Katholieke Universiteit Leuven

Departement Elektrotechniek

ESAT-SISTA/TR 12-173

Acoustic zooming by multi-microphone sound scene manipulation[1]

Toon van Waterschoot[2] [3], Wouter Joos Tirry[4], and Marc Moonen[2]

July 2013
Published in *J. Audio Eng. Soc.*, vol. 61, no. 7/8, Jul/Aug. 2013, pp. 489–507.

[1]This report is available by anonymous ftp from *ftp.esat.kuleuven.be* in the directory *pub/sista/vanwaterschoot/reports/12-173.pdf*

[2]K.U.Leuven, Dept. of Electrical Engineering (ESAT), Research group SCD(SISTA), Kasteelpark Arenberg 10, 3001 Leuven, Belgium, Tel. +32 16 321927, Fax +32 16 321970, WWW: *http://homes.esat.kuleuven.be/∼tvanwate*. E-mail: *toon.vanwaterschoot@esat.kuleuven.be, marc.moonen@esat.kuleuven.be*.

[3]T. van Waterschoot is a Postdoctoral Fellow of the Research Foundation Flanders (FWOVlaanderen). This research work was carried out at the ESAT Laboratory of Katholieke Universiteit Leuven, in collaboration with and funded by NXP Software, and in the frame of KU Leuven Research Council CoE EF/05/006 Optimization in Engineering (OPTEC) and PFV/10/002 (OPTEC), Concerted Research Action GOA-MaNet, the Belgian Programme on Interuniversity Attraction Poles initiated by the Belgian Federal Science Policy Office IUAP P6/04 Dynamical systems, control and optimization (DYSCO) 20072011, Research Project IBBT, and Research Project FWO nr. G.0600.08 Signal processing and network design for wireless acoustic sensor networks. The scientific responsibility is assumed by its authors.

[4]NXP Software, Interleuvenlaan 80, 3001 Leuven, Belgium, Tel. +32 16 390832, Fax +32 16 394490, WWW: *http://www.software.nxp.com/*. E-mail *wouter.tirry@nxp.com*.

Acoustic Zooming by Multimicrophone Sound Scene Manipulation

TOON VAN WATERSCHOOT,[1] *AES Associate Member,* **WOUTER JOOS TIRRY[2], AND**
(toon.vanwaterschoot@esat.kuleuven.be) (wouter.tirry@nxp.com)

MARC MOONEN,[3] *AES Associate Member*
(marc.moonen@esat.kuleuven.be)

[1]*KU Leuven, Department of Electrical Engineering, ESAT-SCD, BE-3001 Leuven, Belgium*
[2]*NXP Software, BE-3001 Leuven, Belgium*
[3]*KU Leuven, Department of Electrical Engineering, ESAT-SCD, BE-3001 Leuven, Belgium*

Zoom control is a key feature of audiovisual capture in both professional and consumer cameras. While a video zoom operation is often not complemented with a corresponding acoustic zoom, psychophysical as well as neuroimaging results suggest that a cross-modal approach to zooming may facilitate multisensory integration. As auditory distance perception is primarily determined by sound intensity, an audiovisual zoom effect may be obtained by matching the levels of different sources in a sound scene with their visually perceived motion during video zooming. In this paper, we propose a general theory for independent sound source level control which can be exploited to attain an acoustic zoom effect. An essential feature of the proposed theory is that it does not consist in an explicit sound source separation, which relieves its potential computational requirements. An efficient implementation using fixed and adaptive spatial and spectral noise reduction algorithms is proposed and evaluated. Experimental results using an array with a small number of low-cost microphones confirm that the proposed approach is particularly suited for consumer audiovisual capture applications.

0 INTRODUCTION

A vital aspect of user control in audiovisual capture is the ability to zoom in on a spatial area in the scene that is of particular interest. Video zooming has become a ubiquitous feature of audiovisual recording devices, both in high-end and consumer cameras. In contrast, only few devices are equipped with an acoustic zoom (AZ) feature, which allows to create the auditory illusion of approaching or receding from a sound source in front of the camera. A first, somewhat pragmatic reason for this discrepancy is presumably that from a technical point of view, it is more straightforward to attain spatial selectivity in the visual domain (i.e., using a zoom lens) than in the auditory domain (which typically requires the use of multiple microphones). A second and more fundamental reason is that a video zoom operation itself may already induce an auditory zoom sensation, even if it is complemented with auditory information corresponding to a stationary sound source. This effect is known as the visual capture or ventriloquist effect, which essentially implies that the perceived spatial location of a sound source may be shifted from its physical location due to concurrent visual stimuli from a different point in space [1]–[2]. While this effect has mainly been studied for scenarios in which static auditory and visual stimuli originate from different azimuthal positions, it has also been found to occur for moving visual stimuli complemented with static auditory stimuli [3],[4] as well as for static auditory and visual stimuli produced at different distances in the same azimuthal direction [5],[6],[7]. Despite the visual capture effect, there are several motivations to aim for a fully cross-modal (i.e., visual and auditory) stimulation when performing a zoom operation in audiovisual applications. First of all, the supposition that the occurrence of congruent visual and auditory stimuli facilitates audiovisual multisensory integration is supported by results from psychophysics [8],[9] as well as from brain imaging [10],[11]. Secondly, cross-modal stimulation has been shown to result in a more robust perception [12],[2], which is particularly relevant if one of both stimuli is corrupted or distorted. Finally, the area of multimodal signal processing has recently reached a degree of maturity that allows research results to be applied to a wide range of audio engineering problems [13].

An AZ effect is based on the alteration of one or more acoustic cues that are related to the auditory perception of sound source distance. It is well known that auditory distance perception is primarily determined by the following

auditory cues (see [14] and references therein): intensity, direct-to-reverberant ratio (DRR), spectral distortion, interaural time and level difference (ITD/ILD), and motion-induced intensity rate of change. Moreover, it was found in [15] that intensity is the dominating auditory cue for distance perception in case the sound source produces a speech signal or in case the sound source is located in the median plane (i.e., the plane defined by a zero azimuthal angle). This observation explains why most existing approaches to AZ essentially aim at a sound level control of the sound source in front of the camera, relative to the sound level of the sources outside the median plane. Perhaps the first approach to AZ was proposed in [16], where it was rightfully pointed out that increasing the frontal directivity of a microphone (array) allows for increasing both the intensity and the DRR of the sound source in front of the camera. A simple analog implementation was proposed in which the directivity of a three-microphone array can be changed continuously from omnidirectional through cardiod to second-order pressure-gradient unidirectional. The perceived AZ effect of this approach was further improved in [17] by adding a fourth microphone to allow for a stereophonic omnidirectional sound capture (when fully zoomed out), which contrasts even more with the second-order pressure-gradient unidirectional output at full zoom-in. Yet another improvement to this approach was recently proposed in [18], where not only the frontal directivity but also the response in other directions (in particular the back direction) is constrained to change smoothly and monotonically along a zoom operation. A fundamentally different approach to AZ was taken in [19], and is based on the use of a particular spherical microphone array that allows for decomposing the sound field into so-called eigenbeams, i.e., spatially orthonormal spherical harmonics that can be combined to approximate any desired output beam pattern (see also [20] for more details). Finally, an AZ algorithm for use in a directional audio coding (DirAC) framework was proposed in [21]. It was shown that the DirAC framework, in which a multichannel audio stream is represented in the time-frequency domain as a single-channel audio stream complemented with side information (more particularly direction of arrival (DOA) and diffuseness), is particularly suited for AZ since the DOA and diffuseness can be controlled directly, and spatial filtering of the single-channel stream can be implemented straightforwardly by short-time spectral attenuation.

While the most recent approaches to AZ based on eigenbeamforming and DirAC are believed to deliver promising results, the corresponding requirements in terms of hardware and computational resources are currently too high for consumer applications. Particularly, the eigenbeamforming approach requires an array with a large number of microphones (e.g., a 24-microphone array was used in [20]), while the DirAC approach operates on so-called B-format microphone recordings obtained with a dedicated microphone array and requires considerable computational resources to execute the time-frequency analysis, processing, and synthesis. The other approaches to AZ in [16],[17] are based on linearly combining a number of signals that cover different spatial areas of the sound scene, using time-varying weights controlled by the status of the zoom operation. In the approach proposed in this paper, we will adopt this linear combination concept, which has two distinct advantages compared to the most recent AZ approaches in [19],[21]. First, it has a moderate computational complexity: given the signals to be mixed, only scalar signal operations are required in their linear combination. Second, it does not require professional or dedicated hardware: the experimental results in this paper are obtained with an array of only three low-cost microphones mounted externally on a mobile phone device.

We will now present the structure of the paper, thereby emphasizing its main contributions. In Section 1, we briefly introduce the mathematical signal model used throughout the paper. In Section 2, we propose a theory that generalizes the linear-combination-based AZ approaches in [16],[17]. We will show that we can independently control the sound levels of an arbitrary number of sources in the sound scene by linearly combining a reference microphone signal with a number of so-called auxiliary signals. These auxiliary signals are themselves linear combinations of the different sound source signals, in which each of the sound source levels has been reweighted as compared to the reference microphone signal. We should stress that the theory for independent sound level control proposed here, is also applicable to interactive sound scene manipulation problems other than AZ, as it explicitly allows the user to set the desired sound source levels. In Section 3, we propose a three-block structure in which the auxiliary signals required for independent sound source level control are calculated. Each of the three functional blocks in the proposed structure is aimed towards reweighting the levels of one particular category of sound sources. Section 4 deals with the design of the spatial and spectral filters operating in each of the three functional blocks mentioned above. In contrast to the analog implementations of the linear-combination-based AZ approach in [16],[17], we fully exploit the strength of digital signal processing (DSP) in the design of the spatial and spectral filters by building upon recent results in DSP-based noise reduction and speech enhancement. In Section 5, we will present experimental results illustrating the effectiveness of the proposed AZ algorithm, and finally Section 6 concludes the paper.

1 MATHEMATICAL SIGNAL MODEL

The sound scene in the acoustic environment in which the audiovisual capture is performed, is assumed to consist of four contributions coming from a different azimuthal direction θ:

1. a front sound source $s_F(t)$, which is considered to be the desired sound source and is located in front of the camera at an angle $\theta_F = 0^\circ$ (by definition),
2. a back sound source $s_B(t)$, which may or may not be a desired sound source, corresponding to the sound produced by the camera operator (if any) at an angle $\theta_B = 180^\circ$,

3. a number of localized interfering sound sources $\{s_I^{(i)}(t)\}_{i=1}^{N_I}$, which are considered to be undesired and originate from directions $\theta_I^{(i)}$ different from the front and back directions,
4. a diffuse noise field, that cannot be attributed to a single sound source or direction, and which is considered to be undesired.

In the sequel, we will consistently use the subscripts F, B, I, and N to denote front, back, localized interfering, and diffuse noise signals, respectively. All sources are assumed to be located in the far field of the capture device, and at a zero elevation angle. Moreover, we will usually assume that $N_I = 1$ and omit the index i in $s_I(t)$. We should stress that the proposed approach can be generalized without difficulty to a three-dimensional sound scene comprising sound sources at nonzero elevation angle and to scenarios involving multiple localized interfering sound sources.

