Chapter 5

Separation

This chapter describes the full Model-based EM Source Separation and Localization (MESSL) system. This system separates and localizes multiple sound sources from a reverberant two-channel recording. It builds on chapter 3's modeling of the interaural parameters of sources and chapter 4's flexible, probabilistic localization model.

Once a model is constructed of a single source's interaural parameters, multiple models can be combined into a mixture over sources and delays. This mixture model reduces the multi-source localization problem to a collection of single source localization problems, with straightforward solutions. It is described in section 5.2. Section 5.2.1 then derives an expectation maximization (EM) algorithm to estimate the maximum-likelihood parameters of this mixture model, and shows that these parameters correspond well with interaural parameters measured in isolation in section 5.3. As a byproduct of fitting this model, the algorithm creates probabilistic spectrogram masks that can be used for source separation.

Fundamentally, MESSL clusters spectrogram points into sources by their interaural phase and level differences. Each source is represented by its centroid in this interaural parametrization. Using these source models, points from each source can be separated from one another; they are assigned to each source probabilistically, according to their similarity to the source's centroid. Using these separations, a more accurate localization can be made; the centroid of each source is re-estimated from the points that are assigned to it. These two steps alternate for a fixed number of iterations or until the separation and localization converge.

This chapter concludes with an extensive set of experiments exploring different aspects of MESSL and comparing it with other separation and localization systems in section 5.4. These experiments show that MESSL's separations

- are best when using more complicated parametrizations, without over-fitting the data
- are slightly better when initialized with ground truth parameters instead of estimating them from a mixture

- have signal-to-distortion ratios (SDRs) 1.4 dB higher when including a model of reverberation
- have SDRs 2.0 dB higher and PESQ scores 0.19 MOS higher than four comparable separation algorithms

They also show that MESSL's localization has a mean absolute error 40% smaller than four comparable localization algorithms. We believe that these performance benefits are due to the use of a model that is structured enough to extract useful information from observations, but flexible enough to accept the variability introduced by interference and reverberation.

This chapter is based on (Mandel et al., 2007; Mandel and Ellis, 2007; Weiss et al., 2008; Mandel et al., 2009). Mandel et al. (2007) first introduced the concept of localization and separation using this mixture model. Mandel and Ellis (2007) added the psychoacoustically relevant interaural level difference cue. Weiss et al. (2008) added source models, which are described in section 5.2.3 and take advantage of known regularities of speech in separating speakers. Mandel et al. (2009) reported the results of many of the experiments included in this chapter.

5.1 Background

MESSL is a synthesis of localization-based clustering and spectral masking. A number of systems have taken a similar approach (Jourjine et al., 2000; Roman et al., 2003; Viste and Evangelista, 2003; Harding et al., 2006; Mouba and Marchand, 2006). Localization in azimuth is a popular cue for segregating sound sources, as discussed in section 2.3.3 and section 2.5. 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 (Hyvärinen, 1999) and other multi-microphone techniques (Buchner et al., 2005), which can separate at most as many sources as there are microphones.

As described in section 2.3.3, many models of mammalian auditory localization have been described in the literature. Most focus on localization within individual critical bands of the auditory system, whereas 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 (Stern et al., 1988, 2006) and simultaneous sources can be localized using a "stencil" filter that embraces multimodality (Liu et al., 2000). Other systems localize the source in each band probabilistically and then combine probabilities across frequency by assuming statistical independence. Nonparametric modeling in this

vein (Nix and Hohmann, 2006; Harding et al., 2006; Brown et al., 2006) 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 Harding et al. (2006); Brown et al. (2006) collect histograms of per-band interaural time differences, Nix and Hohmann (2006) collect 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 (Jourjine et al., 2000; Yilmaz and Rickard, 2004) 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 of Roman et al. (2003) 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 Viste and Evangelista (2003); Mouba and Marchand (2006) 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.

Given the localization of each time-frequency point, most algorithms perform separation through some kind of clustering (Mouba and Marchand, 2006) or histogram peak picking (Jourjine et al., 2000; Viste and Evangelista, 2003), although when the locations are known in advance, classification is also a popular technique (Roman et al., 2003; Harding et al., 2006; Brown et al., 2006). 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 (Mouba and Marchand, 2006).

A separation system that shares many similarities with these systems, but does not explicitly use localization is two-source, frequency-domain, blind source separation (Sawada et al., 2007). 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.

5.2 A localization mixture model

For the purposes of deriving this model we will first 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.

As in chapters 3 and 4, denote the sound source as $s(t)$ and the signals received at the left and right ears as $\ell(t)$ and $r(t)$, respectively. In this chapter, we derive our results for the case where the interaural transfer function is frequency-dependent, as discussed in chapter 3, but this subsumes the frequency-independent case discussed in chapter 4. The two received signals relate to the source by a frequency-dependent delay and gain, in addition to a disruption due to noise. In the experiments in section 5.4.2, we compare models assuming frequency-independence to those assuming frequency-dependence.

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:

$$\ell(t) = s(t - \tau_\ell) * h_\ell(t) * n_\ell(t) \qquad r(t) = s(t - \tau_r) * h_r(t) * n_r(t). \tag{5.1}$$

The ratio of the short-time Fourier transforms of both equations is the interaural spectrogram,

$$\frac{L(\omega, t)}{R(\omega, t)} = 10^{\alpha(\omega,t)/20} e^{j\phi(\omega,t)} \tag{5.2}$$

$$\approx 10^{a(\omega)/20} e^{-j\omega\tau(\omega)} N(\omega, t) \tag{5.3}$$

Equation (5.2) is the ratio of the actual observations at both ears, while equation (5.3) is our model of that ratio.

The interaural spectrogram is parametrized 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, $a(\omega)$. As discussed in chapter 4, we use a top-down approach to localization that mitigates the problem of spatial aliasing.

To measure the difference between the IPD predicted by a delay of τ samples and the observed IPD, we use the phase residual defined in equation (4.24). The residual error can be modeled with a circular probability distribution like the von Mises distribution or higher order maximum entropy distributions, as described in chapter 4, or it can be approximated as a linear distribution using a Gaussian scale mixture model, a Gaussian mixture model in which all of the Gaussians share the same mean (Mandel et al., 2007). We have found, however, that for separation purposes a single Gaussian works well enough

$$p(\phi(\omega, t) \,|\, \tau(\omega), \sigma(\omega)) = \mathcal{N}\big(\hat{\phi}(\omega, t; \tau(\omega)) \,|\, 0, \sigma^2(\omega)\big) \tag{5.4}$$

$$\approx \mathcal{N}\big(\phi(\omega, t) \,|\, \omega\tau(\omega), \sigma^2(\omega)\big). \tag{5.5}$$

The equality is approximate because it only holds when the standard deviation σ is small relative to 2π, in which case the Gaussian and von Mises distributions are very similar.

In accordance with the observations in chapter 3, we model the interaural level difference, measured in dB, by a single Gaussian with frequency-dependent mean and variance

$$p(\alpha(\omega,t)\,|\,\mu(\omega),\eta(\omega)) = \mathcal{N}\big(\alpha(\omega,t)\,|\,\mu(\omega),\eta^2(\omega)\big). \tag{5.6}$$

Because the ILD is modeled with a Gaussian, we use the normal-Wishart distribution, its conjugate, as the prior on its parameters (Gauvain and Lee, 1994). This prior, for diagonal covariance Gaussians, is

$$p(\mu(\omega),\eta(\omega)\,|\,v,m(\omega),c(\omega),d(\omega)) \propto \exp\left(-v\sum_\omega \frac{(\mu(\omega)-m(\omega))^2}{2\eta^2(\omega)}\right)\prod_\omega \eta(\omega)^{\Omega-c(\omega)}\exp\left(-\sum_\omega \frac{d(\omega)}{2\eta^2(\omega)}\right) \tag{5.7}$$

The mean of this normal-Wishart, $m(\omega)$, is learned from impulse responses, as described in section 5.2.2, but the system is not especially sensitive to particular values. In our model, we set $c(\omega)$ and $d(\omega)$ to uninformative values, so they drop out and the prior becomes

$$p(\mu(\omega),\eta(\omega)\,|\,v,m(\omega)) \propto \exp\left(-v\sum_\omega \frac{(\mu(\omega)-m(\omega))^2}{2\eta^2(\omega)}\right). \tag{5.8}$$

In effect, the prior acts as a number of "virtual" observations that are included in the parameter estimates discussed later. The prior precision (inverse variance), u, controls the strength of the prior relative to the observations, i.e. the number of virtual observations.

