9

Direction-of-Arrival and Time-Difference-of-Arrival Estimation

9.1 Introduction

In the previous chapters we have studied how to use a microphone array to enhance a desired target signal and suppress unwanted noise and interference. Another major functionality of microphone array signal processing is the estimation of the location from which a source signal originates. Depending on the distance between the source and the array relatively to the array size, this estimation problem can be divided into two sub problems, i.e., direction-of-arrival (DOA) estimation and source localization.

The DOA estimation deals with the case where the source is in the array's far-field, as illustrated in Fig. 9.1. In this situation, the source radiates a plane wave having the waveform $s(k)$ that propagates through the non-dispersive medium-air. The normal to the wavefront makes an angle θ with the line joining the sensors in the linear array, and the signal received at each microphone is a time delayed/advanced version of the signal at a reference sensor. To see this, let us choose the first sensor in Fig. 9.1 as the reference point and denote the spacing between the two sensors as d. The signal at the second sensor is delayed by the time required for the plane wave to propagate through $d\cos\theta$. Therefore, the time difference (time delay) between the two sensors is given by $\tau_{12} = d\cos\theta/c$, where c is the sound velocity in air. If the angle ranges between 0° and 180° and if τ_{12} is known then θ is uniquely determined, and *vice versa*. Therefore, estimating the incident angle θ is essentially identical to estimating the time difference τ_{12}. In other words, the DOA estimation problem is the same as the so-called time-difference-of-arrival (TDOA) estimation problem in the far-field case.

Although the incident angle can be estimated with the use of two or more sensors, the range between the sound source and the microphone array is difficult (if not impossible) to determine if the source is in the array's far-field. However, if the source is located in the near-field, as illustrated in Fig. 9.2, it is now possible to estimate not only the angle from which the wave ray reaches each sensor but also the distance between the source and each microphone.

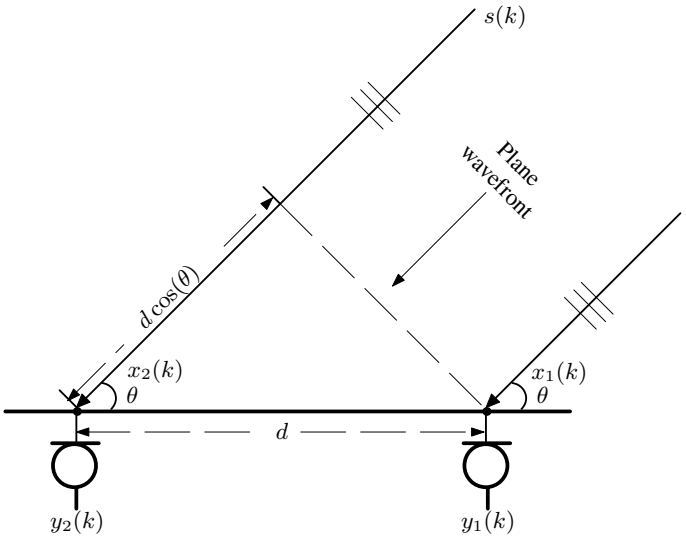

Fig. 9.1. Illustration of the DOA estimation problem in 2-dimensional space with two identical microphones: the source $s(k)$ is located in the far-field, the incident angle is θ, and the spacing between the two sensors is d.

To see this, let us consider the simple example shown in Fig. 9.2. Again, we choose the first microphone as the reference sensor. Let θ_n and r_n denote, respectively, the incident angle and the distance between the sound source and microphone n, $n = 1, 2, 3$. The TDOA between the second and first sensors is given by

$$\tau_{12} = \frac{r_2 - r_1}{c}, \tag{9.1}$$

and the TDOA between the third and first sensors is

$$\tau_{13} = \frac{r_3 - r_1}{c}. \tag{9.2}$$

Applying the cosine rule, we obtain

$$r_2^2 = r_1^2 + d^2 + 2r_1 d\cos(\theta_1) \tag{9.3}$$

and

$$r_3^2 = r_1^2 + 4d^2 + 4r_1 d\cos(\theta_1). \tag{9.4}$$

For a practical array system, the spacing d can always be measured once the array geometry is fixed. If τ_{12} and τ_{13} are available then we can calculate all the unknown parameters θ_1, r_1, r_2, and r_3 by solving the equations from (9.1) to (9.4). Further applying the sine rule, we can obtain an estimate of θ_2

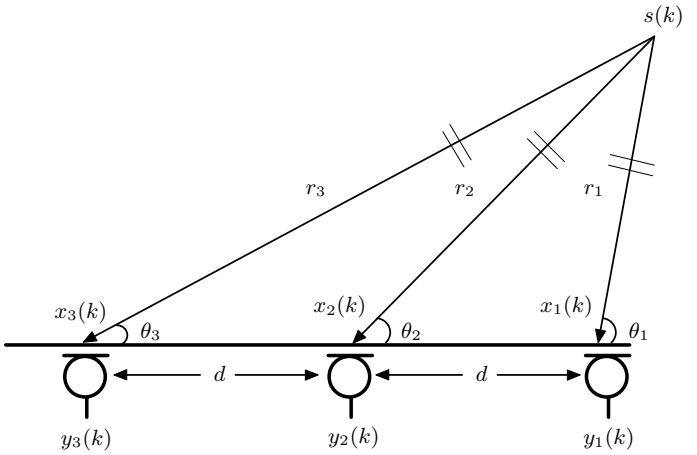

Fig. 9.2. Illustration of the source localization problem with an equispaced linear array: the source $s(k)$ is located in the near-field, and the spacing between any two neighboring sensors is d.

and θ_3. Therefore, all the information regarding the source position relatively to the array can be determined using the triangulation rule once the TDOA information is available. This basic triangulation process forms the foundation for most of the source-localization techniques, even though many algorithms may formulate and solve the problem from a different theoretical perspective [7], [26], [71], [74], [97], [116], [117], [188], [204], [221], [222], [223], [226].

Therefore, regardless if the source is located in the far-field or near-field, the most fundamental step in obtaining the source-origin information is the one of estimating the TDOA between different microphones. This estimation problem would be an easy task if the received signals were merely a delayed and scaled version of each other. In reality, however, the source signal is generally immersed in ambient noise since we are living in a natural environment where the existence of noise is inevitable. Furthermore, each observation signal may contain multiple attenuated and delayed replicas of the source signal due to reflections from boundaries and objects. This multipath propagation effect introduces echoes and spectral distortions into the observation signal, termed as reverberation, which severely deteriorates the source signal. In addition, the source may also move from time to time, resulting in a changing time delay. All these factors make TDOA estimation a complicated and challenging problem.

This chapter discusses the basic ideas underlying TDOA estimation. We will begin our discussion with scenarios where there is only a single source in the sound field. We will then explore what approaches can be used to improve the robustness of TDOA estimation with respect to noise and reverberation. Many fundamental ideas developed for the single-source TDOA estimation

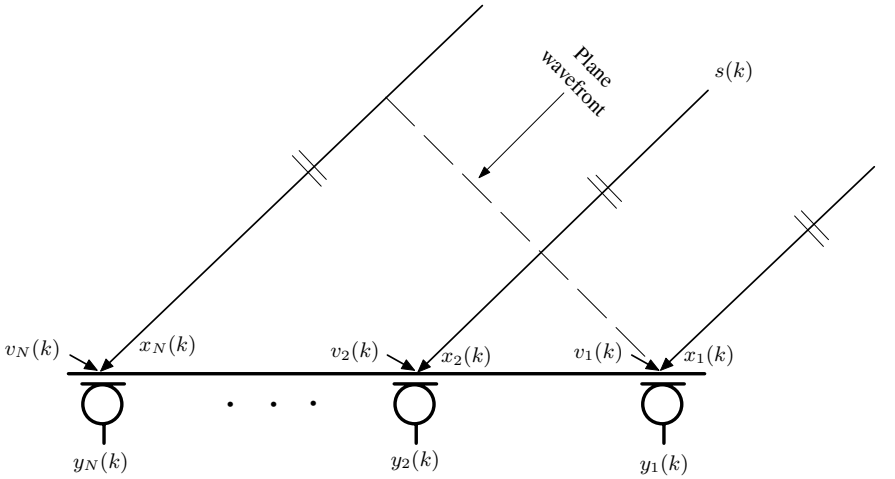

Fig. 9.3. Illustration of the ideal free-field single-source model.

can be extended to the multiple-source situation. To illustrate this, we will discuss the philosophy underlying multiple-source TDOA estimation.

9.2 Problem Formulation and Signal Models

The TDOA estimation problem is concerned with the measurement of time difference between the signals received at different microphones. Depending on the surrounding acoustic environment, we consider two situations: the free-field environment where each sensor receives only the direct-path signal, and reverberant environments where each sensor may receive a large number of reflected signals in addition to the direct path. For each situation, we differentiate the single-source case from the multiple-source scenario since the estimation principles and complexity in these two conditions may not necessarily be the same. So, in total, we consider four signal models: the single-source free-field model, the multiple-source free-field model, the single-source reverberant model, and the multiple-source reverberant model.

9.2.1 Single-Source Free-Field Model

Suppose that there is only one source in the sound field and we use an array of N microphones. In an anechoic open space as shown in Fig. 9.3, the speech source signal $s(k)$ propagates radiatively and the sound level falls off as a function of distance from the source. If we choose the first microphone as the reference point, the signal captured by the nth microphone at time k can be expressed as follows:

$$\begin{aligned} y_n(k) &= \alpha_n s\left(k - t - \tau_{n1}\right) + v_n(k) \\ &= \alpha_n s\left[k - t - \mathcal{F}_n(\tau)\right] + v_n(k) \\ &= x_n(k) + v_n(k),\ n = 1, 2, \ldots, N, \end{aligned} \tag{9.5}$$

where α_n $(n = 1, 2, \ldots, N)$, which range between 0 and 1, are the attenuation factors due to propagation effects, $s(k)$ is the unknown source signal, t is the propagation time from the unknown source to sensor 1, $v_n(k)$ is an additive noise signal at the nth sensor, which is assumed to be uncorrelated with both the source signal and the noise observed at other sensors, τ is the TDOA (also called relative delay) between sensors 1 and 2, and $\tau_{n1} = \mathcal{F}_n(\tau)$ is the TDOA between sensors 1 and n with $\mathcal{F}_1(\tau) = 0$ and $\mathcal{F}_2(\tau) = \tau$. For $n = 3, \ldots, N$, the function $\mathcal{F}_n$ depends not only on τ but also on the microphone array geometry. For example, in the far-field case (plane wave propagation), for a linear and equispaced array, we have

$$\mathcal{F}_n(\tau) = (n-1)\tau, \quad n = 2, \ldots, N, \tag{9.6}$$

and for a linear but non-equispaced array, we have

$$\mathcal{F}_n(\tau) = \frac{\sum_{i=1}^{n-1} d_i}{d_1}\tau, \quad n = 2, \ldots, N, \tag{9.7}$$

where d_i is the distance between microphones i and $i+1$ ($i = 1, \ldots, N-1$). In the near-field case, $\mathcal{F}_n$ depends also on the position of the sound source. Note that $\mathcal{F}_n(\tau)$ can be a *nonlinear* function of τ for a nonlinear array geometry, even in the far-field case (e.g., 3 equilateral sensors). In general τ is not known, but the geometry of the array is known such that the mathematical formulation of $\mathcal{F}_n(\tau)$ is well defined or given. For this model, the TDE (time-delay estimation) problem is formulated as one of determining an estimate $\hat{\tau}$ of the true time delay τ using a set of finite observation samples.