We further assume that the capture device is equipped with N microphones ($N \geq 2$). Each of the microphones may have a different frequency- and angle-dependent response, as defined by

$$A_n(\omega, \theta) = a_n(\omega, \theta)e^{-j\psi_n(\omega,\theta)}, \quad n = 0, \ldots, N-1. \tag{1}$$

The acoustic response (including the effect of the direct path time delay as well as reverberation) of a source at angle θ to each of the microphones is given by

$$F_n(\omega, \theta) = f_n(\omega, \theta)e^{-j\xi_n(\omega,\theta)}, \quad n = 0, \ldots, N-1. \tag{2}$$

For ease of notation, we introduce the joint acoustic and microphone response, defined as:

$$G_n(\omega, \theta) = A_n(\omega, \theta)F_n(\omega, \theta), \quad n = 0, \ldots, N-1. \tag{3}$$

Using the above definitions, we can express each of the N microphone signals $U_n(\omega)$ as a function of the four sound scene contributions in the frequency domain as follows,

$$\begin{aligned} U_n(\omega) = &\underbrace{G_n(\omega, \theta_F)S_F(\omega)}_{U_n^{(F)}(\omega)} + \underbrace{G_n(\omega, \theta_B)S_B(\omega)}_{U_n^{(B)}(\omega)} \\ &+ \underbrace{G_n(\omega, \theta_I)S_I(\omega)}_{U_n^{(I)}(\omega)} + U_n^{(N)}(\omega), \quad n = 0, \ldots, N-1 \end{aligned} \tag{4}$$

where $U_n^{(N)}(\omega)$ denotes the diffuse noise component. In the sequel, we consider the microphone with index $n = 0$ to be the reference microphone. By defining the microphone signal vector, the steering vector, and the diffuse noise component vector as follows:

$$\mathbf{U}(\omega) = [U_0(\omega) \ldots U_{N-1}(\omega)]^T \tag{5}$$

$$\mathbf{G}(\omega, \theta) = [G_0(\omega, \theta) \ldots G_{N-1}(\omega, \theta)]^T \tag{6}$$

$$\mathbf{U}^{(N)}(\omega) = \left[U_0^{(N)}(\omega) \ldots U_{N-1}^{(N)}(\omega)\right]^T \tag{7}$$

we can rewrite (4) in matrix notation,

$$\begin{aligned} \mathbf{U}(\omega) = &\underbrace{\mathbf{G}(\omega, \theta_F)S_F(\omega)}_{\mathbf{U}^{(F)}(\omega)} + \underbrace{\mathbf{G}(\omega, \theta_B)S_B(\omega)}_{\mathbf{U}^{(B)}(\omega)} \\ &+ \underbrace{\mathbf{G}(\omega, \theta_I)S_I(\omega)}_{\mathbf{U}^{(I)}(\omega)} + \mathbf{U}^{(N)}(\omega). \end{aligned} \tag{8}$$

2 INDEPENDENT SOUND SOURCE LEVEL CONTROL BY LINEAR COMBINATION

As discussed in the Introduction, an essential feature of any approach to AZ is the capability of controlling the front sound source level, relative to the level of other sound sources in the scene. For this reason, we explicitly formulate the AZ problem as a sound source level control problem. It will become clear later on, however, that due to the particular implementation of the sound level control method (which typically involves spatial filtering, see Section 3), a DRR increase is obtained as a beneficial side effect.

Consider the following sound level control problem: if the nth microphone signal $u_n(t)$ is decomposed in the time domain following (4),

$$u_n(t) = u_n^{(F)}(t) + u_n^{(B)}(t) + u_n^{(I)}(t) + u_n^{(N)}(t) \tag{9}$$

then the corresponding desired output of the AZ algorithm can be written as follows:

$$\begin{aligned} \zeta_n(t) = &g_F(t)u_n^{(F)}(t) + g_B(t)u_n^{(B)}(t) \\ &+ g_I(t)u_n^{(I)}(t) + g_N(t)u_n^{(N)}(t) \end{aligned} \tag{10}$$

where $g_F(t)$, $g_B(t)$, $g_I(t)$, and $g_N(t)$ denote the user-specified time-varying gain factors for the different sound source components (to be specified later on). Note that one is not necessarily interested in calculating N output signals of the AZ algorithm. Typically, the focus is on obtaining a mono or stereo output, which implies that the relation in (10) only needs to be considered for one or two particular values of n, say n_1 (and n_2). Nevertheless, all N microphone signals will typically be used to obtain an estimate of the signals $\zeta_{n_1}(t)$ and $\zeta_{n_2}(t)$. Also note that we have not included the microphone index n in the notation of the gain factors in (10), since typically the same gain factors will be used for the different output signals of the AZ algorithm.

Intuitively, it could be expected that the AZ algorithm needs to perform some kind of source separation to isolate the different sound source components. However, since we are not interested in the separated sound source components, but rather in a mixture in which these components have been "reweighted" as compared to the microphone signals, an explicit source separation is not required. As a consequence, one of the paradigms of source separation, which states that the number of microphones needs to exceed the number of sources to be separated, can be evaded by replacing some of the microphone signals by auxiliary signals. Suppose that we are able to generate three auxiliary signals $x_n(t)$, $y_n(t)$, and $z_n(t)$, in which the different sound

source components have been arbitrarily reweighted compared to the corresponding components in the microphone signal $u_n(t)$, i.e.,

$$x_n(t) = \gamma^{(F)}_{x_n,u_n} u_n^{(F)}(t) + \gamma^{(B)}_{x_n,u_n} u_n^{(B)}(t) + \gamma^{(I)}_{x_n,u_n} u_n^{(I)}(t) + \gamma^{(N)}_{x_n,u_n} u_n^{(N)}(t) \quad (11)$$

$$y_n(t) = \gamma^{(F)}_{y_n,u_n} u_n^{(F)}(t) + \gamma^{(B)}_{y_n,u_n} u_n^{(B)}(t) + \gamma^{(I)}_{y_n,u_n} u_n^{(I)}(t) + \gamma^{(N)}_{y_n,u_n} u_n^{(N)}(t) \quad (12)$$

$$z_n(t) = \gamma^{(F)}_{z_n,u_n} u_n^{(F)}(t) + \gamma^{(B)}_{z_n,u_n} u_n^{(B)}(t) + \gamma^{(I)}_{z_n,u_n} u_n^{(I)}(t) + \gamma^{(N)}_{z_n,u_n} u_n^{(N)}(t). \quad (13)$$

Here, each of the reweighting factors is by definition equal to the square root of the power ratio of the corresponding sound source components, e.g.,

$$\gamma^{(F)}_{x_n,u_n} = \frac{\sigma_{x_n^{(F)}}}{\sigma_{u_n^{(F)}}} = \sqrt{\frac{E\left\{\left(x_n^{(F)}\right)^2\right\}}{E\left\{\left(u_n^{(F)}\right)^2\right\}}} \quad (14)$$

and similarly for the other reweighting factors. Note that the auxiliary signals, as defined in (11)–(13), cannot be calculated as such, since the individual sound source components $u_n^{(\cdot)}(t)$ are not available. However, in Section 3, we will propose a scheme that enables to compute the auxiliary signals directly from the microphone signals.

The output signal of the AZ algorithm can now be calculated as a linear combination of the nth microphone signal and the auxiliary signals $x_n(t)$, $y_n(t)$, and $z_n(t)$ defined above, i.e.,

$$\zeta_n(t) = a_n^{(0)}(t) u_n(t) + a_n^{(1)}(t) x_n(t) + a_n^{(2)}(t) y_n(t) + a_n^{(3)}(t) z_n(t). \quad (15)$$

By using the relations in (9) and (11)–(13), the expression for the calculated AZ output signal in (15) can be shown to be equivalent to the expression for the desired AZ output signal in (10) if the weights $a_n^{(m)}(t),\ m = 0, \ldots, 3$ satisfy the following relationship:

$$\begin{bmatrix} a_n^{(0)}(t) \\ a_n^{(1)}(t) \\ a_n^{(2)}(t) \\ a_n^{(3)}(t) \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & \gamma^{(F)}_{x_n,u_n} & \gamma^{(F)}_{y_n,u_n} & \gamma^{(F)}_{z_n,u_n} \\ 1 & \gamma^{(B)}_{x_n,u_n} & \gamma^{(B)}_{y_n,u_n} & \gamma^{(B)}_{z_n,u_n} \\ 1 & \gamma^{(I)}_{x_n,u_n} & \gamma^{(I)}_{y_n,u_n} & \gamma^{(I)}_{z_n,u_n} \\ 1 & \gamma^{(N)}_{x_n,u_n} & \gamma^{(N)}_{y_n,u_n} & \gamma^{(N)}_{z_n,u_n} \end{bmatrix}^{-1}}_{\mathbf{\Gamma}^{-1}} \begin{bmatrix} g_F(t) \\ g_B(t) \\ g_I(t) \\ g_N(t) \end{bmatrix} \quad (16)$$

This implies that a unique set of weight trajectories $a_n^{(m)}(t),\ m = 0, \ldots, 3,\ \forall t$ can be calculated that *exactly* produces the desired AZ effect, under the following two conditions:

1. the reweighting matrix $\mathbf{\Gamma}$ is of full rank,
2. the reweighting factors $\gamma^{(\cdot)}_{\cdot,\cdot}$ are known.

The first condition either requires the microphone signal $u_n(t)$ and the auxiliary signals $x_n(t)$, $y_n(t)$, and $z_n(t)$ to be linearly independent (which leads to linearly independent columns in $\mathbf{\Gamma}$) or requires the reweighting of the different sound source components in each of the auxiliary signals $x_n(t)$, $y_n(t)$, and $z_n(t)$ to be linearly independent (which leads to linearly independent rows in $\mathbf{\Gamma}$). As for the second condition, a method for calculating the reweighting factors will be discussed in Section 3.

A remaining issue is how the time-varying gain factors $g_F(t)$, $g_B(t)$, $g_I(t)$, and $g_N(t)$ should be specified to obtain the AZ effect, and how these factors can be related to the status of the camera zoom lens. This issue has been thoroughly investigated in [17] and [21], and therefore will not be treated in detail here. However, we mention two general guidelines for specifying these functions. As a first guideline, one should discriminate between sound sources of which the visually perceived position is altered by the video zoom operation, and sound sources for which this is not the case. For the first category of sound sources, a monotonically increasing gain factor should be specified as the perceived distance to the camera decreases, and vice versa. For the second category, a constant gain factor would typically be specified, as there is no visual information that suggests any movement of these sound sources. However, if the aim is to create the illusion of motion for sound sources outside the scope of the video camera as well, a time-varying gain factor may be specified instead. An example of this approach is to assign a monotonically decreasing gain factor to the back sound source while the camera zooms in on the front sound source, to create the illusion that the camera is effectively moving away from the back sound source while approaching the front sound source. A second guideline relates to the specification of the monotonically increasing gain function of a sound source that is perceived to approach the camera. From a physical point of view, the intensity of an approaching sound source is increased by 6 dB as its distance to the observer is halved. However, it was found in [17] and [21] that such a physically motivated relation does not necessarily result in a perceptually attractive AZ effect.

3 BLOCK STRUCTURE FOR AUXILIARY SIGNAL CALCULATION

We will now propose a block structure for the calculation of the auxiliary signals $x_n(t)$, $y_n(t)$, and $z_n(t)$ required in the AZ algorithm. Initially, we will focus on obtaining a mono AZ output signal, where (15) is evaluated for the reference microphone ($n = 0$) only. Afterward, we will show how this approach can be straightforwardly extended to the stereo case.

The block structure in Fig. 1 consists of three functional blocks:

1. *Fixed beamformer:* the aim of this block is to perform a reweighting of the sound source components of which the source direction is known a priori, i.e., the front and back sound source components. The power ratios of these components should be altered

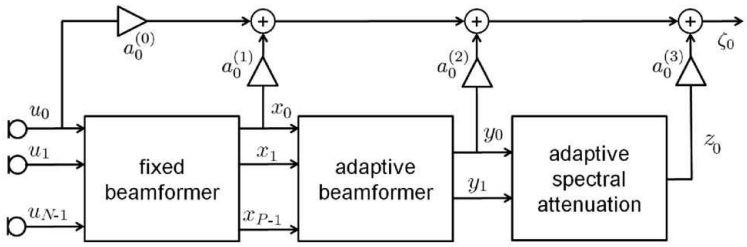


Fig. 1. Block structure for mono AZ algorithm.

by the fixed beamformer, both relative to each other and relative to the other sound source components.

