Acoustics - Spatial properties

Emmanuel Vincent, Sharon Gannot, Tuomas Virtanen

3
Acoustics – Spatial properties

Emmanuel Vincent, Sharon Gannot, and Tuomas Virtanen

In Chapter 2, we presented the spectral properties of sound sources which can be exploited for the separation or enhancement of single-channel signals. In multichannel scenarios, the fact the acoustic scene is observed from different positions in space can also be exploited. In this chapter, we recall basic elements of acoustics and sound engineering, and use them to model multichannel mixtures.

We consider the relationship between a source signal and its spatial image in a given channel in Section 3.1, and examine how it translates in the case of microphone recordings or artificial mixtures in Sections 3.2 and 3.3, respectively. We then introduce several possible models in Section 3.4. We summarize the main concepts and provide links to other chapters and more advanced topics in Section 3.5.

3.1
Formalization of the mixing process

3.1.1
General mixing model

Sturmel *et al.* (2012) proposed the following general two-stage model for audio mixtures. In the first stage, each single-channel point source signal $s_j(t)$ is transformed into an $I \times 1$ multichannel source spatial image signal $\mathbf{c}_j(t)$ by means of a possibly nonlinear spatialization operation $\mathfrak{A}_j$:

$$\mathbf{c}_j(t) = [\mathfrak{A}_j(s_j)](t). \tag{3.1}$$

In the second stage, the source spatial image signals $\mathbf{c}_j(t)$, $j \in \{1 \ldots, J\}$, of all (point and diffuse) sources are added together and passed through a possibly nonlinear post-mixing operation $\mathfrak{A}$ to obtain the $I \times 1$ multichannel mixture signal $\mathbf{x}(t)$:

$$\mathbf{x}(t) = \left[\mathfrak{A} \left(\sum_{j=1}^{J} \mathbf{c}_j \right) \right] (t). \tag{3.2}$$

The linear, time-invariant case is of particular interest. In that case, the spatialization operations $\mathfrak{A}_j$ boil down to linear, time-invariant filters $\mathbf{a}_j(\tau) = [a_{1j}(\tau), \ldots, a_{Ij}(\tau)]^T$

$$\mathbf{c}_j(t) = \sum_{\tau=-\infty}^{+\infty} \mathbf{a}_j(\tau) s_j(t-\tau) \tag{3.3}$$

and the post-mixing operation $\mathfrak{A}$ reduces to identity[1)]

$$\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{c}_j(t). \tag{3.4}$$

The filters with coefficients $a_{ij}(\tau)$ are called *mixing filters* or *impulse responses*.

3.1.2 Microphone recordings vs. artificial mixtures

To investigate model (3.1)–(3.2) further, one must consider how the mixture was obtained in practice. Two different situations arise. *Microphone recordings* refer to the situation when multiple sources which are simultaneously active are captured by a microphone *array*. Typical examples include hands-free phones, audioconferencing systems, or hearing aids. *Artificial mixtures*, by contrast, are generated by mixing individually recorded sound sources using appropriate hardware or software. Most audio media (television, music, cinema…) fall into this category. Certain audio media such as classical music or documentaries result from a more complicated mixing process by which microphone recordings are first conducted and then artificially remixed in a studio. For more information about the recording and mixing strategies used by sound engineers, see Bartlett and Bartlett (2012).

3.2 Microphone recordings

3.2.1 Acoustic impulse responses

In the case of a microphone recording, the mixing process is due to the propagation of sound in the recording environment. This phenomenon is linear and time-invariant provided that the sources are static (not moving), so model (3.3)–(3.4) holds (Kuttruff, 2000). Each *acoustic impulse response* $a_{ij}(\tau)$ represents the propagation of sound from one source j to one microphone i and it is causal, i.e., $a_{ij}(\tau) = 0$ for $\tau < 0$.

1) Because of the linearity of summation, linear post-mixing, if any, is considered to be part of $\mathbf{a}_j(\tau)$.

In *free field*, that is in open air without any obstacle, sound propagation incurs a delay r_{ij}/c and an attenuation $1/\sqrt{4\pi}r_{ij}$ as a function of the distance r_{ij} from the source to the microphone. The acoustic impulse response is given by

$$a_{ij}(\tau) = \frac{1}{\sqrt{4\pi}r_{ij}}\delta\left(\tau - \frac{r_{ij}}{c}f_{\mathrm{s}}\right) \tag{3.5}$$

where c is the sound speed (343 m/s at 20°C), f_{s} the sampling rate, and δ the Dirac function or, more generally, a fractional delay filter. The attenuation due to distance directly affects the signal-to-noise ratio (SNR) (ISO, 2003).

In practice, various obstacles such as walls and furniture must be considered. The propagation of a sound of frequency ν changes depending on the size of the obstacle compared to its wavelength $\lambda = c/\nu$, which varies from $\lambda = 17$ mm at $\nu = 20$ kHz to $\lambda = 17$ m at $\nu = 20$ Hz. Obstacles which are substantially smaller than λ have little or no impact on the delay and attenuation. Obstacles of comparable dimension to λ result in *diffraction*: sound takes more time to pass the obstacle and it is more attenuated than in the free field. This phenomenon is famous for binaural recordings, i.e., recordings obtained from in-ear microphones, where the torso, head, and pinna of the listener act as obstacles (Blauert, 1997). It also explains source directivity, i.e., the fact that the sound emitted by a source varies with spatial direction. Finally, surfaces of dimension larger than λ result in *reflection* of the sound wave in the opposite direction with respect to the surface normal and absorption of part of its power. Many reflections typically occur on different obstacles, which induce multiple propagation paths. The acoustic impulse response between each source and each microphone results from the summation of all those paths.

