Chapter 3
Multichannel Spatial Clustering Using Model-Based Source Separation

Michael I. Mandel and Jon P. Barker

Abstract Recent automatic speech recognition results are quite good when the training data is matched to the test data, but much worse when they differ in some important regard, like the number and arrangement of microphones or the reverberation and noise conditions. Because these configurations are difficult to predict a priori and difficult to exhaustively train over, the use of unsupervised spatial-clustering methods is attractive. Such methods separate sources using differences in spatial characteristics, but do not need to fully model the spatial configuration of the acoustic scene. This chapter will discuss several approaches to unsupervised spatial clustering, with a focus on model-based expectation maximization source separation and localization (MESSL). It will discuss the basic two-microphone version of this model, which clusters spectrogram points based on the relative differences in phase and level between pairs of microphones, its generalization to more than two microphones, and its use to drive minimum variance distortionless response (MVDR) beamforming. These systems are evaluated for speech enhancement as well as automatic speech recognition, for which they are able to reduce word error rates by between 9.9 and 17.1% relative over a standard delay-and-sum beamformer in mismatched train–test conditions.

3.1 Introduction

While automatic speech recognition (ASR) systems using deep neural networks (DNNs) as acoustic models have recently provided remarkable improvements in recognition performance [23], their discriminative nature makes them prone to overfitting the conditions used to train them. For example, in the recent REVERB challenge [27], far-field multichannel ASR systems consistently performed more accurately in the simulated conditions that matched their training than in the real recording conditions that did not. In order to address generalization, DNN acoustic

M.I. Mandel (✉)
Brooklyn College (CUNY), 2900 Bedford Ave, Brooklyn, NY 11210, USA
e-mail: mim@sci.brooklyn.cuny.edu

J.P. Barker
University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, UK

S. Watanabe et al. (eds.), *New Era for Robust Speech Recognition*,
DOI 10.1007/978-3-319-64680-0_3

models should be trained on data that reflect the conditions in which the model will be operating. One common approach to such training is the use of multicondition data [32], in which the recognizer is trained on speech mixed with many different kinds of noise, in the hope that the noise at test time will resemble one of the training noises. Multicondition training provides benefits for both Gaussion-mixture-model (GMM)- and DNN-based acoustic models [39]. DNN enhancement systems can similarly be trained explicitly to generalize across source positions for a fixed microphone array [24], or even to generalize across microphone spacings in linear arrays [47].

While explicit generalization to new spatial configurations of microphones, sources, and rooms is expensive to include in discriminative training procedures, it can be naturally factored out of the data through beamforming. Traditional beamforming assumes a known array geometry, which hinders generalization to new conditions, but unsupervised localization-based clustering avoids this assumption. Successful systems of this type have been introduced for two-microphone separation [37, 45, 61], and in larger ad hoc microphone arrays for localization [33], calibration [18], and construction of time–frequency (T–F) masks [4]. It can be applied to distributed microphone arrays [22], but this chapter describes three similar systems for performing unsupervised spatial clustering and beamforming with compact microphone arrays [29, 37, 49].

These spatial-clustering approaches are based on the idea of time–frequency masking, a technique for suppressing unwanted sound sources in a mixture by applying different attenuations to different T–F points in a spectrogram [58]. The time–frequency masking technique is also discussed in Chap. 2. Clustering T–F points results in groups of points with similar spatial characteristics. Arranging the weight of each T–F point's membership in each group results in a T–F mask that can be used to isolate an individual source. This mask-based approach is in contrast to traditional approaches to blind source separation (BSS), which aim to model all sources at all time–frequency points. A good overview of BSS methods for audio is presented in [57], including various types of additional information that can be utilized to aid more in the source separation process.

3.2 Multichannel Speech Signals

Let a signal of interest in the time domain be denoted by $x_1[n]$. If it is recorded with $I-1$ other signals, $x_i[n]$, at J microphones, with $y_j[n]$ the signal at the jth microphone, then

$$y_j[n] = \sum_{i=1}^{I} \sum_{l=1}^{L} h_{ij}[l] x_i[n-l] + u_j[n] \tag{3.1}$$

$$= \sum_{i=1}^{I} (h_{ij} * x_i)[n] + u_j[n] \tag{3.2}$$

where $h_{ij}[n]$ is the impulse response between source i and microphone j and $u_j[n]$ are noise terms. In the time–frequency domain, assuming that impulse responses are shorter than the Fourier transform analysis window, this relation becomes

$$Y_j(t,f) = \sum_{i=1}^{I} H_{ij}(f)X_i(t,f) + U_j(t,f), \tag{3.3}$$

where $Y_j(t,f)$, $H_{ij}(f)$, $X_i(t,f)$, and $U_j(t,f)$ are all complex scalar values.

The impulse responses $H_{ij}(f)$ capture the communication channel between source and microphone, which includes all of the paths that the sound from the source can take to get to the microphone. This includes the direct sound path, paths coming from a distinct direction that have experienced one or more specular reflections off walls, and paths that come from no distinct direction after having bounced or scattered off many walls, resulting in diffuse reverberation. In general, this channel is time-varying, but many models, including the spatial-clustering methods described below, make the assumption that it is time-invariant, i.e., that the sources, microphones, and reflectors are fixed in space.

3.2.1 Binaural Cues Used by Human Listeners

Human listeners are able to attend to and understand the speech of a talker of interest even when it co-occurs with speech from several competing talkers. They are able to do this using certain cues from individual impulse responses, but mainly by utilizing differences between impulse responses of the same source at the two ears [38]. By comparing the two observed signals to each other, it is easier to differentiate between the effects of the original sound source and the channel on the observations. Performing this same task on single-channel observations requires a strong prior model of the sound source, the channel, or both.

The difference between two microphone channels that human listeners utilize comes from the ratio of two complex spectrograms,

$$C_{jj'}(t,f) = \frac{Y_j(t,f)}{Y_{j'}(t,f)}. \tag{3.4}$$

The log magnitude of this quantity is known as the interaural level difference (ILD),

$$\alpha_{jj'}(t,f) = 20\log_{10}|C_{jj'}(t,f)| = 20\log_{10}|Y_j(t,f)| - 20\log_{10}|Y_{j'}(t,f)|. \tag{3.5}$$

When $Y_j(t,f)$ and $Y_{j'}(t,f)$ are dominated by the contribution of a single source, $X_{i^*}(t,f)$, where i^* is the index of that dominant source at (t,f),

$$\alpha_{jj'}(t,f) \approx 20\log_{10}|H_{i^*j}(t,f)||X_{i^*}(t,f)| - 20\log_{10}|H_{i^*j'}(t,f)||X_{i^*}(t,f)| \quad (3.6)$$

$$= 20\log_{10}|H_{i^*j}(t,f)| - 20\log_{10}|H_{i^*j'}(t,f)|. \quad (3.7)$$

Note that this quantity is entirely independent of the source signals $X_{i^*}(t,f)$ because it is common to both channels. The property of a single source dominating each individual time–frequency point is known as W-disjoint orthogonality [61], and has been observed in binaural recordings of anechoic speech signals. In addition, in single-channel source separation systems, the log magnitude of a mixture of signals is commonly approximated as the log magnitude of the most energetic signal [46], known as the log-max approximation. This approximation holds for multiple channels as well, as long as the same source is the loudest in all channels. Both of these approximations support the idea of a single source dominating each time–frequency point, which will be used heavily in the remainder of this chapter. Note that different points can be dominated by different sources, so that each source has a set of points at which it dominates all other sources, including the noise.

The phase of $C_{jj'}(t,f)$ is known as the interaural phase difference (IPD). Again assuming W-disjoint orthogonality,

$$\phi_{jj'}(t,f) = \angle C_{jj'}(t,f) = \angle Y_j(t,f) - \angle Y_{j'}(t,f) + 2\ell\pi \quad (3.8)$$

$$\approx \angle H_{i^*j}(t,f)X_{i^*}(t,f) - \angle H_{i^*j'}(t,f)X_{i^*}(t,f) + 2\ell\pi \quad (3.9)$$

$$= \angle H_{i^*j}(t,f) - \angle H_{i^*j'}(t,f) + 2\ell'\pi, \quad (3.10)$$

where ℓ and ℓ' are integers that capture the 2π ambiguity in phase measurements. In the case that $h_{j'}[n]$ is related to $h_j[n]$ by a pure delay of $\Delta_{\tau_{jj'}}$ samples, then through basic Fourier transform properties

$$h_{ij'}[n] = h_{ij}[n] * \delta[n - \Delta_{\tau_{jj'}}], \quad (3.11)$$

$$\therefore\ \phi_{jj'}(t,f) = \angle\exp(\iota 2\pi f\Delta_{\tau_{jj'}}/f_s) = 2\pi f\Delta_{\tau_{jj'}}/f_s + 2\pi\ell, \quad (3.12)$$

where $\iota = \sqrt{-1}$ is the imaginary unit, ℓ is an unknown integer, and f_s is the sampling rate. This pure delay, $\Delta_{\tau_{jj'}}$, is known as the interaural time difference (ITD) and models the nonzero time it takes for a sound to physically traverse a microphone array. As can be seen in (3.12), this ITD corresponds to an IPD that increases linearly with frequency.

