Latent Gaussian Activity Propagation

Using Smoothness and Structure to Separate and Localize Sounds in Large Noisy Environments

Daniel D. Johnson, Daniel Gorelik, Ross Mawhorter, Kyle Suver, Weiqing Gu,
Department of Mathematics, Harvey Mudd College

Steven Xing, Cody Gabriel, Peter Sankhagowit
Intel Corporation

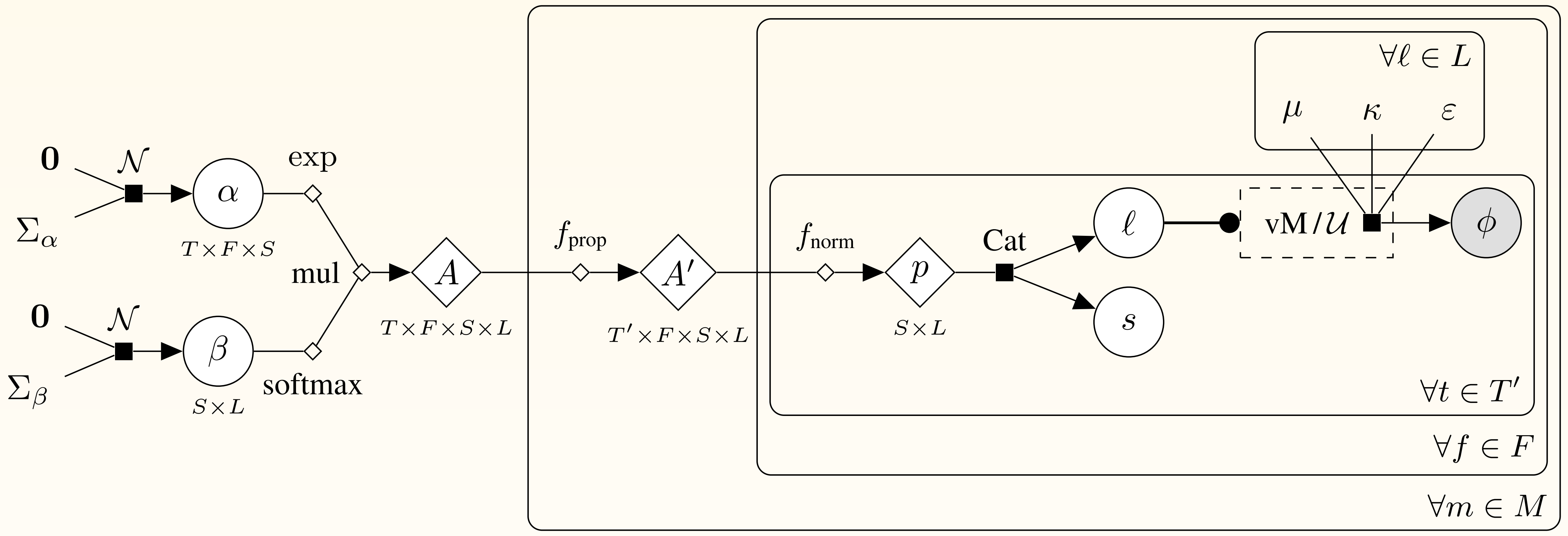
Motivation

Realistically reproducing the audio of an event in virtual space requires that sounds be separated and localized accurately. Such applications, however, may require microphones to be placed far away from each other and include significant background noise.

We propose a method for simultaneously separating and localizing sounds by performing Bayesian maximum a posteriori inference in an approximate probabilistic model of sound propagation and inter-microphone phase differences.

Model

Let T be a set of time bins, F a set of frequency bins, L a set of candidate locations where sounds could originate, and M a set of microphone pairs placed at known locations in 2D space. We hypothesize a set S of sources, and model each observed time-frequency bin as being assigned a latent dominating source and location that determines its phase difference.



