Combining Localization Cues and Source Model Constraints for Binaural Source Separation

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

LabROSA, Dept. of Electrical Engineering
Columbia University
New York NY 10027 USA

Abstract

We describe a system for separating multiple sources from a two-channel recording based on interaural cues and prior knowledge of the statistics of the underlying source signals. The proposed algorithm effectively combines information derived from low level perceptual cues, similar to those used by the human auditory system, with higher level information related to speaker identity. We combine a probabilistic model of the observed interaural level and phase differences with a prior model of the source statistics and derive an EM algorithm for finding the maximum likelihood parameters of the joint model. The system is able to separate more sound sources than there are observed channels in the presence of reverberation. In simulated mixtures of speech from two and three speakers the proposed algorithm gives a signal-to-noise ratio improvement of 1.7 dB over a baseline algorithm which uses only interaural cues. Further improvement is obtained by incorporating eigenvoice speaker adaptation to enable the source model to better match the sources present in the signal. This improves performance over the baseline by 2.7 dB when the speakers used for training and testing are matched. However, the improvement is minimal when the test data is very different from that used in training.

Key words: source separation, binaural, source models, eigenvoices, EM

Email address:
ronw@ee.columbia.edu,mim@ee.columbia.edu,dpwe@ee.columbia.edu (Ron J. Weiss, Michael I. Mandel, Daniel P. W. Ellis)

1. Introduction

Human listeners are often able to attend to a single sound source in the presence of background noise and other competing sources. This is partially a result of the human auditory system's ability to isolate sound sources that arrive from different spatial locations, an effect of the fact that humans have two ears (Cherry, 1953). Localization is derived from low-level acoustic cues based on the time and level differences of the sounds arriving at a listener's ears (Blauert, 1997). The use of these perceptual localization cues has had much success in the development of binaural source separation algorithms (Yilmaz and Rickard, 2004; Mandel and Ellis, 2007). Unlike competing source separation approaches such as independent component analysis, localization-based algorithms are often able to separate mixtures containing more than two sources despite utilizing only binaural observations.

In contrast to binaural source separation based on the same principles used by the human auditory system, the most successful approaches to separating sources given a single channel observation have been model-based systems which rely on pre-trained models of source statistics (Cooke et al., 2010). Such monaural source separation algorithms generally require relatively large, speaker-dependent (SD) models to obtain high quality separation. These supervised methods therefore have the disadvantage of requiring that the identities of all sources be known in advance and that sufficient data be available to train models for each them. In contrast, most binaural separation algorithms based on localization cues operate without any prior knowledge of the signal content. The only assumption they make is that the sources be spatially distinct from one another. However, it is to be expected that incorporating some prior knowledge about the source characteristics would be able to further improve separation performance.

In this paper we describe a system for source separation that combines inference of localization parameters with model-based separation methods and show that the additional constraints derived from the source model help to improve separation performance. In contrast to typical model-based monaural separation algorithms, which require complex SD source models to obtain high quality separation, the proposed algorithm is able to achieve high quality separation using significantly simpler source models and without requiring that the models be specific to a particular speaker.

The remainder of this paper is organized as follows: Section 2 reviews previous work related to the algorithms we describe in this work. Section 3

describes the proposed signal model for binaural mixtures and section 4 describes how this model is used for source separation. Experimental results comparing the proposed systems to other state of the art algorithms for binaural source separation are reported in section 5.

2. Previous Work

In this paper we propose an extension of the Model-based Expectation Maximization Source Separation and Localization (MESSL) algorithm (Mandel et al., 2010), which combines a cross-correlation approach to source localization with spectral masking for source separation. MESSL is based on a model of the interaural phase and level differences derived from the observed binaural spectrograms. This is similar to the Degenerate Unmixing Estimation Technique (DUET) algorithm for separating underdetermined mixtures (Yilmaz and Rickard, 2004) and other similar approaches to source localization (Nix and Hohmann, 2006) which are based on clustering localization cues across time and frequency. These systems work in an unsupervised manner by searching for peaks in the two dimensional histogram of interaural level difference (ILD) and interaural time, or phase, difference (ITD or IPD) to localize sources. In the case of DUET, source separation is based on the assumption that each point in the spectrogram is dominated by a single source. Different regions of the mixture spectrogram are associated with different spatial locations to form time-frequency masks for each source.

Harding et al. (2006) and Roman et al. (2003) take a similar but supervised approach, where training data is used to learn a classifier to differentiate between sources at different spatial locations based on features derived from the interaural cues. Unlike the unsupervised approach of Yilmaz and Rickard (2004) and Nix and Hohmann (2006), this has the disadvantage of requiring labeled training data. MESSL is most similar to the unsupervised separation algorithms, and is able to jointly localize and separate spatially distinct sources using a parametric model of the interaural parameters estimated directly from a particular mixture.

A problem with all of these methods is the fact that, as we will describe in the next section, the localization cues are often ambiguous in some frequency bands. Such regions can be ignored if the application is limited to localization, but the uncertainty leads to reduced separation quality when using spectral masking. Under reverberant conditions the localization cues are additionally obscured by the presence of echoes which come from all directions. Binaural

source separation algorithms that address reverberation have been proposed by emphasizing onsets and suppressing echoes in a process inspired by the auditory periphery (Palomäki et al., 2004), or by preprocessing the mixture using a dereverberation algorithm (Roman and Wang, 2006).

In this paper we describe two extensions to the unsupervised MESSL algorithm which incorporate a prior model of the underlying anechoic source signal which does not suffer from the same underlying ambiguities as the interaural observations and therefore is able to better resolve the individual sources in these regions. Like the supervised separation methods described above, this approach has the disadvantage of requiring training data to learn the source prior (SP) model, but as we will show in section 5, such a prior can significantly improve performance even if it is not perfectly matched to the test data. Furthermore, because the source prior model is trained using anechoic speech, it tends to de-emphasize reverberant noise and therefore improves performance over the MESSL baseline, despite the fact that it does not explicitly compensate for reverberation in a manner similar to Palomäki et al. (2004) or Roman and Wang (2006).

The idea of combining localization with source models for separation has been studied previously in Wilson (2007) and Rennie et al. (2003). Given prior knowledge of the source locations, Wilson (2007) describes a complementary method for binaural separation based on a model of the magnitude spectrum of the source signals. This approach combines a model of the IPD based on known source locations with factorial model-based separation as in Roweis (2003) where each frame of the mixed signal is explained by the combination of models for each of the underlying source signals. The system described in Wilson (2007) models all sources using the same source-independent (SI) Gaussian mixture model (GMM) trained on clean speech from multiple talkers. Such a model generally results in very poor separation due to the lack of temporal constraints and lack of source-specific information available to disambiguate the sources (Weiss and Ellis, 2010). In this case, however, the localization model is able to compensate for these shortcomings. Per-source binary masks are derived from the joint IPD and source model and shown to improve performance over separation systems based on localization cues alone.

Rennie et al. (2003) take a similar approach to combining source models with known spatial locations for separation using microphone arrays. Instead of treating the localization and source models independently, they derive a model of the complex speech spectrum based on a prior on the speech

magnitude spectrum that takes into account the effect of phase rotation consistent with a source signal arriving at the microphone array from a particular direction. Like the other systems described above, Rennie et al. (2003) is able to separate more sources than there are microphones.

These systems have some disadvantages when compared to the extensions to MESSL described in this paper. The primary difference is that they depend on prior knowledge of the source locations whereas MESSL and its extensions are able to jointly localize and separate sources. Rennie et al. (2005) describe an extension to Rennie et al. (2003) that is able to estimate the source locations as well, bringing it closer to our approach. A second difference is that these systems use a factorial model to model the interaction between different sources. In Wilson (2007) this leads to inference that scales exponentially with the number of underlying sources. Although the signal models in Rennie et al. (2003, 2005) are similar, they are able to manage this complexity using an approximate variational learning algorithm. In contrast, exact inference in the model we propose in this paper is linear in the number of sources.

In the next section, we describe the baseline MESSL algorithm and two closely related extensions to incorporate a prior distribution over the source signal statistics: MESSL-SP (Source Prior) which uses the same SI model for all sources as in Weiss et al. (2008), and MESSL-EV (Eigenvoice) which uses eigenvoice speaker adaptation (Kuhn et al., 2000) to learn source-specific parameters to more accurately model the source signals. In both cases, the information extracted from the interaural cues and source model serve to reinforce each other. We show that it is possible to obtain significant improvement in separation performance of speech signals in reverberation over a baseline system employing only interaural cues. As in Wilson (2007) and Rennie et al. (2003), the improvement is significant even when the source models used are quite weak, and only loosely capture the spectral shapes characteristic of different speech sounds. The use of speaker-adapted models in MESSL-EV is sometimes able to improve performance even more, a further improvement over the source-independent models used by other similar systems.

