Model-Based Expectation Maximization Source Separation and Localization

Michael I. Mandel, *Student Member, IEEE,* Ron J. Weiss, *Student Member IEEE,*
Daniel P. W. Ellis, *Senior Member, IEEE*

***Abstract*—This paper describes a system, referred to as MESSL, for separating and localizing multiple sound sources from an underdetermined reverberant two-channel recording. By clustering individual spectrogram points based on their interaural phase and level differences, MESSL generates masks that can be used to isolate individual sound sources. We first describe a probabilistic model of interaural parameters that can be evaluated at individual spectrogram points. By creating a mixture of these models over sources and delays, the multi-source localization problem is reduced to a collection of single source problems. We derive an expectation maximization algorithm for computing the maximum-likelihood parameters of this mixture model, and show that these parameters correspond well with interaural parameters measured in isolation. As a byproduct of fitting this mixture model, the algorithm creates probabilistic spectrogram masks that can be used for source separation. In simulated anechoic and reverberant environments, separations using MESSL produced on average a signal-to-distortion ratio 1.6 dB greater and PESQ results 0.27 mean opinion score units greater than four comparable algorithms.**

I. Introduction

Humans are very good at focusing their attention on the speech of a single speaker, even in the presence of other speakers and background noise [1]. This ability is greatly diminished, however, when listening with only one ear, especially in reverberant environments [2]. In order for an automatic speech recognizer to focus on a speaker of interest, it must similarly be able to separate that speaker from background noise and stereo recordings likely contain significantly more information relevant to this task than monaural recordings. This paper thus describes a system for simultaneously separating and localizing multiple sound sources from a stereo mixture made in a reverberant environment, performing underdetermined, convolutive source separation.

This system separates sources by modelling each one probabilistically, thus we call it Model-based EM Source Separation and Localization (MESSL). Each source in a mixture is described by a probabilistic model of interaural parameters. By using models that can be evaluated at each point in the spectrogram independently, we can successfully learn multiple models from a broadband mixture and at the same time identify the regions of the spectrogram that best fit each model. In Section II we discuss the construction of such models for individual sources. We then describe in Section III an expectation maximization (EM) algorithm for estimating the parameters of a mixture of these models that reduces the multi-source parameter estimation problem to a collection of single-source estimation problems using probabilistic masking.

This EM algorithm iteratively refines both its estimates of the model parameters and the regions of the spectrogram dominated by each source. In the expectation step of the algorithm, spectrogram points are assigned to models probabilistically, based on the agreement between the observation at that point and each model's prediction. In the maximization step, parameters for each source model are re-estimated from the set of spectrogram points assigned to that model. Currently, these models include the interaural time difference (ITD) via the interaural phase difference (IPD), and the interaural level difference (ILD), but the flexibility of the framework and its probabilistic nature allow the addition of other cues such as source models [3]. This flexibility also lets the system separate mixtures made with both human-like dummy head recordings and free field microphones, although in these experiments we focus on dummy head recordings.

In addition to separating sources well, these models are transparent in their representation of these sources, i.e. the model parameters are interesting in and of themselves. Estimating interaural level and time differences for a single source in an anechoic recording is straightforward, but MESSL estimates these same parameters from a reverberant recording in the presence of other sources, a much more difficult task. This topic is discussed in more detail in Section IV.

Since our system uses spectral masking to separate sources, it assumes that each point in the spectrogram is dominated by no more than one source. This assumption, called W-disjoint orthogonality [4], has been shown generally to hold for simultaneous speech signals.

In Section V we describe a number of experiments that measure the performance of various parts of our system and the system as a whole. We examine the effect of increasing model flexibility (and hence complexity), the effect of cheating by replacing different parts of our system with ground truth parameters, and the use of a "garbage" source to collect reverberation. We also compare our system to four similar state-of-the-art source separation systems in a number of conditions.

These experiments indicate that our model parameterization can separate sources well, but estimating those parameters from a mixture can be difficult and that our most complex model is best, especially when using a garbage source. Under the signal-to-distortion ratio metric [5], MESSL showed an average advantage of 1.6 dB when compared to similar systems.

M. Mandel, R. Weiss, and D. Ellis are with the Department of Electrical Engineering, Columbia University, New York, NY, 10027 USA e-mail: {mim,ronw,dpwe}@ee.columbia.edu

Manuscript received January 1, 2008; revised January 2, 2008.

A. Background

MESSL is a synthesis of localization-based clustering and spectral masking. A number of systems have taken a similar approach [6]–[10]. Localization in azimuth is a popular cue for segregating sound sources [11]. Spectral masking, sometimes called time-frequency masking, binary masking, or ideal binary masking, allows the separation of an arbitrary number of sources from a mixture, by assuming that a single source is active at every time-frequency point. This is in contrast to independent component analysis [12] and other multi-microphone techniques [13], which can separate at most as many sources as there are microphones.

Many models of mammalian auditory localization have been described in the literature, see [11] for a review. Most focus on localization within individual critical bands of the auditory system and are either based on cross-correlation [14] or the equalization-cancellation model [15], [16]. We are more concerned in this work with the way in which localization estimates are combined across bands. The main difficulty in such combination is that the cross-correlation of bandlimited signals is multimodal and the correct mode must be selected. MESSL's localization facility is based on the insight that this multimodality is an artifact of bottom-up processing and that a top-down approach that tests a set of candidate interaural time differences yields an unambiguous answer at all frequencies.

In this bottom-up paradigm, a single source can be localized using the "straightness" of cross-correlations across frequency [17], [18] and simultaneous sources can be localized using a "stencil" filter that embraces multimodality [19]. Other systems localize the source in each band probabilistically and then combine probabilities across frequency by assuming statistical independence. Nonparametric modeling in this vein [9], [20], [21] employs histograms of interaural parameters collected over a large amount of training data, which can be compared to the observation and to one another when normalized properly. While [9], [21] collect histograms of per-band interaural time differences, [20] collects histograms of interaural phase difference, which avoids multimodality and facilitates the analysis of moments. This is the general approach that MESSL takes, although it uses a parametric model of both the noise in interaural parameters and the connection between ITD and frequency, avoiding the need for training data, making it more robust to reverberation, and making it easier to deploy in unfamiliar environments.

When using localization to separate multiple sources in a sound mixture, it is important to be able to localize, in some way, individual time-frequency points. The simplification made by DUET [4], [6] is to ignore high frequencies in which the cross-correlation is multimodal and to convert interaural phase differences to interaural time differences only for frequencies in which the cross-correlation is unimodal. DUET combines these ITD estimates with similar ILD estimates at individual time-frequency points in the spectrogram and then identifies sources as peaks in the two dimensional ILD-ITD histogram. The localization subsystem in [7] aggregates its cross-correlation across frequency and time to estimate the ITD of multiple sources. It then selects the mode in each frequency band's cross-correlation that is closest to one of these global ITD estimates. The localization subsystems of [8], [10] both use ILD to resolve local ITD ambiguities. Using a model of the relationship between ILD and azimuth, they map ILDs to a given direction, then choose the ITD mode that is most consistent with this direction.

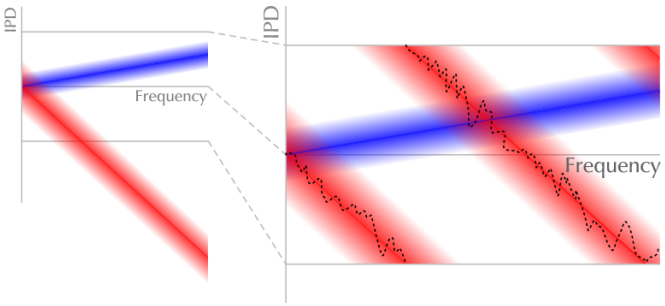


Fig. 1. A caricature of our probabilistic model of interaural phase difference (IPD) as a function of frequency, eq (5). On the left are the probability density functions (PDFs) of two competing model predictions, eq (3), where the IPDs are centered around their respective $\omega\tau$ values. On the right, those PDFs are constrained to the interval $(-\pi, \pi]$ so that the likelihood of the observation, eq (2), the dotted line, can be evaluated under each.

Given the localization of each time-frequency point, most algorithms perform separation through some kind of clustering [10] or histogram peak picking [6], [8], although when the locations are known in advance, classification is also a popular technique [7], [9], [21]. MESSL uses a clustering approach, but simultaneously learns the interaural parameters of each source. This allows it to jointly localize and separate the sources, making it more noise robust than approaches that first commit to azimuth estimates based on noisy observations and then perform clustering in azimuth space [10].

A separation system that shares many similarities with these systems, but does not explicitly use localization is two-source, frequency-domain, blind source separation [22]. It performs an ICA-like clustering in each frequency band separately and then clusters the separation masks across frequency to unpermute them. In the two-microphone case, the per-frequency clustering uses features that are similar to MESSL's, containing the same information as IPD and ILD, but more easily generalizing to multiple microphones. This system is purely for separation, no correspondence is enforced between the parameters estimated at each frequency, as would be implied by a model of localization.

While MESSL jointly localizes and separates sound sources, the current paper only evaluates its separation performance. Localization results are similar to those reported previously [23] and are highly dependent on, if not indistinguishable from, the localization algorithm used to initialize it. We therefore believe that separation is a more challenging and discriminative task for comparing these systems.