Figure 3.1 provides a schematic illustration of the shape of an acoustic impulse response and Fig. 3.2 shows a real acoustic impulse response. The real response differs from the illustration as it exhibits both positive and negative values, but its magnitude follows the same overall shape. Three parts can be seen. The first peak is the line of sight, called *direct path* (3.5). It is followed by a few disjoint reflections on the closest obstacles called *early echoes*. Many reflections then simultaneously occur and form an exponentially decreasing tail called *late reverberation* or simply *reverberation*. The boundary τ_{c} between early echoes and reverberation, called *mixing time*, depends on the acoustic properties of the room. A typical value is 50 ms after the direct path. One can then decompose each acoustic impulse response $\mathbf{a}_j(\tau)$ into the sum of a direct part $\mathbf{a}_j^{\mathrm{dir}}(\tau)$ and an indirect part due to echoes and reverberation $\mathbf{a}_j^{\mathrm{rev}}(\tau)$. Similarly, each source spatial image can be decomposed as $\mathbf{c}_j(t) = \mathbf{c}_j^{\mathrm{dir}}(t) + \mathbf{c}_j^{\mathrm{rev}}(t)$. Note that early echoes are sometimes considered as part of $\mathbf{c}_j^{\mathrm{dir}}(t)$ instead of $\mathbf{c}_j^{\mathrm{rev}}(t)$ when defining the task of dereverberation (see Chapter 15).

3.2.2 Main properties of acoustic impulse responses

Acoustic impulse responses manifest themselves by a modification of the phase and power spectrum of the emitted signal and they smear the signal across time. Although they typically have thousands of coefficients, they can be described by three main

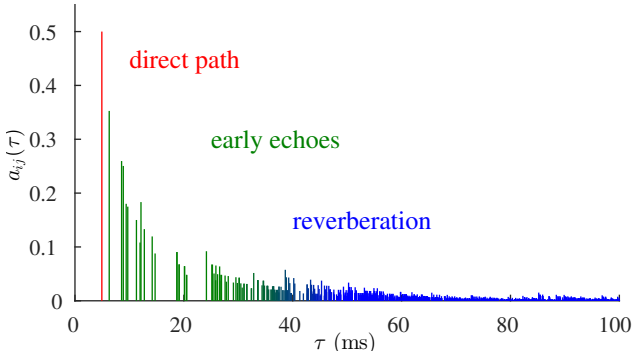

Figure 3.1 Schematic illustration of the shape of an acoustic impulse response $a_{ij}(\tau)$ for a room of dimensions 8.00 $\times$ 5.00 $\times$ 3.10 m, a RT60 of 230 ms, and a source distance of $r_{ij} = 1.70$ m. All reflections are depicted as Dirac impulses.

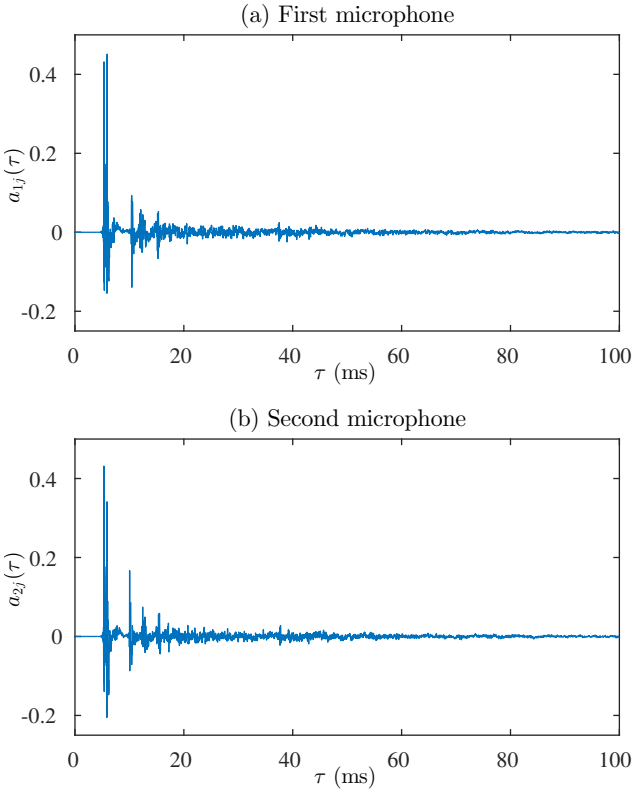

Figure 3.2 First 100 ms of a pair of real acoustic impulse responses $\mathbf{a}_j(\tau)$ from the Aachen Impulse Response Database (Jeub *et al.*, 2009) recorded in a meeting room with the same characteristics as in Fig. 3.1 and sampled at 48 kHz.

Table 3.1 Range of RT60 reported in the literature for different environments (Ribas *et al.*, 2016).