3. Binaural mixed signal model

We model the mixture of I spatially distinct source signals $\{\mathrm{x}_i(t)\}_{i=1..I}$ based on the binaural observations $\mathrm{y}^\ell(t)$ and $\mathrm{y}^r(t)$ corresponding to the signals

arriving at the left and right ears respectively. For a sufficiently narrowband source in an anechoic environment, the observations will be related to a given source signal primarily by the gain and delay that characterize the direct path from the source location. However, in reverberant environments this assumption is confused by the addition of convolutive noise arising from the room impulse response. In general the observations can be modeled as follows in the time domain:

$$\mathrm{y}^{\ell}(t) = \sum_i \mathrm{x}_i(t - \tau_i^{\ell}) * \mathrm{h}_i^{\ell}(t) \tag{1}$$

$$\mathrm{y}^{r}(t) = \sum_i \mathrm{x}_i(t - \tau_i^{r}) * \mathrm{h}_i^{r}(t) \tag{2}$$

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

3.1. Interaural model

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

$$\mathrm{X}_{IS}(\omega, t) \triangleq \frac{\mathrm{Y}^{\ell}(\omega, t)}{\mathrm{Y}^{r}(\omega, t)} = 10^{\alpha(\omega,t)/20} e^{j\phi(\omega,t)} \tag{3}$$

where $\mathrm{Y}^{\ell}(\omega, t)$ and $\mathrm{Y}^{r}(\omega, t)$ are the short-time Fourier transforms of $\mathrm{y}^{\ell}(t)$ and $\mathrm{y}^{r}(t)$, respectively. For a given time-frequency cell, the interaural level difference (ILD) in decibels between the two channels is $\alpha(\omega, t)$, and $\phi(\omega, t)$ is the corresponding interaural phase difference (IPD).

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

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

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

where $|\mathrm{H}(\omega)|$ is the magnitude of $\mathrm{H}(\omega)$, which is defined analogously to $\mathrm{Y}(\omega, t)$, and the subscript i is the index of the particular source dominant at that cell,

and thus depends on ω and t. These quantities have the advantage of being independent of the source signal, which is why the baseline MESSL model does not require knowledge of the distribution of $x_i(t)$.

A necessary condition for the accurate modeling of the observation is that the interaural time difference (ITD) $\tau_i^\ell - \tau_i^r$ be much smaller than the window function used in calculating $\mathrm{X}_{IS}(\omega, t)$. In the experiments described in section 5, we use a window length of 64 ms and a maximum ITD of about 0.75 ms. Similarly, $\mathrm{h}_i^{\ell,r}(t)$ must be shorter than the window. This assumption does not generally hold in reverberation because a typical room impulse response has a duration of at least a few hundred milliseconds. However, we ignore this for the purposes of our model and note that effect of violating this assumption is to increase the variance in the ILD model. We model the ILD for source i as a Gaussian distribution whose mean and variance will be learned directly from the mixed signal:

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

where θ stands for the otherwise unspecified model parameters.

The model for the IPD requires some additional considerations. It is difficult to learn the IPD for a given source directly from the mixed signal because $\phi(\omega, t)$ is only observed modulo 2π. This is a consequence of spatial aliasing that results at high frequencies if the ITD is large enough that $|\omega(\tau^\ell - \tau^r)| > \pi$ (Yilmaz and Rickard, 2004). Because of this the observed IPD cannot always be mapped directly to a unique time difference. However, a particular ITD will correspond unambiguously to a single phase difference. This is illustrated in figure 1. This motivates a top down approach where the observed IPD will be tested against the predictions of a set of predefined time differences. The difference between the IPD predicted by an ITD of τ samples and the observed IPD is measured by the phase residual:

$$\tilde{\phi}_\tau(\omega,t) = \arg\left(e^{j\phi(\omega,t)}\,e^{-j\omega\tau}\right) \tag{7}$$

which is always in the interval $(-\pi, \pi]$. Given a predefined set of such τs, the IPD distribution for a given source has the form of a Gaussian mixture model with one mixture component for each time difference:

$$P\big(\phi(\omega,t),\,i\,|\,\theta\big) = \sum_\tau \psi_{i\tau}\,\mathcal{N}\big(\tilde{\phi}_\tau(\omega,t);\,0,\,\varsigma_i^2\big) \tag{8}$$

where $\psi_{i\tau} = P\big(i,\tau\big)$ are the mixing weights for source i and delay τ.

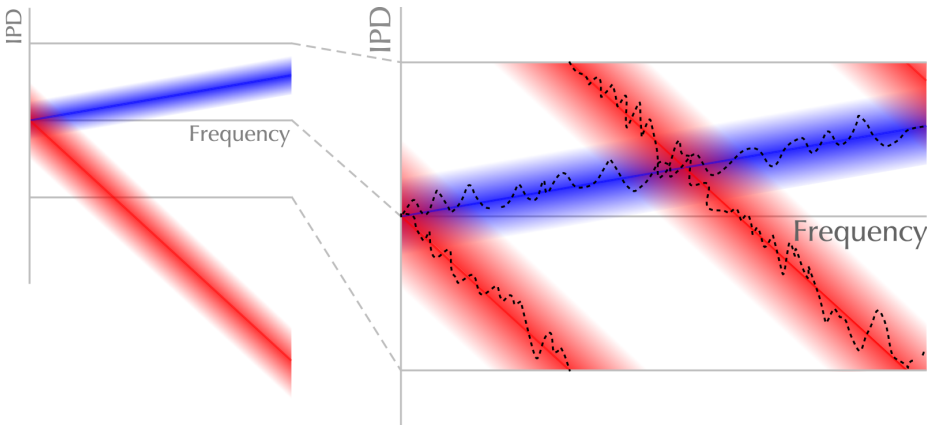

Figure 1: Illustration of spatial aliasing in our model of the interaural phase difference (IPD). The left pane shows the predicted IPD distribution for two distinct sources centered on their respective values of ω, t. The right pane demonstrates the observed IPDs for the two sources (dotted lines) with the distributions overlaid. The IPDs are observed modulo 2π due to the periodicity of the complex sinusoid in equation (3). For small interaural time difference (blue) this is not a problem, however if the ITD is large (red) the IPD wraps around from $-\pi$ to π. This is especially problematic in mixtures because the wrapping results in additional ambiguity when the IPDs for the different sources intersect.

An example of the ILD and IPD observations used by the interaural model is shown in figure 2. The contributions of the two sources are clearly visible in both the ILD and IPD observations. The target source, which is located at an angle of 0° relative to the microphones, has an ILD close to zero at all frequencies while the ILD of the other source becomes increasingly negative at higher frequencies. This trend is typical of a source off to one side, since the level difference, which results from the "shadowing" effect of the head or baffle between the microphones, increases when the wavelength of the sound is small relative to the size of the baffle. Similarly, the IPD for the target source has an IPD close to zero at all frequencies while the IPD for the other source varies with frequency, with the phase wrapping clearly visible at about 1, 3, and 5 kHz.

3.2. Source model

We extend the baseline MESSL model described in the previous section to incorporate prior knowledge of the source statistics. This makes it possible

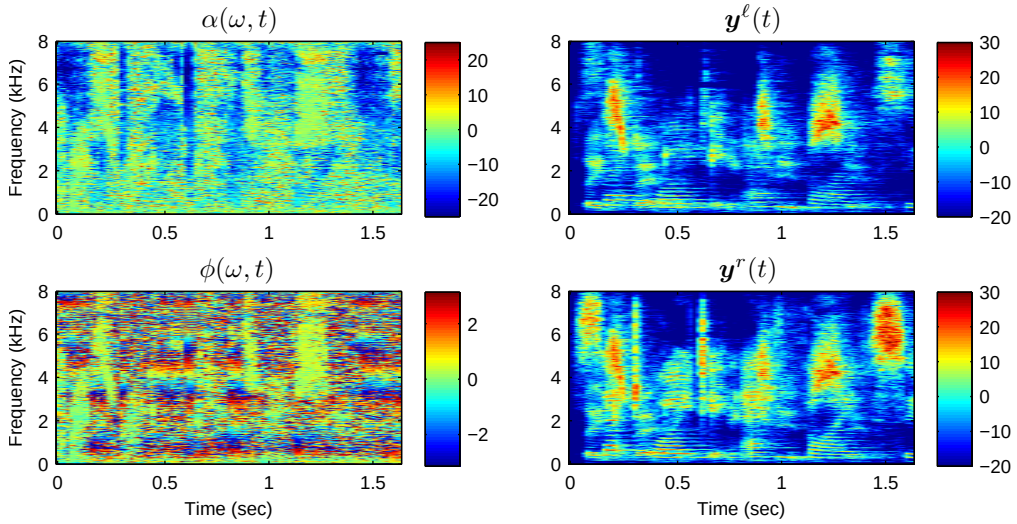

Figure 2: Observed variables in the MESSL-EV model derived from a mixture of two sources in reverberation separated by 60 degrees. The left column shows example ILD (top) and IPD (bottom) observations. The right column shows the left and right spectrograms modeled using the source model.

to model the binaural observations directly:

$$y^{\ell}(\omega, t) \approx x_i(\omega, t) + h_i^{\ell}(\omega) \tag{9}$$

$$y^{r}(\omega, t) \approx x_i(\omega, t) + h_i^{r}(\omega) \tag{10}$$

where $x_i(\omega, t) \triangleq 20 \log_{10} |\mathrm{X}_i(\omega, t)|$, and $y^{\ell}(\omega, t)$, $y^{r}(\omega, t)$, and $h_i(\omega)$ are defined analogously. An example of these observations derived from a mixture of two sources in reverberation is shown in the right column of figure 2.

For simplicity we model the distribution of the source signal $x_i(\omega, t)$ using a Gaussian mixture model with diagonal covariances. The likelihood of a frame of one frame of the signal, $\boldsymbol{x}_i(t)$, can therefore be written as follows:

$$P\big(\boldsymbol{x}_i(t)\big) = \sum_c \pi_{ic}\, \mathcal{N}\big(\boldsymbol{x}_i(t);\, \boldsymbol{\mu}_{ic},\, \Sigma_{ic}\big) \tag{11}$$

where c indexes the different source mixture components (states), and $\pi_{ic} = P\big(c \,|\, i\big)$ are the mixing weights for source i and component c.