2. *Adaptive beamformer:* this block serves to perform a reweighting of the localized interfering sound source component(s), which necessarily requires an adaptive beamforming algorithm since the interfering sound source direction is unknown.
3. *Adaptive spectral attenuation:* using this block we aim at a reweighting of the diffuse noise field component, by exploiting its assumed spectral diversity relative to the localized sound source components.

Note that in the mono output case, some of the auxiliary signals (more particularly $x_n(t)$ and $y_n(t)$ with $n > 0$) are not explicitly used to calculate the AZ output signal $\zeta_0(t)$. However, these signals are used internally in the adaptive beamformer and adaptive spectral attenuation algorithms. More particularly, the signals $x_n(t), n = 1, \ldots, P-1$ at the output of the fixed beamformer will be constructed to be "noise reference signals" (see Section 4.1), i.e., signals in which the desired (front and possibly back) sound sources have been suppressed and that are used subsequently in the adaptive beamformer to estimate the localized interfering sound source component in the primary output signal $x_0(t)$ of the fixed beamformer. The signal $y_1(t)$ is then constructed to be a "diffuse noise reference" that is used by the adaptive spectral attenuation algorithm to estimate the diffuse noise component in the primary output signal $y_0(t)$ of the adaptive beamformer.

Because of the above discrimination between the primary output signals $x_0(t)$ and $y_0(t)$, and the other beamformer output signals $x_n(t)$ and $y_n(t)$ with $n > 0$, a stereo AZ output signal should preferably not be created by calculating $\zeta_0(t)$ and $\zeta_1(t)$ using (15) for $n = 0$ and $n = 1$. Instead, we propose the block structure shown in Fig. 2 for the stereo case. Here, the stereo AZ output signals are calculated as follows:

$$\zeta_0(t) = a_0^{(0)}(t)u_0(t) + a_0^{(1)}(t)x_0(t) + a_0^{(2)}(t)y_0(t) + a_0^{(3)}(t)z_0(t) \tag{17}$$

$$\zeta_1(t) = a_1^{(0)}(t)u_1(t) + a_1^{(1)}(t)x_0(t) + a_1^{(2)}(t)y_0(t) + a_1^{(3)}(t)z_0(t) \tag{18}$$

Obviously, in case $N > 2$, one should select for $u_0(t)$ and $u_1(t)$ the two microphone signals that are best suited to deliver a stereo image. Note that, due to the particular structure in Fig. 2, the AZ weight calculation for the secondary output signal $\zeta_1(t)$ should be slightly altered into

$$\begin{bmatrix} a_1^{(0)}(t) \\ a_1^{(1)}(t) \\ a_1^{(2)}(t) \\ a_1^{(3)}(t) \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & \gamma_{x_0,u_1}^{(F)} & \gamma_{y_0,u_1}^{(F)} & \gamma_{z_0,u_1}^{(F)} \\ 1 & \gamma_{x_0,u_1}^{(B)} & \gamma_{y_0,u_1}^{(B)} & \gamma_{z_0,u_1}^{(B)} \\ 1 & \gamma_{x_0,u_1}^{(I)} & \gamma_{y_0,u_1}^{(I)} & \gamma_{z_0,u_1}^{(I)} \\ 1 & \gamma_{x_0,u_1}^{(N)} & \gamma_{y_0,u_1}^{(N)} & \gamma_{z_0,u_1}^{(N)} \end{bmatrix}^{-1}}_{\boldsymbol{\Gamma}^{-1}} \begin{bmatrix} g_F(t) \\ g_B(t) \\ g_I(t) \\ g_N(t) \end{bmatrix} \tag{19}$$

while the weights for the primary output signal $\zeta_0(t)$ can be calculated using (16) with $n = 0$.

As already mentioned in Section 2, the AZ weight calculation in (16) and (19) requires the knowledge of the reweighting factors $\gamma_{\cdot,\cdot}^{(\cdot)}$. We will now illustrate how the reweighting factors can be calculated if the above AZ block structure is used to generate the auxiliary signals.

The input-output relation of the three functional blocks in the AZ block structure can be described in the frequency domain as follows. The fixed beamformer is assumed to be composed of P filter-and-sum beamformers such that it can be specified by an $N \times P$ transfer function matrix $\mathbf{W}_1(\omega)$, i.e.,

$$\mathbf{X}(\omega) = \mathbf{W}_1^H(\omega)\mathbf{U}(\omega) \tag{20}$$

where

$$\mathbf{X}(\omega) = [X_0(\omega) \ldots X_{P-1}(\omega)]^T \tag{21}$$

$$\mathbf{W}_1(\omega) = \begin{bmatrix} W_{1,(0,0)}(\omega) & \ldots & W_{1,(0,P-1)}(\omega) \\ \vdots & \ddots & \vdots \\ W_{1,(P-1,0)}(\omega) & \ldots & W_{1,(P-1,P-1)}(\omega) \end{bmatrix} \tag{22}$$

and $\mathbf{U}(\omega)$ was defined in (5). The adaptive beamformer can be specified by a $P \times 1$ transfer function vector $\mathbf{W}_2(\omega)$ that defines the relation between the adaptive beamformer input and its primary output signal, i.e.,

$$Y_0(\omega) = \mathbf{W}_2^H(\omega)\mathbf{X}(\omega) \tag{23}$$

where

$$\mathbf{W}_2(\omega) = \left[W_{2,(0)}(\omega) \ldots W_{2,(P-1)}(\omega) \right]^T. \tag{24}$$

As explained earlier, the adaptive beamformer secondary output signal should ideally be an estimate of the diffuse noise component in the adaptive beamformer primary output signal. The most straightforward approach is to choose

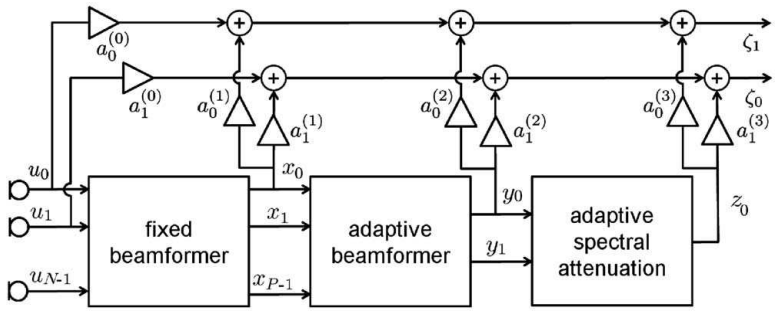


Fig. 2. Block structure for stereo AZ algorithm.

the secondary output signal to be equal to one of the noise references at the output of the fixed beamformer, e.g.,

$$Y_1(\omega) = X_1(\omega). \tag{25}$$

Alternatively, one could attempt to remove the localized interfering sound source component from the adaptive beamformer secondary output signal, however, this approach is not explicated here. The adaptive spectral attenuation can finally be specified using a scalar transfer function $W_3(\omega)$, i.e.,

$$Z_0(\omega) = W_3(\omega)Y_0(\omega). \tag{26}$$

Using the above input-output relations, we can derive expressions for the different localized sound source components in the primary auxiliary signals $X_0(\omega)$, $Y_0(\omega)$, and $Z_0(\omega)$ as a function of the corresponding dry sound source signals $S_F(\omega)$, $S_B(\omega)$, and $S_I(\omega)$,

$$X_0^{(c)}(\omega) = \mathbf{e}_1^T \mathbf{W}_1^H(\omega)\mathbf{G}(\omega, \theta_c)S_c(\omega) \tag{27}$$

$$Y_0^{(c)}(\omega) = \mathbf{W}_2^H(\omega)\mathbf{W}_1^H(\omega)\mathbf{G}(\omega, \theta_c)S_c(\omega) \tag{28}$$

$$Z_0^{(c)}(\omega) = W_3(\omega)\mathbf{W}_2^H(\omega)\mathbf{W}_1^H(\omega)\mathbf{G}(\omega, \theta_c)S_c(\omega) \tag{29}$$

where c represents the component F, B, or I, and $\mathbf{e}_1^T \triangleq [1, 0, \ldots, 0]_{1\times P}$. Similarly, the diffuse noise component in the primary auxiliary signals can be expressed as a function of the diffuse noise components in the microphone signals,

$$X_0^{(N)}(\omega) = \mathbf{e}_1^T \mathbf{W}_1^H(\omega)\mathbf{U}^{(N)}(\omega) \tag{30}$$

$$Y_0^{(N)}(\omega) = \mathbf{W}_2^H(\omega)\mathbf{W}_1^H(\omega)\mathbf{U}^{(N)}(\omega) \tag{31}$$

$$Z_0^{(N)}(\omega) = W_3(\omega)\mathbf{W}_2^H(\omega)\mathbf{W}_1^H(\omega)\mathbf{U}^{(N)}(\omega). \tag{32}$$

We will now make the following assumptions, with the aim of simplifying the calculation of the reweighting factors $\gamma_{\cdot,\cdot}^{(\cdot)}$:

1. the diffuse noise spectrum does not overlap with the spectra of the different localized sound sources,

$$\nexists\omega : S_c(\omega) \neq 0 \,\&\, U_n^{(N)}(\omega) \neq 0, \quad n = 0, \ldots, N-1,\ c = F, B, I \tag{33}$$

2. the diffuse noise power in each of the microphone signals is equal,

$$\sigma^2_{u_0^{(N)}} = \ldots = \sigma^2_{u_{N-1}^{(N)}} \tag{34}$$

3. the joint acoustic and microphone responses have a constant magnitude within the bandwidth and in the direction of the different sound source components, i.e.,

$$\begin{aligned} \forall\omega :\quad & S_c(\omega) \neq 0, U_n^{(N)}(\omega) \neq 0 \\ & \Rightarrow |G_n(\omega, \theta_c)| \equiv |G_n(\theta_c)|, \\ & n = 0, \ldots, N-1,\ c = F, B, I \end{aligned} \tag{35}$$

4. the fixed and adaptive beamformers have a constant magnitude within the bandwidth and in the direction of the different sound source components, i.e.,

$$\begin{aligned} & \forall\omega : S_c(\omega) \neq 0, U_n^{(N)}(\omega) \neq 0 \\ & \Rightarrow \begin{cases} |W_{1,(m,n)}(\omega)| \equiv |W_{1,(m,n)}|, \\ |W_{2,(n)}(\omega)| \equiv |W_{2,(n)}|, \end{cases} \\ & m = 0, \ldots, N-1, n = 0, \ldots, P-1, \\ & c = F, B, I \end{aligned} \tag{36}$$

5. the adaptive spectral attenuation magnitude response is constant within the bandwidth of the different sound source components,

$$\forall\omega : S_c(\omega) \neq 0 \Rightarrow |W_3(\omega)| \equiv |W_3^{(c)}|, \quad c = F, B, I \tag{37}$$

$$\forall\omega : U_n^{(N)}(\omega) \neq 0 \Rightarrow |W_3(\omega)| \equiv |W_3^{(N)}|, \quad n = 0, \ldots, N-1. \tag{38}$$