II. The Interaural Spectrogram of a Single Source

For the purposes of deriving this model we will examine the situation where one sound source arrives at two spatially distinct microphones or ears. We then generalize this to the assumption that at most one source is present at each time-frequency point in a spectrogram, but that different sources

could be present at different points.

Denote the sound source as $s(t)$ and the signals received at the left and right ears as $\ell(t)$ and $r(t)$, respectively. For a sufficiently narrowband source, the two received signals relate to the source by some delay and gain, in addition to a disruption due to noise. For a wideband source, this delay and gain can vary with frequency, which, in the time domain, manifests as a short impulse response at each ear. In the experiments in Section V-B, we compare models assuming frequency dependence to those assuming frequency-independence.

For analytical convenience, we assume a noise process that is convolutive in the time domain, making it additive in both the log-magnitude and phase domains. Thus the transfer function is modeled as a single large, deterministic coefficient at a certain delay and small, randomly changing coefficients at all other delays. We have found [24] that a reverberant noise model is still able to localize sources in the presence of additive noise, the noise model typically chosen by other authors.

If we combine the frequency-dependent gains and delays into two short impulse responses, $h_\ell(t)$ and $h_r(t)$, the various signals are related by:

$$\begin{aligned} \ell(t) &= s(t-\tau_\ell) * h_\ell(t) * n_\ell(t) \\ r(t) &= s(t-\tau_r) * h_r(t) * n_r(t). \end{aligned} \tag{1}$$

The ratio of the short-time Fourier transforms, $\mathcal{F}\{\cdot\}$, of both equations is the interaural spectrogram,

$$\frac{L(\omega,t)}{R(\omega,t)} = 10^{\alpha(\omega,t)/20} e^{j\phi(\omega,t)} \tag{2}$$

$$\approx 10^{a(\omega)/20} e^{-j\omega\tau(\omega)} N(\omega,t) \tag{3}$$

where $N(\omega,t) = \frac{N_\ell(\omega,t)}{N_r(\omega,t)} = \frac{\mathcal{F}\{n_\ell(t)\}}{\mathcal{F}\{n_r(t)\}}$, $\tau(\omega) = \tau_\ell - \tau_r + \angle H(\omega)$, $a(\omega) = 20\log_{10}|H(\omega)|$, and $H(\omega) = \frac{\mathcal{F}\{h_\ell(t)\}}{\mathcal{F}\{h_r(t)\}}$. Equation (2) is the ratio of the actual observations at both ears, while (3) is our model of that ratio. For this model to hold, τ must be much smaller than the window over which the Fourier transform is taken. Our experiments used a dummy-head microphone where position-dependent delay differences were limited to about 0.75ms, while the window length was 64ms (1024 samples at a sampling rate of 16 kHz). Similarly, $h(t)$ must be smaller than the window, but because distinguishing between $h(t)$ and $n(t)$ is an ill-posed problem, parts of $h(t)$ beyond one window's length can be considered part of $n(t)$, with a corresponding increase in the noise variance.

The interaural spectrogram is parameterized by $\phi(\omega,t)$, the interaural phase difference (IPD) at frequency ω and time t, and $\alpha(\omega,t)$, the interaural level difference (ILD) measured in dB. We model these with the frequency-dependent interaural time difference (ITD), $\tau(\omega)$, and the frequency-dependent interaural level difference, $a(\omega)$. All IPD observations are constrained to the interval $(-\pi,\pi]$, and spatial aliasing prevents IPD from being mapped directly to a unique τ. Every τ, however, maps unambiguously to a phase difference at every frequency. Using a top-down approach, then, it is possible to test how well any τ fits an observed IPD, even when multiple sources are present and even at only a subset of frequencies.

Figure 1 shows a caricature of this top-down process. The two shaded bars correspond to two competing τ predictions and the dotted line to the actual observation. On the left hand side of the figure is the model predictions of (3), on the right hand side are the observations of (2), the dotted line. Even though the phase wrapping makes the bottom-up observed-IPD-to-ITD mapping ambiguous, the top-down ITD-to-IPD mapping is unambiguous and it can be seen that the observation is much more likely under one of the predictions than the other. Note, however, that the phase wrapping does cause ambiguity at certain frequencies, where the two τs predict the same IPD, as can be seen in the right half of the figure, where the bars cross. A similar effect can be seen in the horizontal lines of lower probability in the example masks of Figure 5(b)–(f).

To measure the difference between the IPD predicted by a delay of τ samples and the observed IPD, we define the phase residual $\hat{\phi}$ as

$$\hat{\phi}(\omega,t;\tau) = \arg\left(e^{j\phi(\omega,t)} e^{-j\omega\tau(\omega)}\right) \tag{4}$$

which is always in the interval $(-\pi,\pi]$. Without this calculation, phase circularity becomes a problem when observations originate from delays that are not close to 0. With this calculation, phase circularity is no longer a problem for delays that are close to the observations' true delay, but only for those that are farther away. The residual error can be modeled with a circular probability distribution like the von Mises distribution [24], or approximated as a linear distribution using a Gaussian scale mixture model [23]. We have found, however, that a single Gaussian works well enough in practice

$$p(\phi(\omega,t) \,|\, \tau(\omega), \sigma(\omega)) = \mathcal{N}\big(\hat{\phi}(\omega,t;\tau(\omega)) \,|\, 0, \sigma^2(\omega)\big) \tag{5}$$

$$\approx \mathcal{N}\big(\phi(\omega,t) \,|\, \omega\tau(\omega), \sigma^2(\omega)\big). \tag{6}$$

The equality is approximate because it only holds when the standard deviation σ is small relative to 2π, in which case the linear Gaussian is very similar to a von Mises distribution. Even when the standard deviation is small, though, the distribution of the IPD can be thought of as a Gaussian with mean $\omega\tau(\omega)$ only if the mean is subtracted from samples in a way that respects phase wrapping, hence the need for (4).

From observations measured in dB, the interaural level difference similarly appears to be well modeled by a single Gaussian with frequency-dependent mean and variance

$$p(\alpha(\omega,t) \,|\, \mu(\omega), \eta^2(\omega)) = \mathcal{N}\big(\alpha(\omega,t) \,|\, \mu(\omega), \eta^2(\omega)\big). \tag{7}$$

We combine the ILD and IPD models by assuming that they are conditionally independent, given their respective parameters

$$\begin{aligned} &p(\phi(\omega,t), \alpha(\omega,t) \,|\, \Theta) = \\ &\quad \mathcal{N}\big(\hat{\phi}(\omega,t) \,|\, \xi(\omega), \sigma^2(\omega)\big) \cdot \mathcal{N}\big(\alpha(\omega,t) \,|\, \mu(\omega), \eta^2(\omega)\big), \end{aligned} \tag{8}$$

where Θ represents all of the model parameters. Note that this assumption of conditional independence applies only to the noise that corrupts the measurements, it does not contradict the well known correlation between ILD and ITD in actual head-related transfer functions, which should be enforced instead in the means of these Gaussians, $\xi(\omega)$ and $\mu(\omega)$. In this work, we model this correlation by enforcing a prior on the ILD based on the initialization of the ITD. Because the ILD is modeled with a Gaussian, we use the normal-Wishart distribution, its

conjugate, as the prior [25]. The mean of this normal-Wishart is set from the initial ITD using data on their relationship collected from synthetic impulse responses, but the system is not especially sensitive to particular values.

Equation (8) can be used to evaluate the likelihood of an observation at any point in a spectrogram under the model specified by the parameters Θ. Points are assumed to be independent of one another as well, so such a likelihood can be computed for any set of points by multiplying the individual points' likelihoods. Such a set of points could be a spectrogram frame, as in traditional cross-correlation, but could just as easily be a frequency band or an arbitrarily shaped region in the spectrogram.

III. Parameter estimation from mixtures

The parameters of the model described above cannot be estimated directly from a mixture of sources, where different regions of the spectrogram are dominated by different sources, because the sources have different distributions over IPD and ILD. Only points from the same source and at the same delay are assumed to be distributed identically. The parameters of each of these distributions, then, could only be estimated if somehow the source and delay of each point in the spectrogram were known.

This is a classic missing data problem and the maximum-likelihood model parameters can be estimated with an expectation maximization (EM) algorithm. For each source in a multisource mixture, this EM algorithm selects the regions of the spectrogram that best fit its parameters, and then re-estimates its parameters from only those regions. Instead of using hard, binary masks, however, EM uses soft, likelihood-based masks. It treats the source dominating each spectrogram point as a hidden variable, i, and uses the expected probability of source membership to estimate source parameters.

The delay of the source that dominates each spectrogram point, $\tau(\omega)$, also includes a hidden variable. We model it as the sum of two components,

$$\tau(\omega) = \tau + \omega^{-1}\xi(\omega). \tag{9}$$

The first term is a frequency-independent delay that is used for localization. To make the inference tractable, the hidden variable τ is modeled as a discrete random variable, where the set of allowable delays is specified *a priori*. The parameter $\xi(\omega)$ is an offset in $(-\pi, \pi]$ that allows for minor deviations from this frequency-independent model. It replaces 0 as the mean of the IPD Gaussian in (5). Both i and τ are combined into the hidden variable $z_{i\tau}(\omega, t)$, which is 1 if spectrogram point (ω, t) comes from both source i and delay τ and 0 otherwise. Each observation must come from some source and delay, so $\sum_{i,\tau} z_{i\tau}(\omega, t) = 1$.

