Konrad Kowalczyk, Oliver Thiergart, Maja Taseska,
Giovanni Del Galdo, Ville Pulkki, and Emanuël A.P. Habets

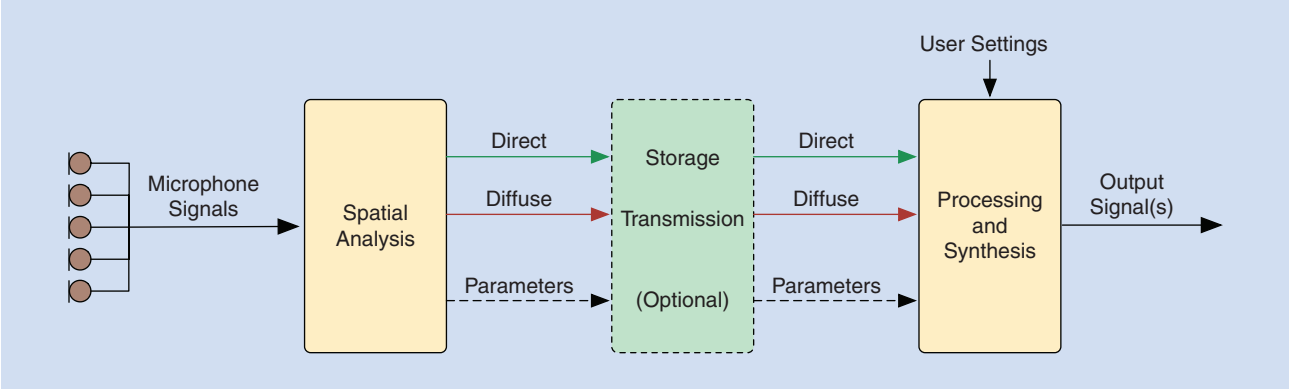
Parametric Spatial Sound Processing

A flexible and efficient solution to sound scene acquisition, modification, and reproduction

Flexible and efficient spatial sound acquisition and subsequent processing are of paramount importance in communication and assisted listening devices such as mobile phones, hearing aids, smart TVs, and emerging wearable devices (e.g., smart watches and glasses). In application scenarios where the number of sound sources quickly varies, sources move, and nonstationary noise and reverberation are commonly encountered, it remains a challenge to capture sounds in such a way that they can be reproduced with a high and invariable sound quality. In addition, the objective in terms of what needs to be captured, and how it should be reproduced, depends on the application and on the user's preferences. Parametric spatial sound processing has been around for two decades and provides a flexible and efficient solution to capture, code, and transmit, as well as manipulate and reproduce spatial sounds.

Instrumental to this type of processing is a parametric model that can describe a sound field in a compact and general way. In most cases, the sound field can be decomposed into a direct sound component and a diffuse sound component. These two components together with parametric side information such as the direction-of-arrival (DOA) of the direct sound component or the position of the sound source, provide a perceptually motivated description of the acoustic scene [1]–[3]. In this article, we provide an overview of recent advances in spatial sound capturing, manipulation, and reproduction based on such parametric descriptions of the sound field. In particular, we focus on two established parametric descriptions presented in a unified way and show how the signals and parameters can be obtained using multiple microphones. Once the sound field is analyzed, the sound scene can be transmitted, manipulated, and synthesized depending on the application. For example,

Digital Object Identifier 10.1109/MSP.2014.2369531
Date of publication: 12 February 2015



[FIG1] **A high-level overview of the parametric spatial sound processing scheme.**

sounds can be extracted from a specific direction or from a specific arbitrary two-dimensional or even three-dimensional region of interest. Furthermore, the sound scene can be manipulated to create an acoustic zoom effect in which direct sounds within the listening angular range are amplified depending on the zoom factor, while other sounds are suppressed. In addition, the signals and parameters can be used to create surround sound signals. As the manipulation and synthesis are highly application dependent, we focus in this article on three illustrative assisted listening applications: spatial audio communication, virtual classroom, and binaural hearing aids.

INTRODUCTION

Communication and assisted listening devices commonly use multiple microphones to create one or more signals, the content of which highly depends on the application. For example, when smart glasses are used to record a video, the microphones can be used to create a surround sound recording that consists of multiple audio signals. A compact yet accurate representation of the sound field at the recording position makes it possible to render the sound field on an arbitrary reproduction setup in a different location. On the other hand, when the device is used in hands-free or speech recognition mode, the microphones can be used to extract the user's speech while reducing background noise and interfering sounds. In the last few decades, sophisticated solutions for these applications were developed.

Spatial recordings are commonly made using specific microphone setups. For instance, there are several stereo recording techniques in which different positioning of the microphones of the same or different types (e.g., cardioid or omnidirectional microphones) is exploited to make a stereo recording that can be reproduced using loudspeakers. When more loudspeakers are available for spatial sound rendering, the microphone recordings are often specifically mixed for a given reproduction setup. These classical techniques do not provide the flexibility required in many modern applications where the reproduction setup is not known in advance. Signal enhancement, on the other hand, is commonly achieved by filtering, and subsequently summing the available microphone signals. Classical spatial filters often require information on the second-order statistics (SOS) of the desired and undesired signals (cf. [4] and [5]). For real-time applications, the SOS need to be estimated online, and the quality of the output signal highly depends on the accuracy of these estimates. To date, major challenges remain, such as:

1) achieving a sufficiently fast response to changes in the sound scene (such as moving and emerging sources) and to changes in the acoustic conditions
2) providing sufficient flexibility in terms of spatial selectivity
3) ensuring a high-quality output signal at all times
4) providing solutions with a manageable computational complexity.

Although the use of multiple microphones provides, at least in theory, a major advantage over a single microphone, the adoption of multimicrophone techniques in practical systems has not been particularly popular until very recently. Possible reasons for this could be that in real-life scenarios, these techniques provided insufficient improvement over single-microphone techniques, while significantly increasing the computational complexity, the system calibration effort, and the manufacturing costs. In the last few years, the smartphone and hearing aid industries made a significant step forward in using multiple microphones, which has recently become a standard for these devices.

Parametric spatial sound processing provides a unified solution to both the spatial recording and signal enhancement problems, as well as to other challenging sound processing tasks such as adding virtual sound sources to the sound scene. As illustrated in Figure 1, the parametric processing is performed in two successive steps that can be completed on the same device or on different devices. In the first step, the sound field is analyzed in narrow frequency bands using multiple microphones to obtain a compact and perceptually meaningful description of the sound field in terms of direct and diffuse sound components and some parametric information (e.g., DOAs and positions). In the second step, the input signals and possibly the parameters are modified, and one or more output signals are synthesized. The modification and synthesis can be user, application, or scenario dependent. Parametric spatial sound processing is also common in audio coding (cf. [6]) where parametric information is extracted directly from the loudspeaker channels instead of the microphone signals.

The described scheme also allows for an efficient transmission of sound scenes to the far-end side [1], [7] for loudspeaker

reproduction with arbitrary setups or for binaural reproduction [8]. Hence, instead of transmitting many microphone signals and carrying out the entire processing at the receiving side, only two signals (i.e., the direct and diffuse signals) need to be transmitted together with the parametric information. These two signals enable synthesis of the output signals on the receiving side for the reproduction system at hand, and additionally allow the listener to arbitrarily adjust the spatial responses. Note that in the considered approach, the same audio and parametric side information is sent, irrespective of the number of loudspeakers used for reproduction.

As an alternative to the classical filters used for signal enhancement, where an enhanced signal is created as a weighted sum of the available microphone signals, an enhanced signal can be created by using the direct and diffuse sound components and the parametric information. This approach can be seen as a generalization of the parametric filters used in [9]–[12] where the filters are calculated based on instantaneous estimates of an underlying parametric sound field model. As will be discussed later in this article, these parameters are typically estimated in narrow frequency bands, and their accuracy depends on the resolution of the time-frequency transform and the geometry of the microphone array. If accurate parameter estimates with a sufficiently high time-frequency resolution are available, parametric filters can quickly adapt to changes in the acoustic scene. The parametric filters have been applied to various challenging acoustic signal processing problems related to assisted listening, such as directional filtering [10], dereverberation [11], and acoustic zooming [13]. Parametric filtering approaches have been used also in the context of binaural hearing aids [14], [15].