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

$$p(\phi(\omega,t),\alpha(\omega,t)\,|\,\Theta) = \mathcal{N}\big(\hat{\phi}(\omega,t)\,|\,\xi(\omega),\sigma^2(\omega)\big)\cdot\mathcal{N}\big(\alpha(\omega,t)\,|\,\mu(\omega),\eta^2(\omega)\big), \tag{5.9}$$

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.

Equation 5.9 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. This is a generalization of the localization system of the previous chapter to include the interaural level difference.

5.2.1 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 maximum-likelihood model parameters can be estimated with an expectation maximization (EM) algorithm. For each source in a multi-source 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,

$$\tau(\omega) = \tau + \frac{\xi(\omega)}{\omega}. \tag{5.10}$$

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 equation (5.4). 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 5.4.2 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

$$\Theta \equiv \{\xi_{i\tau}(\omega), \sigma_{i\tau}(\omega), \mu_i(\omega), \eta_i(\omega), \psi_{i\tau}\} \tag{5.11}$$

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

$$\mathcal{L}(\Theta) = \sum_{\omega,t} \log p(\phi(\omega,t), \alpha(\omega,t) \mid \Theta) \tag{5.12}$$

$$= \sum_{\omega,t} \log \sum_{i,\tau} \Big[p(\phi(\omega,t), \alpha(\omega,t) \mid z_{i\tau}(\omega,t), \Theta) \cdot \; p(z_{i\tau}(\omega,t) \mid \Theta) \Big] \tag{5.13}$$

$$= \sum_{\omega,t} \log \sum_{i,\tau} \Big[\mathcal{N}\big(\hat{\phi}(\omega,t;\tau) \mid \xi_{i\tau}(\omega), \sigma^2_{i\tau}(\omega)\big) \cdot \mathcal{N}\big(\alpha(\omega,t) \mid \mu_i(\omega), \eta_i^2(\omega)\big) \cdot \psi_{i\tau} \Big]. \tag{5.14}$$

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 and the prior over ILD parameters, we define the auxiliary function to maximize with respect to Θ,

$$Q(\Theta \mid \Theta_s) = \sum_{\omega,t} \sum_{i,\tau} p\big(z_{i\tau}(\omega,t) \mid \phi(\omega,t), \alpha(\omega,t), \Theta_s\big) \cdot \log p\big(z_{i\tau}(\omega,t), \phi(\omega,t), \alpha(\omega,t) \mid \Theta\big) + \sum_i \log p(\mu_i(\omega), \eta_i(\omega) \mid v_i, m_i(\omega)) + k \tag{5.15}$$

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

$$\nu_{i\tau}(\omega,t) \equiv p(z_{i\tau}(\omega,t) \mid \phi(\omega,t), \alpha(\omega,t), \Theta_s) \tag{5.16}$$

$$\propto p(z_{i\tau}(\omega,t), \phi(\omega,t), \alpha(\omega,t) \mid \Theta_s) \tag{5.17}$$

$$= \psi_{i\tau} \cdot \mathcal{N}\big(\hat{\phi}(\omega,t;\tau) \mid \xi_{i\tau}(\omega), \sigma^2_{i\tau}(\omega)\big) \cdot \mathcal{N}\big(\alpha(\omega,t) \mid \mu_i(\omega), \eta_i^2(\omega)\big). \tag{5.18}$$

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

$$\langle x \rangle_{t,\tau} \equiv \frac{\sum_{t,\tau} x\, \nu_{i\tau}(\omega,t)}{\sum_{t,\tau} \nu_{i\tau}(\omega,t)} \tag{5.19}$$

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

$$\mu_i(\omega) = \langle \alpha(\omega, t) \rangle_{t,\tau} \tag{5.20}$$

$$\eta_i^2(\omega) = \left\langle \left(\alpha(\omega, t) - \mu_i(\omega) \right)^2 \right\rangle_{t,\tau} \tag{5.21}$$

$$\xi_{i\tau}(\omega) = \left\langle \hat{\phi}(\omega, t; \tau) \right\rangle_t \tag{5.22}$$

$$\sigma_{i\tau}^2(\omega) = \left\langle \left(\hat{\phi}(\omega, t; \tau) - \xi_{i\tau}(\omega) \right)^2 \right\rangle_t \tag{5.23}$$

$$\psi_{i\tau} = \frac{1}{\Omega T} \sum_{\omega,t} \nu_{i\tau}(\omega, t). \tag{5.24}$$

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

$$\sigma_{i\tau}^2 = \left\langle \left(\hat{\phi}(\omega, t; \tau) - \xi_{i\tau}(\omega) \right)^2 \right\rangle_{\omega,t} \tag{5.25}$$

$$\sigma_i^2 = \left\langle \left(\hat{\phi}(\omega, t; \tau) - \xi_{i\tau}(\omega) \right)^2 \right\rangle_{\omega,t,\tau} \tag{5.26}$$

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

$$M_i(\omega, t) \equiv \sum_{\tau} \nu_{i\tau}(\omega, t). \tag{5.27}$$

Note that this mask represents probabilities, see section 5.2.4 for an experiment in converting these probabilities to more Wiener filter-like coefficients.

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 5.4.3 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 (Aarabi, 2002), $\psi_{i\tau}$ 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.

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

5.2.2 Modeling reverberation

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.

As discussed in section 3.1.3.2, reverberation tends to diminish ILD observations. This is because of the isotropic nature of reverberation, which leads to an ILD of 0 dB for purely reverberant signals. The ILD prior introduced in equation (5.8) can be used to help the ILD estimates converge to the ILD of the direct-path portion of the signal, pushing more reverberation into the garbage source, and can help distinguish between reverberation and a direct-path source with an ILD of 0 dB.

When the ILD prior is included in the estimation, equations (5.20) and (5.21) become, respectively

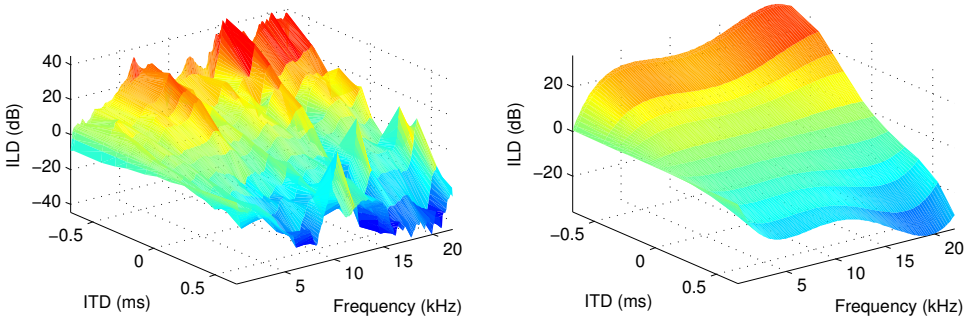
$$\mu_i(\omega) = \frac{v_i m_i(\omega) + \sum_{t,\tau} \alpha(\omega,t)\nu_{i\tau}(\omega,t)}{v_i + \sum_{t,\tau} \nu_{i\tau}(\omega,t)} \tag{5.28}$$

$$\eta_i^2(\omega) = \frac{v_i\big(m_i(\omega) - \mu_i(\omega)\big)^2 + \sum_{t,\tau}\big(\alpha(\omega,t) - \mu_i(\omega)\big)^2 \nu_{i\tau}(\omega,t)}{\sum_{t,\tau} \nu_{i\tau}(\omega,t)}. \tag{5.29}$$

$$\tag{5.30}$$

Note the source-dependence of the ILD prior parameters. In practice, the prior precision, v_i is the same for all sources, but the prior mean, $\mu_i(\omega)$, is different for each source, set from the ITD initialization. The exact value of the prior mean was estimated from a set of anechoic head-related transfer functions using a regression on ITD, 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 facilitates generalization across individuals. Figure 5.1 shows both the prior as a function of ITD and frequency and the original data from which it was learned. We only employ the ILD prior in combination with the garbage source, as a pilot study found that that was when it was most useful.