The parameters of the various Gaussians are estimated in the M step along with the marginal class memberships, $\psi_{i\tau} \equiv p(i, \tau)$, the estimate of the joint probability of any spectrogram point's being in source i at delay τ. Estimates of the time-delay of arrival for each source can be computed from $\psi_{i\tau}$. Since τ only takes on discrete values in our model, $\psi_{i\tau}$ is a two-dimensional matrix of the probability of being in each discrete state.

In the most general form of the model, we include all possible dependencies of the parameters on dimensions of the data. In Section V-B we compare different parameter tying schemes, which reduce these dependencies. The IPD parameters for each source, $\sigma_{i\tau}(\omega)$ and $\xi_{i\tau}(\omega)$, depend on τ as well as ω. The ILD parameters for each source, $\mu_i(\omega)$ and $\eta_i(\omega)$ depend on ω, but are independent of τ. Both IPD and ILD parameters depend on the source from which they come, i. Let

$$\Theta \equiv \{\xi_{i\tau}(\omega), \sigma_{i\tau}(\omega), \mu_i(\omega), \eta_i(\omega), \psi_{i\tau}\} \tag{10}$$

be the collection of the parameters of all of the models. By marginalizing over the hidden variable $z_{i\tau}(\omega, t)$, we arrive at the total log likelihood for a given observation

$$\begin{aligned}
\mathcal{L}(\Theta) &= \sum_{\omega,t} \log p(\phi(\omega,t), \alpha(\omega,t) \mid \Theta) && (11)\\
&= \sum_{\omega,t} \log \sum_{i,\tau} \Big[p(\phi(\omega,t), \alpha(\omega,t) \mid z_{i\tau}(\omega,t), \Theta) \\
&\qquad \cdot p(z_{i\tau}(\omega,t) \mid \Theta) \Big] && (12)\\
&= \sum_{\omega,t} \log \sum_{i,\tau} \Big[\mathcal{N}\big(\hat{\phi}(\omega,t;\tau) \mid \xi_{i\tau}(\omega), \sigma^2_{i\tau}(\omega)\big) \\
&\qquad \cdot \mathcal{N}\big(\alpha(\omega,t) \mid \mu_i(\omega), \eta_i^2(\omega)\big) \cdot \psi_{i\tau} \Big]. && (13)
\end{aligned}$$

This is basically a Gaussian mixture model, with one Gaussian per (i, τ) combination and $\psi_{i\tau}$ as the mixing weights. The number of sources to compare must be specified *a priori*.

From this total log likelihood, we define the auxiliary function to maximize with respect to Θ,

$$\begin{aligned}
Q(\Theta \mid \Theta_s) = k + \sum_{\omega,t} \sum_{i,\tau} \Big[& p\big(z_{i\tau}(\omega,t) \mid \phi(\omega,t), \alpha(\omega,t), \Theta_s\big) \\
& \cdot \log p\big(z_{i\tau}(\omega,t), \phi(\omega,t), \alpha(\omega,t) \mid \Theta\big) \Big] \qquad (14)
\end{aligned}$$

where Θ_s is the estimate of the parameters Θ after s iterations of the algorithm and k is independent of Θ. Maximum-likelihood parameter estimation then proceeds in two steps, the E step, in which the expectation of $z_{i\tau}(\omega, t)$ is computed given the observations and the parameter estimate Θ_s, and the M step, in which Q is maximized with respect to Θ given the expected value of $z_{i\tau}(\omega, t)$.

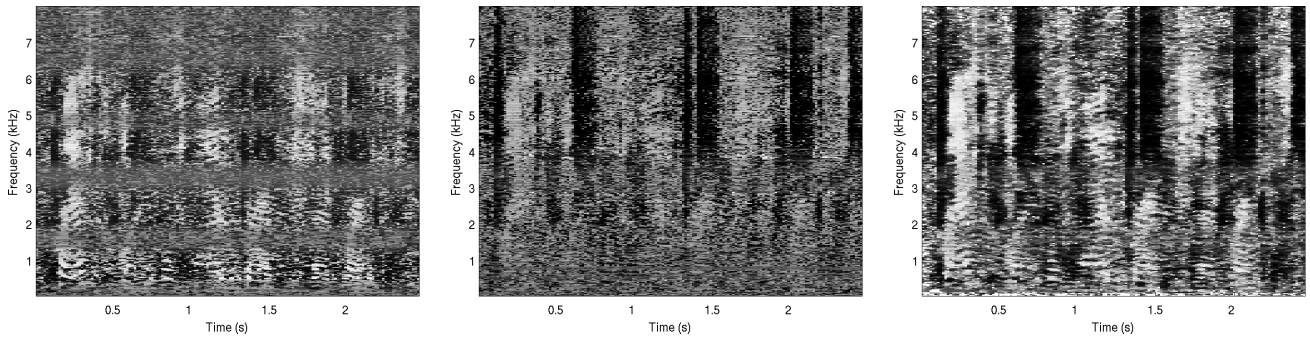
In the E step, we compute

$$\begin{aligned}
\nu_{i\tau}(\omega,t) &\equiv p(z_{i\tau}(\omega,t) \mid \phi(\omega,t), \alpha(\omega,t), \Theta_s) && (15)\\
&\propto p(z_{i\tau}(\omega,t), \phi(\omega,t), \alpha(\omega,t) \mid \Theta_s) && (16)\\
&= \psi_{i\tau} \cdot \mathcal{N}\big(\hat{\phi}(\omega,t;\tau) \mid \xi_{i\tau}(\omega), \sigma^2_{i\tau}(\omega)\big) \\
&\qquad \cdot \mathcal{N}\big(\alpha(\omega,t) \mid \mu_i(\omega), \eta_i^2(\omega)\big). && (17)
\end{aligned}$$

Because $z_{i\tau}(\omega, t)$ is a binary random variable, this probability is equal to its expectation, hence this is the "expectation" step. This expectation is then used in the M step to calculate maximum-likelihood parameters as weighted means of sufficient statistics. Let the operator

$$\langle x \rangle_{t,\tau} \equiv \frac{\sum_{t,\tau} x\, \nu_{i\tau}(\omega,t)}{\sum_{t,\tau} \nu_{i\tau}(\omega,t)} \tag{18}$$

be the weighted mean over the specified variables, in this case t and τ. This notation makes it convenient to specify the indices



(a) IPD-only mask (b) ILD-only mask (c) Combined mask

Fig. 2. Contribution of (a) IPD and (b) ILD to (c) MESSL mask using frequency-dependent $\Theta_{\Omega\Omega}$ parameters. White is 1, black 0. Two speakers at 0° and 75° in reverberation. Notice that the IPD mask is mainly informative for low frequencies and the ILD mask for high frequencies, but not exclusively. For this same example, Figures 3 and 4 show the ground truth and estimated parameters for interaural level and phase differences, respectively.

over which to take the mean, i.e. the indices over which to tie parameters. Without tying, and with an uninformative ILD prior the model parameter updates are

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

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

$$\xi_{i\tau}(\omega) = \left\langle \hat{\phi}(\omega, t; \tau) \right\rangle_t \tag{21}$$

$$\sigma_{i\tau}^2(\omega) = \left\langle \big(\hat{\phi}(\omega, t; \tau) - \xi_{i\tau}(\omega)\big)^2 \right\rangle_t \tag{22}$$

$$\psi_{i\tau} = \frac{1}{\Omega T} \sum_{\omega,t} \nu_{i\tau}(\omega, t). \tag{23}$$

To tie parameters, the weighted mean is taken across more variables. For example, different versions of the IPD variance can be derived

$$\sigma_{i\tau}^2 = \left\langle \big(\hat{\phi}(\omega, t; \tau) - \xi_{i\tau}(\omega)\big)^2 \right\rangle_{\omega,t} \tag{24}$$

$$\sigma_i^2 = \left\langle \big(\hat{\phi}(\omega, t; \tau) - \xi_{i\tau}(\omega)\big)^2 \right\rangle_{\omega,t,\tau} \tag{25}$$

In addition to the model parameters that are output by the algorithm, it is also possible to compute probabilistic masks for each of the sources by marginalizing over delay

$$M_i(\omega, t) \equiv \sum_{\tau} \nu_{i\tau}(\omega, t). \tag{26}$$

Note that this mask represents probabilities and preliminary experiments indicate that separation performance can be improved by converting these probabilities to more Wiener filter-like coefficients. We do not do so in this paper, and this conversion should be explored in future work.

Under weak assumptions, this algorithm is guaranteed to converge to a local maximum of the likelihood, but since the total log likelihood is not convex, it is still sensitive to initialization. Conveniently, however, it is also very flexible in its initialization. Since it can start with the E step or the M step, it can be initialized with data in the form of either model parameters or masks. Even a subset of the model parameters can be used to initialize the algorithm, from which the rest can be bootstrapped. In Section V-C we compare the results of initializing the algorithm in various ways.

Unless otherwise mentioned, we initialize $\psi_{i\tau}$ from a cross-correlation based method while leaving all the other parameters in a symmetric, non-informative state. If the ILD prior (described below) is used, we initialize the ILD with the same mean as its prior and a standard deviation of 10 dB. From these parameters, we compute the first E step mask. Using estimates of τ for each source from PHAT-histogram [26], $\psi_{i\tau}$ is initialized to be centered at each cross-correlation peak and to fall off away from that. Specifically, $p(\tau \mid i)$, which is proportional to $\psi_{i\tau}$, is set to be approximately Gaussian, with its mean at each cross-correlation peak and a standard deviation of one sample.