PARAMETRIC SOUND FIELD MODELS

BACKGROUND

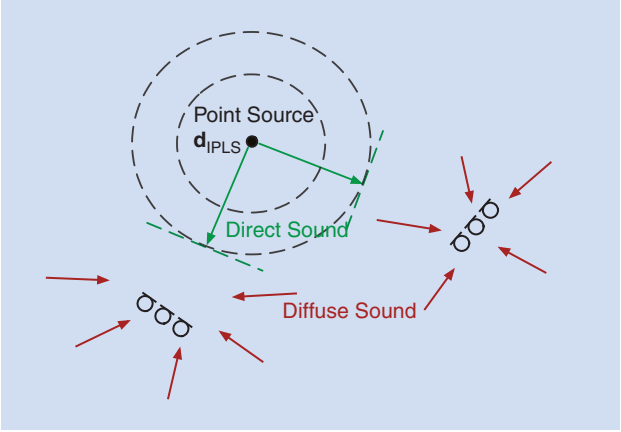
Many parametric models have originally been developed with the aim to subsequently capture, transmit, and reproduce high-quality spatial audio; examples include directional audio coding (DirAC) [1], microphone front ends for spatial audio coders [16], and high angular resolution plane wave expansion (HARPEX) [17]. These models were developed based on observations about the human perception of spatial sound, aiming to recreate perceptually important spatial audio attributes for the listener. For example, in the basic form of DirAC [1], the model parameters are the DOA of the direct sound and the diffuseness that is directly related to the power ratio between the direct signal power and the diffuse signal power. Using a pressure signal and this parametric information, a direct signal and a diffuse signal could be reconstructed at the far-end side. The direct signal is attributed to a single plane wave at each frequency, and the diffuse signal is attributed to spatially extended sound sources, concurrent sound sources (e.g., applause from an audience or cafeteria noise), and room reverberation that occurs due to multipath acoustic wave propagation when sound is captured in an enclosed environment. A similar sound field model that consists of the direct and diffuse sound has been applied in spatial audio scene coding (SASC) [2] and in [3] for sound reproduction with arbitrary reproduction systems and for sound scene manipulations. On the other hand, in [16] the model parameters include the interchannel level difference and the interchannel coherence [18] that were estimated using two microphones and were previously used in various spatial audio coders [6]. These model parameters are sent to the far-end side together with a so-called downmix signal to generate multiple loudspeaker channels for sound reproduction. In this case, the downmix signal and parameters are compatible with those used in different spatial audio coders. In contrast to DirAC and SASC, HARPEX assumes that the direct signal at a particular frequency is composed only of two plane waves.

Besides offering a compact and flexible way to transmit and reproduce high-quality spatial audio, independent of the reproduction setup, parametric processing is highly attractive for sound scene manipulations and signal enhancement. The extracted model parameters can be used to compute parametric filters that can, for instance, achieve directional filtering [10] and dereverberation [11]. The parametric filters represent spectral gains applied to a reference microphone signal, and can in principle provide arbitrary directivity patterns that can adapt quickly to the acoustic scene provided that the sound field analysis is performed with a sufficiently high time-frequency resolution. For this purpose, the short-time Fourier transform (STFT) is considered a good choice as it often offers a sufficiently sparse signal representation to assume a single dominant directional wave in each time-frequency bin. For instance, the assumption that the source spectra are sufficiently sparse is commonly made in speech signal processing [19]. The sources that exhibit sufficiently small spectrotemporal overlap fulfill the so-called W-disjoint orthogonality condition. This assumption is, however, violated when concurrent sound sources with comparable powers are active in one frequency band. Another family of parametric approaches emerged within the area of computational auditory scene analysis [20], where the auditory cues are utilized for instance to derive time-frequency masks that can be used to separate different source signals from the captured sound.

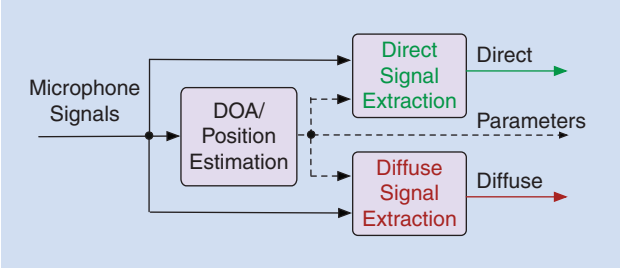
Clearly, the choice of an underlying parametric model depends on the specific application and on the way the extracted parameters and the available audio signals are used to generate the desired output. In this article, we focus on geometry-based parametric models that take into account both direct and diffuse sound components, allowing for high-quality spatial sound acquisition, which can be subsequently used both for transmission and reproduction purposes, as well as to derive flexible parametric filters for sound scene manipulation and signal enhancement for assisted listening.

GEOMETRIC MODELS

In the following, we consider the time-frequency domain with k and n denoting the frequency and time indices, respectively. For each (k, n), we assume that the sound field is a superposition of a single spherical wave and a diffuse sound field. The spherical wave models the direct sound of the point-source in a reverberant environment, while the diffuse field models room reverberation and spatially extended sound sources. As shown in Figure 2, the



[FIG2] A geometric sound field model: the direct sound emitted by a point source arrives at the array with a certain DOA, and the point-source position can be estimated when the DOA estimates from at least two arrays are available.



[FIG3] A block diagram for spatial analysis.

spherical wave is emitted by an isotropic point-like source (IPLS) located at a time-frequency-dependent position $\mathbf{d}_{\text{IPLS}}(k,n)$. The magnitude of the pressure of the spherical wave is inversely proportional to the distance traveled, which is known in physics as the inverse distance law. The diffuse sound is assumed to be spatially isotropic and homogenous, which means that diffuse sound arrives from all directions with equal power and that its power is position independent. Finally, it is assumed that the direct sound and diffuse sound are uncorrelated.

The direct and diffuse sounds are captured with one or more microphone arrays (depending on the application) that are located in the far field of the sound sources. Therefore, at the microphone array(s), the spherical wave can be approximated by a plane wave arriving from direction $\theta(k,n)$. In the following, we will differentiate between two related geometrical models: the DOA-based model and the position-based model. In the DOA-based model, the DOA and direct sound are estimated with a single microphone array, while in the position-based model, the position of the IPLS is estimated using at least two spatially distributed arrays, and the sound is captured using one or more microphones.

Under the aforementioned assumptions, the signals received at the omnidirectional microphones of an M-element microphone array can be written as

$$\mathbf{x}(k,n) = \mathbf{x}_{\text{s}}(k,n) + \mathbf{x}_{\text{d}}(k,n) + \mathbf{x}_{\text{n}}(k,n), \tag{1}$$

where the vector $\mathbf{x}(k,n) = [X(k,n,\mathbf{d}_1),\ldots,X(k,n,\mathbf{d}_M)]^{\text{T}}$ contains the M microphone signals in the time-frequency domain, where $\mathbf{d}_{1\ldots M}$ are the microphone positions. Without loss of generality, the first microphone located at $\mathbf{d}_1$ is used as a reference microphone. The vector $\mathbf{x}_s(k,n) = [X_s(k,n,\mathbf{d}_1),\ldots,X_s(k,n,\mathbf{d}_M)]^{\text{T}}$ is the captured direct sound at the different microphones and $\mathbf{x}_{\text{d}}(k,n) = [X_{\text{d}}(k,n,\mathbf{d}_1),\ldots,X_{\text{d}}(k,n,\mathbf{d}_M)]^{\text{T}}$ is the captured diffuse sound. Furthermore, $\mathbf{x}_{\text{n}}(k,n)$ contains the slowly time-varying noise signals (for example, the microphone self-noise). The direct sound at the different microphones can be related to the direct sound at the reference microphone via the array propagation vector $\mathbf{g}(k,\theta)$, which can be expressed as

$$\mathbf{x}_{\text{s}}(k,n) = \mathbf{g}(k,\theta)X_{\text{s}}(k,n,\mathbf{d}_1). \tag{2}$$

The mth element of the array propagation vector $\mathbf{g}(k,\theta) = [g(k,n,\mathbf{d}_1),\ldots,g(k,n,\mathbf{d}_M)]^{\text{T}}$ is the relative transfer function of the direct sound from the mth to the first microphone, which depends on the DOA $\theta(k,n)$ of the direct sound from the point of view of the array. For instance, for a uniform linear array of omnidirectional microphones $g(k,n,\mathbf{d}_m) = \exp\{j\kappa\|\mathbf{d}_m - \mathbf{d}_1\|\sin\theta\}$ where j denotes the imaginary unit, κ is the wavenumber, and $\|\mathbf{d}_m - \mathbf{d}_1\|$ is the distance between positions $\mathbf{d}_m$ and $\mathbf{d}_1$.

In this article, we will demonstrate how this geometric model can be effectively utilized to support a number of assisted listening applications. In the considered applications, the desired output signal of a loudspeaker (or headphone) channel $Y_i(k,n)$ is given as a weighted sum of the direct and diffuse sound at the reference microphone, i.e.,

$$Y_i(k,n) = G_i(k,n)X_{\text{s}}(k,n,\mathbf{d}_1) + Q_i(k)X_{\text{d}}(k,n,\mathbf{d}_1) \tag{3a}$$
$$= Y_{s,i}(k,n) + Y_{\text{d},i}(k,n), \tag{3b}$$

where i is the index of the output channel, and $G_i(k,n)$ and $Q_i(k)$ are the application-dependent weights. It is important to note that $G_i(k,n)$ depends on the DOA $\theta(k,n)$ of the direct sound or on the position $\mathbf{d}_{\text{IPLS}}(k,n)$. To synthesize a desired output signal two steps are required: 1) extract the direct and diffuse sound components and estimate the parameters (i.e., DOAs or positions), and 2) determine the weights $G_i(k,n)$ and $Q_i(k)$ using the estimated parameters and application-specific requirements. The first step is commonly referred to as the *spatial analysis* and is discussed next. In this article, the second step is referred to as the *application-specific synthesis*.