Environment		RT60 (s)
Car		0.05
Work	Office	0.25–0.43
	Meeting room	0.23–0.70
Home	Living room	0.44–0.74
	Bedroom	0.39–0.68
	Bathroom	0.41–0.75
	Kitchen	0.41–0.83
Public spaces	Classroom	0.20–1.27
	Lecture room	0.64–1.25
	Restaurant	0.50–1.50

properties. The *reverberation time* (RT60) is the duration over which the envelope of the reverberant tail decays by 60 decibels (dB). It depends on the size and the absorption level of the room (including obstacles) and it represents the time scale of smearing. Table 3.1 reports typical RT60 values for various environments. The *direct-to-reverberant ratio* (DRR) is ratio of the power of direct and indirect sound. It varies with the size and the absorption of the room, but also with the distance between the source and the microphone according to the curves in Fig. 3.3. It governs the amount of smearing of the signal. The distance beyond which the power of indirect sound becomes larger than that of direct sound is called *critical distance*. Finally, the *direct-to-early ratio*, that is the power of direct sound divided by the remaining power in the first τ_c samples, quantifies the modification of the power spectrum of the signal induced by early echoes. It is low when the microphone and/or the source is close to an obstacle such as a table or a window, and higher otherwise. The later two properties are not systematically reported in the literature, yet all three properties are important to characterize multichannel mixtures. Also, as we shall see, the RT60 values in Table 3.1 are larger than usually considered in the literature until recently.

Another useful property of acoustic impulse responses is the statistical dependency between impulse responses corresponding to the same source. Due to the summation of many propagation paths, reverberation can be statistically modeled using the law of large numbers as a zero-mean Gaussian noise signal with decaying amplitude (Polack, 1993). This Gaussian noise signal is characterized by its normalized correlation called *interchannel coherence* (IC). On average over all possible absolute positions of the sources and the microphone array in the room, the IC between two channels i and i' has the following closed-form expression (Kuttruff, 2000; Gustafsson *et al.*, 2003):

$$\omega_{ii'}(f) = \frac{\mathbb{E}\{c_{ij}^{\text{rev}}(\cdot, f) c_{i'j}^{\text{rev}\,*}(\cdot, f)\}}{\sqrt{\mathbb{E}\{|c_{ij}^{\text{rev}}(\cdot, f)|^2\}}\sqrt{\mathbb{E}\{|c_{i'j}^{\text{rev}}(\cdot, f)|^2\}}} = \frac{\sin(2\pi\nu_f \ell_{ii'}/c)}{2\pi\nu_f \ell_{ii'}/c} \tag{3.6}$$

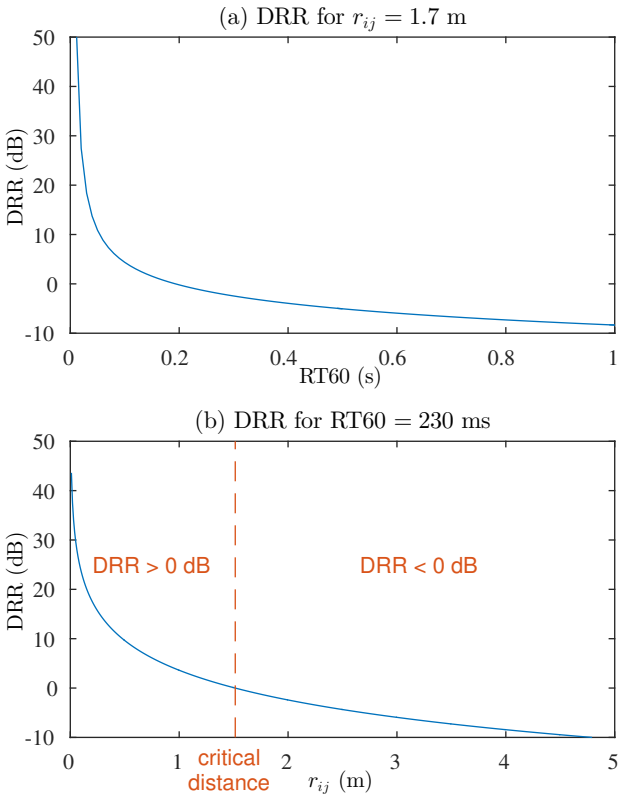

Figure 3.3 DRR as a function of the RT60 and the source distance r_{ij} based on Eyring's formula (Gustafsson *et al.*, 2003). These curves assume that there is no obstacle between the source and the microphone, so that the direct path exists. The room dimensions are the same as in Fig. 3.1.

where $\ell_{ii'}$ denotes the distance between the microphones, ν_f the center frequency of frequency bin f, and the expectation operator is taken over all directions of space. These scalar ICs can be grouped into an $I \times I$ coherence matrix $\mathbf{\Omega}(f) = [\omega_{ii'}(f)]_{ii'}$. Interestingly, the IC does not depend on the source nor on the room: it is large for small arrays and low frequencies and it decreases with microphone distance and frequency, as shown in Fig. 3.4. This result holds not only on average, but also in any practical setup provided that the RT60 is large enough so that the reverberant sound field is approximately diffuse, that is for all environments listed in Table 3.1 except cars.

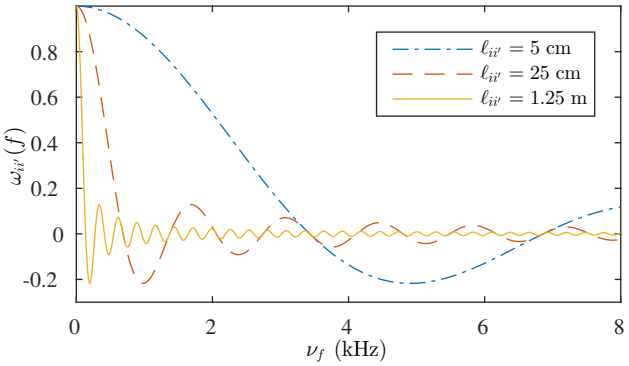

Figure 3.4 IC $\omega_{ii'}(f)$ of the reverberant part of an acoustic impulse response as a function of microphone distance $\ell_{ii'}$ and frequency ν_f.