In order to model reverberation, we introduce a "garbage" source into MESSL that is initialized to have a uniform $p(\tau \mid i)$, a uniform distribution across IPD, and an ILD with 0 mean across frequency. This garbage source is designed to account for spectrogram points that are not well described by any of the other source models. While the direct-path signal has interaural cues consistent with the specific direction of the source, reverberation has a diffuse character that may not fit a source model particularly well. Thus a single garbage source should be able to account for the reverberation from all of the sources in a mixture, regardless of their locations. The garbage source also allows the parameters of the other sources to be estimated more accurately, as they are no longer forced to include poorly fitting points into their parameter estimation.

The ILD prior affects the estimation of the ILD parameters $\mu_i(\omega)$ and η_i in (19) and (20). In effect, the prior acts as a number of "virtual" observations that are included in (19) and (20). The prior precision (inverse variance) controls the strength of the prior relative to the observations, i.e. the number of virtual observations. The exact value of the prior mean was estimated from a set of synthetic binaural room impulse responses, using a regression on ITD and frequency and interaction terms up to the third order. The fact that this mean is only able to capture broad features of the relationship between ILD and ITD makes it better able to generalize across individuals. We only employ the ILD prior when also using the garbage source, as a pilot study found that that was when it was most useful.

A note about computational complexity. The running time of this algorithm is linear in the number of points in the

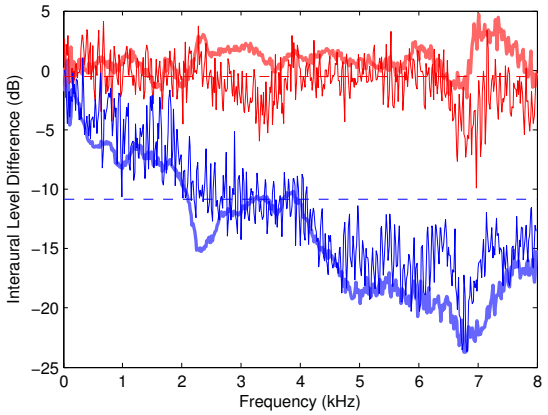


Fig. 3. Example ILD for two sources at 0° (upper lines) and 75° (lower lines) in reverberation. Thick lines are ground truth direct-path ILD measured in isolation, thin lines are estimated from a mixture by MESSL using frequency-dependent Θ_G parameters, and the dashed lines are estimated by MESSL using frequency-independent Θ_{11} parameters.

spectrogram, the number of sources, the number of discrete values of τ that are used, and the number of EM iterations. Running time on a 1.86 GHz Intel Xeon processor was approximately 80 seconds to separate 2 sources from a 2.5-second (40,000 sample) mixture using a τ grid of 61 elements and 16 EM iterations. This processing rate is approximately 32 times slower than real time. To separate 3 sources under the same conditions took approximately 110 seconds, or 45 times slower than real time.

IV. Example parameter estimates

We now present an example separation illustrating some of MESSL's properties. The example mixture includes two speakers in a reverberant environment. The target speaker is female and is located at 0°, saying, "Presently, his water brother said breathlessly." The interfering speaker is male and is located at 75°, saying, "Tim takes Sheila to see movies twice a week." It is taken directly from the experiments in Section V-D without any modification and is used in Figures 2, 3, 4, and 5. Sound files from this example are also available on the project's webpage[1]. Two MESSL parameterizations were used in this example, one frequency-dependent and the other frequency-independent. Both use the garbage source and ILD prior.

Figure 2 shows the contributions of the IPD and ILD to the full MESSL mask when using the frequency-dependent parameterization. Note that as in human hearing, the IPD mask is generally informative at low frequencies, while the ILD mask is generally informative at high frequencies. Unlike in human hearing, however, the IPD is effective at separating the sources up to 6 kHz and beyond, albeit with periodic interruptions. These interruptions occur when both sources' ITD models predict the same IPD at a given frequency. The IPDs from any pair of ITDs will coincide at a certain set of harmonically related frequencies. See Figure 1 for an example of this ambiguity. In the combined mask, these bands of uncertainty remain present at low frequencies, but are resolved at high frequencies by the ILD.

[1]http://labrosa.ee.columbia.edu/projects/messl

The parameters estimated by MESSL are not only effective at separating sources, but are also interesting in and of themselves. Figure 3 compares the ILD that MESSL estimates to the ground truth direct-path ILD measured in isolation. Notice that the estimate closely follows the ground truth for both sources even though neither source was ever observed alone. The parameters that MESSL estimates are different from the ILD measured for a single source in reverberation, as reverberation tends to reduce the magnitude of ILD cues [27]. MESSL's ILD prior helps it to estimate the correct direct-path parameters and the garbage source absorbs reverberant time-frequency points that do no fit these parameters well.

Similarly, Figure 4 compares the probability density function of IPD that MESSL estimates to the ground truth IPD measured in isolation. Even without depending on frequency, this model accurately estimates the IPD of both sources, again without ever observing either source alone. Marginalizing over the delay yields the probability of a given IPD under the model

$$p(\phi(\omega,t)\,|\,i) = \sum_\tau \psi_{i\tau}\mathcal{N}\big(\hat{\phi}(\omega,t;\tau)\,|\,\xi_{i\tau}(\omega),\sigma^2_{i\tau}(\omega)\big). \quad (27)$$

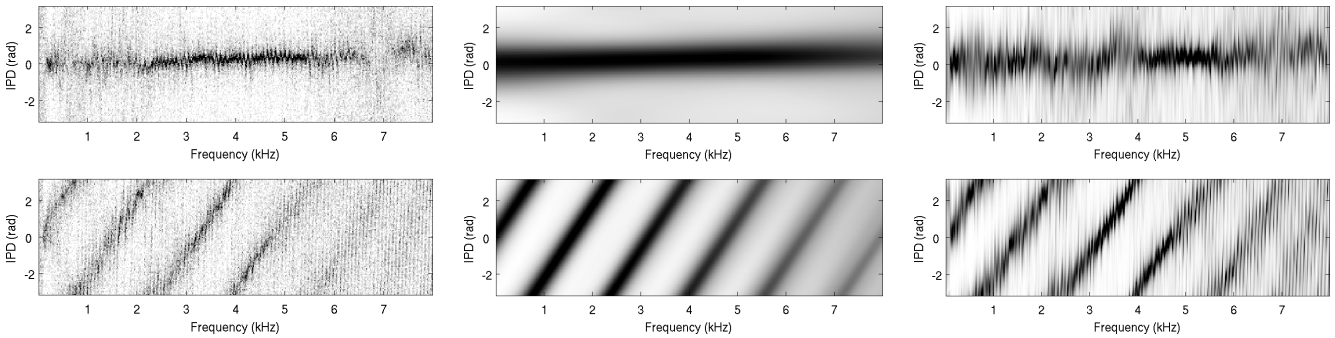
Each component in this mixture is a line with a different slope, as illustrated in Figure 1, although for Figure 4 there are 61 lines rather than two. In the frequency-independent case, the width of each line is constant in IPD. At lower frequencies, these mixture components are very close together and so the variance of the mixture is lower. At higher frequencies, however, the various mixture components are spaced farther apart due to their proportionality to ω, and so the model is less informative.

Even though the IPD of the source at 0° is not distributed exactly around 0 IPD, the model is able to approximate it by mixing together components for a few delays that are close to the true ITD. Thus, the marginal IPD distribution (27) is still able to vary with frequency, even when the parameters $\xi_{i\tau}$ and $\sigma_{i\tau}$ do not, as can be seen in Figure 4(b). Also, learning τ-dependent, but frequency-independent $\xi_{i\tau}$ parameters can more favorably align the straight lines that sum to the best IPD model than setting them to 0.

Figure 4(c) shows the extra information captured by the frequency-dependent parameters. This extra information mostly takes the form of rapid fluctuations of the mean and variance of the IPD with frequency, particularly at high frequencies. It is not clear to us what these represent, possibly aspects of the room impulse response like early echoes, and further investigation is warranted.

V. Speech separation experiments

We perform three experiments in order to examine MESSL's performance and compare it to four other well-known algorithms. The basic form of the experiments is as follows. Single sentence utterances recorded with a microphone close to the speaker are convolved with binaural impulse responses recorded from a KEMAR dummy head, simulating speech originating from a particular direction. A number of these utterances, simulated at different azimuthal directions in the horizontal plane, are then mixed together. The target speaker is always directly in front of the dummy head (0°) while the others are up to 90° to either side. These stereo mixtures are given to



(a) Before mixing (b) Estimated by MESSL with Θ_{11} (c) Estimated by MESSL with Θ_G

Fig. 4. Interaural phase differences (IPD) as a function of frequency for two speakers at 0° and 75° in reverberation. Black is higher probability. (a) Histogram of each source's IPD before mixing, (b) PDF of each source's IPD estimated from the mixture by MESSL using frequency-independent parameters, Θ_{11}. (c) PDF of each source's IPD estimated from the mixture by MESSL using frequency-dependent parameters, $\Theta_{\Omega\Omega}$.

the algorithms to separate along with the number of sources that were mixed together. The algorithms attempt to extract the target signal, which is mixed down to mono and evaluated on its signal-to-distortion ratio.

