Chapter 12
Multimicrophone MMSE-Based Speech Source Separation

Shmulik Markovich-Golan, Israel Cohen and Sharon Gannot

Abstract Beamforming methods using a microphone array successfully utilize spatial diversity for speech separation and noise reduction. Adaptive design of the beamformer based on various minimum mean squared error (MMSE) criteria significantly improves performance compared to fixed, and data-independent design. These criteria differ in their considerations to noise minimization and desired speech distortion. Three common data-dependent beamformers, namely, matched filter (MF), MWF and LCMV are presented and analyzed. Estimation methods for implementing the various beamformers are surveyed. Simple examples of applying the various beamformers to simulated narrowband signals in an anechoic environment and to speech signals in a real-life reverberant environment are presented and discussed.

12.1 Introduction

In this chapter we introduce multimicrophone methods for speech separation and noise reduction methods, which are based on *beamforming*. Traditionally, beamforming methods are adopted from classical array processing techniques, in which a *beam* of high response is *steered* towards the desired source, while suppressing other directions. These methods were mainly applied in communications and radar domains. They usually assume free-field propagation, i.e., the angle-of-arrival fully determines the source position, although several design methods take multi-path propagation into account.

S. Markovich-Golan (✉)
Communication and Devices Group, Intel Corporation, Petah Tikva, Israel
e-mail: shmulik.markovich-golan@intel.com

S. Markovich-Golan · S. Gannot
Faculty of Engineering, Bar Ilan University, 5290002 Ramat-Gan, IsraelS. Gannot
e-mail: sharon.gannot@biu.ac.il

I. Cohen
Department of Electrical Engineering, Technion, 32000 Haifa, Israel
e-mail: icohen@ee.technion.ac.il

S. Makino (ed.), *Audio Source Separation*, Signals and Communication
Technology, https://doi.org/10.1007/978-3-319-73031-8_12

Statistically optimal beamformers are powerful multichannel filtering tools that optimize a certain design criteria while adapting to the received data, hence usually referred to as *data-dependent* approaches. A plethora of optimization criteria were proposed. The MVDR beamformer, also referred to as Capon beamformer [1], minimizes the noise power at the output of the beamformer subject to a unit gain constraint in the look direction. Frost [2] presented an adaptive implementation of the MVDR beamformer for wideband signals. Griffiths and Jim [3] proposed the GSC which is an efficient decomposition of the MVDR beamformer into two branches: one satisfying the constraint on the desired source, and the other for minimizing the noise (and interference).

Several researchers, e.g. [4] (see also Van Veen and Buckley [5]) have proposed modifications to the MVDR beamformer to deal with multiple linear constraints, denoted linearly constrained minimum variance (LCMV). Their work was motivated by the desire to apply further control to the array beampattern, beyond that of a steer-direction gain constraint. Hence, the LCMV can be applied to construct a beampattern satisfying certain constraints for a set of directions, while minimizing the array response in all other directions. Breed and Strauss [6] proved that the LCMV extension has also an equivalent GSC structure, which decouples the constraining and the minimization operations. The multichannel Wiener filter (MWF) is another important beamforming criterion, which minimizes the minimum mean squared error (MMSE) between the desired signal and the array output. It can be shown that the MWF decomposes into an MVDR beamformer followed by a single-channel Wiener post-filter [7]. A comprehensive analysis of beamformer criteria can be found in [5, 8].

Speech signals usually propagate in acoustic enclosures, e.g., rooms, and not in free-field. In the presence of obstacles, the sound wave is subject to diffractions and reflections, depending on its wavelength. Due to the typically small absorbtion coefficients of the obstacles, many successive wave reflections occur before their power decay. This induces multiple propagation paths between each source and each microphone, each with a different delay and attenuation factor. This phenomenon is often referred to as *reverberation*. The AIR (and its respective ATF) encompasses all these reflections and is usually a very long (few thousands taps) and time-varying filter.

Due to this intricate propagation regime, resorting to beampatterns as a function of the angle-of-arrival implies a reduction of a complex multi-parameter problem to an arbitrary single parameter problem. Classical beamformers, that construct their steering-vector under the assumption of free-field propagation, are often prone to performance degradation when applied in reverberant environments. It is therefore very important to take the reverberation effects into account while designing beamformers.

To circumvent the simplified free-field assumption, it was proposed [9, 10] to substitute the delay-only steering vector by the (normalized) ATFs relating the source and the microphones. This concept was later extended to the multiple sources scenario for extracting the desired source(s) from a mixture of desired and interference sources [11].

The MWF is also widely applied for speech enhancement, especially in its more flexible form, the speech distortion weighted-MWF (SDW-MWF) [12], which introduces a tradeoff factor that controls the amount of speech distortion versus the level of the residual noise.

A recent review paper [13] surveys many beamforming design criteria and their relation to BSS techniques.

12.2 Background

In this section we formulate the problem of speaker separation using spatial filtering methods. The signals and their propagation models are defined in Sect. 12.2.1. The following Sects. 12.2.2 and 12.2.3 are dedicated to defining spatial filters and criteria for evaluating their performances, respectively.

12.2.1 Generic Propagation Model

The speech sources are typically modeled in the STFT as quasi-stationary complex random processes with zero mean and time-varying variance, with stationarity time of the order of tens of milliseconds. Let us consider the case of J speech sources, denoted:

$$s_j(n, f) \sim \mathcal{N}\left(0, \phi_{s_j}(n, f)\right) \tag{12.1}$$

for $j = 1, \ldots, J$ where $\phi_{s_j}(n, f)$ denotes the time-varying signals spectra, and the indices $n = 0, 1, \ldots,$ and $f = 0, 1, \ldots, F - 1$ stand for the time-frame and frequency-bin index, and F denotes the STFT window length.

Given a microphone array comprising M microphones, the received microphone signals are given in an $M \times 1$ vector notation by

$$\mathbf{x}(n, f) = \sum_{j=1}^{J} \mathbf{c}_j(n, f) + \mathbf{u}(n, f) \tag{12.2}$$

where $\mathbf{c}_j(n, f)$ for $j = 1, \ldots, J$ denotes the J vectors of speech sources as received by the microphone array and $\mathbf{u}(n, f)$ denotes the $M \times 1$ dimensional vector comprising the noise components received at the microphones. Modeling the speech sources as coherent point sources and modeling the AIR as time-invariant convolution system, the speech components at the microphones are modeled in the STFT domain as a simple multiplication

$$\mathbf{c}_j(n, f) \triangleq \mathbf{a}_j(f) s_j(n, f) \tag{12.3}$$

where

$$\mathbf{a}_j(f) = \left[a_{j1}(f), \cdots a_{jI}(f) \right]^T \tag{12.4}$$

denotes the $M \times 1$ dimensional vector of ATFs relating the j-th source and the microphone array. Note that we assume that the STFT length is longer than the *effective length* of the AIR, such that convolution in the time-domain can be approximated as multiplication in the STFT domain (theoreticaly, only cyclic-convolution in the time-domain transforms to multiplication in the STFT domain [14]). Note that we assume that the AIR are time-invariant, i.e., the sources and the enclosure are static. This assumption can be relaxed to slowly time-varying environments, in which case the separating algorithm needs to adapt faster than the the system variations. However, for brevity we consider here time-invariant systems. The covariance matrices of the sources are given by:

$$\boldsymbol{\Phi}_{c_j}(n, f) \triangleq \mathrm{E}\left[\mathbf{c}_j(n, f)\mathbf{c}_j^H(n, f)\right] = \mathbf{a}_j(f)\mathbf{a}_j^H(f)\phi_{s_j}(n, f). \tag{12.5}$$

The noise-field is also assumed stationary, and the covariance matrix of its components at the microphone array is defined as:

$$\boldsymbol{\Phi}_u(f) = \mathrm{E}\left[\mathbf{u}(n, f)\mathbf{u}^H(n, f)\right]. \tag{12.6}$$

Note that the noise-stationarity assumption can be relaxed to slowly time-varying statistics, however, for ease of notation and derivation we assume that the noise is stationary.

12.2.2 Spatial Filtering

The spatial filter which is designed to extract the j-th speech source is denoted by $\mathbf{w}_j(n, f)$. Its corresponding output is defined by:

$$y_j(n, f) \triangleq \mathbf{w}_j^H(n, f)\mathbf{x}(n, f). \tag{12.7}$$

Note that generally the spatial filter may vary over time. By substituting (12.2) into (12.7), the output of the j-th spatial filter is decomposed into different components:

$$y_j(n, f) = \sum_{j'=1}^{J} d_{j,j'}(n, f) + v_j(n, f) \tag{12.8}$$

where

$$d_{j,j'}(n, f) \triangleq \mathbf{w}_j^H(n, f)\mathbf{c}_{j'}(n, f) \tag{12.9}$$

is the component that corresponds to the j'-th source at the output of the j-th spatial filter and

$$v_j(n,f) \triangleq \mathbf{w}_j^H(n,f)\mathbf{u}(n,f) \tag{12.10}$$

is the noise component at the j-th output. The aim of the j-th spatial filter is to maintain the j-th speech source, i.e., $d_{j,j}(n,f) \approx s_j(n,f)$, attenuate the other speech sources, i.e., $d_{j,j'}(n,f) \approx 0$ for $j' \neq j$, and reduce the noise, i.e., $v_j(n,f) \approx 0$. Note that aiming to obtain the *dry* signal of the j-th source (the original source before the convolution with the AIR) is a cumbersome task, and that in many practical scenarios obtaining the desired source as picked up by one of the microphones, which is denoted the *reference* microphone, is sufficient. Let us assume that the first microphone is selected as the reference microphone, and therefore the desired source at the output of the j-th spatial filter is $d_{j,j}(n,f) = a_{j1}(f)s_j(n,f)$. The RTFs relating the received components of the j'-th source at all microphones with its component at the reference microphone is defined as [10]:

$$\tilde{\mathbf{a}}_{j'}(f) \triangleq \frac{\mathbf{a}_{j'}(f)}{a_{j'1}(f)} \tag{12.11}$$

for $j' = 1, \ldots, J$. In the following sections, for the sake of clarity, we present the derivations of the various spatial filtering criteria using the ATF vectors rather than the RTF vectors.

12.2.3 Second-Order Moments and Criteria

Let us consider the output of the j-th spatial filter which aims to extract the j-th source while reducing noise and other interfering speakers, and define criteria to evaluate its performance. The difference between the output of the spatial filter and the desired signal is denoted as the error signal. The variance of the error signal, also known as the mean squared error (MSE) which we denote as $\chi_j(n,f)$, can be decomposed to its various components:

$$\begin{aligned}\chi_j(n,f) &\triangleq \mathrm{E}\left[\left|y_j(n,f) - s_j(n,f)\right|^2\right] \\ &= \delta_j(n,f) + \sum_{j' \neq j} \psi_{d_{j,j'}}(n,f) + \psi_{v\,j}(n,f)\end{aligned} \tag{12.12}$$

where

$$\delta_j(n,f) \triangleq \mathrm{E}\left[\left|s_j(n,f) - d_{j,j}(n,f)\right|^2\right] = |1 - \mathbf{w}_j^H(n,f)\mathbf{a}_j(f)|^2 \phi_{s_j}(n,f) \tag{12.13}$$

is the distortion of the j-th source component,

$$\psi_{d_{j,j'}}(n,f) \triangleq \mathrm{E}\left[\left|d_{j,j'}(n,f)\right|^2\right] = \left|\mathbf{w}_j^H(n,f)\mathbf{a}_{j'}(f)\right|^2 \phi_{s_{j'}}(n,f) \tag{12.14}$$

is the variance of the residual j'-th signal component, for $j' \neq j$ and

$$\psi_{v_j}(n,f) \triangleq \mathrm{E}\left[\left|v_j(n,f)\right|^2\right] = \mathbf{w}_j^H(n,f)\boldsymbol{\Phi}_u(f)\mathbf{w}_j(n,f) \tag{12.15}$$

is the variance of the residual noise component.