Under these assumptions, the signal powers of the different sound source components in the microphone and auxiliary signals can be estimated as follows:

$$\sigma^2_{u_n^{(c)}} = |G_n(\theta_c)|^2\sigma^2_{s_c}, \quad n = 0, \ldots, N-1,\ c = F, B, I$$

$$\sigma^2_{x_0^{(c)}} = |\mathbf{e}_1^T \mathbf{W}_1^H \mathbf{G}(\theta_c)|^2\sigma^2_{s_c}, \quad c = F, B, I \tag{39}$$

$$\sigma^2_{y_0^{(c)}} = |\mathbf{W}_2^H \mathbf{W}_1^H \mathbf{G}(\theta_c)|^2\sigma^2_{s_c}, \quad c = F, B, I \tag{40}$$

$$\sigma^2_{z_0^{(c)}} = |W_3^{(c)}|^2 |\mathbf{W}_2^H \mathbf{W}_1^H \mathbf{G}(\theta_c)|^2 \sigma^2_{s_c}, \quad c = F, B, I \tag{41}$$

$$\sigma^2_{x_0^{(N)}} = \|\mathbf{W}_1 \mathbf{e}_1\|_2^2 \sigma^2_{u_0^{(N)}} \tag{42}$$

$$\sigma^2_{y_0^{(N)}} = \|\mathbf{W}_1 \mathbf{W}_2\|_2^2 \sigma^2_{u_0^{(N)}} \tag{43}$$

$$\sigma^2_{x_0^{(N)}} = |W_3^{(N)}|^2 \|\mathbf{W}_1 \mathbf{W}_2\|_2^2 \sigma^2_{u_0^{(N)}} \tag{44}$$

and consequently, the reweighting factors can be calculated as:

$$\gamma^{(c)}_{x_0,u_n} = \frac{|\mathbf{e}_1^T \mathbf{W}_1^H \mathbf{G}(\theta_c)|}{|G_n(\theta_c)|}, \quad n = 0, \ldots, N-1, \ c = F, B, I \tag{45}$$

$$\gamma^{(c)}_{y_0,u_n} = \frac{|\mathbf{W}_2^H \mathbf{W}_1^H \mathbf{G}(\theta_c)|}{|G_n(\theta_c)|}, \quad n = 0, \ldots, N-1, \ c = F, B, I \tag{46}$$

$$\gamma^{(c)}_{z_0,u_n} = \frac{|W_3^{(c)}||\mathbf{W}_2^H \mathbf{W}_1^H \mathbf{G}(\theta_c)|}{|G_n(\theta_c)|}, \quad n = 0, \ldots, N-1, \ c = F, B, I \tag{47}$$

$$\gamma^{(N)}_{x_0,u_n} = \|\mathbf{W}_1 \mathbf{e}_1\|_2, \quad n = 0, \ldots, N-1 \tag{48}$$

$$\gamma^{(N)}_{y_0,u_n} = \|\mathbf{W}_1 \mathbf{W}_2\|_2, \quad n = 0, \ldots, N-1 \tag{49}$$

$$\gamma^{(N)}_{z_0,u_n} = |W_3^{(N)}| \|\mathbf{W}_1 \mathbf{W}_2\|_2, \quad n = 0, \ldots, N-1 \tag{50}$$

Two important statements should be specified here. First, it is clear that the assumptions in (33)–(38) are too restrictive for most realistic sound scenes. We could relax these assumptions by "upgrading" the linear combination approach in (15) to a filter-and-sum approach (i.e., by replacing the multipliers in Figs. 1 and 2 with filters), and considering frequency-dependent reweighting factors $\gamma^{(\cdot)}_{\cdot,\cdot}(\omega)$. However, we explicitly choose not to do so, with the aim of keeping the proposed AZ algorithm conceptually and computationally simple. Second, we should note that from a computational point of view, calculating the reweighting factors online using the relations in (45)–(50) appears not to be an attractive option. A cheaper approach consists in setting the values of the reweighting factors off-line, making use of the fixed beamformer response (known a priori) and of heuristics about the behavior of the adaptive beamformer and spectral attenuation response (see Section 5).

4 SPATIAL AND SPECTRAL FILTER DESIGN

4.1 Fixed Beamformer Design

As explained in Section 3, the fixed beamformer creates a primary output signal $X_0(\omega)$ that spatially enhances the front sound source signal, as well as $P-1$ other output signals $X_n(\omega), n = 1, \ldots, P-1$ that serve as "noise references" for the adaptive beamformer. This is achieved by implementing the fixed beamformer as a parallel combination of P filter-and-sum beamformers. We will first discuss the design of the first, so-called front source beamformer (FSB), and afterwards we will explain the design of the $P-1$ remaining beamformers, also denoted as the blocking matrix (BM).

Depending on the kind of spatial enhancement one wants to achieve for the front sound source, different fixed beamformer design methods may be employed for the FSB, e.g., an array pattern synthesis approach [22],[23] or a differential [24] or superdirective [25] design method. Here, we will adopt a superdirective (SD) design method, which is recommendable when the aim is to maximize the directivity factor of the microphone array, i.e., to maximize the array gain in the presence of a diffuse noise field. The frequency-domain SD design equation for the FSB is specified following Doclo and Moonen [25],

$$\mathbf{W}_1(\omega)\mathbf{e}_1 = \frac{\left(\tilde{\Phi}_{\mathbf{U}}^{(N)} + \delta \mathbf{I}_N\right)^{-1} \mathbf{G}(\omega, \theta_F)}{\mathbf{G}^H(\omega, \theta_F)\left(\tilde{\Phi}_{\mathbf{U}}^{(N)} + \delta \mathbf{I}_N\right)^{-1} \mathbf{G}(\omega, \theta_F)} \tag{51}$$

where $\mathbf{G}(\omega, \theta_F)$ denotes the front sound source steering vector as defined in (6), $\mathbf{I}_N$ represents the $N \times N$ identity matrix, δ is a regularization parameter, and $\tilde{\Phi}_{\mathbf{U}}^{(N)}$ denotes the normalized diffuse noise correlation matrix, which can be calculated from the joint acoustic and microphone responses as follows:

$$\tilde{\Phi}_{\mathbf{U}}^{(N)} = \begin{bmatrix} \tilde{\Phi}^{(N)}_{U_0,U_0} & \cdots & \tilde{\Phi}^{(N)}_{U_0,U_{N-1}} \\ \vdots & \ddots & \vdots \\ \tilde{\Phi}^{(N)}_{U_{N-1},U_0} & \cdots & \tilde{\Phi}^{(N)}_{U_{N-1},U_{N-1}} \end{bmatrix} \tag{52}$$

with

$$\tilde{\Phi}^{(N)}_{U_m,U_n} = \frac{1}{2\pi} \int_0^{2\pi} G_m(\omega, \theta) G_n^*(\omega, \theta) d\theta. \tag{53}$$

The BM in the fixed beamformer consists of $P-1$ filter-and-sum beamformers that each operate on one particular subset of microphone signals. In this way, $P-1$ noise reference signals are created, in which the power of the desired signal components is maximally reduced as compared to the power of these components in the microphone signals. The required number of noise references depends on the number of localized sound sources one aims to suppress in the adaptive beamformer primary output signal. Considering the signal model introduced in Section 1, we should choose $P - 1 = N_I$ to suppress all localized interfering sound sources, or $P - 1 = N_I + 1$ if the back sound source should also be suppressed by the adaptive beamformer. However, in case the bandwidth of the different localized sound sources does not (significantly) overlap, the actual required number of noise references may be reduced. In this way, the number of adaptive filters and hence the computational complexity of the adaptive beamformer can be reduced significantly. Since we have restricted our attention to the case $N_I = 1$, we will stick to the choice $P = 2$ (even when a back sound source is also present), thus employing a BM consisting of only one filter-and-sum beamformer designed using the complete set of available microphone signals.

In the context of the BM design, we consider the back sound source (if any) to be an undesired signal (which should be cancelled by the adaptive beamformer). It should hence be included in the noise references, and so the BM design reduces to a front-canceling beamformer (FCB) design. Again, one of several different fixed beamformer design methods can be employed. We propose an array pattern synthesis method, different from the existing methods in [22],[23]. In general, we can specify the frequency-domain FCB design at a set of angles $\theta = \{\theta_0, \ldots, \theta_{M-1}\}$ by the following linear system of equations:

$$\underbrace{\begin{bmatrix} G_0^*(\omega, \theta_0) & \ldots & G_{N-1}^*(\omega, \theta_0) \\ \vdots & \ddots & \vdots \\ G_0^*(\omega, \theta_{M-1}) & \ldots & G_{N-1}^*(\omega, \theta_{M-1}) \end{bmatrix}}_{\tilde{\mathbf{G}}^H(\omega)} \mathbf{W}_1(\omega)\mathbf{e}_2 = \underbrace{\begin{bmatrix} P_0^*(\omega) \\ \vdots \\ P_{M-1}^*(\omega) \end{bmatrix}}_{\mathbf{P}^*(\omega)} \tag{54}$$

where $P_m(\omega), m = 0, \ldots, M-1$ denotes the desired response at frequency ω and angle θ_m, and $\mathbf{e}_2^T \triangleq [0, 1, \ldots, 0]_{1\times P}$. The least-squares (LS) optimal solution (also having minimum norm in case $M < N$) is then given by

$$\mathbf{W}_1(\omega)\mathbf{e}_2 = \left(\tilde{\mathbf{G}}^H(\omega)\right)^\dagger \mathbf{P}^*(\omega) \tag{55}$$

where $(\cdot)\dagger$ denotes the Moore-Penrose inverse. More specifically, to obtain an FCB design we should specify a zero response in the front direction and a nonzero response in any other direction. Preferably the latter direction should be the back direction to avoid that the design would actually correspond to a front-back-canceling beamformer design. As a consequence, the number of equations in (54) is $M = 2$, and the specification angles correspond to $\theta_0 = \theta_F$ and $\theta_1 = \theta_B$. Finally, the desired response vector is equal to $\mathbf{P}^*(\omega) = [0, 1]^H$.

4.2 Adaptive Beamformer Design

The adaptive beamformer may be implemented using a generalized sidelobe canceller (GSC) algorithm [26], a multichannel Wiener filtering (MWF) algorithm [27]–[28], or another multi-channel adaptive algorithm. Here, we employ the speech-distortion-weighted multichannel Wiener filtering (SDW-MWF) algorithm [27],[29],[30]–[31] which includes the GSC and MWF as special cases. The objective of the SDW-MWF is to jointly minimize the energy of the undesired components (B, I, N) and the distortion of the desired component (F) in the enhanced signal $Y_0(\omega)$, i.e.,

$$\min_{\mathbf{W}_2(\omega)} E\left\{\left|\mathbf{W}_2^H(\omega)\left[\mathbf{X}^{(B)}(\omega) + \mathbf{X}^{(I)}(\omega) + \mathbf{X}^{(N)}(\omega)\right]\right|^2\right\} + \tfrac{1}{\mu} E\left\{\left|X_0^{(F)}(\omega) - \mathbf{W}_2^H(\omega)\mathbf{X}^{(F)}(\omega)\right|^2\right\} \tag{56}$$

resulting in the adaptive beamformer estimate

$$\mathbf{W}_2(\omega) = \left[\mathbf{\Phi}_x{}^{(F)}(\omega) + \mu\,\mathbf{\Phi}_x{}^{(B,I,N)}(\omega)\right]^{-1} \mathbf{\Phi}_x{}^{(F)}(\omega)\mathbf{e}_1 \tag{57}$$

where the correlation matrices of the desired and undesired components in the adaptive beamformer input signal are defined as:

$$\mathbf{\Phi}_x{}^{(F)}(\omega) = E\left\{\left[\mathbf{X}^{(F)}(\omega)\right]\left[\mathbf{X}^{(F)}(\omega)\right]^H\right\} \tag{58}$$

$$\mathbf{\Phi}_x{}^{(B,I,N)}(\omega) = E\left\{\left[\mathbf{X}^{(B)}(\omega) + \mathbf{X}^{(I)}(\omega) + \mathbf{X}^{(N)}(\omega)\right]\left[\mathbf{X}^{(B)}(\omega) + \mathbf{X}^{(I)}(\omega) + \mathbf{X}^{(N)}(\omega)\right]^H\right\}. \tag{59}$$

The parameter μ can be tuned to trade off energy reduction of the undesired components versus distortion of the desired component. Several recursive implementations of the SDW-MWF filter estimate in (57) have been proposed, in which the adaptive SDW-MWF filter update is based on a generalized singular value decomposition (GSVD) [27], a QR decomposition (QRD) [30], a time-domain stochastic gradient method [31], or a frequency-domain stochastic gradient method [29]. A common feature of these implementations is that the correlation matrices $\mathbf{\Phi}_x{}^{(F)}(\omega)$ and $\mathbf{\Phi}_x{}^{(B,I,N)}(\omega)$ are explicitly estimated before the SDW-MWF filter estimate in (57) is computed, by making use of a voice activity detector (VAD).

