Source separation based on binaural cues and source model constraints

Ron J. Weiss, Michael I. Mandel, Daniel P. W. Ellis [*]

LabROSA, Department of Electrical Engineering
Columbia University, New York, NY, 10027
{ronw,mim,dpwe}@ee.columbia.edu

Abstract

We describe a system for separating multiple sources from a two-channel recording based on interaural cues and known characteristics of the source signals. We combine a probabilistic model of the observed interaural level and phase differences with a prior model of the source statistics and derive an EM algorithm for finding the maximum likelihood parameters of the joint model. The system is able to separate more sound sources than there are observed channels. In simulated reverberant mixtures of three speakers the proposed algorithm gives a signal-to-noise ratio improvement of 2.1 dB over a baseline algorithm using only interaural cues.

Index Terms: source separation, binaural, source models, EM

1. Introduction

The use of signal models is a common approach to source separation when only a single channel observation is available. However, for good performance, such algorithms require relatively large, speaker-dependent models [1]. Separation becomes considerably easier when multiple channels are observed, as shown by the way the binaural system of human listeners allows them to focus on a particular sound source in an environment containing distracting sources. In fact, by leveraging the information present in binaural signals it is possible to separate sources without prior knowledge of their content [2, 3]. However, it is to be expected that such prior knowledge can further improve separation performance. In this paper we describe a system for source separation that combines inference of localization parameters with model-based separation methods and show that the additional constraints derived from the source model do help to improve separation performance.

Independent component analysis (ICA) is another well known approach to source separation given multi-channel observations. Conventional ICA does not work well when the number of sources is greater than the number of channels, but in [4], Sawada et al. describe a two-stage frequency-dependent blind source separation approach (2S-FD-BSS) that combines ideas from model-based separation with ICA to be able to separate underdetermined mixtures. In the first stage, blind source separation is performed separately on each frequency band of a spectrogram using a probabilistic model of the mixing matrices. The separated signals from each band are then associated with the corresponding signals from other bands using k-means clustering on the posterior probabilities of each source, then further refined by aligning bands with nearby and harmonically-related frequencies.

Instead of using the independence-maximizing technique of ICA, our approach is based on a model of the interaural time and level differences derived from the binaural observation. This is similar to the DUET algorithm [2] for separating undetermined mixtures by clustering localization cues across time and frequency. Given prior knowledge of the source locations, Wilson describes in [5] a complementary method based on source model inference. In this paper we propose an algorithm to jointly estimate both the localization and source model parameters. We extend the Model-based EM Source Separation and Localization (MESSL) algorithm of [3] to incorporate additional constraints obtained from a prior source model. We call this extension MESSL-SP (Source Prior). We show that it is possible to obtain significant improvement in separation performance of speech signals in reverberation over a baseline system employing only interaural cues. The improvement is significant even when the source models used are quite weak. The information extracted from the interaural cues and source model serve to reinforce each other.

2. Signal Model

We model the mixture of I spatially distinct source signals $\{s_i(t)\}_{i=1..I}$ based on the binaural observations $\ell(t)$ and $r(t)$ corresponding to the signals arriving at the left and right ears respectively. In anechoic environments, the observations will be related to a given source signal primarily by the gain and delay that characterize the direct path from the source location. However, in reverberant environments this assumption is confused by the addition of convolutive noise arising from the room impulse response. In general the observations can be modeled as follows:

$$\ell(t) = \sum_i s_i(t - \tau_i^\ell) * h_i^\ell(t) \tag{1}$$

$$r(t) = \sum_i s_i(t - \tau_i^r) * h_i^r(t) \tag{2}$$

where τ_i is the delay characteristic of the direct path for source i and $h_i^{\ell,r}(t)$ are the corresponding "channel" responses for the left and right channels respectively that approximate the room impulse response and additional filtering due to the head related transfer function, excluding the primary delay.

2.1. Interaural model

We model the observations in the short-time spectral domain using the interaural spectrogram $X_{IS}(\omega, t)$.

$$X_{IS}(\omega, t) \triangleq \frac{L(\omega, t)}{R(\omega, t)} = 10^{\alpha(\omega,t)/20} e^{j\phi(\omega,t)} \tag{3}$$

[*] This work was supported by the National Science Foundation (NSF) under Grants No. IIS-0238301 and IIS-0535168, EU Project AMIDA, and the Fu Foundation School of Engineering and Applied Science via a Presidential Fellowship. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

where $L(\omega, t)$ and $R(\omega, t)$ are the short-time Fourier transforms of $\ell(t)$ and $r(t)$, respectively. For a given time-frequency cell, the interaural level difference (ILD) between the two channels is $\alpha(\omega, t)$, and $\phi(\omega, t)$ is the corresponding interaural phase difference (IPD).