Note that the ILD prior does make assumptions about the relationship between ILD and ITD. These assumptions concede generality for performance in a particular condition of interest. Specifically, the modeled relationship between ILD and ITD assumes a particular separation between the microphones or ears, and the fact that there is an object of a particular size and shape between the two microphones. These assumptions make sense



(a) Actual ILD-ITD relationship (b) Approximate

Figure 5.1: The ILD prior captures the dependence of ILD on ITD. (a) The data that the ILD prior is learned from: anechoic HRTFs, (b) The learned prior. Note that unlike other plots, the maximum frequency here is 22 kHz.

for dummy head recordings and even for in-ear recordings of many individuals, because of the smoothness of the relationship and the weakness of the prior. They would need to be modified for free-field recordings, however, which typically include a much smaller ILD, as there is no frequency-dependent absorber shielding the far-side microphone.

5.2.3 Source modeling

One of the advantages of the MESSL system is that it is agnostic to the source signals in a mixture. This gives it the power to separate any sources, as long as they have been mixed under realistic interaural conditions. This generality comes at a cost, however. Just as assuming a specific relationship between the ILD and ITD can increase separation performance in the particular situations that we are interested in while forfeiting some amount of generality, so too can making certain assumptions about source signals.

Specifically, as described by Weiss et al. (2008), when separating speech sources, MESSL can be extended with a model of these sources. Speakers are modeled in the STFT domain using relatively simple Gaussian mixtures, typically having 32 or 64 components. Such models include many phonetic characteristics, but are generally insufficient for speech recognition. Even so, they capture enough information about the spectral shapes that appear typically in speech to improve source separation performance.

There are two related speech models that have been incorporated into MESSL. The first is speaker-independent, and it is referred to as MESSL-SP, for Source Prior. The second is speaker-dependent, with the speaker adaptation taking place in the eigenvoice space (Kuhn et al., 1998), and it is referred to as MESSL-EV, for Eigenvoice. The experiments in section 5.4.4 compare the performance of MESSL-SP to other parametrizations of MESSL. For experiments with MESSL-EV, see (Weiss, 2009, Chap. 5).

In addition to modeling the ways that each speaker's utterance changes over time, we also model the static acoustic channel through which they are speaking. While the main MESSL

system is built on the difference between the left and right channels, to explicitly model both channels requires their sum as well. In practice we use two parallel channel models, the standard MESSL model of their difference, and a source prior model of both their difference and sum. We are in some sense over-counting the evidence from the channel differences, although we have not noticed any adverse effects from this.

Mathematically, then, we model the signals at the left and right ears as

$$L(\omega,t) = S_i(\omega,t) + h_i^\ell(\omega) \qquad R(\omega,t) = S_i(\omega,t) + h_i^r(\omega) \tag{5.31}$$

where $L(\omega,t)$, $R(\omega,t)$, and $S_i(\omega,t)$ are all measured in dB, $S_i(\omega,t)$ is the STFT of source i, and $h_i^\ell(\omega)$ and $h_i^r(\omega)$ are the channel responses for source i at the left and right ears. Using the speech models, source i is modeled as

$$p(S_i(\omega,t)) = \sum_c \pi_{ic}\mathcal{N}(S_i(\omega,t) \mid \mu_{ic}(\omega), \sigma_{ic}(\omega)) \tag{5.32}$$

where $S_i(\omega,t)$ is the energy at frequency ω and time t from source i, and c indexes the GMM component in the source model. In the case of MESSL-SP, μ_{ic} and σ_{ic} do not depend on i, but in the case of MESSL-EV, μ_{ic} and σ_{ic} are functions of the eigenvoice parameters.

This second channel model is necessary, however, to avoid source permutation problems. For the speaker-independent model and the initial stages of the eigenvoice model, the two speaker models are identical. This symmetry makes the models uninformative in separating sources, and it must be broken so that the source priors can contribute to the separation. We introduce the channel model to break this symmetry, but it also can be broken by enforcing temporal constraints through e.g. a hidden Markov model. Temporal constraints, however, greatly increase the model complexity.

The channel is modeled using low-order discrete cosine transform (DCT) terms, $h_i = B\mathbf{h}_i$. The DCT matrix that we are using, B, only includes the lowest L frequencies from the complete DCT matrix, and is defined as

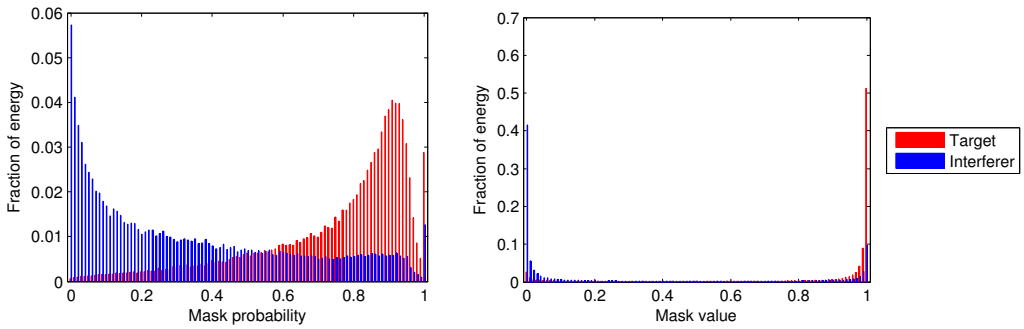
$$B_{\omega j} = \frac{2}{\sqrt{\Omega}} \cos\left(\frac{\pi}{\Omega}(\omega + \tfrac{1}{2})j\right) \quad i = 0 \ldots \Omega - 1, \quad j = 0 \ldots L - 1. \tag{5.33}$$

The full model of each observed channel given that it comes from a particular source and GMM component is thus

$$p(L(\omega,t) \mid i, c, \Theta) = \mathcal{N}(L(\omega,t) \mid \mu_{ic}(\omega) + B_{\omega\cdot}\mathbf{h}_i^\ell, \sigma_{ic}^2(\omega)) \tag{5.34}$$

$$p(R(\omega,t) \mid i, c, \Theta) = \mathcal{N}(R(\omega,t) \mid \mu_{ic}(\omega) + B_{\omega\cdot}\mathbf{h}_i^r, \sigma_{ic}^2(\omega)) \tag{5.35}$$

where $B_{\omega\cdot}$ is the row of the B matrix corresponding to frequency ω, and $\mu_{ic} = \mu_c(\mathbf{w}_i)$ and $\sigma_{ic}^2 = \sigma_c^2(\mathbf{w}_i)$ according to the eigenvoice parametrization.



(a) (b)

Figure 5.2: Histograms of MESSL mask values weighted by target and masker energies (a) before and (b) after warping through sigmoid function with $\alpha = 0.5$, $\beta = 14.4$.

The total likelihood, then, in schematic form, including the hidden variables, i, τ, and c, and the observations of the interaural parameters and the power spectra at the two ears is

$$p(\phi, \alpha, L, R \mid \Theta) = \sum_{i\tau c} p(\phi, \alpha, L, R, i, \tau, c \mid \Theta) \tag{5.36}$$

$$= \sum_{i\tau c} p(\phi \mid i, \tau, \Theta)\, p(\alpha \mid i, \Theta)\, p(L \mid i, c, \Theta)\, p(R \mid i, c, \Theta)\, p(c \mid i)\, p(i, \tau) \tag{5.37}$$

where (ω, t) has been omitted for brevity from the observations ϕ, α, L, and R. To convert this into a total log likelihood equation like equation (5.14), a dependence on the hidden variable c must be added to the indicator variable z, which becomes $z_{i\tau c}(\omega, t)$.