4.3 Adaptive Spectral Attenuation Design

The adaptive spectral attenuation block is included in the proposed AZ structure with the aim of reducing the diffuse noise energy in the adaptive beamformer primary output signal. To this end, the short-term magnitude spectra of the reference microphone signal, $|U_0(\omega, l)|$, and the adaptive beamformer primary and secondary output signals, $|Y_0(\omega, l)|$ and $|Y_1(\omega, l)|$, are estimated by means of a discrete Fourier transform (DFT), with l denoting the time frame index. An instantaneous spectral gain function is then calculated as follows, see, e.g., [32]–[33, Ch. 10]

$$G_{\text{inst}}(\omega, l) = \frac{|U_0(\omega, l)| - \beta_n C(\omega, l)|Y_1(\omega, l)|}{|Y_0(\omega, l)| + \varepsilon} \tag{60}$$

where the subtraction factor $\beta_n \in [0, 1]$ determines the amount of spectral attenuation and the regularization factor ε is a small constant which prevents division by zero. Since the adaptive beamformer secondary output signal $Y_1(\omega, l)$ is equal to the noise reference $X_1(\omega, l)$ at the output of the fixed beamformer (see (25)), a spectral coherence function $C(\omega, l)$ that relates the magnitude spectra of the diffuse noise components in the fixed beamformer primary and secondary output signals needs to be estimated and taken into account in (60). The instantaneous gain function in (60) is then lowpass filtered and clipped, before being applied to the speech estimate, i.e.,

$$G_{\text{lp}}(\omega, l) = (1 - \alpha)G_{\text{lp}}(\omega, l-1) + \alpha G_{\text{inst}}(\omega, l) \tag{61}$$

$$G(\omega, l) = \max\left\{G_{\text{lp}}(\omega, l), \xi_n\right\} \tag{62}$$

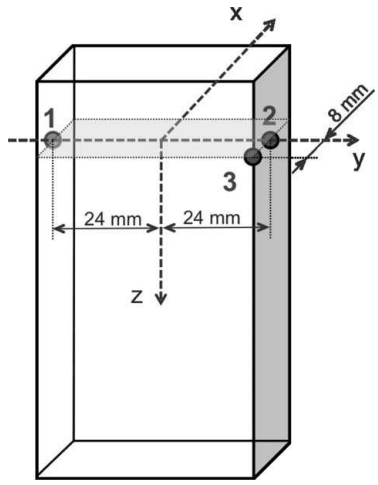


Fig. 3. Microphone positioning on mobile device used in audio recordings.

$$|Z(\omega, l)| = G(\omega, l)|Y_0(\omega, l)| \tag{63}$$

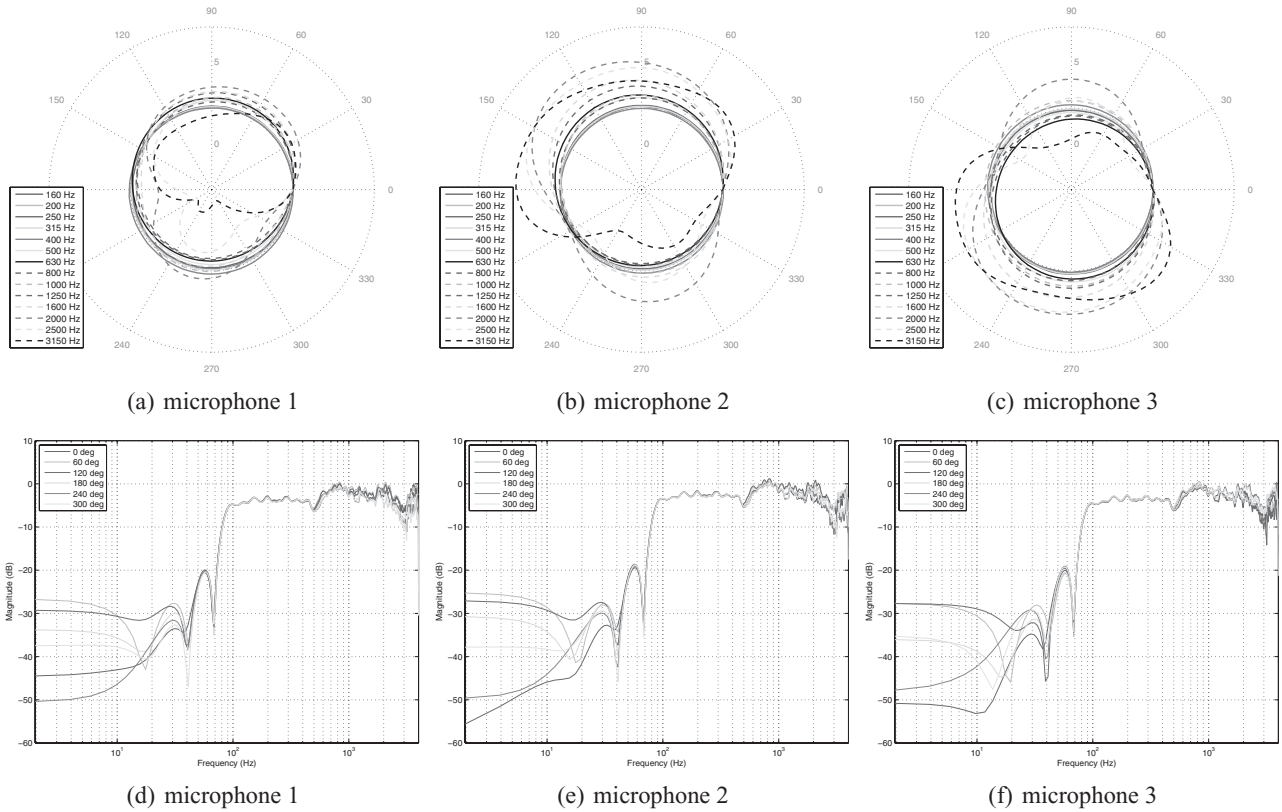
where α denotes the lowpass filter pole and $\xi_n = 1 - \beta_n$ is the clipping level. The enhanced signal magnitude spectrum $|Z(\omega, l)|$ is subsequently transformed back to the time domain by applying an inverse DFT (IDFT), and by using the phase spectrum of the adaptive beamformer primary output signal $Y_0(\omega, l)$.

5 EXPERIMENTAL EVALUATION

The AZ algorithm will be evaluated by discussing the experimental results obtained in two different sound scenes. With the aim of obtaining realistic yet reproducible results, the experiments are based on simulations that employ signals and impulse responses measured with a consumer device in a realistic acoustic environment. To this end, we have performed a number of recordings using a mobile phone with three low-cost microphones mounted externally. The microphones are positioned on the device as shown in Fig. 3, with the following 3-D coordinates (in mm): $(0, -24, 0)$ for microphone 1, $(0, 24, 0)$ for microphone 2, and $(-8, 24, 0)$ for microphone 3. The motivation for using a three-microphone array is that this is the minimal configuration that allows the combination of a broadside and an end-fire microphone pair. The broadside microphone pair (positions 1 and 2) is well suited for increasing the directivity towards the front sound source and for obtaining a stereo effect, while the end-fire pair (positions 2 and 3) allows for designing a BM having a broadband suppression in the front direction. The positive x-axis in Fig. 3 points by definition in a direction with azimuthal angle $\theta = 0$ and elevation angle $\phi = 0$. The mobile phone is approximately centered in the acoustic environment, which is a room of about 50 m^3 with a moderate reverberation time.

A first set of audio recordings consists of different types of diffuse noise signals, which were created by playing back a white noise signal with four loudspeakers pointing towards the corners of the room. A second set of audio recordings was created by playing back a white noise signal using a loudspeaker at a distance of 2 m from the device (and directed toward the device), and measuring the resulting microphone signals. Such recordings were obtained for 72 different azimuthal loudspeaker angles ($\theta = 0^\circ, 5^\circ, 10^\circ, \ldots, 355^\circ$). From these recordings, the joint acoustic and microphone responses of the device have been estimated, the results of which are shown in Figs. 4. Here, the spatial responses are calculated in fourteen 3-octave bands with center frequencies as indicated in the plot legends. The frequency responses are calculated for six azimuthal angles that are uniformly distributed in the horizontal plane. While the audio recordings are performed at a 48 kHz sampling rate, we have plotted the joint acoustic and microphone responses for a frequency range up to 4 kHz because we will eventually run the AZ algorithm at an 8 kHz sampling rate. The use of these estimated responses in the experimental evaluation is twofold. First of all, we have simulated all localized sound sources (i.e., front, back, and localized interfering sound sources) by filtering dry audio signals with one of the estimated responses (which obviously restricts the azimuthal angles θ_F, θ_B, and θ_I to be an integer multiple of 5°). Second, the estimated responses have been used in the fixed beamformer design, which will be discussed in more detail below. At first sight, this may appear to be an impractical decision since the fixed beamformer design would then depend on the exact knowledge of the response of the acoustic environment (i.e., the measurement room). However, by restricting the fixed beamformer length to a value ($L_{\mathrm{FSB}} = L_{\mathrm{BM}} = 64$ at an 8 kHz sampling rate) that is significantly smaller than the measurement room reverberation time, we are also restricting the resolution of the frequency grid used in the beamformer design ($64/8000 = 0.008$ Hz^{-1}) to be far below the resolution of 0.2 Hz^{-1} [34] that would be needed to include the reverberant characteristics of the measurement room response. Consequently, the fixed beamformer design will primarily depend on the microphone contribution in the joint acoustic and microphone responses.