3.3 Artificial mixtures

In the case of artificial mixtures, mixing is typically performed in four steps (Sturmel *et al.*, 2012). In the first step, the sound engineer applies a series of effects to each source. In the second step, the source is transformed into a multichannel spatial image $\mathbf{c}_j(t)$. In the third step, the spatial images of all sources are summed to obtain the so-called "master". In the last step, additional effects which depend on the distribution medium are applied to the master to provide the mixture $\mathbf{x}(t)$ known as "artistic mix" or "commercial mix". Steps 1 and 2 and steps 3 and 4 are formalized in equations (3.1) and (3.2), respectively. The overall mixing process then results from the effects chosen by the sound engineer. Example effects are listed in Table 3.2.

The inversion of nonlinear effects has been sparsely studied and shown to be difficult even when the nonlinearity is known (Gorlow and Reiss, 2013). For this reason, it is desirable to express the mixing process in linear form. It turns out that this is feasible under two conditions. First, the effects used to transform the source signal into its spatial image in step 2 must be linear. This condition often holds since panning or convolution by simulated or real reverberant impulse responses are typically used in this step and they are linear. The nonlinear effects possibly applied in step 1 can then be considered as part of the original source signal. Second, the nonlinear effects applied to the master in step 4 must be amenable to time-varying linear filtering. This condition generally holds too since dynamic compression and equalization are often the only effects applied at this stage and they can be modeled as a linear time-varying filter whose coefficients depend on the master signal. This time-varying filter may then be equivalently be applied to all source images before summation. The mixture then becomes equal to the sum of the source images as in (3.4) and each source image can be expressed similarly to (3.3), except that the mixing filters are time-varying. If convolution by reverberant impulse responses is used in step 2 and the amount of nonlinearity in step 4 is limited, the mixing filters share similar characteristics with

Table 3.2 Example artificial mixing effects, from Sturmel *et al.* (2012).

Linear instantaneous effects	Gain Panning (instantaneous mixing)
Linear convolutive effects	Equalization Reverberation Delay
Nonlinear effects	Dynamic compression Chorus Distortion

the acoustic impulse responses reviewed above.

3.4 Impulse response models

Given the physical properties of mixing filters described above, we can now build models for multichannel source separation and enhancement. Throughout the rest of this book, we assume linear mixing and static sources. The additional issues raised by moving sources or time-varying mixing are discussed in Chapter 19.

Time-domain modeling of the mixing filters as finite impulse response (FIR) filters of a few thousand coefficients was popular in the early stages of research (Nguyen Thi and Jutten, 1995; Ehlers and Schuster, 1997; Gupta and Douglas, 2007) and has gained new interest recently with sparse decomposition-based approaches (Lin *et al.*, 2007; Benichoux *et al.*, 2014; Koldovský *et al.*, 2015). However, the large number of coefficients to be estimated and the integration with time-frequency domain models for the sources result in costly algorithms (Kowalski *et al.*, 2010).

Most methods today model both the sources and the mixing filters in the time-frequency domain. Exact modeling using the theoretical tools in Section 2.3.1 is feasible but uncommon and it is discussed in Chapter 19. In the following, we present three approximate models which can be applied both to microphone recordings and artificial mixtures. For each model, we also explain how the parameters may be constrained in the specific case of microphone recordings. Similar constraints may be designed for artificial mixtures.

3.4.1 Narrowband approximation

3.4.1.1 Definition

Let us denote by $\mathbf{c}_j(n,f)$ and $s_j(n,f)$ the short-time Fourier transform (STFT) of $\mathbf{c}_j(t)$ and $s_j(t)$, respectively. The most common model is based on the narrowband approximation. Under the conditions discussed in Section 2.3.2, time-domain filtering can be approximated by complex-valued multiplication in the STFT domain:

$$\mathbf{c}_j(n,f) = \mathbf{a}_j(f)\, s_j(n,f) \tag{3.7}$$

where the $I \times 1$ vector $\mathbf{a}_j(f)$ is called *mixing vector*. Each element $a_{ij}(f)$ of the mixing vector is the discrete Fourier transform (DFT) associated with $a_{ij}(\tau)$ called *transfer function* or *acoustic transfer function*. The mixing vectors of all sources are sometimes concatenated into an $I \times J$ matrix $\mathbf{A}_j(f) = [\mathbf{a}_1(f), \ldots, \mathbf{a}_J(f)]$ called the *mixing matrix*.

3.4.1.2 Steering vector – Near field vs. far field

When the source position is known, geometrical (soft or hard) constraints can be set on $\mathbf{a}_j(f)$ to ensure that it is close to the *steering vector* $\mathbf{d}_j(f)$ which encodes the direct path (Parra and Alvino, 2002; Knaak *et al.*, 2007). In the case of a microphone recording, the steering vector for source j is given by

$$\mathbf{d}_j(f) = \begin{bmatrix} \frac{1}{\sqrt{4\pi} r_{1j}} e^{-2\jmath\pi r_{1j}\nu_f/c} \\ \vdots \\ \frac{1}{\sqrt{4\pi} r_{Ij}} e^{-2\jmath\pi r_{Ij}\nu_f/c} \end{bmatrix} \tag{3.8}$$

where each element is the DFT of the free-field acoustic impulse response (3.5) from the source to microphone i. This expression is mainly applied in the *near field*, that is when the source-to-microphone distances r_{ij} are smaller or comparable to the microphone distances $\ell_{ii'}$. In the *far field*, the attenuation factors $1/\sqrt{4\pi} r_{ij}$ become approximately equal so the following expression of the steering vector (up to a multiplicative factor) is used instead:

$$\mathbf{d}_j(f) = \begin{bmatrix} e^{-2\jmath\pi r_{1j}\nu_f/c} \\ \vdots \\ e^{-2\jmath\pi r_{Ij}\nu_f/c} \end{bmatrix}. \tag{3.9}$$

Note that, in either case, the steering vector depends both on the *direction of arrival* (DOA) and the distance of the source relative to the array.