The E step of the algorithm is now analogous to the original MESSL E step, but with this new z variable. When updating the ILD and IPD parameters, $z_{i\tau c}(\omega, t)$ and $v_{i\tau c}(\omega, t)$ can be averaged over their c dependence, yielding the original $z_{i\tau}(\omega, t)$ and $v_{i\tau}(\omega, t)$. This is possible because the arguments to the weighted sums in equations (5.20) to (5.24) do not depend on c. The M step of the algorithm is unchanged except for the addition of independent estimates for the μ_{ic}, σ_{ic}, and $\mathbf{h}_i^{\ell,r}$ variables. These updates are beyond the scope of this section, but are derived by Weiss (2009, Appendix 3.4).

Two other systems combining localization and speech models have been described recently (Wilson, 2007; Rennie et al., 2003, 2007). These systems' main disadvantage is that the computational complexity of their factorial GMMs is exponential in the number of sources, whereas MESSL's is linear in the number of sources.

5.2.4 Converting probabilities to masks

The masks estimated in equation (5.27) represent the probability of each time-frequency point's coming from a given source. The optimal separation filter in the minimum mean squared error (MMSE) sense, however, is the Wiener filter, which depends on both the

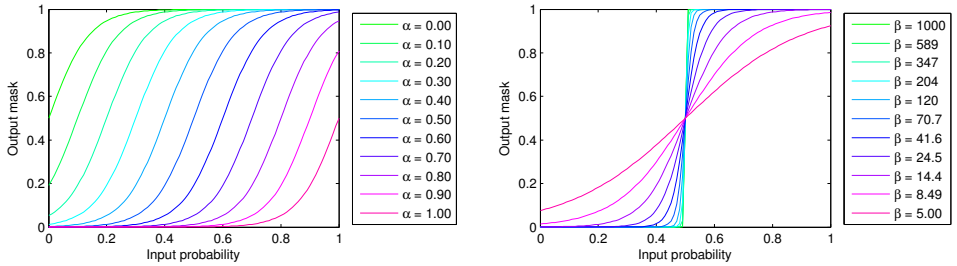


Figure 5.3: Sigmoid functions of equation (5.38) with different parameters. (a) Varying the α parameter, (b) varying the β parameter.

power in the signal and the power in the noise as shown in equation (3.8). In essence, the optimal MMSE mask is not the probability that a given point originated from the target source, but the fraction of the energy at that point that originated from the target source.

Typically, for speech mixtures this Wiener mask is very close to a binary mask because of the W-disjoint orthogonality property of speech (Yilmaz and Rickard, 2004). A single speaker contributes almost all of the energy at almost every time-frequency point. This constraint especially holds for the highest energy points, as seen in figure 3.8. See figure 5.2 for histograms of the different probabilities that MESSL assigns to target and interference energy. This figure shows the fraction of target and masker energy that is present at different mask values, averaged over all of the two-source reverberant separations.

Because MESSL does not model this energy ratio explicitly, we can attempt to convert its probabilistic masks into Wiener masks with a post-processing function. Here were evaluate functions of single time-frequency points that map probabilities into masks. We further constrain these functions to be the same for all ω and t. And even more specifically, we consider the parametric family of sigmoid functions for this warping

$$f(p) = \frac{1}{1 + \exp(-\beta(p - \alpha))} \tag{5.38}$$

where p is the probability estimated by MESSL, and α and β are the sigmoid's parameters that we adjust to optimize our separation metric. See figure 5.3 for examples of a number of sigmoid functions with different parameter settings.

In order to compare the merits of different parameter settings, we perform a grid search over α and β. This experiment uses artificially generated impulse responses from an image model (Allen and Berkley, 1979) room simulator using the CIPIC anechoic HRTFs. These impulse responses were computed for a source 1 meter from the listener every 15° from $-90°$ to 90°, with five repetitions at each location. MESSL was run over these 480 two-source, reverberant mixtures using the frequency-dependent parametrization with the garbage source and ILD prior. The masks from these separations were then run through all of the warpings under investigation and the signal-to-noise ratio improvement (discussed further in chapter 6) of each one was measured. These SNRIs were then averaged together

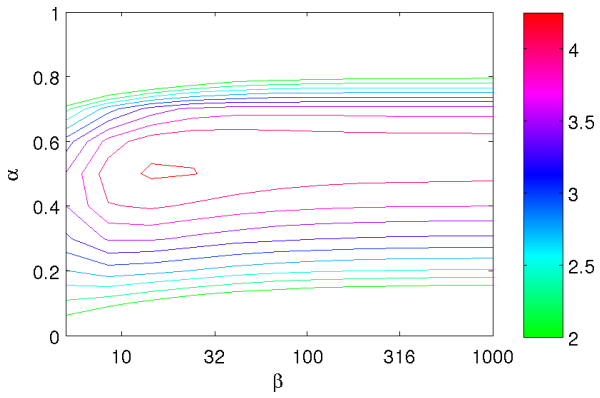


Figure 5.4: Contour plot of signal-to-noise ratio improvement (in dB) as a function of α and β parameters in the sigmoidal nonlinearity. Aggregated over the 480 mixtures in a two-source reverberant condition. The maximum is 4.29 dB at $\alpha = 0.5$, $\beta = 14.4$.

across all of the mixtures and the aggregate two-dimensional SNRI surface is shown in figure 5.4.

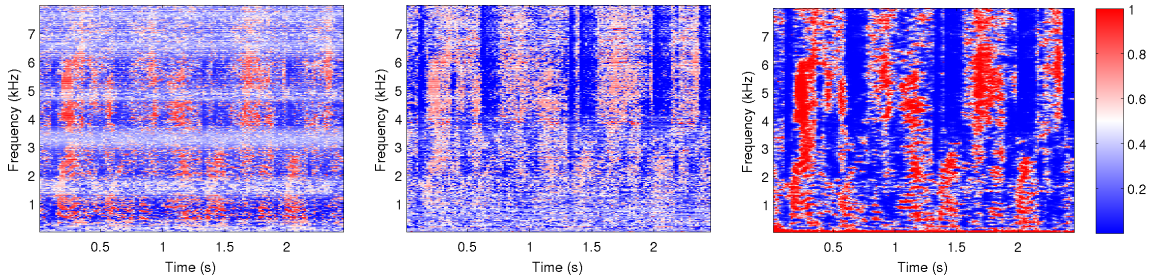
From this figure, it can be seen that the optimal setting, in terms of SNRI, is $\alpha = 0.5$ and $\beta = 14.4$. More interestingly, however, is to examine the way that the surface changes around this peak. The SNRI falls off for smaller values of β, i.e. shallower sigmoid slopes. For very large β, the sigmoid becomes a step function which binarizes the mask and SNRI falls off slightly. This indicate that the binarized version of the mask yields a slightly lower SNRI than the soft mask. In terms of α, the best SNRI comes from slightly larger values, although performance also has a steeper roll-off as α increases than as it decreases.

5.3 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 5.4.5 without any modification and is used in figures 5.10 and 5.5 to 5.7. Sound files from this example are also available on the project's webpage[1]. Two MESSL parametrizations were used in this example, one frequency-dependent and the other frequency-independent. Both use the garbage source and ILD prior.

Figure 5.5 shows the contributions of the IPD and ILD to the full MESSL mask when using the frequency-dependent parametrization. Note that as in human hearing, the IPD mask is

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



(a) IPD (b) ILD (c) Combined

Figure 5.5: Contribution of IPD and ILD to MESSL mask using frequency-dependent parameters, garbage source, and probability-to-mask warping. 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 5.6 and 5.7 show the ground truth and estimated parameters for ILD and IPD, respectively.

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 4.8 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.