We assume that the channel responses $h_i^{\ell,r}$ will be relatively smooth across frequency, and that they will be constant across the entire mixture,

i.e. the sources and the sensors remain stationary. The channel response is parametrized in the DCT domain, giving $h_i^\ell(\omega) = B(\omega, :)\,\mathbf{h}_i^\ell$ where B is a matrix of DCT basis vectors, $B(\omega, :)$ is the row of B corresponding to frequency ω, and $\mathbf{h}_i^\ell$ is a vector of weights, the projection of the channel onto the DCT basis. This allows $h_i^{\ell,r}$ to be modeled using many fewer DCT coefficients than the number of frequency bands Ω.

Combining this model of the channel response with the source model gives the following likelihoods for the left and right channel spectrograms:

$$P\big(y^\ell(\omega,t)\,|\,i,\,c,\,\theta\big) = \mathcal{N}\big(y^\ell(\omega,t);\,\mu_{ic}(\omega) + B(\omega,:)\,\mathbf{h}_i^\ell,\,\sigma_{ic}^2(\omega)\big) \tag{12}$$

$$P\big(y^r(\omega,t)\,|\,i,\,c,\,\theta\big) = \mathcal{N}\big(y^r(\omega,t);\,\mu_{ic}(\omega) + B(\omega,:)\,\mathbf{h}_i^r,\,\sigma_{ic}^2(\omega)\big) \tag{13}$$

where $\sigma_{ic}^2(\omega)$ is the diagonal entry of Σ_{ic} corresponding to frequency ω.

3.2.1. Speaker-independent source prior

Because the number of observations in a typical mixture is generally very small compared to the amount of data needed to reliably train a signal model describing the distribution of $\boldsymbol{x}_i(t)$, we use a speaker-independent prior source model trained in advance on data from a variety of speakers. When using such a model, the GMM parameters in equation (11) are independent of i and the distributions in equations (13) and (12) for each source are only differentiated by the source-dependent channel parameters. These distributions are therefore initially uninformative because $\mathbf{h}_i^\ell$ and $\mathbf{h}_i^r$ are initialized to zero, in which case equations (13) and (12) evaluate to the same likelihood for each source. However, when the interaural model and source prior model are combined, the binaural cues begin to disambiguate the sources and the estimated channel responses help to differentiate the source models. We refer to the combination of the interaural model and source prior model in this configuration as MESSL-SP.

3.2.2. Eigenvoice adaptation

Alternatively, we can use model adaptation to take advantage of the source-dependent characteristics of the different sources despite the lack of sufficient observed data to robustly estimate source-dependent distributions. Model adaptation is a widely studied topic in automatic speech recognition. Kuhn et al. (2000) propose the "eigenvoice" technique for rapid speaker adaptation when the amount of adaptation data is limited, as little as a single utterance containing only a few seconds of speech. When incorporating

eigenvoice adaptation into the combined interaural and source models, we refer to the model as MESSL-EV.

The eigenvoice idea is to represent the means of a speaker-dependent GMM as a linear combination of a "mean voice", essentially corresponding to the SI model, and a set of basis vectors U. The likelihood of component c under such an adapted model for source i can be written as follows:

$$P\big(\boldsymbol{x}_i(t) \,|\, c,\, \mathbf{w}_i\big) \;=\; \mathcal{N}\big(\boldsymbol{x}_i(t);\, \boldsymbol{\mu}_c(\mathbf{w}_i),\, \bar{\Sigma}_c\big) \tag{14}$$

$$\boldsymbol{\mu}_{ci} \;=\; \boldsymbol{\mu}_c(\mathbf{w}_i) \;=\; \bar{\boldsymbol{\mu}}_c + \sum_k w_{ik}\, \hat{\boldsymbol{\mu}}_{ck} \;=\; \bar{\boldsymbol{\mu}}_c + U_c\, \mathbf{w}_i \tag{15}$$

where $\bar{\boldsymbol{\mu}}_c$ and $\bar{\Sigma}_c$ are the mean and covariance, respectively, of the SI model, $\hat{\boldsymbol{\mu}}_{ck}$ is the kth basis vector for mixture component c, and $U_c = [\hat{\boldsymbol{\mu}}_{c1},\, \hat{\boldsymbol{\mu}}_{c2},\, \ldots,\, \hat{\boldsymbol{\mu}}_{cK}]$.

Essentially, the high dimensional model parameters for a particular speaker are represented as a function of the low dimensional adaptation parameters $\mathbf{w}_i$, which typically contains only a few tens of dimensions. The bulk of the knowledge of speakers characteristics is embedded in the predefined speaker basis vectors U. Adaptation is just a matter of learning the ideal combination of bases, essentially projecting the observed signal onto the space spanned by U.

The eigenvoice bases U are learned from a set of pre-trained SD models using principal component analysis. For each speaker in the training data, a supervector of model parameters, $\boldsymbol{\mu}_i$, is constructed by concatenating the set of Gaussian means for all mixture components in the model. Parameter supervectors are constructed for all M speaker models and used to construct a parameter matrix $P = [\boldsymbol{\mu}_1,\, \boldsymbol{\mu}_2,\, \ldots,\, \boldsymbol{\mu}_M]$ that spans the space of speaker variation. The mean voice $\bar{\boldsymbol{\mu}}$ is found by taking the mean across columns of P. Performing the singular value decomposition on $P - \bar{\boldsymbol{\mu}}$ then yields orthonormal basis vectors for the eigenvoice space, U.

Although the ordering of components in the parameter supervectors is arbitrary, care must be taken to ensure that the ordering is consistent for all speakers. A simple way to guarantee this consistency is to use an identical initialization for learning all of the underlying speaker models. We therefore bootstrap each SD model using the SI model described above to ensure that each mixture component of the SD models corresponds directly to the same component in the SI model.

A more detailed discussion of eigenvoice adaptation is beyond the scope of this paper. Its application to model-based source separation is explored in detail in Weiss and Ellis (2010) and Weiss (2009).

3.3. Putting it all together

Combining the model of the interaural signals with the source model gives the complete likelihood of the model including the hidden variables:

$$\begin{aligned} P\big(\phi(\omega,t),\, \alpha(\omega,t),\, y^\ell(\omega,t),\, y^r(\omega,t),\, i,\, \tau,\, c \,|\, \theta\big) \\ = P\big(i,\, \tau\big)\, P\big(\phi(\omega,t) \,|\, i,\, \tau,\, \theta\big)\, P\big(\alpha(\omega,t) \,|\, i,\, \theta\big) \\ P\big(c \,|\, i\big)\, P\big(y^\ell(\omega,t) \,|\, i,\, c,\, \theta\big)\, P\big(y^r(\omega,t) \,|\, i,\, c,\, \theta\big) \end{aligned} \quad (16)$$

This equation explains each time-frequency point of the mixed signal as being generated by a single source i at a given delay τ using a particular component c in the source model. The graphical model corresponding to this factorization is shown in figure 3. This figure only includes the observations and those parameters that are estimated to match a particular mixture. We describe the parameter estimation and source separation process in the following section. For simplicity we omit the parameters that remain fixed during separation, including π_c, $\bar{\boldsymbol{\mu}}_c$, U_c, and Σ_c, which are learned offline from a corpus of training data. It is also important to note that the figure depicts the full MESSL-EV model. If eigenvoice adaptation is not used, then $\mathbf{w}_i$ is clamped to zero and the model reduces to the original MESSL-SP model as described in Weiss et al. (2008).

Note that all time-frequency points are conditionally independent given the model parameters. The total likelihood of the observations can therefore be written as follows:

$$P\big(\boldsymbol{\phi},\, \boldsymbol{\alpha},\, \boldsymbol{y}^\ell,\, \boldsymbol{y}^r \,|\, \theta\big) = \prod_{\omega t} \sum_{i \tau c} P\big(\phi(\omega,t),\, \alpha(\omega,t),\, y^\ell(\omega,t),\, y^r(\omega,t),\, i,\, \tau,\, c \,|\, \theta\big) \quad (17)$$

The combined model is essentially the product of three independent mixtures of Gaussians, corresponding to the IPD, ILD, and source models. For conciseness we will drop the (ω, t) where convenient throughout the remainder of this paper.

4. Parameter estimation and source separation

The model described in the previous section can be used to separate sources because it naturally partitions the mixture spectrogram into regions dominated by different sources. Given estimates of the source-specific model

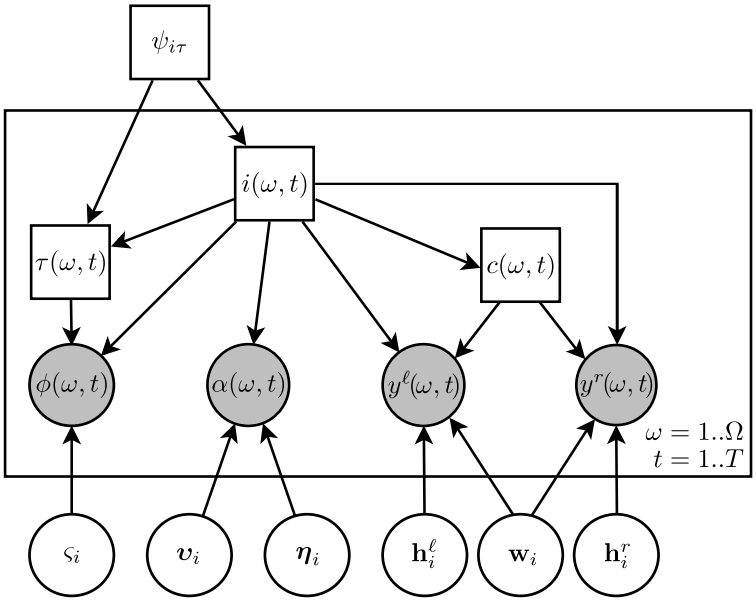

Figure 3: MESSL-EV graphical model of a mixture spectrogram. Each time-frequency point is explained by a source i, a delay τ, and a source model component c. Square nodes represent discrete variables and round nodes represent continuous variables Shaded nodes correspond to observed quantities.

parameters $\theta = \{\psi_{i\tau}, \varsigma_i^2, \boldsymbol{\upsilon}_i, \boldsymbol{\eta}_i^2, \mathbf{w}_i, \mathbf{h}_i^{\ell}, \mathbf{h}_i^{r}\}$, the responsibilities at each time-frequency point can be easily computed. Similarly, given knowledge of the responsibilities, it is straightforward to estimate the model parameters. However, because neither of these quantities are generally known in advance, neither can be computed directly. We derive an expectation-maximization algorithm to iteratively learn both the parameters and responsibilities of time-frequency points for each source in a particular mixture.