3.4.2 Relative transfer function and interchannel cues

3.4.2.1 Definition

The transfer functions $a_{ij}(f)$ have a specific phase and amplitude for each channel i. In theory, this could be exploited to perform source localization and separation even in a single-channel or monaural setting (Blauert, 1997; Asari *et al.*, 2006) by disambiguating $a_{ij}(f)$ from $s_j(n,f)$ in 3.7. In practice, however, the phase spectrum of the source is unknown and its magnitude spectrum is rarely known to the required level of precision[2)]. This has motivated researchers to disregard monaural cues and

2) Contrary to a widespread belief, human audition relies more on head movements than monaural cues to solve ambiguous spatial percepts (Wallach, 1940; Wightman and Kistler, 1999).

consider the differences between channels instead.

Taking the first channel as a reference, the *relative mixing vector* for source j is defined as (Gannot *et al.*, 2001; Markovich *et al.*, 2009)

$$\tilde{\mathbf{a}}_j(f) = \frac{1}{a_{1j}(f)}\mathbf{a}_j(f). \tag{3.10}$$

The elements $\tilde{a}_{ij}(f)$ of this vector are called *relative transfer functions* (RTFs). They can be interpreted as transfer functions relating the channels of the source spatial image to each other. Note that $\tilde{\mathbf{a}}_j(f)$ is defined only when $a_{1j}(f) \neq 0$, which is sometimes not true in low DRR conditions. An alternative definition was given by Affes and Grenier (1997) and Sawada *et al.* (2007):

$$\bar{\mathbf{a}}_j(f) = \frac{e^{-\jmath\angle a_{1j}(f)}}{\|\mathbf{a}_j(f)\|_2}\mathbf{a}_j(f). \tag{3.11}$$

By taking all channels into account, this definition increases the chance that the relative mixing vector is defined and it makes it more invariant to the magnitude of the reference channel. For generalizations of this concept, see Li *et al.* (2015).

The RTFs encode the *interchannel level difference* (ILD), also known as the *interchannel intensity difference*, in decibels and the *interchannel phase difference* (IPD) in radians between pairs of microphones as a function of frequency:

$$\mathrm{ILD}_{ij}(f) = 20\log_{10}|\tilde{a}_{ij}(f)| \tag{3.12}$$

$$\mathrm{IPD}_{ij}(f) = \angle\tilde{a}_{ij}(f). \tag{3.13}$$

Figure 3.5 illustrates these two quantities as a function of frequency. The ILD and the IPD appear to cluster around the ILD and the IPD associated with the direct path, but they can exhibit significant deviations due to early echoes and reverberation.

The *interchannel time difference* (ITD) in seconds is sometimes considered instead of the IPD:

$$\mathrm{ITD}_{ij}(f) = \frac{\angle\tilde{a}_{ij}(f)}{2\pi\nu_f}. \tag{3.14}$$

Note however that the ITD is unambiguously defined only below the frequency c/ℓ_{i1}. With a sampling rate of 16 kHz, this requires a microphone distance ℓ_{i1} of less than 4.3 cm. For larger distances or higher frequencies, *spatial aliasing* occurs: several candidate ITDs correspond to a given IPD up to an integer multiple of 2π, therefore the ITD can be measured only up to an integer multiple of $1/\nu_f$. In a binaural setting, the ILD spans a large range and it can be exploited to disambiguate multiple ITDs corresponding to the same IPD. With free-field microphone arrays, this is hardly feasible as the ILD is smaller in the far field and varies a lot more with reverberation. One must then integrate the IPD information across frequency to recover the ITD.

These interchannel quantities play a key role in human hearing and, consequently, in hearing aids. For more details, see Chapter 18.

