Chapter 12
Steered Beamforming Approaches for Acoustic Source Localization

Jacek P. Dmochowski and Jacob Benesty

Abstract Multiple microphone devices aimed at hands-free capabilities and speech recognition via acoustic beamforming require reliable estimates of the position of the acoustic source. Two-stage approaches based on the time-difference-of-arrival (TDOA) are computationally simple but lack robustness in practical environments. This chapter presents a family of broadband source localization algorithms based on parameterized spatiotemporal correlation, including the popular and robust steered response power (SRP) algorithm. Before forming the conventional spatial correlation matrix, the vector of microphone signals is time-aligned with respect to a hypothesized source location. It is shown that this parametrization easily generalizes classical narrowband techniques to the broadband setting. Methods based on minimum information entropy and temporally constrained minimum variance are developed. In order to ease the high computational demands imposed by a location parameterized scheme, a sparse representation of the parameterized spatial correlation matrix is proposed.

12.1 Introduction

The localization of acoustic sources represents both a classical parameter estimation problem and an important component of more general problems such as hands-free voice communication and speech recognition. The problem is significantly more difficult than that of narrowband source localization [1], [2], as the desired signal is wideband and its propagation to the array is convolutive.

Jacek P. Dmochowski
City College of New York, NY, USA, e-mail: jdmochowski@ccny.cuny.edu

Jacob Benesty
INRS-EMT, QC, Canada, e-mail: benesty@emt.inrs.ca

I. Cohen et al. (Eds.): Speech Processing in Modern Communication, STSP 3, pp. 307–337.
springerlink.com

Simply put, the localization problem is to estimate the position of the source using only the signals observed at an array of microphones. The basic premise behind localization is that sources at different locations will exhibit different relative delays – the differences in propagation time from one microphone to the next. This physical property forms the basis for virtually all localization methods, and also necessitates the use of multiple microphones. While the theoretical minimum number of microphones is indeed two, it will be shown in the remaining sections of this chapter that it is advantageous to include additional microphones in the localization scheme.

The other classical acoustic localization techniques consist of two steps. In the first step, the relative delays across the various microphone pairs are estimated using the generalized cross-correlation (GCC) method described in [3]. The estimates of the relative delays are then mapped to the source location using one of a handful of techniques ranging from maximum-likelihood estimation to spherical interpolation [4], [5], [6]. In practice, the estimates of the relative delays are quite noisy. The second step typically involves a non-linear transformation of these relative delays, leading to performance limitations in harsh environments.

To that end, the methods presented in this chapter refrain from making a hard decision on a single relative delay between each microphone pair. Instead, a spatial statistic is formed for each potential location; having computed such a statistic for each candidate location, the algorithms designate the estimate of the source by the location which either minimizes or maximizes the statistic, depending on the context. Instead of parameterizing the relative delay, these techniques parameterize the cross-correlation functions with the location of the source.

Acoustic source localization is an active research area. We very briefly mention here some recent trends. Particle filtering methods formulate the localization problem in a state space and model the location probabilistically using a Markov process where the current "state" (i.e., location and velocity) evolve as a function of the previous states and estimates of the source location [7], [8], [9], [10], [11]. The particle filtering approach is useful for tracking purposes. While the majority of localization algorithms have focused on the inter-microphone delays, recently approaches utilizing energy measurements have also been proposed [12], [13], [14]. Localization methods employing a set of distributed microphones have also been studied [15], [16]. In this paradigm, the array consists of a set of networked multimedia devices with embedded microphones. Finally, methods based on blind multiple-input multiple-output (MIMO) identification of the acoustic impulse responses are presented in [17], [18]. This chapter is not meant to serve as a comprehensive treatment of the state-of-the-art. Rather, we present herein a group of algorithms based on parameterized spatial correlation which generalize classical narrowband techniques to the broadband setting.

The signal model used throughout the chapter is described in Section 12.2. Section 12.3 briefly covers the fundamentals of spatial and spatiotemporal

filtering. Section 12.4 presents the parameterized spatial correlation matrix as the fundamental structure in acoustic source localization and a family of methods based on this structure are detailed in Section 12.5. A sparse representation of the parameterized spatial correlation matrix is described in Section 12.6. A linearly constrained minimum variance approach based on the parameterized spatiotemporal correlation matrix is presented in Section 12.7. Section 12.8 summarizes the current limitations of the state-of-the-art, and concluding statements are made in Section 12.9.

12.2 Signal Model

Assume that an array of N microphones samples the sound field in both space and time; the output of sensor n at time k is modeled as

$$y_n(k) = \alpha_n(\mathbf{r}_\mathrm{s}) s\left[k - \tau - \mathcal{F}_{1n}(\mathbf{r}_\mathrm{s})\right] + v_n(k), \tag{12.1}$$

where $\alpha_n(\mathbf{r}_\mathrm{s}), n = 1, 2, \ldots, N$, models the attenuation of the source signal at sensor n as a function of the source location $\mathbf{r}_\mathrm{s} = \begin{bmatrix} r_\mathrm{s} & \phi_\mathrm{s} & \theta_\mathrm{s} \end{bmatrix}^T$ (superscript ^{T} denotes the transpose of a vector or a matrix), where r_s, ϕ_s, and θ_s denote the range, elevation, and azimuth, respectively, in a spherical coordinate system, s is the source signal, τ is the propagation time (in samples) from the source to sensor 1, $\mathcal{F}_{nm}(\mathbf{r}_\mathrm{s})$ is a function that relates the source position to the relative delay between microphones n and m, and v_n is the additive noise at sensor n. In the free-field case,

$$\alpha_n(\mathbf{r}_\mathrm{s}) \propto \frac{1}{\|\mathbf{r}_n - \mathbf{r}_\mathrm{s}\|}, \tag{12.2}$$

where $\mathbf{r}_n$ is the position of the nth microphone. In the majority of microphone array applications, the distance from the desired source to the array is large relative to the extent of the spatial aperture; in other words, the source is located in the array's far-field and the propagation of the signal is effectively that of a plane wave. As a result, it may be reasonably assumed that the attenuation coefficients are uniform across the array:

$$\alpha_n(\mathbf{r}_\mathrm{s}) = 1, \quad \forall n, \mathbf{r}_\mathrm{s}. \tag{12.3}$$

Moreover, the relative delays across the array are also independent of the range:

$$\mathcal{F}_{nm}(\phi_\mathrm{s}, \theta_\mathrm{s}) = \frac{1}{c}\boldsymbol{\zeta}^T(\phi_\mathrm{s}, \theta_\mathrm{s})(\mathbf{x}_m - \mathbf{x}_n), \tag{12.4}$$

where

$$\boldsymbol{\zeta}(\phi_\mathrm{s}, \theta_\mathrm{s}) = \begin{bmatrix} \sin\phi_\mathrm{s}\cos\theta_\mathrm{s} & \sin\phi_\mathrm{s}\sin\theta_\mathrm{s} & \cos\phi_\mathrm{s} \end{bmatrix}^T \tag{12.5}$$

is a unit vector which points in the direction of propagation of the source, and $\mathbf{x}_n = \left[x_n \; y_n \; z_n \right]^T$ is the position vector of the nth microphone in Cartesian co-ordinates. The dimensionality of the location space is further reduced to one if one assumes that the source and the array lie on a plane; in other words, when $\phi_\mathrm{s} = \frac{\pi}{2}$:

$$\begin{aligned} \mathcal{F}_{nm}(\phi_\mathrm{s}, \theta_\mathrm{s}) &= \mathcal{F}_{nm}(\theta_\mathrm{s}) \\ &= \frac{1}{c} \boldsymbol{\zeta}^T (\theta_\mathrm{s}) (\mathbf{x}_m - \mathbf{x}_n) \\ &= \frac{1}{c} \left[\cos\theta_\mathrm{s}(x_m - x_n) + \sin\theta_\mathrm{s}(y_m - y_n) \right]. \end{aligned} \tag{12.6}$$

Throughout the chapter, we assume that the source rests in the array's far-field. Moreover, to ease notation, we further assume that the source is at least approximately on the same plane as the microphones. While the latter is a rather lofty assumption, the argument of the relative delay function may be made multivariate to generalize to the case of an elevated source. The methods presented in the remainder of this chapter apply equally to both the two-dimensional elevation-azimuth space and the one-dimensional azimuth space. If the source is located in the array's near-field, $\mathcal{F}_{nm}$ becomes a function of the range r_s and the resulting methods may also determine the source range.

Lastly, it should be mentioned that in all practical acoustic environments, each microphone picks up a convolution of the source with a room impulse response. Since only the direct-path component conveys location information, the reverberant components are not included in the desired signal term and may be lumped into the additive noise terms $v_n(k)$, thus making the additive noise temporally correlated with the desired signal. It will be shown that this temporal correlation is one of the challenges to robust localization in real environments.

12.3 Spatial and Spatiotemporal Filtering

Before delving into the steered-beamforming approach to acoustic source localization, we briefly cover the fundamentals of spatiotemporal filtering.

The principle behind linear filtering is to collect the signal across an aperture (i.e., a discrete set of samples), apply a weight to each collected sample, and then sum the weighted samples to form the filter output. Spatiotemporal filtering, more commonly known as broadband beamforming, is a direct application of this principle. In addition to filtering the signal in the spatial aperture, we store the previous $L-1$ samples of each microphone to form a spatiotemporal aperture of size NL:

$$\bar{\mathbf{y}}(k) = \left[\, \mathbf{y}^T(k)\ \mathbf{y}^T(k-1) \cdots\ \mathbf{y}^T(k-L+1) \,\right]^T, \tag{12.7}$$

where

$$\mathbf{y}(k) = \left[\, y_1(k)\ y_2(k) \cdots\ y_N(k) \,\right]^T \tag{12.8}$$

is the spatial aperture at time k.