SPATIAL ANALYSIS

To facilitate flexible sound field manipulation with high-quality audio signals, it is crucial to accurately estimate the components describing the sound field, specifically the direct and diffuse sound components, as well as the DOAs or positions. Such spatial analysis based on the microphone signals is depicted in Figure 3. The direct and diffuse sound components can be estimated using single-channel or multichannel filters. To compute these filters, we may exploit knowledge about the DOA estimate of the direct sound or compute additional parameters as discussed in the following.

SIGNAL EXTRACTION

SINGLE-CHANNEL FILTERS

A computationally efficient estimation of the direct and the diffuse components is possible using single-channel filters. Such processing is applied for instance in DirAC [1], where the direct and diffuse signals are estimated by applying a spectral gain to a single microphone signal. The direct sound is then estimated as

$$\hat{X}_{\mathrm{s}}(k,n,\mathbf{d}_1) = W_{\mathrm{s}}(k,n)X(k,n,\mathbf{d}_1), \tag{4}$$

where $W_{\mathrm{s}}(k,n)$ is a single-channel filter, which is multiplied with the reference microphone signal to obtain the direct sound at $\mathbf{d}_1$. An optimal filter $W_{\mathrm{s}}(k,n)$ can be found, for instance, by minimizing the mean-squared error between the true and estimated direct sound, which yields the well-known Wiener filter (WF). If we assume no microphone noise, the WF for extracting the direct sound is given by $W_{\mathrm{s}}(k,n) = 1 - \Psi(k,n)$. Here, $\Psi(k,n)$ is the diffuseness, which is defined as

$$\Psi(k,n) = \frac{1}{1+\mathrm{SDR}(k,n)}, \tag{5}$$

where $\mathrm{SDR}(k,n)$ is the signal-to-diffuse ratio (SDR) (power ratio of the direct sound and the diffuse sound). The diffuseness is bounded between zero and one, and describes how diffuse the sound field is at the recording position. For a purely diffuse field, the SDR is zero leading to the maximum diffuseness $\Psi(k,n) = 1$. In this case, the WF, $W_{\mathrm{s}}(k,n)$, equals zero and thus, the estimated direct sound in (4) equals zero as well. In contrast, when the direct sound is strong compared to the diffuse sound, the SDR is high and the diffuseness in (5) approaches zero. In this case, the WF $W_{\mathrm{s}}(k,n)$ approaches one and thus, the estimated direct sound in (4) is extracted as the microphone signal. The SDR or diffuseness, required to compute the WF, is estimated using multiple microphones as will be explained in the section "Parameter Estimation."

The diffuse sound $X_{\mathrm{d}}(k,n,\mathbf{d}_1)$ can be estimated in the same way as the direct sound. In this case, the optimal filter is found by minimizing the mean-squared error between the true and estimated diffuse sound. The resulting WF is given by $W_{\mathrm{d}}(k,n) = \Psi(k,n)$. Instead of using the WF, the square root of the WF is often applied to estimate the direct sound and diffuse sound (cf. [1]). In the absence of sensor noise, the total power of the estimated direct and diffuse sound components is then equal to the total power of the received direct and diffuse sound components.

In general, extracting the direct and diffuse signals with single-channel filters has several limitations:

1) Although the required SDR or diffuseness are estimated using multiple microphones (as will be discussed later), only a single microphone signal is utilized for the filtering. Hence, the available spatial information is not fully exploited.

2) The temporal resolution of single-channel filters may be insufficient in practice to accurately follow rapid changes in the sound scene. This can cause leakage of the direct sound into the estimated diffuse sound.

3) The WFs defined earlier do not guarantee a distortionless response for the estimated direct and diffuse sounds, i.e., they may alter the direct and diffuse sounds, respectively.

4) Since the noise, such as the microphone self-noise or the background noise, is typically not considered when computing the filters, it may leak into the estimated signals and deteriorate the sound quality.

Limitations 1 and 4 are demonstrated in Figure 4(a), (b), and (d), where the spectrograms of the input (reference microphone) signal and both extracted components for the noise only (before time frame 75), castanet sound (between time frame 75 and time frame 150), and speech (latter frames) are shown. The noise is clearly visible in the estimated diffuse sound and slightly visible in the estimated direct sound. Furthermore, the onsets of the castanets leak into the estimated diffuse signal, while the reverberant sound from the castanets and the speech leaks into the estimated direct signal.

MULTICHANNEL FILTERS

Many limitations of single-channel filters can be overcome by using multichannel filters. In this case, the direct and diffuse signals are estimated via a weighted sum of multiple microphone signals. The direct sound is estimated with

$$\hat{X}_{\mathrm{s}}(k,n,\mathbf{d}_1) = \mathbf{w}_{\mathrm{s}}^{\mathrm{H}}(k,n)\mathbf{x}(k,n), \tag{6}$$

where $\mathbf{w}_{\mathrm{s}}(k,n)$ is a complex weight vector containing the filter weights for the M microphones and $(\cdot)^{\mathrm{H}}$ denotes the conjugate transpose. A filter $\mathbf{w}_{\mathrm{s}}(k,n)$ can be found for instance by minimizing the mean-squared error between the true and estimated direct sound, similarly as in the single-channel case. Alternatively, the filter weights can be found by minimizing the diffuse sound and noise at the filter output while providing a distortionless response for the direct sound, which assures that the direct sound is not altered by the filter. This filter is referred to as the linearly constrained minimum variance (LCMV) [21] filter, which can be obtained by solving

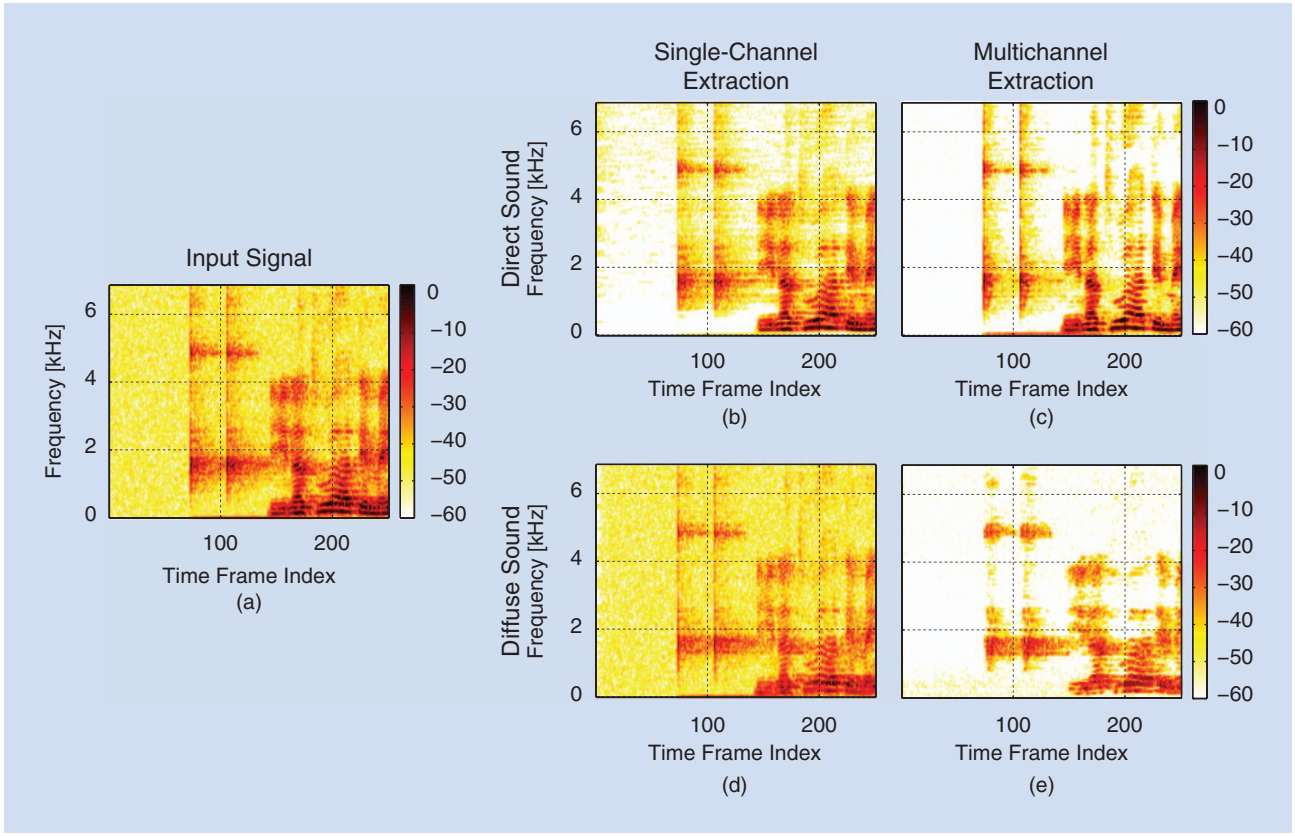
$$\begin{aligned}\mathbf{w}_{\mathrm{s}}(k,n) = \ & \underset{\mathbf{w}}{\operatorname{argmin}}\ \mathbf{w}^{\mathrm{H}}[\boldsymbol{\Phi}_{\mathrm{d}}(k,n)+\boldsymbol{\Phi}_{\mathrm{n}}(k)]\mathbf{w} \\ & \text{subject to } \mathbf{w}^{\mathrm{H}}(k,n)\mathbf{g}(k,\theta) = 1,\end{aligned} \tag{7}$$