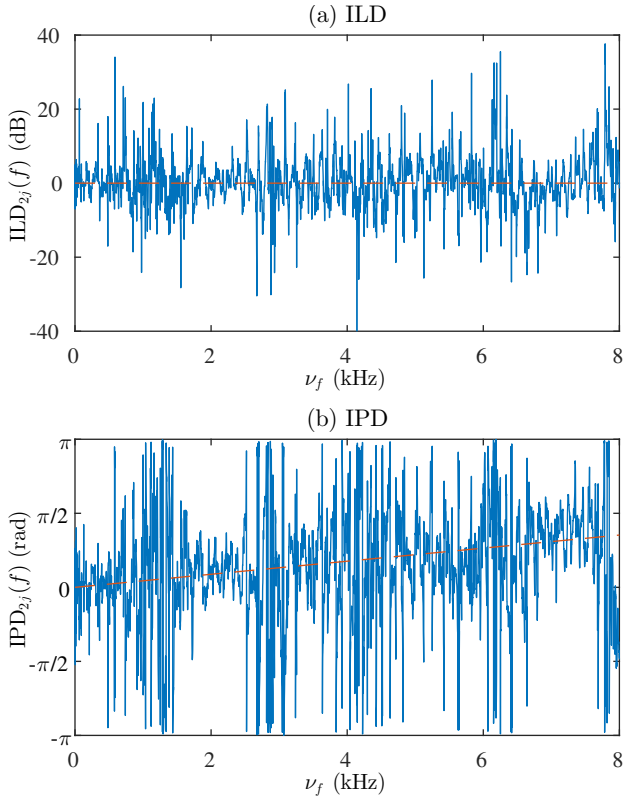

Figure 3.5 ILD and IPD corresponding to the pair of real acoustic impulse responses in Fig. 3.2. Dashed lines denote the theoretical ILD and IPD in the free field, as defined by the relative steering vector $\tilde{\mathbf{d}}_j(f)$.

3.4.2.2 **Relative steering vector**

Similarly to Section 3.4.1.2, geometrical constraints can be set on $\tilde{\mathbf{a}}_j(f)$ to ensure that it is close to the *relative steering vector* $\tilde{\mathbf{d}}_j(f) = \mathbf{d}_j(f)/d_{1j}(f)$ (Yılmaz and Rickard, 2004; Sawada *et al.*, 2007; Reindl *et al.*, 2013), as observed in Fig. 3.5. In the far field, the relative steering vector for source j is given by

$$\tilde{\mathbf{d}}_j(f) = \begin{bmatrix} 1 \\ e^{-2\jmath\pi\Delta_{2j}\nu_f} \\ \vdots \\ e^{-2\jmath\pi\Delta_{Ij}\nu_f} \end{bmatrix} \tag{3.15}$$

where

$$\Delta_{ij} = \frac{r_{ij} - r_{1j}}{c} \tag{3.16}$$

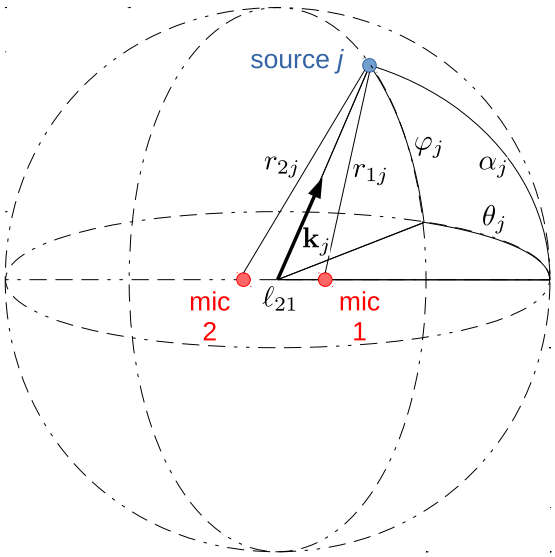

Figure 3.6 Geometrical illustration of the position of a far-field source j with respect to a pair of microphones on the horizontal plane, showing the azimuth θ_j, the elevation φ_j, the angle of arrival α_j, the microphone distance ℓ_{21}, the source-to-microphone distances r_{1j} and r_{2j}, and the unit-norm vector $\mathbf{k}_j$ pointing to the source.

is the ITD in the free field called *time difference of arrival* (TDOA). The TDOA in the far field depends only on the source DOA (not on its distance). More precisely, denoting by θ_j and φ_j the azimuth and elevation of source j with respect to the array center, as represented in Fig. 3.6, and by $\mathbf{k}_j = [\cos\theta_j \cos\varphi_j, \sin\theta_j \cos\varphi_j, \sin\varphi_j]^T$ the unit-norm vector pointing to the source, the TDOA can be computed as

$$\Delta_{ij} = \frac{\mathbf{k}_j^T(\mathbf{m}_i - \mathbf{m}_1)}{c} \tag{3.17}$$

with $\mathbf{m}_i$ the Cartesian coordinates of microphone i with respect to the array center. In the case when there are two microphones or all microphones are spatially aligned with each other, it can also be expressed as

$$\Delta_{ij} = \frac{\ell_{i1} \cos\alpha_j}{c} \tag{3.18}$$

with α_j the *angle of arrival* with respect to the microphone axis. The TDOA can also be defined in the near field according to (3.16), but its expression and that of the relative steering vector also depend on the source distance in that case.

3.4.3 Full-rank covariance model

3.4.3.1 Definition

We recall that the above models are valid only for point sources under the conditions in Section 2.3.2. For practical mixing filter lengths on the order of several hundred milliseconds and STFT analysis window lengths on the order of 50 ms, these conditions do not fully hold so the time-domain mixing process (3.3) is only roughly approximated by STFT-domain multiplication (3.7). One approach which partly overcomes this issue is to move from a linear (first-order) model to a second-order model of the signals.

Considering all signals of interest as wide-sense stationary processes within each time frame n, we denote by $\mathbf{\Sigma}_{\mathbf{c}_j}(n,f) = \mathbb{E}\{\mathbf{c}_j(n,f)\mathbf{c}_j^H(n,f)\}$ the $I \times I$ covariance matrix of $\mathbf{c}_j(n,f)$ across channels. Under the narrowband approximation, it can be shown that

$$\mathbf{\Sigma}_{\mathbf{c}_j}(n,f) = \sigma_{s_j}^2(n,f)\mathbf{R}_j(f) \tag{3.19}$$

where $\sigma_{s_j}^2(n,f) = \mathbb{E}\{|s_j(n,f)|^2\}$ is the variance of $s_j(n,f)$ and the $I \times I$ rank-1 matrix $\mathbf{R}_j(f) = \mathbf{a}_j(f)\mathbf{a}_j^H(f)$ is called the *spatial covariance matrix* (Févotte and Cardoso, 2005; Vincent *et al.*, 2009). The rank-1 property implies that the channels of $\mathbf{c}_j(n,f)$ are coherent, i.e., perfectly correlated.