A linear spatiotemporal filter $\mathbf{h}$ is then applied to the aperture to form the beamformer output:

$$z(k) = \mathbf{h}^T \bar{\mathbf{y}}(k), \tag{12.9}$$

where $z(k)$ is the output of the broadband beamformer and

$$\mathbf{h} = \left[\, \mathbf{h}_0^T\ \mathbf{h}_1^T \cdots\ \mathbf{h}_{L-1}^T \,\right]^T \tag{12.10}$$

is the spatiotemporal filter which consists of L spatial filters $\mathbf{h}_l$, where each spatial filter spatially filters the array signals at time $k-l$:

$$\mathbf{h}_l = \left[\, h_{l,1}\ h_{l,2} \cdots\ h_{l,N} \,\right]^T, \tag{12.11}$$

where $h_{l,n}$ is the coefficient applied to $y_n(k-l)$.

The variance of the beamformer output then follows as

$$\begin{aligned} E\left[z^2(k)\right] &= E\left\{ \left[\mathbf{h}^T \bar{\mathbf{y}}(k)\right]^2 \right\} \\ &= \mathbf{h}^T \mathbf{R}_{\bar{y}} \mathbf{h}, \end{aligned} \tag{12.12}$$

where $E[\cdot]$ denotes mathematical expectation and

$$\mathbf{R}_{\bar{y}} = E\left[\bar{\mathbf{y}}(k) \bar{\mathbf{y}}^T(k)\right] \tag{12.13}$$

is the spatiotemporal correlation matrix (STCM) of the observed signals.

If the signal of interest is narrowband, the temporal aperture length may be taken to be $L = 1$, and the filter consists of a weight for each spatial sample (i.e., microphone). In this case, the STCM simplifies to the spatial correlation matrix (SCM), which is the fundamental structure in narrowband localization methods:

$$\mathbf{R}_y = E\left[\mathbf{y}(k) \mathbf{y}^T(k)\right]. \tag{12.14}$$

12.4 Parameterized Spatial Correlation Matrix (PSCM)

In source localization applications, one is interested in how the location of the source affects the observed second orders statistics (SOS) at the array. In

other words, one would like to parameterize the SCM and STCM with the source location. One way of achieving this is to time-align the microphone signals *prior* to forming the correlation matrix. This is in contrast to the narrowband approach, where a linear weighting is applied to the SCM in order to properly phase-delay the sensor signals. In the broadband case, the time-aligning is performed directly in the correlation matrix.

To that end, consider forming the *parameterized spatial correlation matrix* (PSCM) according to

$$\mathbf{R}_y(\theta) = E\left[\mathbf{y}(k,\theta)\mathbf{y}^T(k,\theta)\right], \tag{12.15}$$

where

$$\mathbf{y}(k,\theta) = \left[\, y_1(k)\; y_2\left[k+\mathcal{F}_{12}(\theta)\right] \cdots\; y_N\left[k+\mathcal{F}_{1N}(\theta)\right]\right]^T \tag{12.16}$$

is the spatial aperture time aligned with respect to location θ. Substituting (12.1) into (12.16) yields the nth element of $\mathbf{y}(k,\theta)$:

$$y_n(k,\theta) = s\left[k-\tau-\mathcal{F}_{1n}\left(\theta_{\mathrm{s}}\right)+\mathcal{F}_{1n}\left(\theta\right)\right] + v_n\left[k+\mathcal{F}_{1n}\left(\theta\right)\right]. \tag{12.17}$$

Thus, we may write the time-aligned spatial aperture as

$$\mathbf{y}(k,\theta) = \mathbf{s}(k-\tau,\theta) + \mathbf{v}(k,\theta), \tag{12.18}$$

where

$$\mathbf{s}(k,\theta) = \left[\, s(k)\; s\left[k-\mathcal{F}_{12}(\theta_{\mathrm{s}})+\mathcal{F}_{12}(\theta)\right] \cdots\; s\left[k-\mathcal{F}_{1N}(\theta_{\mathrm{s}})+\mathcal{F}_{1N}(\theta)\right]\right]^T$$

and

$$\mathbf{v}(k,\theta) = \left[\, v_1(k)\; v_2\left[k+\mathcal{F}_{12}(\theta)\right] \cdots\; v\left[k+\mathcal{F}_{1N}(\theta)\right]\right]^T.$$

Notice that when $\theta = \theta_{\mathrm{s}}$,

$$\mathbf{s}(k,\theta_{\mathrm{s}}) = s(k)\mathbf{1}_N, \tag{12.19}$$

where $\mathbf{1}_N$ is a vector of N ones. Conversely, the parameterized noise vector is given by

$$\mathbf{v}(k,\theta_{\mathrm{s}}) = \left[\, v_1(k)\; v_2\left[k+\mathcal{F}_{12}(\theta_{\mathrm{s}})\right] \cdots\; v\left[k+\mathcal{F}_{1N}(\theta_{\mathrm{s}})\right]\right]^T.$$

Assuming that the signal and noise vectors are uncorrelated:

$$E\left[\mathbf{s}(k,\theta)\mathbf{v}^T(k,\theta)\right] = \mathbf{0}_{N\times N}, \tag{12.20}$$

where $\mathbf{0}_{N\times N}$ is an N-by-N matrix of zeros, the PSCM may be written as

$$\mathbf{R}_y(\theta) = \mathbf{R}_s(\theta) + \mathbf{R}_v(\theta), \tag{12.21}$$

where

$$\mathbf{R}_s(\theta) = E\left[\mathbf{s}(k-\tau,\theta)\mathbf{s}^T(k-\tau,\theta)\right] \tag{12.22}$$

and

$$\mathbf{R}_v(\theta) = E\left[\mathbf{v}(k,\theta)\mathbf{v}^T(k,\theta)\right]. \tag{12.23}$$

From (12.23), the parameterized noise correlation matrix $\mathbf{R}_v(\theta)$ will typically be full-rank, regardless of the parameter θ. Conversely, $\mathbf{R}_s(\theta)$ is rank-one if $\theta = \theta_\mathrm{s}$. This property allows us to localize the source by observing the nature of the PSCM as θ is varied across the space of locations.

12.5 Source Localization Using Parameterized Spatial Correlation

Informally speaking, the elements of the PSCM are highly correlated (i.e., larger in magnitude) when the hypothesized parameter θ matches the true parameter θ_s. There are thus several ways of processing the various PSCMs to identify the actual source location. The remainder of this chapter focuses on how to choose a function $g(\mathbf{M})$, where $\mathbf{M}$ is a matrix, such that

$$\begin{aligned}\hat{\theta}_\mathrm{s} &= \arg\max_\theta g\left[\mathbf{R}_y(\theta)\right] \\ &\approx \theta_\mathrm{s}.\end{aligned} \tag{12.24}$$

A convenient aspect of the parametrization of the SCM is that it allows for the extension of the classical narrowband methods such as minimum variance, subspace, and linear prediction [2] to the broadband signal case. This will be clear in the forthcoming subsections.

12.5.1 Steered Response Power

The steered response power (SRP) method [20], [21] is the simplest method based on the PSCM. Note that the correlated nature of $\mathbf{R}_s(\theta_\mathrm{s})$ will increase the off-diagonal values comprising the correctly-steered PSCM $\mathbf{R}_y(\theta_\mathrm{s})$. Thus, one way of localizing the source is to simply sum the elements of the PSCM as a function of the parameter θ. The location which yields the largest sum of elements is designated as the source:

$$\hat{\theta}_\mathrm{s} = \arg\max_\theta \mathbf{1}_N^T\mathbf{R}_y(\theta)\mathbf{1}_N. \tag{12.25}$$

Note that this is equivalent to applying a fixed filter given by

$$\begin{aligned}\mathbf{h}(\theta) &= \mathbf{h} \\ &= \mathbf{1}_N, \ \forall\theta,\end{aligned} \tag{12.26}$$

to the parameterized aperture $\mathbf{y}(k,\theta)$,

$$z(k,\theta) = \mathbf{h}^T\mathbf{y}(k,\theta), \tag{12.27}$$

and computing the resulting output power at each parameter:

$$E\left[z^2(k,\theta)\right] = \mathbf{1}_N^T\mathbf{R}_y(\theta)\mathbf{1}_N. \tag{12.28}$$

Since the filter $\mathbf{h}$ is independent of both the data and the parameter, the SRP approach is computationally attractive.

Figure 12.1 depicts the SRP spatial spectra averaged over one minute of synthetically convolved speech at four levels of room reverberation; the reflection coefficients of the walls, ceiling, and floor are adjusted to generate 60 dB decay times T_{60} of 0, 100, 200, and 300 ms. The reverberation times are measured using the reverse time integrated method of [22]. The image method for simulating room acoustics is employed in order to generate the synthetic impulse responses [23]. The simulated room has the following properties:

- dimensions given by 304.8cm-by-457.2cm-by-381cm,
- a four-element uniform linear array (ULA) with an inter-microphone spacing of $d = 0.0425$ cm with the center of the array located at $(152.4, 19.05, 101.6)$,
- an isotropic speech source located at $(254, 190.5, 101.6)$ cm, and
- a spatially and temporally white Gaussian noise field with an SNR of 30 dB measured with respect to the convolved speech.

The location estimates are computed every 128 ms frame over a 60 second female speech signal (i.e., English). The sampling rate is 48 kHz. In order to increase spatial resolution, the cross-correlation functions are upsampled by a factor of 20 before forming the PSCM.

In order to ease complexity, the simulations assumed (correctly) that the source and array lie on the same plane. Thus, the parameter space was one-dimensional, and ranged from 0° to 179° degrees azimuth in increments of one degree, measured with respect to the array axis such that an angle of 90° indicates array broadside.

Notice that the absolute levels of the spatial spectra are irrelevant to performance. The key factor is the level of the false direction-of-arrivals (DOAs) relative to that of the true DOA, which in this case is 120 degrees azimuth. The bias of the DOA estimator increases as the room reverberation becomes stronger.