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

$$z_{i\tau c}(\omega, t) = \frac{P\big(\phi,\, \alpha,\, y^{\ell},\, y^{r},\, i,\, \tau,\, c \,|\, \theta_j\big)}{\sum_{i\tau c} P\big(\phi,\, \alpha,\, y^{\ell},\, y^{r},\, i,\, \tau,\, c \,|\, \theta_j\big)} \tag{18}$$

This is easily computed using the factorization in equation (16).

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

$$\mathcal{L}\big(\theta \,|\, \theta_j\big) = k + \sum_{\omega t} \sum_{i\tau c} z_{i\tau c}(\omega, t)\, \log P\big(\phi,\, \alpha,\, y^\ell,\, y^r,\, i,\, \tau,\, c \,|\, \theta\big) \tag{19}$$

where k is a constant that is independent of θ.

The maximum likelihood model parameters are weighted means of sufficient statistics of the data. First, we define the operator

$$\langle x \rangle_{t,\tau} \triangleq \frac{\sum_{t,\tau} z_{i\tau c}(\omega, t) x}{\sum_{t,\tau} z_{i\tau c}(\omega, t)} \tag{20}$$

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

$$\varsigma_i^2 = \left\langle \tilde{\phi}_\tau^2(\omega, t) \right\rangle_{\omega,t,\tau,c} \tag{21}$$

$$\upsilon_i(\omega) = \langle \alpha(\omega, t) \rangle_{t,\tau,c} \tag{22}$$

$$\eta_i^2(\omega) = \left\langle (\alpha(\omega, t) - \upsilon_i(\omega))^2 \right\rangle_{t,\tau,c} \tag{23}$$

$$\psi_{i\tau} = \frac{1}{\Omega T} \sum_{\omega t c} z_{i\tau c}(\omega, t) \tag{24}$$

Unlike the interaural parameters, the source model parameters are tied across frequency to ensure that each time frame is explained by a single component in the source prior. The updated parameters can be found by solving the following set of simultaneous equations for $\mathbf{w}_i, \mathbf{h}_i^\ell$, and $\mathbf{h}_i^r$:

$$\sum_{tc} U_c^T M_{ict}\, \bar{\Sigma}_c^{-1} \left(2\left(\bar{\boldsymbol{\mu}}_c + U_c \mathbf{w}_i\right) + B\left(\mathbf{h}_i^r + \mathbf{h}_i^\ell\right)\right) = \sum_{tc} U_c^T M_{ict}\, \bar{\Sigma}_c^{-1} \left(\boldsymbol{y}^\ell(t) + \boldsymbol{y}^r(t)\right) \tag{25}$$

$$\sum_{tc} B^T M_{ict}\, \bar{\Sigma}_c^{-1} \left(\bar{\boldsymbol{\mu}}_c + U_c \mathbf{w}_i + B\, \mathbf{h}_i^\ell\right) = \sum_{tc} B^T M_{ict}\, \bar{\Sigma}_c^{-1}\, \boldsymbol{y}^\ell(t) \tag{26}$$

$$\sum_{tc} B^T M_{ict}\, \bar{\Sigma}_c^{-1} \left(\bar{\boldsymbol{\mu}}_c + U_c \mathbf{w}_i + B\, \mathbf{h}_i^r\right) = \sum_{tc} B^T M_{ict}\, \bar{\Sigma}_c^{-1}\, \boldsymbol{y}^r(t) \tag{27}$$

where M_{ict} is a diagonal matrix whose diagonal entries correspond to a soft mask encoding the posterior probability of component c from source i dominating the mixture at frame t:

$$M_{ict} \triangleq \operatorname{diag}\left(\sum_{\tau} \boldsymbol{z}_{i\tau c}(:,t)\right) \tag{28}$$

This EM algorithm is guaranteed to converge to a local maximum of the likelihood surface, but because the total likelihood in equation (17) is not convex, the quality of the solution is sensitive to initialization. We initialize $\psi_{i\tau}$ using an enhanced cross-correlation based localization method while leaving all the other parameters in a symmetric, non-informative state. From those parameters, we compute the first E step mask.

Initial estimates of τ are obtained for each source from the PHAT-histogram (Aarabi, 2002), which estimates the time delay between $\mathrm{x}^{\ell}(t)$ and $\mathrm{x}^{r}(t)$ by whitening the signals and then computing their cross-correlation. Then, $\psi_{i\tau}$ is initialized to be centered at each cross-correlation peak and to fall off away from that. Specifically, $P\big(\tau \,|\, i\big)$, 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. The remaining IPD, ILD, and source model parameters are estimated from the data in the M-step following the initial E-step.

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 the ILD parameters $\boldsymbol{\upsilon}_i$ and $\boldsymbol{\eta}_i$ which are a function of frequency. To address this problem, we use a bootstrapping approach where initial EM iterations are performed with a frequency-independent ILD model, and frequency-dependence is gradually introduced. Note that the number of EM iterations is specified in advance, and is set to 16 in the experiments described in the following section. 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.

Figure 4 shows the interaural parameters estimated from the observations in figure 2 using the EM algorithm described in this section. The algorithm

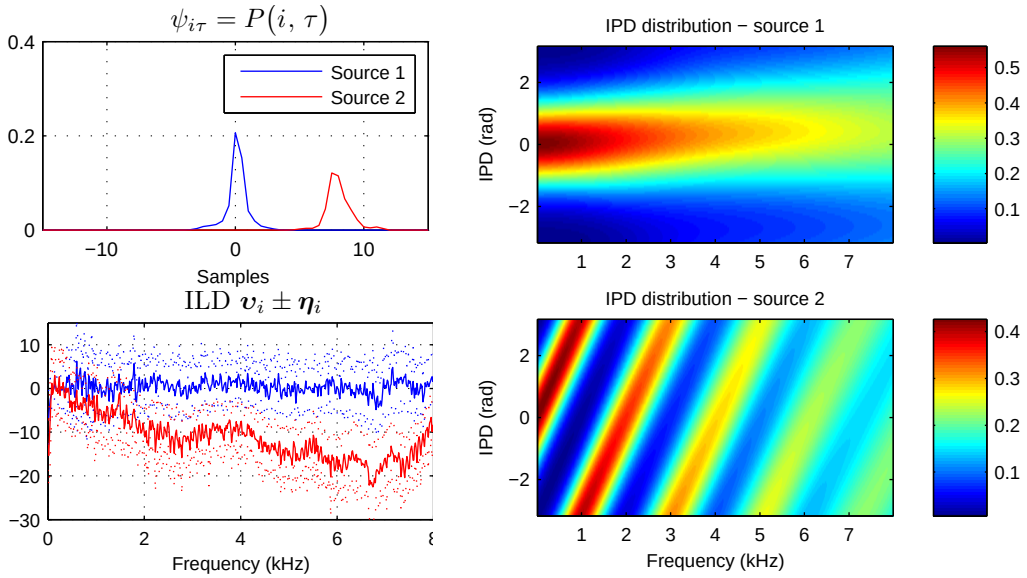

Figure 4: Interaural model parameters estimated by the EM algorithm given the observations shown in figure 2. In the bottom left plot, solid lines indicate υ_i, the mean ILD for each source, while dotted lines indicate $\upsilon_i \pm \eta_i$.

does a good job localizing the sources, as shown in the plot of $\psi_{i\tau}$. The ILD distribution (bottom left) accurately characterizes the true distribution as well. As described earlier, the ILD of the source facing the microphones head on (source 1), is basically flat across the entire frequency range while that of source 2 becomes more negative with frequency. Similarly, the per-source IPD distributions shown in the right hand column closely match the predictions made earlier as well. These distributions consist of a mixture of Gaussians calculated by marginalizing over all possible settings of τ as in equation (8). Since $\psi_{i\tau}$ contains non-zero probability mass for multiple τ settings near the correct location for each source, there is some uncertainty as to the exact source locations. The mixture components are spaced further apart at high frequencies because of their proportionality to ω. This is why the distributions are quite tight at low frequencies, but get gradually broader with increasing frequency.