A. Shared experimental details

1) Data sources: The utterances used are from the TIMIT acoustic-phonetic continuous speech corpus [28], a dataset of utterances spoken by 630 native American English speakers. Of the 6300 utterances in the database, we randomly selected 15 of approximately the same duration to use in our evaluation. Each utterance is approximately 3 seconds long, and we used the first 2.5 seconds to avoid ending silence. Before convolution with the binaural impulse responses, all utterances were normalized to have the same root mean square energy. All of the experiments either include two or three simultaneous speakers, a single target and one or two distractors. In the three speaker case, the two distractors were situated symmetrically about the target.

The binaural impulse responses come from two sources. For simulating anechoic signals, we use the head-related impulse responses from [29], an effort to record such impulse responses for many individuals. We use the measurements of the KEMAR dummy with small ears, although the dataset contains impulse responses for around 50 individuals. We restrict our attention to the 50 impulse responses measured in the horizontal plane, both in front and in back of the dummy. These 50 impulse responses were measured more densely near the median plane and more sparsely to the sides.

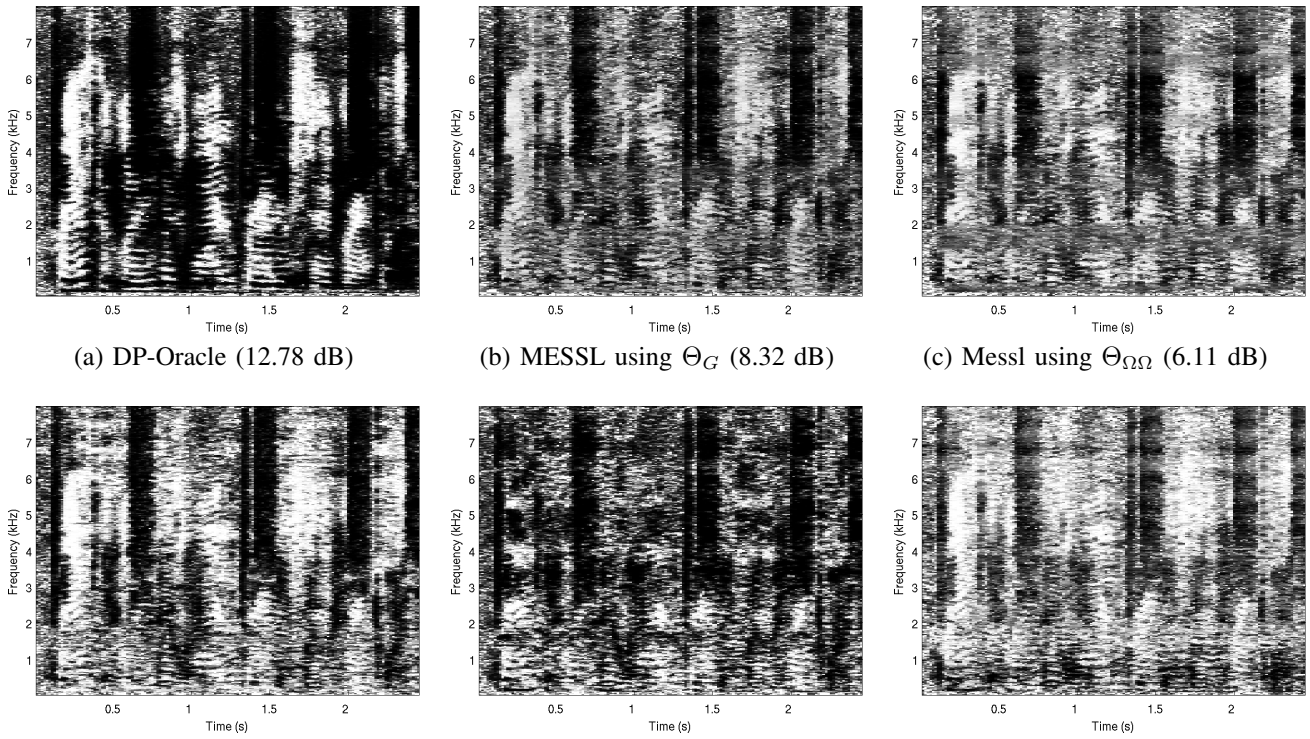
The second set of binaural impulse responses comes from [27]. They were recorded in a real classroom with a reverberation time of around 565 ms. These measurements were also made on a KEMAR dummy, although a different actual dummy was used. We used the measurements taken in the middle of the classroom, with the source 1 meter from the listener, at 7 different angles spaced evenly between 0° and 90°, i.e. all the way to the right. The recording equipment was taken down and setup on three different occasions, so there are three recordings at each location, for a total of 21 binaural impulse responses. For three-source measurements, we transpose the two ears to simulate the source on the left.

For each configuration of the simulated sources, i.e. each off-axis binaural impulse response, we randomly select five different sets of utterances. Thus, for anechoic mixtures, 10 different sets of utterances are mixed at each angular separation, five in front and five behind the listener. For reverberant mixtures, 15 different sets of utterances are mixed at each angular frequency, five for each of the three repetitions of the impulse response measurement. Each configuration was repeated for two and three speakers, for a total of 240 different anechoic mixtures and 180 different reverberant mixtures.

2) Evaluation metrics: We evaluate separation performance using the signal-to-distortion ratio (SDR) [5]. This metric is the ratio of the energy in the original signal to the energy in interference from other signals and other unexplained artifacts. Any energy in the estimated signal that can be explained with a linear combination of delayed versions of the target signal (up to 32 ms) counts towards the target energy. Similarly, any energy that can be explained with a linear combination of delayed versions of the interferer signals counts towards interferer energy. Any energy that cannot be explained by either of these projections is deemed to be an artifact, most notably reverberation from any of the sources. Using the same structure, we can also compute the signal-to-interferer ratio (SIR), the ratio of the target to interferer energy, which ignores artifacts like reverberation.

We also evaluate the speech quality of the separations using the Perceptual Evaluation of Speech Quality, or PESQ [30, Sec. 10.5.3.3]. This measure is highly correlated with the Mean Opinion Score (MOS) of human listeners asked to evaluate the quality of speech examples. MOS ranges from -0.5 to 4.5, with 4.5 representing the best possible quality. Although it was initially designed for use in evaluating speech codecs, PESQ can also be used to evaluate speech enhancement systems [31]. The PESQ results for each algorithm in Table III should be compared to those of the unseparated mixtures in the last row of that table.

3) Control "algorithms": All algorithms are compared against three control masks, two ground truth 0 dB masks, and a random mask. These masks are included to provide upper and lower bounds on the separation performance achievable using spectral masking. Since our test mixtures are created synthetically, we use knowledge of the original separate sources to create ground truth binary masks. Optimal masking separation is achieved by a mask that is 1 at every spectrogram



(a) DP-Oracle (12.78 dB) (b) MESSL using Θ_G (8.32 dB) (c) Messl using $\Theta_{\Omega\Omega}$ (6.11 dB)

(d) Sawada (6.87 dB) (e) DUET (5.48 dB) (f) Mouba (5.58 dB)

Fig. 5. Example masks from the various algorithms compared in Section V-D (white is 1, black 0). Two speakers at 0° and 75° in reverberation. SDR for each mask is in parentheses. Not shown: TRINICON (5.02 dB), Random mask (−3.52 dB). Note that (b) is the same as Figure 2(c).

point in which the target signal is at least as loud as the combined interferers and 0 everywhere else [4]. Because this information is not available in an actual source separation application, this mask is referred to as the Oracle mask. In our notation, the Oracle and DP-Oracle masks are the same except for their treatment of reverberation from the target source.

The Oracle mask considers reverberation from the target source to be part of the desired signal. In reality, however, reverberation is an unwanted addition to the signal and should be considered interference rather than target. Thus, the DP-Oracle mask only considers direct-path energy from the target source as desirable. For anechoic mixtures, the two masks are equivalent. Since we simulate source location by convolution with impulse responses, we are able to isolate the direct-path of each impulse response by forcing all of the samples after 10 ms to be 0. This eliminates both early reflections and reverberation while still coloring the signal with the anechoic head-related transfer function, making it comparable to the separated sources.

The lower-bound control mask is one in which each spectrogram point is assigned uniformly at random to one of the sources. Its SDR is typically the same as the SDR of the mixture.

B. Model complexity

The first experiment examines the tying of various parameters to determine the amount of complexity our test data can support. Tying parameters means that model parameters that are similar in some way are forced to take the same value, for example the IPD variance could be forced to be the same across frequency as in (24) and (25). Parameters can be tied across frequency, across delay, across sources, or assumed to have a fixed value, although certain combinations make more sense than others. This tying is easy to enforce in the M step equations (19)-(23), by averaging together tied parameters.

In this experiment, we compare the signal-to-distortion ratio of MESSL with different amounts of parameter tying when separating two sources in reverberation. See Table I for the list of parameters used for each version of the algorithm. The parameter sets are named by their complexity in ILD and IPD, respectively, 0 being simplest, 1 being more complex, and Ω being most complex, i.e. frequency-dependent. For example, the model with parameters $\Theta_{\Omega 0}$ uses a complex ILD model, but a simple IPD model. The model with parameters Θ_{01} on the other hand uses a simple ILD model and a moderately complex IPD model. See Table I for the specific parameterization used in each condition and the separation results.