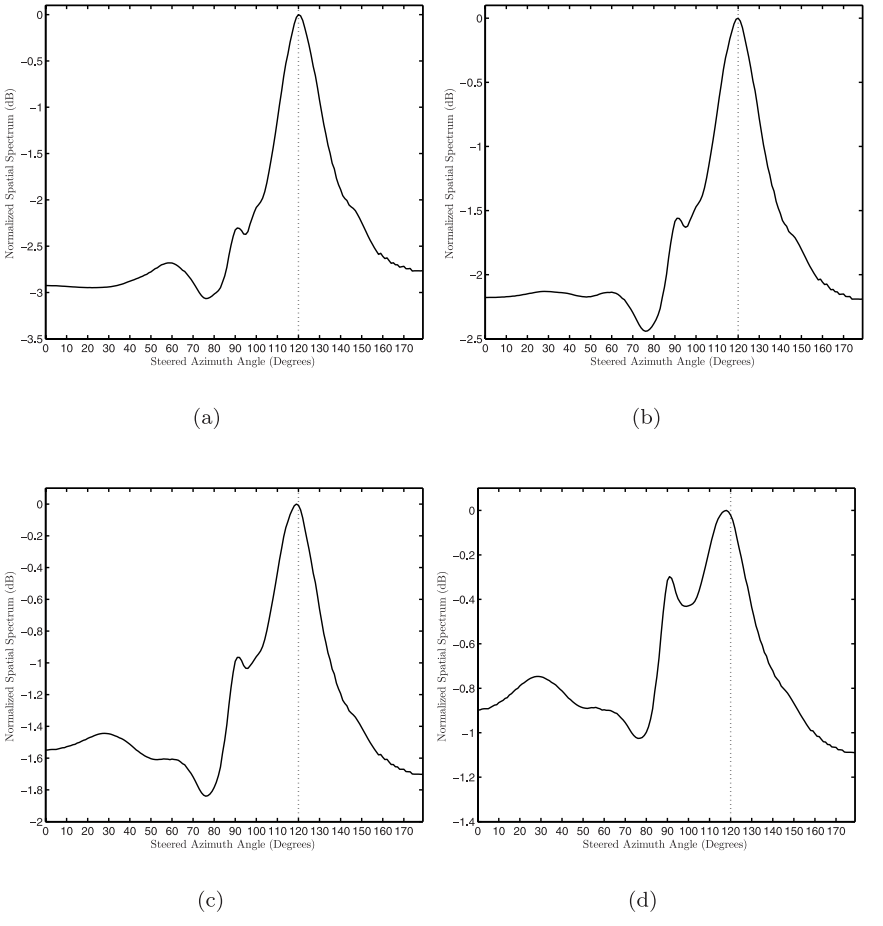

Fig. 12.1 Ensemble averaged SRP spatial spectra: (a) 0 ms, (b) 100 ms, (c) 200 ms, and (d) 300 ms.

12.5.2 Minimum Variance Distortionless Response

The SRP method is commonly employed due to its simplicity – the various PSCMs need to be formed and the elements summed. A more sophisticated method, proposed by Krolik and Swingler [24], attempts to apply a location-dependent filter $\mathbf{h}(\theta)$ to the PSCM in order to minimize the contribution of $\mathbf{R}_v(\theta)$ to $\mathbf{R}_y(\theta)$.

To that end, consider constraining $\mathbf{h}(\theta_s)$ to have unity gain in response to the properly-aligned source vector:

$$\mathbf{h}^T(\theta_\mathrm{s})\mathbf{s}(k,\theta_\mathrm{s}) = s(k), \tag{12.29}$$

which is equivalent to

$$\mathbf{h}^T(\theta_\mathrm{s})\mathbf{1}_N = 1, \tag{12.30}$$

while utilizing the remaining degrees of freedom in $\mathbf{h}(\theta_\mathrm{s})$ to minimize the beamformer output power $\mathbf{h}^T(\theta_\mathrm{s})\mathbf{R}_y(\theta_\mathrm{s})\mathbf{h}(\theta_\mathrm{s})$. This brings about the following optimization problem, which needs to be solved for every parameter θ:

$$\mathbf{h}_\mathrm{mvdr}(\theta) = \arg\min_{\mathbf{h}(\theta)} \mathbf{h}^T(\theta)\mathbf{R}_y(\theta)\mathbf{h}(\theta) \ \text{ subject to } \ \mathbf{h}^T(\theta)\mathbf{1}_N = 1. \tag{12.31}$$

This is an application of the celebrated minimum variance distortionless response (MVDR) method [25] to the steered-beamforming problem. Using the method of Lagrange multipliers, the optimal weights are found as

$$\mathbf{h}_\mathrm{mvdr}(\theta) = \frac{\mathbf{R}_y^{-1}(\theta)\mathbf{1}_N}{\mathbf{1}_N^T\mathbf{R}_y^{-1}(\theta)\mathbf{1}_N}. \tag{12.32}$$

Consequently, the location estimate is given by

$$\begin{aligned}\hat{\theta}_\mathrm{s} &= \arg\max_\theta \mathbf{h}_\mathrm{mvdr}^T(\theta)\mathbf{R}_y(\theta)\mathbf{h}_\mathrm{mvdr}(\theta) \\ &= \arg\max_\theta \left[\mathbf{1}_N^T\mathbf{R}_y^{-1}(\theta)\mathbf{1}_N\right]^{-1}.\end{aligned} \tag{12.33}$$

It is very interesting to see that while SRP sums the elements of the PSCM, MVDR inverts the sum of the PSCM's inverse's elements.

Figure 12.2 depicts the MVDR spatial spectra for the scenario described in the previous subsection. Notice that the spectra bear a strong resemblance to that of the SRP method.

12.5.3 Maximum Eigenvalue

In addition to selecting a weight vector that minimizes the contribution of the noise, one may also attempt to find a weight vector that simply maximizes the output power of the steered beamformer at each direction in order to identify the location θ at which the steered power $E\left[z^2(k,\theta)\right]$ is the largest [27].

This weight selection may be written as the following constrained optimization problem:

$$\mathbf{h}_\mathrm{maxeig}(\theta) = \arg\max_{\mathbf{h}(\theta)} \mathbf{h}^T(\theta)\mathbf{R}_y(\theta)\mathbf{h}(\theta) \ \text{ subject to } \ \mathbf{h}^T(\theta)\mathbf{h}(\theta) = 1. \tag{12.34}$$

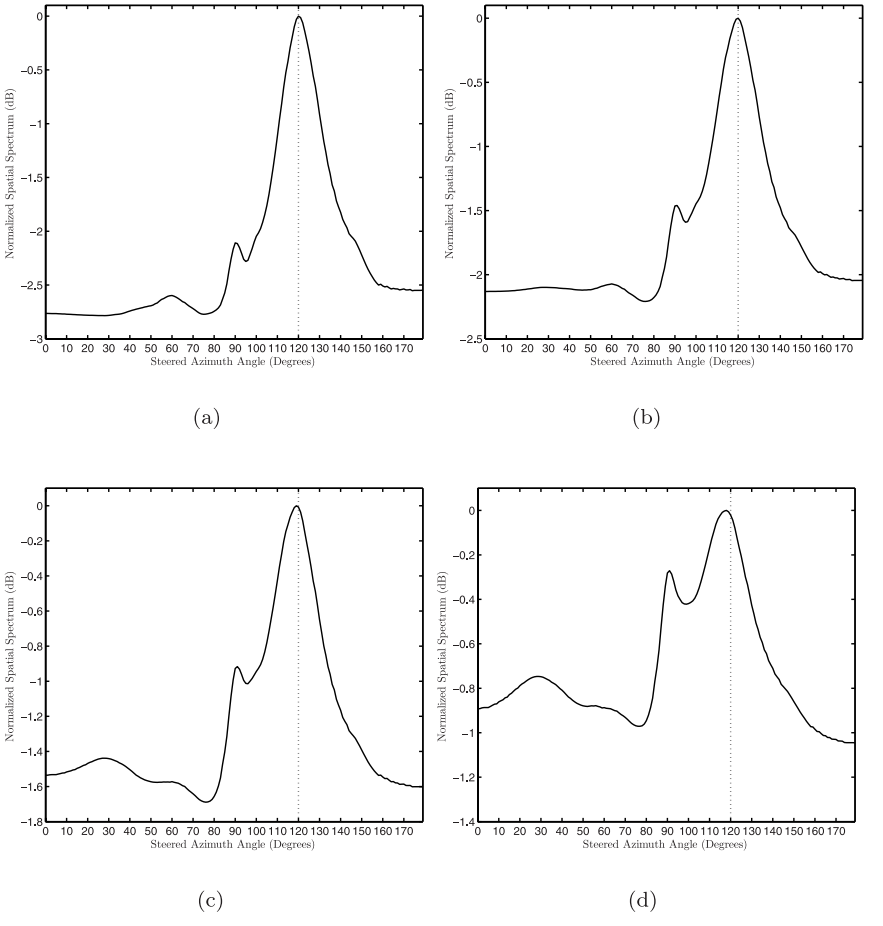

Fig. 12.2 Ensemble averaged MVDR spatial spectra: (a) 0 ms, (b) 100 ms, (c) 200 ms, and (d) 300 ms.

The solution to this optimization problem is the vector that maximizes the Rayleigh quotient

$$\frac{\mathbf{h}^T(\theta)\mathbf{R}_y(\theta)\mathbf{h}(\theta)}{\mathbf{h}^T(\theta)\mathbf{h}(\theta)}.$$

The solution is well known and given by the principal eigenvector of $\mathbf{R}_y(\theta)$:

$$\mathbf{h}_{\text{maxeig}}(\theta) = \mathbf{u}_1(\theta), \tag{12.35}$$

where $\mathbf{u}_1(\theta)$ is the principal eigenvector of $\mathbf{R}_y(\theta)$. Note again that the principal eigenvector must be found for all PSCMs. The location estimate follows as

$$\begin{aligned}\hat{\theta}_{\mathrm{s}} &= \arg\max_{\theta} \mathbf{h}_{\mathrm{maxeig}}^T(\theta)\mathbf{R}_y(\theta)\mathbf{h}_{\mathrm{maxeig}}(\theta) \\ &= \arg\max_{\theta} \lambda_1(\theta),\end{aligned} \tag{12.36}$$

where $\lambda_1(\theta)$ is the maximum eigenvalue of $\mathbf{R}_y(\theta)$.

Figure 12.3 depicts the maximum eigenvalue spatial spectra. Once again, the PSCM based adaptive weighting has little impact on the resulting spatial spectra.