For being able to produce audiovisual simulation results, we recorded the dry signal to be used as the front sound source using a camcorder device. This signal is a male speech signal and has been recorded in a room with a very short reverberation time. The audiovisual capture consists of three phases: a far camera shot with a duration of 5 s, a camera zoom operation with a duration of 10 s, and a close camera shot with a duration of 11 s. In the first simulation scenario, the sound scene is composed of the front speech source at $\theta_F = 0^\circ$ and two additional components: a localized interfering music signal at $\theta_I = 90^\circ$ and a diffuse babble noise signal, both having a level of 5 dB above the front speech source level. Note that this corresponds to a relatively noisy scenario, which is intended to clearly illustrate the potential benefit of applying the AZ algorithm. In the second simulation scenario, the sound scene consists of the same front speech source, the same localized interfering and diffuse noise signals, which are now scaled to have a



(a) microphone 1 (b) microphone 2 (c) microphone 3

(d) microphone 1 (e) microphone 2 (f) microphone 3

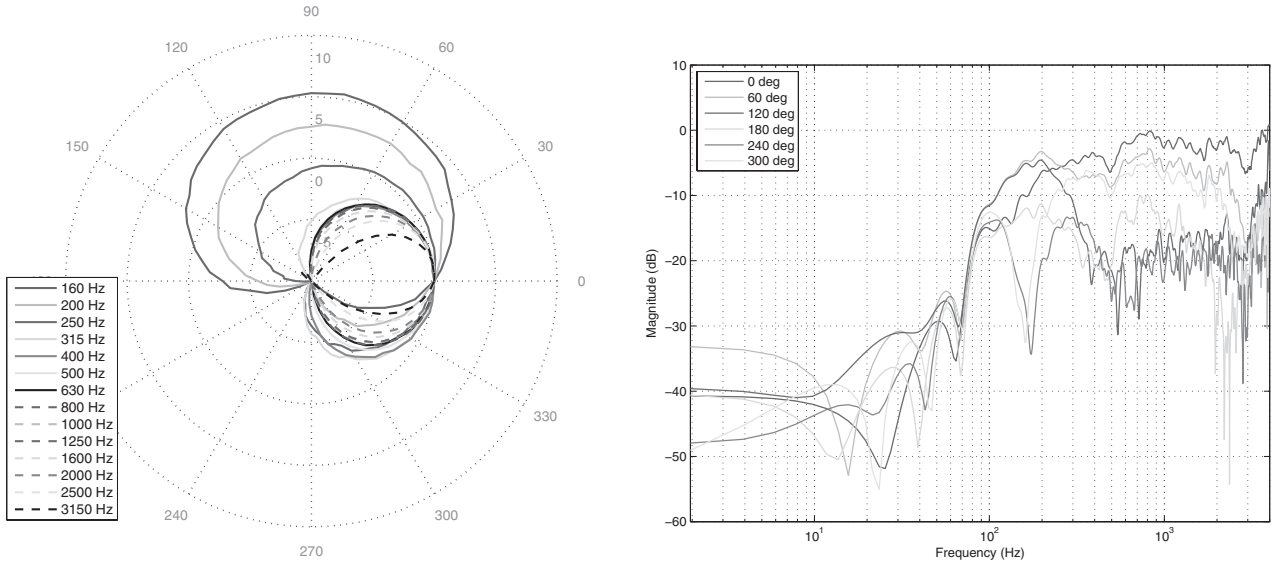
Fig. 4. Estimated joint acoustic and microphone responses of mobile device used in audio recordings: spatial responses (top row) and frequency responses (bottom row).

sound level of 10 dB below the front speech source level, and an additional back sound source. The back sound source is a female speaker at $\theta_B = 180^\circ$, with a level of 20 dB above the front speech source. This rather unbalanced situation is a very common one if the camera operator is speaking during an audiovisual capture, and is due to the typically small distance between the operator and the microphone(s).

We will now discuss the spatial and spectral filter design, as well as the choice of the relevant design parameters. The SD design of the FSB as specified in (51) requires two parameters to be set: the regularization parameter δ and the filter length L_{FSB}. The choice of the regularization parameter requires a trade-off between robustness to microphone mismatch and sensor noise on the one hand, and beamformer directivity on the other hand [25]. The choice of the filter length essentially determines the frequency grid at which the SD design is executed. Indeed, the SD design equation (51) is evaluated at $L_{\mathrm{FSB}}/2$ frequencies that are uniformly distributed in the Nyquist interval, after which the frequency-domain FSB coefficients are transformed to length- L_{FSB} time-domain filters. A larger filter length leads to a beamformer with a higher directivity and a flatter magnitude response (which may be important in the context of assumption (36)), but obviously results in a higher computational load. We will use the values $\delta = 10^{-6}$ and $L_{\mathrm{FSB}} = 64$, which leads to an FSB with a spatial and frequency response as shown in Fig. 5. Note that an excellent directivity and a relatively flat magnitude response is obtained above 300 Hz. At lower frequencies, the spatial response is drawn to the left due to the asymmetry in the microphone configuration.

Next to the FSB, the fixed beamformer also constitutes a BM that is designed to cancel the front sound source, see (55). Only a single design parameter needs to be decided on, namely the filter length L_{BM}. Similar to the FSB design, the BM design is executed on a uniform frequency grid, the resolution of which is determined by the specified time-domain filter length. The choice of this parameter is a compromise between the degree of suppression in the front direction and the flatness of the magnitude response on the one hand, and the computational complexity of the filtering operations on the other hand. The choice $L_{\mathrm{BM}} = 64$ results in the BM responses shown in Fig. 6. The average magnitude response in the front direction is around -10 dB for frequencies above 150 Hz, and drops down to -30 dB above 700 Hz. Note that the BM magnitude response is not as flat as the FSB magnitude response, and exhibits some comb filtering behavior.

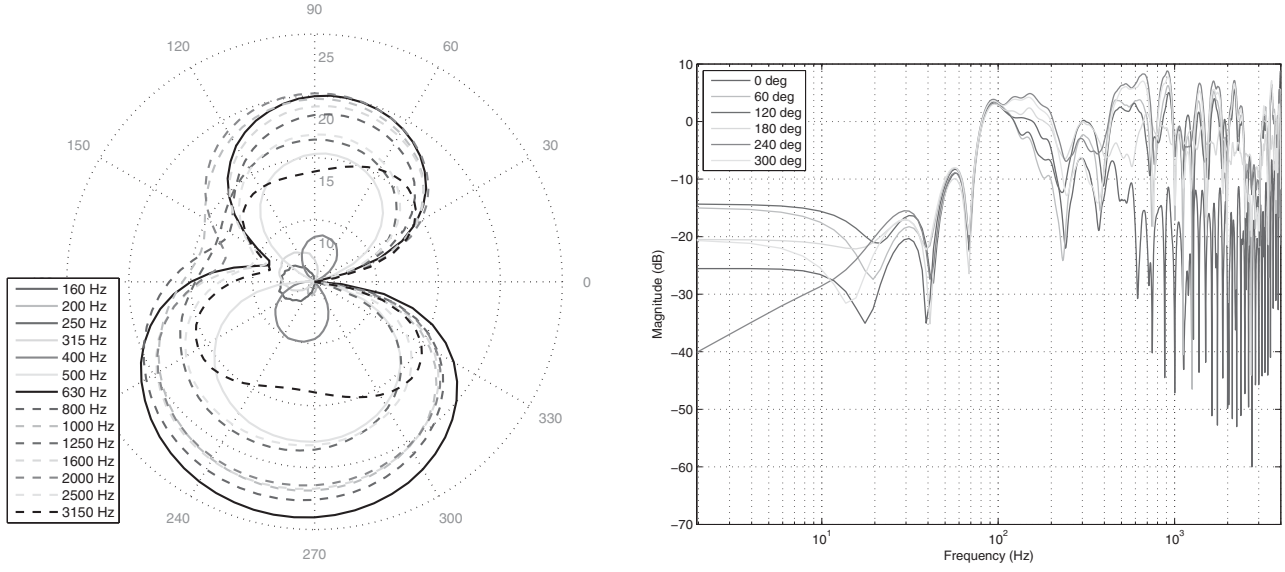
The design of the SDW-MWF adaptive beamformer as given in (57)–(59) requires some more implementation decisions to be taken. First of all, we have to decide on a fundamental structural issue, that is, whether or not to use a feedforward filter $W_{2,\,(0)}(\omega)$ that includes the fixed beamformer primary output signal $X_0(\omega)$ as an additional noise reference in the adaptive beamformer. With the aim of avoiding too large suppression of the desired front sound source by the adaptive beamformer, we decide not to use a feedforward filter $W_{2,\,(0)}(\omega)$, in which case the SDW-MWF reduces



(a) (b)

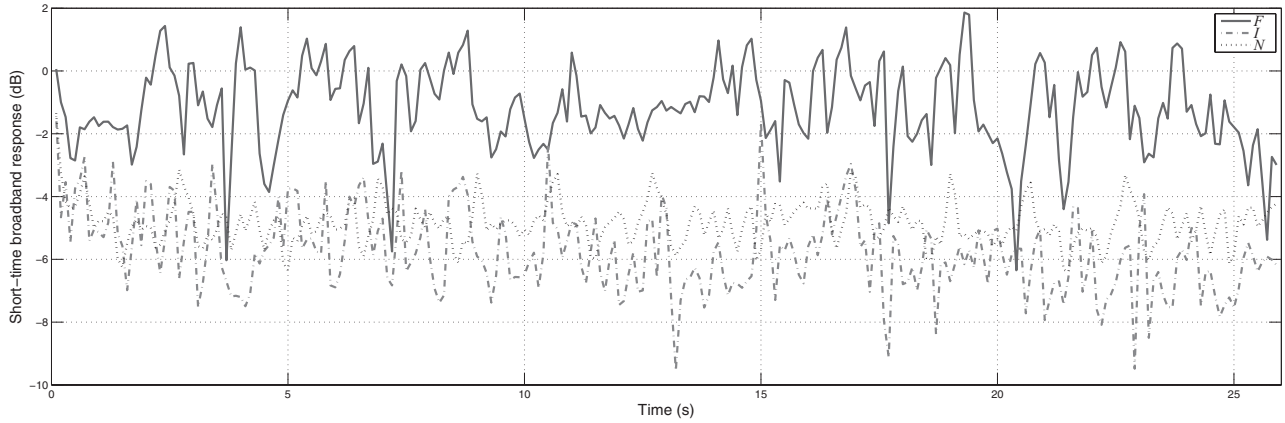
Fig. 5. Superdirective FSB design: (a) spatial response and (b) frequency response.

a speech-distortion-regularized GSC (SDR-GSC) [35]. A second design choice concerns the value of the trade-off parameter μ. Even in the SDR-GSC case, it makes sense to explicitly minimize the distortion of the desired front sound source (by choosing $1/\mu > 0$), because this signal may leak into the noise references due to an imperfect front sound source cancellation in the BM. On the other hand, μ should be chosen large enough so that the noise reduction performance is not affected. The value $\mu = 100$ appears to be a good compromise for this specific scenario, as it provides about 5 dB suppression of localized interferers while hardly affecting the desired front sound source (see Fig. 7 discussed below). A more thorough and general treatment on the choice of μ can be found in [29],[35]. Third, the adaptive beamformer filter length is set to $L_{\mathrm{AB}} = 128$, which is again a trade-off between spatial and spectral resolution and computational complexity. Finally, a recursive implementation of (57)–(59) needs to be sought for. We adopt the stochastic gradient frequency-domain algorithm proposed in [29], with a constrained filter update involving a diagonal correlation matrix approximation (denoted as "Algo 2 – constrained" in [29]). The effect of the adaptive beamformer in the two simulation scenarios is depicted in Fig. 7. Here, we display the short-time broadband response of the adaptive beamformer as a function of time, for each of the individual sound source components (F, I, N, and possibly B). The short-time broadband response is calculated as the ratio of the power of the adaptive beamformer primary

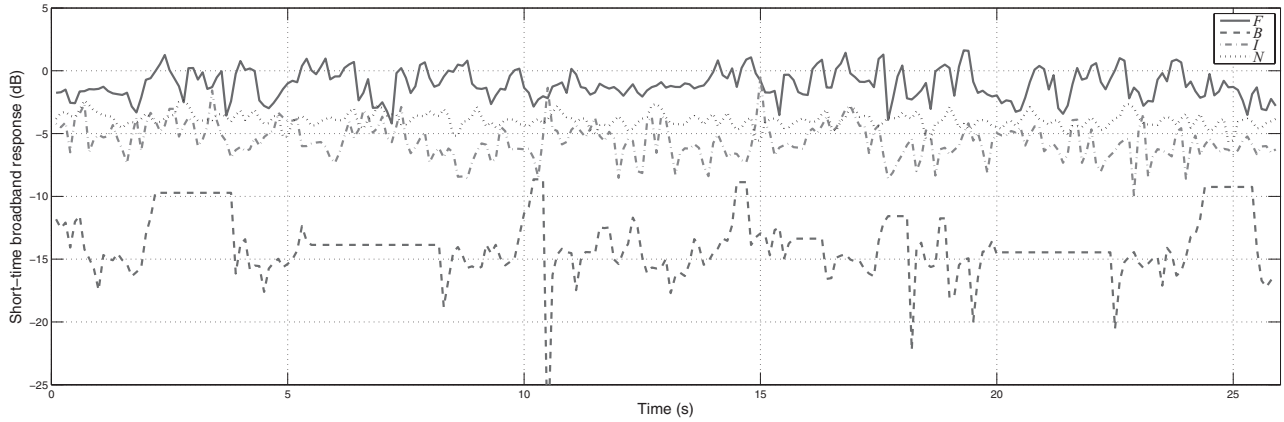


(a) (b)

Fig. 6. Front-canceling BM design: (a) spatial response and (b) frequency response.