9.2.2 Multiple-Source Free-Field Model

Still in the anechoic environments, if there are multiple sources in the sound field, the signal received at the nth sensor becomes

$$\begin{aligned} y_n(k) &= \sum_{m=1}^{M} \alpha_{nm} s_m\left[k - t_m - \mathcal{F}_n(\tau_m)\right] + v_n(k) \\ &= x_n(k) + v_n(k),\ n = 1, 2, \ldots, N, \end{aligned} \tag{9.8}$$

where M is the total number of sound sources, α_{nm} $(n = 1, 2, \ldots, N,\ m = 1, 2, \ldots, M)$, are the attenuation factors due to propagation effects, $s_m(k)$ $(m = 1, 2, \ldots, M)$ are the unknown source signals, which are assumed to be mutually independent with each other, t_m is the propagation time from the unknown source m to sensor 1 (reference sensor), $v_n(k)$ is an additive noise

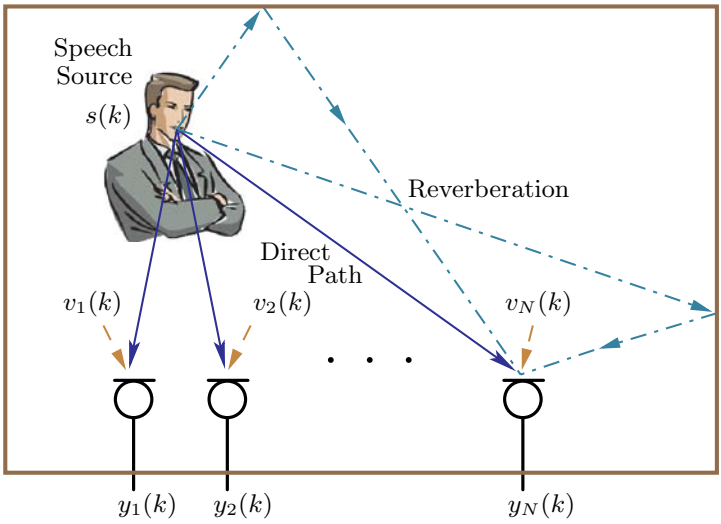

Fig. 9.4. Illustration of the single-source reverberant model.

signal at the nth sensor, which is assumed to be uncorrelated with not only all the source signals but also with the noise observed at other sensors, τ_m is the TDOA between sensors 2 and 1 due to the mth source, and $\mathcal{F}_n(\tau_m)$ is the TDOA between sensors n and 1 for the source m. For this model, the objective of TDOA estimation is to determine all the parameters τ_m, $m = 1, 2, \ldots, M$ using microphone observations.

9.2.3 Single-Source Reverberant Model

While the ideal free-field models have the merit of being simple, they do not take into account the multipath effect. Therefore, such models are inadequate to describe a real reverberant environment and we need a more comprehensive and more informative alternative to model the effect of multipath propagation, leading to the so-called reverberant models, which treat the acoustic impulse response with an FIR filter. If there is only one source in the sound filed as illustrated in Fig. 9.4, the problem can be modeled as a SIMO (single-input multiple-output) system and the nth microphone signal is given by

$$\begin{aligned} y_n(k) &= g_n * s(k) + v_n(k), \\ &= x_n(k) + v_n(k), \quad n = 1, 2, \ldots, N, \end{aligned} \tag{9.9}$$

where g_n is the channel impulse response from the source to microphone n. In vector/matrix form, (9.9) is re-written as

$$\mathbf{y}_n(k) = \mathbf{G}_n \mathbf{s}(k) + \mathbf{v}_n(k), \quad n = 1, 2, \ldots, N, \tag{9.10}$$

where

$$
\begin{aligned}
\mathbf{y}_n(k) &= \left[\, y_n(k) \cdots y_n(k-L+1) \,\right]^T, \\
\mathbf{G}_n &= \begin{bmatrix} g_{n,0} & \cdots & g_{n,L-1} & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & g_{n,0} & \cdots & g_{n,L-1} \end{bmatrix}, \\
\mathbf{s}(k) &= \left[\, s(k)\ s(k-1) \cdots s(k-L+1) \cdots s(k-2L+2) \,\right]^T, \\
\mathbf{v}_n(k) &= \left[\, v_n(k) \cdots v_n(k-L+1) \,\right]^T,
\end{aligned}
$$

and L is the length of the longest channel impulse response of the SIMO system. Again, it is assumed that $v_n(k)$ is uncorrelated with both the source signal and the noise observed at other sensors.

In comparison with the free-field model, the TDOA τ in this reverberant model is an implicit or hidden parameter. With such a model, the TDOA can only be obtained after the SIMO system is "blindly" identified (since the source signal is unknown), which looks like a more difficult problem but is fortunately not insurmountable.

9.2.4 Multiple-Source Reverberant Model

If there are multiple sources in the sound filed, the array can be modeled as a MIMO (multiple-input multiple-output) system with M inputs and N outputs. At time k, we have

$$
\mathbf{y}(k) = \mathbf{G}\mathbf{s}_{ML}(k) + \mathbf{v}(k), \tag{9.11}
$$

where

$$
\begin{aligned}
\mathbf{y}(k) &= \left[\, y_1(k)\ y_2(k) \cdots y_N(k) \,\right]^T, \\
\mathbf{G} &= \left[\, \mathbf{G}_1\ \mathbf{G}_2 \cdots \mathbf{G}_M \,\right], \\
\mathbf{G}_m &= \begin{bmatrix} g_{1m,0} & g_{1m,1} & \cdots & g_{1m,L-1} \\ g_{2m,0} & g_{2m,1} & \cdots & g_{2m,L-1} \\ \vdots & \vdots & \ddots & \vdots \\ g_{Nm,0} & g_{Nm,1} & \cdots & g_{Nm,L-1} \end{bmatrix}_{N\times L}, \\
& \quad m = 1, 2, \ldots, M, \\
\mathbf{v}(k) &= \left[\, v_1(k)\ v_2(k) \cdots v_N(k) \,\right]^T, \\
\mathbf{s}_{ML}(k) &= \left[\, \mathbf{s}_1^T(k)\ \mathbf{s}_2^T(k) \cdots \mathbf{s}_M^T(k) \,\right]^T, \\
\mathbf{s}_m(k) &= \left[\, s_m(k)\ s_m(k-1) \cdots s_m(k-L+1) \,\right]^T.
\end{aligned}
$$

and g_{nm} $(n = 1, 2, \ldots, N,\ m = 1, 2, \ldots, M)$ is the impulse response of the channel from source m to microphone n. Similar to the multiple-source free-field model, we assume that all the source signals are mutually independent,

and $v_n(k)$ is uncorrelated with not only all the source signals but also with the noise observed at other sensors.

For this model, in order to estimate the TDOA, we have to "blindly" identify the MIMO system, which can be an extremely difficult problem.

9.3 Cross-Correlation Method

We are now ready to investigate the algorithms for TDOA estimation. Let us start with the most simple and straightforward method: cross-correlation (CC). Consider the single-source free-field model with only two sensors, i.e., $N = 2$. The cross-correlation function (CCF) between the two observation signals $y_1(k)$ and $y_2(k)$ is defined as

$$r_{y_1y_2}^{\mathrm{CC}}(p) = E\left[y_1(k)y_2(k+p)\right]. \tag{9.12}$$

Substituting (9.5) into (9.12), we can readily deduce that

$$\begin{aligned} r_{y_1y_2}^{\mathrm{CC}}(p) = \alpha_1\alpha_2 r_{ss}^{\mathrm{CC}}(p-\tau) + \alpha_1 r_{sv_2}^{\mathrm{CC}}(p+t) + \\ \alpha_2 r_{sv_1}(p-t-\tau) + r_{v_1v_2}(p). \end{aligned} \tag{9.13}$$

If we assume that $v_n(k)$ is uncorrelated with both the signal and the noise observed at the other sensor, it can be easily checked that $r_{y_1y_2}^{\mathrm{CC}}(p)$ reaches its maximum at $p = \tau$. Therefore, given the CCF, we can obtain an estimate of the TDOA between $y_1(k)$ and $y_2(k)$ as

$$\hat{\tau}^{\mathrm{CC}} = \arg\max_p r_{y_1y_2}^{\mathrm{CC}}(p), \tag{9.14}$$

where $p \in [-\tau_{\max}, \tau_{\max}]$, and $\tau_{\max}$ is the maximum possible delay.

In digital implementation of (9.14), some approximations are required because the CCF is not known and must be estimated. A normal practice is to replace the CCF defined in (9.12) by its time-averaged estimate. Suppose that at time instant k we have a set of observation samples of x_n, $\{x_n(k), x_n(k+1), \cdots, x_n(k+K-1)\}$, $n = 1, 2$, the corresponding CCF can be estimated as either

$$\hat{r}_{y_1y_2}^{\mathrm{CC}}(p) = \begin{cases} \dfrac{1}{K}\displaystyle\sum_{i=0}^{K-p-1} y_1(k+i)y_2(k+i+p), & p \geq 0 \\ \hat{r}_{y_2y_1}^{\mathrm{CC}}(-p), & p < 0 \end{cases}, \tag{9.15}$$

or

$$\hat{r}_{y_1y_2}^{\mathrm{CC}}(p) = \begin{cases} \dfrac{1}{K-p}\displaystyle\sum_{i=0}^{K-p-1} y_1(k+i)y_2(k+i+p), & p \geq 0 \\ \hat{r}_{y_2y_1}^{\mathrm{CC}}(-p), & p < 0 \end{cases}, \tag{9.16}$$

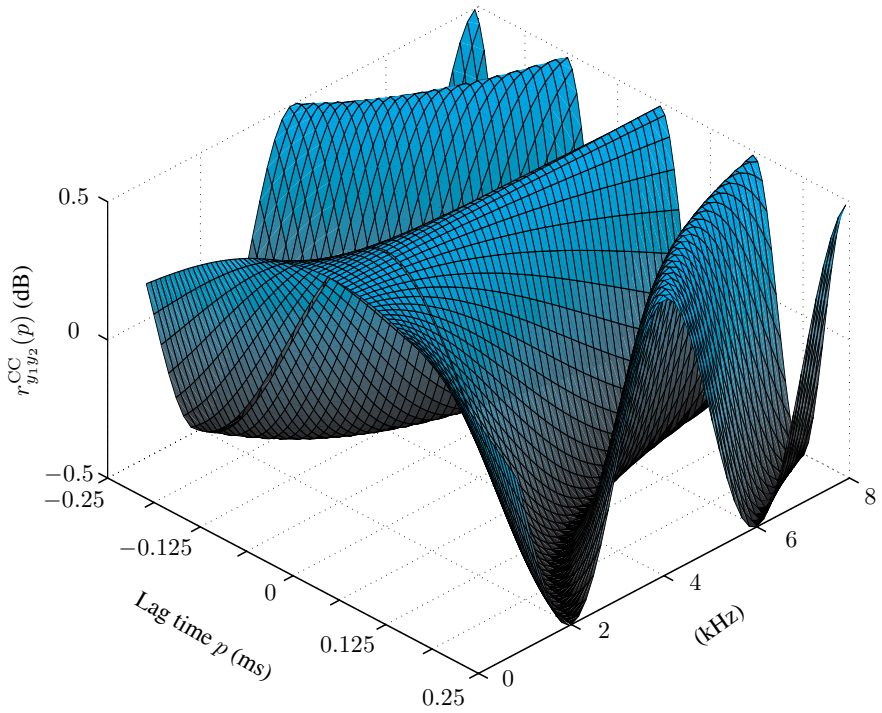

Fig. 9.5. CCF between $y_1(k)$ and $y_2(k)$: the source is a narrowband signal, the incident angle is $\theta = 0^\circ$, there is no noise, i.e., $v_n(k) = 0$, the spacing between the two sensors is $d = 8$ cm, and the sampling frequency is 16 kHz.

where K is the block size. The difference between (9.15) and (9.16) is that the former leads to a biased estimator, while the latter is an unbiased one. However, since it has a lower estimation variance and is asymptotically unbiased, the former has been widely adopted in many applications.

The CC method is simple to implement. However, its performance is often affected by many factors such as signal self correlation, reverberation, etc. Many efforts have been devoted to improving this method, which will be discussed in the next section. But before we finish this section, we would like to point out one potential problem that is often neglected in TDOA estimation: spatial aliasing. In Chapter 3, we have shown that spatial aliasing may cause ambiguity for the array to distinguish signals propagating from different directions. Similarly, this problem will also affect the TDOA estimation. To see this, let us consider a simple example where the source is a narrowband signal in the form of

$$s(k) = \cos(2\pi f k). \tag{9.17}$$

If we neglect both the propagation attenuation and noise effects in (9.5), we get

$$y_1(k) = \cos[2\pi f(k-t)], \tag{9.18}$$

$$y_2(k) = \cos[2\pi f(k-t-\tau)]. \tag{9.19}$$

Substituting (9.18) and (9.19) into (9.12), we easily obtain

$$r_{y_1y_2}^{\mathrm{CC}}(p) = E\left[y_1(k)y_2(k+p)\right] = \frac{1}{2}\cos[2\pi f(p-\tau)]. \tag{9.20}$$

Figure 9.5 plots the CCF for different frequencies. The spacing between the two microphones is 8 cm. Assuming that the sound velocity is 320 m/s, one can easily check, based on the analysis given in Section 3.3, that when $f > 2$ kHz, there will be spatial aliasing for beamforming. From Fig. 9.5, we see that when $f > 2$ kHz, the CCF experiences multiple peaks in the range of $[-\tau_{\max}, \tau_{\max}]$ ($\tau_{\max}$ is the maximum possible TDOA and can be determined from the spacing between the two microphones), which makes it difficult to search for the correct TDOA. In microphone-array applications, the source is usually speech, which consists of rich frequency components. In order to avoid the spatial aliasing problem and improve TDOA estimation, one should low-pass filter the microphone signal before feeding it to the estimation algorithms. The cutoff frequency can be calculated using the sensor spacing, i.e., $f_{\mathrm{c}} = c/2d$.