12.5.4 Broadband MUSIC

The multiple signal classification (MUSIC) method is a classical method for the localization of narrowband signal sources [28], [29]. It is a subspace method which exploits the distinct eigenstructure of the SCM. In this section, it is shown that the parametrization of the SCM leads to a similar eigenstructure in the PSCM, thus allowing for the generalization of the MUSIC method to broadband signals [30].

Recall that the PSCM's signal component $\mathbf{R}_s(\theta)$ is rank-one if $\theta = \theta_{\mathrm{s}}$. Moreover, since the PSCM is a correlation matrix, it is positive-definite and thus has a spectral representation. Consider the case when $\theta = \theta_{\mathrm{s}}$; the PSCM may then be written as

$$\mathbf{R}_y(\theta_{\mathrm{s}}) = \sigma_s^2 \mathbf{1}_N \mathbf{1}_N^T + \mathbf{R}_v(\theta_{\mathrm{s}}), \tag{12.37}$$

where $\sigma_s^2 = E\left[s^2(k)\right]$ is the variance of the source signal. For the time being, assume that the parameterized noise correlation matrix may be written as

$$\mathbf{R}_v(\theta_{\mathrm{s}}) = \sigma_v^2 \mathbf{I}_{N\times N}, \tag{12.38}$$

where $\mathbf{I}_{N\times N}$ is the N-by-N identity matrix. It then follows that any vector $\mathbf{u}$ orthogonal to $\mathbf{1}_N$:

$$\mathbf{1}_N^T \mathbf{u} = 0 \tag{12.39}$$

is an eigenvector of $\mathbf{R}_y(\theta_{\mathrm{s}})$, since

$$\mathbf{R}_y(\theta_{\mathrm{s}})\mathbf{u} = \sigma_v^2 \mathbf{u}. \tag{12.40}$$

Since the dimensionality of $\mathbf{1}_N$ is N, there are $N-1$ such eigenvectors, which are termed the *noise eigenvectors* as their corresponding eigenvalues are all equal to the noise variance σ_v^2. Moreover, the remaining eigenvector must

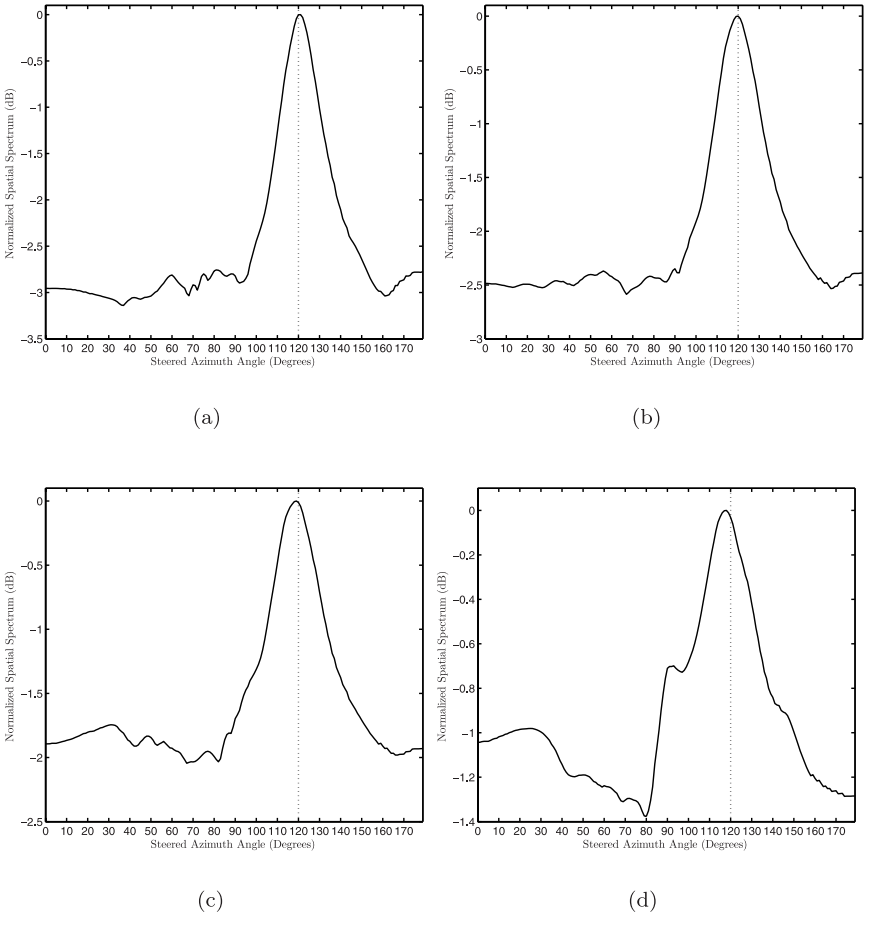

Fig. 12.3 Ensemble averaged maximum eigenvalue spatial spectra: (a) 0 ms, (b) 100 ms, (c) 200 ms, and (d) 300 ms.

necessarily be the principal eigenvector since it is not orthogonal to $\mathbf{1}_N$ – this eigenvector is termed the signal eigenvector, and the spectral representation of the location-matched PSCM is given by

$$\mathbf{R}_y(\theta_\mathrm{s}) = \lambda_1 \mathbf{u}_1(\theta_\mathrm{s})\mathbf{u}_1^T(\theta_\mathrm{s}) + \sigma_v^2 \sum_{i=2}^{N} \mathbf{u}_i(\theta_\mathrm{s})\mathbf{u}_i^T(\theta_\mathrm{s}), \tag{12.41}$$

where $\mathbf{u}_1(\theta_\mathrm{s})$ is the principal signal eigenvector of $\mathbf{R}_y(\theta_\mathrm{s})$ and $\mathbf{u}_i(\theta_\mathrm{s}), \quad i = 2, \ldots, N$ are the noise eigenvectors.

Notice that when the parameter θ does not match the actual location θ_{s}, the eigenvalue spectrum of the PSCM will effectively be spread out by the mismatch between θ and θ_{s}.

Putting all of this together, one may look for the source by examining the $N-1$ lower eigenvalues of the PSCM and test them for orthogonality with $\mathbf{1}_N$. At the actual source location θ_{s}, these $N-1$ lower eigenvalues theoretically yield

$$\frac{1}{\mathbf{1}_N^T \sum_{i=2}^{N} \mathbf{u}_i(\theta_{\mathrm{s}})\mathbf{u}_i^T(\theta_{\mathrm{s}})\mathbf{1}_N} = \infty. \tag{12.42}$$

Thus, the broadband MUSIC location estimate is given by

$$\hat{\theta}_{\mathrm{s}} = \arg\max_{\theta} \left[\mathbf{1}_N^T \left(\sum_{i=2}^{N} \mathbf{u}_i(\theta)\mathbf{u}_i^T(\theta) \right) \mathbf{1}_N \right]^{-1}. \tag{12.43}$$

Figure 12.4 depicts the broadband MUSIC spatial spectra. It is evident that the spectra show increased spatial resolution, analogous to the narrowband MUSIC method. In the presence of reverberation, the resulting spectra reveal false peaks corresponding to both multipath components and the autocorrelation of the speech signal.

12.5.5 Minimum Entropy

Given a random variable x with probability density function (pdf) $p(x)$, the entropy of the random variable is given by [31]

$$\begin{aligned} H(x) &= -E\left[\ln p(x)\right] \\ &= -\int_{-\infty}^{\infty} p(x) \ln p(x) dx, \end{aligned} \tag{12.44}$$

and quantifies the level of uncertainty associated with the random variable x. High probability values of x contribute less in the $-\ln p(x)$ term but are weighted more due to the $p(x)$ term. Conversely, low probability values of x have a large $-\ln p(x)$ contribution but these contributions are weighted less by the low value of $p(x)$. The value of $H(x)$ quantifies the average value of the uncertainty about x. The measure generalizes in a straightforward manner to a random vector $\mathbf{x}$ with joint pdf $p(\mathbf{x})$; the joint entropy of $\mathbf{x}$ is given by

$$\begin{aligned} H(\mathbf{x}) &= -E\left[\ln p(\mathbf{x})\right] \\ &= -\int_{-\infty}^{\infty} p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}. \end{aligned} \tag{12.45}$$

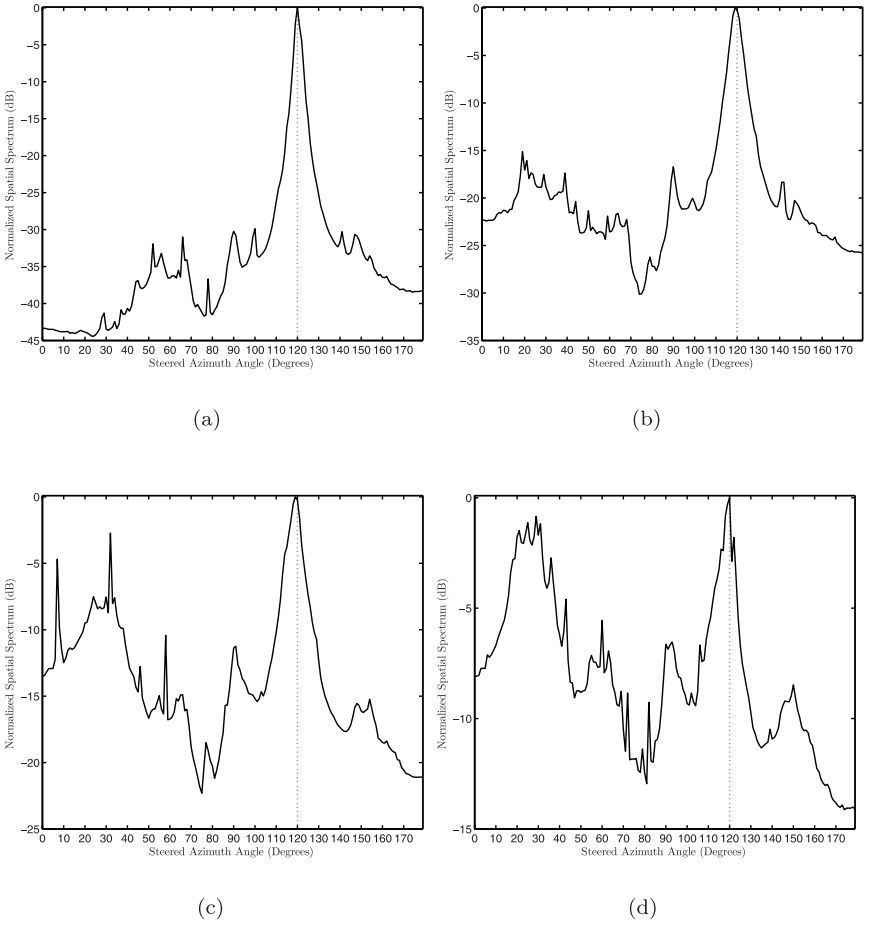

Fig. 12.4 Ensemble averaged MUSIC spatial spectra: (a) 0 ms, (b) 100 ms, (c) 200 ms, and (d) 300 ms.