where the propagation vector $\mathbf{g}(k,\theta)$ depends on the array geometry and DOA $\theta(k,n)$ of the direct sound. Here, $\boldsymbol{\Phi}_{\mathrm{d}}(k,n)$ is the power spectral density (PSD) matrix of the diffuse sound, which can be written using the aforementioned assumptions as

$$\boldsymbol{\Phi}_{\mathrm{d}}(k,n) = \mathrm{E}\{\mathbf{x}_{\mathrm{d}}(k,n)\mathbf{x}_{\mathrm{d}}^{\mathrm{H}}(k,n)\} \tag{8a}$$

$$= \phi_{\mathrm{d}}(k,n)\boldsymbol{\Gamma}_{\mathrm{d}}(k), \tag{8b}$$

where $\phi_{\mathrm{d}}(k,n)$ is the power of the diffuse sound and $\boldsymbol{\Gamma}_{\mathrm{d}}(k)$ is the diffuse sound coherence matrix. The (m',m)th element of $\boldsymbol{\Gamma}_{\mathrm{d}}(k)$ is the spatial coherence between the signals received at microphones m and m', which is known a priori when assuming a specific diffuse field characteristic. For instance, for a spherically isotropic diffuse field and omnidirectional microphones, the spatial coherence is a sinc function depending on



[FIG4] Spectograms of (a) the input signal, (b) the direct signal estimated using a single-channel filter, (c) the direct signal estimated using a multichannel filter, (d) the diffuse signal estimated using a single-channel filter, and (e) the diffuse signal estimated using a multichannel filter.

the microphone spacing and frequency [22]. Therefore, $\mathbf{\Phi}_{\mathrm{d}}(k,n)$ in (7) can be computed with (8b) when the diffuse sound power $\phi_{\mathrm{d}}(k,n)$ is known. The PSD matrix of the noise $\mathbf{\Phi}_{\mathrm{n}}(k)$ in (7) is commonly estimated during silence, i.e., when the sources are inactive, assuming that the noise is stationary. The estimation of $\phi_{\mathrm{d}}(k,n)$ and $\mathbf{\Phi}_{\mathrm{n}}(k)$ is explained in more detail in the next section. Note that the filter $\mathbf{w}_{\mathrm{s}}(k,n)$ is recomputed for each time-frequency bin with the geometric parameters estimated for that bin. The solution is computationally feasible since there exists a closed-form solution to the optimization problem in (7) [21].

To estimate the diffuse sound $\hat{X}_{\mathrm{d}}(k,n,\mathbf{d}_1)$, a multichannel filter that suppresses the direct sound and minimizes the noise while capturing the diffuse sound can be applied. Such a filter can be obtained by solving

$$\mathbf{w}_{\mathrm{d}}(k,n) = \arg\min_{\mathbf{w}} \mathbf{w}^{\mathrm{H}}\mathbf{\Phi}_{\mathrm{n}}(k)\mathbf{w} \text{ subject to}$$
$$\mathbf{w}^{\mathrm{H}}(k,n)\mathbf{g}(k,\theta) = 0 \text{ and } \mathbf{w}^{\mathrm{H}}(k,n)\mathbf{a}(k,n) = 1. \quad (9)$$

The first linear constraint ensures that the direct sound is strongly suppressed by the filter. The second linear constraint ensures that we capture the diffuse sound as desired. Note that there exist different definitions for the vector $\mathbf{a}(k,n)$. In [23], $\mathbf{a}(k,n)$ corresponds to the propagation vector of a notional plane wave arriving from a direction $\theta_0(k,n)$, which is far away from the DOA $\theta(k,n)$ of the direct sound. With this definition, $\mathbf{w}_{\mathrm{d}}(k,n)$ represents a multichannel filter that captures the diffuse sound mainly from direction $\theta_0(k,n)$, while attenuating the direct sound from direction $\theta(k,n)$. In [24], $\mathbf{a}(k,n)$ corresponds to the mean relative transfer function of the diffuse sound between the array microphones. With this approach, $\mathbf{w}_{\mathrm{d}}(k,n)$ represents a multichannel filter that captures the diffuse sound from all directions except for the direction $\theta(k,n)$ from which the direct sound arrives. Note that the optimization problem (9) has a closed-form solution [21], which can be computed when the DOA $\theta(k,n)$ of the direct sound is known.

Figure 4(c) and (e) depict the spectrograms of the direct sound and diffuse sound that were extracted using the multichannel LCMV filters for the example scenario consisting of noise, castanets, and speech. As can be observed, the direct sound extracted using the multichannel filter is less noisy and contains less diffuse sound compared to the direct sound extracted using the single-channel filter. Moreover, the diffuse sound extracted using the multichannel filer contains no onsets of the direct sound (clearly visible for the onsets of the castanets in time frames 75–150) and a significantly reduced noise level. As expected, the multichannel filters provide more accurate decomposition of the sound field into a direct and a diffuse signal component. The estimation accuracy strongly influences the performance of the discussed parametric processing approaches.

PARAMETER ESTIMATION

For the computation of the filters described in the previous section, the required parameters need to be estimated. In single-channel extraction, one parameter needs to be estimated, specifically the signal-to-diffuse ratio $\mathrm{SDR}(k,n)$ or the diffuseness $\Psi(k,n)$. In the case of multichannel signal extraction, the required parameters include the DOA $\theta(k,n)$ of the direct sound, the diffuse sound power $\phi_\mathrm{d}(k,n)$, and the PSD matrix $\mathbf{\Phi}_\mathrm{n}(k)$ of slowly time-varying noise. In addition, the DOA or the position of the direct sound sources, respectively, are required to control the application-specific processing and synthesis. It should be noted that the quality of the extracted and synthesized sounds is largely influenced by the accuracy of the estimated parameters.

The estimation of the DOA of a direct sound component is a well-addressed topic in literature and different approaches for this task are available. Common approaches to estimate the DOAs in the different frequency bands are ESPRIT and root MUSIC (cf. [21] and the references therein).

For estimating the SDR, two different approaches are common in practice, depending on which microphone array geometry is used. For linear microphone arrays, the SDR is typically estimated based on the spatial coherence between the signals of two array microphones [25]. The spatial coherence is given by the normalized cross-correlation between two microphone signals in the frequency domain. When the direct sound is strong compared to the diffuse sound (i.e, the SDR is high), the microphone signals are strongly correlated (i.e., the spatial coherence is high). On the other hand, when the diffuse sound is strong compared to the direct sound (i.e., the SDR is low), the microphone signals are less correlated.