For a pure delay between two microphones, $\Delta_{\tau_{jj'}} = f_s d_{jj'}/c$, where c is the speed of sound in air, approximately 340 m/s, and $d_{jj'}$ is the distance between microphones j and j'. When $\ell = 0$, it is trivial to map from an observed IPD to an unobserved ITD using $\Delta_{\tau_{jj'}} = \phi_{jj'}f_s/2\pi f$. This is only possible when $f < c/2d_{jj'}$. For frequencies close to or above this critical value, it is less straightforward to map from IPD to ITD because ℓ must be estimated in some way, a problem known as spatial aliasing [16, 43]. This estimation becomes more difficult in the presence of noise and reverberation. For human listeners, ITD can be measured directly in an anechoic chamber to establish the critical frequency. For a set of 45 subjects, [2] found the average maximum ITD to be 646 µs, corresponding to a distance of 22.0 cm, for which spatial aliasing begins at approximately 800 Hz. Thus this problem is clearly relevant to the processes of source localization and separation in humans.

Figure 3.1 shows example interaural parameters for a recording from the CHiME-3 dataset [6], which collected speech interactions with a six-microphone tablet device in several noisy environments. The top row shows log magnitude spectrograms for the individual channels, 0–3. The microphone for channel 0 was located very close to the talker's mouth, so has a much higher signal-to-noise ratio than the other channels. This can be seen from the lower noise level, after the entire mixture has been attenuated to maintain a consistent speech level. The microphone for channel 2 was facing away from the talker on the back side of the tablet device, leading to a much lower signal-to-noise ratio than the other channels. This can be seen in the lower speech levels relative to channels 1 and 3.

These differences in the levels of the speech and noise signals in each channel lead to characteristic ILDs between pairs of channels. For example, between channels 1 and 2, there is a clear ILD for the speech that distinguishes it from the noise. The IPD, on the other hand, does not discriminate between them. For channels 1 and 3, differences in time of arrival cause the IPD to be much more useful in discrimination between target and noise than the ILD is. Spatial-clustering systems take advantage of these differences to identify time–frequency points from the same source and group them together.

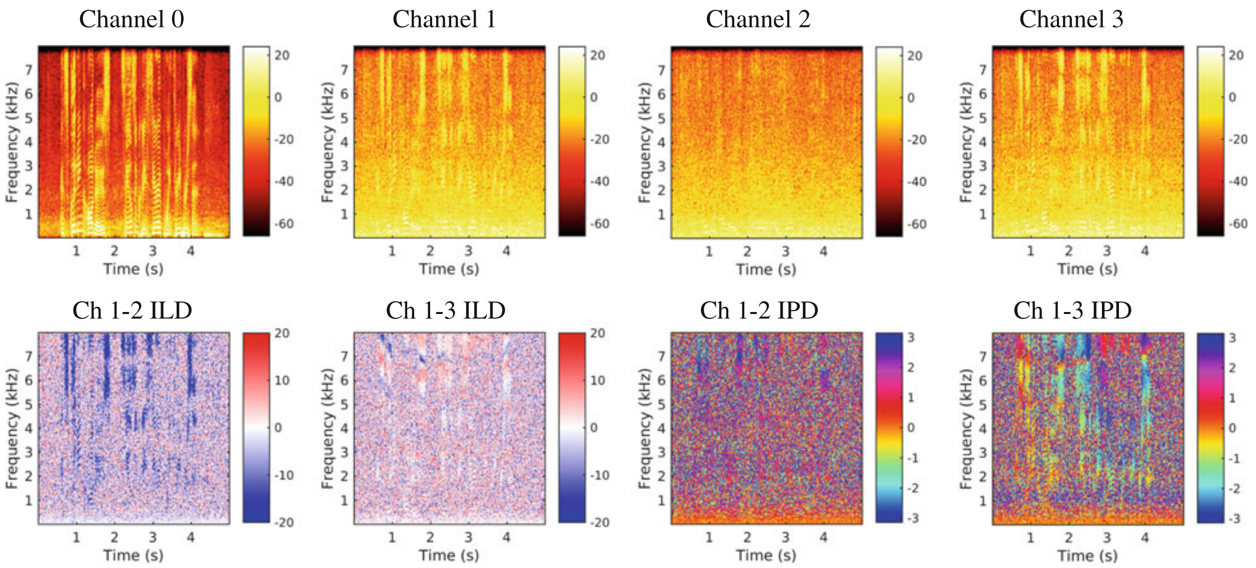
3.2.2 Parameters for More than Two Channels

For recordings with more than two channels, the spatial parameters can be generalized in two ways. The first is a direct generalization of the interaural computation, arranging the $C_{jj'}(t,f)$ terms from (3.4) into a matrix, $\mathbf{C}(t,f)$. In this matrix, the phase term is the difference between the phases in channels j and j' at time–frequency point (t,f), and the log magnitude is the difference between the log magnitudes at that point.

The linearly constrained minimum variance (LCMV) beamformer [8] uses a slightly different matrix to characterize the relationship between the microphone channels. In particular, it uses quantities such as the spatial covariance matrices

$$\Phi_{UU}(f) = \mathbb{E}_t[\mathbf{U}(t,f)\mathbf{U}^\dagger(t,f)] \qquad \Phi_{YY}(f) = \mathbb{E}_t[\mathbf{Y}(t,f)\mathbf{Y}^\dagger(t,f)], \tag{3.13}$$

where $\mathbf{U}(t,f) = [U_1(t,f), \ldots, U_J(t,f)]^\top$ and $\mathbf{Y}(t,f) = [Y_1(t,f), \ldots, Y_J(t,f)]^\top$. In these computations, the phase of element j,j' at time–frequency point (t,f) is again the difference in phases between $Y_j(t,f)$ and $Y_{j'}(t,f)$, but the log magnitude is the



Channel 0

Channel 1

Channel 2

Channel 3

Ch 1-2 ILD

Ch 1-3 ILD

Ch 1-2 IPD

Ch 1-3 IPD

Fig. 3.1 Example multichannel recording from the CHiME-3 development dataset [6]. Real recording of talker F01 in pedestrian noise saying, “Excluding autos, sales increased zero point three percent.” *Top row* shows individual channels. Channel 0 is the reference close-talking mic, channel 2 is rear-facing on the device, and channels 1 and 3 are front-facing. *Bottom row* shows interaural differences between channel 1 and channels 2 and 3

sum of the log magnitudes of $Y_j(t,f)$ and $Y_{j'}(t,f)$. If, however, it is assumed that there are no acoustic obstructions between any of the microphones in the array and the source is far away from the array (i.e., equidistant from all microphones), then

$$|Y_j(t,f)| = |Y_{j'}(t,f)| = 1, \tag{3.14}$$

$$\therefore \left|\frac{Y_j(t,f)}{Y_{j'}(t,f)}\right| = \left|Y_j(t,f)Y_{j'}^*(t,f)\right| = 1, \tag{3.15}$$

and the two sets of cues are equivalent. When the magnitudes are not unity, as in the CHiME-3 setup for channel 2, the perceptually motivated $\mathbf{C}(t,f)$ matrix is not Hermitian symmetric because it contains the ratios of the channel observations, while the LCMV-related observation matrices and parameters are Hermitian symmetric.

3.3 Spatial-Clustering Approaches

Due to spatial aliasing, it is not possible to unambiguously map from noisy IPD estimates to the ITD at individual time–frequency points. The ambiguity can be resolved using spatial-clustering approaches. There are two main approaches: narrowband and wideband.

Narrowband spatial clustering (e.g., [49]) takes advantage of the fact that at almost all frequencies, sounds from two sources located at different positions will have different interaural parameters (phase and level differences). It typically does not make strong predictions about what those parameters will be, just that they will be different for different sources, thus permitting the separation of mixtures that include spatial aliasing. Once separation is performed in each individual frequency band, the sources identified in each band must be permuted to "match up" with one another in a second step.

In contrast, wideband models, such as [29] and [37], make stronger predictions about the connection between the interaural parameters at each frequency. In so doing, they are able to pool information across frequencies and avoid the potentially error-prone step of source alignment. The cost of this approach is that it must make certain assumptions about the form of the relationship across frequencies, and a failure of the observations to meet these assumptions could cause the failure of the separation process. In addition, care must be taken in developing this model so that it is robust to spatial aliasing.

All of these algorithms (i.e., [29, 37, 49]) have a similar structure. They first define each source by a set of frequency-dependent model parameters, $\Theta_i(f)$. They then alternate between two steps, an assignment of individual time–frequency points to source models using a soft or hard mask, $z_i(t,f)$, and an update of the source model parameters $\Theta_i(f)$ based on the observations at the points assigned to that source. This follows the expectation and maximization steps of the expectation maximization (EM) algorithm [15] in the case of soft masks or the two alternating steps of the k-means algorithm [34] in the case of hard masks.