A measure closely related to entropy is the mutual information, which quantifies the level of dependence between two random variables x and y:

$$I(x;y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\ln\frac{p(x,y)}{p(x)p(y)}dxdy, \tag{12.46}$$

where $p(x,y)$ is the joint distribution of x and y. One can show that the mutual information may be written as

$$\begin{aligned} I(x;y) &= H(x) - H(x|y) \\ &= H(y) - H(y|x) \\ &= H(x) + H(y) - H(x,y), \end{aligned} \tag{12.47}$$

where $H(x|y)$ is the entropy of x conditioned on y:

$$H(x|y) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\ln p(x|y)dxdy, \tag{12.48}$$

and $H(x,y)$ is the joint entropy of x and y:

$$H(x,y) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\ln p(x,y)dxdy. \tag{12.49}$$

Theorem 12.1. *The joint entropy of the random variables x and y attains a minimum when $p(x|y) = \delta(x-y) = p(y|x)$. In other words, joint entropy is minimized when the two random variables are equal at each ensemble.*

Proof. Assuming that $p(x|y) = \delta(x-y)$, the joint entropy of x and y may be written as

$$\begin{aligned} H(x,y) &= -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)\ln p(x,y)dxdy \\ &= -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x)p(y|x)\ln\left[p(x)p(y|x)\right]dxdy \\ &= -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x)\delta(x-y)\ln\left[p(x)\delta(x-y)\right]dxdy \\ &= -\int_{-\infty}^{\infty} p(y)\ln\left[p(y)\right]dy \\ &= H(y) \\ &= H(x) \\ &= H. \end{aligned} \tag{12.50}$$

Since the joint entropy is given by

$$H(x,y) = H(x) + H(y) - I(x;y), \tag{12.51}$$

it follows that

$$I(x;y) = H. \tag{12.52}$$

To prove that $H(x,y) \geq H(x)$, suppose that we could find two random variables x and y such that

$$H(x,y) < H(x). \tag{12.53}$$

This would imply that

$$H(x) + H(y) - I(x;y) < H(x), \tag{12.54}$$

or

$$I(x;y) > H(y). \tag{12.55}$$

However, since the mutual information is given by

$$I(x;y) = H(y) - H(y|x) \tag{12.56}$$

and since $H(y|x) = -E\left[\ln p(y|x)\right] \geq 0$, this implies a contradiction. Therefore, $H(x,y)$ is minimized when $p(x|y) = p(y|x) = \delta(x-y)$.

Applying this result to the problem of source localization, recall that $\mathbf{s}(k,\theta) = s(k)\mathbf{1}_N$ when $\theta = \theta_{\mathrm{s}}$. Thus, when steered to the true location, the elements of the random vector $\mathbf{s}(k,\theta)$ are fully dependent and their joint entropy is minimized. As a result, one can localize the source by scanning the location space for the location that minimizes the joint entropy of the $\mathbf{y}(k,\theta)$. Notice that the noise component is assumed to be incoherent across the array and thus varying the parameter theoretically does not reduce the entropy of $\mathbf{y}(k,\theta)$.

12.5.5.1 Gaussian Signals

In order to compute the minimum entropy estimate of the source location, one must assume a distribution for the random vector $\mathbf{y}(k,\theta)$. An obvious choice is the multivariate Gaussian distribution; the random vector $\mathbf{x}$ follows a multivariate Gaussian distribution with a mean vector of $\mathbf{0}_N$ and a covariance matrix $\mathbf{R}$ if its joint pdf is given by

$$p(\mathbf{x}) = \frac{1}{\left(\sqrt{2\pi}\right)^N \det^{1/2}(\mathbf{R})} e^{-1/2\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}}, \tag{12.57}$$

where $\det(\cdot)$ denotes the determinant of a matrix. The joint entropy of a Gaussian random vector is given by [32]

$$H(\mathbf{x}) = \frac{1}{2}\ln\left[(2\pi e)^N \det(\mathbf{R})\right]. \tag{12.58}$$

Thus, the entropy of a jointly distributed Gaussian vector is proportional to the determinant of the covariance matrix. Applying this to the source localization problem, the minimum entropy estimate of the location θ_{s} is given by [32]

$$\begin{aligned}\hat{\theta}_{\mathrm{s}} &= \arg\min_{\theta} H\left[\mathbf{y}(k,\theta)\right] \\ &= \arg\min_{\theta} \det\left[\mathbf{R}_y(\theta)\right]. \end{aligned} \tag{12.59}$$

It is interesting to link the minimum entropy approach to the eigenvalue methods presented earlier. To that end, notice that the determinant of a positive definite matrix is given by the product of its eigenvalues; thus, the minimum entropy approach may also be written as

$$\hat{\theta}_{\mathrm{s}} = \arg\min_{\theta} \prod_{n=1}^{N} \lambda_n(\theta). \tag{12.60}$$

Figure 12.5 depicts the minimum entropy spatial spectra. The minimum entropy estimator also shows increased resolution compared to the methods based on steered beamforming (i.e., SRP, MVDR, and maximum eigenvalue). However, as the level of reverberation is increased, spurious peaks are introduced into the spectra.

12.5.5.2 Laplacian Signals

The speech signal is commonly modeled by a Laplacian distribution whose heavier tail models the dynamic nature of speech; the univariate Laplacian distribution is given by

$$p(x) = \frac{\sqrt{2}}{2\sigma_x} e^{\frac{-\sqrt{2}|x|}{\sigma_x}}. \tag{12.61}$$

An N-dimensional zero-mean random vector is said to follow a jointly Laplacian distribution if its joint PDF is given by

$$p(\mathbf{x}) = 2(2\pi)^{-N/2}\det^{-1/2}(\mathbf{R})\left(\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}\right)^{P/2} K_P\left(\sqrt{2\mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}}\right), \tag{12.62}$$

where $P = \frac{2-N}{2}$ and $K_P(\cdot)$ is the modified Bessel function of the third kind:

$$K_P(a) = \frac{1}{2}\left(\frac{a}{2}\right)^P \int_0^{\infty} z^{-P-1} e^{-z-\frac{a^2}{4z}} dz, \quad a > 0. \tag{12.63}$$

It then follows that the joint entropy of a Laplacian distributed random vector is given by

$$H(\mathbf{x}) = \frac{1}{2}\ln\left[\frac{(2\pi)^N}{4}\det\mathbf{R}\right] - \frac{P}{2}E\left[\ln(\eta/2)\right] - E\left[\ln K_P\left(\sqrt{2\eta}\right)\right], \tag{12.64}$$

where $\eta = \mathbf{x}^T\mathbf{R}^{-1}\mathbf{x}$ and the expectation terms apparently lack closed forms.

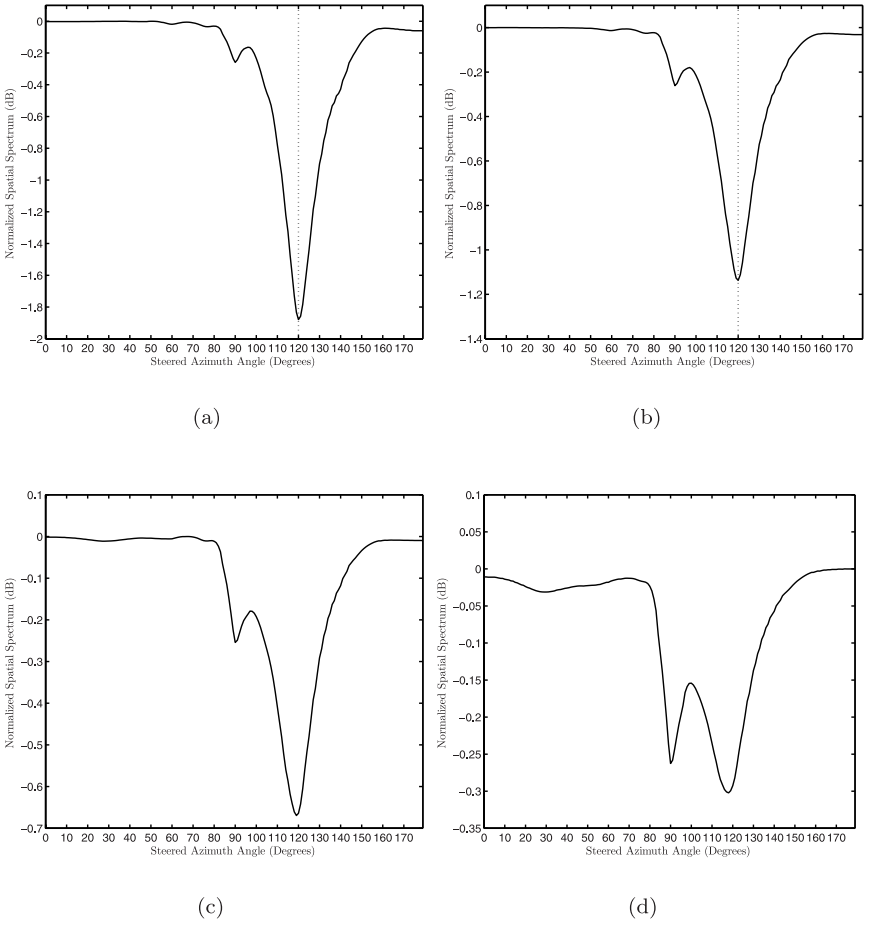

Fig. 12.5 Ensemble averaged minimum entropy spatial spectra (Gaussian assumption): (a) 0 ms, (b) 100 ms, (c) 200 ms, and (d) 300 ms.