The parameters estimated by MESSL are not only effective at separating sources, but are also interesting in and of themselves. Figure 5.6 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 (Shinn-Cunningham et al., 2005). 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 5.7 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

$$p(\phi(\omega,t)\,|\,i) = \sum_{\tau} \psi_{i\tau}\mathcal{N}\big(\hat{\phi}(\omega,t;\tau)\,|\,\xi_{i\tau}(\omega),\sigma^2_{i\tau}(\omega)\big). \tag{5.39}$$

Each component in this mixture is a line with a different slope, as illustrated in figure 4.8, although for figure 5.7 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

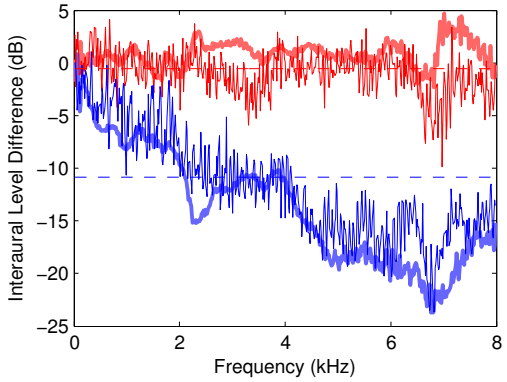


Figure 5.6: Example ILD for two sources at 0° (upper lines) and 75° (lower lines) in reverberation. Thick lines are ground truth direct-path ILD measured in isolation, thin lines are estimated from a mixture by MESSL using frequency-dependent Θ_G parameters, and the dashed lines are estimated by MESSL using frequency-independent Θ_{11} parameters.

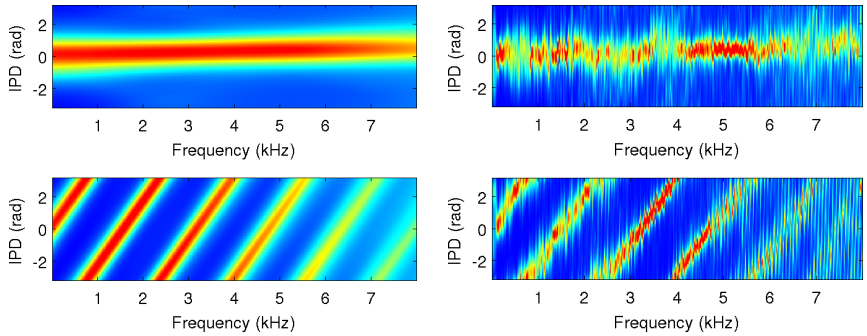
components are very close together and so the variance of the mixture is lower. At higher frequencies, however, the various mixture components are spaced farther apart due to their proportionality to ω, and so the model is less informative.

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, equation (5.39), 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 5.7(a). 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.

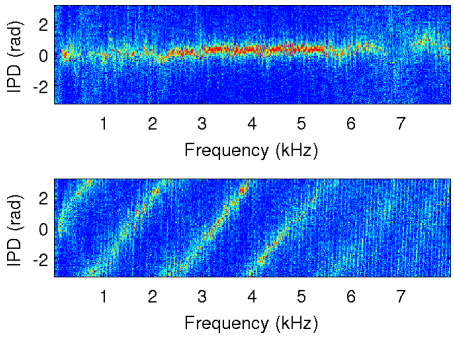
Figure 5.7(b) shows the extra information captured by the frequency-dependent parameters. This extra information mostly takes the form of rapid fluctuations of the mean and variance of the IPD with frequency, particularly at high frequencies. It is not clear to us what these represent, possibly aspects of the room impulse response like early echoes. Further investigation is warranted.

5.4 Speech separation experiments

We perform five 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



(a) Frequency-independent (b) Frequency-dependent



(c) Pre-mixing histogram

Figure 5.7: Interaural phase differences (IPD) as a function of frequency for two speakers at 0° and 75° in reverberation. Red is higher probability. (a) PDF of each source's IPD estimated from the mixture by MESSL using frequency-independent parameters, Θ_{11}, (b) PDF of each source's IPD estimated from the mixture by MESSL using frequency-dependent parameters, $\Theta_{\Omega\Omega}$, (c) Histogram of each source's IPD before mixing.

side. These stereo mixtures are given to 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.

5.4.1 Shared experimental details

A number of experimental details are shared across all of the experiments in this chapter. These include the data sources used, the evaluation metric, and the control "algorithms" against which the separations are performed. Note that the data sources are also the same as in the localization experiments in chapter 4.

5.4.1.1 Data sources

The utterances used are from the TIMIT acoustic-phonetic continuous speech corpus (Garofolo et al., 1993), 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 (Algazi et al., 2001b), 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 (Shinn-Cunningham et al., 2005). 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.

5.4.1.2 Evaluation metrics

We evaluate separation performance using the signal-to-distortion ratio (SDR) (Vincent et al., 2006). Specifically, this is the function `bss_eval_sources.m` used in the Signal Separation Evaluation Campaign (SiSEC 2008) (Vincent et al., 2009). 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 masker signals counts towards masker 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-interference 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 (PESQ) (Loizou, 2007, 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 (Di Persia et al., 2008). The PESQ results for each algorithm in table 5.5 should be compared to those of the unseparated mixtures in the last row of that table.

5.4.1.3 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 point in which the target signal is at least as loud as the combined maskers and 0 everywhere else (Yilmaz and Rickard, 2004). 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.

5.4.2 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 equations (5.25) and (5.26).

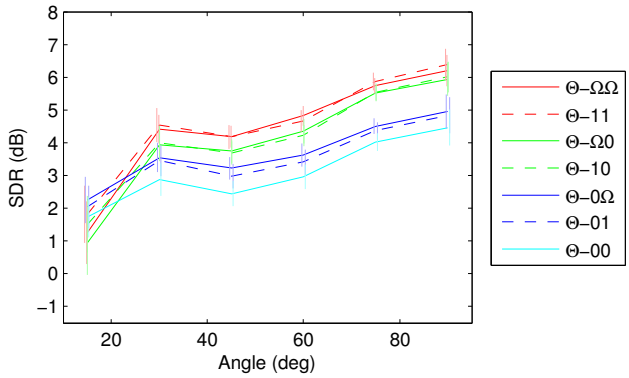


Figure 5.8: Signal-to-distortion ratio for complexity experiment, two speakers in reverberation. Each point is the average of 15 mixtures with the same angular separation. Error bars show 1 standard error. Note that the random and DP-Oracle masks bound performance between 11.79 and −0.78 dB SDR, respectively, independent of angle.

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 (5.20) to (5.24), 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 5.1 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 5.1 for the specific parametrization used in each condition and the separation results.

The version of MESSL using Θ_{00} was published by Mandel et al. (2007) and referred to as EM−ILD by Mandel and Ellis (2007). 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 by Mandel and Ellis (2007), they use both IPD and ILD, but their IPD model again only uses a single σ per source. The others were introduced by (Mandel et al., 2009) under the same naming convention used here. The versions using Θ_{01} and $\Theta_{0\Omega}$ have τ-dependent IPD mean and variance, but no ILD. The last two versions use the full IPD and ILD models. With Θ_{11}, both are frequency-independent and with $\Theta_{\Omega\Omega}$ both are frequency-dependent.

It should be noted that initializing models with a large 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

Table 5.1: 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_{11}	μ_i	η_i	$\xi_{i\tau}$	$\sigma_{i\tau}$	4.57
$\Theta_{\Omega\Omega}$	$\mu_i(\omega)$	$\eta_i(\omega)$	$\xi_{i\tau}(\omega)$	$\sigma_{i\tau}(\omega)$	4.45
Θ_{10}	μ_i	η_i	0	σ_i	4.16
$\Theta_{\Omega 0}$	$\mu_i(\omega)$	$\eta_i(\omega)$	0	σ_i	4.07
$\Theta_{0\Omega}$	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

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.