A key assumption in the MESSL signal model is that each time-frequency point is dominated by a single source. This implies the following approximations for the observed ILD and ITD:

$$\alpha(\omega, t) \approx 20 \log_{10} \frac{|H_i^\ell(\omega, t)|}{|H_i^r(\omega, t)|} \tag{4}$$

$$\phi(\omega, t) \approx \omega(\tau_i^\ell - \tau_i^r) \tag{5}$$

These variables have the advantage of being independent of the source signal, so the resulting model does not depend on the distribution of $s_i(t)$.

We model the ILD for source i as a Gaussian distribution whose mean and variance will be learned from the mixed signal:

$$P(\alpha(\omega, t) \mid i, \theta) = \mathcal{N}(\alpha(\omega, t);\, \mu_i(\omega), \eta_i^2(\omega)) \tag{6}$$

where θ stands for the otherwise unspecified model parameters. Because $\phi(\omega, t)$ is only observed modulo 2π, it is difficult to learn the IPD for a given source directly from the mixed signal. To simplify this process, we define a grid of time differences τ to search over, and the IPD distribution for a given source has the form of a Gaussian mixture model (GMM) with one mixture component for each time difference on the grid:

$$P(\phi(\omega, t) \mid i, \tau, \theta) = \mathcal{N}(\phi(\omega, t);\, \omega\tau, \sigma_i^2) \tag{7}$$

$$P(\phi(\omega, t), i \mid \theta) = \sum_\tau \psi_{i\tau} P(\phi(\omega, t) \mid i, \tau, \theta) \tag{8}$$

where $\psi_{i\tau}$ are the mixing weights for source i and delay τ.

2.2. Source model

We extend the baseline MESSL model described above to incorporate prior knowledge of the source statistics. This makes it possible to model the binaural observations directly:

$$\hat{L}(\omega, t) \approx \hat{S}_i(\omega, t) + \hat{H}_i^\ell(\omega, t) \tag{9}$$

$$\hat{R}(\omega, t) \approx \hat{S}_i(\omega, t) + \hat{H}_i^r(\omega, t) \tag{10}$$

where $\hat{X} \triangleq 20 \log_{10} |X|$. Each source $\hat{S}_i$ is modeled using a GMM:

$$P(\hat{\mathbf{S}}_i(:, t)) = \sum_c \pi_{ic} \mathcal{N}(\hat{\mathbf{S}}_i(:, t);\, \boldsymbol{\nu}_{ic}, \boldsymbol{\varsigma}_{ic}^2) \tag{11}$$

We assume that $\hat{H}_i^{\ell,r}$ will be constant across the entire mixture and that they will be relatively smooth across frequency. Therefore, we parametrize the channel response in the DCT domain, giving $\hat{H}_i^\ell(\omega, t) = B\mathbf{h}_i^\ell$ where B is a matrix of DCT basis vectors. This allows $\hat{H}_i^{\ell,r}$ to be modeled using many fewer DCT coefficients than the number of frequency bands Ω.

Combining this channel model with the source priors above give the following likelihoods for the left and right channel spectrograms:

$$P(\hat{L}|i, c, \theta) = \mathcal{N}(\hat{L}(\omega, t);\, \nu_{ic}(\omega) + B(\omega, :)h_i^\ell, \varsigma_{ic}^2(\omega)) \tag{12}$$

$$P(\hat{R}|i, c, \theta) = \mathcal{N}(\hat{R}(\omega, t);\, \nu_{ic}(\omega) + B(\omega, :)h_i^r, \varsigma_{ic}^2(\omega)) \tag{13}$$

where $B(\omega, :)$ is the row of B corresponding to frequency ω.

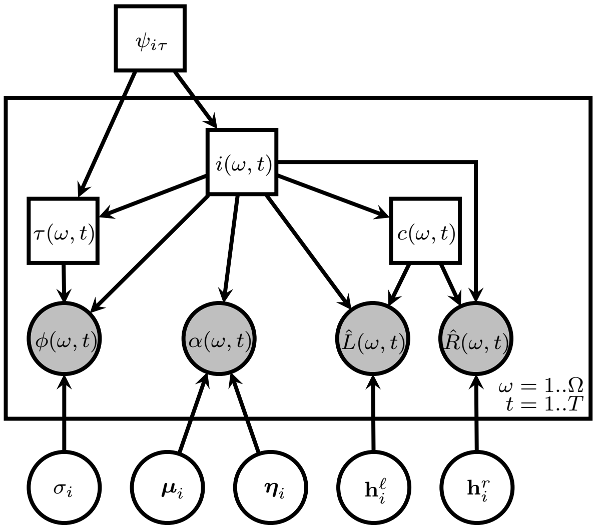

Figure 1: MESSL-SP graphical model. Each time-frequency point is explained by a source i, a delay τ, and a component c.