The estimated source model parameters are shown in figure 5. As with the ILD and IPD parameters, the source model parameters are initialized to an uninformative state. However, as the binaural cues begin to disambiguate

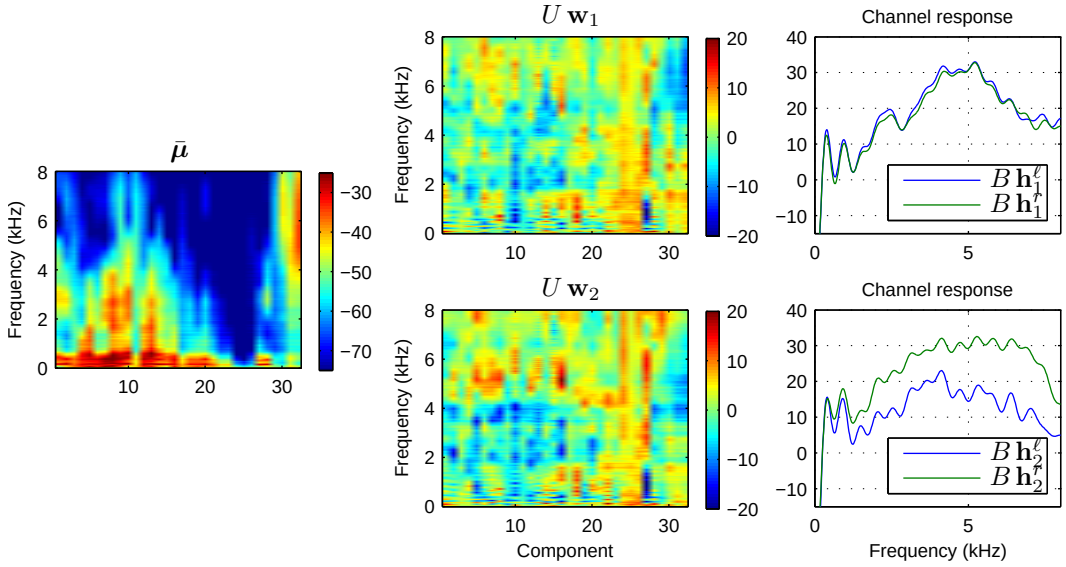

Figure 5: Source model parameters estimated by the EM algorithm given the observations shown in figure 2. The overall model for source i is the sum of the speaker-independent means, $\bar{\boldsymbol{\mu}}$, the source-adapted term $U\mathbf{w}_i$ based on the eigenvoice model of inter-speaker variability, and the channel response at each ear, $B\,\mathbf{h}_i^{\ell,r}$.

the sources, the learned channel responses and source adaptation parameters help to differentiate the source models. By the time the algorithm has converged, the source models have become quite different, with $\mathbf{w}_i$ learning the characteristics unique to each source under the predefined eigenvoice model that are common to both left and right observations, e.g. the increased energy near 6 kHz in many components for source 2. $\mathbf{h}_i^{\ell,r}$ similarly learns the magnitude responses of the filters applied to each channel. Note that the overall shapes of $B\,\mathbf{h}_i^{\ell,r}$ reflect the effects of the HRTFs applied to the source in creating the mixture. These are unique to the particular mixture and were not present in the training data used to learn $\bar{\mu}$ and U. The difference between the channel response at each ear, $B\,\mathbf{h}_i^\ell - B\,\mathbf{h}_i^r$, reflects the same interaural level differences as the ILD parameters, $\boldsymbol{\upsilon}_i$ in figure 4.

Although the parameters learned by the MESSL model are interesting in their own right, they cannot separate the sources directly. After the EM algorithm converges, we derive a time-frequency mask from the posterior

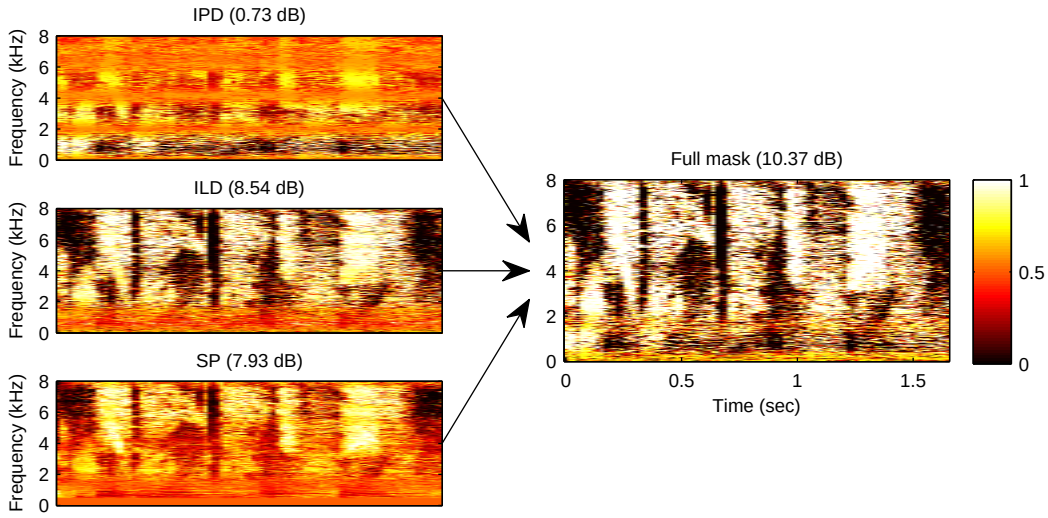

Figure 6: Contribution of the IPD, ILD, and source model to the final mask learned using the full MESSL-EV algorithm on the mixtures from figure 2. The SNR improvement computed using equation (32) is shown in parenthesis.

probability of the hidden variables for each source:

$$\mathrm{M}_i(\omega, t) = \sum_{\tau c} z_{i\tau c}(\omega, t) \tag{29}$$

Estimates of clean source i can then be obtained by multiplying the short-time Fourier transform of each channel of the mixed signal by the mask for the corresponding source. This assumes that the mask is identical for both channels.

$$\hat{\mathrm{X}}_i^{\ell}(\omega, t) = \mathrm{M}_i(\omega, t)\, \mathrm{Y}^{\ell}(\omega, t) \tag{30}$$

$$\hat{\mathrm{X}}_i^{r}(\omega, t) = \mathrm{M}_i(\omega, t)\, \mathrm{Y}^{r}(\omega, t) \tag{31}$$

Figure 6 shows an example mask derived from the proposed algorithm. To demonstrate the contributions of the different types of observations in the signal model to the overall mask, we also plot masks isolating the IPD, ILD, and source models. These masks are found by leaving unrelated terms out of the factored likelihood and computing "marginal" posteriors, i.e. the full model is used to learn the parameters, but in the final EM iteration

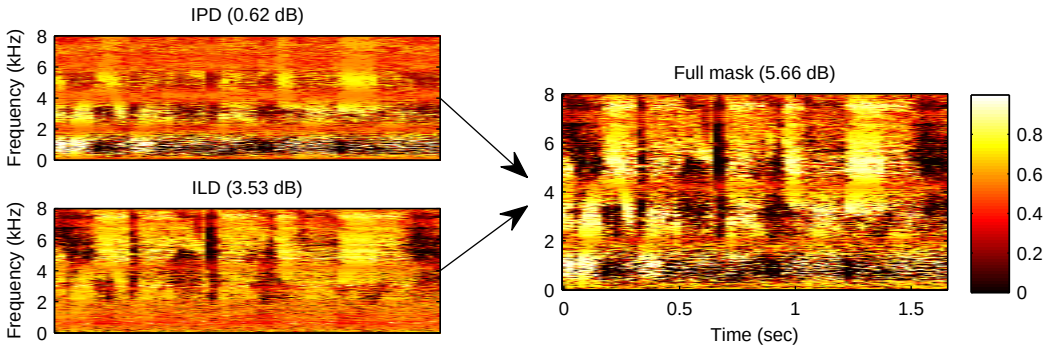

Figure 7: Contribution of the IPD and ILD to the final mask learned using the baseline MESSL separation algorithm using only the interaural signal model on the mixtures from figure 2. The SNR improvement computed using equation (32) is shown in parenthesis.

the contributions of each underlying model to the complete likelihood in equation (16) are treated independently to compute three different posterior distributions.

The IPD and ILD masks make qualitatively different contributions to the final mask, so they serve as a good complement to one another. The IPD mask is most informative in low frequencies, and has characteristic subbands of uncertainty caused by the spatial aliasing described earlier. The ILD mask primarily adds information in high frequencies above 2 kHz and so it is able to fill in many of the regions where the IPD mask is ambiguous. This poor definition in low frequencies is because the per-source ILD distributions shown in figure 4 have significant overlap below 2 kHz. These observations are consistent with the use of the ITD and ILD cues for sound localization in human audition (Wightman and Kistler, 1992).

Finally, the source model mask is qualitatively quite similar to the ILD mask, with some additional detail below 2 kHz. This is not surprising because both the ILD and source models capture related features of the mixed signal. We expect that the additional constraints from the prior knowledge built into the source model should allow for more accurate estimation than the ILD model alone, however it is not clear that this is the case based on this figure.

To better illustrate the contribution of the source model, figure 7 shows the mask estimated from the same data using the baseline MESSL algorithm of Mandel and Ellis (2007) which is based only on the interaural model. The

MESSL mask is considerably less confident (i.e. less binary) than that of the MESSL-EV mask in figure 6. The contribution of the IPD mask is quite similar in both cases. The difference in quality between the two systems is a result of the marked difference in the ILD contributions. The improvement in the MESSL-EV case can be attributed to the addition of the source model, which, although not as informative on its own, is able to indirectly improve the estimation of the ILD parameters. This is because the source model introduces correlations across frequency, that are only loosely captured by the ILD model during initial iterations. This is especially true in the higher frequencies which are highly correlated in speech signals. By modeling each frame with GMM components with a different spectral shape, the source model is able to decide which time-frequency regions are a good fit to each source based on how well the observations in each frame match the source prior distribution. It is able to isolate the sources based on how speech-like they are, using prior knowledge such as the high-pass shape characteristic of fricatives and characteristic resonance structure of vowels, etc. In contrast, the ILD model treats each frequency band independently and is prone to source permutations if poorly initialized. Although the bootstrapping process described earlier alleviates these problems to some extent, the source model's ability to emphasize time-frequency regions consistent with the underlying speech model further reduces this problem and significantly improves the quality of the interaural parameters and thus the overall separation.