9.4 The Family of the Generalized Cross-Correlation Methods

The generalized cross-correlation (GCC) algorithm proposed by Knapp and Carter [145] is the most widely used approach to TDOA estimation. Same as the CC method, GCC employs the free-field model (9.5) and considers only two microphones, i.e., $N = 2$. Then the TDOA estimate between the two microphones is obtained as the lag time that maximizes the CCF between the filtered signals of the microphone outputs [which is often called the generalized CCF (GCCF)]:

$$\hat{\tau}^{\mathrm{GCC}} = \arg\max_{\tau} r_{y_1y_2}^{\mathrm{GCC}}(p), \tag{9.21}$$

where

$$\begin{aligned} r_{y_1y_2}^{\mathrm{GCC}}(p) &= F^{-1}\left[\Psi_{y_1y_2}(f)\right] \\ &= \int_{-\infty}^{\infty} \Psi_{y_1y_2}(f) e^{j2\pi fp} df \\ &= \int_{-\infty}^{\infty} \vartheta(f)\phi_{y_1y_2}(f) e^{j2\pi fp} df \end{aligned} \tag{9.22}$$

is the GCC function, $F^{-1}[\cdot]$ stands for the inverse discrete-time Fourier transform (IDTFT),

$$\phi_{y_1y_2}(f) = E\left[Y_1(f)Y_2^*(f)\right] \tag{9.23}$$

is the cross-spectrum with

$$Y_n(f) = \sum_k y_n(k)e^{-j2\pi fk}, \quad n = 1, 2,$$

$\vartheta(f)$ is a frequency-domain weighting function, and

$$\Psi_{y_1y_2}(f) = \vartheta(f)\phi_{y_1y_2}(f) \tag{9.24}$$

is the generalized cross-spectrum.

There are many different choices of the frequency-domain weighting function $\vartheta(f)$, leading to a variety of different GCC methods.

9.4.1 Classical Cross-Correlation

If we set $\vartheta(f) = 1$, it can be checked that the GCC degenerates to the cross-correlation method discussed in the previous section. The only difference is that now the CCF is estimated using the discrete Fourier transform (DFT) and the inverse DFT (IDFT), which can be implemented efficiently thanks to the fast Fourier transform (FFT).

We know from the free-field model (9.5) that

$$Y_n(f) = \alpha_n S(f)e^{-j2\pi f[t-\mathcal{F}_n(\tau)]} + V_n(f), \quad n = 1, 2. \tag{9.25}$$

Substituting (9.25) into (9.24) and noting that the noise signal at one microphone is uncorrelated with the source signal and the noise signal at the other microphone by assumption, we have

$$\Psi_{y_1y_2}^{\mathrm{CC}}(f) = \alpha_1\alpha_2 e^{-j2\pi f\tau} E\left[|S(f)|^2\right]. \tag{9.26}$$

The fact that $\Psi_{y_1y_2}^{\mathrm{CC}}(f)$ depends on the source signal can be detrimental for TDOA estimation since speech is inherently non-stationary.

9.4.2 Smoothed Coherence Transform

In order to overcome the impact of fluctuating levels of the speech source signal on TDOA estimation, an effective way is to pre-whiten the microphone outputs before their cross-spectrum is computed. This is equivalent to choosing

$$\vartheta(f) = \frac{1}{\sqrt{E\left[|Y_1(f)|^2\right]E\left[|Y_2(f)|^2\right]}}, \tag{9.27}$$

which leads to the so-called smoothed coherence transform (SCOT) method [36]. Substituting (9.25) and (9.27) into (9.24) produces the SCOT cross-spectrum:

$$\begin{aligned}\Psi_{y_1y_2}^{\mathrm{SCOT}}(f) &= \frac{\alpha_1\alpha_2 e^{-j2\pi f\tau} E\left[|S(f)|^2\right]}{\sqrt{E\left[|Y_1(f)|^2\right] E\left[|Y_2(f)|^2\right]}} \\ &= \frac{\alpha_1\alpha_2 e^{-j2\pi f\tau} E\left[|S(f)|^2\right]}{\sqrt{\alpha_1^2 E\left[|S(f)|^2\right] + \sigma_{v_1}^2(f)} \cdot \sqrt{\alpha_2^2 E\left[|S(f)|^2\right] + \sigma_{v_2}^2(f)}} \\ &= \frac{e^{-j2\pi f\tau}}{\sqrt{1+\frac{1}{\mathrm{SNR}_1(f)}} \cdot \sqrt{1+\frac{1}{\mathrm{SNR}_2(f)}}}, \end{aligned} \tag{9.28}$$

where

$$\begin{aligned}\sigma_{v_n}^2(f) &= E\left[|V_n(f)|^2\right], \\ \mathrm{SNR}_n(f) &= \frac{\alpha_n^2 E\left[|S(f)|^2\right]}{E\left[|V_n(f)|^2\right]}, \quad n = 1, 2.\end{aligned}$$

If the SNRs are the same at the two microphones, then we get

$$\Psi_{x_1x_2}^{\mathrm{SCOT}}(f) = \left[\frac{\mathrm{SNR}(f)}{1+\mathrm{SNR}(f)}\right] \cdot e^{-j2\pi f\tau}. \tag{9.29}$$

Therefore, the performance of the SCOT algorithm for TDOA estimation would vary with the SNR. But when the SNR is large enough,

$$\Psi_{x_1x_2}^{\mathrm{SCOT}}(f) \approx e^{-j2\pi f\tau}, \tag{9.30}$$

which implies that the estimation performance is independent of the power of the source signal. So, the SCOT method is theoretically superior to the CC method. But this superiority only holds when the noise level is low.

9.4.3 Phase Transform

It becomes clear by examining (9.22) that the TDOA information is conveyed in the phase rather than the amplitude of the cross-spectrum. Therefore, we can simply discard the amplitude and only keep the phase. By setting

$$\vartheta(f) = \frac{1}{|\phi_{y_1y_2}(f)|}, \tag{9.31}$$

we get the phase transform (PHAT) method [145]. In this case, the generalized cross-spectrum is given by

$$\Psi_{y_1y_2}^{\mathrm{PHAT}}(f) = e^{-j2\pi f\tau}, \tag{9.32}$$

which depends only on the TDOA τ. Substituting (9.32) into (9.22), we obtain an ideal GCC function:

$$r_{y_1y_2}^{\mathrm{PHAT}}(p) = \int_{-\infty}^{\infty} e^{j2\pi f(p-\tau)} df = \begin{cases} \infty, & p = \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{9.33}$$

As a result, the PHAT method performs in general better than the CC and SCOT methods for TDOA estimation with respect to a speech sound source.

The GCC methods are computationally efficient. They induce very short decision delays and hence have a good tracking capability: an estimate is produced almost instantaneously. The GCC methods have been well studied and are found to perform fairly well in moderately noisy and non-reverberant environments [37], [128]. In order to improve their robustness to additive noise, many amendments have been proposed [25], [174], [175], [222]. However, these methods still tend to break down when room reverberation is high. This is insightfully explained by the fact that the GCC methods model the surrounding acoustic environment as an ideal free field and thus have a fundamental weakness in their ability to cope with room reverberation.

9.5 Spatial Linear Prediction Method

In this section, we explore the possibility of using multiple microphones (more than 2) to improve the TDOA estimation in adverse acoustic environments. The fundamental underlying idea is to take advantage of the redundant information provided by multiple sensors. To illustrate the redundancy, let us consider a three-microphone system. In such a system, there are three TDOAs, namely τ_{12}, τ_{13}, and τ_{23}. Apparently, these three TDOAs are not independent but are related as follows: $\tau_{13} = \tau_{12} + \tau_{23}$. Such a relationship was used in [144] and a Kalman filtering based two-stage TDE algorithm was proposed. Recently, with a similar line of thoughts, several fusion algorithms have been developed [55], [93], [172]. In what follows, we present a TDOA estimation algorithm using spatial linear prediction [14], [39], which takes advantage of the TDOA redundancy among multiple microphones in a more intuitive way.

Consider the free-field model in (9.5) with a linear array of N ($N \geq 2$) microphones. If the source is in the far-field and we neglect the noise terms, it can be easily checked that

$$y_n[k + \mathcal{F}_n(\tau)] = \alpha_n s(k - t), \quad \forall n = 1, 2, \ldots, N. \tag{9.34}$$

Therefore, $y_1(k)$ is aligned with $y_n[k + \mathcal{F}_n(\tau)]$. From this relationship, we can defined the forward spatial prediction error signal

$$e_1(k, p) = y_1(k) - \mathbf{y}_{\mathrm{a},2:N}^T(k, p)\mathbf{a}_{2:N}(p), \tag{9.35}$$

where p, again, is a dummy variable for the hypothesized TDOA τ,

$$\mathbf{y}_{\mathrm{a},2:N}(k, p) = \left[y_2[k + \mathcal{F}_2(p)] \cdots y_N[k + \mathcal{F}_N(p)] \right]^T, \tag{9.36}$$

is the aligned (subscript a) signal vector, and

$$\mathbf{a}_{2:N}(p) = \left[a_2(p)\ a_3(p) \cdots a_N(p) \right]^T$$

contains the forward spatial linear prediction coefficients. Minimizing the mean-square value of the prediction error signal

$$J_1(p) = E\left[e_1^2(k,p)\right] \tag{9.37}$$

leads to the linear system

$$\mathbf{R}_{\mathrm{a},2:N}(p)\mathbf{a}_{2:N}(p) = \mathbf{r}_{\mathrm{a},2:N}(p), \tag{9.38}$$

where

$$\begin{aligned}\mathbf{R}_{\mathrm{a},2:N}(p) &= E\left[\mathbf{y}_{\mathrm{a},2:N}(k,p)\mathbf{y}_{\mathrm{a},2:N}^T(k,p)\right] \\ &= \begin{bmatrix} \sigma_{y_2}^2 & r_{\mathrm{a},y_2y_3}(p) & \cdots & r_{\mathrm{a},y_2y_N}(p) \\ r_{\mathrm{a},y_3y_2}(p) & \sigma_{y_3}^2 & \cdots & r_{\mathrm{a},y_3y_N}(p) \\ \vdots & \vdots & \ddots & \vdots \\ r_{\mathrm{a},y_Ny_2}(p) & r_{\mathrm{a},y_Ny_3}(p) & \cdots & \sigma_{y_N}^2 \end{bmatrix}\end{aligned} \tag{9.39}$$

is the spatial correlation matrix of the aligned signals with

$$\begin{aligned}\sigma_{y_n}^2 &= E\left[y_n^2(k)\right], \quad n = 1,2,\ldots,N, \\ r_{\mathrm{a},y_iy_j}(p) &= E\left\{y_i[k+\mathcal{F}_i(p)]y_j[k+\mathcal{F}_j(p)]\right\}, \quad i,j = 1,2,\ldots,N,\end{aligned}$$

and

$$\mathbf{r}_{\mathrm{a},2:N}(p) = \left[\, r_{\mathrm{a},y_1y_2}(p)\ r_{\mathrm{a},y_1y_3}(p)\cdots r_{\mathrm{a},y_1y_N}(p)\,\right]^T.$$

Substituting the solution of (9.38), which is

$$\mathbf{a}_{2:N}(p) = \mathbf{R}_{\mathrm{a},2:N}^{-1}(p)\mathbf{r}_{\mathrm{a},2:N}(p),$$

into (9.35) gives the minimum forward prediction error

$$e_{1,\min}(k,p) = y_1(k) - \mathbf{y}_{\mathrm{a},2:N}^T(k,p)\mathbf{R}_{\mathrm{a},2:N}^{-1}(p)\mathbf{r}_{\mathrm{a},2:N}(p). \tag{9.40}$$

Accordingly, we have

$$J_{1,\min}(p) = E\left\{e_{1,\min}^2(k,p)\right\} = \sigma_{y_1}^2 - \mathbf{r}_{\mathrm{a},2:N}^T(p)\mathbf{R}_{\mathrm{a},2:N}^{-1}(p)\mathbf{r}_{\mathrm{a},2:N}(p). \tag{9.41}$$

Then we can argue that the lag time p inducing a minimum $J_{1,\min}(p)$ would be the TDOA between the first two microphones:

$$\hat{\tau}^{\mathrm{FSLP}} = \arg\min_p J_{1,\min}(p), \tag{9.42}$$

where the superscript "FSLP" stands for forward spatial linear prediction.