3.3.1 Binwise Clustering and Alignment

The narrowband approach is exemplified by Sawada et al. [49]. Instead of clustering vectors of ILD and IPD measurements, the observed multichannel spectrogram signals are clustered directly. Building upon the notation of (3.3) to make a vectorial observation at each time–frequency point, let

$$\mathbf{H}_{i^*}(f) = [H_{i^*1}(f), \dots, H_{i^*J}(f)]^\top, \tag{3.16}$$

$$\mathbf{Y}(t,f) = [Y_1(t,f), \dots, Y_J(t,f)]^\top \tag{3.17}$$

$$\approx [H_{i^*1}(f)X_{i^*}(t,f), \dots, H_{i^*J}(f)X_{i^*}(t,f)]^\top = \mathbf{H}_{i^*}(f)X_{i^*}(t,f). \tag{3.18}$$

Independent processing is performed at each frequency, with no dependence on the frequency itself, so we will drop the f index from our notation in the remainder of this section. In order to separate the contribution of the target source from its spatial characteristics, the multichannel observations are magnitude-normalized at each time–frequency point:

$$\tilde{\mathbf{Y}}(t) = \frac{\mathbf{Y}(t)}{\|\mathbf{Y}(t)\|} = \frac{\mathbf{H}_{i^*}}{\|\mathbf{H}_{i^*}\|} \frac{X_{i^*}(t)}{|X_{i^*}(t)|}. \tag{3.19}$$

This normalization removes the magnitude of the source signal, but not its phase, $X_{i^*}(t)/|X_{i^*}(t)|$, which must be accounted for in the clustering procedure.

These observations are then clustered in a way that is similar to the line orientation separation technique (LOST) [40]. In this model, sources correspond to directions in a multidimensional complex space. These directions are represented by complex unit vectors, $\mathbf{a}_i$, and the distance between sources and observations is measured by projecting the observation onto the source direction. These distances are assumed to follow a circular complex Gaussian distribution with scalar variance σ_i^2,

$$p(\mathbf{Y}(t) \mid \mathbf{a}_i, \sigma_i) = \frac{1}{(\pi\sigma_i^2)^{J-1}} \exp\left(\frac{1}{\sigma_i^2} \|\tilde{\mathbf{Y}}(t) - (\mathbf{a}_i^H \tilde{\mathbf{Y}}(t))\mathbf{a}_i\|^2\right). \tag{3.20}$$

Note that, as required, this likelihood is invariant to a scalar phase applied to all channels of $\mathbf{Y}(t)$, because it is applied identically to $\tilde{\mathbf{Y}}(t)$ and $(\mathbf{a}_i^H \tilde{\mathbf{Y}}(t))\mathbf{a}_i$ and then removed by the magnitude computation. Thus this likelihood is invariant to the original phase of the source, $X_{i^*}(t)/|X_{i^*}(t)|$. For the same reason, it is also invariant to an additional scalar phase applied to the impulse response, so without loss of generality, we assume that $\angle H_{i^*1} = 0$. Similarly, a scalar phase applied to $\mathbf{a}_i$ will cancel out, so without loss of generality, we assume that $\angle[\mathbf{a}_i]_1 = 0$.

If considered in relation to the interaural parameters described above, it can be seen that for two channels, $\angle H_{i^*2} = \phi_{12}$, i.e., this parametrization is equivalent to the IPD. In addition,

$$\frac{[\mathbf{H}_{i^*}]_1}{\|\mathbf{H}_{i^*}\|} = \frac{H_{i^*1}}{\sqrt{|H_{i^*1}|^2 + |H_{i^*2}|^2}} = \sqrt{\frac{|H_{i^*1}|^2}{|H_{i^*1}|^2 + |H_{i^*2}|^2}} \tag{3.21}$$

$$= \sqrt{\frac{1}{1 + |H_{i^*2}|^2/|H_{i^*1}|^2}} = \sqrt{\frac{1}{1 + 10^{\alpha_{12}/10}}}, \tag{3.22}$$

showing that this parametrization is also equivalent to a pointwise transformation of the ILD. For every channel that is added beyond the second, this formulation adds an additional degree of freedom in phase and another in level for each source model. This linear growth is unlike the quadratic growth in degrees of freedom displayed by the spatial covariance matrix of each source in (3.13). This behavior implies that this parametrization can model point sources, but perhaps not diffuse sources, which require the full spatial covariance.

3.3.1.1 Cross-Frequency Source Alignment

In the narrowband clustering formulation, the frequency bands are processed independently and the source clusters can have a different arbitrary ordering in each band. Further processing is therefore required to assign clusters to sources in a consistent manner across frequency. Earlier techniques, e.g., [48], solved this same problem for frequency-domain independent component analysis (ICA) by correlating the extracted sources' magnitudes in adjacent frequency bands. For masking-based approaches, however, [49] found that performing the same sort of correlational alignment using the posterior probabilities from the masks, $z_i(t,f)$, yielded better alignments and thus better separation performance.

To perform this alignment exhaustively would take $O(J!F^2)$ time, where J is the number of sources and F the number of frequency bands. This is quite expensive, but [49] describes several heuristics for reducing the cost. The first is to perform a global exemplar-based clustering of the source posteriors across frequency. Instead of comparing all frequencies to each other, the posteriors at each frequency are compared to those of J exemplars, which reduces the cost to $O(J!FJ)$. While J and $J!$ are relatively small (typically $J < 5$), [49] suggests a greedy approach to the alignment calculation between a given pair of source sets, leading to an overall cost of $O(J^2FJ)$. This initial rough alignment is then refined using a fine-grained alignment based on comparing frequency bands that are either close to each other or harmonically related.

Overall, being a narrowband approach, this system is quite flexible in modeling impulse responses that vary a great deal across frequency. Such flexibility is not always required, however, and sacrifices some amount of noise robustness that comes from pooling information across frequencies. Instead, narrowband approaches tend to require longer temporal observations with stationary sources to achieve good separation performance. In addition, a good solution to the alignment problem requires careful tuning of the above heuristics. This can be difficult, for example, for wideband speech, where activity in frequencies up to 4 kHz containing sonorant phonemes is uncorrelated or even negatively correlated with activity in frequencies above 4 kHz containing obstruent phonemes, as can be seen in Fig. 3.1.

3.3.2 Fuzzy c-Means Clustering of Direction of Arrival

An example of a wideband approach is that of [29], which combines ideas from [10, 28, 53]. This approach performs clustering based solely on IPD converted to ITD using the Stepwise Phase dIfference REstoration (SPIRE) method [53], which resolves the spatial aliasing issue for certain kinds of arrays. SPIRE uses closely spaced pairs of microphones within a larger array to estimate the phase-wrapping terms in (3.12). Specifically, by sorting microphone pairs from the smallest separation to the largest, SPIRE identifies the unknown ℓ term in (3.12), expanded as

$$\phi_k = 2\pi f \Delta_{\tau_k}/f_s + 2\pi\ell_k = 2\pi d_k f/c + 2\pi\ell_k, \tag{3.23}$$

where k indexes the microphone pair and all terms indexed by k are specific to time–frequency point (t,f). For two different microphone pairs, most of these quantities are identical, allowing the correct ℓ_k to be identified recursively by

$$(\phi_{k-1} + 2\pi\ell_{k-1})\frac{d_k}{d_{k-1}} - \pi \le \phi_k + 2\pi\ell_k \le (\phi_{k-1} + 2\pi\ell_{k-1})\frac{d_k}{d_{k-1}} + \pi. \tag{3.24}$$

Once these IPD terms are identified for each time–frequency point, they can be directly converted to ITDs by

$$\Delta_{\tau_k}(t,f) = \frac{f_s}{2\pi f}(\phi_k(t,f) - 2\pi\ell_k(t,f)). \tag{3.25}$$

The scalar ITDs for the outermost microphone pair, $\Delta_{\tau_K}(t,f)$, are then clustered using an alternating approach similar to the GMM expectation maximization described above. Specifically, the parameters in this clustering are the direction for each source, denoted θ_i, and the soft cluster assignment for each time–frequency point, $z_i(t,f)$. These two quantities are updated using

$$z_i(t,f) = \frac{\|\Delta_{\tau_K}(t,f) - \theta_i\|^{2/(\gamma-1)}}{\sum_{i'} \|\Delta_{\tau_K}(t,f) - \theta_{i'}\|^{2/(\gamma-1)}}, \qquad \theta_i = \frac{\sum_{t,f} z_i^{\gamma}(t,f)\Delta_{\tau_K}(t,f)}{\sum_{t,f} z_i^{\gamma}(t,f)}, \tag{3.26}$$

where $\gamma > 1$ is a user-defined parameter controlling the softness of the likelihoods. Aside from this γ parameter, which effectively scales the log-likelihood, these updates are equivalent to GMM EM with a spherical unit variance.

Because it is wideband, this approach is able to pool information across frequency and requires fewer temporal observations than narrowband approaches. The use of the microphone pair with the widest spacing for localization provides the most precise estimates. In order to do so, however, it makes the assumption that the ITD is a pure delay between microphones, which appears in the form of (3.25). This is generally not the case in reverberant environments when early specular reflections disrupt this relationship. It also implies that sounds come from point sources and have no diffuse component, which is also unlikely in reverberant environments.