For evaluating the distortion level at the enhanced j-th signal, we define the signal-to-distortion ratio (SDR) as the power ratio of the desired speech component and its distortion:

$$\begin{aligned}\mathrm{SDR}_{\mathrm{o},j}(n,f) &\triangleq \frac{\phi_{s_j}(n,f)}{\delta_j(n,f)} \\ &= \frac{1}{\left|1-\mathbf{w}_j^H(n,f)\mathbf{a}_j(f)\right|^2}.\end{aligned} \tag{12.16}$$

To evaluate the noise reduction of the spatial filter we define the signal-to-noise ratio (SNR) improvement, denoted $\Delta\mathrm{SNR}_j$, which is the ratio of the SNR at the output and at the input, denoted $\mathrm{SNR}_{\mathrm{o},j}$ and $\mathrm{SNR}_{\mathrm{i},j}$:

$$\mathrm{SNR}_{\mathrm{i},j}(n,f) \triangleq \frac{\mathrm{trace}\left(\boldsymbol{\Phi}_{c_j}(n,f)\right)}{\mathrm{trace}\left(\boldsymbol{\Phi}_u(f)\right)} \tag{12.17a}$$

$$\mathrm{SNR}_{\mathrm{o},j}(n,f) \triangleq \frac{\psi_{dj,j}(n,f)}{\psi_{vj}(f)} \tag{12.17b}$$

$$\begin{aligned}\Delta\mathrm{SNR}_j(n,f) &\triangleq \frac{\mathrm{SNR}_{\mathrm{o},j}(n,f)}{\mathrm{SNR}_{\mathrm{i},j}(n,f)} \\ &= \frac{\left|\mathbf{w}_j^H(n,f)\mathbf{a}_j(f)\right|^2 / \mathbf{w}_j^H(n,f)\boldsymbol{\Phi}_u(f)\mathbf{w}_j(n,f)}{\|\mathbf{a}_j(f)\|^2/\mathrm{trace}\left(\boldsymbol{\Phi}_u(f)\right)}.\end{aligned} \tag{12.17c}$$

Note that the last expression of $\Delta\mathrm{SNR}_j$ is obtained by substituting the expressions from (12.5), (12.14), (12.15), (12.17a), and (12.17b).

The interfering speakers reduction is evaluated by using the SIR improvement, denoted as $\Delta\mathrm{SIR}_{jj'}$, defined for pairs of desired speaker and interfering speaker, denoted j and j' respectively, as the ratio the output SIR and the input SIR, denoted as $\mathrm{SIR}_{\mathrm{o},jj'}$ and $\mathrm{SIR}_{\mathrm{i},jj'}$:

$$\mathrm{SIR}_{\mathrm{i},jj'}(n,f) \triangleq \frac{\mathrm{trace}\left(\boldsymbol{\Phi}_{c_j}(n,f)\right)}{\mathrm{trace}\left(\boldsymbol{\Phi}_{c_{j'}}(n,f)\right)} \tag{12.18a}$$

$$\mathrm{SIR}_{\mathrm{o},jj'}(n,f) \triangleq \frac{\psi_{djj}(n,f)}{\psi_{djj'}(n,f)} \tag{12.18b}$$

$$\begin{aligned}\Delta\mathrm{SIR}_{jj'}(n,f) &\triangleq \frac{\mathrm{SIR}_{\mathrm{o},jj'}(n,f)}{\mathrm{SIR}_{\mathrm{i},jj'}(n,f)}\\ &= \frac{\left|\mathbf{w}_j^H(n,f)\mathbf{a}_j(f)\right|^2 / \left|\mathbf{w}_j^H(n,f)\mathbf{a}_{j'}(f)\right|^2}{\|\mathbf{a}_j(f)\|^2/\|\mathbf{a}_{j'}(f)\|^2}.\end{aligned} \tag{12.18c}$$

Note that the last expression of $\Delta\mathrm{SIR}_{jj'}$ is obtained by substituting the expressions from (12.5), (12.14), (12.18a) and (12.18b).

Finally, in order to evaluate the total interference and noise reduction, the signal-to-interference-and-noise ratio (SINR) improvement, denoted $\Delta\mathrm{SINR}_j$, is defined as the ratio of the SINR at the output and at the input, denoted $\mathrm{SINR}_{\mathrm{o},j}$ and $\mathrm{SINR}_{\mathrm{i},j}$:

$$\mathrm{SINR}_{\mathrm{i},j}(n,f) = \frac{\|\mathbf{a}_j(f)\|^2\phi_{s_j}(n,f)}{\sum_{j'\neq j}\|\mathbf{a}_{j'}(f)\|^2\phi_{s_{j'}}(n,f) + \mathrm{trace}\left(\boldsymbol{\Phi}_u(f)\right)} \tag{12.19a}$$

$$\mathrm{SINR}_{\mathrm{o},j}(n,f) = \frac{\psi_{dj,j}(n,f)}{\sum_{j'\neq j}\psi_{d_{j,j'}}(n,f) + \psi_{v_j}(f)} \tag{12.19b}$$

$$\begin{aligned}\Delta\mathrm{SINR}_j(n,f) = &\frac{\left|\mathbf{w}_j^H(n,f)\mathbf{a}_j(f)\right|^2}{\|\mathbf{a}_j(f)\|^2}\cdot\\ &\cdot\frac{\sum_{j'\neq j}\|\mathbf{a}_{j'}(f)\|^2\phi_{s_{j'}}(n,f) + \mathrm{trace}\left(\boldsymbol{\Phi}_u(f)\right)}{\sum_{j'\neq j}\left|\mathbf{w}_j^H(n,f)\mathbf{a}_{j'}(f)\right|^2\phi_{s_{j'}}(n,f) + \mathbf{w}_j^H(n,f)\boldsymbol{\Phi}_u(f)\mathbf{w}_j(n,f)}.\end{aligned} \tag{12.19c}$$

12.3 Matched Filter

In this section, the *matched filter* spatial filtering method is presented. Its design criterion is defined and explained in Sect. 12.3.1, and its performance is analyzed in Sect. 12.3.2. The MF based spatial filter was first introduced in [15], where it was implemented in the time domain.

12.3.1 Design

As suggested by its name, the matched filter is designed to match the ATFs of the desired source (here denoted as the j-th source). Formally, it is defined as:

$$\mathbf{w}_j^{\mathrm{MF}}(f) \triangleq \frac{\mathbf{a}_j(f)}{\|\mathbf{a}_j(f)\|^2} \tag{12.20}$$

where the scaling is designed to maintain a distortionless response towards the desired source, i.e.,

$$\left(\mathbf{w}_j^{\mathrm{MF}}(f)\right)^H \mathbf{a}_j(f) = 1 \tag{12.21}$$

and therefore $d_{jj'}(n, f) = s_j(n, f)$.

This criterion, can be shown optimal in the sense of maximizing the SNR at the output for the case of a single source contaminated by spatially white noise. The main advantage of this spatial filter lies in its simplicity as it is independent of the noise and interferences properties. In the special case of a desired source signal arriving from the far-field regime in an anechoic environment, the matched filter reduces to the well known delay-and-sum (DS) beamformer.

12.3.2 Performance

As stated in the previous section, the matched filter is designed to pass the j-th source undistorted. Hence, by substituting (12.21) in (12.13) we obtain that the distortion equals zero

$$\delta_j^{\mathrm{MF}}(n, f) = 0 \tag{12.22}$$

and by following (12.16) the SDR of the j-th source is infinite, i.e.:

$$\mathrm{SDR}_{\mathrm{o},j}^{\mathrm{MF}}(n, f) \rightarrow \infty. \tag{12.23}$$

Since the MF is designed independently of the noise and interference sound fields, the SIR and SNR improvements are accidental. The spectrogram of the j'-th interfering source at the output of the j-th output, $\psi_{d_{j,j'}}^{\mathrm{MF}}(n, f)$, and the corresponding SIR improvement of the j-th source with respect to the j'-th interfering source are:

$$\psi_{d_{j,j'}}^{\mathrm{MF}}(n, f) = \frac{\|\mathbf{a}_{j'}(f)\|^2}{\|\mathbf{a}_j(f)\|^2} \left|\rho_{jj'}(f)\right|^2 \phi_{s_{j'}}(n, f) \tag{12.24a}$$

$$\varDelta \mathrm{SIR}_{jj'}^{\mathrm{MF}}(f) = \frac{1}{\left|\rho_{jj'}(f)\right|^2} \tag{12.24b}$$

where $\rho_{jj'}(f)$ is defined as the normalized projection of the desired source ATF onto the interfering source ATF (per frequency-bin):

$$\rho_{jj'}(f) \triangleq \frac{\mathbf{a}_j^H(f)\mathbf{a}_{j'}(f)}{\|\mathbf{a}_j(f)\| \cdot \|\mathbf{a}_{j'}(f)\|}. \tag{12.25}$$

Note that from the Cauchy-Schwarz inequality the reciprocal of the SIR improvement expression is bounded by $0 \leq |\rho_{jj'}(f)|^2 \leq 1$, therefore the SIR improvement is bounded by $1 \leq \Delta\text{SIR}_{jj'}^{\text{MF}}(f) < \infty$. The SIR improvement will reach its upper-bound with the j'-th source being nulled by the MF of the j-th source if their corresponding ATFs are orthogonal (i.e., $\rho_{jj'}(f) = 0$).

By substituting (12.20) in (12.15) and (12.17c) the spectrum of the noise at the j-th output, and the corresponding SNR improvement are given by:

$$\psi_{v_j}^{\text{MF}}(f) = \frac{\mathbf{a}_j^H(f)\boldsymbol{\Phi}_u(f)\mathbf{a}_j(f)}{\|\mathbf{a}_j(f)\|^4} \tag{12.26a}$$

$$\Delta\text{SNR}_j^{\text{MF}}(f) = \frac{\|\mathbf{a}_j(f)\|^2 \cdot \text{trace}\left(\boldsymbol{\Phi}_u(f)\right)}{\mathbf{a}_j^H(f)\boldsymbol{\Phi}_u(f)\mathbf{a}_j(f)}. \tag{12.26b}$$

12.4 Multichannel Wiener Filter

In this section we present the MWF and analyze its performance.