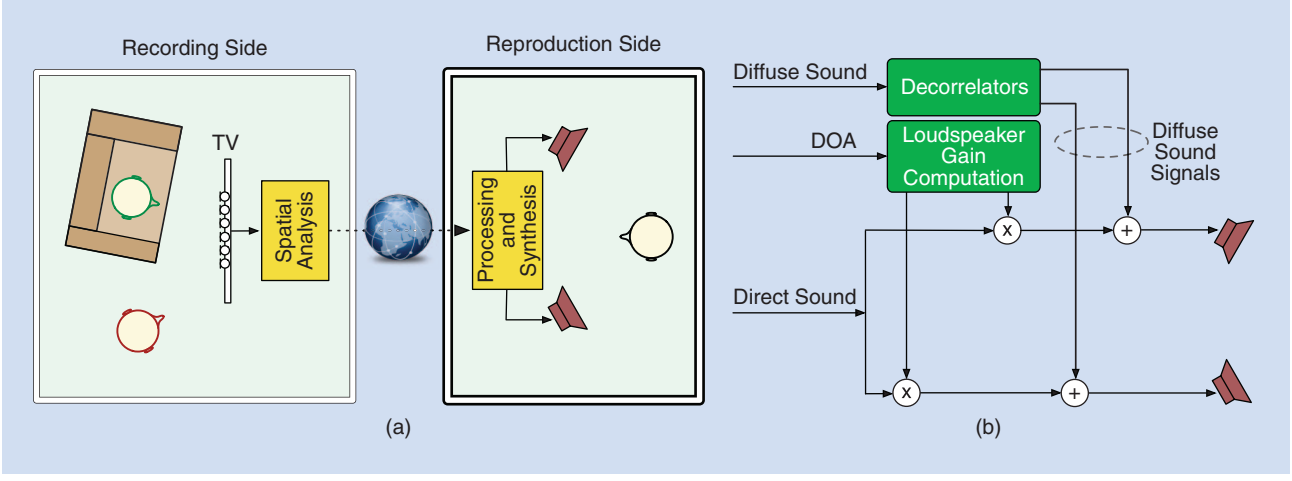
Alternatively, when a planar microphone array is used, the SDR can be estimated based on the so-called active sound intensity vector [26]. This vector points in the direction in which the acoustic energy flows. When only the direct sound arriving at the array from a specific DOA is present, the intensity vector constantly points in this direction and does not change its direction unless the sound source moves. In contrast, when the sound field is entirely diffuse, the intensity vector fluctuates quickly over time and points towards random directions as the diffuse sound is arriving from all directions. Thus, the temporal variation of the intensity vector can be used as a measure for the SDR and diffuseness, respectively [26]. Note that, as in [1], the inverse direction of the intensity vector can also be used to estimate the DOA of the direct sound. The intensity vector can be determined from an omnidirectional pressure signal and the particle velocity vector as described in [26], where the later signals can be computed from the planar microphone array as explained, for instance, in [11].

Various approaches have been described in the literature to estimate the slowly time-varying noise PSD matrix $\mathbf{\Phi}_\mathrm{n}(k)$. Assuming that the noise is stationary, which is a reasonable assumption in many applications (e.g., when the noise represents microphone self-noise or a stationary background noise), the noise PSD matrix can be estimated from the microphone signals during periods where only the noise is present in the microphone signals, which can be detected using a voice activity detector. To estimate the diffuse power $\phi_\mathrm{d}(k,n)$, we employ the spatial filter $\mathbf{w}_\mathrm{d}(k,n)$ in (9) that provides an estimate of the diffuse sound $X_\mathrm{d}(k,n,\mathbf{d}_1)$. Computing the mean power of $\hat{X}_\mathrm{d}(k,n,\mathbf{d}_1)$ yields an estimate of the diffuse power.

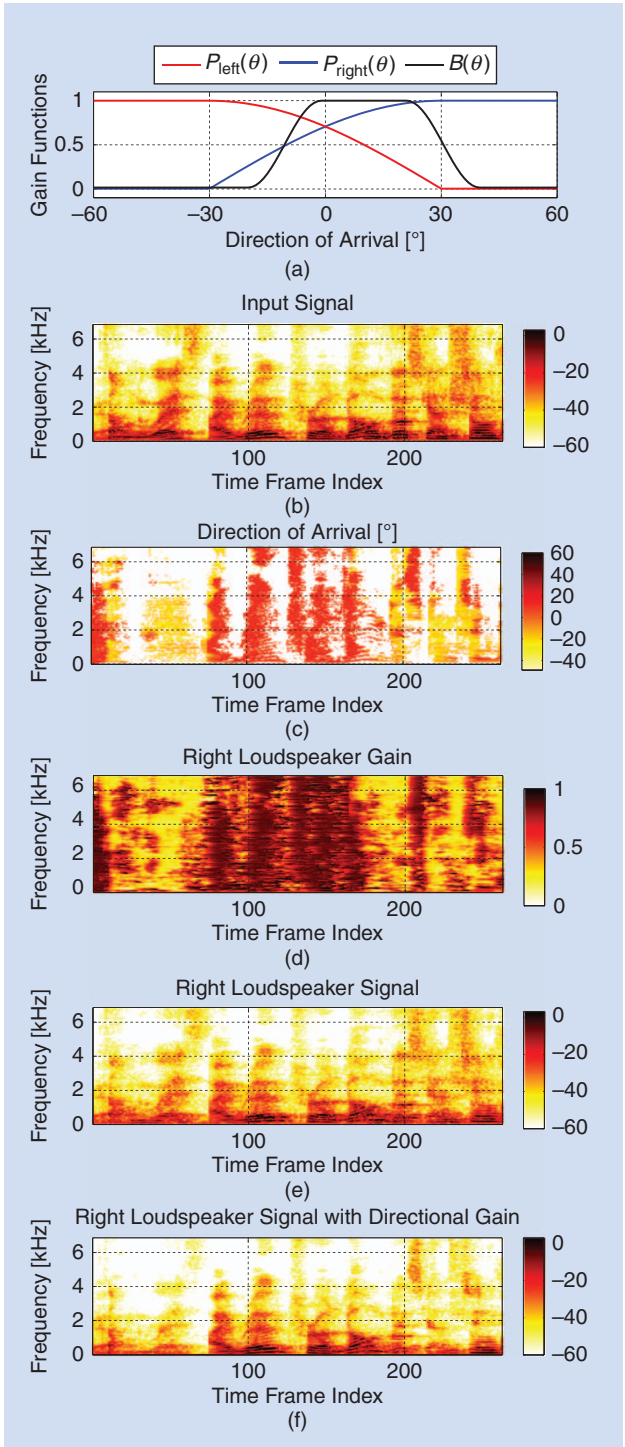
Finally, note that for some applications, such as the virtual classroom application described in the next section, the estimation of the IPLS positions from which the direct sounds originate may also be required to perform the application-specific synthesis. To determine the IPLS positions, the DOAs at different positions in the room are estimated using multiple distributed microphone arrays. The IPLS position can then be determined by triangulating the estimated DOAs, as done in [27] and illustrated in Figure 2.

APPLICATION-SPECIFIC SYNTHESIS

The compact description of the sound field in terms of a direct signal component, a diffuse signal component, and sound field parameters, as shown in Figure 1, can contribute to assisted listening in a variety of applications. While the spatial analysis yielding estimates of the model parameters and the direct and diffuse signal components at a reference microphone is application independent, the processing and synthesis is application dependent. For this



[FIG5] Spatial audio communication application: (a) communication scenario and (b) rendering of the loudspeaker signals.



[FIG6] The results in a communication scenario: (a) applied gain functions, (b) spectrogram of the input signal, (c) estimated directions of arrival, (d) gains applied to the direct sound for the right loudspeaker channel, (e) spectrogram of the right loudspeaker signal, and (f) spectrogram of the right loudspeaker signal after applying $B(\theta)$ defined in (a).

purpose, we adjust the gains $G_i(k,n)$ and $Q_i(k)$ in (3) depending on the application and as desired by the user. For spatial audio rendering, $G_i(k,n)$ and $Q_i(k)$ are used to generate the different output channels for a given reproduction setup, whereas for signal enhancement applications, $G_i(k,n)$ and $Q_i(k)$ are used to realize parametric filters that extract a signal of the desired sound source while reducing undesired and diffuse sounds. In all cases, the gains are computed using the estimated sound field parameters, and are used to obtain a weighted sum of the estimated direct and diffuse components, as given by (3). In the following, we present an overview of different applications in which the output signals are obtained using this approach.

SPATIAL AUDIO COMMUNICATION

Using spatial audio communication, we can allow participants in different locations to communicate with each other in a natural way. The sound acquisition and reproduction should provide good speech intelligibility, as well as a natural and immersive sound. Spatial cues are highly beneficial for understanding speech of a desired talker in multitalker and adverse listening situations [18]. Therefore, accurate spatial sound reproduction is expected to enable the human brain to better segregate spatially distributed sounds, which in turn could lead to better speech intelligibility. In addition, flexible spatial selectivity offered by adjusting the time-frequency dependent gains of the transmitted signals based on the geometric side information, enables the listener to focus even better on one or more talkers. These two features make the parametric methods particularly suited to immersive audio-video teleconferencing, where hands-free communication is typically desired. In hands-free communication (that is without any tethered microphones), the main challenge is to ensure the high quality of the reproduced audio signals captured from distance, and to recreate plausible spatial cues at the listeners ears. Note that for full-duplex communication, multichannel acoustic echo control would additionally be required to remove the acoustic coupling between the loudspeakers and the microphones [5]. However, the acoustic echo cancelation problem is beyond the scope of this article.