The version of MESSL using Θ_{00} was published in [23] and referred to as EM−ILD in [32]. It uses only IPD and has a single σ per source. The versions using Θ_{10} and $\Theta_{\Omega 0}$ were referred to as EM+1ILD and EM+ILD in [32], they use both IPD and ILD, but their IPD model again only uses a single σ per source. The others are introduced for the first time in this paper. The versions using Θ_{01} and $\Theta_{0\Omega}$ have τ-dependent IPD mean and variance, but no ILD. The last three versions use the full IPD and ILD models. With Θ_{11}, both are frequency-independent and with $\Theta_{\Omega\Omega}$ both are frequency-dependent. The Θ_G parameters are the same as the $\Theta_{\Omega\Omega}$ parameters, but also include the garbage source and ILD prior.

It should be noted that initializing models with a large

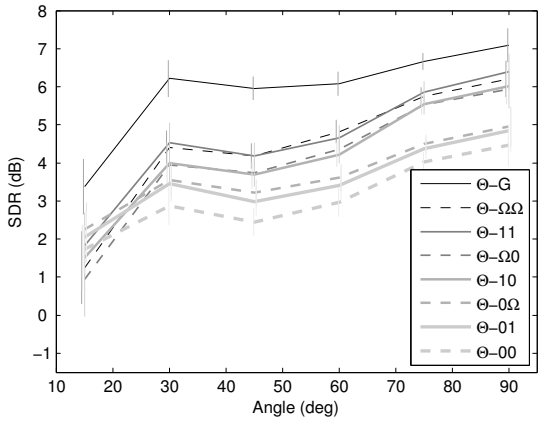


Fig. 6. Signal-to-distortion ratio for complexity experiment, two speakers in reverberation. Each point is the average of 15 mixtures with the same angular separation. Error bars show 1 standard error. Note that the random and DP-Oracle masks bound performance between 11.79 and -0.78 dB SDR, respectively, independent of angle.

number of parameters requires some care to avoid source permutation errors and other local maxima. This is most important with regards to parameter tying across frequency. To address this problem, we use a bootstrapping approach where initial EM iterations are performed with frequency-independent models, and frequency-dependence is gradually introduced. Specifically, for the first half of the total number of iterations, we tie all of the parameters across frequency. For the next iteration, we tie the parameters across two groups, the low and high frequencies, independently of one another. For the next iteration, we tie the parameters across more groups, and we increase the number of groups for subsequent iterations until in the final iteration, there is no tying across frequency and all parameters are independent of one another, but still consistent.

1) Results: A summary of the results of this experiment can be seen in Table I. The most complex model, using Θ_G achieves the best separation by 1.3 dB. Note that the models after that are paired as performance is comparable with frequency-dependent and frequency-independent parameters. Comparing $\Theta_{\Omega 0}$ with Θ_{00}, the frequency-dependent ILD increases the signal-to-distortion ratio of the target by 1.0 dB. The more complete model of interaural phase difference present in $\Theta_{0\Omega}$ provides an extra 0.6 dB of separation over Θ_{00}. Their combination is not quite additive, $\Theta_{\Omega\Omega}$ increases SDR by 1.4 dB over Θ_{00}.

A graph of MESSL's performance using each of these parameter sets versus interferer azimuth can be seen in Figure 6. First note that all of the algorithms perform similarly compared to the two controls (which are not shown). Second, note that they also all perform worse as the separation decreases. Third, note that the ILD improves separation, except at the smallest angles. And finally, note that the models that use the more complicated parameterization of the IPD, including a τ-dependent mean and variance, are able to realize proportionally larger improvements at larger separations than those that use the simpler parameterization with zero mean and a single σ per source.

TABLE I
SDR FOR DIFFERENT MODEL COMPLEXITIES, SEPARATING TWO SPEAKERS IN REVERBERATION, AVERAGED OVER 15 MIXTURES AT EACH OF 6 ANGULAR SEPARATIONS.

Name	ILD mean	ILD std	IPD mean	IPD std	SDR (dB)
Θ_G	$\mu_i(\omega)$	$\eta_i(\omega)$	$\xi_{i\tau}(\omega)$	$\sigma_{i\tau}(\omega)$	5.87
$\Theta_{\Omega\Omega}$	$\mu_i(\omega)$	$\eta_i(\omega)$	$\xi_{i\tau}(\omega)$	$\sigma_{i\tau}(\omega)$	4.45
Θ_{11}	μ_i	η_i	$\xi_{i\tau}$	$\sigma_{i\tau}$	4.57
$\Theta_{\Omega 0}$	$\mu_i(\omega)$	$\eta_i(\omega)$	0	σ_i	4.07
Θ_{10}	μ_i	η_i	0	σ_i	4.16
$\Theta_{0\Omega}$	0	∞	$\xi_{i\tau}(\omega)$	$\sigma_{i\tau}(\omega)$	3.69
Θ_{01}	0	∞	$\xi_{i\tau}$	$\sigma_{i\tau}$	3.51
Θ_{00}	0	∞	0	σ_i	3.08

C. Initialization

Our second experiment compares the normal version of MESSL with other versions that are allowed to "cheat" using ground truth information. This experiment provides insight into the ability of the non cheating model to extract parameters from mixtures and the ways in which the imperfections in parameter estimates from mixtures hurt separation. Note that all models use frequency-dependent parameters for this experiment, with no garbage source and no ILD prior.

The ground truth parameters were extracted using MESSL on "mixtures" of a single source at a time. White noise was passed through each pair of reverberant binaural impulse responses and then fed into MESSL, allowing it to make the best possible estimates of the ILD parameters $\mu_i(\omega)$ and $\eta_i(\omega)$ and the IPD parameters $\xi_{i\tau}(\omega)$, $\sigma_{i\tau}(\omega)$, and $\psi_{i\tau}$. To perform actual separations, the ideal parameters for each source in the mixture were combined in initializing MESSL. The DP-Oracle mask was used for initializing from a ground truth mask.

Seven different initializations were compared in this experiment on the 90 mixtures of the two-speaker, reverberant condition. See Table II for an enumeration of the parameters used in each one and their separation results. MESSL can be initialized from many different parameters, including masks, ILD, ITD, or any combination thereof. ITD can be estimated from a mixture using cross-correlation based methods like PHAT-histogram [26], and masks can be estimated using e.g. monaural signal-based source separators. ILD is more difficult to estimate directly from a mixture in practice. In the other experiments in this paper, we only initialize MESSL's ITD parameters from a non-cheating estimate, which appears on the second to last line of the table.

The top section of the table shows the performance of initializations that include ground truth interaural parameters in various combinations. From the top of the table down, these are: ground truth IPD and ILD information along with DP-Oracle masks, ground truth IPD and ILD information, only IPD information, and only ILD information. Initializations including ground truth ILD are run for as few iterations as possible, because on a separate parameter tuning data set their performance decreased with each iteration. This property indicates that it is the estimation of ILD parameters that is limiting the system's performance, not the separation based on an ILD estimate. This is not the case for ITD estimates, for which separation improves with each iteration as the parameters

TABLE II
SDR FOR MESSL WITH DIFFERENT INITIALIZATIONS AND NO GARBAGE SOURCE OR ILD PRIOR, SEPARATING TWO SPEAKERS IN REVERBERATION, AVERAGED OVER 15 MIXTURES AT EACH OF 6 ANGULAR SEPARATIONS. KEY: *1*: UNINFORMATIVE INIT., *gt*: GROUND TRUTH INIT., *xcorr*: ITD INIT. FROM CROSS-CORRELATION PEAKS.

ILD	IPD	Mask	Iterations	SDR (dB)
gt	gt	gt	2	6.40
gt	gt	1	2	6.29
1	gt	1	16	5.59
gt	xcorr	1	2	5.37
1	xcorr	gt	16	4.67
1	xcorr	1	16	4.43
1	1	gt	16	4.17

adapt to a particular mixture, even after initialization with ground truth ITD parameters. One possible explanation for this is that ILD is reduced by reverberation, while ITD is not [27].

The bottom section of the table shows the performance of initializations that do not include ground truth interaural parameters. From the top of this section, these are: estimated ITD with ground truth DP-Oracle masks, only estimated ITD, and only DP-Oracle masks. These systems were all run for 16 iterations because their performance tended to improve every iteration.

Initial model parameters were used in the first E step to calculate $\nu_{i\tau}(\omega, t)$. The algorithm then proceeded normally, discarding the initial parameters and replacing them with estimates made directly from the mixture and $\nu_{i\tau}(\omega, t)$. When an initial mask was supplied, however, it survived until the second E step. This is because in the first E step, it is used as a prior in estimating $\nu_{i\tau}(\omega, t)$, which also varies with τ, and only re-estimated after the first M step. Thus two iterations are required for a fair evaluation.

1) Results: The results of this experiment can be seen in Table II. Considering the top portion of the table, all initializations that use ground truth interaural parameters are better able to separate the sources than those estimating initial parameters from the mixture. IPD parameters seem to be slightly more helpful than ILD parameters, increasing SDR by 1.2 dB versus 0.9 dB when compared with the $\Theta_{\Omega\Omega}$ parameters, shown in the second to last row in the table. In combining the two cues, the performance gain is approximately additive, increasing SDR 1.9 dB beyond what is achieved with IPD alone. Including the ground truth mask only increases SDR by another 0.1 dB.