12.4.1 Design

Considering the problem of enhancing the j-th source, recall that the MSE of an arbitrary spatial filter $\mathbf{w}(f)$ is denoted $\chi_j(n, f)$ and is defined by (12.12). The MSE is comprised of the following components: (a) distortion (denoted $\delta_j(n, f)$); (b) residual interferers spectra (denoted $\psi_{d_{j,j'}}(n, f)$, for $j' \neq j$); (c) residual noise spectrum (denoted $\psi_{v_j}(n, f)$). The MWF is designed to minimize the MSE expression:

$$\begin{aligned}\mathbf{w}_j^{\text{WF}}(n, f) &\triangleq \text{argmin}_{\mathbf{w}} \chi_j(n, f)\\ &= \frac{\left(\sum_{j'\neq j} \boldsymbol{\Phi}_{c_{j'}}(n, f) + \boldsymbol{\Phi}_u(f)\right)^{-1} \mathbf{a}_j(f)}{\mathbf{a}_j^H(f)\left(\sum_{j'\neq j} \boldsymbol{\Phi}_{c_{j'}}(n, f) + \boldsymbol{\Phi}_u(f)\right)^{-1} \mathbf{a}_j(f) + 1/\phi_{s_j}(n, f)}.\end{aligned} \tag{12.27}$$

Note that computing the MWF in (12.27) requires knowledge of: (a) power spectral density (PSD) of the desired source (denoted $\phi_{s_j}(n, f)$); (b) ATFs of the desired source (denoted $\mathbf{a}_j(f)$); (c) PSD matrices of the interferes (denoted $\boldsymbol{\Phi}_{c_{j'}}(n, f)$ for $j' \neq j$); (d) PSD matrix of the noise (denoted $\Phi_u(f)$). The estimation of these parameters is discussed in details in Sect. 12.6.

Although, practical methods exist for estimating the required parameters and implementing the MWF for the single source case, some relaxation is required when considering the multiple sources case. The *long-term averaged SOS* MWF is defined similarly to (12.27) by replacing the instantaneous PSD and PSD matrices of the desired source and interferers, respectively, with long-term averages:

$$\overline{\mathbf{w}}_j^{\mathrm{WF}}(f) = \frac{\left(\sum_{j'\neq j} \overline{\boldsymbol{\Phi}}_{c_{j'}}(f) + \boldsymbol{\Phi}_u(f)\right)^{-1} \mathbf{a}_j\,(f)}{\mathbf{a}_j^H\,(f)\left(\sum_{j'\neq j} \overline{\boldsymbol{\Phi}}_{c_{j'}}(f) + \boldsymbol{\Phi}_u(f)\right)^{-1} \mathbf{a}_j\,(f) + 1/\overline{\phi}_{s_j}(f)} \tag{12.28}$$

where

$$\bar{\phi}_{s_j}(f) \triangleq \frac{1}{\sum_n 1_{s_j}(n,f)} \sum_n 1_{s_j}(n,f)\mathrm{E}\left[|s_j(n,f)|^2\right] \tag{12.29a}$$

$$\overline{\boldsymbol{\Phi}}_{c_{j'}} \triangleq \frac{1}{\sum_n 1_{s_{j'}}(n,f)} \sum_n 1_{s_{j'}}(n,f)\mathrm{E}\left[\mathbf{c}_{j'}(n,f)\mathbf{c}_{j'}^H(n,f)\right] \tag{12.29b}$$

and $1_{s_{j'}}(n,f)$ denotes an indicator function which equals 1 for time-frequency bins in which the j'-th source is active, for $j' \in \{1,\ldots,J\}$.

12.4.2 Performance

By substituting (12.27) in (12.19c), the SINR improvement of the MWF can be shown to be

$$\begin{aligned}\Delta\mathrm{SINR}_j^{\mathrm{WF}}(n,f) = &\frac{1}{\|\mathbf{a}_j(f)\|^2}\left(\sum_{j'\neq j}\|\mathbf{a}_{j'}(f)\|^2\phi_{s_{j'}}(n,f) + \mathrm{trace}\left(\boldsymbol{\Phi}_u(f)\right)\right)\\ &\times \mathbf{a}_j^H(f)\left(\sum_{j'\neq j}\phi_{s_{j'}}(n,f)\mathbf{a}_{j'}(f)\mathbf{a}_{j'}^H(f) + \boldsymbol{\Phi}_u(f)\right)^{-1}\mathbf{a}_j(f).\end{aligned} \tag{12.30}$$

Note that the MWF allows to introduce distortion to the desired source, as long as it minimizes the variance of the total error between the desired source and the MWF output. The latter distortion may become high for example in low SIR cases when the number of interfering speech sources is larger than the number of microphones, i.e., $J-1>M$.

12.5 Multichannel LCMV

The criterion for designing the LCMV spatial filter is defined in Sect. 12.5.1, and its performance is analyzed in Sect. 12.5.2.

12.5.1 Design

Let us consider the design of the LCMV spatial filter which enhances the j-th speech source. The LCMV is designed to satisfy a set of J linear constraints, one for each speech source, that are defined by:

$$\mathbf{A}^H(f)\mathbf{w}_j^{\mathrm{LCMV}}(f) = \mathbf{g}_j \tag{12.31}$$

where

$$\mathbf{A}(f) \triangleq \left[\mathbf{a}_1(f), \cdots, \mathbf{a}_J(f)\right] \tag{12.32}$$

is the source ATFs matrix and

$$\mathbf{g}_j \triangleq \left[\mathbf{0}_{1\times(j-1)}\ 1\ \mathbf{0}_{1\times(J-j)}\right]^T \tag{12.33}$$

is the desired response for each of the sources and $\mathbf{w}_j^{\mathrm{LCMV}}(f)$ denotes the LCMV spatial filter at the f-th frequency-bin. Note that the desired response for the j-th source is $g_{j,j} = 1$, i.e., pass the j-th source undistorted, and the desired response for all other sources is $g_{j,j'} = 0$ for $j' \neq j$, i.e., null all other source.

The LCMV spatial filter is defined as the optimal solution of the following criterion:

$$\mathbf{w}_j^{\mathrm{LCMV}}(f) \triangleq \mathrm{argmin}_{\mathbf{w}} \mathbf{w}^H \boldsymbol{\Phi}_u(f)\mathbf{w};\ \text{ s.t. } \mathbf{A}^H(f)\mathbf{w} = \mathbf{g}_j \tag{12.34}$$

which aims to minimize the power of the noise at the output of the spatial filter (defined by (12.15)) while satisfying the linear constraints set, defined in (12.31). The closed-form solution of the optimization problem in (12.34) is given by:

$$\mathbf{w}_j^{\mathrm{LCMV}}(f) = \boldsymbol{\Phi}_u^{-1}(f)\mathbf{A}(f)\left(\mathbf{A}^H(f)\boldsymbol{\Phi}_u^{-1}(f)\mathbf{A}(f)\right)^{-1}\mathbf{g}_j. \tag{12.35}$$

An alternative form for implementing the LCMV, denoted GSC [3], conveniently separates the tasks of constraining the spatial filter and minimizing the noise variance. Additionally, the GSC can be efficiently implemented as a time-recursive procedure which tracks the noise statistics and, adapts to it, and converges to the optimal LCMV solution.

The performance and behavior of the LCMV are different than those of the MWF. On the one hand, the MWF gives equal weight to the three sources of error, i.e. distortion, interfering speakers and noise, when designing the spatial filter the sum of the three is minimized at the output. On the other hand, the LCMV maintains the desired signal undistorted and nulls (zero response) towards the interfering speech signals at the output. The remaining degrees of freedom (DoF) are designed to minimize the noise at the output. By doing so, conceptually, the LCMV gives significantly higher weights to the distortion and interfering speech components compared to the weight of the noise component. In [16] the multiple speech distortion weighted-MWF (MSDW-MWF) criterion which generalizes both MWF and LCMV criteria is defined. The latter enables component specific weights to each of the error sources at the output of the spatial filter. It extends the SDW-MWF to the multiple speakers case.

12.5.2 Performance

By design, the LCMV satisfies a set of J linear constraints, one per speech source. The constraint that corresponds to the j-th desired source is designed to maintain a distortionless response towards this source, and therefore the distortion equals zero

$$\delta_j^{\mathrm{LCMV}}(n, f) = 0 \tag{12.36}$$

and correspondingly the SDR is infinite

$$\mathrm{SDR}_{\mathrm{o},j}^{\mathrm{LCMV}}(n, f) \rightarrow \infty. \tag{12.37}$$

Similarly, as the rest of the $J - 1$ constraints are associated with interfering speech sources and are designed to null them out, their corresponding SIRs are infinite:

$$\varDelta\mathrm{SIR}_{jj'}^{\mathrm{LCMV}}(f) \rightarrow \infty. \tag{12.38}$$

By substituting (12.35) in (12.15) the noise variance at the output of the LCMV and the corresponding SNR improvement are:

$$\psi_{vj}^{\mathrm{LCMV}}(f) = \mathbf{g}_j^H \left(\mathbf{A}^H(f)\boldsymbol{\Phi}_u^{-1}\mathbf{A}(f)\right)^{-1} \mathbf{g}_j \tag{12.39a}$$

$$\varDelta\mathrm{SINR}_j^{\mathrm{LCMV}}(f) = \frac{\mathbf{g}_j^H \left(\mathbf{A}^H(f)\boldsymbol{\Phi}_u^{-1}\mathbf{A}(f)\right)^{-1} \mathbf{g}_j}{\sum_{j'\neq j} \|\mathbf{a}_{j'}(f)\|^2 \phi_{s_{j'}}(n, f) + \mathrm{trace}\,(\boldsymbol{\Phi}_u(f))}. \tag{12.39b}$$

12.6 Parameters Estimation

12.6.1 Multichannel SPP Estimators

Speech presence probability (SPP) is a fundamental and crucial component of many speech enhancement algorithms, among them are the spatial filters described in the previous sections. In the latter, SPP governs the adaptation of various components which contribute to the calculation of the spatial filter. Specifically, it can be used to govern the estimation of noise and speech covariance matrices (see Sect. 12.6.2) and of RTFs (see Sect. 12.6.3).

The problem of estimating SPP is derived from the classic detection problem, also known as the radar problem, and its goal is to identify the temporal-spectral activity pattern of speech contaminated by noise. Explicitly, determining if a time-frequency bin contains a noisy speech component or just noise. Contrary to the VAD problem where low resolution is sufficient, high-resolution activity estimation in both time and frequency is required here for proper enhancement. Most single-channel SPP estimators are based on non-stationarity of speech as opposed to the stationarity of the noise. However, in low SNR cases the accuracy of the estimation degrades.

When utilized for controlling the gain in single-channel postfiltering, the estimated SPP is "tuned" to have a tendency towards speech. This relates to the single-channel processing tradeoff between speech distortion and noise reduction, and to the common understanding that speech distortion and artifacts are more detrimental for human listeners than increased noise level. In difference to its use for single-channel enhancement, where the effect of SPP errors (i.e., false-alarms and miss-detections) is short-term (in time), in spatial processing the consequences of such errors can be grave and spreads over a longer period. Miss-detections of speech, and its false classification as noise might lead to a major distortion, also known as the self cancellation phenomenon. On the other hand false-alarms, i.e., time-frequency bins containing noise which are mistakenly classified as desired speech, result in increased noise level at the output of the spatial filter, since it is designed to pass them through.