3.3.3 Binaural Model-Based EM Source Separation and Localization (MESSL)

The binaural model-based EM source separation and localization (MESSL) algorithm [37] explicitly models IPD and ILD observations using a Gaussian mixture model. In order to avoid spatial aliasing, MESSL models the ITD as a discrete random variable, and the IPD as a mixture over these ITDs, computing the source assignment variables as $z_i(t,f) = \sum_\tau z_{i\tau}(t,f)$. Intuitively, while an IPD does not correspond to a unique ITD in the presence of spatial aliasing, every ITD does correspond to a unique IPD, and so, by comparing the likelihoods of a set of ITDs, the most likely explanation for a set of observed IPDs can be found. The probability distribution for the Gaussian observations for a source i and discrete delay Δ_τ is

$$\begin{aligned} &p(\phi(t,f), \alpha(t,f) \mid i, \tau, \Theta) \\ &\quad = p(\phi(t,f) \mid \tau, \xi_{i\tau}(f), \sigma_{i\tau}(f)) \cdot p(\alpha(t,f) \mid \mu_i(f), \eta_i(f)). \end{aligned} \tag{3.27}$$

The distributions of individual features are given by

$$p(\phi(t,f) \mid \tau, \xi_{i\tau}(f), \sigma_{i\tau}(f)) = \mathcal{N}\big(\hat{\phi}(t,f;\tau,\xi_{i\tau}(f)) \mid 0, \sigma^2(f)\big), \tag{3.28}$$

$$\text{where } \hat{\phi}(t,f;\tau,\xi_{i\tau}(f)) = \angle \exp\left(\iota(\phi(t,f) - 2\pi f \Delta_\tau / f_s - \xi(f))\right), \tag{3.29}$$

$$p(\alpha(t,f) \mid \mu_i(f), \eta_i(f)) = \mathcal{N}\big(\alpha(t,f) \mid \mu_i(f), \eta_i^2(f)\big). \tag{3.30}$$

The phase residual, $\hat{\phi}(t,f;\tau,\xi_{i\tau}(f))$, computes the distance between the observed phase difference, $\phi(t,f)$, and the phase difference that should be observed from source i at frequency f, namely $2\pi f \Delta_\tau / f_s + \xi_{i\tau}(f)$. The first term in this expression is the phase difference predicted at frequency f by the ITD model with delay Δ_τ, and the second term is a frequency-dependent phase offset parameter, which permits variations from the pure delay model caused by early echoes. Furthermore, this difference is constrained to lie in the interval $(-\pi,\pi]$. Note that Δ_τ comes from a discrete grid of delays on which the above expressions must be evaluated, so that their likelihoods may be compared to one another. This step is computationally expensive and could be avoided by using a more sophisticated optimization scheme to find the most likely ITD.

These likelihoods are then used in the expectation and maximization steps of the EM algorithm. In the expectation step, the assignment of time–frequency points to sources is computed by

$$z_{i\tau}(t,f) = \frac{p(\phi(t,f),\alpha(t,f) \mid i,\tau,\Theta)p(i,\tau)}{\sum_{i'\tau'} p(\phi(t,f),\alpha(t,f) \mid i',\tau'\Theta)p(i',\tau')}. \tag{3.31}$$

In the maximization step, the model parameters are all updated by taking weighted sums of sufficient statistics of the observations using the assignments as weights. For details, see [37].

As a wideband method, MESSL can pool localization information across frequency, requiring temporally shorter observations than narrowband methods. Its statistical formulation permits the incorporation of additional parameters, like the IPD means, $\xi_{i\tau}(f)$, that can model early echoes in addition to the direct-path pure delay. Because the model is so flexible, however, it requires careful initialization to avoid local minima that do effectively separate the sources. It also permits the use of a prior on the ILD means given the ITDs.

This flexibility has facilitated several extensions of MESSL. Weiss et al. [59] combined the spatial separation of MESSL with a probabilistic source model. Instead of estimating a single maximum likelihood setting of parameters, [14] used variational Bayesian inference to estimate posterior distributions over the MESSL parameters. Instead of a grid of ITDs, [54] used random sampling to extract the best IPD-ILD parameters for a multichannel configuration.

3.3.4 Multichannel MESSL

Multichannel MESSL [5] models every pair of microphones separately using the binaural model described in Sect. 3.3.3. These models are coordinated through a global T–F mask for each source. For a spatial-clustering system to be as flexible as possible, it should not require calibration information for the microphone array. This flexibility will allow it to be used in applications from ad hoc microphone arrays to databases of user-generated content that lack specifications of the hardware that produced the recordings: source separation that is blind to the microphone array geometry. Without calibration, model parameters are difficult to translate between microphone pairs, but T–F masks are much more consistent across pairs and can be used to coordinate sources and models. This is the strategy adopted by multichannel MESSL, which maximizes the following total log-likelihood for J microphones:

$$\begin{aligned}\mathscr{L}(\Theta) &= \frac{2}{J}\sum_{j<j'=1}^{J}\mathscr{L}(\Theta_{jj'}) \qquad (3.32)\\ &= \frac{2}{J}\sum_{j<j'=1}^{J}\sum_{tf}\log\sum_{i\tau}\Big[p(z_{i\tau}(t,f)\mid\Theta_{jj'})\cdot p(\phi_{jj'}(t,f),\alpha_{jj'}(t,f)\mid z_{i\tau}(t,f),\Theta_{jj'})\Big].\end{aligned}$$

Averaging over all pairs in this way assumes that all microphone pairs are independent of one another, whereas in reality only $J-1$ are. This false assumption leads to an overconfidence in the likelihoods that is compensated by the $2/J$ term. This factor has much the same effect as the γ coefficient in (3.26) for the fuzzy c-means clustering approach. Preliminary experiments showed that using all pairs of microphones with this correction factor led to higher-quality separations than designating a single microphone as a reference and using $J-1$ pairs. The E and M steps for the model then proceed almost as in the two-channel algorithm. In the E step, the likelihood of the observations for each microphone pair is calculated under each source model. These likelihoods are then multiplied across microphone pairs and normalized across sources to give the final global posterior masks. In the M step, these global masks are used to reestimate the parameters of each pairwise model.

Initializing the multichannel model requires initializing the pairwise models and coordinating the source models across microphone pairs. We explored two different initializations. The first used the PHAT-histogram approach [1] to find the dominant peaks in cross-correlations between pairs of channels, followed by several iterations of binaural MESSL to estimate a mask for each source. These masks were then used to align the sources across microphone pairs. This approach has the advantage of being self-contained. The second initialization used a T–F mask derived from level differences between a beamformer output and a reference microphone. In the experiments below on CHiME-3 data, this was between the output of BeamformIt [3] and channel 2, the microphone facing away from the talker. The initial mask is then constructed from the 30% of points where the beamformer output is maximally greater in energy than the reference. This initialization has the advantage of automatically aligning the source models across microphone pairs, but can fail if the baseline beamformer fails in localization or separation.

Multichannel MESSL modeling all pairs of microphones has enough parameters to model both point sources and diffuse sources. The models can be arranged into a $J \times J$ matrix of sorts to reflect the observations $\mathbf{C}(t,f)$, where each pair of microphones corresponds to an entry in the matrix. This parametrization comes at the cost of a running time that is quadratic in the number of microphones. Preliminary experiments to reduce this computational complexity showed that subsampling the microphone pairs could trade off separation performance for complexity. For the six microphone recordings in CHiME-3, this was not necessary, so we will leave this investigation for future work.

3.4 Mask-Smoothing Approaches

One widely recognized problem that arises in mask-based separation is musical noise due to isolated false positive T–F points in the mask. Several approaches have attempted to alleviate this problem by applying a separate smoothing process after mask estimation [13, 19, 35, 56]. This section discusses the incorporation of these smoothing procedures into the spatial clustering process itself.

3.4.1 Fuzzy Clustering with Context Information

[29] introduced a mask-smoothing approach based on a heuristic modification of the source assignments $z_i(t,f)$ after each expectation step, following an approach first applied in image segmentation [12]. In particular, they defined

$$\bar{z}_i(t,f) = \frac{1}{|N(t,f)|} \sum_{t',f' \in N(t,f)} z_i(t',f'), \tag{3.33}$$

where $N(t,f)$ is a set of time–frequency indices for points that neighbor point (t,f). In [30], N is a rectangular neighborhood of 15 frequency bands and 9 time frames centered on the target point, equivalent to a rectangle of size 118 Hz by 90 ms. This averaged mask is applied in the expression for the update of the source memberships,

$$\tilde{z}_i(t,f) = \frac{z_i^{\gamma}(t,f)\bar{z}_i^{\beta}(t,f)}{\sum_{i'} z_{i'}^{\gamma}(t,f)\bar{z}_{i'}^{\beta}(t,f)}, \tag{3.34}$$

where β is a parameter controlling the relative contribution of the smoothed masks. [29] ran an initial iteration of the separation process with $\beta = 0$ to provide an unbiased estimate of θ_i, followed by five iterations with $\beta = 10$ to provide robustness to noise and reverberation. After iteration, a median filter was run over the masks to further reduce spurious classifications and musical noise.