If there are only two microphones, i.e., $N = 2$, it can be easily checked that the FLSP algorithm is identical to the CC method. However, as the number of microphones increases, the FLSP approach can take advantage of the redundant information provided by the multiple microphones to improve the

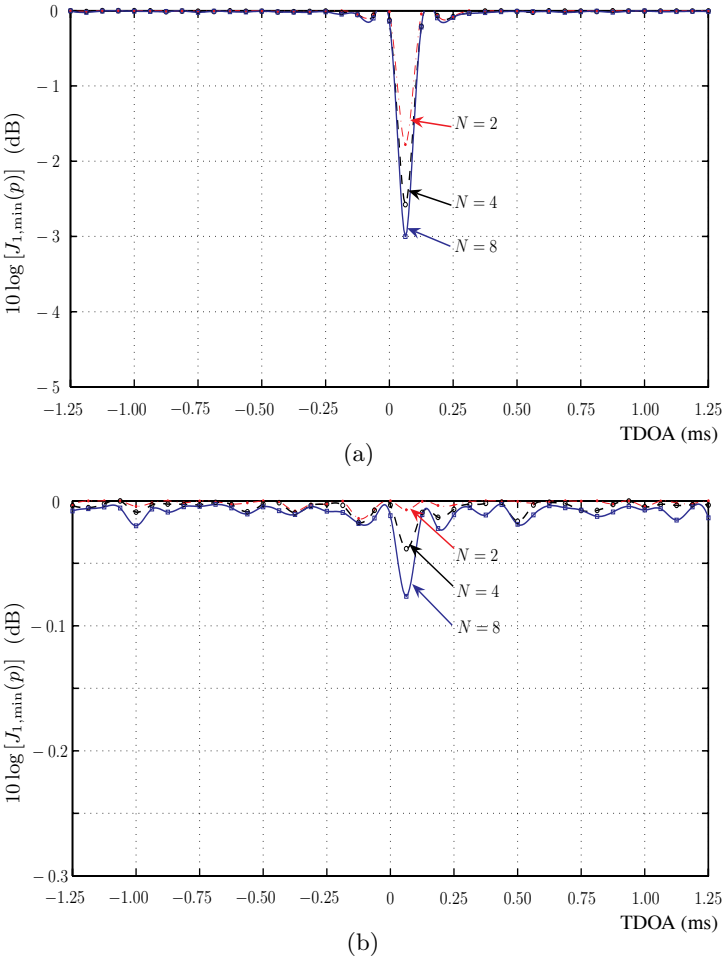

Fig. 9.6. Comparison of $J_{1,\min}(p)$ for different number of microphones. (a) SNR = 10 dB and (b) SNR = −5 dB. The source (speech) is in the array's far-field, the sampling frequency is 16 kHz, the incident angle is $\theta = 75.5°$, and the true TDOA is $\tau = 0.0625$ ms.

TDOA estimation. To illustrate this, we consider a simple simulation example where we have an equispaced linear array consisting of 10 omnidirectional microphones. The spacing between any two neighboring sensors is 8 cm. A sound source located in the far-field radiates a speech signal (female) to the array, with an incident angle of $\theta = 75.5°$. At each microphone, the signal is corrupted by a white Gaussian noise. The microphone signals are digitized with a sampling rate of 16 kHz. Figure 9.6 plots the cost function $J_{1,\min}(p)$ computed from a frame (128 ms in length) of data in two SNR conditions. When SNR = 10 dB, it is seen that the system can achieve correct estimation of the true

TDOA with only two microphones. However, as the number of microphone increases, the valley of the cost function becomes better defined, which will enable an easier search of the minimum. When SNR drops to -5 dB, this time the estimate with two microphones is incorrect. But when 4 or more microphones are employed, the system produces a correct estimate. Both situations clearly indicate that the TDOA estimation performance increases with the number of microphones

Similarly, the TDOA estimation can be developed using backward prediction or interpolation with any one of the N microphone outputs being regarded as the reference signal [39], which will be left to the reader's investigation.

9.6 Multichannel Cross-Correlation Coefficient Algorithm

It is seen from the previous section that the key to the spatial prediction based techniques is the use of the spatial correlation matrix. A more natural way of using the spatial correlation matrix in TDOA estimation is through the so-called multichannel cross-correlation coefficient (MCCC) [14], [39], which measures the correlation among the outputs of an array system and can be viewed as a seamless generalization of the classical cross-correlation coefficient to the multichannel case and where there are multiple random processes.

Following (9.36), we define a new signal vector

$$\mathbf{y}_{\mathrm{a}}(k,p) = \begin{bmatrix} y_1(k) & y_2[k+\mathcal{F}_2(p)] & \cdots & y_N[k+\mathcal{F}_N(p)] \end{bmatrix}^T. \tag{9.43}$$

Similar to (9.39), we can now write the corresponding spatial correlation matrix as

$$\begin{aligned} \mathbf{R}_{\mathrm{a}}(p) &= E\left[\mathbf{y}_{\mathrm{a}}(k,p)\mathbf{y}_{\mathrm{a}}^T(k,p)\right] \\ &= \begin{bmatrix} \sigma_{y_1}^2 & r_{\mathrm{a},y_1y_2}(p) & \cdots & r_{\mathrm{a},y_1y_N}(p) \\ r_{\mathrm{a},y_2y_1}(p) & \sigma_{y_2}^2 & \cdots & r_{\mathrm{a},y_2y_N}(p) \\ \vdots & \vdots & \ddots & \vdots \\ r_{\mathrm{a},y_Ny_1}(p) & r_{\mathrm{a},y_Ny_2}(p) & \cdots & \sigma_{y_N}^2 \end{bmatrix}. \end{aligned} \tag{9.44}$$

The spatial correlation matrix $\mathbf{R}_{\mathrm{a}}(p)$ can be factored as

$$\mathbf{R}_{\mathrm{a}}(p) = \mathbf{\Sigma}\tilde{\mathbf{R}}_{\mathrm{a}}(p)\mathbf{\Sigma}, \tag{9.45}$$

where

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{y_1} & 0 & \cdots & 0 \\ 0 & \sigma_{y_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \sigma_{y_N} \end{bmatrix}$$

is a diagonal matrix,

$$\tilde{\mathbf{R}}_{\mathrm{a}}(p) = \begin{bmatrix} 1 & \rho_{\mathrm{a},y_1y_2}(p) & \cdots & \rho_{\mathrm{a},y_1y_N}(p) \\ \rho_{\mathrm{a},y_2y_1}(\tau) & 1 & \cdots & \rho_{\mathrm{a},y_2y_N}(p) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{\mathrm{a},y_Ny_1}(p) & \rho_{\mathrm{a},y_Ny_2}(p) & \cdots & 1 \end{bmatrix}$$

is a symmetric matrix, and

$$\rho_{\mathrm{a},y_iy_j}(p) = \frac{r_{\mathrm{a},y_iy_j}(p)}{\sigma_{y_i}\sigma_{y_j}}, \quad i,j = 1,2,\ldots,N,$$

is the correlation coefficient between the ith and jth aligned microphone signals.

Since the matrix $\tilde{\mathbf{R}}_{\mathrm{a}}(p)$ is symmetric and positive semi-definite, and its diagonal elements are all equal to one, it can be shown that [14], [39]

$$0 \leq \det\left[\tilde{\mathbf{R}}_{\mathrm{a}}(p)\right] \leq 1, \tag{9.46}$$

where $\det(\cdot)$ stands for determinant.

If there are only two channel, i.e., $N = 2$, it can be easily checked that the squared correlation coefficient is linked to the normalized spatial correlation matrix by

$$\rho^2_{\mathrm{a},y_1y_2}(p) = 1 - \det\left[\tilde{\mathbf{R}}_{\mathrm{a}}(p)\right]. \tag{9.47}$$

Then by analogy, the squared MCCC among the N aligned signals $y_n[k + \mathcal{F}_n(p)]$, $n = 1,2,\ldots,N$, is constructed as

$$\begin{aligned} \rho^2_{\mathrm{a},y_1:y_N}(p) &= 1 - \det\left[\tilde{\mathbf{R}}_{\mathrm{a}}(p)\right] \\ &= 1 - \frac{\det\left[\mathbf{R}_{\mathrm{a}}(p)\right]}{\prod_{n=1}^{N}\sigma^2_{y_n}}. \end{aligned} \tag{9.48}$$

The MCCC has the following properties (presented without proof) [14], [39]:

1. $0 \leq \rho^2_{\mathrm{a},y_1:y_N}(p) \leq 1$;
2. if two or more signals are perfectly correlated, then $\rho^2_{\mathrm{a},y_1:y_N}(p) = 1$;
3. if all the signals are completely uncorrelated with each other, then $\rho^2_{\mathrm{a},y_1:y_N}(p) = 0$;
4. if one of the signals is completely uncorrelated with the $N-1$ other signals, then the MCCC will measure the correlation among those $N-1$ remaining signals.

Using the definition of the MCCC, we deduce an estimate of the TDOA between the first two microphone signals as

$$\hat{\tau}^{\mathrm{MCCC}} = \arg\max_p \rho^2_{\mathrm{a},y_1:y_N}(p), \tag{9.49}$$

which is equivalent to computing

$$\begin{aligned}\hat{\tau}^{\mathrm{MCCC}} &= \arg\max_p \left\{1 - \det\left[\tilde{\mathbf{R}}_{\mathrm{a}}(p)\right]\right\} \\ &= \arg\max_p \left\{1 - \frac{\det\left[\mathbf{R}_{\mathrm{a}}(p)\right]}{\prod_{n=1}^{N}\sigma_{y_n}^2}\right\} \\ &= \arg\min_p \det\left[\tilde{\mathbf{R}}_{\mathrm{a}}(p)\right] \\ &= \arg\min_p \det\left[\mathbf{R}_{\mathrm{a}}(p)\right]. \end{aligned} \tag{9.50}$$

To illustrate the TDOA estimation with the MCCC algorithm, we study the same example that was used in Section 9.5. The cost function $\det\left[\mathbf{R}_{\mathrm{a}}(p)\right]$ computed for the same frame of data used in Fig 9.6 is plotted in Fig 9.7. It is clearly seen that the algorithm achieves better estimation performance as more microphones are used.

To investigate the link between the MCCC and FSLP methods, let us revisit the spatial prediction error function given by (9.41). We define

$$\begin{aligned}\mathbf{a}(p) &= \left[\, a_1(p)\ a_2(p)\ \cdots\ a_N(p)\,\right]^T \\ &= \left[\, a_1(p)\ \mathbf{a}_{2:N}^T(p)\,\right]^T. \end{aligned} \tag{9.51}$$

Then, for $a_1(p) = -1$, the forward spatial prediction error signal (9.35) can be written as

$$e_1(k,p) = -\mathbf{y}_{\mathrm{a}}^T(k,p)\mathbf{a}(p), \tag{9.52}$$

and (9.37) can be expressed as

$$\begin{aligned}J_1(p) &= E\left[e_1^2(k,p)\right] + \mu\left[\mathbf{u}^T\mathbf{a}(p) + 1\right] \\ &= \mathbf{a}^T(p)\mathbf{R}_{\mathrm{a}}(p)\mathbf{a}(p) + \mu\left[\mathbf{u}^T\mathbf{a}(p) + 1\right], \end{aligned} \tag{9.53}$$

where μ is a Lagrange multiplier introduced to force $a_1(p)$ to have the value -1 and

$$\mathbf{u} = \left[\, 1\ 0\ \cdots\ 0\,\right]^T.$$

Taking the derivative of (9.53) with respect to $\mathbf{a}(p)$ and setting the result to zero yields

$$\frac{\partial J_1(p)}{\partial \mathbf{a}(p)} = 2\mathbf{R}_{\mathrm{a}}(p)\mathbf{a}(p) + \mu\mathbf{u} = \mathbf{0}_{N\times 1}. \tag{9.54}$$

Solving (9.54) for $\mathbf{a}(p)$ produces

$$\mathbf{a}(p) = -\frac{\mu\mathbf{R}_{\mathrm{a}}^{-1}(p)\mathbf{u}}{2}. \tag{9.55}$$

Substituting (9.55) into (9.53) leads to

$$J_1(p) = \mu\left[1 - \frac{\mathbf{u}^T\mathbf{R}_{\mathrm{a}}^{-1}(p)\mathbf{u}}{4}\mu\right], \tag{9.56}$$