Several contributions extend SPP estimation to utilize spatial information when using an array of microphones. Here we present some of these methods. In [17] which is presented in Sect. 12.6.1.1, the single channel Gaussian signal model of both speech and noise is extended to multichannel input, yielding a multichannel SPP. In Sect. 12.6.1.2, the work of [18], suggesting to incorporate the spatial information embedded in the direct-to-reverberant ratio (DRR) into the speech a priori probability (SAP), is presented. Thereby utilizing the coherence property of the speech source, assuming diffuse noise. Multichannel SPP incorporating spatial diversity can be utilized to address complex scenarios of multiple speakers. In [19, 20], the authors extend the previous DRR based SAP and incorporate estimated speaker positions to distinguish between different speakers, see Sect. 12.6.1.3.

12.6.1.1 Multichannel Gaussian Variables Model Based SPP

All derivations in this section refer to a specific time-frequency bin (n, f) and are replicated for all time-frequency bins. For brevity, the time and frequency indexes are omitted in the rest of this section. The received microphone signals

$$\mathbf{x} = \mathbf{c} + \mathbf{u} \tag{12.40}$$

and the speech and noise components thereof, are modeled as Gaussian random variables:

$$\mathbf{c} \sim \mathscr{N}_c\left(\mathbf{0}, \boldsymbol{\Phi}_c\right) \tag{12.41a}$$

$$\mathbf{u} \sim \mathscr{N}_c\left(\mathbf{0}, \boldsymbol{\Phi}_u\right) \tag{12.41b}$$

where $\boldsymbol{\Phi}_c = \phi_s \mathbf{a}\mathbf{a}^H$ is the covariance matrix of the speech image at the microphone signals. Consequently, a multichannel Gaussian model is adopted for the noise only, and noisy speech hypothesis:

$$\mathbf{x}|\mathscr{H}_u \sim \mathscr{N}_c\left(\mathbf{0}, \boldsymbol{\Phi}_u\right) \tag{12.42a}$$

$$\mathbf{x}|\mathscr{H}_s \sim \mathscr{N}_c\left(\mathbf{0}, \boldsymbol{\Phi}_c + \boldsymbol{\Phi}_u\right). \tag{12.42b}$$

It can be shown [17] that the SPP, defined as:

$$p \triangleq P\left(\mathscr{H}_s|\mathbf{x}\right) \tag{12.43}$$

can be formulated as

$$p = \frac{\Lambda}{1+\Lambda} \tag{12.44}$$

where Λ is the generalized likelihood ratio, which in our case equals

$$\Lambda = \frac{1-q}{q} \cdot \frac{1}{1+\operatorname{tr}\left\{\boldsymbol{\Phi}_u^{-1}\boldsymbol{\Phi}_c\right\}} \cdot \exp\left\{\frac{\boldsymbol{x}^H\boldsymbol{\Phi}_u^{-1}\boldsymbol{\Phi}_c\boldsymbol{\Phi}_u^{-1}\boldsymbol{x}}{1+\operatorname{tr}\left\{\boldsymbol{\Phi}_u^{-1}\boldsymbol{\Phi}_c\right\}}\right\}. \tag{12.45}$$

and q is the a priori speech absence probability. Define the multichannel SNR as

$$\xi \triangleq \operatorname{tr}\left\{\boldsymbol{\Phi}_u^{-1}\boldsymbol{\Phi}_c\right\} \tag{12.46}$$

and also define

$$\beta \triangleq \mathbf{x}^H\boldsymbol{\Phi}_u^{-1}\boldsymbol{\Phi}_c\boldsymbol{\Phi}_u^{-1}\mathbf{x}. \tag{12.47}$$

Substituting (12.45), (12.46) and (12.47) in (12.44) yields the multichannel SPP:

$$p = \left\{ 1 + \frac{q}{1-q} \cdot (1+\xi) \cdot \exp\left\{ -\frac{\beta}{1+\xi} \right\} \right\}^{-1}. \tag{12.48}$$

Note that the single-channel SPP (of the first microphone) can be derived as a special case of the multichannel SPP by substituting

$$\xi_1 = \frac{\Phi_{c,11}}{\Phi_{u,11}} \tag{12.49a}$$

$$\beta_1 = \gamma_1 \cdot \xi_1 \tag{12.49b}$$

with $\gamma_1 \triangleq \frac{|x_1|^2}{\Phi_{u,11}}$ defined as the posterior SNR and $\Phi_{c,11}$, $\Phi_{u,11}$ denote the speech and noise variances at the first microphone, respectively. The multichannel SPP can be interpreted as a single channel SPP applied to the output of an MVDR spatial filter designed to minimize the noise while maintaining a distortionless response towards the speech, with corresponding covariance matrices of $\boldsymbol{\Phi}_u$ and $\boldsymbol{\Phi}_c$, respectively.

The improvement of using the multichannel SPP depends on the spatial properties of the noise and of the speech. Two interesting special cases are the spatially white noise case and the coherent noise case. In the first case of a spatially white noise, the noise covariance matrix equals $\boldsymbol{\Phi}_u = \phi_u \mathbf{I}$ where $\mathbf{I}$ is the identity matrix. For this case the multichannel SNR equals $M \cdot \xi_1$ and is higher than the single-channel SNR by a factor of the number of microphones (assuming that the SNRs at all microphones are equal). In the second case of a coherent noise, the noise covariance matrix equals $\boldsymbol{\Phi}_u = \mathbf{a}_u \mathbf{a}_u^H \phi_{u,\mathrm{c}} + \phi_{u,\mathrm{nc}} \mathbf{I}$, where $\mathbf{a}_u$ and $\phi_{u,\mathrm{c}}$ are the vector of ATFs relating the coherent interference and the microphone signals and its respective variance. It is further assumed that the microphones also contain spatially white noise components with variance $\phi_{u,\mathrm{nc}}$. In this case, perfect speech detection is obtained, i.e., $p|\mathscr{H}_s \to 1$ and $p|\mathscr{H}_u \to 0$ regardless of the coherent noise power, assuming that the ATFs vectors of the speech and the coherent noise are not parallel and that the spatially white sensors noise power $\phi_{u,\mathrm{nc}}$ is sufficiently low.

12.6.1.2 Coherence Based SAP

As presented in Sect. 12.6.1.1, computing the SPP requires the speech a priori probability (SAP), denoted q. The SAP can be either set to a constant [21] or derived from the received signals and updated adaptively according to past estimates of SPP and SNR [22, 23] (also known as the *decision-directed* approach). In [19] the multichannel generalization of the SPP (see Sect. 12.6.1.1 and [17]) is adopted and it is proposed to incorporate coherence information in the SAP.

Let us consider a scenario where a single desired speech component contaminated by a diffuse noise is received by a pair of omnidirectional microphones. The diffuse noise field can be modeled as an infinite number of equal power statistically independent interferences uniformly distributed over a sphere surrounding the microphone array. A well known result [24] is that the coherence of diffuse noise components received by a pair of microphones is

$$\gamma_{\text{diff}}(\ell, \lambda) = \text{sinc}\left(\frac{2\pi\ell}{\lambda}\right) \tag{12.50}$$

were λ is the wavelength and ℓ is the microphones spacing. The direct to diffuse ratio (DDR) is defined as the SNR in this case, i.e., the power ratio of the directional speech received by the microphone and the diffuse noise. Heuristically, high and low DDR values are transformed into low and high SAP, respectively. The estimation of the DDR is based on a sound field model where the sound pressure at any position and time-frequency bin is modelled as a superposition of a direct sound represented by a single monochromatic plane wave and an ideal diffuse field, for more details please refer to [19, 25]. The DDR is estimated by:

$$\mathit{\Gamma} = \text{Re}\left\{\frac{\gamma_{\text{diff}} - \hat{\gamma}}{\hat{\gamma} - \exp(\text{j}\hat{\theta})}\right\} \tag{12.51}$$

where

$$\gamma \triangleq \frac{\text{E}\left[x_1 x_2^*\right]}{\sqrt{\text{E}\left[|x_1|^2\right] \cdot \text{E}\left[|x_2|^2\right]}} \tag{12.52}$$

is the coherence between the microphone signals and $\theta \triangleq \angle\left(c_1 \cdot c_2^*\right)$ is the phase between the speech components received by the microphones. The coherence is computed from estimates of the auto-PSDs and cross-PSD of the microphones (see Sect. 12.6.2), and the phase θ is approximated from the phase of the cross-PSD by:

$$\hat{\theta} = \angle\left(\text{E}\left[x_1 x_2^*\right]\right) \tag{12.53}$$

assuming that both SNR and DDR are high.

12.6.1.3 Multiple Speakers Position Based SPP

Consider the J speakers scenario in which the microphone signals can be formulated as:

$$\mathbf{x} = \sum_{j=1}^{J} \mathbf{c}_j + \mathbf{u}. \tag{12.54}$$

In [26], the authors propose to use a MWF for extracting a desired source from a multichannel convolutive mixture of sources. By incorporating position estimates into the SPP and classifying the dominant speaker per time-frequency point, the "interference" components' PSD matrix, comprising noise and interfering speakers, and the desired speaker components' PSD matrix are estimated and utilized for constructing a spatial filter. Speaker positions are derived by triangulation of DOA estimates obtained from distributed sub-arrays of microphones with known positions.

Individual sources SPPs are defined as:

$$p_j \triangleq p\left(\mathscr{H}_{s_j}|\mathbf{x}\right) = p\left(\mathscr{H}_{s_j}|\mathbf{x}, \mathscr{H}_s\right) p \tag{12.55}$$

where $\mathscr{H}_{s_j}$ denotes the hypothesis that the j-th speaker is active (per time-frequency point), and p is the previously defined SPP (for any speaker activity).

The conditional SPPs given the microphone signals are replaced by conditional SPPs given an estimate position of the dominant active speaker, denoted $\hat{\boldsymbol{\Theta}}$, i.e., it is assumed that:

$$p\left(\mathscr{H}_{s_j}|\hat{\boldsymbol{\Theta}}, \mathscr{H}_s\right) \approx p\left(\mathscr{H}_{s_j}|\mathbf{x}, \mathscr{H}_s\right). \tag{12.56}$$

The estimated position, given that a specific speaker is active, is modeled as a mixture of Gaussian variables centered at the sources' positions:

$$p\left(\hat{\boldsymbol{\Theta}}|\mathscr{H}_s\right) = \sum_{j=1}^{J} \pi_j \mathscr{N}\left(\hat{\boldsymbol{\Theta}}; \boldsymbol{\mu}_j, \boldsymbol{\Omega}_j\right) \tag{12.57}$$

where $\boldsymbol{\mu}_j$, $\boldsymbol{\Omega}_j$ and π_j are the mean, covariance and mixing coefficient of Gaussian vector distribution which corresponds to the estimated position of the j-th source, for $j = 1, \ldots, J$. The parameters of the distribution of $\hat{\boldsymbol{\Theta}}$ are estimated by an expectation maximization (EM) procedure given a batch of estimated positions. For a detailed explanation please refer to [26].