As the observed signal is likely to be small compared to the amount of data needed to reliably train a prior signal model, we use $\pi_{ic}, \boldsymbol{\nu}_{ic}$ and $\boldsymbol{\varsigma}_{ic}$ trained on clean data, and learn from the mixture only the source-dependent channel and location parameters.

Combining the model of the interaural signals with the source model gives the total likelihood of the model:

$$\begin{aligned} &P(\phi(\omega,t), \alpha(\omega, t), \hat{L}(\omega, t), \hat{R}(\omega, t), i, \tau, c \mid \theta) \\ &\quad = P(i, \tau)\, P(\phi(\omega, t) \mid i, \tau, \theta)\, P(\alpha(\omega, t) \mid i, \theta) \\ &\qquad P(c \mid i)\, P(\hat{L}(\omega, t) \mid i, c, \theta)\, P(\hat{R}(\omega, t) \mid i, c, \theta) \end{aligned} \tag{14}$$

This equation explains each time-frequency point of the mixed signal as coming from a single source at a given delay and a particular component in the source prior. The graphical model corresponding to this factorization is shown in figure 1. For conciseness we will drop the (ω, t) where convenient in the remainder of the paper.

3. Source Separation

If we knew which time-frequency points were responsible for each source it would be trivial to separate them. But inference of these posteriors requires knowledge of the unknown model parameters $\theta = \{\psi_{i\tau}, \sigma_i, \boldsymbol{\mu}_i, \boldsymbol{\eta}_i^2, \mathbf{h}_i^\ell, \mathbf{h}_i^r\}$. To solve this problem we derive an expectation-maximization algorithm to iteratively learn the parameters and posteriors.

The E-step consists of evaluating the posterior responsibilities for each time-frequency point given the estimated parameters for iteration j, θ_j. We introduce a hidden variable representing the posterior of i, τ, c:

$$z_{i\tau c}(\omega, t) = \frac{P(\phi, \alpha, L, R, i, \tau, c, |\theta_j)}{\sum_{i\tau c} P(\phi, \alpha, L, R, i, \tau, c, |\theta_j)} \tag{15}$$

The M-step consists of maximizing the expectation of the total log-likelihood given the current parameters θ_j:

$$\mathcal{Q}(\theta|\theta_j) = \sum_{\omega t} z_{i\tau c}(\omega, t) \log P(\phi, \alpha, L, R, i, \tau, c|\theta) \tag{16}$$

As shown in (14), the log likelihood is factored into conditionally independent terms which can be maximized independently.

The updates for the ITD and ILD parameters are derived in [6] and [3] respectively, but we reproduce them here for completeness. First, we define the operator

$$\langle x \rangle_{t,\tau} \triangleq \frac{\sum_{t,\tau} z_{i\tau c}(\omega,t) x}{\sum_{t,\tau} z_{i\tau c}(\omega,t)} \tag{17}$$

as the mean over the specified variables, t and τ in this case, weighted by $z_{i\tau c}(\omega,t)$. The updates for the interaural parameters can then be written as follows:

$$\sigma_i^2 = \left\langle \phi^2(\omega,t) \right\rangle_{\omega,t,\tau,c} \tag{18}$$

$$\mu_i(\omega) = \left\langle \alpha(\omega,t) \right\rangle_{t,\tau,c} \tag{19}$$

$$\eta_i^2(\omega) = \left\langle (\alpha(\omega,t) - \mu_i(\omega))^2 \right\rangle_{t,\tau,c} \tag{20}$$

$$\psi_{i\tau} = \frac{1}{\Omega T} \sum_{\omega t c} z_{i\tau c}(\omega,t) \tag{21}$$

Unlike the interaural parameters, the source prior parameters are tied across frequency to ensure that each time frame is explained by a single component in the source prior. The updates are as follows:

$$\mathbf{h}_i^\ell = \Big(\sum_{tc} B^T \Sigma_{ict}^{-1} B\Big)^{-1} \sum_{tc} B^T \Sigma_{ict}^{-1} (\hat{\mathbf{L}}(:,t) - \boldsymbol{\nu}_{ic}) \tag{22}$$

$$\mathbf{h}_i^r = \Big(\sum_{tc} B^T \Sigma_{ict}^{-1} B\Big)^{-1} \sum_{tc} B^T \Sigma_{ict}^{-1} (\hat{\mathbf{R}}(:,t) - \boldsymbol{\nu}_{ic}) \tag{23}$$

where Σ_{ict} encodes the posterior probability of component c from source i dominating the mixture at frame t:

$$\Sigma_{ict} \triangleq \text{diag}\left(\frac{\varsigma_{ic}^2}{\sum_\tau \mathbf{z}_{i\tau c}(:,t)}\right) \tag{24}$$