Forward model:

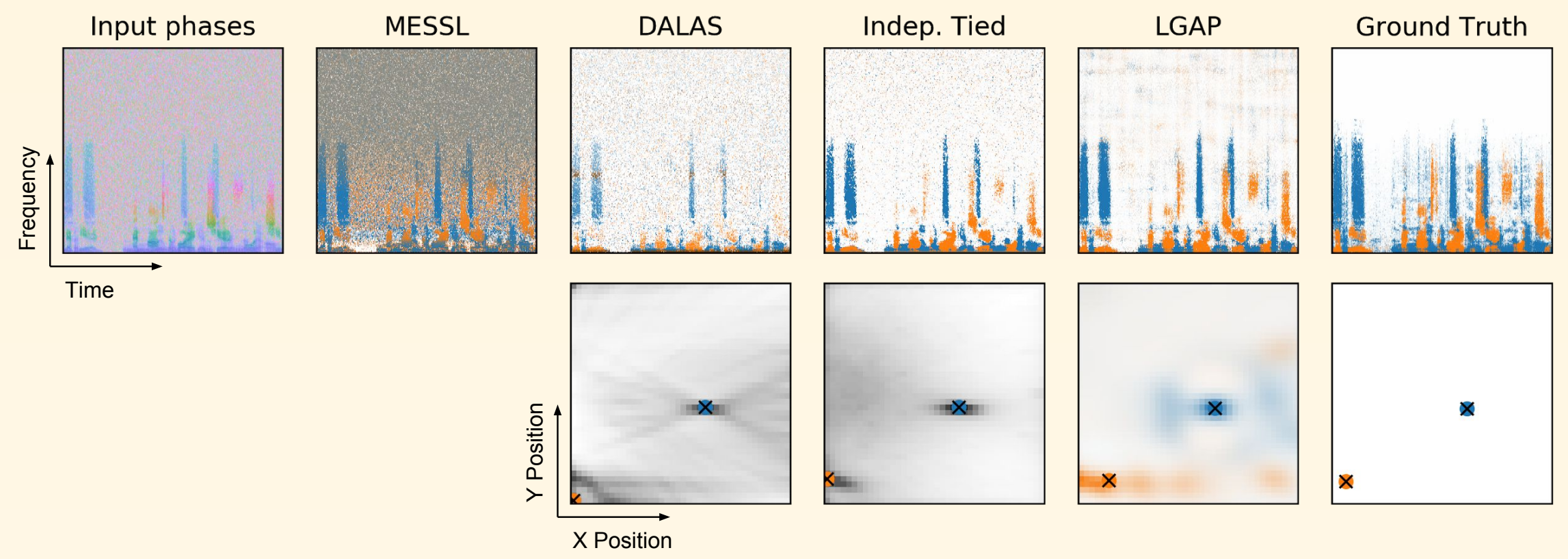
1. Random-valued multidimensional arrays α (representing source activity over time and frequency) and β (representing source location) are sampled from smooth factorized Gaussian process priors.
2. α and β are combined into a nonnegative source-location activity array A.
3. A is propagated through time for each microphone pair by adding time delays based on the distance from each source, yielding A'.
4. Dominating sources s and locations ℓ are sampled proportional to (normalized) values in A' for each microphone pair and time-frequency bin.
5. Phase differences ϕ are drawn from a mixture between a von Mises distribution (centered on the ideal delay for location ℓ's angle of incidence) and a uniform distribution (to approximate noise).

Inference: Given phase differences ϕ, we approximate the MAP values for α and β using gradient ascent on the likelihood $\log P(\alpha,\beta|\phi) = \log P(\phi|\alpha,\beta) + \log P(\alpha,\beta) - \log(\phi)$. The value of β then gives location estimates, and the distribution $P(s|\alpha,\beta,\phi)$ gives approximate time-frequency masks for each source.