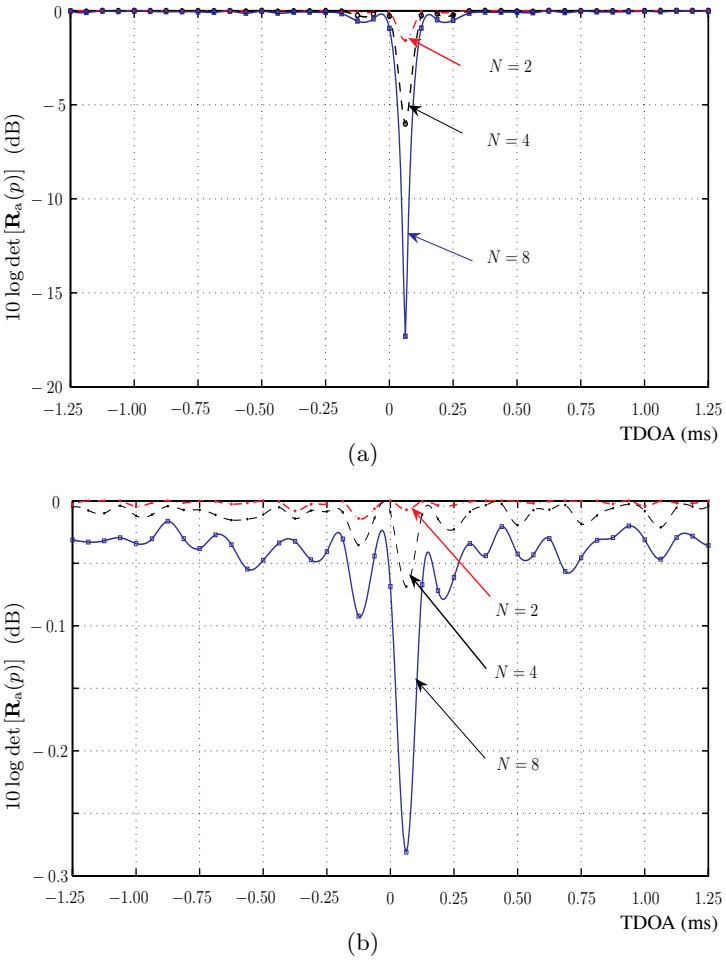

Fig. 9.7. Comparison of det $[\mathbf{R}_a(p)]$ for an equispaced linear array with different number of microphones. (a) SNR = 10 dB and (b) SNR = −5 dB. The source is in the array's far-field, the sampling frequency is 16 kHz, the incident angle is $\theta = 75.5°$, and the true TDOA is $\tau = 0.0625$ ms.

from which we know that

$$J_{1,\min}(p) = \frac{1}{\mathbf{u}^T \mathbf{R}_a^{-1}(p)\mathbf{u}}. \tag{9.57}$$

Substituting (9.45) into (9.57) and using the fact that

$$\mathbf{\Sigma}^{-1}\mathbf{u} = \frac{\mathbf{u}}{\sigma_{y_1}},$$

we have

$$J_{1,\min}(p) = \frac{\sigma_{y_1}^2}{\mathbf{u}^T \tilde{\mathbf{R}}_{\mathrm{a}}^{-1}(p)\mathbf{u}}. \tag{9.58}$$

Note that $\mathbf{u}^T\tilde{\mathbf{R}}_{\mathrm{a}}^{-1}(p)\mathbf{u}$ is the $(1,1)$th element of the matrix $\tilde{\mathbf{R}}_{\mathrm{a}}^{-1}(p)$, which is computed using the adjoint method as the $(1,1)$th cofactor of $\tilde{\mathbf{R}}_{\mathrm{a}}(p)$ divided by the determinant of $\tilde{\mathbf{R}}_{\mathrm{a}}(p)$, i.e.,

$$\mathbf{u}^T\tilde{\mathbf{R}}_{\mathrm{a}}^{-1}(p)\mathbf{u} = \frac{\det\left[\tilde{\mathbf{R}}_{\mathrm{a,2:N}}(p)\right]}{\det\left[\tilde{\mathbf{R}}_{\mathrm{a}}(p)\right]}, \tag{9.59}$$

where $\tilde{\mathbf{R}}_{\mathrm{a,2:N}}(p)$ is the lower-right submatrix of $\tilde{\mathbf{R}}_{\mathrm{a}}(p)$ by removing the first row and the first column. By substituting (9.59) into (9.58), we get

$$J_{1,\min}(p) = \sigma_{y_1}^2 \cdot \frac{\det\left[\tilde{\mathbf{R}}_{\mathrm{a}}(p)\right]}{\det\left[\tilde{\mathbf{R}}_{\mathrm{a,2:N}}(p)\right]}. \tag{9.60}$$

Therefore, the FSLP estimate of τ is found as

$$\begin{aligned}\hat{\tau}^{\mathrm{FSLP}} &= \arg\min_p J_{1,\min}(p) \\ &= \arg\min_p \frac{\det\left[\tilde{\mathbf{R}}_{\mathrm{a}}(p)\right]}{\det\left[\tilde{\mathbf{R}}_{\mathrm{a,2:N}}(p)\right]}.\end{aligned} \tag{9.61}$$

Comparing (9.50) to (9.61) reveals a clear distinction between the two methods in spite of high similarity. In practice, the FSLP method may suffer numerical instabilities since the calculation of the FSLP cost function (9.60) involves the division by $\det\left[\tilde{\mathbf{R}}_{\mathrm{a,2:N}}(p)\right]$, while the MCCC method is found to be fairly stable. If we compare Figs. 9.6 and 9.7, one can notice that the peak of the MCCC cost function is better defined than that of the FSLP function, which indicates that the MCCC algorithm is superior to the FSLP method.

It is worth pointing out that the microphone outputs can be pre-whitened before computing their MCCC as was done in the PHAT algorithm in the two-channel scenario. By doing so, the TDOA estimation algorithms become more robust to the volume variation of the speech source signal.

9.7 Eigenvector-Based Techniques

Another way to use the spatial correlation matrix for TDOA estimation is through the eigenvector-based techniques. These techniques were originally developed in radar for DOA estimation [184], [192], [198], and have been recently extended to processing a broadband speech using microphone arrays

[58]. We start with the narrowband formulation since it is much easier to comprehend. We consider the single-source free-field model in (9.5) with N microphones. For ease of analysis, we assume that the source is in the array's far-field, all the attenuation coefficients α_n are equal to 1, and the observation noises $v_n(k)$, $n = 1, 2, \ldots, N$, are mutually independent Gaussian random processes with the same variance.

9.7.1 Narrowband MUSIC

If we transform both sides of (9.5) into the frequency domain, the nth microphone output can be written as

$$\begin{aligned} Y_n(f) &= X_n(f) + V_n(f) \\ &= S(f)e^{-j2\pi[t+\mathcal{F}_n(\tau)]f} + V_n(f), \end{aligned} \tag{9.62}$$

where $Y_n(f)$, $X_n(f)$, $V_n(f)$, and $S(f)$ are, respectively, the DTFT of $y_n(k)$, $x_n(k)$, $v_n(k)$, and $s(k)$. Let us define the following frequency-domain vector:

$$\vec{\mathbf{y}} = \left[Y_1(f)\ Y_2(f) \cdots Y_N(f) \right]^T. \tag{9.63}$$

Substituting (9.62) into (9.63), we get

$$\begin{aligned} \vec{\mathbf{y}} &= \vec{\mathbf{x}} + \vec{\mathbf{v}} \\ &= \boldsymbol{\varsigma}(\tau)S(f)e^{-j2\pi tf} + \vec{\mathbf{v}}, \end{aligned} \tag{9.64}$$

where

$$\boldsymbol{\varsigma}(\tau) = \left[e^{-j2\pi\mathcal{F}_1(\tau)f} \quad e^{-j2\pi\mathcal{F}_2(\tau)f} \quad \cdots \quad e^{-j2\pi\mathcal{F}_N(\tau)f} \right]^T,$$

and $\vec{\mathbf{v}}$ is defined similarly to $\vec{\mathbf{y}}$. It follows that the output covariance matrix can be written as

$$\mathbf{R}_Y = E\left(\vec{\mathbf{y}}\vec{\mathbf{y}}^H\right) = \mathbf{R}_X + \sigma_V^2\mathbf{I}, \tag{9.65}$$

where

$$\mathbf{R}_X = \sigma_S^2\boldsymbol{\varsigma}(\tau)\boldsymbol{\varsigma}^H(\tau), \tag{9.66}$$

and $\sigma_S^2 = E[|S(f)|^2]$ and $\sigma_V^2 = E[|V_1(f)|^2] = \cdots = E[|V_N(f)|^2]$ are, respectively, the signal and noise variances. It can be easily checked that the positive semi-definite matrix $\mathbf{R}_X$ is of rank 1. Therefore, if we perform the eigenvalue decomposition of $\mathbf{R}_Y$, we obtain

$$\mathbf{R}_Y = \mathbf{B}\boldsymbol{\Lambda}\mathbf{B}^H, \tag{9.67}$$

where

$$\begin{aligned}\mathbf{\Lambda} &= \mathrm{diag}[\,\lambda_{Y,1} \quad \lambda_{Y,2} \quad \cdots \quad \lambda_{Y,N}\,] \\ &= \mathrm{diag}[\,\lambda_{X,1}+\sigma_V^2 \quad \sigma_V^2 \quad \cdots \quad \sigma_V^2\,]\end{aligned} \tag{9.68}$$

is a diagonal matrix consisting of the eigenvalues of $\mathbf{R}_Y$,

$$\mathbf{B} = \left[\,\mathbf{b}_1\ \mathbf{b}_2 \cdots \mathbf{b}_N\,\right], \tag{9.69}$$

$\mathbf{b}_n$ is the eigenvector associated with the eigenvalue $\lambda_{Y,n}$, and $\lambda_{X,1}$ is the only non-zero positive eigenvalue of $\mathbf{R}_X$.

Therefore, for $n \geq 2$, we have

$$\mathbf{R}_Y\mathbf{b}_n = \lambda_{Y,n}\mathbf{b}_n = \sigma_V^2\mathbf{b}_n. \tag{9.70}$$

We also know that

$$\mathbf{R}_Y\mathbf{b}_n = \left[\sigma_S^2\boldsymbol{\varsigma}(\tau)\boldsymbol{\varsigma}^H(\tau) + \sigma_V^2\mathbf{I}\right]\mathbf{b}_n. \tag{9.71}$$

The combination of (9.70) and (9.71) indicates that

$$\sigma_S^2\boldsymbol{\varsigma}(\tau)\boldsymbol{\varsigma}^H(\tau)\mathbf{b}_n = \mathbf{0}, \tag{9.72}$$

which is equivalent to

$$\boldsymbol{\varsigma}^H(\tau)\mathbf{b}_n = 0 \tag{9.73}$$

or

$$\mathbf{b}_n^H\boldsymbol{\varsigma}(\tau) = 0 \tag{9.74}$$

This is to say that the eigenvectors associated with the $N-1$ lowest eigenvalues of $\mathbf{R}_Y$ are orthogonal to the vector corresponding to the actual TDOA. This remarkable observation forms the cornerstone for almost all eigenvector-based algorithms. If we form the following cost function

$$J_{\mathrm{MUSIC}}(p) = \frac{1}{\displaystyle\sum_{n=2}^{N}\left|\mathbf{b}_n^H\boldsymbol{\varsigma}(p)\right|^2}, \tag{9.75}$$

where the subscript "MUSIC" stands for MUltiple SIgnal Classification (MUSIC) [198]. The lag time p that gives the maximum of $J_{\mathrm{MUSIC}}(p)$ corresponds to the TDOA τ:

$$\hat{\tau}^{\mathrm{MUSIC}} = \arg\max_p J_{\mathrm{MUSIC}}(p). \tag{9.76}$$

9.7.2 Broadband MUSIC

While the narrowband formulation of the MUSIC algorithm is straightforward to follow, it does not work well for microphone arrays because speech is nonstationary in nature. Even during the presence of speech, each frequency band may not permanently be occupied with speech, and for a large percentage of the time the band may consist of noise only. One straightforward way of circumventing this issue is to fuse the cost function given in (9.75) across all the frequency bands before searching for the TDOA. This fusion method will, in general, make the peak less well defined, thereby degrading the estimation performance. A more natural broadband MUSIC formulation has been recently developed [58]. This broadband MUSIC is derived based on the spatial correlation matrix defined in Section 9.5. Let us rewrite the alignment signal vector given in (9.43),

$$\mathbf{y}_{\mathrm{a},1:N}(k,p) = \left[\, y_1[k+\mathcal{F}_1(p)]\; y_2[k+\mathcal{F}_2(p)] \cdots y_N[k+\mathcal{F}_N(p)] \,\right]^T, \quad (9.77)$$