3.4.2 MESSL in a Markov Random Field

With the same motivation, [36] proposed embedding the MESSL algorithm into a grid-shaped pairwise Markov random field (MRF) to simultaneously estimate model parameters and smooth T–F masks. This MRF penalizes the assignment of neighboring T–F points to different sources, smoothing the masks and reducing musical noise. The combined model is referred to as MESSL-MRF. In image segmentation applications, these models have been shown to be effective in combining evidence across neighboring pixels, e.g., [7]. While exact inference in these models is intractable, a number of approximation methods have been shown to be effective, including graph-cuts and loopy belief propagation (LBP) [52]. In addition, learning the parameters of an MRF model is also typically intractable, but it has been shown that approximate learning using expectation maximization can provide a reasonable approximation in practice for segmenting noisy images [20, 62]. MRFs have been used in several speech separation systems recently for both single- [25, 31] and multichannel approaches [26].

3.4.2.1 Pairwise Markov Random Fields

An MRF is an undirected graphical model, representing the joint probability of several random variables as a product of potential functions over subsets of those variables [7]. Depending on the structure of the graph, certain quantities can be estimated much more efficiently because of this factorization. This section focuses on pairwise MRFs, in which only pairwise interactions between variables are nonzero and thus only pairwise potential functions are necessary. In such models, the joint distribution of random variables $z_1, z_2, \dots, z_N$ can be written as

$$p(z_1, z_2, \dots, z_N) = \frac{1}{Z} \prod_{kk'} \psi_{kk'}(z_k, z_{k'}) \prod_k \psi_k(z_k), \tag{3.35}$$

where $\psi_k(z_k)$ is the potential function of variable z_k by itself, perhaps induced by a corresponding observation, and $\psi_{kk'}(z_k, z_{k'})$ is the pairwise potential function between z_k and $z_{k'}$, representing compatibilities between their various configurations. Using the sum–product variant of the belief propagation algorithm [41], it is possible to estimate the distribution of each individual variable when all of the others are marginalized away. In the case of tree-structured graphs, belief propagation can compute these quantities exactly. In the case of graphs with loops, it can only approximate these quantities, but it has been shown that such approximations perform well in practice [60].

3.4.2.2 MESSL-MRF

We propose smoothing MESSL masks by using the MESSL likelihood as the local potential in a grid-shaped pairwise MRF. In the context of such a model, z_k is the random variable representing the source number responsible for the majority of the energy at time–frequency point k.[1] If there are I sound sources, then z_k is a discrete I-dimensional multinomial random variable. In the experiments below, I was 2. The grid-shaped MRF then has potentials between every T–F point and its four direct neighbors in time and frequency. Thus the potential function $\psi_{kk'}(z_k, z_{k'})$ represents the compatibility between source z_k dominating T–F point k and source $z_{k'}$ dominating T–F point k'. We set the compatibility potentials, $\psi_{kk'}(z_k, z_{k'})$, to

$$\psi_{kk'}(z_k, z_{k'}) = \exp(-\beta\delta(z_k, z_{k'})), \tag{3.36}$$

where $\delta(z_k, z_{k'})$ is the discrete Dirac delta function, which is 1 when $z_k = z_{k'}$ and 0 otherwise, and β is a parameter that we tuned on a separate validation dataset. While simple, this potential is standard in MRF approaches to image segmentation.

More sophisticated compatibility potentials are possible and can be learned from training data. In particular, at low frequencies, ground truth masks tend to be more correlated across time because of the presence of strong lower harmonics. At high frequencies, they are more correlated across frequency because of wideband bursts and frication noise. Thus a frequency-dependent compatibility potential could be useful, but we leave this approach for future work.

In MESSL-MRF, the local potential is defined as

$$\psi_{tf}(z_{tf}) = \sum_{\tau} z_{i\tau}(t,f), \tag{3.37}$$

where we have changed the notation back from indexing hidden variables by k to t,f, and $z_{i\tau}(t,f)$ is defined in (3.31). We find the maximum likelihood parameters Θ from the test data using the EM algorithm [15, 20, 62]. Although learning in this MRF is intractable, it can be approximated by inserting the MRF belief propagation step between the E and M steps of a standard EM algorithm. In MESSL, it thus becomes a mask-smoothing step. MESSL's E step computes $z_{i\tau}(t,f)$, which defines the local potential $\psi_{tf}(z_{tf})$ in (3.37). From these, LBP is run until convergence to compute the soft masks, $b_{tf}(z_{tf})$, which are used to compute updated posteriors

$$\bar{z}_{i\tau}(t,f) = z_{i\tau}(t,f)\frac{b_{tf}(z_{tf})}{\sum_{\tau'} z_{i\tau'}(t,f)}. \tag{3.38}$$

And these are used in the standard MESSL M-step updates.

[1] For the purposes of the MESSL-MRF discussion, the indices k and k' are a shorthand for the T–F coordinates (t_k, f_k) and $(t_{k'}, f_{k'})$.

This approach has a similar effect to the context incorporation described in Sect. 3.4.1, namely encouraging neighboring points to belong to the same source. The probabilistic formulation of MESSL-MRF allows it to easily incorporate prior information about the relationships between neighboring points. It permits the substitution of the solution algorithm from loopy belief propagation [52] if desired. It also makes clear the approximations being made and their effect on the solution. The cost of the approach, however, is that to maintain these desirable properties, it must not utilize too large a neighborhood in its smoothing. Large neighborhoods in grid-shaped graphical models reduce the benefits of factorizing the joint distribution and lead to longer convergence times, if convergence is achieved at all.

3.5 Driving Beamforming from Spatial Clustering

Beamforming is the process of combining signals recorded from a microphone array into a single estimate of a target signal. This estimate is typically driven by an optimality criterion. One popular criterion for fixed (nonadaptive) filter-and-sum beamforming is that of minimum variance distortionless response (MVDR) [8], which aims to minimize the output power of the beamformer while preserving signals from a target "look" direction. For signals recorded as in (3.3), a filter-and-sum beamformer can be represented as a frequency-dependent vector, $\mathbf{w}(f)$, and the signal estimated by the beamformer is

$$\hat{X}_1(t,f) = \mathbf{w}^H(f)\mathbf{Y}(t,f). \tag{3.39}$$

For a steering vector $\mathbf{d}(f)$, which should have unity gain, the MVDR beamformer is

$$\mathbf{w}^*(f) = \min_{\mathbf{w}} E\left\{|\mathbf{w}^H\mathbf{X}(t,f)|^2\right\} \text{s.t.} \mathbf{w}^H\mathbf{d}(f) = 1. \tag{3.40}$$

Recently, [50] showed that this can be solved without the use of an explicit steering vector by

$$\mathbf{w}^*(f) = \frac{\Phi_{UU}^{-1}(f)\Phi_{HH}(f)e_{\text{ref}}}{\operatorname{tr}\left(\Phi_{UU}^{-1}(f)\Phi_{HH}(f)\right)} = \frac{(\Phi_{UU}^{-1}(f)\Phi_{YY}(f) - I)e_{\text{ref}}}{\operatorname{tr}\left(\Phi_{UU}^{-1}(f)\Phi_{YY}(f)\right) - J}, \tag{3.41}$$

where I is the $J \times J$ identity matrix, and e_{ref} is a vector of zeros with a single one selecting a reference microphone. This method allows the MVDR beamformer to be estimated without the use of an explicit steering vector, but still requires the estimation of $\Phi_{UU}(f)$, the noise spatial covariance, and either $\Phi_{YY}(f)$, the mixture spatial covariance, or $\Phi_{XX}(f) \propto \Phi_{HH}(f)$, the target source spatial covariance (note that the constant of proportionality divides out in (3.41)). In our experiments, the denominator of these expressions was sometimes close to zero or even negative for

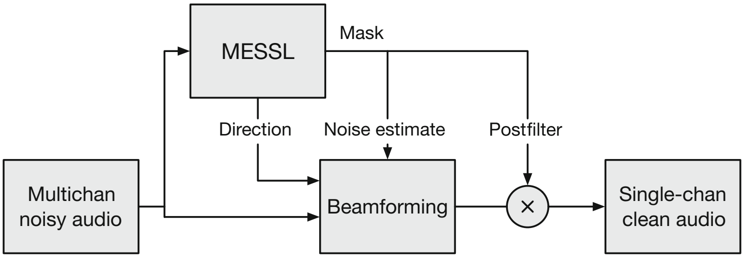


Fig. 3.2 Three ways that spatial-clustering outputs can drive minimum variance distortionless response beamforming: IPD parameters for look direction and masks for noise estimation and/or nonlinear postfiltering

a small set of frequencies, causing a large gain in the output at those frequencies and poor overall sound quality. We overcame this issue by enforcing that it be at least 1.