5.4.2.1 Results

A summary of the results of this experiment can be seen in table 5.1. The most complex models, using Θ_{11} and $\Theta_{\Omega\Omega}$, achieve the best SDR. Note that the models are paired as performance is comparable with frequency-dependent and frequency-independent parameters. Because the performance of Θ_{11} and $\Theta_{\Omega\Omega}$ is so close, and because $\Theta_{\Omega\Omega}$ might allow more flexibility in more challenging conditions, we will use it in subsequent experiments. Comparing $\Theta_{\Omega 0}$ with Θ_{00}, 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 masker azimuth can be seen in figure 5.8. 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 parametrization 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 parametrization with zero mean and a single σ per source.

5.4.3 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 $\Theta_{\Omega\Omega}$ parameters for this experiment, with no garbage source, no ILD prior, and no probability-to-mask warping.

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 5.2 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 (Aarabi, 2002), 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 chapter, 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 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 (Shinn-Cunningham et al., 2005).

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 is only re-estimated after the first M step. Thus two iterations are required for a fair evaluation.

Table 5.2: 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

5.4.3.1 Results

The results of this experiment can be seen in table 5.2. 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.

5.4.4 Model extensions

In this section we bring together the various extensions to the basic MESSL system that have been discussed in sections 5.2.2 to 5.2.4. The three extensions are the garbage source used in conjunction with the ILD prior, the speaker-independent source prior using a 32-component mixture model, and the probability-to-mask warping. All eight combinations of these three extensions were evaluated on the two-speaker, reverberant mixtures.

Results are shown in figure 5.9 and table 5.3. The best performing system uses the garbage source, ILD prior, and the probability-to-mask warping, but not the source prior. Note that

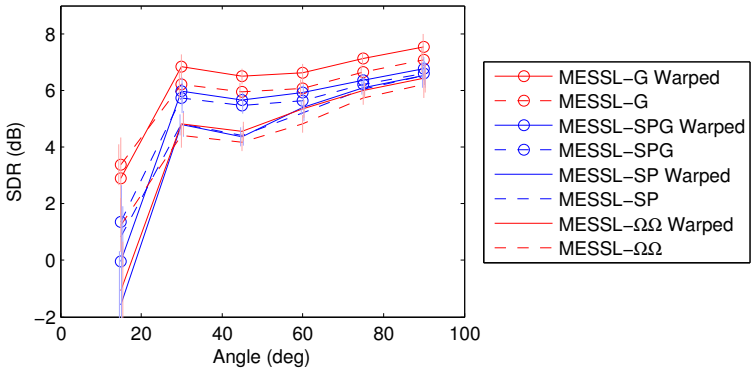


Figure 5.9: SDR as a function of angle for extended variants of MESSL. Uses combinations of a garbage source (marker), source priors (color), and probability-to-mask warping (line style). For example, the solid blue line with markers represents MESSL with a source prior, with the warping, and with the garbage source.

all of these extensions increase SDR above that of the best system in the previous section, MESSL-ΩΩ.

Of the three extensions, the garbage source increases the SDR the most consistently. The source prior helps performance a small amount, mostly when the garbage source is not present. The source prior appears to perform especially poorly when sources are close together. This is most likely due to the uncertainty of the interaural parameter-based separations, which are necessary before the source prior's channel model can start to discriminate between the two sources well.

The probability-to-mask warping is very effective with the garbage source and at larger angles, but it hurts performance at small angles, which diminishes and sometimes negates its advantage in table 5.3. The effect of the warping is to push the masks closer to binary, making them more "certain". This is helpful in most cases, where MESSL is, as it were, less certain than it should be, but it is detrimental when sources are close together as MESSL is rightly ambivalent. The warping helps less when applied to MESSL-SP variants, probably because both the source prior and the warping push the masks closer to being binary, so their effects are not additive.

5.4.5 Separation comparison with other algorithms

The fourth experiment compares MESSL's separation performance with four other well known source separation algorithms, DUET (Jourjine et al., 2000), the algorithm of Sawada et al. (Sawada et al., 2007), which we refer to as Sawada, the algorithm of Mouba and Marchand (Mouba and Marchand, 2006), which we refer to as Mouba, and TRINICON-based blind source separation using second order statistics (Buchner et al., 2005). We implemented the first three of these algorithms ourselves and tested them on mixtures of

Table 5.3: SDR for extended variants of MESSL. Uses combinations of a garbage source, source priors, and probability-to-mask warping.

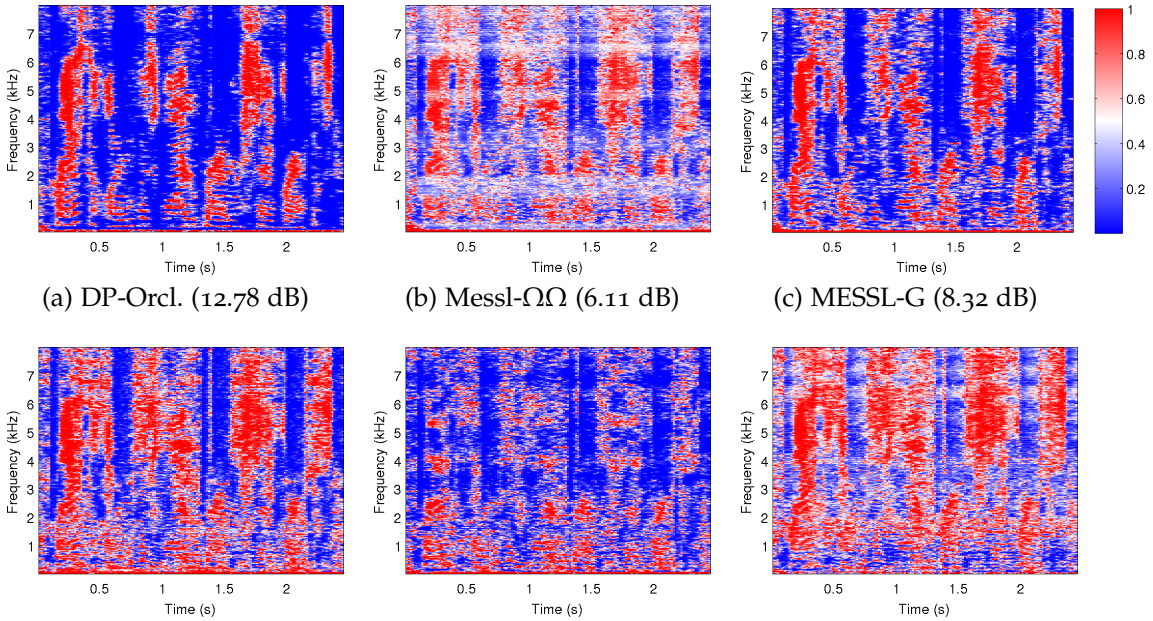
	Garbage Src	Source Priors	Warping	Avg
MESSL-G Warped	+	−	+	6.26
MESSL-G	+	−	−	5.90
MESSL-SPG Warped	+	+	+	5.11
MESSL-SPG	+	+	−	5.17
MESSL-SP Warped	−	+	+	4.27
MESSL-SP	−	+	−	4.64
MESSL-ΩΩ Warped	−	−	+	4.35
MESSL-ΩΩ	−	−	−	4.43

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) (Jourjine et al., 2000; Yilmaz and Rickard, 2004) 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 5.10(e) for an example mask estimated by DUET.

Two-stage frequency-domain blind source separation (Sawada et al., 2007) 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 5.10(d) for an example mask estimated by this algorithm.