Let us consider such a teleconferencing scenario with two active talkers at the recording side, as illustrated in Figure 5. The goal is to recreate the spatial cues from the recording side at the listener side over an arbitrary, user-defined multichannel loudspeaker setup. At the recording side, one of the talkers is sitting on a couch located in front of a TV screen at a distance of 1.5 m and angle 10° with respect to array broadside direction, while the other is located to the left (at –20°) at roughly the same distance. The TV has a built-in camera and is equipped with a six-element linear array with inter-microphone spacing of 2.5 cm that captures the reverberant speech and noise (with SNR = 45 dB); the reverberation time is 350 ms. At the reproduction side, the ith loudspeaker signal is obtained as a weighted sum of the direct and diffuse signals, as given by (3). To recreate the original spatial impression of the recording side (without additional sound scene manipulation), the following gains suffice $G_i(k,n) = P_i(k,n,\theta)$ and $Q_i(k) = 1$, where $P_i(k,n,\theta)$ is the panning gain for reproducing the direct sound from the correct direction, which depends on the selected panning scheme and the loudspeaker setup. As an example, the vector-base amplitude panning (VBAP) [28] gain factors for a stereo reproduction system with loudspeakers positioned at $\pm 30°$ are

depicted in Figure 6(a). To reproduce the diffuse sound, the signals $Y_{\mathrm{d},i}(k,n)$ are decorrelated such that $Y_{\mathrm{d},i}(k,n)$ and $Y_{\mathrm{d},j}(k,n)$ for $i \neq j$ are uncorrelated [29]. Note that the less correlation between the loudspeaker channels, the more enveloping the perceived sound is. The described processing for synthesizing the loudspeaker signals is depicted in Figure 5(b).

When sound scene manipulation, such as directional filtering [10] and dereverberation [11], is also desired, an additional gain $B(k,n,\theta)$ can be applied to modify the direct signal. In this case, the ith loudspeaker channel gain for the direct sound can be expressed as

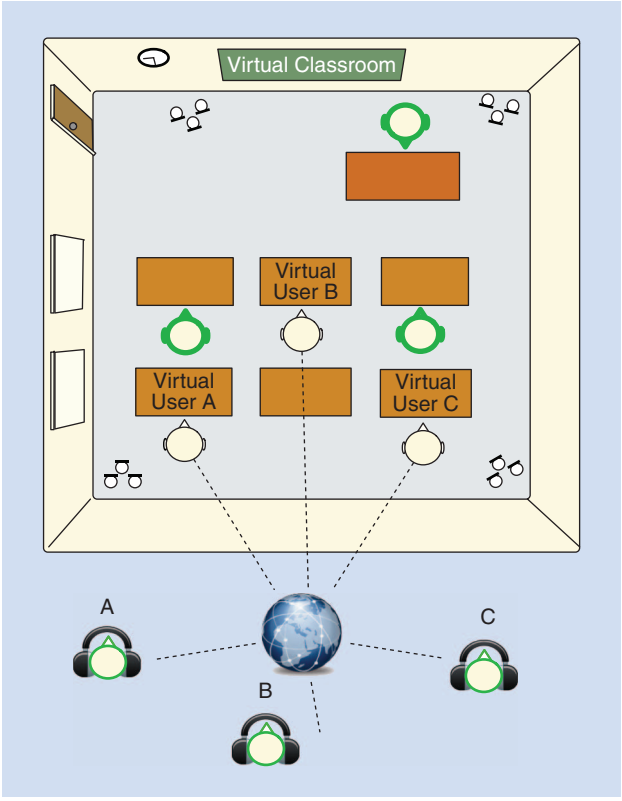
$$G_i(k,n) = P_i(k,n,\theta)B(k,n,\theta), \tag{10}$$

where $B(k,n,\theta)$ is the desired gain for the sound arriving from $\theta(k,n)$. In principle, $B(k,n,\theta)$ can be defined freely to provide any desired directivity pattern; an example directivity gain function is shown in Figure 6(a). In addition, the diffuse sound gain $Q_i(k)$ can be adjusted to control the level of reproduced ambient sound. For instance, dereverberation is achieved by selecting $Q_i(k) < 1$.

The results for the considered teleconferencing scenario are illustrated in Figure 6. Depicted in Figure 6(a)–(c) are the gain functions, the spectrogram of an input signal, and the DOAs estimated using ESPRIT. Figure 6(d) and (e) illustrate the spatial reproduction and depict the panning gains $P_{\mathrm{right}}(k,n,\theta)$ used for the right loudspeaker and the spectrogram of the resulting signal. Lower weights can be observed when the source on the left side is active than for the source in the right, which is expected from the panning curve $P_{\mathrm{right}}(k,n,\theta)$ depicted in Figure 6(a). Note that the exact values for the respective DOAs should be $P_{\mathrm{right}} = 0.26$ for $-20°$ and $P_{\mathrm{right}} = 0.86$ for $10°$. Next we illustrate an example of sound scene manipulation. If the listener prefers to extract the signal of the talker sitting on a sofa, while reducing the other talker, a suitable gain function $B(k,n,\theta)$ can be designed to preserve the sounds coming from the sofa and attenuate sounds arriving from other directions; an example of such a gain function is shown in Figure 6(a). Additionally, setting the diffuse gain to a low value, for example $Q_i(k) = 0.25$, reduces the power level of the diffuse sound, thereby increasing the SDR during reproduction. The spectrogram of the manipulated output signal is shown in Figure 6(f), where the power of the interfering talker and reverberation are significantly reduced.

VIRTUAL CLASSROOM

The geometric model with IPLS positions as parametric information can facilitate assisted listening by creating binaural signals for any desired position in the acquired sound scene, regardless of where the microphone arrays are located. Let us consider the virtual classroom scenario in Figure 7 as an example, although the same concept also applies to other applications such as teleconference systems in dedicated rooms, assisted listening in museums, augmented reality, and many others. A teacher tutors in a typical classroom environment, where only some students are physically present, while the rest participates in the class remotely, for example, from home. As illustrated in Figure 7, the sound scene is



[FIG7] A virtual classroom scenario.

captured using several distributed microphone arrays, with known positions. The goal is to assist a remote student to virtually participate in a class from his preferred position, for instance close to the teacher, in between the teacher and another student involved in the discussion, or at his favorite desk, by synthesizing the binaural signals for the desired virtual listener (VL) location $\mathbf{d}_{\mathrm{VL}}$. These binaural signals are generated at the reproduction side based on the received audio and position information, such that the student could listen to the synthesized sound over headphones on a laptop or any mobile device that can play multimedia content.

The processing to achieve this goal is in essence similar to that utilized in the virtual microphone (VM) technique [12], [27], [30], where the goal was to generate the signal of a VM that sounds perceptually similar to the signal that would be recorded with a physical microphone located at the same position. The technique has been shown successful in synthesizing the VM signals in arbitrary positions in a room [27], [30]. However, in the virtual classroom application, instead of generating the signals of nonexisting microphones with physical characteristics, we directly aim to generate the binaural signals for headphone reproduction. The overall gain for the direct sound in the ith channel can be divided into three components:

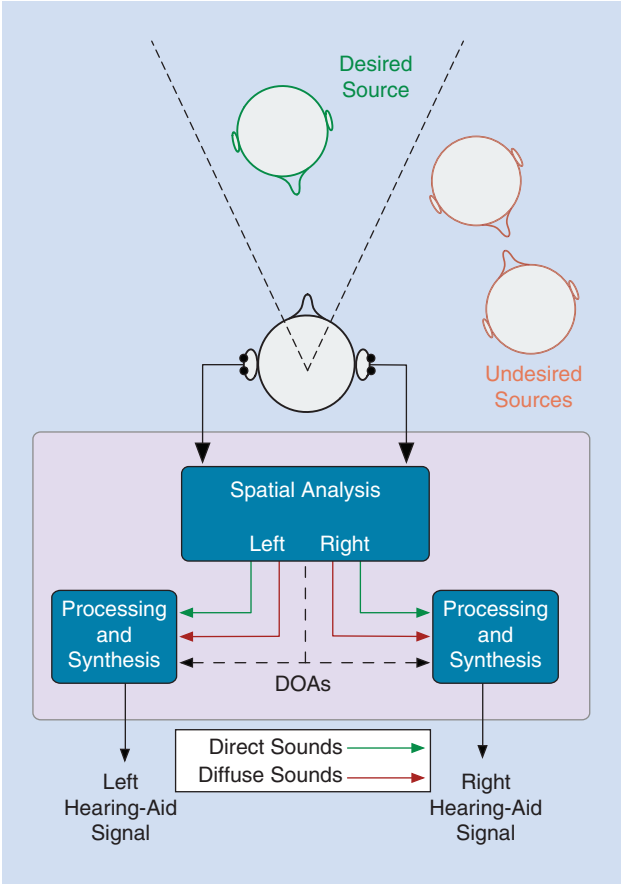
$$G_i(k,n) = D_{\mathrm{s}}(k,n)H_{\mathrm{HRTF},i}(k,n)B(k,n,\mathbf{d}_{\mathrm{IPLS}}). \tag{11}$$