Considering the bottom portion of the table, initializing with just the ground truth mask separates sources more poorly than the baseline algorithm initialized from PHAT-histogram localization. When combined, however, the ground truth mask provides a slight improvement in separation. The difficulty in starting with just a mask is that the $\psi_{i\tau}$ extracted using the mask is peaked at the proper delay, but assigns significant probability to the other delays as well. It takes between 8 and 12 iterations before the values in $\psi_{i\tau}$ begin to resemble those coming from the PHAT-histogram initialization. When starting with $\psi_{i\tau}$ already reasonably initialized, the mask helps performance, possibly in estimating the ILD parameters.

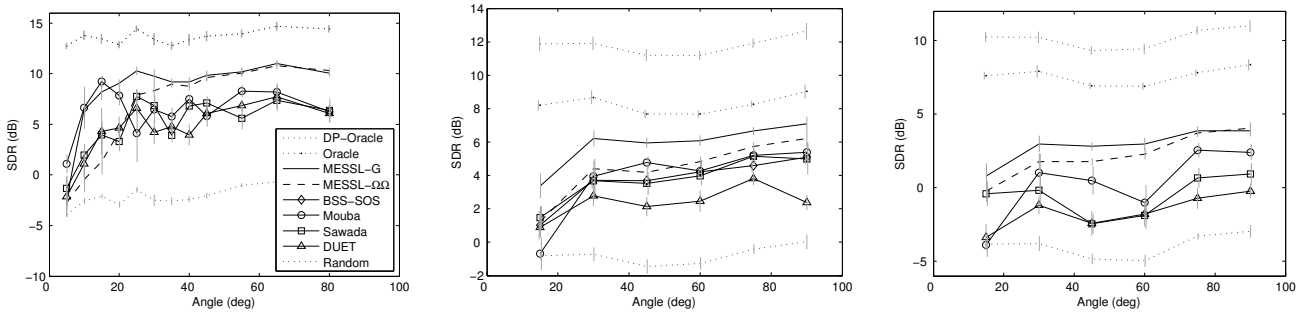
D. Comparison with other algorithms

The third experiment compares MESSL with four other well known source separation algorithms, DUET [6], the algorithm of Sawada et al. [22], which we refer to as Sawada, the algorithm of Mouba and Marchand [10], which we refer to as Mouba, and TRINICON-based blind source separation using second order statistics [13]. We implemented the first three of these algorithms ourselves and tested them on mixtures of two and three sources in reverberant and anechoic environments. TRINICON was run on our mixtures by the authors of that paper using their original code.

The Degenerate Unmixing Estimation Technique (DUET) [4], [6] creates a two-dimensional histogram of the interaural level and time differences observed over an entire spectrogram. It then smooths the histogram and finds the I largest peaks, which should correspond to the I sources. DUET assumes that the interaural level and time differences are constant at all frequencies and that there is no spatial aliasing, conditions that can be met to a large degree with free-standing microphones close to one another. With dummy head recordings, however, the ILD varies with frequency and the microphones are spaced far enough apart that there is spatial aliasing above about 1 kHz. Frequency-varying ILD scatters observations of the same source throughout the histogram as does spatial aliasing, making sources harder to localize and isolate. See Figure 5(e) for an example mask estimated by DUET.

Two-stage frequency-domain blind source separation [22] is a combination of ideas from model-based separation and independent component analysis (ICA) that can separate underdetermined mixtures. In the first stage, blind source separation is performed on each frequency band of a spectrogram separately using a probabilistic model of mixing coefficients. In the second stage, the sources in different bands are unpermuted using k-means clustering on the posterior probabilities of each source and then refined by matching sources in each band to those in nearby and harmonically related bands. The first stage encounters problems when a source is not present in every frequency and the second encounters problems if sources' activities are not similar enough across frequency. See Figure 5(d) for an example mask estimated by this algorithm.

The algorithm of Mouba and Marchand [10], like MESSL, uses EM clustering to separate sources from binaural recordings. This algorithm needs access to certain coefficients describing the relationship between ILD, ITD, and azimuth, which can be extracted offline from head-related transfer functions. It it not particularly sensitive to the exact values of these coefficients, however, so the same values generally work for different heads. Using these coefficients, the algorithm maps the ILD at each point in the spectrogram to an azimuth, with which it disambiguates each IPD-to-ITD mapping. The ITD is then mapped to azimuth at each spectrogram point and these azimuth values (after weighting by the energy at that spectrogram point) are clustered using a Gaussian mixture model. The means of the Gaussians are the estimated source locations and the posterior probability of each azimuth coming from each Gaussian is used to construct a spectral mask. See Figure 5(f) for an example mask estimated by Mouba and Marchand's algorithm.



(a) Anechoic, 3 speakers (b) Reverberant, 2 speakers (c) Reverberant, 3 speakers

Fig. 7. SDR as a function of angle for three conditions. Anechoic conditions are averaged over 10 mixtures at each of 12 angles, reverberant over 15 mixtures at each of 6 angles. Error bars show 1 standard error.

Blind Source Separation using Second Order Statistics (BSS-SOS) using Triple-N ICA for convolutive mixtures (TRINICON) [13] is very different from the other algorithms compared in this experiment. Instead of using time-frequency masking to separate sources, it learns a convolutive linear system to unmix the signals. The three Ns referred to are nongaussianity, nonstationarity, and nonwhiteness, properties of signals that TRINICON takes advantage of in separation. The unmixing system that it learns minimizes the correlation between the separated signals at multiple time lags. A number of algorithms in this framework have been derived, trading off computational complexity for separation performance. Note that this algorithm can only separate critically-determined and over-determined mixing systems, so we only evaluate it on two-source mixtures.

1) Results: The results of this experiment can be seen in Figure 7, where the SDR is shown as a function of separation angle for two and three sources in anechoic and reverberant environments. From the graphs, it is clear that the performance of all of the algorithms decreases as the sources get closer together and their spatial cues become more similar. This is not the case with the ground truth masks, suggesting that an algorithm that relied more on spectral cues (perhaps from source models) as opposed to spatial cues might be able to separate the two speakers equally well at any separation. Note that the DP-Oracle mask achieves much higher SDR and PESQ scores than the Oracle mask, due to its exclusion of all reverberation.

MESSL isolates the target source better than the other algorithms in all conditions and at all angles except for the smallest angles in anechoic conditions. Adding a second distractor source does not affect performance when the separation is large, but hurts performance for separations less than 40°. The garbage source and ILD prior increase performance quite a bit in reverberation, and when the sources are close together in anechoic mixtures.

In all conditions except A2, Mouba and Marchand's algorithm performs second best. Qualitatively, it appears to do a good job of resolving 2π ambiguities when the phase has wrapped one or two times. This is the case for most of the spectrum when azimuths are small and for lower frequencies when azimuths are larger. This frequency range generally includes most of the speech energy of interest, giving it good SDR performance and a significant advantage over DUET. At very small separations, it is difficult for the clustering to resolve two separate peaks in the azimuth histogram, so performance appears to be best for angles that are neither too big nor too small.

Generally, Sawada et al.'s algorithm performs third best. Visually inspecting some of the masks it creates, it generally works quite well, especially at high frequencies where adjacent frequency bands are highly correlated. It seems to have some difficulty unpermuting sources at low frequencies, however, where adjacent bands are less well-correlated. These problems are exacerbated in reverberation and with more sources. Comparing Figure 5(b) and (d), it is interesting to note that MESSL and this algorithm both exhibit "bands of ambiguity" in frequencies for which two sources' ITDs predict the same IPD (e.g. 1.5–2 kHz).

In anechoic conditions, TRINICON dramatically outperforms the masking-based systems in terms of SDR, even the DP-Oracle mask. In reverberation, however, it performs slightly worse than Sawada et al.'s algorithm. According to the signal-to-interferer ratio (SIR), which measures only the ratio of energy directly attributable to the target and interference (i.e. ignoring most reverberation), TRINICON outperforms MESSL even in reverberation, achieving 12.79 dB SIR on average in the two-source reverberant condition, compared to MESSL's 11.07 dB. We believe that the inversion of the ranking between SIR and SDR is due to the introduction of a substantial amount of reverberation by TRINICON (which is penalized under SDR but does not affect SIR). Another interesting performance characteristic is that in reverberation TRINICON achieves a much better SDR for the interfering source, 5.64 dB, than the target source, 3.72 dB. We believe that this is because the on-axis target source is easier to cancel than the off-axis interfering source. Under the PESQ metric, TRINICON's estimated speech quality is second only to MESSL's.

DUET performs relatively well in the anechoic case, but worse in reverberation. It is not affected much by the presence of a second distractor source. It performs much better in this experiment than in previous experiments we have run, possibly because the sources are more balanced here, while before they sometimes had different energies.

The PESQ results follow the SDR results quite closely. As would be expected, the PESQ scores for anechoic mixtures

TABLE III
SDR AND PESQ METRICS COMPARING SYSTEMS IN ANECHOIC AND REVERBERANT ENVIRONMENTS WITH 2 AND 3 SPEAKERS, E.G. "A2" INDICATES ANECHOIC, 2 SPEAKER CONDITION. ANECHOIC CONDITIONS AVERAGED OVER 120 MIXTURES EACH, REVERBERANT OVER 90 MIXTURES EACH.