Duong *et al.* (2010) and Sawada *et al.* (2013) proposed to consider the spatial covariance matrix $\mathbf{R}_j(f)$ as a *full-rank* matrix instead. This more flexible model applies to longer impulse responses and to diffuse sources. In such conditions, the sound emitted by each source reaches the microphones from several directions at once, such that the channels of $\mathbf{c}_j(n,f)$ become incoherent. The entries $(\mathbf{R}_j(f))_{ii'}$ of $\mathbf{R}_j(f)$ encode not only the ILD and the IPD, but also the IC[3)]

$$\mathrm{IC}_{ii'j}(f) = \frac{\mathbb{E}\{c_{ij}(\cdot,f)c_{i'j}^*(\cdot,f)\}}{\sqrt{\mathbb{E}\{|c_{ij}(\cdot,f)|^2\}}\sqrt{\mathbb{E}\{|c_{i'j}(\cdot,f)|^2\}}}. \tag{3.20}$$

Indeed, they can be expressed as $\mathrm{ILD}_{ij}(f) = 10\log_{10}(|(\mathbf{R}_j(f))_{ii}|/|(\mathbf{R}_j(f))_{11}|)$, $\mathrm{IPD}_{ij}(f) = \angle(\mathbf{R}_j(f))_{i1}$, and $\mathrm{IC}_{ii'j}(f) = (\mathbf{R}_j(f))_{ii'}/\sqrt{(\mathbf{R}_j(f))_{ii}}\sqrt{(\mathbf{R}_j(f))_{i'i'}}$. The quantity $|\mathrm{IC}_{ii'j}(f)|^2$ is referred to as the magnitude squared coherence (MSC).

3.4.3.2 Parametric covariance models

When the source position and the room characteristics are known, geometrical (soft or hard) constraints can be set on $\mathbf{R}_j(f)$. The average value of the spatial covariance matrix over all possible absolute positions of the sources and the microphone array in the room is equal to (Duong *et al.*, 2010)

$$\mathbf{D}_j(f) = \mathbf{d}_j(f)\mathbf{d}_j^H(f) + \sigma_{\mathrm{rev}}^2(f)\mathbf{\Omega}(f) \tag{3.21}$$

3) Note that the IC is defined for the full spatial image in (3.20) instead of the reverberant part only in (3.6).

with $\mathbf{d}_j(f)$ the steering vector in (3.8), $\mathbf{\Omega}(f)$ the covariance matrix of a diffuse sound field in (3.6), and $\sigma^2_{\text{rev}}(f)$ the power of early echoes and reverberation. The matrix $\mathbf{D}_j(f)$ generalizes the concept of steering vector to the second-order case. Duong *et al.* (2013) showed that, for moderate or large RT60, $\mathbf{R}_j(f)$ is close to $\mathbf{D}_j(f)$. Nikunen and Virtanen (2014) alternatively constrained $\mathbf{R}_j(f)$ as the weighted sum of rank-1 matrices of the form $\mathbf{d}_j(f)\mathbf{d}_j^H(f)$ uniformly spanning all possible incoming sound directions on the 3D sphere. Ito *et al.* (2015) proposed similar linear subspace constraints for diffuse sources.

3.5 Summary

In this chapter, we described the various types of mixtures encountered in audio and argued that, in most cases, they boil down to a linear mixing model. We examined the properties of impulse responses and reviewed the most common impulse response models.

These models are essentially used for multichannel separation and enhancement in Part III of the book. Specifically, the narrowband approximation and RTFs are used in Chapters 10, 11, 12, 13, and full-rank models in Chapter 14. For specific use of binaural properties, see Chapter 18. Advanced topics such as handling moving sources or microphones, convolution in the STFT domain, and learning the manifold of impulse responses are discussed in Chapter 19.

Bibliography

Affes, S. and Grenier, Y. (1997) A signal subspace tracking algorithm for microphone array processing of speech. *IEEE Transactions on Speech and Audio Processing*, **5** (5), 425–437.

Asari, H., Pearlmutter, B.A., and Zador, A.M. (2006) Sparse representations for the cocktail party problem. *The Journal of Neuroscience*, **26** (28), 7477–7490.

Bartlett, B. and Bartlett, J. (2012) *Practical Recording Techniques: the Step-by-step Approach to Professional Recording*, Focal Press, 6th edn..

Benichoux, A., Simon, L.S.R., Vincent, E., and Gribonval, R. (2014) Convex regularizations for the simultaneous recording of room impulse responses. *IEEE Transactions on Signal Processing*, **62** (8), 1976–1986.

Blauert, J. (1997) *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT Press.

Duong, N.Q.K., Vincent, E., and Gribonval, R. (2010) Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Transactions on Audio, Speech, and Language Processing*, **18** (7), 1830–1840.

Duong, N.Q.K., Vincent, E., and Gribonval, R. (2013) Spatial location priors for Gaussian model based reverberant audio source separation. *EURASIP Journal on Advances in Signal Processing*, **2013**, 149.