In the experiments discussed below, we explore the use of spatial clustering, and specifically MESSL, in driving MVDR beamforming in several ways, as illustrated in Fig. 3.2. Masks from spatial clustering can be used to estimate the noise spatial covariance $\Phi_{UU}(f)$, model parameters from spatial clustering can be used to compute a steering vector $\mathbf{d}(f)$, and the masks can also be used as a nonlinear postfilter applied to the output of the beamformer. This use of spatial clustering to drive MVDR beamforming was suggested by Cermak et al. [10], [11], and Kühne et al. [29].

The complement of the mask for a single source, $z_i(t,f)$, can be used as a frequency-dependent noise activity detector to estimate $\Phi_{UU}(f)$ as

$$\Phi_{UU}(f) \approx \frac{\sum_{t=1}^{T} (1 - z_i(t,f))\,\mathbf{X}(t,f)\mathbf{X}^H(t,f)}{\sum_{t=1}^{T} (1 - z_i(t,f))}. \tag{3.42}$$

Alternatively, [10] models and separates $I - 1$ noise sources individually, and computes $\Phi_{UU}(f)$ from the sum of these noise sources. To avoid speech damage, observations can be excluded from this sum from frames in which more than 40% of frequencies are predicted to be speech. To ensure that $\Phi_{UU}(f)$ is invertible, a certain number of frames from the beginning and end of the signal can be included in estimating it. We have found that the first M frames and the last $2M$ frames of an utterance work well for this empirically.

The steering vector can also be computed from the output of spatial clustering. From the estimated ITDs, assuming a pure delay,

$$\mathbf{d}^{(i)}(f) = [1, \exp(-\iota 2\pi f \Delta_{\tau_{12}^{(i)}}(f)/f_s), \ldots, \exp(-\iota 2\pi f \Delta_{\tau_{1J}^{(i)}}(f)/f_s)]. \tag{3.43}$$

Another possibility [11] is to find the $\mathbf{d}^{(i)}(f)$ that produces the best resynthesis of the observation from an estimate of the target signal, captured by the cost function

$$\mathscr{L}(\mathbf{d}) = \mathbb{E}_t\left[(\mathbf{x}(t,f) - \mathbf{d}(f)z_i(t,f)y_1(t,f))^2\right], \tag{3.44}$$

which is solved by

$$\mathbf{d}^{(i)}(f) = \frac{\sum_t \mathbf{x}(t,f)z_i(t,f)y_1^*(t,f)}{\sum_t |z_i(t,f)y_1(t,f)|^2}. \tag{3.45}$$

And, finally, it is possible to use the IPD estimates from multichannel MESSL to directly compute a full-rank Φ_{HH} for use in (3.41). While ILD is not useful for beamforming, as shown in (3.15), it is close to 1 for arrays without acoustic obstructions between microphones. Using just the IPD,

$$\Phi^{(i)}_{H_jH_{j'}}(f) = \frac{\phi_{ijj'f}}{|\phi_{ijj'f}|} \text{ for } \phi_{ijj'f} = \mathbb{E}_\tau\left[\exp(-\iota 2\pi f(\Delta_\tau + \xi^{(i)}_{jj'\tau}(f))/f_s)\right], \tag{3.46}$$

where $\Delta_\tau + \xi^{(i)}_{jj'\tau}(f)$ is the fine-grained mean of the IPD Gaussian between microphones j and j' for source i at the ITD indexed by τ. This approach takes advantage of MESSL's frequency-varying IPD estimates and does not assume a pure delay between microphones, as the first steering-vector formulation does.

Finally, masks estimated through spatial filtering can be used as nonlinear postfilters for the output of the MVDR beamformer. Suppressing points where $z_i(t,f) = 0$ to silence leads to musical noise, which can be avoided by suppressing them by some maximum amount. We found that using a maximum suppression of $-9\,\text{dB} = 0.355$ gave good noise suppression without causing noticeable musical noise.

3.6 Automatic Speech Recognition Experiments

This section describes experiments that examine the performance of MVDR beamforming driven by spatial clustering as a means of adapting far-field DNN-based automatic speech recognition to mismatched conditions. In particular, it uses the baseline recognizer from the AMI Meeting Corpus [9, 44], and tests it on the CHiME-3 corpus. These two conditions are mismatched in many ways, including signal-to-noise ratio, amount of reverberation, the distance to the microphone array, and the number and arrangement of the microphones. These experiments show that spatial clustering can provide significant recognition performance gains towards overcoming mismatched far-field conditions.

The recognizer was trained on the AMI Meeting Corpus, which contains speech recorded on an 8-microphone circular array of diameter 10 cm. We used the multiple-distant-mic (MDM) condition processed by the BeamformIt tool [3], which performs delay-and-sum beamforming using time-varying source localization. We used the AMI Full-ASR partition training set (about 78 h of speech) proposed in [51] and the corresponding Kaldi recipe with the provided automatic segmentations (version 1.6.1). The final acoustic model was a fully connected DNN that takes as input 40-dimensional log mel filterbank features with first and second time derivatives [55]. This DNN was trained on labels aligned by a GMM–hidden-Markov-model (HMM) model trained on mel-frequency cepstral coefficient (MFCC) features followed by linear discriminant analysis [21] and semi-tied covariance transforms [17], and discriminatively trained using the boosted maximum mutual information [42] criterion. The number of tied states was roughly 4000.

This recognizer was tested on the live-data portion of the CHiME-3 [6] dataset, which records speech input to a simulated tablet device in noisy environments. It used a 6-microphone rectangular array built around the edge of the tablet, to which a talker whose mouth was 30–50 cm away read sentences from the Wall Street Journal corpus (WSJ0). The recordings were made in four different noisy environments with an estimated signal-to-noise ratio averaging around 0 dB. The acoustic model described above was used with the default CHiME-3 language model. Thus the training and test sets differed significantly in the number of microphones, array geometry, amount of reverberation, microphone array distance, amount and type of noise, speaking style, and vocabulary. MESSL was used only on the development and test sets, not in training. The variant of multichannel MESSL used in the experiments had fully frequency-dependent parameters and smoothed its masks using MESSL-MRF.

For estimating the noise spatial covariance matrices $\Phi_{UU}(f)$, we compared using MESSL's masks to using the 400–800 ms of audio preceding the speech of each utterance, assumed to be noise only, which is the approach taken by the baseline CHiME-3 system (see [6]). For estimating the steering vector, we compared an estimate of $\Phi_{HH}(f)$ based on MESSL's IPD parameters to a derivation from (3.41) using Φ_{YY}. For a nonlinear postfilter, we compared the use of MESSL's masks to apply a gain to each T–F point of the beamformed signal to the use of the unmodified output of the beamformer.

3.6.1 Results

Table 3.1 shows the results of these experiments. The best system on the development set is shown in row 15 and used the MESSL noise estimate, the MESSL postfilter, cross-correlation initialization for MESSL, and the mixture spatial covari-

Table 3.1 Word error rates for recognizer trained on AMI data and tested on enhanced CHiME-3 real recordings

	Noise est	Postfilt	MESSL init	Look dir	WER (%)	
					Dev	Test
1	Prev	None	–	Mix	29.2	48.6
2	Prev	None	BeamformIt	MESSL	26.1	39.7
3	Prev	None	Xcorr	MESSL	24.6	40.2
4	Prev	MESSL	BeamformIt	MESSL	22.8	35.4
5	Prev	MESSL	BeamformIt	Mix	23.2	39.5
6	Prev	MESSL	Xcorr	MESSL	20.8	35.6
7	Prev	MESSL	Xcorr	Mix	22.5	40.1
8	MESSL	None	BeamformIt	MESSL	26.7	43.9
9	MESSL	None	BeamformIt	Mix	22.4	32.4
10	MESSL	None	Xcorr	MESSL	23.1	41.3
11	MESSL	None	Xcorr	Mix	22.1	34.8
12	MESSL	MESSL	BeamformIt	MESSL	23.9	39.5
13	MESSL	MESSL	BeamformIt	Mix	20.8	**30.0**
14	MESSL	MESSL	Xcorr	MESSL	20.4	36.1
15	MESSL	MESSL	Xcorr	Mix	**19.7**	32.6
16	–	None	–	–	22.7	36.2
17	–	MESSL	–	–	20.6	31.0

Key: Noise estimates from the previous 400–800 ms (Prev) or MESSL mask. Postfilter not used (None) or MESSL mask. MESSL initialized from BeamformIt or cross-correlation (Xcorr). Look direction from mixture (Mix) or from MESSL IPD. Bottom: BeamformIt baselines. Rows that are discussed in the text are shaded with $N = 27{,}119$, system 15 is significantly better than system 14 on the dev set ($p < 0.05$) and system 13 is significantly better than system 17 on the test set ($p < 0.01$) according to a one-sided binomial test

ance for (3.41). The columns of the table are ordered by the increase in word error rate (WER) on the development set caused by changing one of these parameters from this best setting. The rows of the table are ordered by the settings in each column. The parameter with the largest effect on this system is the noise estimate. Using the preceding 800 ms instead of the MESSL mask to estimate the noise results in a 2.75% absolute (14.0% relative) increase in the development set WER (row 7 vs. 15). The second largest effect comes from the postfilter. Removing the postfilter results in a 2.4% absolute (12.2% relative) increase in WER (row 11 vs. 15). The last two parameters have smaller effects on the development set. Initializing MESSL from BeamformIt instead of using cross-correlations results in a 1.1% absolute (5.4% relative) increase in WER (row 13 vs. 15). Using the look direction from the MESSL IPD instead of the mixture results in a 0.7% absolute (3.7% relative) increase in WER (row 14 vs. 15).