This work is further extended in [19], where a MWF is designed to extract sources arriving from a predefined "spot", i.e., a bounded area, while suppressing all other sources outside of the spot. This method is denoted by *spotforming*.

12.6.2 Covariance Matrix Estimators

The noise covariance matrix can be estimated by recursively averaging instantaneous covariance matrices weighted according to the SPP:

$$\begin{aligned}\widehat{\boldsymbol{\Phi}}_{\boldsymbol{u}}(n, f) =& \lambda'_u(n, f)\widehat{\boldsymbol{\Phi}}_{\boldsymbol{u}}(n-1, f) \\ &+ \left(1 - \lambda'_u(n, f)\right)\mathbf{x}(n, f)\mathbf{x}^H(n, f).\end{aligned} \tag{12.58}$$

where

$$\lambda'_u(n,f) \triangleq (1 - p(n,f))\,\lambda_u + p(n,f) \tag{12.59}$$

is a time-varying recursive averaging factor and λ_u is selected such that its corresponding estimation period ($\frac{1}{1-\lambda_u}$ frames) is shorter than the stationarity time of the noise. Alternatively, a hard binary weighting, obtained by applying a threshold to the SPP, can be used instead of the soft weighting.

The hypothesis that speaker j is present and the corresponding SPP are denoted in Sect. 12.6.1.1 as $\mathscr{H}_{s_j}(n,f)$ and $p_j(n,f)$, respectively. Similarly to (12.58), the covariance matrix of the spatial image of source j, denoted $\boldsymbol{\Phi}_{c_j}(n,f)$, can be estimated by

$$\begin{aligned}\widehat{\boldsymbol{\Phi}}_{\boldsymbol{c}_j}(n,f) &= \lambda'_{c_j}(n,f)\widehat{\boldsymbol{\Phi}}_{\boldsymbol{c}_j}(n-1,f)\\ &\quad + (1-\lambda'_{c_j}(n,f))(\mathbf{x}(n,f)\mathbf{x}^H(n,f) - \widehat{\boldsymbol{\Phi}}_{\boldsymbol{u}}(n-1,f))\end{aligned} \tag{12.60}$$

where

$$\lambda'_{c_j}(n,f) \triangleq \big(1 - p_j(n,f)\big)\,\lambda_c + p_j(n,f) \tag{12.61}$$

is a time-varying recursive-averaging factor, and λ_c is selected such that its corresponding estimation period ($\frac{1}{1-\lambda_c}$ frames) is shorter than the *coherence time* of the AIRs of speaker j, i.e. the time period over which the AIRs are assumed to be time-invariant. Note that: (1) usually the estimation period is longer than the speech nonstationarity time, therefore, although the spatial structure of $\boldsymbol{\Phi}_{c_j}(n,f)$ is maintained, the estimated variance is an average of the speech variances over multiple time periods, denoted $\bar{\phi}_{s_j}(n,f)$, rather than $\phi_{s_j}(n,f)$, the actual time-varying variance of the speaker ; (2) the estimate $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{c}_j}(n,f)$ keeps its past value when speaker j is absent.

12.6.3 Procedures for Semi-blind RTF Estimation

Two common approaches for RTF estimation are the covariance subtraction [27, 28] and the covariance whitening [11, 29] methods. Here, for brevity we assume a single speaker scenario. Both of these approaches rely on estimated noisy speech and noise-only covariance matrices, i.e. $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{xj}}(n,f)$ (where $\boldsymbol{\Phi}_{\boldsymbol{xj}}(n,f) = \boldsymbol{\Phi}_{c_j}(n,f) + \boldsymbol{\Phi}_u(f)$) and $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{u}}(n,f)$. Given the estimated covariance matrices, covariance subtraction estimates the speaker RTF by

$$\tilde{\mathbf{a}}_{j,\mathrm{CS}}(f) \triangleq \frac{1}{\mathbf{i}_1^H(\widehat{\boldsymbol{\Phi}}_{\boldsymbol{c}_j}(n,f) - \widehat{\boldsymbol{\Phi}}_{\boldsymbol{u}}(n,f))\mathbf{i}_1}(\widehat{\boldsymbol{\Phi}}_{\boldsymbol{c}_j}(n,f) - \widehat{\boldsymbol{\Phi}}_{\boldsymbol{u}}(n,f))\mathbf{i}_1 \tag{12.62}$$

where $\mathbf{i}_1 = [\,1\ \mathbf{0}_{1\times M-1}\,]^T$ is an $M \times 1$ selection vector for extracting the component of the reference microphone, here assumed to be the first micrphone.

The covariance whitening approach estimates the RTF by: (1) applying the generalized eigenvalue decomposition (GEVD) to $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{xj}}(n, f)$ with $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{u}}(n, f)$ as the whitening matrix; (2) de-whitening the eigenvector corresponding to the strongest eigenvalue, denoted $\dot{\mathbf{a}}_j(f)$, namely $\widehat{\boldsymbol{\Phi}}_{\boldsymbol{u}}(n, f)\dot{\mathbf{a}}_j(f)$; (3) normalizing the de-whitened eigenvector by the reference microphone component. Explicitly:

$$\tilde{\mathbf{a}}_{j,\mathrm{CW}}(f) \triangleq (\mathbf{i}_1^H \widehat{\boldsymbol{\Phi}}_{\boldsymbol{u}}(n, f)\dot{\mathbf{a}}_j(f))^{-1} \widehat{\boldsymbol{\Phi}}_{\boldsymbol{u}}(n, f)\dot{\mathbf{a}}_j(f). \tag{12.63}$$

A preliminary analysis and comparison of the covariance subtraction and covariance whitening methods can be found in [30].

Other methods utilize the speech nonstationarity property, assuming that the noise has slow time-varying statistics. In [10], the problem of estimating the RTF of microphone i is formulated as a least squares (LS) problem where the l-th equation utilizes $\widehat{\phi}^l_{x_i x_1}$(f), the estimated cross-PSD of microphone i and the reference microphone in the l-th time segment. This cross-PSD satisfies:

$$\widehat{\phi}^l_{xj,i1}(f) = \tilde{a}_{j,i}(f)\widehat{\phi}^l_{xj,11}(f) + \widehat{\phi}_{\dot{u}i,xj1}(f) + \varepsilon^l_{j,i}(f) \tag{12.64}$$

where we use the relation $\mathbf{x}(n, f) = \tilde{\mathbf{a}}_j(f)x_1(n, f) + \dot{\mathbf{u}}(n, f)$. The unknowns are $\tilde{a}_{j,i}(f)$, i.e. the required RTF, and $\widehat{\phi}_{\dot{u}i,xj1}(f)$, which is a nuisance parameter. $\varepsilon^l_{j,i}(f)$ denotes the error term of the l-th equation. Multiple LS problems, one for each microphone, are solved for estimating the vector RTF. Note that, the latter method, also known as the *nonstationarity*-based RTF estimation, does not require a prior estimate of the noise covariance, since it simultaneously solves for RTF and the noise statistics. Similarly, a weighted least squares (WLS) problem with exponential weighting can be defined and implemented using a recursive least squares (RLS) algorithm [31]. Considering speech sparsity in the STFT domain, in [28] the SPPs were incorporated into the weights of the WLS problem, resulting in a more accurate solution.

12.7 Examples

In the following section some simple examples are used to present the behaviors and differences between the MF, MWF and LCMV spatial filters. The case of a narrowband signal in an anechoic environment is presented in Sect. 12.7.1, and the case of two speech sources in a reverberant environment is presented in Sect. 12.7.2.

12.7.1 Narrowband Signals at an Anechoic Environment

Consider the case of $J = 2$ narrowband sources occupying the f-th frequency-bin propagating in an anechoic environment and received by a uniform linear array (ULA) array comprising M microphones with microphone spacing ℓ. Define a spherical-coordinate system with the origin coincides with the center of the ULA, and rotated such that the ULA is placed along the elevation angle $\theta = \pm 90^\circ$ and azimuth angle of $\phi = 0^\circ$. The sources are positioned in the far-field at a large distance from the microphones and on the same plane as the microphones. The DOA of the sources with respect to the microphones array are denoted θ_j for $j = 1, 2$. By adopting the far-field free-space propagation model the ATF vectors of the sources are given by:

$$\mathbf{a}_j^0(f) = \left[1,\ \exp\left(-\mathrm{j}2\pi \frac{\ell \sin(\theta_j)}{\lambda}\right),\ \cdots,\ \exp\left(-\mathrm{j}2\pi \frac{(M-1)\ell \sin(\theta_j)}{\lambda}\right)\right]^T \tag{12.65}$$

for $j = 1, 2$ where λ is the wavelength corresponding to the f-th frequency-bin. The wavelength can be expressed as:

$$\lambda \triangleq \frac{\nu F}{f f_s} \tag{12.66}$$

with the continuous frequency which corresponds to the discrete frequency-bin f is $\frac{f f_s}{F}$ where f_s is the sample-rate, F is the length of STFT window and $\nu \approx 343$ m/s is the sound velocity. An additive white Gaussian noise with covariance matrix of

$$\boldsymbol{\Phi}_u^0(n, f) = \phi_u \mathbf{I}. \tag{12.67}$$

is contaminating the received microphone signals. The setup of the sources and microphones is depicted in Fig. 12.1. Note that the ATF vector is independent of the azimuth angle ϕ, and therefore the beampattern and all performance measures of any spatial filter in this case will have a cylindrical symmetry. Next, we compare the performance of various spatial filters that are applied in this problem, namely MF, MWF and LCMV. The performance criteria that we use are SNR, SIR, SINR and SDR which are evaluated empirically from the signals. For the MF spatial filter we derive simplified expressions for the performance criteria, whereas for the MWF and the LCMV spatial filters we use the previously defined generic scenario expressions.