A closed-form minimum entropy location estimator for the Laplacian case is thus not available; however, in practice, the assumption of ergodicity for the signals $y_n(k,\theta),\ n = 1,2,\ldots,N$ allows us to form an empirical minimum entropy estimator. To that end, consider first forming a time-averaged estimate of the PSCM:

$$\hat{\mathbf{R}}_y(\theta) = \frac{1}{K}\sum_{k'=1}^{K} \mathbf{y}(k',\theta)\mathbf{y}^T(k',\theta), \tag{12.65}$$

where $\mathbf{y}(k', \theta)$ is the k'th parameterized observation vector and there are K total observations. Next, the two terms lacking closed forms are estimated according to

$$E\left[\ln(\eta/2)\right] \approx \frac{1}{K}\sum_{k'=1}^{K} \ln\left[\frac{1}{2}\mathbf{y}^T\left(k', \theta\right)\hat{\mathbf{R}}(\theta)\mathbf{y}\left(k', \theta\right)\right], \quad (12.66)$$

$$E\left[\ln K_P\left(\sqrt{2\eta}\right)\right] \approx \frac{1}{K}\sum_{k'=1}^{K} \ln K_P\sqrt{2\mathbf{y}^T\left(k', \theta\right)\hat{\mathbf{R}}(\theta)\mathbf{y}\left(k', \theta\right)}. \quad (12.67)$$

The empirical joint entropy is then found by substituting the time-averaged quantities of (12.65)–(12.67) into the theoretical expression (12.64). As before, the parameter θ which minimizes the resulting joint entropy is chosen as the location estimate [32].

Notice that the minimum entropy estimators consider more than just second-order features of the observation vector $\mathbf{y}(k, \theta)$. The performance of this and all previously described algorithms depends ultimately on the sensitivity of the statistical criterion (i.e., joint Laplacian entropy) to the parameter θ, particularly to parameters which lead to a large noise presence in $\mathbf{y}(k, \theta)$.

12.6 Sparse Representation of the PSCM

In real applications, the computational complexity of a particular algorithm needs to be taken into account. The advantage of the algorithms presented in this chapter is the utilization of additional microphones to increase robustness to noise and reverberation. On the other hand, all algorithms inherently require a search of the parameter space to determine the optimal θ. In this section, we propose a sparse representation of the PSCM in terms of the observed cross-correlation functions across the array.

In practice, the cross-correlation functions across all microphone pairs are computed for a frame of incoming data. This is typically performed in the frequency-domain by taking the inverse Fourier transform of the cross-spectral density (CSD):

$$R_{y_n y_m}(\tau) = \frac{1}{L}\sum_{l=0}^{L-1} Y_n^*(l)Y_m(l)e^{j2\pi\frac{l}{L}\tau}, \quad (12.68)$$

where $Y_n^*(l)Y_m(l)$ is the instantaneous estimate of the CSD between channels n and m at discrete frequency $\frac{l}{L}$, superscript * denotes complex conjugate, and

$$Y_n(l) = \sum_{k=0}^{L-1} y_n(k) e^{-j2\pi \frac{l}{L} k} \tag{12.69}$$

is the L-point fast Fourier transform of the signal at microphone n evaluated at discrete frequency $\frac{l}{L}$. From the N^2 cross-correlation functions, the various PSCMs must be constructed and then evaluated to determine the location estimate $\hat{\theta}$:

$$[\mathbf{R}_y(\theta)]_{nm} = R_{y_n y_m} [\mathcal{F}_{nm}(\theta)] . \tag{12.70}$$

Thus, the task is to construct $\mathbf{R}_y(\theta)$ from the cross-correlation functions $R_{y_n y_m}(\tau)$.

Notice that the cross-correlation functions are computed prior to forming the various PSCMs. Moreover, for a given microphone pair, the cross-correlation function usually exhibits one or more distinct peaks across the relative delay space. Instead of taking into account the entire range of τ, it is proposed in [33] to only take into account the highly-correlated lags when forming the PSCM.

The conventional search technique relies on the forward mapping between the parameter θ and the resulting relative delay τ:

$$\tau_{nm} = \mathcal{F}_{nm}(\theta) \tag{12.71}$$

is the relative delay experience between microphones n and m if the source is located at θ. The problem with forming the PSCMs using the forward mapping is that the entire parameter space must be traversed before the optimal parameter is selected. Moreover, there is no *a priori* information about the parameter that can be utilized in reducing the search. Consider instead the inverse mapping from the relative delay τ to the set of locations which experience that relative delay at a given microphone pair:

$$\mathcal{F}_{nm}^{-1}(\tau) = \{\theta | \mathcal{F}_{nm}(\theta) = \tau\} . \tag{12.72}$$

For the microphone pair (n, m), define the set $\mathcal{C}_{nm}(p)$ which is composed of the $2p$ lags directly adjacent to the peak value of $R_{y_n y_m}(\tau)$:

$$\mathcal{C}_{nm}(p) = \{\hat{\tau}_{nm} - p, \ldots, \hat{\tau}_{nm} - 1, \hat{\tau}_{nm}, \hat{\tau}_{nm} + 1 \ldots, \hat{\tau}_{nm} + p\} , \tag{12.73}$$

where

$$\hat{\tau}_{nm} = \arg \max_{\tau} R_{y_n y_m}(\tau). \tag{12.74}$$

The set $\mathcal{C}_{nm}(p)$ hopefully contains the most correlated lags of of the cross-correlation function between microphones n and m. Consider nonlinearly processing the cross-correlation functions such that

Table 12.1 Localization using the sparse PSCM.

Compute:
for all microphone pairs (n, m)
$R_{y_n y_m}(\tau) = \frac{1}{L}\sum_{l=0}^{L-1} Y_n^*(l) Y_m(l) e^{j2\pi \frac{l}{L}\tau}$
$\hat{\tau}_{nm} = \arg\max_\tau R_{y_n y_m}(\tau)$
$\mathcal{C}_{nm}(p) = \{\hat{\tau}_{nm} - p, \ldots, \hat{\tau}_{nm} - 1, \hat{\tau}_{nm}, \hat{\tau}_{nm} + 1 \ldots, \hat{\tau}_{nm} + p\}$

Initialization:
for all θ, $\mathbf{R}_y(\theta) = \mathbf{0}_{N\times N}$

Search:
for all microphone pairs (n, m)
for all $\tau \in C_{nm}(p)$
look up $\mathcal{F}_{nm}^{-1}(\tau)$
for all $\theta \in \mathcal{F}_{nm}^{-1}(\tau)$
update : $[\mathbf{R}_y(\theta)]_{nm} = [\mathbf{R}_y(\theta)]_{nm} + R_{y_n y_m}(\tau)$

$\hat{\theta} = \arg\max_\theta f\left[\mathbf{R}_y(\theta)\right]$

$$R'_{y_n y_m}(\tau) = \begin{cases} R_{y_n y_m}(\tau), \tau \in \mathcal{C}_{nm}(p) \\ 0, \qquad \text{otherwise} \end{cases}. \tag{12.75}$$

The resulting elements of the PSCM are given by

$$\begin{aligned} \left[\mathbf{R}'_y(\theta)\right]_{nm} &= R'_{y_n y_m}\left[\mathcal{F}_{nm}(\theta)\right] \\ &= \begin{cases} R_{y_n y_m}\left[\mathcal{F}_{nm}(\theta)\right], & \mathcal{F}_{nm}(\theta) \in \mathcal{C}_{nm}(p) \\ 0, & \text{otherwise} \end{cases}. \end{aligned} \tag{12.76}$$

The modified PSCM $\mathbf{R}'_y(\theta)$ is now sparse provided that the sets $\mathcal{C}_{nm}(p)$ represent a small subset of the feasible relative delay space for each microphone pair.

Table 12.1 describes the general procedure for implementing a localization algorithm based on the sparse representation of the PSCM. As a comparison, Table 12.2 describes the corresponding algorithm but this time employing the forward mapping from location to relative delay. The conventional search involves iterating across the typically large location space. On the other hand, the sparse approach introduces a need to identify the peak lag of each cross-correlation function, albeit avoiding the undesirable location search.

Table 12.2 Localization using the PSCM.

Compute:
for all microphone pairs (n, m)
$R_{y_n y_m}(\tau) = \frac{1}{L} \sum_{l=0}^{L-1} Y_n^*(l) Y_m(l) e^{j2\pi \frac{l}{L}\tau}$

Initialization:
for all θ, $\mathbf{R}_y(\theta) = \mathbf{0}_{N \times N}$

Search:
for all locations θ
for all microphone pairs (n, m)
look up $\tau = \mathcal{F}_{nm}(\theta)$
update : $[\mathbf{R}_y(\theta)]_{nm} = [\mathbf{R}_y(\theta)]_{nm} + R_{y_n y_m}(\tau)$

$\hat{\theta} = \arg\max_\theta f[\mathbf{R}_y(\theta)]$

12.7 Linearly Constrained Minimum Variance

All approaches described thus far have focused on the relationship between the location of the acoustic source and the resulting relative delays observed across multiple microphones. Such techniques are purely spatial in nature. Notice that the resulting algorithms consider a temporally instantaneous aperture, in that previous samples are not appended to the vector of received spatial samples.

A truly spatiotemporal approach to acoustic source localization encompasses both spatial and temporal discrimination: that is, the aperture consists of a block of temporal samples for each microphone pair. The advantage of including previous temporal samples in the processing of each microphone is that the resulting algorithm may distinguish the desired signal from the additive noise by exploiting any temporal differences between them. This is the essence of the linearly constrained minimum variance (LCMV) adaptive beamforming method proposed by Frost in 1972 [34], which is equivalent to the generalized sidelobe canceller of [35], both of which are nicely summarized in [36]. The application of the LCMV scheme to the source localization algorithm is presented in [37].