Prior Work

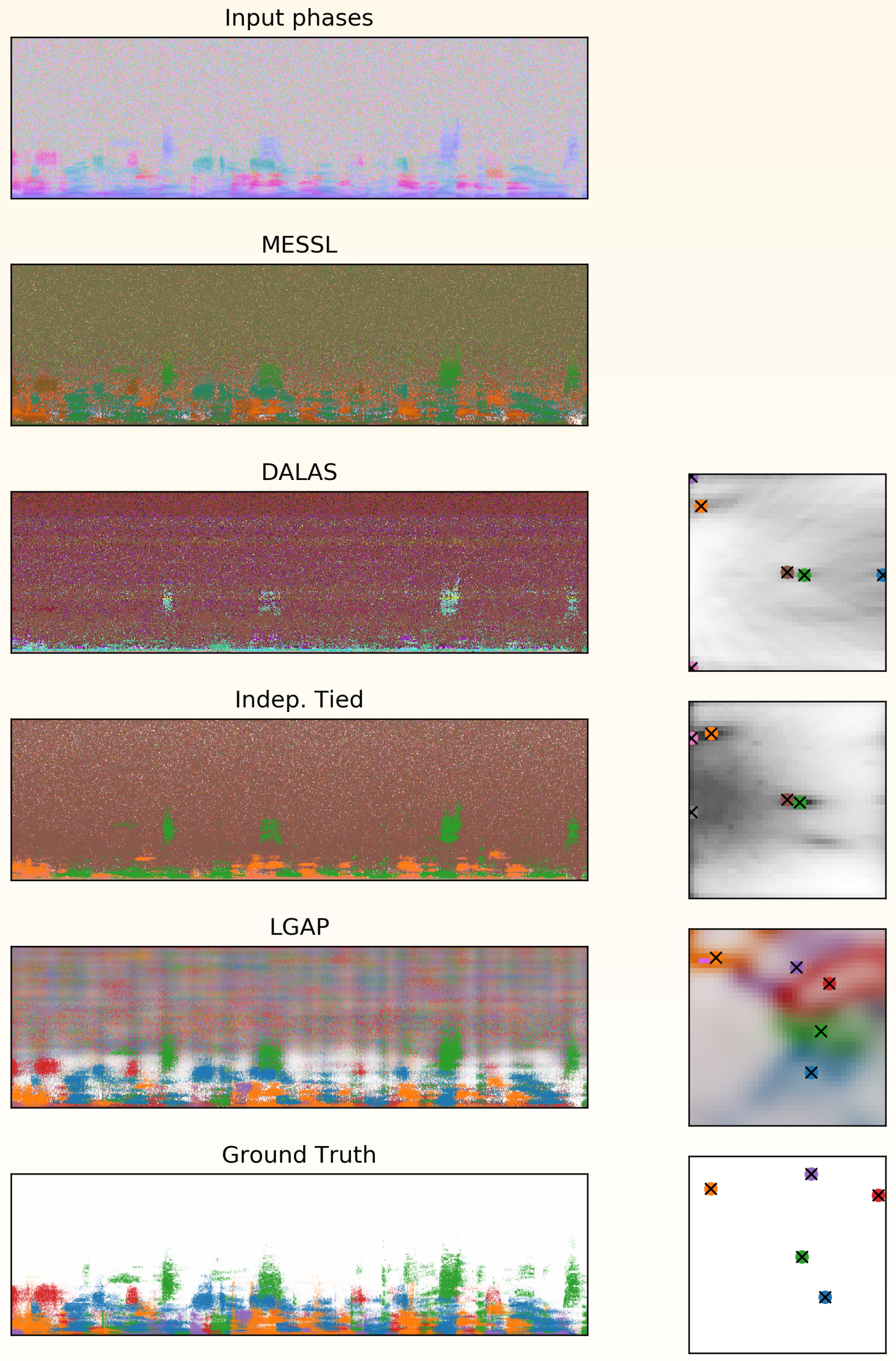
The DUET method [Rickard, 2007] uses *phase differences* in the STFTs of adjacent microphones to partially separate sounds, assuming only one source dominates each time-frequency bin. MESSL [Mandel et al., 2009] extends DUET by using these phase differences to infer a consistent angle of incidence for each source, obtaining a more consistent separation. DALAS [Dorfan et al., 2015] combines angle estimates from multiple distant microphones to obtain location estimates, but does not use temporal or frequential structure of the sounds to help separate them.