Baseline systems using BeamformIt are shown in the bottom two rows. The MESSL postfilter decreases WER by 2.1% absolute (9.3% relative) (row 16 vs. 17). Without a postfilter, two MESSL-MVDR systems (rows 9 and 11) achieve lower development and test WERs than the corresponding baseline (row 16), showing that

MESSL can be used to effectively drive beamforming. With the postfilter, the same two systems (rows 13 and 15) perform comparably to the baseline (row 17). The MESSL-MVDR system that performs best on the development set (row 15) reduces the WER on the test set by 3.6% absolute (9.9% relative) compared to the plain BeamformIt baseline. Consistent differences in performance have been seen on test and development sets for CHiME-3 [6], which might suggest looking directly for the best system on the test set, in which case, the best MESSL-MVDR system (row 13) reduces the WER by 6.2% absolute (17.1% relative).

3.6.2 Example Separations

Figure 3.3 shows example outputs of several of the systems described above for the input mixture shown in Fig. 3.1. The leftmost column shows a noisy input channel (channel 1) and the close microphone recording for reference. The remaining plots show system outputs. The top row of the figure shows the effect of the postfilter mask, with system 11 using no postfilter, system 15 using a postfilter with 9 dB maximum suppression, and the unnumbered system in the rightmost plot showing 40 dB maximum suppression. System 15 gave the best performance on the development set, and it can be seen that using too little postfilter suppression leaves too much noise in the output, while using too much suppression leads to artifacts such as musical noise. These artifacts, including the lack of noise suppression at the lowest frequencies, are due to the postfilter being purely based on spatial characteristics of the recordings. The incorporation of a speech-aware model into the mask estimation procedure could mitigate these artifacts, permitting a greater maximum suppression to be used with the postfilter.

The bottom row of plots in Fig. 3.3 shows the output of systems that differ in a single component from the best system (number 15), paralleling the discussion in Sect. 3.6.1. System 14 is the same as system 15, except that it uses MESSL's estimate of the look direction, based on its IPD model. It can be seen that in this separation, MESSL's look direction estimate leads to a residual noise that is more uniform across frequency in regions where the speech is inactive, although with slightly more noise at frequencies between 500 and 1000 Hz. System 13 uses BeamformIt to initialize MESSL instead of cross-correlations between channels. Its performance looks quite similar to that of system 15 for this separation. System 7 uses the 400–800 ms of noise preceding the speech to estimate the noise parameters. Its output on this example contains slightly more residual noise than system 15s, for example around 1200 Hz.

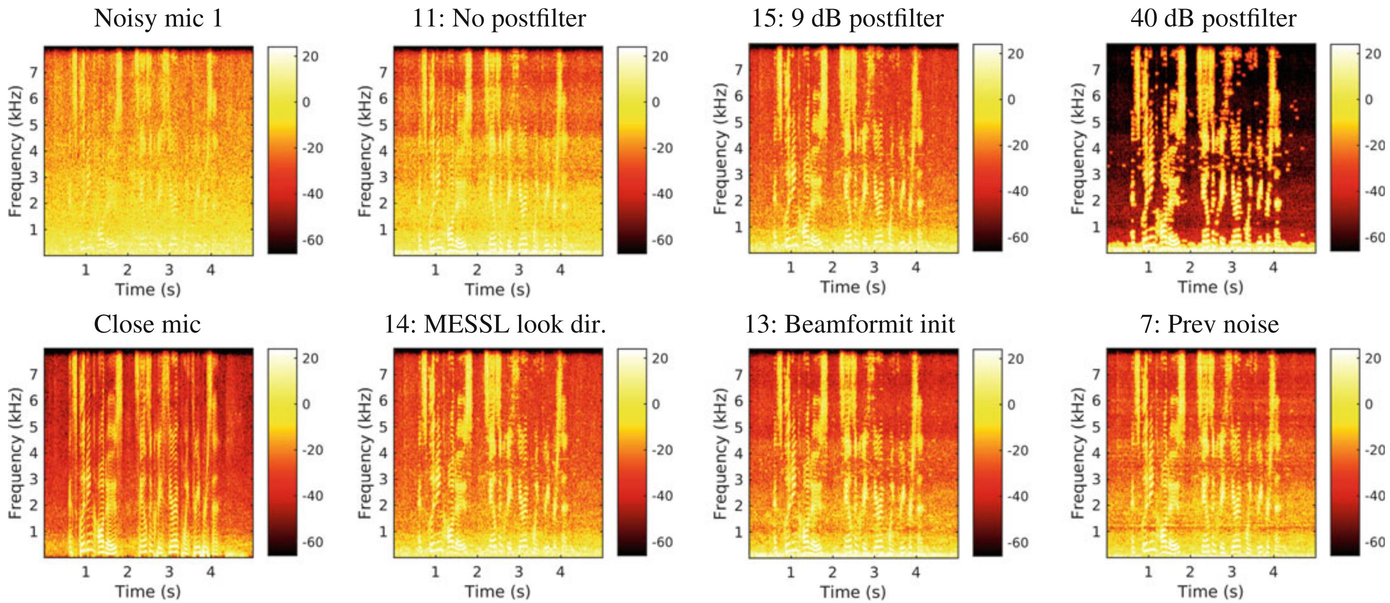


Fig. 3.3 Enhanced versions of the recording shown in Fig. 3.1 from several systems described in Table 3.1. The numbers above each plot identify the row in Table 3.1 corresponding to each system

3.7 Conclusion

This chapter has described the use of multichannel spatial clustering to drive minimum variance distortionless response beamforming. By clustering time–frequency points based on their spatial characteristics, these systems are able to generalize to quite different recording conditions. Experiments recognizing data from CHiME-3 with a recognizer trained on AMI show that there are several ways of utilizing the outputs of spatial clustering with MVDR beamforming, including incorporating its mask into the noise spatial covariance estimate, using the mask as a post–filter, or using estimated interaural phase differences to form the target spatial covariance matrix. In the future, generalizing the speech models of [59] from binaural to multichannel recordings could improve performance further.

Acknowledgements The work reported here was carried out during the 2015 Jelinek Memorial Summer Workshop on Speech and Language Technologies at the University of Washington, Seattle, and was supported by Johns Hopkins University via NSF Grant No. IIS 1005411, and gifts from Google, Microsoft Research, Amazon, Mitsubishi Electric, and MERL. It is also based upon work supported by the NSF under Grant No. IIS 1409431. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. Aarabi, P.: Self-localizing dynamic microphone arrays. IEEE Trans. Syst. Man Cybern. C **32**(4), 474–484 (2002)
2. Algazi, V.R., Duda, R.O., Thompson, D.M., Avendano, C.: The CIPIC HRTF database. In: Proceedings of WASPAA, pp. 99–102 (2001)
3. Anguera, X., Wooters, C., Hernando, J.: Acoustic beamforming for speaker diarization of meetings. IEEE Trans. Audio Speech Language Process. **15**(7), 2011–2022 (2007)
4. Araki, S., Sawada, H., Mukai, R., Makino, S.: Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors. Signal Process. **87**, 1833–1847 (2007)
5. Bagchi, D., Mandel, M.I., Wang, Z., He, Y., Plummer, A., Fosler-Lussier, E.: Combining spectral feature mapping and multi-channel model-based source separation for noise-robust automatic speech recognition. In: Proceedings of ASRU (2015)
6. Barker, J., Marxer, R., Vincent, E., Watanabe, S.: The third "CHiME" speech separation and recognition challenge: dataset, task and baselines. In: Proceedings of ASRU (2015)
7. Besag, J.: On the statistical analysis of dirty pictures (with discussion). J. R. Stat. Soc. B **48**(3), 259–302 (1986)
8. Capon, J.: High-resolution frequency-wavenumber spectrum analysis. Proc. IEEE **57**(8), 1408–1418 (1969)
9. Carletta, J.: Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. Lang. Resour. Eval. **41**(2), 181–190 (2007)
10. Cermak, J., Araki, S., Sawada, H., Makino, S.: Blind speech separation by combining beamformers and a time frequency binary mask. In: Proceedings of IWAENC, Paris (2006)
11. Cermak, J., Araki, S., Sawada, H., Makino, S.: Blind source separation based on a beamformer array and time frequency binary masking. In: Proceedings of ICASSP, vol. 1, pp. 145–148. IEEE, New York (2007)