The spatial correlation matrix is given by

$$\begin{aligned} \mathbf{R}_{\mathrm{a}}(p) &= E\left[\mathbf{y}_{\mathrm{a},1:N}(k,p)\mathbf{y}_{\mathrm{a},1:N}^T(k,p)\right] \\ &= \mathbf{R}_s(p) + \sigma_v^2\mathbf{I}, \end{aligned} \quad (9.78)$$

where the source signal covariance matrix $\mathbf{R}_s(p)$ is given by

$$\mathbf{R}_s(p) = \begin{bmatrix} \sigma_s^2 & r_{ss,12}(p,\tau) & \cdots & r_{ss,1N}(p,\tau) \\ r_{ss,21}(p,\tau) & \sigma_s^2 & \cdots & r_{ss,2N}(p,\tau) \\ \vdots & \vdots & \ddots & \vdots \\ r_{ss,N1}(p,\tau) & r_{ss,N2}(p,\tau) & \cdots & \sigma_s^2 \end{bmatrix}, \quad (9.79)$$

and

$$r_{ss,ij}(p,\tau) = E\left\{s[k-t-\mathcal{F}_i(\tau)+\mathcal{F}_i(p)]s[k-t-\mathcal{F}_j(\tau)+\mathcal{F}_j(p)]\right\}. \quad (9.80)$$

If $p = \tau$, we easily check that

$$\mathbf{R}_s(\tau) = \sigma_s^2 \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}, \quad (9.81)$$

which is a matrix of rank 1. If $p \neq \tau$, the rank of this matrix depends on the characteristics of the source signal. If the source signal is a white process, we can see that $\mathbf{R}_s(p)$ is a diagonal matrix with $\mathbf{R}_s(p) = \mathrm{diag}\left[\,\sigma_s^2 \;\; \sigma_s^2 \;\; \cdots \;\; \sigma_s^2\,\right]$. In this particular case, $\mathbf{R}_s(p)$ is of full rank. In general, if $p \neq \tau$, $\mathbf{R}_s(p)$ is positive semi-definite, and its rank is greater than 1. Let us perform the eigenvalue decomposition of $\mathbf{R}_{\mathrm{a}}(p)$ and $\mathbf{R}_s(p)$. Let

$\lambda_{s,1}(p) \geq \lambda_{s,2}(p) \geq \cdots \geq \lambda_{s,N}(p)$ denote the N eigenvalues of $\mathbf{R}_s(p)$. Then the N eigenvalues of $\mathbf{R}_{\mathrm{a}}(p)$ are given by

$$\lambda_{y,n}(p) = \lambda_{s,n}(p) + \sigma_v^2. \tag{9.82}$$

Further let $\mathbf{b}_1(p), \mathbf{b}_2(p), \cdots, \mathbf{b}_N(p)$ denote their associated eigenvectors (since $\mathbf{R}_{\mathrm{a}}(p)$ is symmetric and Toeplitz, all the eigenvectors are real-valued), then

$$\mathbf{R}_{\mathrm{a}}(p)\mathbf{B}(p) = \mathbf{B}(p)\mathbf{\Lambda}(p), \tag{9.83}$$

where

$$\mathbf{B}(p) = \left[\, \mathbf{b}_1(p)\ \mathbf{b}_2(p) \cdots \mathbf{b}_N(p) \,\right], \tag{9.84}$$

$$\mathbf{\Lambda}(p) = \mathrm{diag}\left[\, \lambda_{y,1}(p)\ \lambda_{y,2}(p) \cdots \lambda_{y,N}(p) \,\right]. \tag{9.85}$$

When $p = \tau$, we already know that $\mathbf{R}_{\mathrm{s}}(\tau)$ is of rank 1. Therefore, for $n \geq 2$, we have

$$\mathbf{R}_{\mathrm{a}}(\tau)\mathbf{b}_n(\tau) = \left[\mathbf{R}_{\mathrm{s}}(\tau) + \sigma_v^2\mathbf{I}\right]\mathbf{b}_n(\tau) = \sigma_v^2\mathbf{b}_n(\tau), \tag{9.86}$$

which implies

$$\mathbf{b}_n^T(p)\mathbf{R}_{\mathrm{a}}(p)\mathbf{b}_n(p) = \begin{cases} \sigma_v^2, & p = \tau \\ \lambda_{y,n}(p) \geq \sigma_v^2, & p \neq \tau \end{cases}. \tag{9.87}$$

Therefore, if we form the following function

$$J_{\mathrm{BMUSIC}}(p) = \frac{1}{\sum_{n=2}^{N} \mathbf{b}_n^T(p)\mathbf{R}_{\mathrm{a}}(p)\mathbf{b}_n(p)}, \tag{9.88}$$

the peak of this cost function will correspond to the true TDOA τ:

$$\hat{\tau}^{\mathrm{BMUSIC}} = \arg\max_p J_{\mathrm{BMUSIC}}(p). \tag{9.89}$$

Although the forms of the broadband and narrowband MUSIC algorithms look similar, they are different in many aspects, such as

- the broadband algorithm can take either broadband or narrowband signals as its inputs, while the narrowband algorithm can only work for narrowband signals;
- in the narrowband case, we only need to perform the eigenvalue decomposition once, but in the broadband situation we will have to compute the eigenvalue decomposition for all the spatial correlation matrices $\mathbf{R}_{\mathrm{a}}(p)$, $-\tau_{\max} \leq p \leq \tau_{\max}$;
- in the narrowband case, when $p = \tau$, the objective function $J_{\mathrm{MUSIC}}(p)$ approaches infinity, so the peak is well defined. However, in the broadband situation, the maximum of the cost function $J_{\mathrm{BMUSIC}}(p)$ is $1/[(N-1)\sigma_v^2]$, which indicates that the peak may be less well-defined.

9.8 Minimum Entropy Method

So far, we have explored the use of the cross-correlation information between different channels for TDOA estimation. The correlation coefficient, regardless if it is computed between two or multiple channels, is a second-order-statistics (SOS) measure of dependence between random Gaussian variables. However for non-Gaussian source signals such as speech, higher order statistics (HOS) may have more to say about their dependence. This section discusses the use of HOS for TDOA estimation through the concept of entropy.

Entropy is a statistical (apparently HOS) measure of randomness or uncertainty of a random variable; it was introduced by Shannon in the context of communication theory [203]. For a random variable y with a probability density function (PDF) $p(y)$ (note here we choose not to distinguish random variables and their realizations), the entropy is defined as [52]

$$\begin{aligned} H(y) &= -\int p(y) \ln p(y) dy \\ &= -E\left[\ln p(y)\right]. \end{aligned} \tag{9.90}$$

The entropy (in the continuous case) is a measure of the structure contained in the PDF [146]. As far as the multivariate random variable $\mathbf{y}_{\mathrm{a}}(k,p)$ given by (9.43) is concerned, the joint entropy is

$$H\left[\mathbf{y}_{\mathrm{a}}(k,p)\right] = -\int p\left[\mathbf{y}_{\mathrm{a}}(k,p)\right] \ln p\left[\mathbf{y}_{\mathrm{a}}(k,p)\right] d\mathbf{y}_{\mathrm{a}}(k,p). \tag{9.91}$$

It was then argued in [19] that the time lag p that gives the minimum of $H\left[\mathbf{y}_{\mathrm{a}}(k,p)\right]$ corresponds to the TDOA between the two microphones:

$$\hat{\tau}^{\mathrm{ME}} = \arg\min_{p} H\left[\mathbf{y}_{\mathrm{a}}(k,p)\right], \tag{9.92}$$

where the superscript "ME" refers to the minimum entropy method.

9.8.1 Gaussian Source Signal

If the source is Gaussian, so are the microphone outputs in the absence of noise. Suppose that the aligned microphone signals are zero mean and joint Gaussian random signals. Their joint PDF is then given by

$$p\left[\mathbf{y}_{\mathrm{a}}(k,p)\right] = \frac{\exp\left[-\eta_{\mathrm{a}}(k,p)/2\right]}{\sqrt{(2\pi)^N \det\left[\mathbf{R}_{\mathrm{a}}(p)\right]}}, \tag{9.93}$$

where

$$\eta_{\mathrm{a}}(k,p) = \mathbf{y}_{\mathrm{a}}^T(k,p)\mathbf{R}_{\mathrm{a}}^{-1}(p)\mathbf{y}_{\mathrm{a}}(k,p). \tag{9.94}$$

By substituting (9.93) into (9.91), the joint entropy can be computed [19] as

$$H\left[\mathbf{y}_{\mathrm{a}}(k,p)\right] = \frac{1}{2}\ln\left\{(2\pi e)^N \det\left[\mathbf{R}_{\mathrm{a}}(p)\right]\right\}. \tag{9.95}$$

Consequently, (9.92) becomes

$$\hat{\tau}^{\mathrm{ME}} = \arg\min_{p} \det\left[\mathbf{R}_{\mathrm{a}}(p)\right]. \tag{9.96}$$

It is clear from (9.50) and (9.96) that minimizing the entropy is equivalent to maximizing the MCCC for Gaussian source signals.

9.8.2 Speech Source Signal

Speech is a complicated random process and there is no rigorous mathematical formula for its entropy. But in speech research, it was found that speech can be fairly well modeled by a Laplace distribution [85], [186].

The univariate Laplace distribution with mean zero and variance σ_y^2 is given by

$$p(y) = \frac{\sqrt{2}}{2\sigma_y} e^{-\sqrt{2}|y|/\sigma_y}, \tag{9.97}$$

and the corresponding entropy is [52]

$$H(y) = 1 + \ln\left(\sqrt{2}\ \sigma_y\right). \tag{9.98}$$

Suppose that $\mathbf{y}_{\mathrm{a}}(k,p)$ has a multivariate Laplace distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{R}_{\mathrm{a}}(p)$ [147], [67]:

$$p\left[\mathbf{y}_{\mathrm{a}}(k,p)\right] = \frac{2\left[\eta_{\mathrm{a}}(k,p)/2\right]^{Q/2} K_Q\left[\sqrt{2\eta_{\mathrm{a}}(k,p)}\right]}{\sqrt{(2\pi)^N \det\left[\mathbf{R}_{\mathrm{a}}(p)\right]}}, \tag{9.99}$$

where $Q = (2-N)/2$ and $K_Q(\cdot)$ is the modified Bessel function of the third kind (also called the modified Bessel function of the second kind) given by

$$K_Q(a) = \frac{1}{2}\left(\frac{a}{2}\right)^Q \int_0^\infty z^{-Q-1} \exp\left(-z - \frac{a^2}{4z}\right) dz, \quad a > 0. \tag{9.100}$$

The joint entropy is

$$H\left[\mathbf{y}_{\mathrm{a}}(k,p)\right] = \frac{1}{2}\ln\left\{\frac{(2\pi)^N \det\left[\mathbf{R}_{\mathrm{a}}(p)\right]}{4}\right\} - \frac{Q}{2} E\left\{\ln\left[\frac{\eta_{\mathrm{a}}(k,p)}{2}\right]\right\} - E\left\{\ln K_Q\left[\sqrt{2\eta_{\mathrm{a}}(k,p)}\right]\right\}. \tag{9.101}$$

The two quantities $E\left\{\ln\left[\eta_{\mathrm{a}}(k,p)/2\right]\right\}$ and $E\left\{\ln K_Q\left[\sqrt{2\eta_{\mathrm{a}}(k,p)}\right]\right\}$ do not seem to have a closed form. So a numerical scheme needs to be developed to estimate them. One possibility to do this is the following. Assume that all

processes are ergodic. As a result, ensemble averages can be replaced by time averages. If there are K samples for each element of the observation vector $\mathbf{y}_{\mathrm{a}}(k,p)$, the following estimators were proposed in [19]:

$$E\left\{\ln\left[\eta_{\mathrm{a}}(k,p)/2\right]\right\} \approx \frac{1}{K}\sum_{k=0}^{K-1}\ln\left[\eta_{\mathrm{a}}(k,p)/2\right], \tag{9.102}$$

$$E\left\{\ln K_Q\left[\sqrt{2\eta_{\mathrm{a}}(k,p)}\right]\right\} \approx \frac{1}{K}\sum_{k=0}^{K-1}\ln K_Q\left[\sqrt{2\eta_{\mathrm{a}}(k,p)}\right]. \tag{9.103}$$

The simulation results presented in [19] show that the ME algorithm performs in general comparably to or better than the MCCC algorithm. Apparently the ME algorithm is computationally intensive. But the idea of using entropy expands our horizon of knowledge in pursuit of new TDOA estimation algorithms.