Two Sources



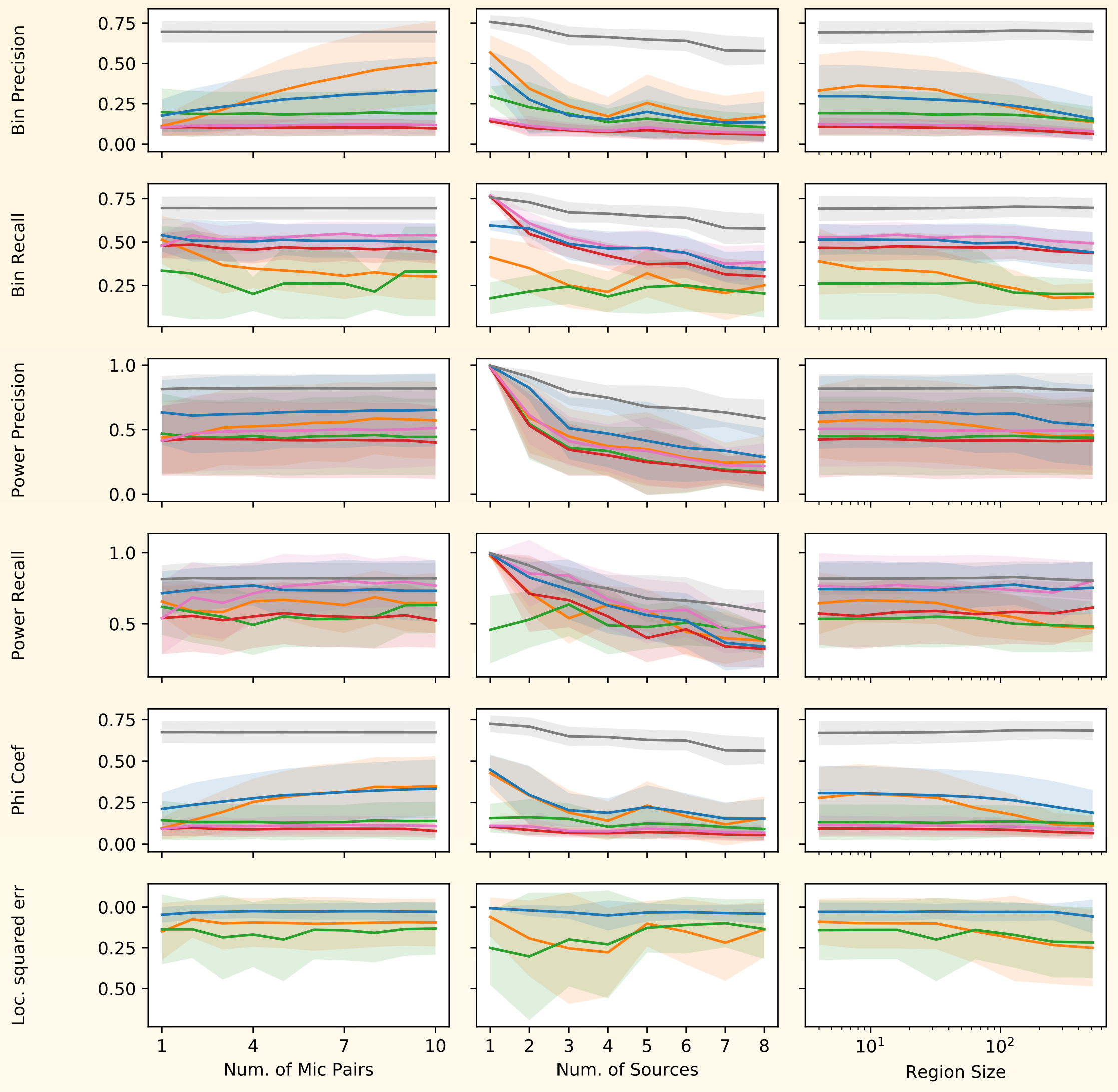
Two sound sources, five microphone pairs along left side of a 32 m by 32 m square. Top row: Spectrograms for center microphone pair colored by source, over four seconds (horizontal axis) up to 22050 Hz (vertical axis). Bottom row: location estimates, where shading represents confidence and crosses represent predicted point locations.

Five Sources



Five sound sources, five microphone pairs, over a 32 m by 32 m square. Left: spectrograms colored by source, right: location estimates. Only LGAP is able to separate and localize all five sources.

Metrics



Metrics for each method: — LGAP, — Indep. Tied, — DALAS, — MESSL (All mics), — MESSL (Best mic), — Ideal Mask (from ground truth). Bin precision/recall measure separation performance across bins, power precision/recall measure separation weighted by source power, and location squared error measures accuracy of location estimates.

Conclusions

LGAP maintains both high precision and high recall and obtains reliably accurate location estimates. Apart from smoothness, the method does not depend on source statistics, and can combine information from distant microphones. It thus has the potential to be used for a variety of applications, especially for sounds that are distributed across large real-world environments.

References

Yuval Dorfan, Dani Cherkassky, and Sharon Gannot. Speaker localization and separation using incremental distributed expectation-maximization. In *Signal Processing Conference (EUSIPCO), 2015 23rd European*, pages 1256–1260. IEEE, 2015.

Michael I Mandel, Ron J Weiss, and Daniel P W Ellis. Model-based expectation-maximization source separation and localization. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(8), 2009.

Scott Rickard. *The DUET Blind Source Separation Algorithm*, pages 217–241. Springer Netherlands, Dordrecht, 2007. ISBN 978-1-4020-6479-1. doi: 10.1007/978-1-4020-6479-1_8. URL https://doi.org/10.1007/978-1-4020-6479-1_8.