12. Chuang, K.S., Tzeng, H.L., Chen, S., Wu, J., Chen, T.J.: Fuzzy c-means clustering with spatial information for image segmentation. Comput. Med. Imaging Graph. **30**(1), 9–15 (2006)
13. Cobos, M., Lopez, J.: Maximum a posteriori binary mask estimation for underdetermined source separation using smoothed posteriors. IEEE Trans. Audio Speech Language Process. **20**(7), 2059–2064 (2012)
14. Deleforge, A., Forbes, F., Horaud, R.: Variational EM for binaural sound-source separation and localization. In: Proceedings of ICASSP, pp. 76–79 (2013)
15. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B **39**, 1–38 (1977)
16. Dmochowski, J., Benesty, J., Affes, S.: On spatial aliasing in microphone arrays. IEEE Trans. Signal Process. **57**(4), 1383–1395 (2009)
17. Gales, M.: Semi-tied covariance matrices for hidden Markov models. IEEE Trans. Audio Speech Language Process. **7**(3), 272–281 (1999)
18. Gaubitch, N.D., Kleijn, W.B., Heusdens, R.: Auto-localization in ad-hoc microphone arrays. In: Proceedings of ICASSP, pp. 106–110. IEEE, New York (2013)
19. Grais, E., Erdogan, H.: Spectro-temporal post-smoothing in NMF based single-channel source separation. In: Proceedings of EUSIPCO, pp. 584–588 (2012)
20. Gu, D.B., Sun, J.: EM image segmentation algorithm based on an inhomogeneous hidden MRF model. IEEE Vis. Image Signal Process. **152**(2), 184–190 (2004)
21. Haeb-Umbach, R., Ney, H.: Linear discriminant analysis for improved large vocabulary continuous speech recognition. In: Proceedings of ICASSP, pp. 13–16 (1992)
22. Himawan, I., McCowan, I., Sridharan, S.: Clustered blind beamforming from ad-hoc microphone arrays. IEEE Trans. Audio Speech Language Process. **19**(4), 661–676 (2011)
23. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., Kingsbury, B.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. **29**(6), 82–97 (2012)
24. Jiang, Y., Wang, D., Liu, R.: Binaural deep neural network classification for reverberant speech segregation. In: Proceedings of Interspeech, pp. 2400–2403 (2014)
25. Kim, M., Smaragdis, P.: Single channel source separation using smooth nonnegative matrix factorization with Markov random fields. In: Proceedings of MLSP, pp. 1–6 (2013)
26. Kim, M., Smaragdis, P., Ko, G.G., Rutenbar, R.A.: Stereophonic spectrogram segmentation using Markov random fields. In: Proceedings of MLSP, pp. 1–6 (2012)
27. Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Habets, E., Sehr, A., Kellermann, W., Gannot, S., Maas, R., Haeb-Umbach, R., Leutnant, V., Raj, B.: The REVERB challenge: a common evaluation framework for dereverberation and recognition of reverberant speech. In: Proceedings of WASPAA, New Paltz, NY (2013)
28. Kühne, M., Togneri, R., Nordholm, S.: Smooth soft mel-spectrographic masks based on blind sparse source separation. In: Proceedings of Interspeech (2007)
29. Kühne, M., Togneri, R., Nordholm, S.: Adaptive beamforming and soft missing data decoding for robust speech recognition in reverberant environments. In: Proceedings of Interspeech, pp. 976–979 (2008)
30. Kühne, M., Togneri, R., Nordholm, S.: A novel fuzzy clustering algorithm using observation weighting and context information for reverberant blind speech separation. Signal Process. **90**(2), 653–669 (2010)
31. Liang, S., Liu, W., Jiang, W.: Integrating binary mask estimation with MRF priors of cochleagram for speech separation. IEEE Signal Process. Lett. **19**(10), 627–630 (2012)
32. Lippmann, R., Martin, E., Paul, D.: Multi-style training for robust isolated-word speech recognition. In: Proceedings of ICASSP, vol. 12, pp. 705–708 (1987)
33. Liu, Z., Zhang, Z., He, L.W., Chou, P.: Energy-based sound source localization and gain normalization for ad hoc microphone arrays. In: Proceedings of ICASSP, vol. 2, pp. 761–764. IEEE, New York (2007)

34. Lloyd, S.P.: Least squares quantization in PCM. IEEE Trans. Inf. Theory **28**(2), 129–137 (1982)
35. Madhu, N., Breithaupt, C., Martin, R.: Temporal smoothing of spectral masks in the cepstral domain for speech separation. In: Proceedings of ICASSP, pp. 45–48 (2008)
36. Mandel, M.I., Roman, N.: Enforcing consistency in spectral masks using Markov random fields. In: Proceedings of EUSIPCO (2015)
37. Mandel, M.I., Weiss, R.J., Ellis, D.P.W.: Model-based expectation maximization source separation and localization. IEEE Trans. Audio Speech Language Process. **18**(2), 382–394 (2010)
38. Middlebrooks, J.C., Green, D.M.: Sound localization by human listeners. Annu. Rev. Psychol. **42**, 135–159 (1991)
39. Narayanan, A., Wang, D.: Investigation of speech separation as a front-end for noise robust speech recognition. IEEE Trans. Audio Speech Language Process. **22**(4), 826–835 (2014)
40. O'Grady, P.D., Pearlmutter, B.A.: Soft-LOST: EM on a mixture of oriented lines. In: Independent Component Analysis and Blind Signal Separation, vol. 3195, 1270 pp. Springer, Berlin (2004)
41. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Francisco, CA (1988)
42. Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G., Visweswariah, K.: Boosted MMI for model and feature-space discriminative training. In: Proceedings of ICASSP, pp. 4057–4060 (2008)
43. Rafaely, B., Weiss, B., Bachmat, E.: Spatial aliasing in spherical microphone arrays. IEEE Trans. Signal Process. **55**(3), 1003–1010 (2007)
44. Renals, S., Hain, T., Bourlard, H.: Recognition and understanding of meetings: the AMI and AMIDA projects. In: Proceedings of ASRU, Kyoto (2007)
45. Roman, N., Wang, D.L., Brown, G.J.: Speech segregation based on sound localization. J. Acoust. Soc. Am. **114**, 2236–2252 (2003)
46. Roweis, S.: Factorial models and refiltering for speech separation and denoising. In: Proceedings of Eurospeech, Geneva, pp. 1009–1012 (2003)
47. Sainath, T.N., Weiss, R.J., Senior, A., Wilson, K.W., Vinyals, O.: Learning the speech front-end with raw waveform CLDNNS. In: Proceedings of Interspeech (2015)
48. Sawada, H., Mukai, R., Araki, S., Makino, S.: A robust and precise method for solving the permutation problem of frequency-domain blind source separation. IEEE Trans. Speech Audio Process. **12**(5), 530–538 (2004)
49. Sawada, H., Araki, S., Makino, S.: Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. IEEE Trans. Audio Speech Language Process. **19**(3), 516–527 (2011)
50. Souden, M., Benesty, J., Affes, S.: On optimal frequency-domain multichannel linear filtering for noise reduction. IEEE Trans. Audio Speech Language Process. **18**(2), 260–276 (2010)
51. Swietojanski, P., Ghoshal, A., Renals, S.: Hybrid acoustic models for distant and multichannel large vocabulary speech recognition. In: Proceedings of ASRU (2013)
52. Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., Rother, C.: A comparative study of energy minimization methods for Markov random fields with smoothness-based priors. IEEE Trans. Pattern Anal. Mach. Intell. **30**(6), 1068–1080 (2008)
53. Togami, M., Sumiyoshi, T., Amano, A.: Stepwise phase difference restoration method for sound source localization using multiple microphone pairs. In: Proceedings of ICASSP (2007)
54. Traa, J., Kim, M., Smaragdis, P.: Phase and level difference fusion for robust multichannel source separation. In: Proceedings of ICASSP, pp. 6687–6690 (2014)
55. Veselý, K., Ghoshal, A., Burget, L., Povey, D.: Sequence-discriminative training of deep neural networks. In: Proceedings of Interspeech, pp. 2345–2349 (2013)
56. Vincent, E.: An experimental evaluation of Wiener filter smoothing techniques applied to under-determined audio source separation. In: International Conference on Latent Variable Analysis and Signal Separation, pp. 157–164. Springer, Berlin, Heidelberg (2010)

57. Vincent, E., Bertin, N., Gribonval, R., Bimbot, F.: From blind to guided audio source separation: how models and side information can improve the separation of sound. IEEE Signal Process. Mag. **31**(3), 107–115 (2014)
58. Wang, D.: On ideal binary mask as the computational goal of auditory scene analysis. In: Divenyi, P. (ed.) Speech Separation by Humans and Machines, pp. 181–197. Springer US, Boston, MA (2005)
59. Weiss, R., Mandel, M.I., Ellis, D.W.P.: Combining localization cues and source model constraints for binaural source separation. Speech Commun. **53**(5), 606–621 (2011)
60. Yedidia, J., Freeman, W., Weiss, Y.: Generalized belief propagation. In: Advances in Neural Information Processing Systems, pp. 689–695. MIT, Cambridge (2000)
61. Yilmaz, O., Rickard, S.: Blind separation of speech mixtures via time–frequency masking. IEEE Trans. Audio Speech Language Process. **52**(7), 1830–1847 (2004)
62. Zhang, Y., Brady, M., Smith, S.: Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging **20**(1), 45–57 (2001)