Let us revisit the performance criteria of the MF for this case. By substituting the ATF vectors in (12.65), the scalar product of the j-th and j'-th ATF vectors, denoted by $\rho_{jj'}(f)$ in (12.25), can be expressed as:

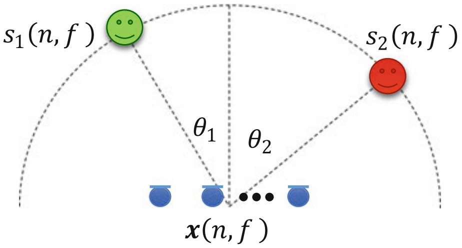


Fig. 12.1 Setup of the narrowband signal at an anechoic environment example

$$\begin{aligned}\rho^0_{jj'}(f) =&\sum_{i=1}^{M} \exp\left(\mathrm{j}2\pi\frac{(i-1)\,\ell\left(\sin\left(\theta_j\right)-\sin\left(\theta_{j'}\right)\right)}{\lambda}\right)\\ =&M\cdot\mathrm{diric}\left(2\pi\frac{\ell\left(\sin\left(\theta_j\right)-\sin\left(\theta_{j'}\right)\right)}{\lambda}\right)\cdot\exp\left(\mathrm{j}\pi\frac{(M-1)\,\ell\left(\sin\left(\theta_j\right)-\sin\left(\theta_{j'}\right)\right)}{\lambda}\right)\end{aligned} \tag{12.68}$$

where

$$\mathrm{diric}\left(2\pi\frac{\ell\left(\sin\left(\theta_j\right)-\sin\left(\theta_{j'}\right)\right)}{\lambda}\right) \triangleq \frac{\sin\left(M\pi\frac{\ell\left(\sin\left(\theta_j\right)-\sin\left(\theta_{j'}\right)\right)}{\lambda}\right)}{M\cdot\sin\left(\pi\frac{\ell\left(\sin\left(\theta_j\right)-\sin\left(\theta_{j'}\right)\right)}{\lambda}\right)} \tag{12.69}$$

is the Dirichlet function which in general has a period of 4π. Note that $\left|\rho^0_{jj'}(f)\right|^2 = M^2$ for

$$\frac{\ell\left(\sin\left(\theta_j\right)-\sin\left(\theta_{j'}\right)\right)}{\lambda} = k \tag{12.70}$$

where $k = 0, \pm 1, \pm 2, \ldots$ is any integer number. Next, since the $\sin(\cdot)$ is bounded by $-1 \leq \sin(\cdot) \leq 1$ the left-hand side of (12.70) is bounded by:

$$-\frac{2\ell}{\lambda} \leq \frac{\ell\left(\sin\left(\theta_j\right)-\sin\left(\theta_{j'}\right)\right)}{\lambda} \leq \frac{2\ell}{\lambda}. \tag{12.71}$$

Hence, in order to avoid the spatial aliasing phenomenon, where undesired directions are passed through the spatial filter without any attenuation, the well-known constraint on the ratio between microphones spacing and the wavelength is given by:

$$\frac{\ell}{\lambda} < \frac{1}{2}. \tag{12.72}$$

Furthermore, note that for

$$M\frac{\ell\left(\sin\left(\theta_j\right)-\sin\left(\theta_{j'}\right)\right)}{\lambda}=k \tag{12.73}$$

for any integer k non-divisible by M with the a zero remainder, i.e. of the form $k \neq \iota M$ where ι is an integer, we obtain that $\left|\rho^0_{jj'}(f)\right|^2 = 0$. Explicitly, in the range of $-\frac{\pi}{2} \leq \theta_{j'} \leq \frac{\pi}{2}$ there are $M-1$ such DOAs that are perfectly attenuated by the MF, also referred to as nulls in the beampattern. By replacing $\rho_{jj'}(f)$ with $\rho^0_{jj'}(f)$ in the power of the residual j'-th interference at the j-th output, see (12.24a), and the corresponding SIR improvement of the MF, see (12.24b), the following simplified expressions are obtained:

$$\psi^{0,\mathrm{MF}}_{d_{j,j'}}(n,f) = \mathrm{diric}^2\left(2\pi\frac{\ell\left(\sin\left(\theta_j\right)-\sin\left(\theta_{j'}\right)\right)}{\lambda}\right)\phi_{s_{j'}}(n,f) \tag{12.74a}$$

$$\Delta\mathrm{SIR}^{0,\mathrm{MF}}_{jj'}(f) = \left(\mathrm{diric}^2\left(2\pi\frac{\ell\left(\sin\left(\theta_j\right)-\sin\left(\theta_{j'}\right)\right)}{\lambda}\right)\right)^{-1}. \tag{12.74b}$$

Considering the spatially non-correlated noise properties, see (12.67), and substituting it in the power of the noise at the output of the j-th source MF, see (12.26a), and the corresponding SNR improvement, see (12.26b), the latter can be expressed in this special case as:

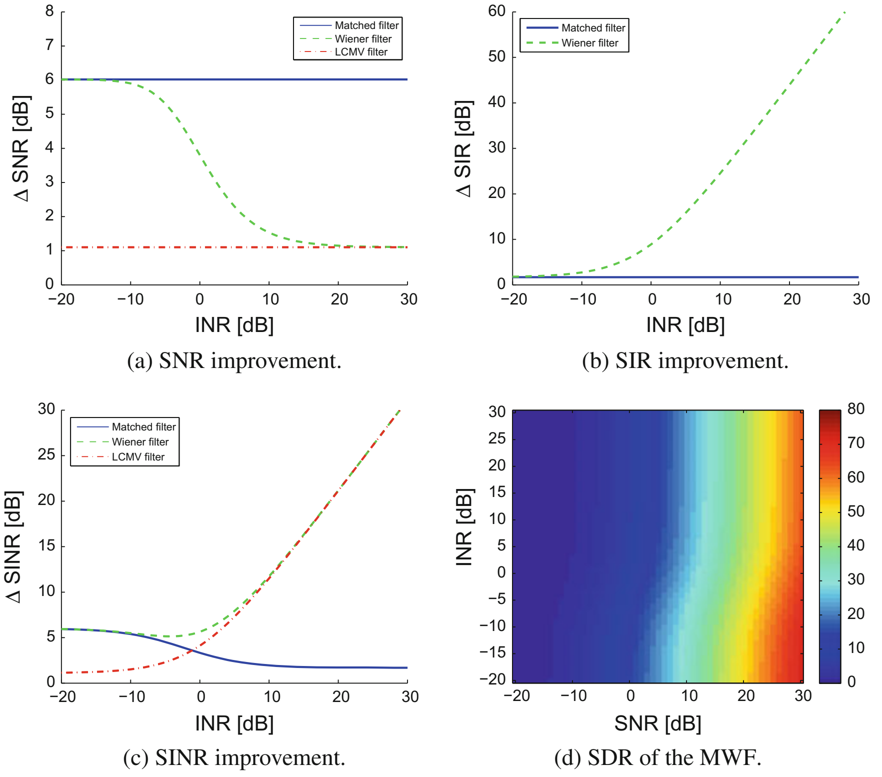
$$\psi^{0,\mathrm{MF}}_{v_j}(f) = \frac{\phi_u(f)}{M} \tag{12.75a}$$

$$\Delta\mathrm{SNR}^{0,\mathrm{MF}}_{j}(f) = M. \tag{12.75b}$$

The corresponding criteria for the MWF and multichannel LCMV are more complicated, and their derivation is omitted.

We compare the spatial filters in a specific scenario of: (a) the microphone array comprises of $M = 4$ microphones with a microphone spacing of $\ell = 10\,\mathrm{cm}$; (b) the desired source is the first source which arrives from $\theta_1 = 0°$. In the following, we investigate the performance dependency on the parameters: (a) SNR and interference-to-noise ratio (INR); (b) interference DOA, θ_2; (c) frequency. For simplicity, we consider two subsets of the above mentioned parameters.

In the first parameters subset, the interference DOA and frequency are set to $\theta_2 = 10^o$ and $\frac{f \cdot f_s}{F} = 1715\,\mathrm{Hz}$, corresponding to $\frac{\ell}{\lambda} = \frac{1}{2}$ (in other figures we explore the performance depending on the frequency or wavelength). For this parameters selection the performance measures of the spatial filters are compared as a function of SNR and INR values that are selected within the range of $[-20\,\mathrm{dB}, 30\,\mathrm{dB}]$. The improvements in SNR, SIR and SINR and the SDR are depicted in Fig. 12.2a–d. We can observe in these figures the consequences of the different design criteria: (a) the MF is designed to maximized the SNR improvement with a spatially white noise (as in this example) and obtains the highest SNR improvement in Fig. 12.2a; (b) the LCMV is designed to null out the interfering sources and therefore obtains an infinite



(a) SNR improvement.

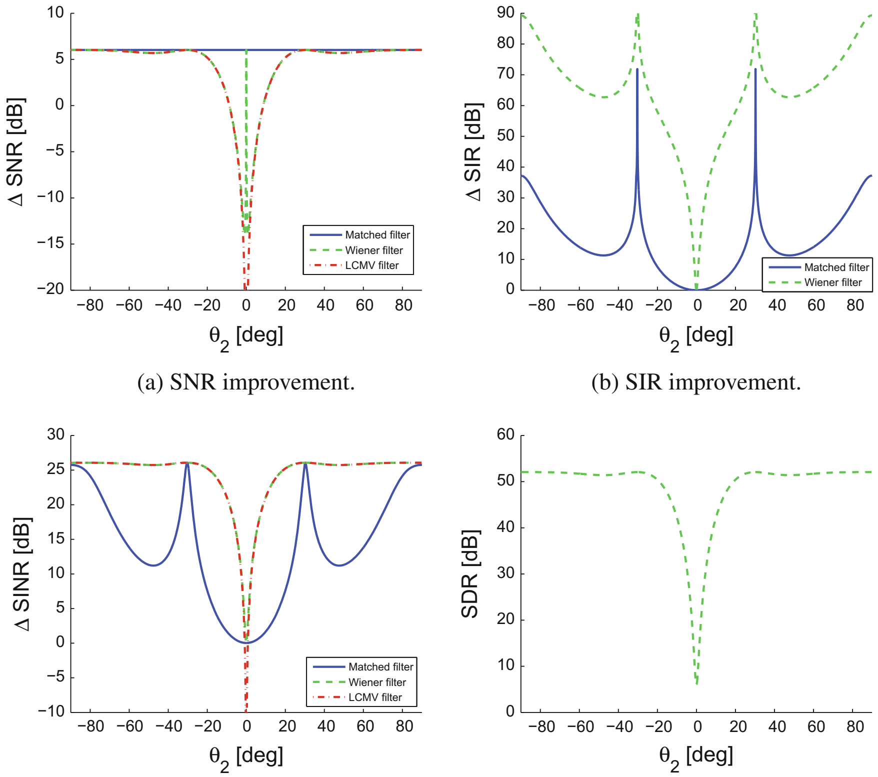
(b) SIR improvement.

(c) SINR improvement.

(d) SDR of the MWF.

Fig. 12.2 Performance comparison depending on input SNR and INR of various spatial filters in the narrowband case with 2 speech sources propagating in freespace and spatially white noise received by a ULA comprising 4 microphones

SIR improvement, which is of course higher than the finite SIR improvement of the other methods that are depicted in Fig. 12.2b; (c) the MWF is designed to maximize the SINR improvement and this is evidently seen in Fig. 12.2c. The MWF aims to maximize the SINR improvement and thus minimize the sum of interference and noise powers at its output. In the limit cases of INR [dB] $\rightarrow -\infty$ where the interference power is negligible and INR [dB] $\rightarrow \infty$ where the noise power is negligible, the MWF coincides with the MF and the LCMV, respectively. This can be clearly seen in Fig. 12.2a–c, where the performance of the MWF converges to that of the MF and LCMV for INR [dB] $\rightarrow -\infty$ and INR [dB] $\rightarrow \infty$, respectively. The MF and LCMV spatial filters are distortionless by design at any input SNR and INR levels, and therefore we do not depict their SDR. The SDR at the output of the MWF as a function of the input SNR and INR is depicted in Fig. 12.2d. The higher is the input SNR the higher is the relative weight of the distortion component compared to the interference and noise components in the MSE in (12.27), which is the MWF design criterion and correspondingly the higher is the SDR of the MWF.



(a) SNR improvement.

(b) SIR improvement.

(c) SINR improvement.

(d) SDR of the MWF.