5. Experiments

In this section we describe a set of experiments designed to evaluate the performance of the proposed algorithm under a variety of different conditions and compare it to two other well known binaural separation algorithms. We assembled a data set consisting of mixtures of two and three speech signals in simulated anechoic and reverberant conditions. The mixtures were formed by convolving anechoic speech utterances with a variety of different binaural impulse responses. We formed two such data sets, one from utterances from the GRID corpus (Cooke et al., 2006) for which training data was available for the source model, and another using the TIMIT corpus (Garofolo et al., 1993) to evaluate the performance on held out speakers using the GRID source models. Although the TIMIT data set contains speech from hundreds of different speakers, it does not contain enough data to adequately train models for each of these speakers. This makes it a good choice for evaluation of the

eigenvoice adaptation technique. In both cases, we used a randomly selected subset of 15 utterances to create each test set. Prior to mixing, the utterances were passed through a first order pre-emphasis filter to whiten their spectra to avoid overemphasizing the low frequencies in our SNR performance metric.

The anechoic binaural impulse responses came from Algazi et al. (2001), a large effort to record head-related transfer functions for many different individuals. We use the measurements for a KEMAR dummy head with small ears, taken at 25 different azimuths at 0° elevation. The reverberant binaural impulse responses were recorded by Shinn-Cunningham et al. (2005) in a real classroom with a reverberation time of around 565 ms. These measurements were also made with a KEMAR dummy head, although a different unit was used. The measurements we used were taken in the center of the classroom, with the source 1 m from the head at 7 different azimuths, each repeated 3 times.

In the synthesized mixtures, the target speaker was located directly in front of the listener, with distractor speakers located off to the sides. The angle between the target and distractors was systematically varied and the results combined for each direction. In the anechoic setting, there were 12 different angles at which we placed the distractors. In the reverberant setting, there were 6 different angles, but 3 different impulse response pairs for each angle, for a total of 18 conditions. Each setup was tested with 5 different randomly chosen sets of speakers and with one and two distractors, for a total of 300 different mixtures. We measure the performance of separation with signal-to-noise ratio improvement, defined for source i as follows:

$$\mathrm{SNRI}_i = 10\log_{10}\frac{\|\mathrm{M}_i\mathrm{X}_i\|^2}{\|\mathrm{X}_i - \mathrm{M}_i\sum_j \mathrm{X}_j\|^2} - 10\log_{10}\frac{\|\mathrm{X}_i\|^2}{\|\sum_{j\neq i}\mathrm{X}_j\|^2} \qquad (32)$$

where X_i is the clean spectrogram for source i, M_i is the corresponding mask estimated from the mixture, and $\|\cdot\|$ is the Frobenius norm operator. This measure penalizes both noise that is passed through the mask and signal that is rejected by the mask.

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.

We compare the proposed separation algorithms to the two-stage frequency-domain blind source separation system from Sawada et al. (2007) (2S-FD-BSS), the Degenerate Unmixing Estimation Technique from Jourjine et al. (2000); Yilmaz and Rickard (2004) (DUET), and the performance using ground truth binary masks derived from oracle knowledge of the clean source signals. The ground truth mask for source i is set to 1 for regions of the spectrogram dominated by that source, i.e. regions with a local SNR greater than 0 dB, and set to zero elsewhere. It represents the ideal binary mask (Wang, 2005) and serves as an upper bound on separation performance.

We also compare three variants of our system: the full MESSL-EV algorithm described in the previous section, the MESSL-SP algorithm from Weiss et al. (2008) that uses a speaker-independent source prior distribution (identical to MESSL-EV but with $\mathbf{w}_i$ fixed at zero), and the baseline MESSL algorithm from Mandel and Ellis (2007) that does not utilize source constraints at all. The MESSL-SP system uses a 32 mixture component, speaker-independent model trained over data from all 34 speakers in the GRID data set. Similarly, the MESSL-EV system uses a 32 component eigenvoice speech model source GMMs trained over all 34 speakers. All 33 eigenvoice bases were retained. Figure 8 shows example masks derived from these systems.

DUET 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 difference are constant at all frequencies and that there is no spatial aliasing, conditions which can be met to a large degree with free-standing microphones close to one another. With dummy head recordings, however, the interaural level difference 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 the sources more difficult to isolate. As shown in figure 8, this manifests itself as poor estimation in frequencies above 4 kHz which the algorithm overwhelmingly assigns to a single source, and spatial aliasing in subbands around 1 and 4 kHz.

The 2S-FD-BSS system uses 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

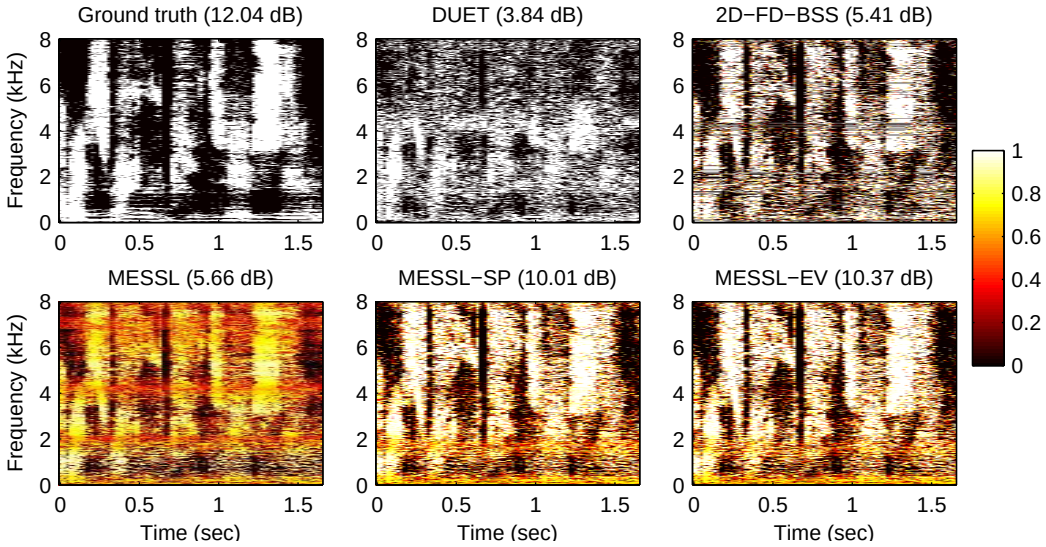

Figure 8: Example binary masks found using the different separation algorithms evaluated in section 5. The mixed signal is composed of two GRID utterances in reverberation separated by 60 degrees.

probabilistic model of mixing coefficients. In the second stage, the sources in different bands are associated with the corresponding signals from other bands using k-means clustering on the posterior probabilities of each source and then further 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. Such second stage errors are visible in the same regions where spatial aliasing causes confusion for the other separation algorithms in the 2S-FD-BSS mask shown in figure 8. In general, such errors tend to happen at low frequencies, where adjacent bands are less well-correlated. In contrast, the failure mode of the MESSL variants is to pass both sources equally when it is unable to sufficiently distinguish between them. This is clearly visible in the regions of the MESSL mask in figure 8 that have posteriors close to 0.5. As a result 2S-FD-BSS is more prone to source permutation errors where significant target energy can be rejected by the mask.

System	A2	R2	A3	R3	Avg
Ground truth	11.83	11.58	12.60	12.26	12.07
MESSL-EV	8.79	**7.85**	**8.20**	**7.54**	**8.09**
MESSL-SP	6.30	7.39	7.08	7.18	6.99
MESSL	7.21	4.37	6.17	3.56	5.33
2S-FD-BSS	**8.91**	6.36	7.94	5.99	7.30
DUET	2.81	0.59	2.40	0.86	1.67

Table 1: Average SNR improvement (in dB) across all distractor angles on mixtures created from the GRID data set. The test cases are described by the number of simultaneous sources (2 or 3) and whether the impulse responses were anechoic or reverberant (A or R).

System	A2	R2	A3	R3	Avg
Ground truth	3.41	3.38	3.10	3.04	3.24
MESSL-EV	**3.00**	**2.65**	**2.32**	**2.24**	**2.55**
MESSL-SP	2.71	2.62	2.22	2.22	2.44
MESSL	2.81	2.39	2.15	1.96	2.33
2S-FD-BSS	2.96	2.50	2.28	2.04	2.44
DUET	2.56	2.03	1.85	1.53	1.99
Mixture	2.04	2.04	1.60	1.67	1.84

Table 2: Average PESQ score (mean opinion score) across all distractor angles on mixtures created from the GRID data set.

5.1. GRID performance

The average performance of the evaluated algorithms on the GRID data set is summarized in tables 1 and 2 using the SNR improvement and PESQ metrics, respectively. Broadly speaking, all algorithms perform better in anechoic conditions than in reverberation and on mixtures of two sources than on mixtures of three sources under both metrics. In most cases MESSL-EV performs best, followed by MESSL-SP and 2S-FD-BSS. 2S-FD-BSS outperforms MESSL-SP in anechoic conditions, however, in reverberation, this trend is reversed and 2S-FD-BSS performs worse. Both of the MESSL variants perform significantly better than the MESSL baseline for reasons described in the previous section. The addition of speaker adaptation in MESSL-EV gives an overall improvement of about 1.1 dB over MESSL-SP and 2.8 dB over MESSL in SNR improvement on average. 2S-FD-BSS generally

performs better than MESSL, but not as well as MESSL-SP and MESSL-EV. The exception is on mixtures of two sources in anechoic conditions where 2S-FD-BSS performs best overall in terms of SNR improvement. Finally, DUET performs worst, especially in reverberation where the IPD/ILD histograms are more diffuse, making it difficult to accurately localize the sources.