The first gain $D_{\mathrm{s}}(k,n)$ is a factor compensating for the wave propagation from $\mathbf{d}_{\mathrm{IPLS}}$ to the VL position $\mathbf{d}_{\mathrm{VL}}$, and from $\mathbf{d}_{\mathrm{IPLS}}$ to $\mathbf{d}_1$ for the direct signal estimated at the reference

microphone position $\mathbf{d}_1$. As in [27], the real factors are typically applied which compensate for the amplitude change following the $1/r$ law, where r is the propagated distance. The second gain $H_{\text{HRTF},i}(k,n)$ is a complex head-related transfer function (HRTF) for the left or right ear, $i \in \{\text{left}, \text{right}\}$, respectively, which depends on the DOA $\theta_{\text{VL}}(k,n)$ with respect to the position and look direction of the VL. Apart from creating a plausible feeling of being present in the actual classroom, the user-defined spatial selectivity can be achieved with the third gain $B(k,n,\mathbf{d}_{\text{IPLS}})$, which enables the amplification or attenuation of directional sounds emitted from $\mathbf{d}_{\text{IPLS}}$ as desired. In principle, any desired spatial selectivity function $B(k,n,\mathbf{d})$ can be defined. For instance, a spatial spot can be defined at a teacher's desk or in front of a blackboard to assist the student in better hearing the teacher's voice. Such a gain function for a circular spot centered around $\mathbf{d}_{\text{spot}}$ with a 1 m radius could be defined as

$$B(k,n,\mathbf{d}_{\text{IPLS}}) = \begin{cases} 1 & r < 1; \\ \dfrac{1}{r^{\alpha}} & \text{otherwise}, \end{cases} \tag{12}$$

where $r = \| \mathbf{d}_{\text{spot}} - \mathbf{d}_{\text{IPLS}}(k,n) \|$ and α controls the spatial selectivity for the sources located outside the spot. In addition, the gain $Q_i(k) \in [0,1]$ applied to the diffuse component enables the student to control the level of the ambient sound. The output



[FIG8] A general parametric spatial sound processing scheme for binaural hearing aids.

diffuse signals $Y_{\text{d},i}(k,n)$ for the left and right headphone channel are decorrelated such that the coherence between $Y_{\text{d,left}}(k,n)$ and $Y_{\text{d,right}}(k,n)$ corresponds to the target coherence in binaural hearing [18], [29]. Finally, it should be noted that since the propagation compensation and the spatial selectivity gains are typically real factors, the phase of the direct and diffuse components are equal to those observed at the reference microphone. However, the complex HRTFs that dependent on the DOAs at the virtual listening position ensure that the spatial cues are correct.

BINAURAL HEARING AIDS

Developments in acoustic signal processing and psychoacoustics have lead to the advancement of digital hearing aids that were first developed in the 1990s. The early devices included the unilateral (i.e., single-ear) and bilateral hearing aids, where two independent unilateral hearing aids are used for the left and right ears, respectively. More recently binaural hearing aids, in which signals and parameters can be exchanged between the left and right hearing aid, have been brought to the market. Binaural hearing aids are advantageous compared to unilateral and bilateral hearing aids as they can further improve speech intelligibility in difficult listening situations, improve the ability to localize sounds, and decrease listening fatigue. Besides dynamic range compression and feedback cancelation, wind and ambient noise reduction, dereverberation and directional filtering are important features of state-of-the-art hearing aids.

Let us consider a situation in which we have one desired talker in front and two interfering talkers at the right side of the hearing-aids user, as illustrated in Figure 8. In such a situation, directional filtering allows a hearing-aid user to perceive sounds arriving from the front more clearly than the sounds from the sides. In addition, one can aim at reducing the amount of diffuse sounds such that the SDR increases.

While many state-of-the-art directional filtering techniques for hearing aids are based on classical differential array processing, some parametric spatial sound processing techniques have been proposed. In [14], the left and right microphone signals were jointly analyzed in the time-frequency domain to determine: 1) the interaural phase difference and interaural level difference that strongly depend on the DOA of the direct sound, and 2) the interaural coherence that measures the degree of diffuseness. Based on these parameters, three gains were computed related to the degree of diffuseness, signal-to-interference ratio, and direction of the sound. Finally, real-valued gains for the left and right microphones were determined based on these gains to reduce reverberation and interfering sounds. According to the authors of [14], the quality of the signal was good but the speech intelligibility improvement for a single interfering talker was unsatisfactory. In [15], the authors used two microphones at each side and adopted the DOA-based geometric model. The DOAs were estimated at low frequencies using the microphones at the left and respectively right side, and at high frequencies using the intermicrophone level differences. Finally, the signal of a single microphone positioned at the left and right, respectively, was modified based on the DOA

estimates and degree of diffuseness. The evaluation of different setups with one desired talker and one interfering talker demonstrated that an improvement in the speech reception threshold (SRT) between 4 and 24 dB could be obtained.

In Figure 8, a general parametric spatial sound processing scheme is illustrated, where spatial analysis provides the DOA estimates, and the direct and diffuse sound estimates for the left and right ear are obtained using different (left or right) reference microphones. The left (and right) output signal can then be computed using (3) with $G_i(k,n) = B(k,n,\theta)H_{\text{ex}}(k)$ for $i \in \{\text{left}, \text{right}\}$, where $B(k,n,\theta)$ defines the desired spatial response that depends on the listening mode, $H_{\text{ex}}(k)$ helps to externalize sounds, and $Q_i(k) = c(k)H_{\text{ex}}(k)$ with $0 \le c(k) < 1$ is a constant used to reduce the diffuse sound and hence increase the SDR at the output. At the cost of an increase in computational complexity and memory use, the proposed scheme can fully exploit all microphones.

While many more examples can be found in the literature, it can readily been seen that the parametric spatial sound processing, using either geometrically or psychoacoustically motivated parametric models, provides a flexible and efficient way to achieve directional filtering. The limited improvement in terms of the SRT reported in [14] could be related to the inherent tradeoff between interference reduction and speech distortion found in most single-channel processing techniques. Further research is required to develop robust and efficient parameter estimators for this application and to study the impact on the SRT. More advanced schemes to modify the spatial response and the DOAs based on the listening mode and the listening situation could be realized using the processing scheme depicted in Figure 8.

CONCLUSIONS

Parametric models have been shown to provide an efficient way to describe sound scenes. While in earlier work multiple microphones were only used to estimate the geometric model parameters, in more recent work it has been shown that they can also be used to estimate the direct and diffuse sound components. As the latter estimates are more accurate than single-channel estimates, the sound quality of the overall system is increased, for instance, by avoiding decorrelating the direct sound that may partially leak into the diffuse sound estimate in single-channel extraction. Depending on the application, the estimated components and parameters can be manipulated before computing one or more output signals by mixing the components together based on the parametric side information. In a spatial audio communication scenario in which the direct and diffuse signals as well as the parameters are transmitted to the far-end side, it is possible to determine at the receiver side which sounds to extract and how to accurately reproduce the recorded spatial sounds over loudspeakers or headphones. By using the position-based model, we have shown how binaural signals can be synthesized at the receiver side that correspond to a desired listening position on the recording side. Finally, we have described how parametric spatial sound processing can be applied to binaural hearing aids to achieve both directional filtering and dereverberation.

To date, the majority of the geometric models assume that at most one direct sound is active per time-frequency. Extensions of these models are currently under development where multiple direct sound components plus diffuse sound components coexist in a single time-frequency instance [23]. Preliminary results have shown that this model can help to further improve the spatial selectivity and sound quality.

We hope that by presenting this unified perspective on parametric spatial sound processing we can help readers to approach other problems encountered in assisted listening from this perspective and to help highlight relations between a family of approaches that may initially seem divergent.

ACKNOWLDGEMENTS

This work has received funding from the European Community's Seventh Framework (FP7/2007-2013) under grant agreement ICT-2011-287760, from the European Research Council under the European Community's Seventh Framework (FP7/2007-2013)/ERC grant agreement 240453, and from the Academy of Finland.

AUTHORS

Konrad Kowalczyk (konrad.kowalczyk@iis.fraunhofer.de) received the B.Eng. and M.Sc. degrees in telecommunications from AGH University of Science and Technology, Krakow, Poland, in 2005 and the Ph.D. degree in electronics and electrical engineering from Queens University, Belfast, United Kingdom, in 2009. From 2009 until 2011, he was a postdoctoral research fellow at the Chair of Multimedia Communications and Signal Processing, Friedrich-Alexander-University Erlangen-Nürnberg, Germany. In 2012, he joined Fraunhofer Institue for Integrated Circuits IIS as an associate researcher for communication acoustics and spatial audio processing. His main research interests include virtual acoustics, sound field analysis, spatial audio, signal enhancement, and array signal processing.

Oliver Thiergart (oliver.thiergart@iis.fraunhofer.de) studied media technology at Ilmenau University of Technology (TUI), Germany, and received his Dipl.-Ing. (M.Sc.) degree in 2008. In 2008, he was with the Fraunhofer Institute for Digital Media Technology IDMT in Ilmenau where he worked on sound field analysis with microphone arrays. He then joined the Audio Department of the Fraunhofer Institute for Integrated Circuits IIS in Erlangen, Germany, where he worked on spatial audio analysis and reproduction. In 2011, he became a member of the International Audio Laboratories Erlangen where he is currently pursuing a Ph.D. degree in the field of parametric spatial sound processing.

