46**
MESSL-SP 32	3.47	3.60	5.12	5.25	4.36
MESSL-SP 16	3.28	3.55	4.94	5.21	4.25
MESSL-SP 8	2.97	3.31	4.47	5.00	3.94
MESSL baseline	**4.74**	3.83	3.36	3.01	3.73
2S-FD-BSS	4.42	**4.82**	4.17	3.30	4.18



5. Discussion

- MESSL outperforms MESSL-SP in anechoic conditions.
  - Because there is no convolutive noise, interaural model alone is often a very good fit to observations.
  - Loose fit of source model in low frequencies causes errors.
- MESSL-SP outperforms MESSL in reverberant conditions.
  - Source model is a good fit to the direct path.
  - Helps resolve ambiguities in interaural parameters resulting from reverberant noise.



- Each observation contributes qualitatively different information to final masks.
  - IPD is very informative in low frequencies, but is uninformative in some high frequency subbands.
  - ILD primarily adds information about high frequencies.
  - Source model introduces correlations across frequency and emphasizes reliable time-frequency regions.

6. References

P. Aarabi. Self-localizing dynamic microphone arrays. *IEEE Transactions on Systems, Man, and Cybernetics*, 32(4), November 2002.

M. Cooke and T. W. Lee. The speech separation challenge, 2006. URL http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm.

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, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.

M. I. Mandel and D. P. W. Ellis. EM localization and separation using interaural level and phase cues. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2007.

H. Sawada, S. Araki, and S. Makino. A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2007.

Interspeech 2008, 22-26 September 2008, Brisbane, Australia