(a)



(b)

Fig. 7. Adaptive beamformer short-time broadband response vs. time for individual sound source components: (a) scenario 1 and (b) scenario 2.

output signal $y_0(t)$ and the average power in the microphone signal pair $u_0(t)$ and $u_1(t)$, estimated in 50% overlapping frames of 100 ms duration. It can be observed from Fig. 7 that in both simulation scenarios the localized interfering sound source component as well as the diffuse noise component is suppressed with approximately 6 dB, while the front sound source component is hardly attenuated. In the second scenario, the suppression of the back sound source component is much larger than the suppression of the other components, since the back sound source component dominates the other components in the noise reference at the BM output. Note that during speech pauses of the back sound source, the short-time broadband response is ill-defined and its value in Fig. 7(b) has been fixed to the last well-defined value before each speech pause.

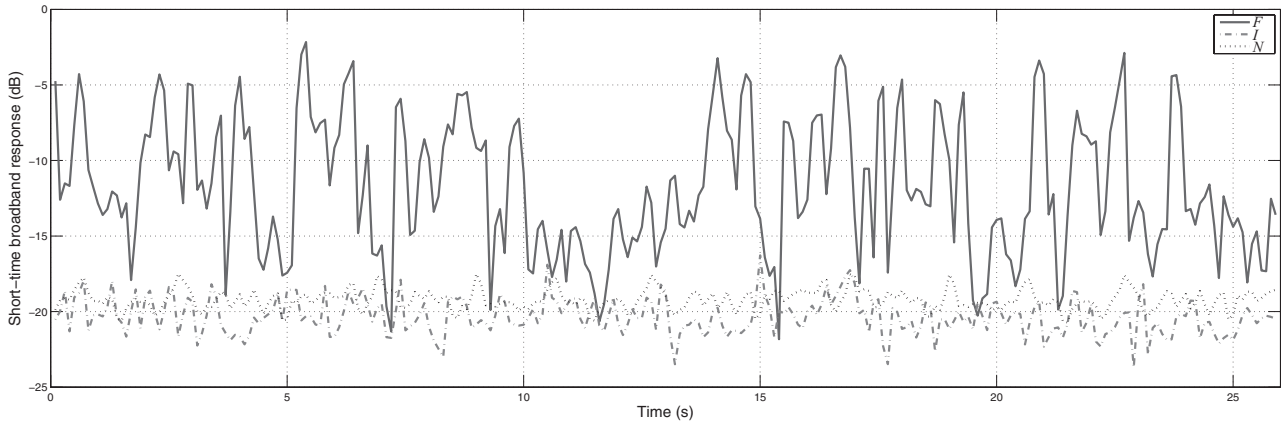
The adaptive spectral attenuation as defined in (60)–(63) is implemented here as a modified version of the spectral postfilter used in the NXP Software LifeVibes NoiseVoid software[1] (version 1.1) for dual-microphone noise reduction. The short-time broadband response of the adaptive spectral attenuation, defined in a similar way as for the adaptive beamformer, is shown in Fig. 8. It is important to note that the amount of suppression of the individual sound source components is remarkably different in the two scenarios. This is due to the fact that the sound sources have significant spectral overlap, and as a consequence the suppression of one sound source component influences the suppression of other components as well. In particular, the suppression of the back sound source component in the second scenario unavoidably leads to an increased suppression of the other components.

The AZ design is finalized by calculating the weight trajectories $a_n^{(m)}(t),\ m = 0, \ldots, 3,\ n = 0, 1,\ \forall t$, which requires the specification of the desired gain factors $g_F(t)$, $g_B(t)$, $g_I(t)$, and $g_N(t)$ and the estimation of the reweighting factors $\gamma_{\cdot\cdot}^{(\cdot)}$. As mentioned in Section 2, we can obtain the desired AZ effect by specifying the front sound source gain factor $g_F(t)$ to be a monotonically increasing function of time during a zoom-in operation. From a physical point of view, this function should be specified as follows:

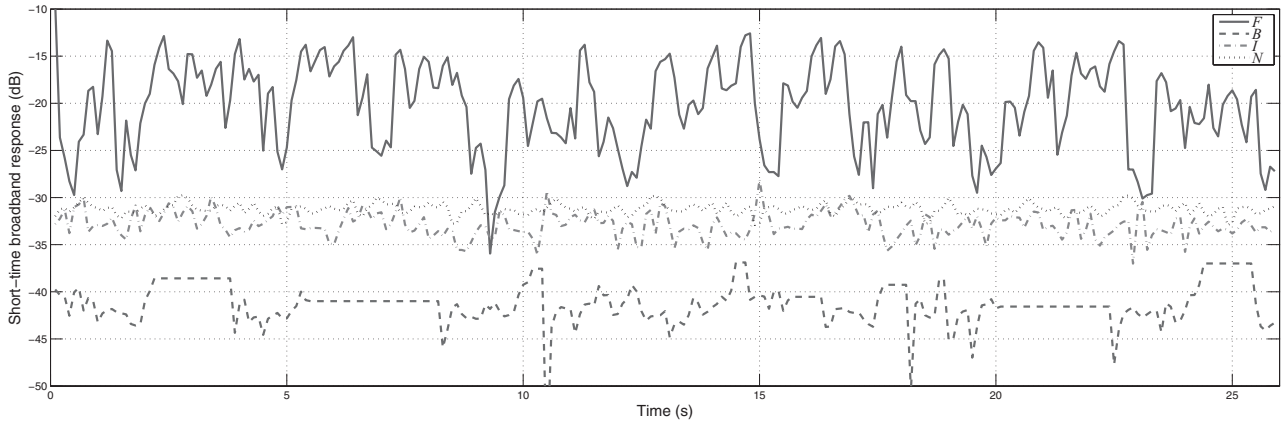
$$g_F(t) = 2^{v_{\text{zoom}}t}, \quad 0 \leq t \leq \frac{d_{\text{zoom}}}{v_{\text{zoom}}} \tag{64}$$

where v_{zoom} and d_{zoom} represent the video zoom speed (i.e., the average speed in m/s at which the camera is perceived to approach the front sound source) and the video

[11] http://www.software.nxp.com/



(a)



(b)

Fig. 8. Adaptive spectral attenuation short-time broadband response vs. time for individual sound source components: (a) scenario 1 and (b) scenario 2.

zoom distance (i.e., the distance over which the camera is perceived to move from the far shot to the close shot). However, informal listening tests involving five experienced audio engineers (including the authors) have revealed that the function in (64) does not always yield perceptually satisfying results. More particularly, a faster level increase at the start of the zoom-in operation, eventually converging to the same final level at close-up, was unanimously found to be more desirable. Therefore, the following function was designed and was used in the simulation results reported below,

$$g_F(t) = 1 + \frac{2^{d_{\text{zoom}}} - 1}{\sqrt{1.2 d_{\text{zoom}}}} \sqrt{1.2 v_{\text{zoom}} t}, \quad 0 \le t \le \frac{d_{\text{zoom}}}{v_{\text{zoom}}}. \tag{65}$$

In Fig. 9, a comparison of the functions in (64) and (65) for an example scenario with $v_{\text{zoom}} = 0.25$ m/s and $d_{\text{zoom}} = 2.5$ m is shown. As stated earlier, we refer to [17] and [21] for a more profound treatment of the specification of $g_F(t)$.

In the first simulation scenario, the localized interfering sound source and the diffuse noise are specified to remain stationary during the zoom operation, i.e.,

$$g_I(t) \equiv g_N(t) \equiv 1. \tag{66}$$

Due to the equivalence in (66), and due to the absence of a back sound source, the weight calculation in (16) can be simplified to solving a 2×2 linear system of equations, thereby explicitly treating the sum of the localized interfering sound source component and the diffuse noise component as a single component in the microphone and auxiliary signals. We thus set $a_n^{(1)}(t) \equiv a_n^{(2)}(t) \equiv 0$, $n = 0, 1$, and calculate the remaining weight trajectories as follows:

$$\begin{bmatrix} a_n^{(0)}(t) \\ a_n^{(3)}(t) \end{bmatrix} = \begin{bmatrix} 1 & \gamma_{z_0,u_n}^{(F)} \\ 1 & \gamma_{z_0,u_n}^{(I+N)} \end{bmatrix}^{-1} \begin{bmatrix} g_F(t) \\ g_{I+N}(t) \end{bmatrix}, \quad n = 0, 1. \tag{67}$$

As explained in Section 3, an off-line calculation of the reweighting factors is often preferable from a computational point of view. To this end, the suppression applied to each of the different sound source components (F, B, I, and N) in the three functional blocks of the AZ algorithm has to be estimated a priori. This estimation should

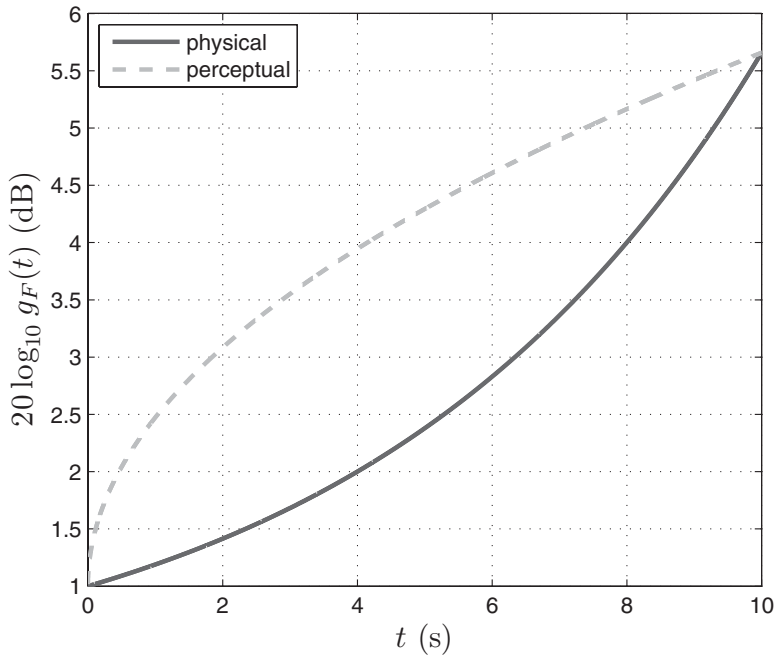


Fig. 9. Physically vs. perceptually "optimal" functions for front sound source level increase during zoom-in at $v_{\text{zoom}} = 0.25$ m/s over $d_{\text{zoom}} = 2.5$ m.

exploit the knowledge of the fixed beamformer response, as well as heuristics about the behavior of the adaptive beamformer and spectral attenuation response. Such heuristics can be inferred from the literature, by adopting average values of the expected noise reduction and desired signal suppression for the particular algorithms used in the adaptive beamformer and spectral attenuation blocks. Here, we assume a 5 dB suppression of the desired (F) component and a 25 dB suppression of the undesired ($I + N$) components in the adaptive spectral attenuation output signal $z_0(t)$ relative to the microphone signals,

$$\gamma_{z_0,u_n}^{(F)} = 10^{-5/20}, \quad n = 0, 1 \tag{68}$$

$$\gamma_{z_0,u_n}^{(I+N)} = 10^{-25/20}, \quad n = 0, 1. \tag{69}$$

The specified gain factors and the resulting weight trajectories for this scenario are displayed in Fig. 10. Finally, Fig. 11 shows the time-domain waveforms of the individual sound source components in the stereo input signal (Figs. 11(a) and 11 (c)) and in the stereo output signal (Figs. 11(b) and 11(d)). It can be observed that the AZ algorithm indeed results in the desired effect, that is, the level of the localized interfering sound source and the diffuse noise is not altered, while the level of the front sound source is gradually increased during the zoom operation (between $t = 5$ s and $t = 15$ s).