We note that unlike the initial results reported in Weiss et al. (2008) MESSL-SP does not perform worse than MESSL on anechoic mixtures. The problems in Weiss et al. (2008) were caused by the channel parameters $\mathbf{h}_i^{\ell,r}$ over-fitting which led to source permutations. To fix this problem in the results reported here, we used a single, flat channel basis for the channel parameters in anechoic mixtures. In reverberant mixtures 30 DCT bases were used.

The poor performance of some of the MESSL systems in table 1 on anechoic mixtures is a result of poor initialization at small distractor angles. An example of this effect can be seen in the left column of figure 9 where the MESSL systems have very poor performance compares to 2S-FD-BSS when the sources are separated by 5 degrees. However, as the sources get further apart, the performance of all of the MESSL systems improves dramatically. The very poor performance at very small angles heavily skews the averages in table 1. This problem did not affect MESSL's performance on reverberant mixtures because the minimum separation between sources on that data is 15 degrees and the initial localization used to initialize MESSL was adequate. Finally, 2S-FD-BSS was unaffected by this problem at small distractor angles because, unlike the other systems we evaluated, it does not directly utilize the spatial locations for separation.

MESSL, 2S-FD-BSS, and DUET all perform significantly better on anechoic mixtures that on reverberant mixtures because the lack of noise from reverberant echoes makes anechoic sources much easier to localize. As described in the previous section, the additional constraints from the source models in MESSL-EV and MESSL-SP help to resolve the ambiguities in the interaural parameters in reverberation so the performance of these systems does not degrade nearly as much in reverberation. In reverberation MESSL-EV and MESSL-SP both improve over the MESSL baseline by over 3 dB. The added benefit from the speaker adaptation in MESSL-EV is limited in reverberation, but is significant in anechoic mixtures. This is likely a result of the fact that the EV model has more degrees of freedom to adapt to the observation. The MESSL-SP system can only adapt a single parameter per source in anechoic conditions due the limited model of channel variation described

System	A2	R2	A3	R3	Avg
Ground truth	12.09	11.86	12.03	11.84	11.95
MESSL-EV	10.08	**8.36**	**8.21**	**7.22**	**8.47**
MESSL-SP	10.00	8.10	7.97	6.96	8.26
MESSL	9.66	5.83	7.12	4.32	6.73
2S-FD-BSS	**10.29**	7.09	6.17	4.86	7.10
DUET	3.87	0.59	3.63	0.62	2.18

Table 3: Average SNR improvement (in dB) across all distractor angles on mixtures created from the TIMIT data set. The test cases are described by the number of simultaneous sources (2 or 3) and whether the impulse responses were anechoic or reverberant (A or R).

System	A2	R2	A3	R3	Avg
Ground truth	3.35	3.33	3.06	3.02	3.19
MESSL-EV	2.99	**2.52**	**2.30**	**2.11**	**2.48**
MESSL-SP	2.98	2.50	2.28	2.10	2.47
MESSL	2.92	2.33	2.24	1.96	2.36
2S-FD-BSS	**3.07**	2.36	1.91	1.76	2.28
DUET	2.59	1.85	2.01	1.48	1.98
Mixture	1.96	1.92	1.53	1.62	1.76

Table 4: Average PESQ score (mean opinion score) across all distractor angles on mixtures created from the TIMIT data set.

above. Finally, we note that the addition of the source model in MESSL-EV and MESSL-SP is especially useful in underdetermined conditions (i.e. A3 and R3) because of the source model's ability to emphasize time-frequency regions consistent with the underlying speech model which would otherwise be ambiguous. In two source mixtures this effect is less significant because the additional clean glimpses of each source allow for more robust estimation of the interaural parameters.

5.2. TIMIT performance

Tables 3 and 4 show the performance of the different separation algorithms on the data set derived from TIMIT utterances. The trends are very similar to those seen in the GRID data set, however performance in general tends to be a bit better in terms of SNR improvement. This is probably because the TIMIT utterances are longer than the GRID utterances, and the additional

Data set	System 1 – System 2	A2	R2	A3	R3	Avg
GRID	MESSL-EV – MESSL	1.58	3.46	2.03	3.98	2.76
	MESSL-EV – MESSL-SP	2.49	0.48	1.12	0.36	1.10
TIMIT	MESSL-EV – MESSL	0.42	2.53	1.09	2.89	1.74
	MESSL-EV – MESSL-SP	0.08	0.26	0.24	0.26	0.21

Table 5: Comparison of the relative performance in terms of dB SNR improvement of MESSL-EV to the MESSL baseline and MESSL-SP on both the GRID data set where the source models are matched to the test data, and on the TIMIT data set where the source models are mismatched to the test data.

observations lead to more robust localization which in turn leads to better separation. The main point to note from the results in table 3 is that the performance improvement of MESSL-EV over the other MESSL variants is significantly reduced when compared to the GRID experiments. This is because of the mismatch between the mixtures and the data used to train the models. However, despite this mismatch, the performance improvement of the MESSL variants that incorporate a prior source model still show a significant improvement over MESSL.

The performance of MESSL-EV relative to the other MESSL variants on both data sets is compared in table 5. On the matched data set, MESSL-EV outperforms MESSL by an average of about 2.8 dB and also outperforms MESSL-SP by an average of 1.1 dB. However, on the mismatched data set the improvement of MESSL-EV is significantly reduced. In fact, the improvement of MESSL-EV over MESSL-SP on this data set is only 0.2 dB on average. This implies that the eigenvoice model of speaker variation is significantly less informative when applied to speakers that are very different from those in the train set. The bulk of MESSL-EV's improvement is therefore due to the speaker-independent portion of the model which is still a good enough model for speech signals in general to improve performance over MESSL, even on mismatched data.

The small improvement in the performance of MESSL-EV when the training and test data are severely mismatched is the result of a number of factors. The primary problem is that a relatively small set of speakers were used to train the GRID eigenvoice bases. In order to adequately capture the full subspace of speaker variation and generalize well to held-out speakers, data from a large number of training speakers, on the order of a few hundred,

are typically required (Weiss, 2009). In these experiments, training data was only available for 34 different speakers.

This lack of diversity in the training data is especially relevant because of the significant differences between the GRID and TIMIT speakers. The speakers in the GRID data set were all speaking British English while TIMIT consists of a collection of American speakers. There are significant pronunciation differences between the two dialects, e.g. British English is generally non-rhotic, which lead to signification differences in the acoustic realizations of common speech sounds and therefore differences between the corresponding speech models. These differences make it impossible to fully capture the nuances of the other dialect without including some speakers of both dialects in the training set. Finally, the likelihood that the eigenvoice model will generalize well to capture speaker-dependent characteristics across both data sets is further decreased because the models themselves were quite small, consisting of only 32 mixture components.

5.3. Performance at different distractor angles

Finally, the results on the TIMIT set are shown as a function of distractor angle in figure 9. Performance of all algorithms generally improves when the sources are better separated in space. In anechoic mixtures of two sources the MESSL variants all perform essentially as well as ground truth masks when the sources are separated by more than 40°. None of the systems are able to approach ideal performance under the other conditions. As noted earlier, 2S-FD-BSS performs best on 2 source anechoic mixtures in tables 1 and 3. As seen in figure 9 this is mainly an effect of very poor performance of the MESSL systems on mixtures with small distractor angles. All MESSL variants outperform 2S-FD-BSS when the are separated by more than about 20°. The poor performance of MESSL when the sources are separated by 5° is a result of poor initialization due to the fact that localization is difficult because the parameters for all sources are very similar. This is easily solved by using better initialization. In fact, it is possible to effectively combine the strengths of both of the ICA and localization systems by using the mask estimated by 2S-FD-BSS to initialize the MESSL systems. This would require starting the separation algorithm with the M-step instead of the E-step as described in section 4, but the flexibility of our model's EM approach allows this. We leave the investigation of the combination of these techniques as future work.

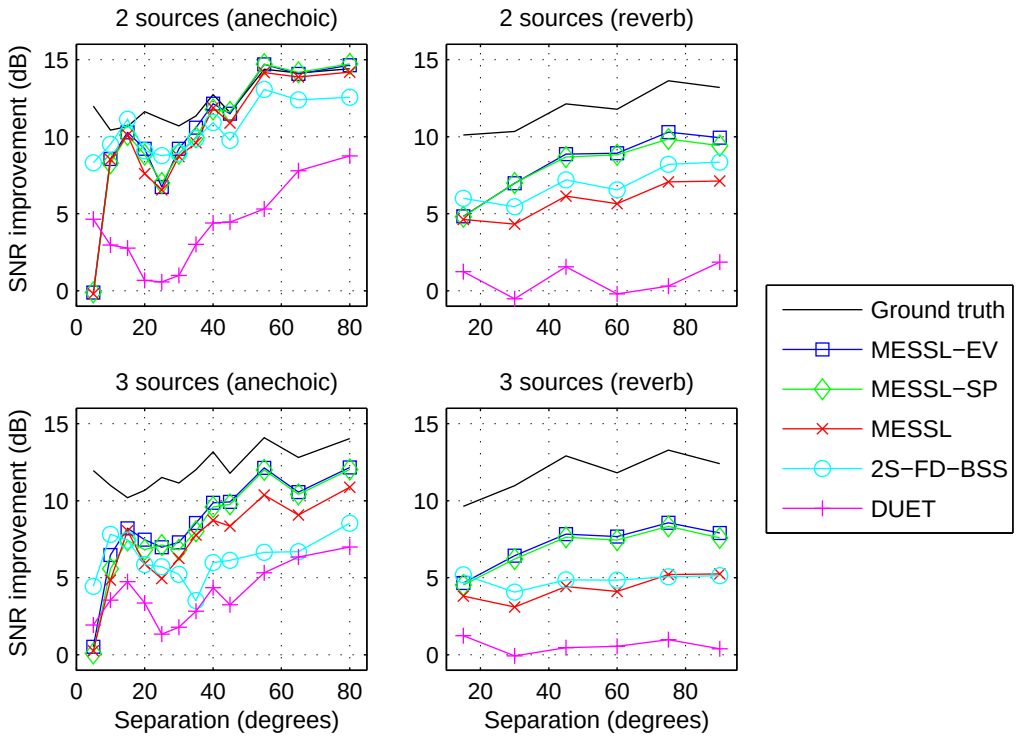

Figure 9: Separation performance on the TIMIT data set as a function of distractor angle.