After the EM algorithm converges, we separate sources by deriving a time-frequency mask from the posterior probability of the hidden variables for each source:

$$m_i(\omega,t) = \sum_{\tau c} z_{i\tau c}(\omega,t) \tag{25}$$

Estimates of clean source i can then be obtained by multiplying the short-time Fourier transform of each channel of the mixed signal by the mask for the corresponding source. This assumes that the mask is identical for both channels. Figure 2 shows an example mask derived from the proposed algorithm.

4. Experiments

To evaluate the performance of the proposed algorithm, we assembled a data set consisting of mixtures of two and three speech signals in simulated anechoic and reverberant conditions. The reverberant mixtures were formed by convolving anechoic speech samples with binaural impulse responses. We used 15 utterances chosen at random from the GRID corpus used in the 2006 Speech Separation challenge [7].

The anechoic binaural impulse responses came from [8], a large effort to record head-related transfer functions for many different individuals. We use the measurements for a KEMAR dummy head with small ears, taken at 25 different azimuths at 0° elevation. The reverberant binaural impulse responses were recorded by Shinn-Cunningham et al. in a real classroom [9]. These measurements were also made with a KEMAR dummy head, although a different unit was used. The measurements we used were taken in the center of the classroom, with the source 1 m from the head at 7 different azimuths, each repeated 3 times.

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

Table 1: Average SNR improvement (in dB) across all distractor angles for each algorithm. The test cases are described by the number of simultaneous sources (2 or 3) and whether the impulse responses were anechoic or reverberant (A or R).

In the synthesized mixtures, the target speaker was located directly in front of the listener, with distractor speakers located off to the sides. The angle between the target and distractors was systematically varied and the results combined for each direction. Each setup was tested with 5 different randomly chosen sets of speakers and with one and two distractors, for a total of 300 different mixtures. We measure the performance of separation with signal-to-noise ratio improvement, defined for source i as follows:

$$10 \log_{10} \frac{||m_i S_i||^2}{||S_i - m_i \sum_j S_j||^2} - 10 \log_{10} \frac{||S_i||^2}{||\sum_{j \neq i} S_j||^2} \tag{26}$$

where m_i and S_i are the estimated mask and spectrogram, respectively, for source i, and $||X||^2$ indicates summing the squared magnitude of X over ω and t. This measure penalizes both noise that is passed through the mask and signal that is rejected by the mask.

We compare the proposed algorithm to the baseline MESSL algorithm from [3] that does not utilize source constraints, the ICA-based separation system from [4] (2S-FD-BSS), and the performance using ground truth binary masks derived from clean source signals. The proposed algorithm (MESSL-SP) was evaluated using prior source models with 8, 16, 32, and 64 GMM components trained over the training data for all 34 speakers. The same speaker-independent prior was used for each source in a given mixture. As a result, the likelihoods in (12) and (13) are initially uninformative because they evaluate to the same likelihood for each source. However, as the binaural cues begin to disambiguate the sources, the learned channel responses help to differentiate the source models.

The average performance of the evaluated algorithms is summarized in table 1. The experimental results are shown as a function of distractor angle in figure 3. Both MESSL systems generally perform slightly better than 2S-FD-BSS in anechoic conditions, except at small distractor angles when localization is difficult because the parameters for all sources are very similar. In reverberation the 2S-FD-BSS system slightly outperforms the MESSL baseline, but performs worse than MESSL-SP. Finally, the performance of MESSL-SP improves as the size of the source prior increases. The best performer overall is MESSL-SP 64, which outperforms the MESSL baseline by about 2 dB on average in reverb.