The algorithm of Mouba and Marchand (Mouba and Marchand, 2006), 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



(a) DP-Orcl. (12.78 dB) (b) Messl-ΩΩ (6.11 dB) (c) MESSL-G (8.32 dB)

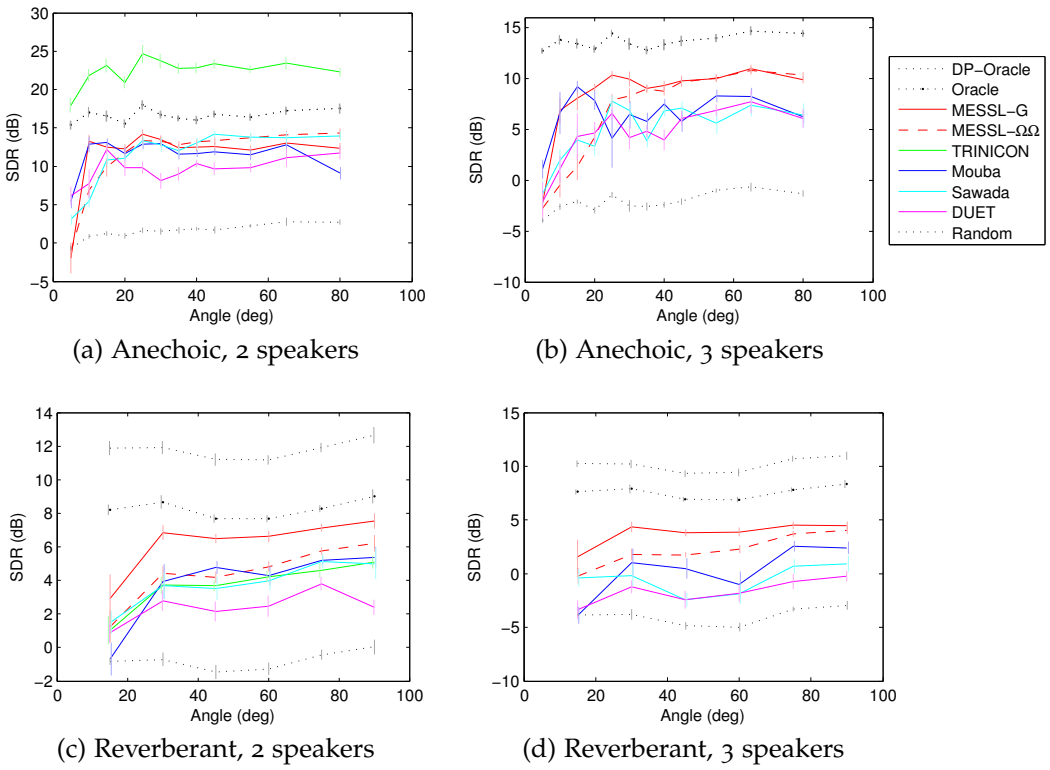
(d) Sawada (6.87 dB) (e) DUET (5.48 dB) (f) Mouba (5.58 dB)

Figure 5.10: Example masks from the algorithms compared in section 5.4.5 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 (c) is the same as figure 5.5(c).

(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 5.10(f) for an example mask estimated by Mouba and Marchand's algorithm.

Blind Source Separation using Second Order Statistics (BSS-SOS) using Triple-N ICA for convolutive mixtures (TRINICON) (Buchner et al., 2005) 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.

Two versions of MESSL are compared in this experiment. The first, referred to as MESSL-ΩΩ, is the basic frequency-dependent parametrization that was the best performing algorithm of section 5.4.2. The second, referred to as MESSL-G, uses the frequency-



(a) Anechoic, 2 speakers

(b) Anechoic, 3 speakers

(c) Reverberant, 2 speakers

(d) Reverberant, 3 speakers

Figure 5.11: SDR as a function of angle for four 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. Note that the y-axis is different for all of these plots.

dependent parametrization along with a garbage source, the ILD prior, and the probability-to-mask warping, which was the best performing algorithm of section 5.4.4.

5.4.5.1 Results

The SDR of these separations are shown as a function of separation angle in figure 5.11 and summarized in table 5.4 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. The source prior did not improve performance at small separations for MESSL-SP, however, possibly because of initialization issues. Note that the DP-Oracle mask achieves much higher SDR and PESQ scores than the Oracle mask, due to its exclusion of all reverberation.

Table 5.4: SDR results (in dB) 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.

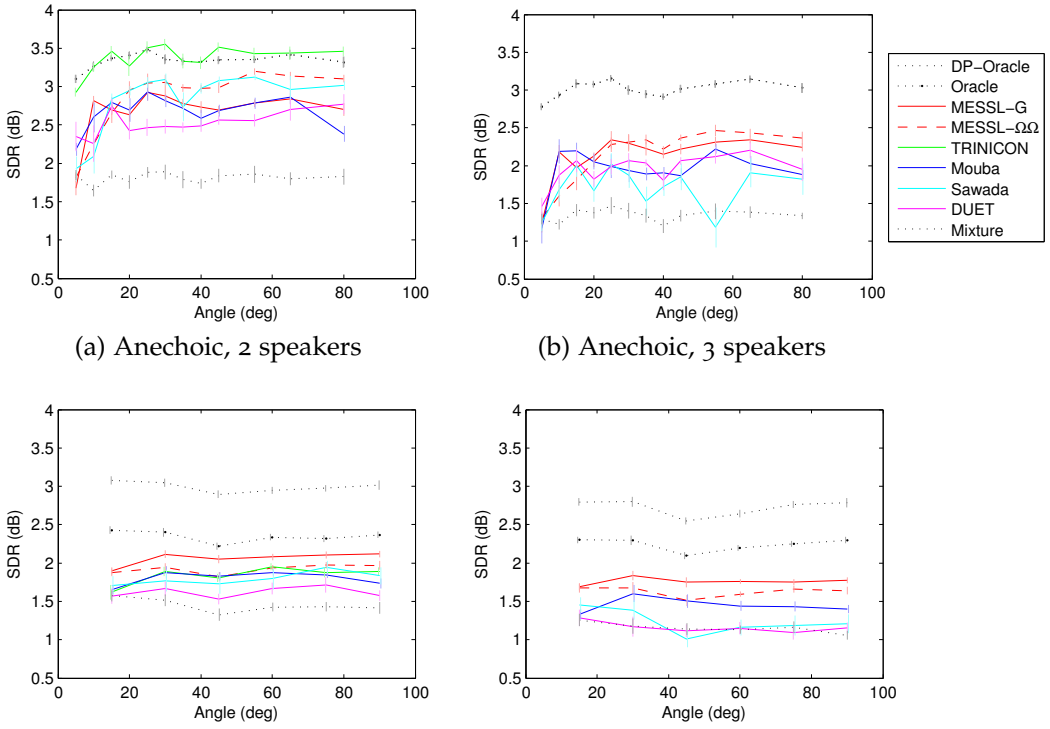
	A2	A3	R2	R3	Avg
DP-Oracle	16.63	13.63	11.79	10.15	13.05
Oracle	16.63	13.63	8.25	7.59	11.52
MESSL-G	11.56	**8.42**	**6.26**	**3.76**	**7.50**
MESSL-ΩΩ	11.30	6.43	4.43	2.23	6.10
Mouba	11.47	6.43	3.81	0.27	5.49
Sawada	11.44	4.98	3.80	−0.55	4.91
TRINICON	**22.47**	—	3.72	—	—
DUET	9.63	4.52	2.41	−1.62	3.73
Random	1.54	−2.13	−0.78	−3.95	−1.33

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 5.10(c) 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-interference 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



(a) Anechoic, 2 speakers

(b) Anechoic, 3 speakers

(c) Reverberant, 2 speakers

(d) Reverberant, 3 speakers

Figure 5.12: PESQ as a function of angle for four 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.