This dependence on spatial localization for adequate source separation highlights a disadvantage of the MESSL family of algorithms, especially as compared to model-based binaural separation algorithms that use factorial model combination (Rennie et al., 2003; Wilson, 2007). As seen in the examples of figures 6 and 7, in MESSL-SP and MESSL-EV the source model is used to help disambiguate uncertainties in the interaural localization model. It does not add any new information about the interaction between the two sources and can only offer incremental improvements over the MESSL baseline. Therefore the addition of the source model does not improve performance when the sources are located very close to each other in space.

In contrast, in Rennie et al. (2003) and Wilson (2007), the factorial source model is used to model the interaction between the sources directly. In these algorithms, the localization cues are used to disambiguate the source model, which, on its own is inherently ambiguous because identical, speaker-

independent models are used for all sources. This makes it impossible for the models to identify which portions of the signal are dominated by each source without utilizing the fact that they arrive from distinct spatial locations. These algorithms therefore suffer from similar problems to MESSL at very small distractor angles where the localization cues are similar for all sources. However, this could be overcome by incorporating additional knowledge about the differences between the distributions of each source signal through the use of speaker-dependent models, or model adaptation as described in this paper. When the sources are close together the binaural separation problem reduces to that of monaural separation, where factorial model based techniques using source-dependent or -adapted models have been very successful (Weiss and Ellis, 2010). MESSL-EV, however, still suffers at small distractor angles despite utilizing source-adapted models.

The advantage of MESSL-EV over the factorial model approach to combining source models with localization cues is that it enables efficient inference because it is not necessary to evaluate all possible model combinations. This is because each time-frequency cell is assumed to be conditionally independent given the latent variables. Because each frequency band is independent given a particular source and mixture component, the sources decouple and all combinations need not be considered. This becomes especially important for dense mixtures of many sources. As the number of sources grows, the factorial approach scales exponentially in terms of the number of Gaussian evaluations required (Roweis, 2003). In contrast, the computational complexity of the algorithms described in this paper scale linearly in the number of sources.

6. Summary

We have presented a system for source separation based on a probabilistic model of binaural observations. A model of the interaural spectrogram that is independent of the source signal is combined with a prior model of the statistics of the underlying anechoic source spectrogram to obtain a hybrid localization and source model based separation algorithm. The joint model explains each point in the mixture spectrogram as being generated by a single source, with a spatial location consistent with a particular time-delay drawn from a set of candidate values, and whose underlying source signal is generated by a particular mixture component in the prior source model. The computational complexity therefore scales linearly in each of these parameters, since the posterior distribution shown in equation (18) takes all possible combinations

of the source, candidate time-delay, and source prior hidden variables into account. Despite the potentially large number of hidden variables, the scaling behavior is favorable compared to separation algorithms based on factorial model combination.

Like other binaural separation algorithms which can separate underdetermined mixtures, the separation process in the proposed algorithm is based on spectral masking. The statistical derivation of MESSL and the variants described in the paper represents an advantage when compared to other algorithms in this family, most of which are constructed based on computational auditory scene analysis heuristics which are complex and difficult to implement.

In the experimental evaluation, we have shown that the proposed model is able to obtain a significant performance improvement over the algorithm that does not rely on a prior source model and another state of the art source separation algorithms based on frequency domain ICA. The improvement is substantial even when the prior on the source statistics is quite limited, consisting of a small speaker-independent model. In this case, the sources are differentiated through the source-specific channel model which compensates for the binaural room impulse responses applied to each of the source signals. Despite the fact that the proposed algorithm does not incorporate an explicit model of reverberation, we have shown that the additional constraints derived from the anechoic source model are able to significantly improve performance in reverberation. The investigation of an model extensions similar to Palomäki et al. (2004) which compensate for early echoes to remove reverberant noise remains as future work.

Finally, we have shown that the addition of source model adaptation based on eigenvoices can further improve performance under some conditions. The performance improvements when using source adaptation are largest when the test data comes from the same sources as were used to train the model. However, when the training and test data are severely mismatched, the addition of source adaptation only boosts performance by a small amount.

7. Acknowledgments

This work was supported by the NSF under Grants No. IIS-0238301 and IIS-0535168, and by EU project AMIDA. 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 Sponsors.

References

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

Algazi, V. R., Duda, R. O., Thompson, D. M., Avendano, C., Oct. 2001. The CIPIC HRTF database. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics. pp. 99–102.

Blauert, J., 1997. Spatial Hearing: Psychophysics of Human Sound Localization. MIT Press.

Cherry, E. C., 1953. Some experiments on the recognition of speech, with one and with two ears. Journal of the Acoustical Society of America 25 (5), 975–979.

Cooke, M., Hershey, J. R., Rennie, S. J., 2010. Monaural speech separation and recognition challenge. Computer Speech and Language 24 (1), 1 – 15.

Cooke, M. P., Barker, J., Cunningham, S. P., Shao, X., 2006. An audio-visual corpus for speech perception and automatic speech recognition. Journal of the Acoustical Society of America 120, 2421–2424.

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

Harding, S., Barker, J., Brown, G. J., 2006. Mask estimation for missing data speech recognition based on statistics of binaural interaction. IEEE Transactions on Audio, Speech, and Language Processing 14 (1), 58–67.

Jourjine, A., Rickard, S., Yilmaz, O., Jun. 2000. Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Vol. 5. pp. 2985–2988.

Kuhn, R., Junqua, J., Nguyen, P., Niedzielski, N., Nov. 2000. Rapid speaker adaptation in eigenvoice space. IEEE Transations on Speech and Audio Processing 8 (6), 695–707.

Loizou, P., 2007. Speech enhancement: theory and practice. CRC press Boca Raton: FL:.

Mandel, M. I., Ellis, D. P. W., Oct. 2007. EM localization and separation using interaural level and phase cues. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). pp. 275–278.

Mandel, M. I., Weiss, R. J., Ellis, D. P. W., Feb. 2010. Model-based expectation-maximization source separation and localization. IEEE Transactions on Audio, Speech, and Language Processing 18 (2), 382–394.

Nix, J., Hohmann, V., 2006. Sound source localization in real sound fields based on empirical statistics of interaural parameters. Journal of the Acoustical Society of America 119 (1), 463–479.

Palomäki, K., Brown, G., Wang, D., 2004. A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. Speech Communication 43 (4), 361–378.

Rennie, S., Aarabi, P., Kristjansson, T., Frey, B. J., Achan, K., 2003. Robust variational speech separation using fewer microphones than speakers. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Vol. 1. pp. I – 88–91.

Rennie, S. J., Achan, K., Frey, B. J., Aarabi, P., 2005. Variational speech separation of more sources than mixtures. In: Proc. Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS). pp. 293–300.

Roman, N., Wang, D., 2006. Pitch-based monaural segregation of reverberant speech. Journal of the Acoustical Society of America 120 (1), 458–469.

Roman, N., Wang, D., Brown, G. J., 2003. A classification-based cocktail party processor. In: Advances in Neural Information Processing Systems.

Roweis, S. T., 2003. Factorial models and refiltering for speech separation and denoising. In: Proc. Eurospeech. pp. 1009–1012.

Sawada, H., Araki, S., Makino, S., Oct. 2007. A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). pp. 139–142.

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

Wang, D., 2005. On ideal binary mask as the computational goal of auditory scene analysis. Springer, Ch. 12, pp. 181–197.

Weiss, R. J., 2009. Underdetermined Source Separation Using Speaker Subspace Models. Ph.D. thesis, Department of Electrical Engineering, Columbia University.

Weiss, R. J., Ellis, D. P. W., Jan. 2010. Speech separation using speaker-adapted eigenvoice speech models. Computer Speech and Language 24 (1), 16–29, Speech Separation and Recognition Challenge.

Weiss, R. J., Mandel, M. I., Ellis, D. P. W., Sep. 2008. Source separation based on binaural cues and source model constraints. In: Proc. Interspeech. Brisbane, Australia, pp. 419–422.

Wightman, F. L., Kistler, D. J., 1992. The dominant role of low-frequency interaural time differences in sound localization. Journal of the Acoustical Society of America 91 (3), 1648–1661.

Wilson, K., 2007. Speech source separation by combining localization cues with mixture models of speech spectra. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. I–33–36.

Yilmaz, O., Rickard, S., Jul. 2004. Blind separation of speech mixtures via time-frequency masking. IEEE Transactions on Signal Processing 52 (7), 1830–1847.