9.9 Adaptive Eigenvalue Decomposition Algorithm

The adaptive eigenvalue decomposition (AED) algorithm approaches the TDOA estimation problem from a different point of view as compared to the methods discussed in the previous sections. Similar to the GCC family, the AED considers only the scenario with a single source and two microphones, but it adopts the real reverberant model instead of the free-field model. It first identifies the two channel impulse responses from the source to the two sensors, and then measures the TDOA by detecting the two direct paths. Since the source signal is unknown, the channel identification has to be a blind method.

Following the single source reverberant model (9.9) and the fact that, in the absence of additive noise,

$$y_1(k) * g_2 = x_1(k) * g_2 = s(k) * g_1 * g_2 = x_2(k) * g_1 = y_2(k) * g_1, \tag{9.104}$$

we deduce the following cross-relation in vector/matrix form at time k:

$$\mathbf{y}^T(k)\mathbf{w} = \mathbf{y}_1^T(k)\mathbf{g}_2 - \mathbf{y}_2^T(k)\mathbf{g}_1 = 0, \tag{9.105}$$

where

$$\begin{aligned}
\mathbf{y}(k) &= \left[\,\mathbf{y}_1^T(k)\ \mathbf{y}_2^T(k)\,\right]^T,\\
\mathbf{w} &= \left[\,\mathbf{g}_2^T\ -\mathbf{g}_1^T\,\right]^T,\\
\mathbf{g}_n &= \left[\,g_{n,0}\ g_{n,1}\ \cdots\ g_{n,L-1}\,\right]^T,\quad n=1,2.
\end{aligned}$$

Multiplying (9.105) by $\mathbf{y}(k)$ from the left-hand side and taking expectation yields

$$\mathbf{R}_{yy}\mathbf{w} = \mathbf{0}_{2L\times 1}, \tag{9.106}$$

where $\mathbf{R}_{yy} = E\left[\mathbf{y}(k)\mathbf{y}^T(k)\right]$ is the covariance matrix of the two microphone signals. This indicates that the vector $\mathbf{w}$, which consists of the two impulse responses, is in the null space of $\mathbf{R}_{yy}$. More specifically, $\mathbf{w}$ is an eigenvector of $\mathbf{R}_{yy}$ corresponding to the eigenvalue 0. If $\mathbf{R}_{yy}$ is rank deficient by 1, $\mathbf{w}$ can be uniquely determined up to a scaling factor, which is equivalent to saying that the two-channel SIMO system can be blindly identified. Using what has been proved in [238], we know that such a two-channel acoustic SIMO system is blindly identifiable using only the second-order statistics (SOS) of the microphone outputs if and only if the following two conditions hold:

- the polynomials formed from $\mathbf{g}_1$ and $\mathbf{g}_2$ are co-prime, i.e., their channel transfer functions share no common zeros;
- the autocorrelation matrix $\mathbf{R}_{ss} = E[\mathbf{s}(k)\mathbf{s}^T(k)]$ of the source signal is of full rank (such that the SIMO system can be fully excited).

In practice, noise always exists and the covariance matrix $\mathbf{R}_{yy}$ is positive definite rather than positive semi-definite. As a consequence, $\mathbf{w}$ is found as the normalized eigenvector of $\mathbf{R}_{yy}$ corresponding to the smallest eigenvalue:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathbf{w}^T\mathbf{R}_{yy}\mathbf{w} \quad \text{subject to} \quad \|\mathbf{w}\| = 1. \tag{9.107}$$

In the AED algorithm, solving (9.107) is carried out in an adaptive manner using a constrained LMS algorithm:

$$\begin{aligned}
&\text{Initialize}\\
&\hat{\mathbf{g}}_n(0) = \left[\begin{array}{cccc}\frac{\sqrt{2}}{2} & 0 & \cdots & 0\end{array}\right]^T, \quad n = 1, 2,\\
&\hat{\mathbf{w}}(0) = \left[\begin{array}{cc}\hat{\mathbf{g}}_2^T(0) & -\hat{\mathbf{g}}_1^T(0)\end{array}\right]^T,\\
&\\
&\text{Compute,} \quad \text{for } k = 0, 1, \ldots\\
&e(k) = \hat{\mathbf{w}}^T(k)\mathbf{y}(k),\\
&\hat{\mathbf{w}}(k+1) = \frac{\hat{\mathbf{w}}(k) - \mu e(k)\mathbf{y}(k)}{\|\hat{\mathbf{w}}(k) - \mu e(k)\mathbf{y}(k)\|},
\end{aligned} \tag{9.108}$$

where the adaptation step size μ is a small positive constant.

After the AED algorithm converges, the time difference between the direct paths of the two identified channel impulse responses $\hat{\mathbf{g}}_1$ and $\hat{\mathbf{g}}_2$ is measured as the TDOA estimate:

$$\hat{\tau}^{\text{AED}} = \arg\max_l |\hat{g}_{1,l}| - \arg\max_l |\hat{g}_{2,l}|. \tag{9.109}$$

9.10 Adaptive Blind Multichannel Identification Based Methods

The AED algorithm provides us a new way to look at the TDOA estimation problem, which was found particularly robust in a reverberant environment. It applies the more realistic real-reverberant model to a two-microphone acoustic system at a time and attempts to blindly identify the two-channel impulse responses, from which the embedded TDOA information of interest is then extracted. Clearly the blind two-channel identification technique plays a central role in such an approach. The more accurately the two impulse responses are blindly estimated, the more precisely the TDOA can be inferred. But for a two-channel system, the zeros of the two channels can be close especially when their impulse responses are long, which leads to an ill-conditioned system that is difficult to identify. If they share some common zeros, the system becomes unidentifiable (using only second-order statistics) and the AED algorithm may not be better than the GCC methods. It was suggested in [120] that this problem can be alleviated by employing more microphones. When more microphones are employed, it is less likely for all channels to share a common zero. As such, blind identification deals with a more well-conditioned SIMO system and the solutions can be globally optimized over all channels. The resulting algorithm is referred as the adaptive blind multichannel identification (ABMCI) based TDOA estimation.

The generalization of blind SIMO identification from two channels to multiple (> 2) channels is not straightforward and in [118] a systematic way was proposed. Consider a SIMO system with N channels whose outputs are described by (9.10). Each pair of the system outputs has a cross-relation in the absence of noise:

$$\mathbf{y}_i^T(k)\mathbf{g}_j = \mathbf{y}_j^T\mathbf{g}_i, \quad i,j = 1,2,\ldots,N. \tag{9.110}$$

When noise is present or the channel impulse responses are improperly modeled, the cross-relation does not hold and an *a priori* error signal can be defined as follows:

$$e_{ij}(k+1) = \frac{\mathbf{y}_i^T(k+1)\hat{\mathbf{g}}_j(k) - \mathbf{y}_j^T(k+1)\hat{\mathbf{g}}_i(k)}{\|\hat{\mathbf{g}}(k)\|}, \quad i,j = 1,2,\ldots,N, \tag{9.111}$$

where $\hat{\mathbf{g}}_i(k)$ is the model filter for the ith channel at time k and

$$\hat{\mathbf{g}}(k) = \left[\hat{\mathbf{g}}_1^T(k)\ \hat{\mathbf{g}}_2^T(k)\cdots\hat{\mathbf{g}}_N^T(k)\right]^T.$$

The model filters are normalized in order to avoid a trivial solution whose elements are all zeros. Based on the error signal defined here, a cost function at time $k+1$ is given by

$$J(k+1) = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} e_{ij}^2(k+1). \tag{9.112}$$

The multichannel LMS (MCLMS) algorithm updates the estimate of the channel impulse responses as follows:

$$\hat{\mathbf{g}}(k+1) = \hat{\mathbf{g}}(k) - \mu \nabla J(k+1), \tag{9.113}$$

where μ is again a small positive step size. As shown in [118], the gradient of $J(k+1)$ is computed as

$$\nabla J(k+1) = \frac{\partial J(k+1)}{\partial \hat{\mathbf{g}}(k)} = \frac{2\left[\bar{\mathbf{R}}_{y+}(k+1)\hat{\mathbf{g}}(k) - J(k+1)\hat{\mathbf{g}}(k)\right]}{\|\hat{\mathbf{g}}(k)\|^2}, \tag{9.114}$$

where

$$\bar{\mathbf{R}}_{y+}(k) = \begin{bmatrix} \sum\limits_{n\neq 1} \bar{\mathbf{R}}_{y_n y_n}(k) & -\bar{\mathbf{R}}_{y_2 y_1}(k) & \cdots & -\bar{\mathbf{R}}_{y_N y_1}(k) \\ -\bar{\mathbf{R}}_{y_1 y_2}(k) & \sum\limits_{n\neq 2} \bar{\mathbf{R}}_{y_n y_n}(k) & \cdots & -\bar{\mathbf{R}}_{y_N y_2}(k) \\ \vdots & \vdots & \ddots & \vdots \\ -\bar{\mathbf{R}}_{y_1 y_N}(k) & -\bar{\mathbf{R}}_{y_2 y_N}(k) & \cdots & \sum\limits_{n\neq N} \bar{\mathbf{R}}_{y_n y_n}(k) \end{bmatrix},$$

and

$$\bar{\mathbf{R}}_{y_i y_j}(k) = \mathbf{y}_i(k)\mathbf{y}_j^T(k), \quad i,j = 1,2,\ldots,N.$$

If the model filters are always normalized after each update, then a simplified MCLMS algorithm is obtained

$$\hat{\mathbf{g}}(k+1) = \frac{\hat{\mathbf{g}}(k) - 2\mu\left[\bar{\mathbf{R}}_{y+}(k+1)\hat{\mathbf{g}}(k) - J(k+1)\hat{\mathbf{g}}(k)\right]}{\left\|\hat{\mathbf{g}}(k) - 2\mu\left[\bar{\mathbf{R}}_{y+}(k+1)\hat{\mathbf{g}}(k) - J(k+1)\hat{\mathbf{g}}(k)\right]\right\|}. \tag{9.115}$$

A number of other adaptive blind SIMO identification algorithms were also developed with faster convergence and lower computational complexity, e.g., [119], [122]. But we would like to refer the reader to [125] and references therein for more details.

After the adaptive algorithm converges, the TDOA τ is determined as

$$\hat{\tau}^{\mathrm{ABMCI}} = \arg\max_l |\hat{g}_{1,l}| - \arg\max_l |\hat{g}_{2,l}|. \tag{9.116}$$

More generally, the TDOA between any two microphones can be inferred as

$$\hat{\tau}_{ij}^{\mathrm{ABMCI}} = \arg\max_l |\hat{g}_{i,l}| - \arg\max_l |\hat{g}_{j,l}|, \quad i,j = 1,2,\ldots,N, \tag{9.117}$$

where we have assumed that in every channel the direct path is always dominant. This is generally true for acoustic waves, which would be considerably attenuated by wall reflection. But sometimes two or more reverberant signals via multipaths of equal delay could add coherently such that the direct-path component no longer dominates the impulse response. Therefore a more robust

way to pick the direct-path component is to identify the Q ($Q > 1$) strongest elements in the impulse responses and choose the one with the smallest delay [120]:

$$\hat{\tau}_{ij}^{\text{ABMCI}} = \min\left[\arg\max_{l}{}^{q}|\hat{g}_{i,l}|\right] - \min\left[\arg\max_{l}{}^{q}|\hat{g}_{j,l}|\right], \quad i,j = 1,2,\ldots,N,\ q = 1,2,\ldots,Q, \tag{9.118}$$

where $\max^q$ computes the qth largest element.

9.11 TDOA Estimation of Multiple Sources

So far, we have assumed that there is only one source in the sound field. In many applications such as teleconferencing and telecollaboration, there may be multiple sound sources active at the same time. In this section, we consider the problem of TDOA estimation for the scenarios where there are more than one source in the array's field of view. Fundamentally, the TDOA estimation in such situations consists of two steps, i.e., determining the number of sources, and estimating the TDOA due to each sound source. Here we assume that the number of sources is known *a priori*, so we focus our discussion on the second step only.