SDR (dB)	A2	A3	R2	R3	Avg
DP-Oracle	16.63	13.63	11.79	10.15	13.05
Oracle	16.63	13.63	8.25	7.59	11.52
MESSL-G	11.91	**8.41**	**5.87**	**2.87**	**7.27**
MESSL-ΩΩ	11.29	6.47	4.45	2.19	6.10
Mouba	11.83	6.34	3.98	0.55	5.67
Sawada	11.44	4.98	3.80	−0.55	4.91
TRINICON	**22.47**	—	3.72	—	—
DUET	9.67	4.38	2.35	−1.42	3.75
Random	1.54	−2.13	−0.78	−3.95	−1.33

PESQ (MOS)	A2	A3	R2	R3	Avg
DP-Oracle	3.34	3.02	2.99	2.72	3.02
Oracle	3.34	3.01	2.35	2.24	2.73
MESSL-G	2.93	**2.29**	**2.07**	**1.73**	**2.26**
MESSL-ΩΩ	2.84	2.12	1.92	1.62	2.13
TRINICON	**3.37**	—	1.84	—	—
Mouba	2.69	1.96	1.81	1.49	1.99
Sawada	2.82	1.71	1.80	1.24	1.89
DUET	2.53	1.95	1.60	1.20	1.82
Mixture	1.81	1.35	1.45	1.15	1.44

are higher than for reverberant mixtures, and they are also higher for two sources than for three. The separations typically maintain the same ordering across conditions, except for TRINICON, which does better than DP-Oracle for anechoic, two-source mixtures. These orderings and this exception are consistent with the SDR results. Of the 1.58 MOS units between the average mixture score and the average DP-Oracle score, MESSL was able to recover approximately half, or 0.82 units.

VI. SUMMARY

This paper has presented a novel source separation procedure using probabilistic models of sources and an expectation maximization parameter estimation procedure. We first built a probabilistic model of a single source that can be evaluated independently at every spectrogram point. We then reduced the multi-source problem to a collection of single source problems by combining this model with the idea of probabilistic masking using an EM algorithm.

We then performed three experiments to test various aspects of the model. We found that the most complex models were the best at separating sources in our experimental conditions. We showed that even from a mixture, MESSL can estimate interaural parameters that are close to those measured in isolation and using these parameters can closely approximate its performance when initialized with "cheating" parameters. In reverberant and anechoic conditions with two and three speakers, MESSL produces SDRs 1.6 dB higher and PESQ results 0.27 MOS higher than comparable algorithms.

There are a number of directions to take this work in the future. The first is to build a more explicit model of reverberation than the garbage source to better distinguish it from direct-path sounds. We would also like to add a model of early echoes to the system so that they could aid in separation and be removed from the reconstructed sources. Other monaural cues could also be used for separation and combined probabilistically with the current binaural separation. Finally, to allow for a real-time implementation we would like to develop an online version of this algorithm. Such a system would propagate sources in time using a dynamics model while simultaneously creating separation masks one frame at a time.

ACKNOWLEDGMENT

The authors would like to thank Prof. W. Kellermann and A. Lombard for evaluating TRINICON on our data and Prof. S. Marchand for sharing his code with us. This work was supported by the Fu Foundation School of Engineering and Applied Science, and by the National Science Foundation (NSF) under Grant No. IIS-0535168. 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.

REFERENCES

[1] C. E. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] W. Koenig, "Subjective effects in binaural hearing," *J. Acoust. Soc. of America*, vol. 22, no. 1, pp. 61–62, 1950.

[3] R. J. Weiss, M. I. Mandel, and D. P. W. Ellis, "Source separation based on binaural cues and source model constraints," in *Proc. Interspeech*, September 2008, pp. 419–422.

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

[5] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[6] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Sig. Process.*, vol. 5, June 2000, pp. 2985–2988.

[7] N. Roman, D. Wang, and G. J. Brown, "A classification-based cocktail party processor," in *Adv. in Neural Info. Process. Systems*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004, pp. 1425–1432.

[8] H. Viste and G. Evangelista, "On the use of spatial cues to improve binaural source separation," in *Proc. Int. Conf. on Digital Audio Effects*, 2003, pp. 209–213.

[9] S. Harding, J. Barker, and G. J. Brown, "Mask estimation for missing data speech recognition based on statistics of binaural interaction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 58–67, 2006.

[10] J. Mouba and S. Marchand, "A source localization / separation / respatialization system based on unsupervised classification of interaural cues," in *Proc. Int. Conf. on Digital Audio Effects*, 2006, pp. 233–238.

[11] S. H. Colburn and A. Kulkarni, "Models of sound localization," in *Sound Source Localization*, A. N. Popper and R. R. Fay, Eds. New York: Springer, 2005, vol. 25, ch. 8, pp. 272–316.

[12] A. Hyvärinen, "Survey on independent component analysis," *Neural Computing Surveys*, vol. 2, no. 94-128, pp. 3–1, 1999.

[13] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 120–134, 2005.

[14] L. Jeffress, "A place theory of sound localization," *Journal of Comparitive and Physiological Psychology*, vol. 41, no. 1, pp. 35–39, 1948.

[15] N. I. Durlach, "Equalization and cancellation theory of binaural masking-level differences," *J. Acoust. Soc. of America*, vol. 35, no. 8, pp. 1206–1218, 1963.

[16] J. Breebaart, S. van de Par, and A. Kohlrausch, "Binaural processing model based on contralateral inhibition. I. model structure," *J. Acoust. Soc. of America*, vol. 110, no. 2, pp. 1074–1088, 2001.

[17] R. M. Stern, A. S. Zeiberg, and C. Trahiotis, "Lateralization of complex binaural stimuli: A weighted-image model," *J. Acoust. Soc. of America*, vol. 84, no. 1, pp. 156–165, 1988.

[18] R. M. Stern, G. J. Brown, and D. Wang, "Binaural sound localization," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. Wang and G. J. Brown, Eds. Wiley-IEEE Press, 2006, ch. 5, pp. 147–185.

[19] C. Liu, B. C. Wheeler, Jr, R. C. Bilger, C. R. Lansing, and A. S. Feng, "Localization of multiple sound sources with two microphones," *J. Acoust. Soc. of America*, vol. 108, no. 4, pp. 1888–1905, 2000.

[20] J. Nix and V. Hohmann, "Sound source localization in real sound fields based on empirical statistics of interaural parameters," *J. Acoust. Soc. of America*, vol. 119, no. 1, pp. 463–479, 2006.

[21] G. J. Brown, S. Harding, and J. P. Barker, "Speech separation based on the statistics of binaural auditory features," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Sig. Process.*, vol. 5, 2006, pp. V949–V952.

[22] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," in *IEEE Workshop on Appl. of Sig. Process. to Audio and Acoust.*, October 2007, pp. 139–142.

[23] M. I. Mandel, D. P. W. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *Adv. in Neural Info. Process. Systems*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007, pp. 953–960.

[24] M. I. Mandel and D. P. W. Ellis, "A probability model for interaural phase difference," in *ISCA Workshop on Statistical and Perceptual Audio Processing (SAPA)*, 2006, pp. 1–6.

[25] J. L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, 1994.

[26] P. Aarabi, "Self-localizing dynamic microphone arrays," *IEEE Trans. Syst., Man, Cybern. C*, vol. 32, no. 4, pp. 474–484, November 2002.

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

[28] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," 1993. [Online]. Available: http://www.ldc.upenn.edu/Catalog/LDC93S1.html

[29] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *IEEE Workshop on Appl. of Sig. Process. to Audio and Acoust.*, Oct 2001, pp. 99–102.

[30] P. C. Loizou, *Speech Enhancement: Theory and Practice*. CRC, 2007.

[31] L. Di Persia, D. Milone, H. Rufiner, and M. Yanagida, "Perceptual evaluation of blind source separation for robust speech recognition," *Signal Processing*, vol. 88, no. 10, pp. 2578–2583, October 2008.

[32] M. I. Mandel and D. P. W. Ellis, "EM localization and separation using interaural level and phase cues," in *IEEE Workshop on Appl. of Sig. Process. to Audio and Acoust.*, October 2007, pp. 275–278.

Michael I. Mandel received the B.S. degree in Computer Science from the Massachusetts Institute of Technology, Cambridge, MA, in 2004 and the M.S. degree in Electrical Engineering from Columbia University, New York, NY, in 2006 where he is currently pursuing the Ph.D. degree in Prof. D. Ellis' Laboratory for the Recognition and Organization of Speech and Audio.

He has published on sound source separation, music similarity, and music recommendation. His research uses machine learning to model sound perception and understanding.

Ron J. Weiss received the B.S. degree in computer engineering in 2004, the M.S. and Ph.D. degrees in electrical engineering in 2005 and 2009, respectively, all from Columbia University, New York, NY.

He is currently a postdoctoral researcher with the Music and Audio Research Laboratory at New York University. His research interests lie at the intersection between audio signal processing and machine learning, focusing on sound source separation and music information retrieval.

Daniel P. W. Ellis received the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge.

He is an Associate Professor in the Electrical Engineering Department, Columbia University, New York. His Laboratory for Recognition and Organization of Speech and Audio (LabROSA) is concerned with all aspects of extracting high-level information from audio, including speech recognition, music description, and environmental sound processing. He also runs the AUDITORY e-mail list of 1700 worldwide researchers in perception and cognition of sound. He worked at MIT, where he was a Research Assistant at the Media Lab, and he spent several years as a Research Scientist at the International Computer Science Institute, Berkeley, CA.