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, and can be seen in figure 5.12 and table 5.5. As would be expected, the PESQ scores for anechoic mixtures 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

Table 5.5: PESQ results (in units of mean opinion score, -0.5 is the worst possible and 4.5 is the best possible) comparing systems in anechoic and reverberant environments with 2 and 3 speakers. Comparable to table 5.4

	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	2.68	**2.14**	**2.06**	**1.76**	**2.16**
MESSL-ΩΩ	2.85	2.13	1.92	1.63	2.13
TRINICON	**3.37**	—	1.84	—	—
Mouba	2.67	1.94	1.80	1.45	1.97
Sawada	2.82	1.71	1.80	1.24	1.89
DUET	2.52	1.95	1.62	1.16	1.81
Mixture	1.81	1.35	1.45	1.15	1.44

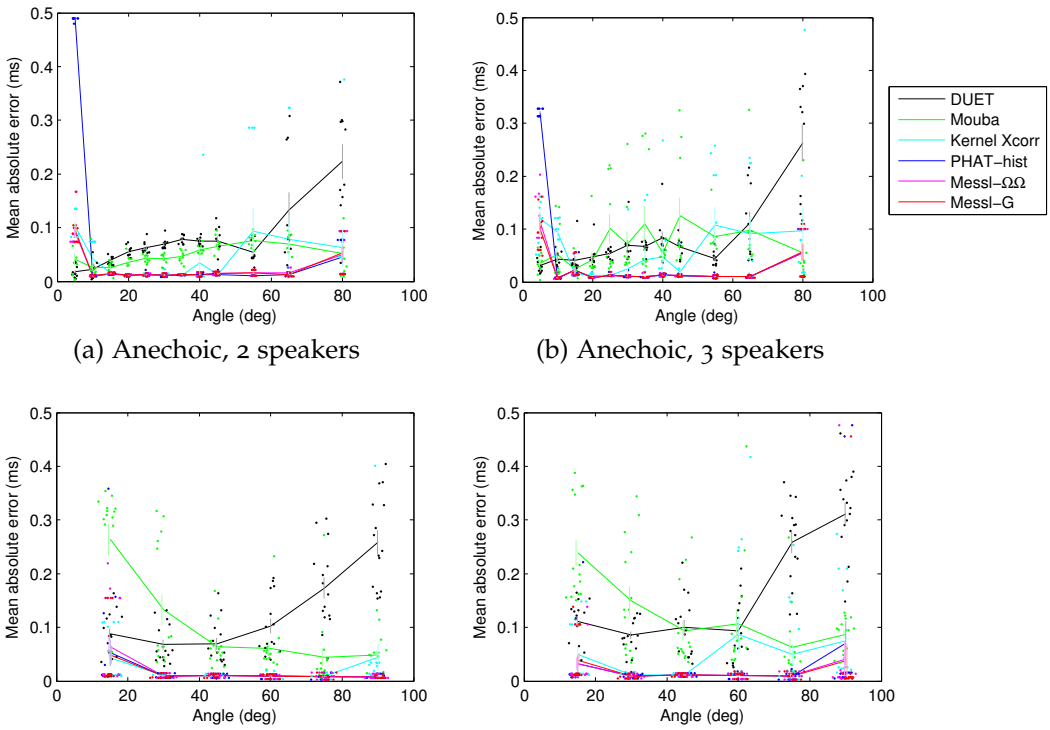
the average mixture score and the average DP-Oracle score, MESSL was able to recover approximately half, or 0.72 units. Note that the probability-to-mask warping decreases MESSL-G's performance under this metric for anechoic mixtures, decreasing its average performance slightly as well.

5.4.6 Localization

We also compare the localization performance of some of the algorithms described in section 5.4.5. Specifically, we compare DUET, Mouba's algorithm, and the two MESSL variants, with and without the garbage source. Sawada's algorithm only performs "localization" separately in each frequency band, it does not estimate a global localization. Unfortunately, we were only able to get the separation estimates from TRINICON, not the localization, so it is not included either. We do include, however, two localization algorithms that were not used in section 5.4.5. These are the two algorithms described in chapter 4, PHAT-histogram (Aarabi, 2002) and the normalized kernel cross-correlation (Kernel Xcorr).

Note that MESSL was initialized from PHAT-histogram, except when PHAT-histogram detected that it could not make a reliable localization. This determination was made by comparing the height of the histogram at the largest peak to the Ith largest peak. In PHAT-histogram's typical failure mode, frames are clustered into too few peaks when the sources were very close together. This results in a spurious peak, much smaller than the true peaks and generally quite far from the true source location, being selected. To detect this situation, the largest peak was compared with the Ith largest peak and if it was 10 times larger then the localization was deemed to have failed. In these cases, MESSL reverted to initializing itself from peaks in a standard cross-correlation between the input signals.

The localization experiments performed here are the same as those performed in chapter 4. The ITDs estimated by each algorithm are compared to the ground truth ITDs of each



(a) Anechoic, 2 speakers

(b) Anechoic, 3 speakers

(c) Reverberant, 2 speakers

(d) Reverberant, 3 speakers

Figure 5.13: Mean absolute localization error as a function of angle for four conditions. Anechoic conditions are averaged over 10 mixtures at each of 12 angles, reverberant over 15 mixtures at each of 6 angles. Each dot is the mean absolute error of all of the sources in one mixture. The maximum localization excursion was ± 0.75 ms.

source, which are estimated by passing 10 s of white noise through the first 10 ms of each impulse response. The mean absolute error of all of the sources, not just the target source, is then computed for each mixture, and each of these mean errors is plotted as a dot in figure 5.13, with the overall mean absolute error at each angle plotted as the line. The mean absolute error for each condition and overall is shown for each algorithm in table 5.6.

These results reveal a number of interesting characteristics. First, unexpectedly, localization performance is not generally better in either anechoic or reverberant conditions. From table 5.6, DUET and Mouba's algorithm both perform better in anechoic conditions, as would be expected. Kernel Xcorr performs about the same in the two conditions. And PHAT-histogram and MESSL perform better in reverberant conditions than anechoic.

Figure 5.13 shows that PHAT-histogram's poor performance in anechoic mixtures is due to poor localization for sources separated by 5°, which is due to the failure mode mentioned above. Because it uses a fall-back cross-correlation based initialization when PHAT-histogram fails, MESSL's performance is generally bounded above by PHAT-histogram. This cross-correlation based initialization is less precise than PHAT-histogram, however,

Table 5.6: Mean absolute localization error (in μs) for algorithms under comparison. The error is averaged over all speakers in two and three speaker mixtures in reverberant and anechoic conditions.

	A2	A3	R2	R3	Avg
DUET	75	77	127	160	110
Mouba	48	71	102	123	86
Kernel Xcorr	41	57	21	47	42
PHAT-hist	54	41	**16**	24	34
Messl-ΩΩ	**23**	24	18	**19**	21
Messl-G	**23**	**23**	**16**	20	**20**

leading to slightly worse localization performance than when PHAT-histogram is functioning properly, but much better localization performance than when PHAT-histogram is in error. Note that this failure at small separations is less of an issue in reverberation because the smallest separation in that dataset is 15°.

Figure 5.13 also shows that DUET works better for smaller separations than for larger separations. This is probably due to the upper frequency threshold being set slightly too high, and spatial aliases disturbing these estimates. In reverberation, even if the threshold is set appropriately, noise can perturb the IPD across the $\pm\pi$ boundary, distorting these estimates.

Mouba's algorithm, on the other hand, works better for larger separations than smaller separations, especially in reverberation. This is probably similar to PHAT-histogram's failure mode. At small separations, the modes of the ITD estimate histogram merge, making it difficult to distinguish between the location of each of the sources.

And finally, the performance of Kernel Xcorr is better for smaller angles. This is probably again due to spatial aliasing, as little is done in Kernel Xcorr to counteract this effect besides pooling estimates over different time-frequency regions. These regions could still be smaller than would be ideal for localization performance, as they were tuned for separation performance.

5.5 Summary

This chapter has presented a novel source separation algorithm that uses probabilistic source models and an expectation maximization parameter estimation procedure. Building on the probabilistic localization model of single sound sources introduced in chapter 4, which can be evaluated independently at every time-frequency point, we constructed a mixture model that reduced the multi-source parameter estimation problem to a collection of single source parameter estimation problems. This approach combines localization with the idea of probabilistic masking using an EM algorithm.

We then performed five experiments to test various aspects of the model. We found that two and a half seconds of audio contained enough information to fit the most complex parametrizations of MESSL, which performed 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 separated sources with signal-to-distortion ratios 2.0 dB higher and PESQ scores 0.19 MOS higher than comparable algorithms, and localized sources with 40% lower mean absolute error.