Many algorithms discussed in Sections 9.3–9.8 can be used or extended for TDOA estimation of multiple sources. Let us take, for example, the CC method. When there are two sources, using the signal model given in (9.8), we can write the CCF between $y_1(k)$ and $y_2(k)$ as

$$\begin{aligned} r_{y_1y_2}^{\text{CC}}(p) = {} & \alpha_{11}\alpha_{21}r_{s_1s_1}^{\text{CC}}(p-\tau_1) + \alpha_{11}\alpha_{22}r_{s_1s_2}^{\text{CC}}(t_1+p-t_2-\tau_2) + \\ & \alpha_{11}r_{s_1v_2}^{\text{CC}}(p+t_1) + \alpha_{12}\alpha_{21}r_{s_2s_1}^{\text{CC}}(p+t_2-t_1-\tau_1) + \\ & \alpha_{12}\alpha 22r_{s_2s_2}^{\text{CC}}(p-\tau_2) + \alpha_{12}r_{s_2v_2}^{\text{CC}}(p+t_2) + \\ & \alpha_{2,1}r_{v_1s_1}^{\text{CC}}(p-t_1-\tau_1) + \alpha_{22}r_{v_1s_2}^{\text{CC}}(p-t_2-\tau_2) + \\ & r_{v_1v_2}^{\text{CC}}(p). \end{aligned} \tag{9.119}$$

Noting that the source signals are assumed to be mutually independent with each other and the noise signal at one sensor is assumed to be uncorrelated with the source signals and the noise at the other microphones, we get

$$r_{y_1y_2}^{\text{CC}}(p) = \alpha_{11}\alpha_{21}r_{s_1s_1}^{\text{CC}}(p-\tau_1) + \alpha_{12}\alpha_{22}r_{s_2s_2}^{\text{CC}}(p-\tau_2). \tag{9.120}$$

The two correlation functions $r_{s_1s_1}^{\text{CC}}(p-\tau_1)$ and $r_{s_2s_2}^{\text{CC}}(p-\tau_2)$ will reach their respective maximum at $p = \tau_1$ and $p = \tau_2$. Therefore, we should expect to see two large peaks in $r_{y_1y_2}^{\text{CC}}(p)$, each corresponding to the TDOA of one source. The same result applies to all the GCC methods [26], [27].

To illustrate the TDOA estimation of two sources using the correlation based method, we consider the simulation example used in Section 9.5 except

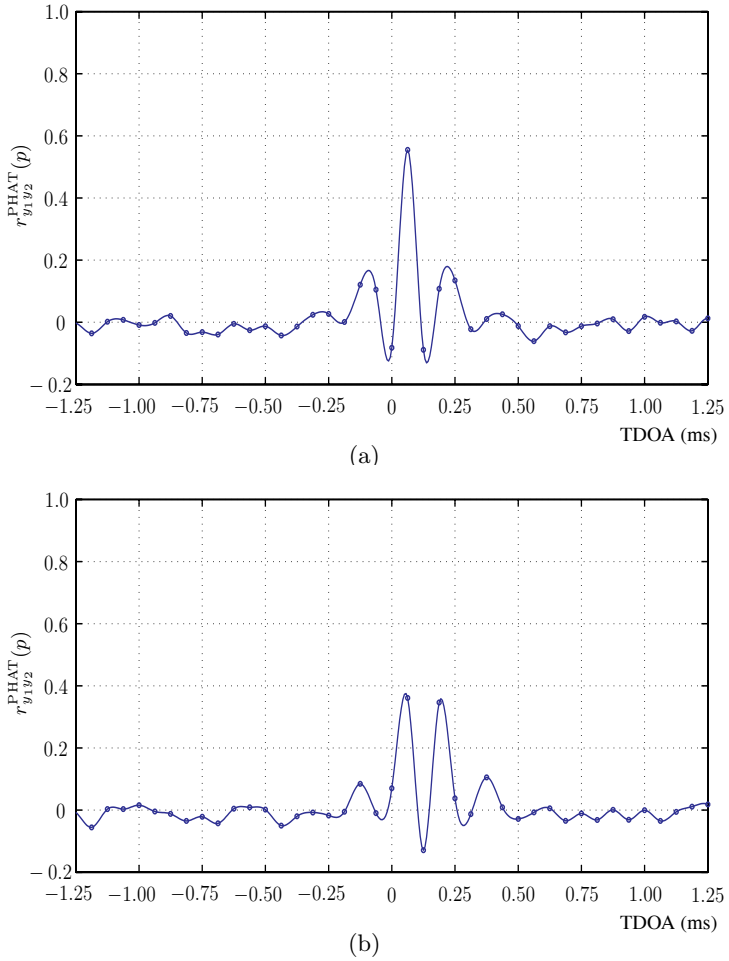

Fig. 9.8. The CCF computed using the PHAT algorithm: (a) there is only one source at $\theta = 75.5^\circ$ and (b) there are two source at $\theta_1 = 75.5^\circ$ and $\theta_2 = 41.4^\circ$ respectively. The microphone noise is white Gaussian with SNR = 10 dB. The sampling frequency is 16 kHz.

that now we have two sources in the far-field and their incident angles are $\theta_1 = 75.5^\circ$ and $\theta_2 = 41.4^\circ$ respectively. Figure 9.8 plots the CCF computed using the PHAT algorithm. We see from Fig. 9.8(b) that there are two large peaks corresponding to the two true TDOAs. However, if we compare Figs. 9.8(b) and (a), one can see that the peaks in the two-source situation are not defined as well as the peak for the single-source scenario. This result should not come as a surprise. From (9.120), we see that the two correlation functions $r^{\mathrm{CC}}_{s_1 s_1}(p - \tau_1)$ and $r^{\mathrm{CC}}_{s_2 s_2}(p - \tau_2)$ interfere with each other. So, one source will behave like noise to the other source, thereby making the TDOA estimation more difficult.

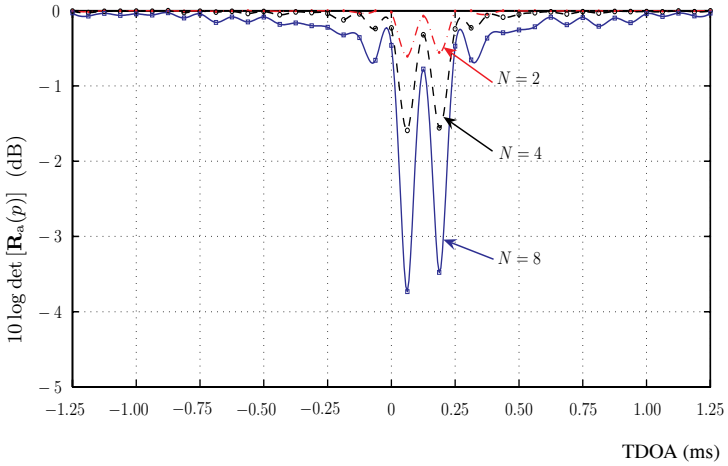

Fig. 9.9. Comparison of det $[\mathbf{R}_a(p)]$ for an equispaced linear array with different number of microphones. There are two source at $\theta_1 = 75.5°$ and $\theta_2 = 41.4°$ respectively. The microphone noise is white Gaussian with SNR = 10 dB. The sampling frequency is 16 kHz.

This problem will become worse as the number of sources in the array's field increases.

Similar to the single-source situation, we can improve the TDOA estimation of multiple sources by increasing the number of microphones. Figure 9.9 plots the cost function computed from the MCCC method with different number of microphones. It is seen that the estimation performance improves with the number of sensors.

The GCC, spatial prediction, MCCC, and entropy based techniques can be directly used to estimate TDOA for multiple sources. The extension of the narrowband MUSIC to the multiple-source situation is also straightforward. Consider the signal model in (9.8) where we neglect the attenuation difference, we have

$$Y_n(f) = \sum_{m=1}^{M} S_m(f) e^{-j2\pi[t_m + \mathcal{F}_n(\tau_m)]f} + V_n(f). \tag{9.121}$$

Following the notation used in Section 9.7.1, we can write $\vec{\mathbf{y}}$ as

$$\vec{\mathbf{y}} = \boldsymbol{\Omega}\vec{\mathbf{s}} + \vec{\mathbf{v}}, \tag{9.122}$$

where

$$\boldsymbol{\Omega} = \left[\boldsymbol{\varsigma}(\tau_1)\ \boldsymbol{\varsigma}(\tau_2) \cdots \boldsymbol{\varsigma}(\tau_M) \right],$$

is a matrix of size $N \times M$, and

$$\vec{\mathbf{s}} = \left[S_1(f)e^{j2\pi t_1 f} \; S_2(f)e^{j2\pi t_2 f} \; \cdots \; S_M(f)e^{j2\pi t_M f} \right]^T.$$

The covariance matrix $\mathbf{R}_Y$ has the form

$$\mathbf{R}_Y = E\left(\vec{\mathbf{y}}\vec{\mathbf{y}}^H\right) = \mathbf{\Omega}\mathbf{R}_S\mathbf{\Omega}^H + \sigma_V^2\mathbf{I}, \tag{9.123}$$

where

$$\mathbf{R}_S = E\left(\vec{\mathbf{s}}\vec{\mathbf{s}}^H\right). \tag{9.124}$$

It is easily seen that the rank of the product matrix $\mathbf{\Omega}\mathbf{R}_S\mathbf{\Omega}^H$ is of M. Therefore, if we perform the eigenvalue decomposition of $\mathbf{R}_Y$ and sort its eigenvalues in descending order, we get

$$\mathbf{\Omega}\mathbf{R}_S\mathbf{\Omega}^H\mathbf{b}_n = \mathbf{0}, \quad n = M+1, \ldots, N, \tag{9.125}$$

where, again, $\mathbf{b}_n$ is the eigenvector associated with the nth eigenvalue of $\mathbf{R}_Y$. This result indicates that

$$\mathbf{b}_n^H\boldsymbol{\varsigma}(\tau_m) = 0, \quad m = 1, 2, \ldots, M, \quad M+1 \leq n \leq N. \tag{9.126}$$

Following the same line of analysis in Section 9.7.1, after the eigenvalue decomposition of $\mathbf{R}_Y$, we can construct the narrowband MUSIC cost function as

$$J_{\text{MUSIC}}(p) = \frac{1}{\sum\limits_{n=M+1}^{N} \left|\mathbf{b}_n^H\boldsymbol{\varsigma}(p)\right|^2}. \tag{9.127}$$

The M largest peaks of $J_{\text{MUSIC}}(p)$ should correspond to the TDOAs τ_m, $m = 1, 2, \ldots, M$.

As we have pointed out earlier, the narrowband MUSIC may not be very useful for microphone array due to the nonstationary nature of speech. The extension of the broadband MUSIC (presented in Section 9.7.2) to multiple-source situation, however, is not straightforward. To see this, let us assume that there are M sources. With some mathematical manipulation, the spatial covariance matrix $\mathbf{R}_{\text{a}}(p)$ can be written as

$$\mathbf{R}_{\text{a}}(p) = \sum_{m=1}^{M} \mathbf{R}_{s_m}(p) + \sigma^2\mathbf{I}. \tag{9.128}$$

Now even when $p = \tau_m$ and $\mathbf{R}_{s_m}(p)$ becomes a matrix of rank 1, the superimposed signal matrix, $\sum_{m=1}^{M} \mathbf{R}_{s_m}(p)$, may still be of rank N. Therefore, the signal and noise subspaces are overlapped and we cannot form a broadband MUSIC algorithm for multiple sources. But in one particular case where all the sources are white, we can still use the estimator in (9.89). In general, for

multiple source TDOA estimation, we would recommend to use the MCCC approach.

Another possible approach for TDOA estimation of multiple sources is to blindly identify the impulse responses of a MIMO system. However, blind MIMO identification is much more difficult than blind SIMO identification, and might be even unsolvable. The research on this problem remains at the state of feasibility investigations. To finish this section, let us mention that recently some algorithms based on the MIMO model of (9.11) have been proposed in [157], [158].

9.12 Conclusions

This chapter presented the problem of DOA and TDOA estimation. We have chosen to focus exclusively on the principles of TDOA estimation since the problem of the DOA estimation is essentially the same as the TDOA estimation. We have discussed the basic idea of TDOA estimation based on the generalized cross-correlation criterion. In practice, the estimation problem can be seriously complicated by noise and reverberation. In order to improve the robustness of TDOA estimation with respect to distortions, we have discussed two basic approaches: exploiting the fact that we can have multiple microphones and using a more practical reverberant signal model, which resulted to a wide range of algorithms such as the spatial prediction, multichannel cross-correlation, minimum entropy, and adaptive blind channel identification techniques. Also discussed in this chapter were the principles for TDOA estimation of multiple sources.