Figure 2 demonstrates the qualitative differences between the errors made by MESSL and 2S-FD-BSS in reverberant conditions. There is significant uncertainty in the MESSL mask

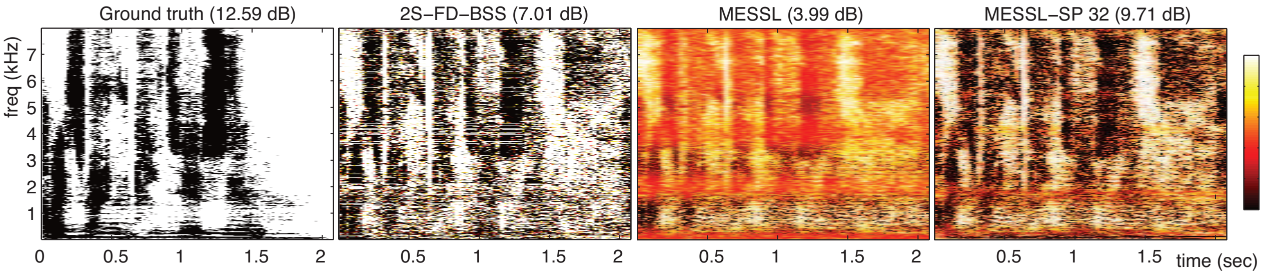

Figure 2: Example binary masks found using different separation algorithms. The mixed signal is composed of two speech sources in reverb separated by 60 degrees. The MESSL-SP 32 mask used a 32 component source prior.

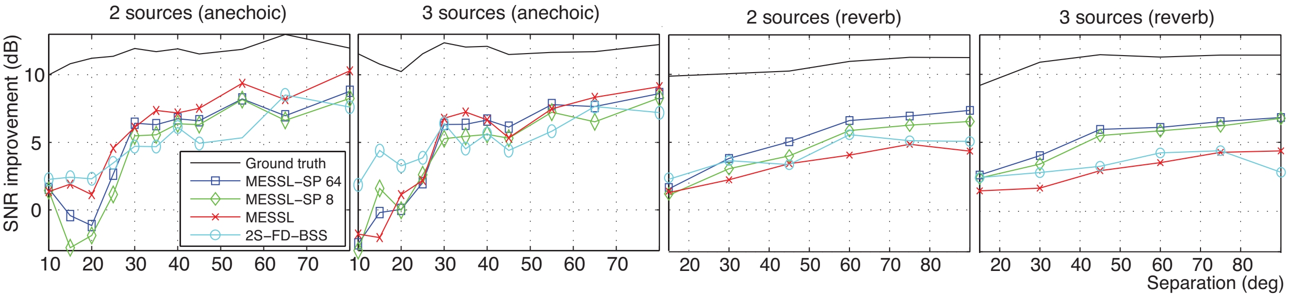

Figure 3: Separation performance as a function of distractor angle.

with many time-frequency points having posteriors close to 0.5 while the 2S-FD-BSS mask is quite close to binary. As a result 2S-FD-BSS is more prone to source permutation errors where significant target energy can be rejected by the mask. In contrast, the failure mode of MESSL is to pass both sources equally when it is unable to sufficiently distinguish between them.

The addition of source constraints hurts performance in anechoic environments. This is because the interaural model alone can be a good fit to anechoic observations, and the limitations of the source model (trained over a variety of speakers, using relatively few components) often become a liability. However, in reverberation the source models are a good fit to the direct path portion of the observations. This helps focus the algorithm on reliable time-frequency regions which in turn enables better inference of the interaural parameters and speeds up convergence relative to the baseline. This is why the MESSL-SP mask is significantly sharper than the MESSL mask in figure 2. The source models also serve to introduce correlations across frequency that are only loosely captured by the ILD model during initial iterations. This is especially true in the higher frequencies which are highly correlated in speech signals.

5. Conclusions

We have presented a system for source separation based on a probabilistic model of binaural observations. We combine a model of the interaural spectrogram with a prior model of the source spectrogram. By learning the frequency response of the combined room impulse response and head-related transfer function filtering applied to the source signal it is possible to obtain a significant performance improvement over the algorithm that does not rely on a prior source model. The improvement is significant even when the prior on the source statistics is quite limited, consisting of a small speaker-independent model.

6. References

[1] R. J. Weiss and D. P. W. Ellis, "Monaural speech separation using source-adapted models," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2007, pp. 114–117.

[2] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[3] 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*, Oct 2007.

[4] 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.

[5] K. Wilson, "Speech source separation by combining localization cues with mixture models of speech spectra," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2007, pp. I–33–36.

[6] M. Mandel, D. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007.

[7] M. Cooke and T. W. Lee, "The speech separation challenge," 2006. [Online]. Available: http://www.dcs.shef.ac.uk/∼martin/SpeechSeparationChallenge.htm

[8] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics*, Oct 2001, pp. 99–102.

[9] B. Shinn-Cunningham, N. Kopco, and T. Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *Journal of the Acoustical Society of America*, vol. 117, pp. 3100–3115, 2005.