Maja Taseska (maja.taseska@audiolabs-erlangen.de) received her B.Sc. degree in electrical engineering at Jacobs University, Bremen, Germany, in 2010, and her M.Sc. degree at the Friedrich-Alexander-University Erlangen-Nürnberg, Germany, in 2012. She then joined the International Audio Laboratories Erlangen, where she is currently pursuing a Ph.D. degree in the field of informed spatial filtering. Her current research interests include informed spatial filtering, source localization and tracking, blind source separation, and noise reduction.

Giovanni Del Galdo (giovanni.delgaldo@iis.fraunhofer.de) studied telecommunications engineering at Politecnico di

Milano, Italy. In 2007, he received his doctoral degree from Technische Universität Ilmenau on the topic of channel modeling for mobile communications. He then joined Fraunhofer Institute for Integrated Circuits IIS working on audio watermarking and parametric representations of spatial sound. In 2012, he was appointed full professor at TU Ilmenau in the research area of wireless distribution systems and digital broadcasting. His current research interests include the analysis, modeling, and manipulation of multidimensional signals, over-the-air testing for terrestrial and satellite communication systems, and sparsity-promoting reconstruction methods.

Ville Pulkki (Ville.Pulkki@aalto.fi) has been working in the field of audio since 1995. In his Ph.D. thesis (2001), he developed a method to position virtual sources for three-dimensional loudspeaker setups after researching the method using psychoacoustic listening tests and binaural computational models of human hearing. Later he worked on the reproduction of recorded spatial sound scenarios, on the measurement of head-related acoustics and on the measurement of room acoustics with laser-induced pressure pulses. Currently he holds a tenure-track assistant professor position in Aalto University and runs a research group with 14 researchers. He is a fellow of the Audio Engineering Society (AES) and received the AES Publication Award. He has also received the Samuel L. Warner Memorial Medal from the Society of Motion Picture and Television Engineers.

Emanuël A.P. Habets (e.habets@ieee.org) is an associate professor at the International Audio Laboratories Erlangen (a joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg and Fraunhofer IIS), and head of the Spatial Audio Research Group at Fraunhofer IIS, Germany. He received the Ph.D. degree in electrical engineering from the Technische Universiteit Eindhoven, The Netherlands, in 2007. From 2007 until 2009, he was a postdoctoral fellow at the Technion–Israel Institute of Technology and at the Bar-Ilan University, Israel. From 2009 until 2010, he was a research fellow at Imperial College London, United Kingdom. Currently, he is an associate editor of *IEEE Signal Processing Letters*, a member of the IEEE Signal Processing Society (SPS) Technical Committee on Audio and Acoustic Signal Processing, a member of the IEEE SPS Standing Committee on Industry Digital Signal Processing Technology, and has been a guest editor of *IEEE Journal of Selected Topics in Signal Processing*. He is the recipient, with I. Cohen and S. Gannot, of the 2014 IEEE SPS Signal Processing Letters Best Paper Award. He is a Senior Member of the IEEE.

REFERENCES

[1] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.*, vol. 55, no. 6, pp. 503–516, June 2007.

[2] M. Goodwin and J.-M. Jot, "Spatial audio scene coding," in *Proc. Audio Engineering Society Convention 125*, Oct. 2008.

[3] Z. Fejzo, S. Hastings, J. D. Johnston, and J.-M. Jot, "Beyond coding: Reproduction of direct and diffuse sound in multiple environments," in *Proc. Audio Engineering Society Convention 129*, Nov. 2010.

[4] S. Gannot and I. Cohen, "Adaptive beamforming and postfiltering," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Berlin, Germany: Springer-Verlag, 2008, ch. 47, pp. 945–978.

[5] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[6] C. Faller, "Parametric coding of spatial audio," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2004.

[7] J. Herre, C. Falch, D. Mahne, G. Del Galdo, M. Kallinger, and O. Thiergart, "Interactive teleconferencing combining spatial audio object coding and DirAC technology," *J. Audio Eng. Soc.*, vol. 59, no. 12, pp. 924–935, 2011.

[8] M.-V. Laitinen and V. Pulkki, "Binaural reproduction for directional audio coding," in *Proc. IEEE Workshop Applications Signal Processing Audio Acoustics, (WASPAA'09)*, Oct. 2009, pp. 337–340.

[9] I. Tashev, M. Seltzer, and A. Acero, "Microphone array for headset with spatial noise suppressor," in *Proc. 9th Int. Workshop Acoustic, Echo, Noise Control (IWAENC)*, Eindhoven, The Netherlands, 2005, pp. 29–32.

[10] M. Kallinger, G. Del Galdo, F. Kuech, D. Mahne, and R. Schultz-Amling, "Spatial filtering using directional audio coding parameters," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Apr. 2009, pp. 217–220.

[11] M. Kallinger, G. Del Galdo, F. Kuech, and O. Thiergart, "Dereverberation in the spatial audio coding domain," in *Proc. Audio Engineering Society Convention 130*, London, U.K., May 2011.

[12] G. Del Galdo, O. Thiergart, T. Weller, and E. A. P. Habets, "Generating virtual microphone signals using geometrical information gathered by distributed arrays," in *Proc. Hands-Free Speech Communication Microphone Arrays (HSCMA)*, Edinburgh, U.K., May 2011, pp. 185–190.

[13] R. Schultz-Amling, F. Kuech, O. Thiergart, and M. Kallinger, "Acoustical zooming based on a parametric sound field representation," in *Proc. Audio Engineering Society Convention 128*, London, U.K., May 2010.

[14] T. Wittkop and V. Hohmann, "Strategy-selective noise reduction for binaural digital hearing aids," *Speech Commun.*, vol. 39, no. 1–2, pp. 111–138, Jan. 2003.

[15] J. Ahonen, V. Sivonen, and V. Pulkki, "Parametric spatial sound processing applied to bilateral hearing aids," in *Proc. Audio Engineering Society Conf.: 45th Int. Conf. Applications Time-Frequency Processing Audio*, Mar. 2012.

[16] C. Faller, "Microphone front-ends for spatial audio coders," in *Proc. Audio Engineering Society Convention 125*, San Francisco, CA, Oct. 2008.

[17] S. Berge and N. Barrett, "High angular resolution planewave expansion," in *Proc. 2nd Int. Symp. Ambisonics Spherical Acoustics*, May 2010.

[18] J. Blauert, Ed., *Communication Acoustics*. New York: Springer, 2005, vol. 1.

[19] S. Rickard and Z. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, Apr. 2002, vol. 1, pp. 529–532.

[20] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing*. New York: Springer, 2008.

[21] H. L. van Trees, *Detection, Estimation, and Modulation Theory*, vol. IV, *Optimum Array Processing*. New York: Wiley, Apr. 2002.

[22] G. W. Elko, "Spatial coherence functions for differential microphones in isotropic noise fields," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. Berlin: Springer, 2001, ch. 4, pp. 61–85.

[23] O. Thiergart, M. Taseska, and E. Habets, "An informed parametric spatial filter based on instantaneous direction-of-arrival estimates," *IEEE/ACM Trans. Audio, Speech, Lang. Processing*, vol. 22, no. 12, pp. 2182–2196, Dec. 2014.

[24] O. Thiergart and E. A. P. Habets, "Extracting reverberant sound using a linearly constrained minimum variance spatial filter," *IEEE Signal Process. Lett.*, vol. 21, no. 5, pp. 630–634, May 2014.

[25] O. Thiergart, G. D. Galdo, and E. A. P. Habets, "On the spatial coherence in mixed sound fields and its application to signal-to-diffuse ratio estimation," *J. Acoust. Soc. Amer.*, vol. 132, no. 4, pp. 2337–2346, 2012.

[26] G. Del Galdo, M. Taseska, O. Thiergart, J. Ahonen, and V. Pulkki, "The diffuse sound field in energetic analysis," *J. Acoust. Soc. Amer.*, vol. 131, no. 3, pp. 2141–2151, Mar. 2012.

[27] O. Thiergart, G. Del Galdo, M. Taseska, and E. Habets, "Geometry-based spatial sound acquisition using distributed microphone arrays," *IEEE Trans. Audio, Speech, Lang. Processing*, vol. 21, no. 12, pp. 2583–2594, Dec. 2013.

[28] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, 1997.

[29] V. Pulkki and J. Merimaa, "Spatial impulse response rendering II: Reproduction of diffuse sound and listening tests," *J. Audio Eng. Soc.*, vol. 54, no. 1/2, pp. 3–20, Feb. 2006.

[30] K. Kowalczyk, O. Thiergart, A. Craciun, and E. A. P. Habets, "Sound acquisition in noisy and reverberant environments using virtual microphones," in *Proc. 2013 IEEE Workshop Applications Signal Processing Audio Acoustics (WASPAA)*, Oct. 2013.

[SP]