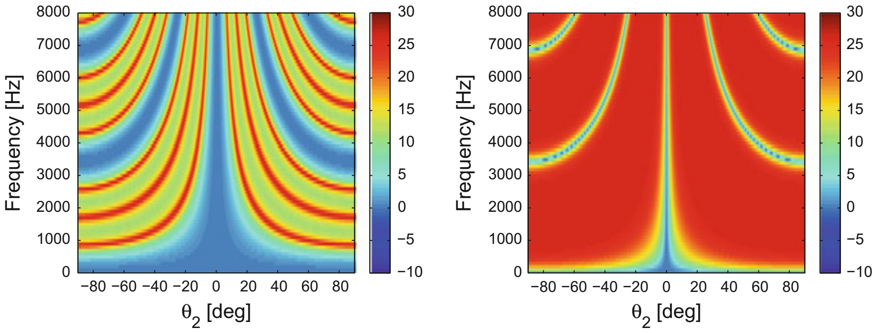
Fig. 12.3 Performance comparison of various spatial filters applied in the narrowband case (at frequency 1715 Hz), where 2 speech sources propagating in freespace and a spatially white additive noise are received by a ULA comprising 4 microphones. The desired source arrives from $\theta_1 = 0°$ and the interfering source arrives from $\theta_2 \int [-90°, 90°]$

In the second parameters subset, the input SNR and INR are both set to 20 dB and the interference DOA and frequency are selected within the range of $[-90°, 90°]$ and [0 Hz, 8000 Hz], respectively. The SNR, SIR and SINR improvement as well as SDR for frequency 1715 Hz (for which $\frac{\ell}{\lambda} = 0.5$) depending the interference source DOA are depicted in Fig. 12.3 for the various spatial filters. As in the previous example and regardless of the DOA of the interference: (a) the MF is optimal in the sense of SNR improvement for a spatially white noise, see Fig. 12.3a; (b) the LCMV is optimal in the sense of SIR improvement, see Fig. 12.3b, as it completely nulls out the interference and obtains infinite SIR improvement, whereas for MF and MWF there is some residual interference at the output for almost all interference DOAs; (c) the MWF is optimal in the sense of SINR improvement, see Fig. 12.3c, although the SINR improvement of the LCMV is very similar for most interference DOAs.

The main difference between the LCMV and MWF can be observed when the interference DOA, θ_2, is close to that of the desired source, $\theta_1 = 0°$. The LCMV, which is designed to null the interference, "struggles" to satisfy its constraints as the interference and desired source DOAs become closer. As a result, the SNR improvement (which is a secondary objective for the LCMV) and correspondingly the SINR improvement are degraded (see Fig. 12.3a, c), and might even become negative (i.e. the spatial noise power at the output might become higher than the noise power at the input and in extreme cases might even become higher than the noise an interference power at the input). Furthermore, the LCMV is not defined for the singular case of $\theta_2 = \theta_1$. In this specific case, the MWF is not able to improve the SIR, however, it is able to improve the SNR. However, note in Fig. 12.3 that as the interference and desired source DOAs become close the SDR degrades. This is because the MWF converges in this case to the MF scaled by a single channel Wiener filter, which introduces more distortion as interference and noise power increases.

Another interesting observation in the SIR improvement (see Fig. 12.3b) is that for some DOAs ($\theta_2 \approx \pm 30°$) the SIR improvement of the MWF and MF also converge to infinity (as the optimal LCMV). The reason for that is that for these DOAs the interfering and desired ATF vectors are orthogonal (i.e. $\rho_{12}^{0}(f) = 0$, see (12.68)) and the corresponding SIR improvement is also infinite.

The SINR improvement of the MF and MWF depending on interference DOA and frequency are depicted in Figs. 12.4a, b. Clearly the SINR improvement of the MWF outperforms that of the MF. For brevity we omit the SINR improvement of the LCMV as it is similar to the improvement of the MWF almost always, except for when the interference DOA approaches the desired source DOA. The red regions in the SINR improvement of the MF in Fig. 12.4a, similarly to the peaks in the SINR improvement of the MF in Fig. 12.3c, correspond to cases where the desired source and interference ATF vectors are orthogonal. Note that positions of these peaks vary over frequency. The blue regions in the SINR improvement of the MF in Fig. 12.4a



(a) Matched filter.

(b) Wiener filter.

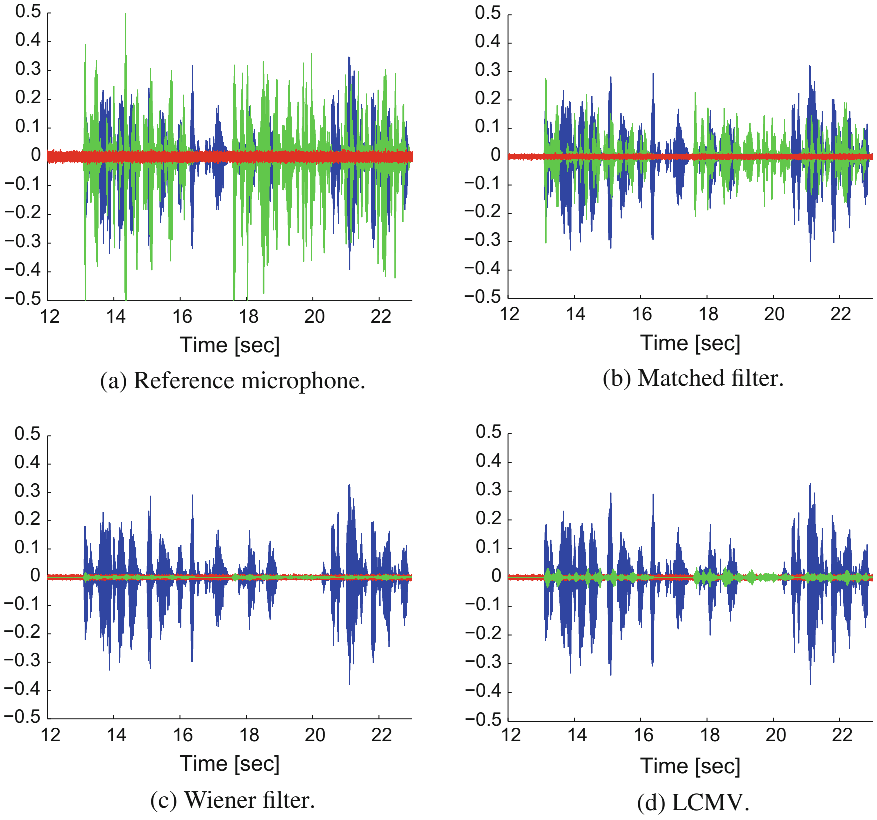
Fig. 12.4 SINR improvement depending on interference DOA and frequency of various spatial filters in the narrowband case with 2 speech sources propagating in freespace and spatially white noise received by a ULA comprising 4 microphones

(except for the one around $\theta_2 \approx \theta_1$) correspond to the spatial aliasing phenomenon. When the interference arrives from the DOA of the desired source or from a DOA which corresponds to a grating lobe of the spatial filter, it cannot be attenuated without degrading the desired source. Hence, the SINR improvement at these DOA is close to zero (blue color). Note that the positions of the grating lobes vary over frequencies. A similar phenomenon can be seen when observing the SINR improvement of the MWF in Fig. 12.4b. For the MWF, however, the areas in which the SINR is close to zero are narrower than in the MF, and the areas of high SINR improvement cover almost the entire interference DOA and frequency ranges.

12.7.2 Speech Signals at a Reverberant Environment

In this section we compare the performance of the various spatial filters in a scenario simulated by convolving recorded speech signals from the WSJCAM0 database [32] with AIRs drawn from a database collected in reverberant enclosures [33]. A ULA comprising $M = 4$ microphones with spacing of $\ell = 8$ cm is picking up signals of $J = 2$ speakers, a female and a male, located at a distance of 1 m from the array at DOAs of $-90°$ and $75°$, respectively, as well as diffuse noise that is generated using a diffuse noise simulator [34]. The SIR is set to 0 dB and the SNR is set to 15 dB.

The signals are transformed to the STFT domain, where MF, MWF and LCMV are designed to enhance the first speaker. Speech-free time-segments and single-talk time-segments of each of the speech sources are used as training segments from which the required parameters for the various spatial filters are estimated: (a) RTFs vector $\tilde{\mathbf{a}}_1(f)$ of the first source for the MF; (b) RTFs vector $\tilde{\mathbf{a}}_1(f)$ and spectrum $\overline{\phi}_{s1}(f)$ of the first source, covariance matrix of the second source $\boldsymbol{\Phi}_{c2}(f)$ and covariance matrix of the noise $\boldsymbol{\Phi}_u(f)$ for the MWF; (c) RTF vectors of both sources $\tilde{\mathbf{a}}_1(f)$,$\tilde{\mathbf{a}}_2(f)$ and covariance matrix of the noise $\boldsymbol{\Phi}_u(f)$ for the LCMV. The output of the spatial filter is transformed back to the time-domain, yielding the enhanced signal. A reference microphone and outputs of the various spatial filters decomposed to their various components (desired speech, interfering speech and noise) are depicted in Fig. 12.5. The corresponding spectrograms of the reference microphone and the outputs of the spatial filters are depicted in Fig. 12.6. The performance measures of each of the spatial filters per frequency-bin in terms of SNR, SIR and SINR improvement as well as SDR are depicted in Fig. 12.7. Considering the SIR and SINR improvements, it is clear from Figs. 12.5 and 12.7b, c, that the MWF is slightly better than the LCMV and that both are significantly better than the MF. While the MWF is expected to obtain the maximal SINR improvement, it is surprising that it outperforms the LCMV in terms of SIR improvement as well. The reason for that lies in the fact that LCMV designates a single constraint for nulling the interfering source, thus assuming a rank-1 model for the interference, while the MWF utilizes the complete covariance matrix of the interfering source, thus allowing to reduce interferences with higher ranks. Although, theoretically, the covariance matrix of coherent point sources is rank-1, in practice, finite window lengths and variations in the AIR (AIR might vary even



(a) Reference microphone.

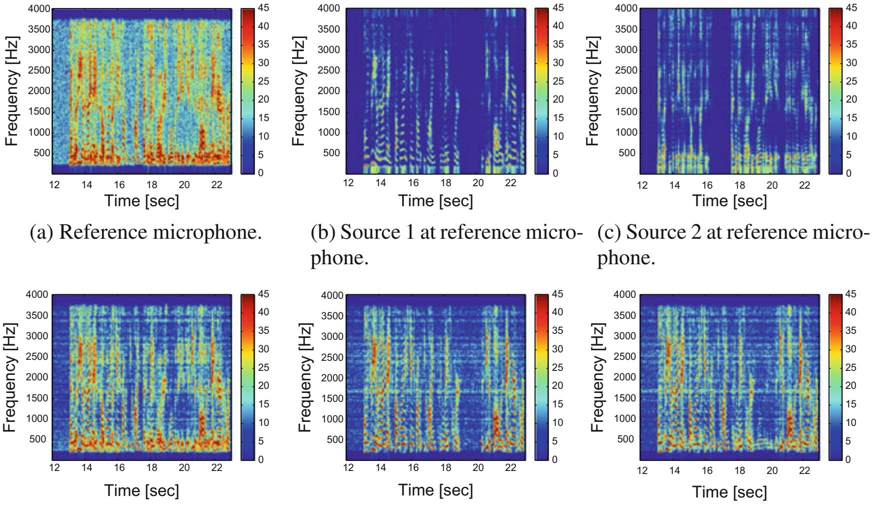
(b) Matched filter.