The parameterized spatiotemporal aperture at the array is written as

$$\bar{\mathbf{y}}(k, \theta) = \left[\mathbf{y}(k, \theta)\ \mathbf{y}(k-1, \theta) \cdots \mathbf{y}(k-L+1, \theta) \right]^T, \qquad (12.77)$$

where we have appended the previous $L-1$ time-aligned blocks of length N to the aperture. With the signal model of (12.1) and assuming uniform attenuation coefficients, the parameterized spatiotemporal aperture is given by

$$\bar{\mathbf{y}}(k,\theta) = \bar{\mathbf{s}}(k-\tau,\theta) + \bar{\mathbf{v}}(k,\theta), \tag{12.78}$$

where

$$\bar{\mathbf{s}}(k,\theta) = \left[\mathbf{s}(k,\theta)\ \mathbf{s}(k-1,\theta)\ \cdots\ \mathbf{s}(k-L+1,\theta)\right]^T,$$
$$\bar{\mathbf{v}}(k,\theta) = \left[\mathbf{v}(k,\theta)\ \mathbf{v}(k-1,\theta)\ \cdots\ \mathbf{v}(k-L+1,\theta)\right]^T.$$

A location-parameterized multichannel finite impulse response (FIR) filter is formed according to

$$\mathbf{h}(\theta) = \left[\mathbf{h}_0^T(\theta)\ \mathbf{h}_1^T(\theta)\ \cdots\ \mathbf{h}_{L-1}^T(\theta)\right]^T, \tag{12.79}$$

where

$$\mathbf{h}_l(\theta) = \left[h_{l1}(\theta)\ h_{l2}(\theta)\ \cdots\ h_{lN}(\theta)\right]^T \tag{12.80}$$

is the spatial filter applied to the block of microphone signals at temporal sample $k-l$.

The question remains as to how to choose the multichannel filter coefficients such that the resulting steered spatiotemporal filter output allows one to better localize the source. In [37], it is proposed to select the weights such that the output of the spatiotemporal filter to a plane wave propagating from location θ is a filtered version of the desired signal:

$$\mathbf{h}^T(\theta)\,\mathbf{s}(k-\tau,\theta) = \sum_{l=0}^{L-1} f_l s(k-\tau-l). \tag{12.81}$$

In order to satisfy the desired criterion of (12.81), the multichannel filter coefficients should satisfy

$$\mathbf{c}_l^T(\theta)\mathbf{h}(\theta) = f_l, \quad l = 0,1,\ldots,L-1, \tag{12.82}$$

where

$$\mathbf{c}_l(\theta) = \left[\mathbf{0}_N^T\ \cdots\ \mathbf{0}_N^T\ \underbrace{\mathbf{1}_N^T}_{l\text{th group}}\ \mathbf{0}_N^T\ \cdots\ \mathbf{0}_N^T\right]^T$$

is a vector of length NL corresponding to the lth constraint, and $\mathbf{0}_N$ is a vector of N zeros. The L constraints of (12.82) may be neatly expressed in matrix notation as

$$\mathbf{C}^T(\theta)\mathbf{h}(\theta) = \mathbf{f}, \tag{12.83}$$

where

$$\mathbf{C}(\theta) = \left[\mathbf{c}_0(\theta)\ \mathbf{c}_1(\theta)\ \cdots\ \mathbf{c}_{L-1}(\theta)\right] \tag{12.84}$$

is the constraint matrix and

$$\mathbf{f} = \left[f_0 \; f_1 \; \cdots \; f_{L-1} \right]^T \tag{12.85}$$

is the constraint vector.

The spatiotemporal filter output is given by

$$z\left(k,\theta\right) = \mathbf{h}^T(\theta)\bar{\mathbf{y}}\left(k,\theta\right). \tag{12.86}$$

For each candidate location θ, we seek to find the multichannel weights $\mathbf{h}(\theta)$ which minimize the total energy of the beamformer output subject to the N linear constraints of (12.84):

$$\hat{\mathbf{h}}\left(\theta\right) = \arg\min_{\mathbf{h}(\theta)} \mathbf{h}^T(\theta)\mathbf{R}_{\bar{y}}(\theta)\mathbf{h}(\theta) \;\; \text{subject to} \;\; \mathbf{C}^T\left(\theta\right)\mathbf{h}(\theta) = \mathbf{f}, \tag{12.87}$$

where

$$\mathbf{R}_{\bar{y}}\left(\theta\right) = E\left\{\bar{\mathbf{y}}(k,\theta)\bar{\mathbf{y}}^T(k,\theta)\right\} \tag{12.88}$$

is the parameterized spatiotemporal correlation matrix (PSTCM), which is given by

$$\mathbf{R}_{\bar{y}}(\theta) = \begin{bmatrix} \mathbf{R}_y(\theta,0) & \mathbf{R}_y(\theta,-1) & \cdots & \mathbf{R}_y(\theta,-L+1) \\ \mathbf{R}_y(\theta,1) & \mathbf{R}_y(\theta,0) & \cdots & \mathbf{R}_y(\theta,-L+2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_y(\theta,L-1) & \mathbf{R}_y(\theta,L-2) & \cdots & \mathbf{R}_y(\theta,0) \end{bmatrix},$$

where it should be pointed out that $\mathbf{R}_y(\theta,0)$ is the PSCM.

The solution to the constrained optimization problem of (12.87) can be found using the method of Lagrange multipliers:

$$\hat{\mathbf{h}}\left(\theta\right) = \mathbf{R}_{\bar{y}}^{-1}(\theta)\mathbf{C}\left(\theta\right)\left[\mathbf{C}^T\left(\theta\right)\mathbf{R}_{\bar{y}}^{-1}(\theta)\mathbf{C}\left(\theta\right)\right]^{-1}\mathbf{f}. \tag{12.89}$$

Having computed the optimal multichannel filter for each potential source location θ, the estimate of the source location is given by

$$\hat{\theta}_{\mathrm{s}} = \arg\max_{\theta} \hat{\mathbf{h}}^T\left(\theta\right)\mathbf{R}_{\bar{y}}(\theta)\hat{\mathbf{h}}\left(\theta\right),$$

meaning that the source estimate is given by the location which emits the most steered (and temporally filtered) energy.

12.7.1 Autoregressive Modeling

It is important to point out that with $L = 1$ (i.e., a purely spatial aperture), the LCMV method reduces to the MVDR method if we select $\mathbf{f} = f_0 = 1$. In

this case, the PSTCM and PSCM are equivalent. Moreover, the constraint imposed on the multichannel filtering is

$$\mathbf{h}^T(\theta)\,\mathbf{s}(k-\tau,\theta) = s(k-\tau), \tag{12.90}$$

meaning that we are attempting to estimate the sample $s(k-\tau)$ from a spatial linear combination of the elements of $\mathbf{s}(k-\tau,\theta)$. Notice that such a procedure neglects any dependence of $s(k)$ on the previous values $s(k-1), s(k-2), \ldots$ of the signal. A signal whose present value is strongly correlated to its previous samples is well-modeled by an autoregressive (AR) process:

$$s(k) = \sum_{l=1}^{q} a_l s(k-l) + w(k), \tag{12.91}$$

where a_l are the AR coefficients, q is the order of the AR process, and $w(k)$ is the zero-mean prediction error.

Applying the AR model to the desired signal in the LCMV localization scheme, the constraint may be written as

$$\mathbf{h}^T(\theta)\,\mathbf{s}(k-\tau,\theta) = \sum_{l=1}^{L-1} a_l s(k-\tau-l), \tag{12.92}$$

where we have substituted

$$\begin{aligned} f_0 &= 0, \\ f_l &= a_l, \quad l = 1, 2, \ldots, q, \\ L-1 &= q. \end{aligned}$$

With the inclusion of previous temporal samples in the aperture, the LCMV scheme is able to temporally focus its steered beam onto the signal with the AR characteristics embedded by the coefficients in the constraint vector $\mathbf{f}$. Thus, the discrimination between the desired signal and noise is now both spatial (i.e., the relative delays differ since the source and interference are located at disparate locations) and temporal (i.e, the AR coefficients of the source and interference or noise generally differ).

It is important to point out that in general, the AR coefficients of the desired signal are not known *a priori*. Thus, the algorithm must first estimate the coefficients from the observed microphone signals. This can either be accomplished using conventional single-channel methods [38] or methods which incorporate the data from multiple channels [39].

Figure 12.6 depicts the ensemble averaged LCMV spatial spectra for the simulated data described previously. A temporal aperture length of $L = 20$ is employed. The AR coefficients are estimated from a single microphone by solving the Yule-Walker equations [38]. The PSTCM is regularized before performing the matrix inversion necessary in the method. The resulting spa-

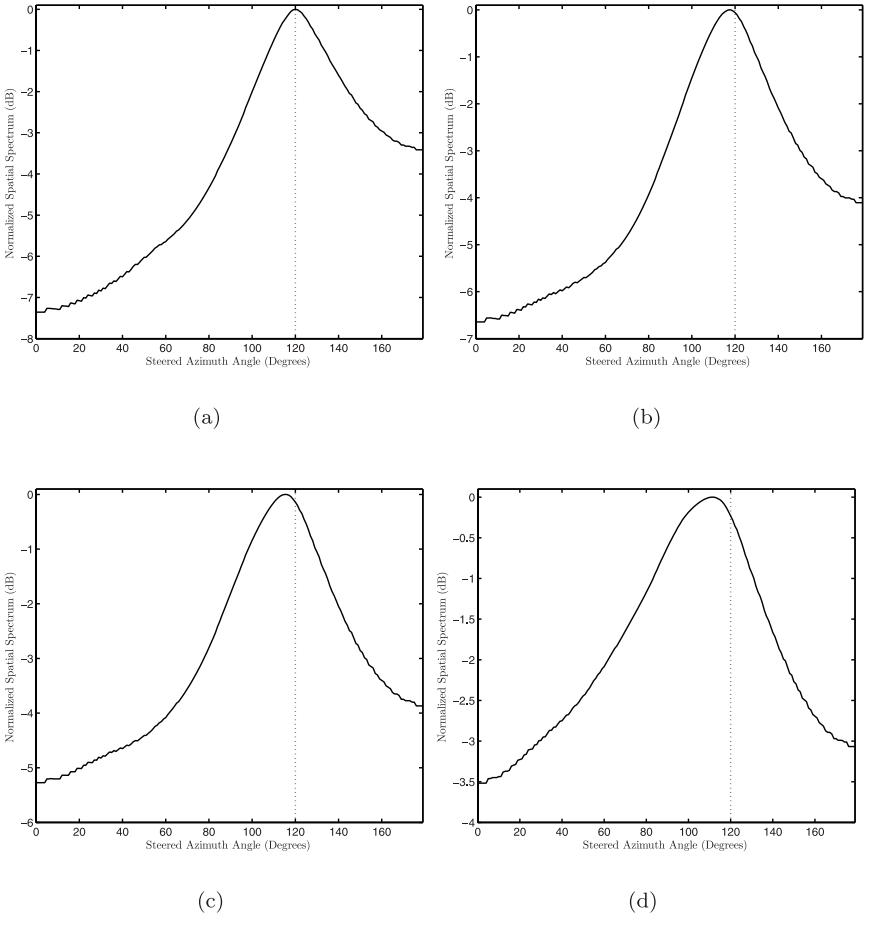

Fig. 12.6 Ensemble averaged LCMV spatial spectra: (a) 0 ms, (b) 100 ms, (c) 200 ms, and (d) 300 ms.

tial spectra are entirely free of any spurious peaks, albeit at the expense of a significant bias error.