Ehlers, F. and Schuster, H.G. (1997) Blind separation of convolutive mixtures and an application in automatic speech recognition in a noisy environment. *IEEE Transactions on Signal Processing*, **45** (10), 2608–2612.

Févotte, C. and Cardoso, J.F. (2005) Maximum likelihood approach for blind audio source separation using time-frequency Gaussian models, in *Proceedings of IEEE Workshop on Applications of Signal Processing to*

Audio and Acoustics, pp. 78–81.

Gannot, S., Burshtein, D., and Weinstein, E. (2001) Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Transactions on Signal Processing*, **49** (8), 1614–1626.

Gorlow, S. and Reiss, J.D. (2013) Model-based inversion of dynamic range compression. *IEEE Transactions on Audio, Speech, and Language Processing*, **21** (7), 1434–1444.

Gupta, M. and Douglas, S.C. (2007) Beamforming initialization and data prewhitening in natural gradient convolutive blind source separation of speech mixtures, in *Proceedings of International Conference on Independent Component Analysis and Signal Separation*, pp. 462–470.

Gustafsson, T., Rao, B.D., and Trivedi, M. (2003) Source localization in reverberant environments: Modeling and statistical analysis. *IEEE Transactions on Speech and Audio Processing*, **11**, 791–803.

ISO (2003) ISO 9921. Ergonomics – Assessment of speech communication.

Ito, N., Vincent, E., Nakatani, T., Ono, N., Araki, S., and Sagayama, S. (2015) Blind suppression of nonstationary diffuse noise based on spatial covariance matrix decomposition. *Journal of Signal Processing Systems*, **79** (2), 145–157.

Jeub, M., Schäfer, M., and Vary, P. (2009) A binaural room impulse response database for the evaluation of dereverberation algorithms, in *Proceedings of IEEE International Conference on Digital Signal Processing*, pp. 1–4.

Knaak, M., Araki, S., and Makino, S. (2007) Geometrically constrained independent component analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, **15** (2), 715–726.

Koldovský, Z., Malek, J., and Gannot, S. (2015) Spatial source subtraction based on incomplete measurements of relative transfer function. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1335–1347.

Kowalski, M., Vincent, E., and Gribonval, R. (2010) Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, **18** (7), 1818–1829.

Kuttruff, H. (2000) *Room acoustics*, Taylor & Francis.

Li, X., Horaud, R., Girin, L., and Gannot, S. (2015) Local relative transfer function for sound source localization, in *Proceedings of European Signal Processing Conference*, pp. 399–403.

Lin, Y., Chen, J., Kim, Y., and Lee, D.D. (2007) Blind channel identification for speech dereverberation using $\ell 1$-norm sparse learning, in *Proceedings of Neural Information Processing Systems*, pp. 921–928.

Markovich, S., Gannot, S., and Cohen, I. (2009) Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, **17** (6), 1071–1086.

Nguyen Thi, H.L. and Jutten, C. (1995) Blind source separation for convolutive mixtures. *Signal Processing*, **45** (2), 209–229.

Nikunen, J. and Virtanen, T. (2014) Direction of arrival based spatial covariance model for blind sound source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **22** (3), 727–739.

Parra, L.C. and Alvino, C.V. (2002) Geometric source separation: Merging convolutive source separation with geometric beamforming. *IEEE Transactions on Speech and Audio Processing*, **10** (6), 352–362.

Polack, J.D. (1993) Playing billiards in the concert hall: The mathematical foundations of geometrical room acoustics. *Applied Acoustics*, **38** (2), 235–244.

Reindl, K., Zheng, Y., Schwarz, A., Meier, S., Maas, R., Sehr, A., and Kellermann, W. (2013) A stereophonic acoustic signal extraction scheme for noisy and reverberant environments. *Computer Speech and Language*, **27** (3), 726–745.

Ribas, D., Vincent, E., and Calvo, J.R. (2016) A study of speech distortion conditions in real scenarios for speaker recognition applications, in *Proceedings of IEEE Spoken Language Technology Workshop*, pp. 13–20.

Sawada, H., Araki, S., Mukai, R., and Makino, S. (2007) Grouping separated frequency components with estimating propagation model parameters in frequency-domain blind source separation. *IEEE Transactions*

on Audio, Speech, and Language Processing, **15** (5), 1592–1604.

Sawada, H., Kameoka, H., Araki, S., and Ueda, N. (2013) Multichannel extensions of non-negative matrix factorization with complex-valued data. *IEEE Transactions on Audio, Speech, and Language Processing*, **21** (5), 971–982.

Sturmel, N., Liutkus, A., Pinel, J., Girin, L., Marchand, S., Richard, G., Badeau, R., and Daudet, L. (2012) Linear mixing models for active listening of music productions in realistic studio conditions, in *Proceedings of the Audio Engineering Society Convention*. Paper 8594.

Vincent, E., Arberet, S., and Gribonval, R. (2009) Underdetermined instantaneous audio source separation via local Gaussian modeling, in *Proceedings of International Conference on Independent Component Analysis and Signal Separation*, pp. 775 – 782.

Wallach, H. (1940) The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology*, **27** (4), 339–368.

Wightman, F.L. and Kistler, D.J. (1999) Resolution of front-back ambiguity in spatial hearing by listener and source movement. *Journal of the Acoustical Society of America*, **105** (5), 2841–2853.

Yılmaz, Ö. and Rickard, S.T. (2004) Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, **52** (7), 1830–1847.