(c) Wiener filter.

(d) LCMV.

Fig. 12.5 Input and output signals of various spatial filters in a simulated scenario with $J = 2$ speech signals contaminated by diffuse noise and received by a $M = 4$ microphones array in a reverberant environment. The signals are decomposed to their components: (1) desired speaker (blue); (2) interfering speaker (green); and (3) noise (red)

when the source is static due to slight variations in the enclosure) increase the matrix rank. Considering the SNR improvement, note that the MWF and LCMV are better than the MF in frequencies lower than 1000 Hz, and that for higher frequencies the MF is better than the MWF and LCMV. This result is attributed to the diffuse noise properties. In low frequencies, where $\frac{\ell}{\lambda} < \frac{1}{2}$ the diffuse noise has a strong coherent component which the data-dependent filters, MWF and LCMV, reduce efficiently. In higher frequencies the diffuse noise becomes spatially uncorrelated, in which case the MF is optimal and outperforms the MWF and LCMV which utilize their DoF to reduce the interfering speech.



(a) Reference microphone.

(b) Source 1 at reference microphone.

(c) Source 2 at reference microphone.

(d) MF output.

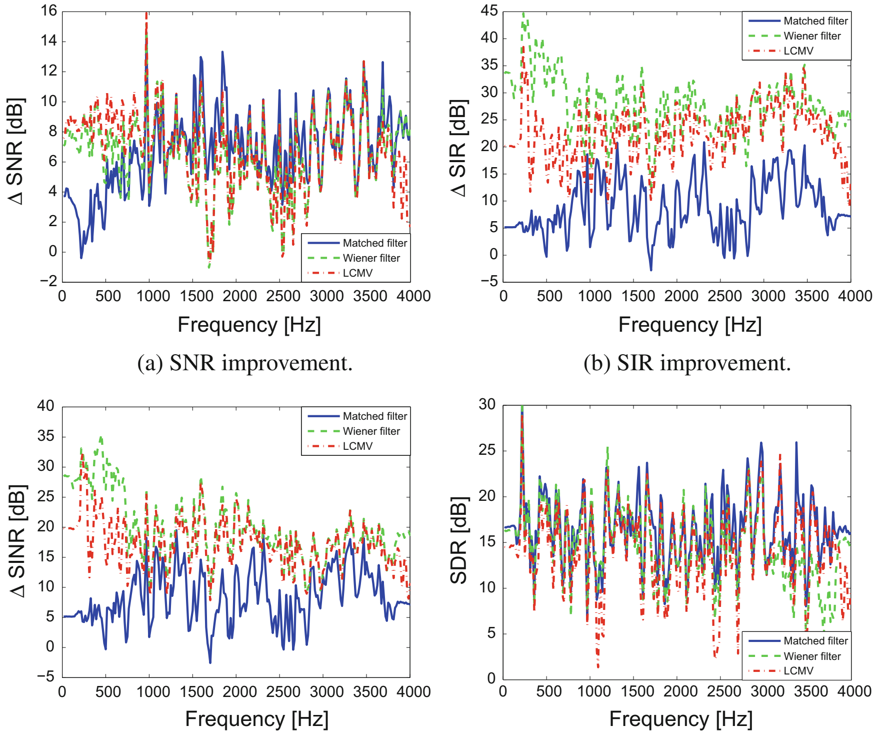
(e) MWF output.

(f) LCMV output.

Fig. 12.6 Input and output spectrograms of various spatial filters aiming to enhance the first source in a simulated scenario with $J = 2$ speech signals contaminated by diffuse noise and received by a $M = 4$ microphones array in a reverberant environment. Input spectrogram of the reference signal and its speech components are respectively depicted in **a**, **b**, **c** and the outputs of the MF, MWF and LCMV spatial filters are respectively depicted in **d**, **e**, **f**

12.8 Summary

MMSE based criteria for designing beamformers, also referred to as spatial-filters, can be used in noise reduction and speech separation tasks. The following methods were presented and analyzed: (1) the MF, which maximizes the SNR at the output without distorting the speech signal, assuming a spatially white noise; (2) the MWF, which minimizes the MSE between the output signal and the desired speech signal, and assigns equal weights to the desired speech distortion, the variance of the interfering speakers at the output, and the noise variance at the output; and (3) the LCMV, which minimizes the noise variance at the output while satisfying a set of constraints designed to maintain the desired speech undistorted and to null out the interfering speakers. Estimation methods for implementing the various beamformers are surveyed. Specifically, methods for estimating the RTFs of speakers and for estimating the spatial covariance matrices of the noise and of the various speaker components were presented. The estimation methods are governed by the multichannel SPP, which was also presented. Some simple examples of applying the various beamformers to simulated narrowband signals in an anechoic environment and to speech signals in a real reverberant environment were presented and discussed.



(a) SNR improvement.

(b) SIR improvement.

(c) SINR improvement.

(d) SDR.

Fig. 12.7 Performance criteria per frequency-bin of various spatial filters in a simulated scenario with $J = 2$ speech signals contaminated by diffuse noise and received by a $M = 4$ microphones array in a reverberant environment

References

1. J. Capon, High-resolution frequency-wavenumber spectrum analysis. Proc. IEEE **57**(8), 1408–1418 (1969)
2. O.L. Frost III, An algorithm for linearly constrained adaptive array processing. Proc. IEEE **60**(8), 926–935 (1972)
3. L.J. Griffiths, C.W. Jim, An alternative approach to linearly constrained adaptive beamforming. IEEE Trans. Antennas Propag. **30**(1), 27–34 (1982)
4. M. Er, A. Cantoni, Derivative constraints for broad-band element space antenna array processors. IEEE Trans. Acoust. Speech Signal Process. **31**(6), 1378–1393 (1983)
5. B.D. Van Veen, K.M. Buckley, Beamforming: a versatile approach to spatial filtering. IEEE Acoust. Speech Signal Process. Mag., 4–24 (1988)
6. B.R. Breed, J. Strauss, A short proof of the equivalence of LCMV and GSC beamforming. IEEE Signal Process. Lett. **9**(6), 168–169 (2002)
7. H.L. Van Trees, *Optimum Array Processing: Estimation Detection, Modulation Theory*, vol. IV (Wiley, New York, 2002)
8. H. Cox, R. Zeskind, M. Owen, Robust adaptive beamforming. IEEE Trans. Acoust. Speech Signal Process. **35**(10), 1365–1376 (1987)

9. S. Affes, Y. Grenier, A signal subspace tracking algorithm for microphone array processing of speech. IEEE Trans. Speech Audio Process. **5**(5), 425–437 (1997)
10. S. Gannot, D. Burshtein, E. Weinstein, Signal enhancement using beamforming and nonstationarity with applications to speech. IEEE Trans. Signal Process. **49**(8), 1614–1626 (2001)
11. S. Markovich, S. Gannot, I. Cohen, Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. IEEE Trans. Audio Speech Lang. Process. **17**(6), 1071–1086 (2009)
12. S. Doclo, A. Spriet, J. Wouters, M. Moonen, Speech distortion weighted multichannel Wiener filtering techniques for noise reduction, in *Speech Enhancement*. Signals and Communication Technology series (Springer, Berlin, 2005), pp. 199–228
13. S. Gannot, E. Vincent, S. Markovich-Golan, A. Ozerov, A consolidated perspective on multimicrophone speech enhancement and source separation. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(4), 692–730 (2017)
14. Y. Avargel, I. Cohen, On multiplicative transfer function approximation in the short-time Fourier transform domain. IEEE Signal Process. Lett. **14**(5), 337–340 (2007)
15. J.L. Flanagan, A.C. Surendran, E.-E. Jan, Spatially selective sound capture for speech and audio processing. Speech Commun. **13**(1–2), 207–222 (1993)
16. S. Markovich-Golan, S. Gannot, I. Cohen, A weighted multichannel Wiener filter for multiple sources scenarios, in *Proceedings of the IEEE Convention of Electrical and Electronics Engineers in Israel (IEEEI)* (Eilat, Israel, 2012)
17. M. Souden, J. Chen, J. Benesty, S. Affes, Gaussian model-based multichannel speech presence probability. IEEE Trans. Audio Speech Lang. Process. **18**(5), 1072–1077 (2010)
18. M. Taseska, E.A. Habets, MMSE-based blind source extraction in diffuse noise fields using a complex coherence-based a priori sap estimator, in *Proceedings of the International Workshop Acoustic Signal Enhancement (IWAENC)* (VDE, 2012), pp. 1–4
19. M. Taseska, E. Habets, Spotforming using distributed microphone arrays, in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, pp. 1–4
20. M. Taseska, E. Habets, Informed spatial filtering for sound extraction using distributed microphone arrays. IEEE/ACM Trans. Audio Speech Lang. Proc. **22**(7), 1195–1207 (2014)
21. T. Gerkmann, C. Breithaupt, R. Martin, Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors. IEEE Trans. Audio Speech Lang. Proc. **16**(5), 910–919 (2008)
22. Y. Ephraim, D. Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. **32**(6), 1109–1121 (1984)
23. I. Cohen, B. Berdugo, Speech enhancement for non-stationary noise environments. Signal Process. **81**(11), 2403–2418 (2001)
24. M. Schroeder, Frequency correlation functions of frequency responses in rooms. J. Acoust. Soc. Am. **34**(12), 1819–1823 (1962)
25. O. Thiergart, G.D. Galdo, E.A. Habets, Signal-to-reverberant ratio estimation based on the complex spatial coherence between omnidirectional microphones, in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 309–312
26. M. Taseska, E.A. Habets, MMSE-based source extraction using position-based posterior probabilities, in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2013), pp. 664–668
27. S. Doclo, M. Moonen, Multimicrophone noise reduction using recursive GSVD-based optimal filtering with ANC postprocessing stage. IEEE Trans. Speech Audio Process. **13**(1), 53–69 (2005)
28. I. Cohen, Relative transfer function identification using speech signals. IEEE Trans. Speech Audio Process. **12**(5), 451–459 (2004)
29. A. Bertrand, M. Moonen, Distributed node-specific LCMV beamforming in wireless sensor networks. IEEE Trans. Signal Process. **60**, 233–246 (2012)

30. S. Markovich-Golan, S. Gannot, Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Brisbane, Australia, 2015)
31. T. Dvorkind, S. Gannot, Time difference of arrival estimation of speech source in a noisy and reverberant environment. Signal Process. **85**(1), 177–204 (2005)
32. T. Robinson, J. Fransen, D. Pye, J. Foote, S. Renals, WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Proceesing (ICASSP)*, vol. 1, 1995, pp. 81–84
33. E. Hadad, F. Heese, P. Vary, S. Gannot, Multichannel audio database in various acoustic environments, in *Proceedings of the International Workshop Acoustic Signal Enhancement (IWAENC)* (IEEE, 2014), pp. 313–317
34. E. Habets, S. Gannot, Generating sensor signals in isotropic noise fields. J. Acoust. Soc. Am. **122**(6), 3464–3470 (2007)