12.8 Challenges

The techniques described in this chapter have been focused on integrating the information from multiple microphone in an optimal fashion to arrive at a robust estimate of the source location. While the algorithms represent

some of the more sophisticated approaches in acoustic source localization, the problem remains challenging due to a number of factors.

- Reverberant signal components act as mirror images of the desired signal but originating from a disparate location. Thus, the additive noise component is strongly correlated to the desired signal in this case. Moreover, the off-diagonal elements of the PSCM at false parameters θ may be large due to the reverberant signal arriving from θ.
- The desired signal (i.e., speech) is non-stationary, meaning that the estimation of necessary statistics is not straightforward.
- The parameter space is large, and the current solutions to broadband source localization require an exhaustive search of the location space.

The reverberation issue is particularly problematic. In the worst-case scenario, a reflected component may arrive at the array with an energy greater than the direct-path. At this point, most localization algorithms will fail, as the key assumption of localization is clearly violated: the true location of the source does not emit more energy than all other locations. Unlike in beamforming, the reverberant signal must be viewed as interference, as the underlying location of the reverberant path is different from that of the source.

12.9 Conclusions

This chapter has provided a treatment of steered beamforming approaches to acoustic source localization. The PSCM and PSTCM were developed as the fundamental structures of algorithms which attempt to process the observations of multiple microphones in such a way that the effect of interference and noise sources is minimized and the estimated source location possesses minimal error.

Purely spatial methods based on the PSCM focus on the relationship between the relative delays across the array and the corresponding source location. By grouping the various cross-correlation functions into the PSCM, well-known minimum variance and subspace techniques may be applied to the source localization problem. Moreover, an information-theoretic approach rooted in minimizing the joint entropy of the time-aligned sensor signals was developed for both Gaussian and Laplacian signals incorporating higher-order statistics in the source localization estimate. While PSCM-based methods are amenable to real-time operation, additional shielding of the algorithm from interference and reverberation may be achieved by extending the aperture to include the previous temporal samples of each microphone. It was shown that the celebrated LCMV method may be applied to the source localization problem by modeling the desired signal as an AR process.

The inclusion of multiple microphones in modern communication devices is relatively inexpensive. The desire for cleaner and crisper speech quality necessitates multiple-channel beamforming methods, which in turn require the localization of the desired acoustic source. By combining the outputs of the microphones via the PSCM and PSTCM, cleaner estimates of the source location may be generated using one of the methods detailed in this chapter.

In addition to the algorithms presented in this chapter, the PSCM and PSTCM provide a neat framework for the development of future localization algorithms aimed at solving the challenging problem of acoustic source localization.

References

1. H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Process. Mag.*, vol. 13, pp 67–94, July 1996.
2. D. H. Johnson, "The application of spectral estimation methods to bearing estimation problems," *Proc. IEEE*, vol. 70, pp. 1018–1028, Sept. 1982.
3. C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, pp. 320–327, Aug. 1976.
4. Y. Huang, J. Benesty, and G. W. Elko, "Microphone arrays for video camera steering," in *Acoustic Signal Processing for Telecommunication*, S. L. Gay and J. Benesty, eds., pp. 240–259. Kluwer Academic Publishers, Boston, MA, 2000.
5. Y. Huang, J. Benesty, and J. Chen, *Acoustic MIMO Signal Processing*. Springer-Verlag, Berlin, Germany, 2006.
6. Y. Huang, J. Benesty, and J. Chen, "Time delay estimation and source localization," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, editors-in-chief, Springer-Verlag, Chapter 51, Part I, pp. 1043–1064, 2007.
7. D. B. Ward and R. C. Williamson,"Particle filter beamforming for acoustic source localization in a reverberant environment," in *Proc. IEEE ICASSP*, 2002, pp. 1777–1780.
8. D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," *IEEE Trans. Speech, Audio Process.*, vol. 11, pp. 826–836, Nov. 2003.
9. E. A. Lehmann, D. B. Ward, and R. C. Williamson, "Experimental comparison of particle filtering algorithms for acoustic source localization in a reverberant room," in *Proc. IEEE ICASSP*, 2003, pp. 177–180.
10. A. M. Johansson, E. A. Lehmann, and S. Nordholm, "Real-time implementation of a particle filter with integrated voice activity detector for acoustic speaker tracking" in *IEEE Asia Pacific Conference on Circuits and Systems APPCCAS*, 2006, pp. 1004–1007.
11. C. E. Chen, H. Wang, A. Ali, F. Lorenzelli, R. E. Hudson, and K. Yao, "Particle filtering approach to localization and tracking of a moving acoustic source in a reverberant room," in *Proc. IEEE ICASSP*, 2006.
12. D. Li and Y. H. Hu, "Least square solutions of energy based acoustic source localization problems," in *Proc. International Conference on Parallel Processing (ICPP)*, 2004, pp. 443–446.
13. K. C. Ho and Ming Sun, "An accurate algebraic closed-form solution for energy-based source localization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, pp. 2542–2550, Nov. 2007.

14. D. Ampeliotis and K. Berberidis, "Linear least squares based acoustic source localization utilizing energy measurements," in *Proc. IEEE SAM*, 2008, pp. 349–352.
15. T. Ajdler, I. Kozintsev, R. Lienhart, and M. Vetterli, "Acoustic source localization in distributed sensor networks," in *Proc. Asilomar Conference on Signals, Systems, and Computers*, 2004, pp. 1328–1332.
16. G. Valenzise, G. Prandi, M. Tagliasacchi, and A. Sarti, "Resource constrained efficient acoustic source localization and tracking using a distributed network of microphones," in *Proc. IEEE ICASSP*, 2008, pp. 2581–2584.
17. H. Buchner, R. Aichner, J. Stenglein, H. Teutsch, and W. Kellennann, "Simultaneous localization of multiple sound sources using blind adaptive MIMO filtering," in *Proc. IEEE ICASSP*, 2005, pp. III-97–III-100.
18. A. Lombard, H. Buchner, and W. Kellermann, "Multidimensional localization of multiple sound sources using blind adaptive MIMO system identification," in *Proc. IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2006, pp. 7–12.
19. D. H. Johnson and D. E. Dudgeon, *Array Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
20. M. Omologo and P. G. Svaizer, "Use of the cross-power-spectrum phase in acoustic event localization," ITC-IRST Tech. Rep. 9303-13, Mar. 1993.
21. J. Dibiase, H.F. Silverman, and M.S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. B. Ward, eds., pp. 157–180, Springer-Verlag, Berlin, 2001.
22. M. R. Schroeder, "New method for measuring reverberation time," *J. Acoust. Soc. Am.*, vol. 37, pp. 409–412, 1965.
23. J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, Apr. 1979.
24. J. Krolik and D. Swingler, "Multiple broad-band source location using steered covariance matrices," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, pp. 1481–1494, Oct. 1989.
25. J. Capon, "High resolution frequency-wavenumber spectrum analysis," *Proc. IEEE*, vol. 57, pp. 1408–1418, Aug. 1969.
26. J. Dmochowski, J. Benesty, and S. Affes, "Direction-of-arrival estimation using the parameterized spatial correlation matrix," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, pp. 1327–1341, May 2007.
27. J. Dmochowski, J. Benesty, and S. Affes, "Direction-of-arrival estimation using eigenanalysis of the parameterized spatial correlation matrix," in *Proc. IEEE ICASSP*, 2007, pp. I-1–I-4.
28. R. O. Schmidt, *A Signal Subspace Approach to Multiple Emitter Location and Spectral Estimation*. Ph. D. dissertation, Stanford Univ., Stanford, CA, Nov. 1981.
29. R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. AP-34, pp. 276–280, Mar. 1986.
30. J. Dmochowski, J. Benesty, and S. Affes, "Broadband MUSIC: opportunities and challenges for multiple acoustic source localization," in *Proc. IEEE WASPAA*, 2007, pp. 18–21.
31. C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.
32. J. Benesty, Y. Huang, and J. Chen, "Time delay estimation via minimum entropy," *IEEE Signal Process. Lett.*, vol. 14, pp. 157–160, Mar. 2007.
33. J. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, pp. 2510–2516, Nov. 2007.
34. O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, pp. 926–935, Aug. 1972.

35. L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propagat.*, vol. AP-30, pp. 27–34, Jan. 1982.
36. B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, pp. 4–24, Apr. 1988.
37. J. Dmochowski, J. Benesty, and S. Affes, "Linearly constrained minimum variance source localization and spectral estimation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, pp. 1490–1502, Nov. 2008.
38. S. L. Marple Jr., *Digital Spectral Analysis with Applications*. Englewood Cliffs, NJ: Prentice Hall, 1987.
39. N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "Statistical analysis of the autoregressive modeling of reverberant speech," *J. Acoust. Soc. Am.*, vol. 120, pp. 4031–4039, Dec. 2006.