In the second simulation scenario, we will again treat the sum of the localized interfering sound source component and the diffuse noise component as a single component, and re-use the gain factors specified in (65) and (66). For the back sound source, we specify a constant gain factor $g_B(t) \equiv 10^{-15/20}$, with the aim of compensating for the level imbalance between the front and back sound source components in the microphone signals. In this scenario, we can reduce the weight calculation to the solution of a 3×3 linear system of equations by fixing $a_n^{(2)}(t) \equiv 0,\ n = 0, 1$, and

$$\begin{bmatrix} a_n^{(0)}(t) \\ a_n^{(1)}(t) \\ a_n^{(3)}(t) \end{bmatrix} = \begin{bmatrix} 1 & \gamma_{x_0,u_n}^{(F)} & \gamma_{z_0,u_n}^{(F)} \\ 1 & \gamma_{x_0,u_n}^{(B)} & \gamma_{z_0,u_n}^{(B)} \\ 1 & \gamma_{x_0,u_n}^{(I+N)} & \gamma_{z_0,u_n}^{(I+N)} \end{bmatrix}^{-1} \begin{bmatrix} g_F(t) \\ g_B(t) \\ g_{I+N}(t) \end{bmatrix}, \quad n = 0, 1. \tag{70}$$

The reweighting factors are specified heuristically as follows:

$$\gamma_{x_0,u_n}^{(F)} = 1, \quad n = 0, 1 \tag{71}$$

$$\gamma_{x_0,u_n}^{(B)} = 10^{-15/20}, \quad n = 0, 1 \tag{72}$$

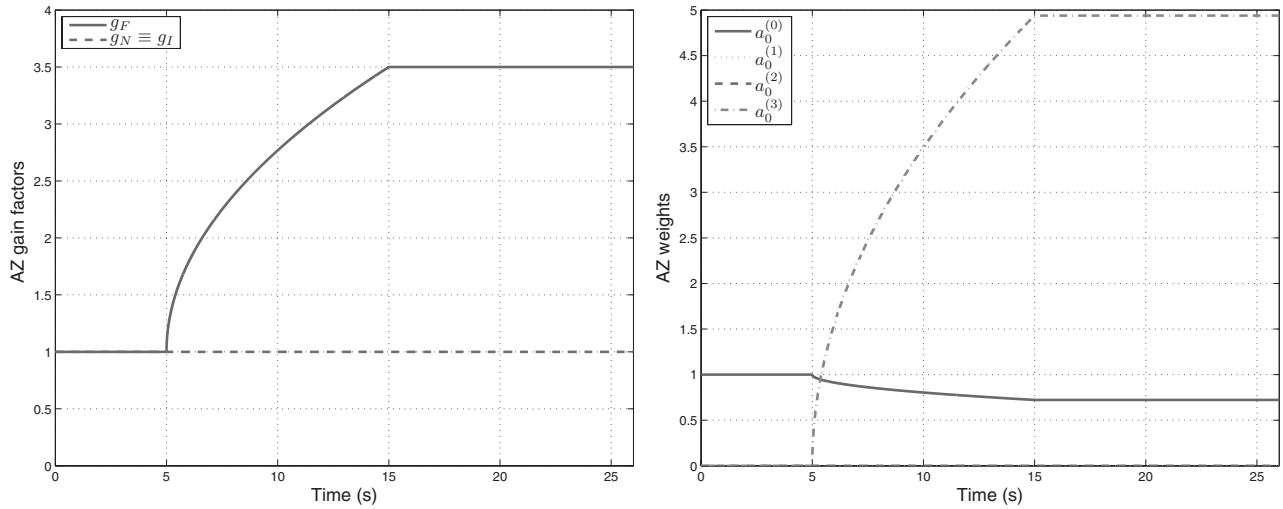
$$\gamma_{x_0,u_n}^{(I+N)} = 10^{-5/20}, \quad n = 0, 1 \tag{73}$$

$$\gamma_{z_0,u_n}^{(F)} = 10^{-15/20}, \quad n = 0, 1 \tag{74}$$

$$\gamma_{z_0,u_n}^{(B)} = 10^{-40/20}, \quad n = 0, 1 \tag{75}$$

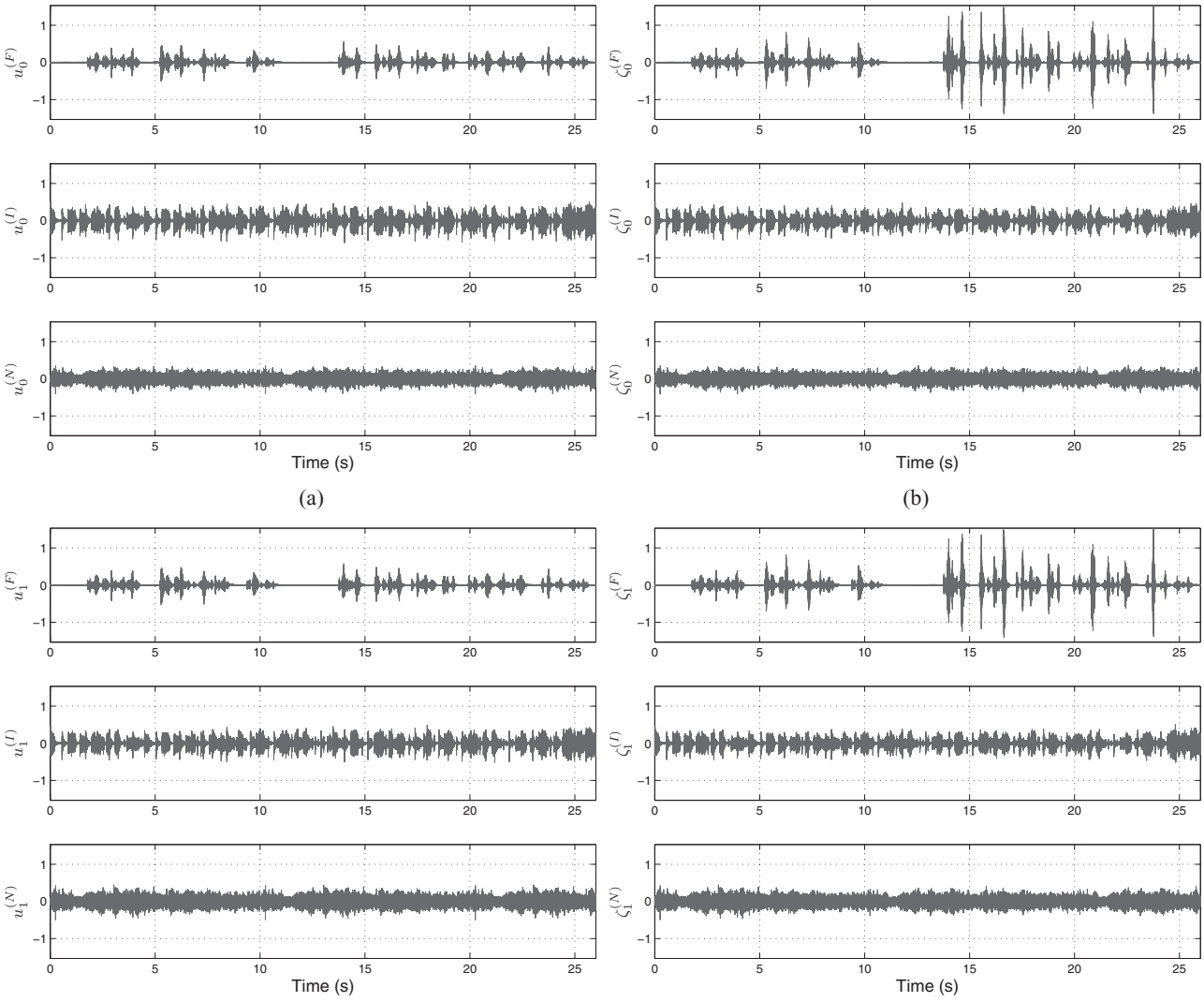
$$\gamma_{z_0,u_n}^{(I+N)} = 10^{-30/20}, \quad n = 0, 1. \tag{76}$$

The first three values are inspired by the fact that the FSB is designed to have a much higher directivity in the front direction than in other directions (specifically, the back direction). The last three values are based on the expected broadband suppression of the undesired sound source components at the output of the adaptive spectral attenuation block. Indeed, as we observed in Fig. 8(b), we can expect



(a) (b)

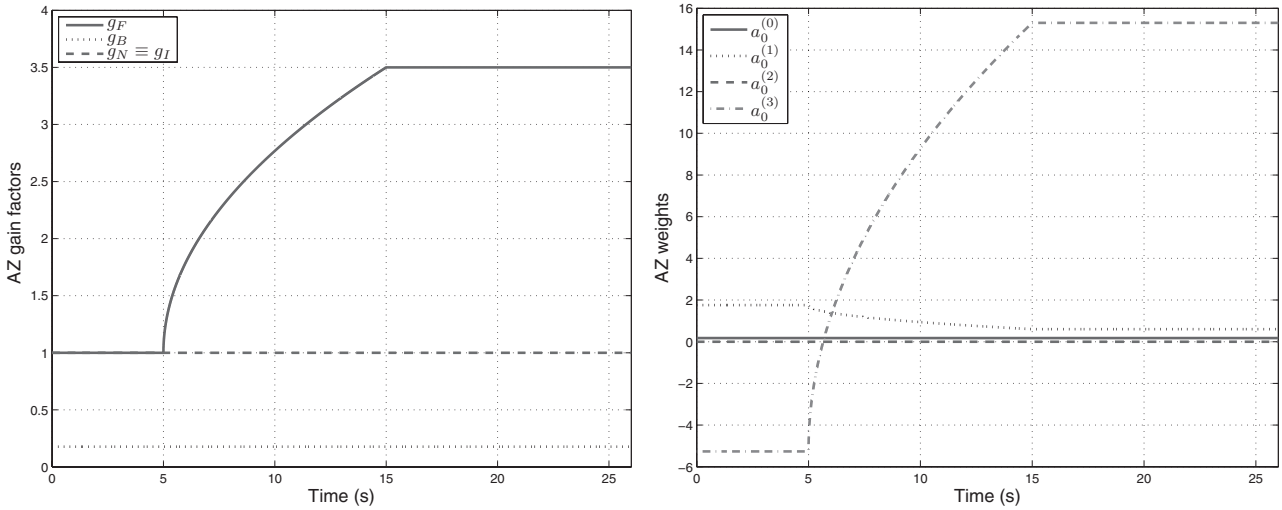
Fig. 10. AZ simulation scenario 1: (a) specified gain factors and (b) calculated weight trajectories.



(a) (b)

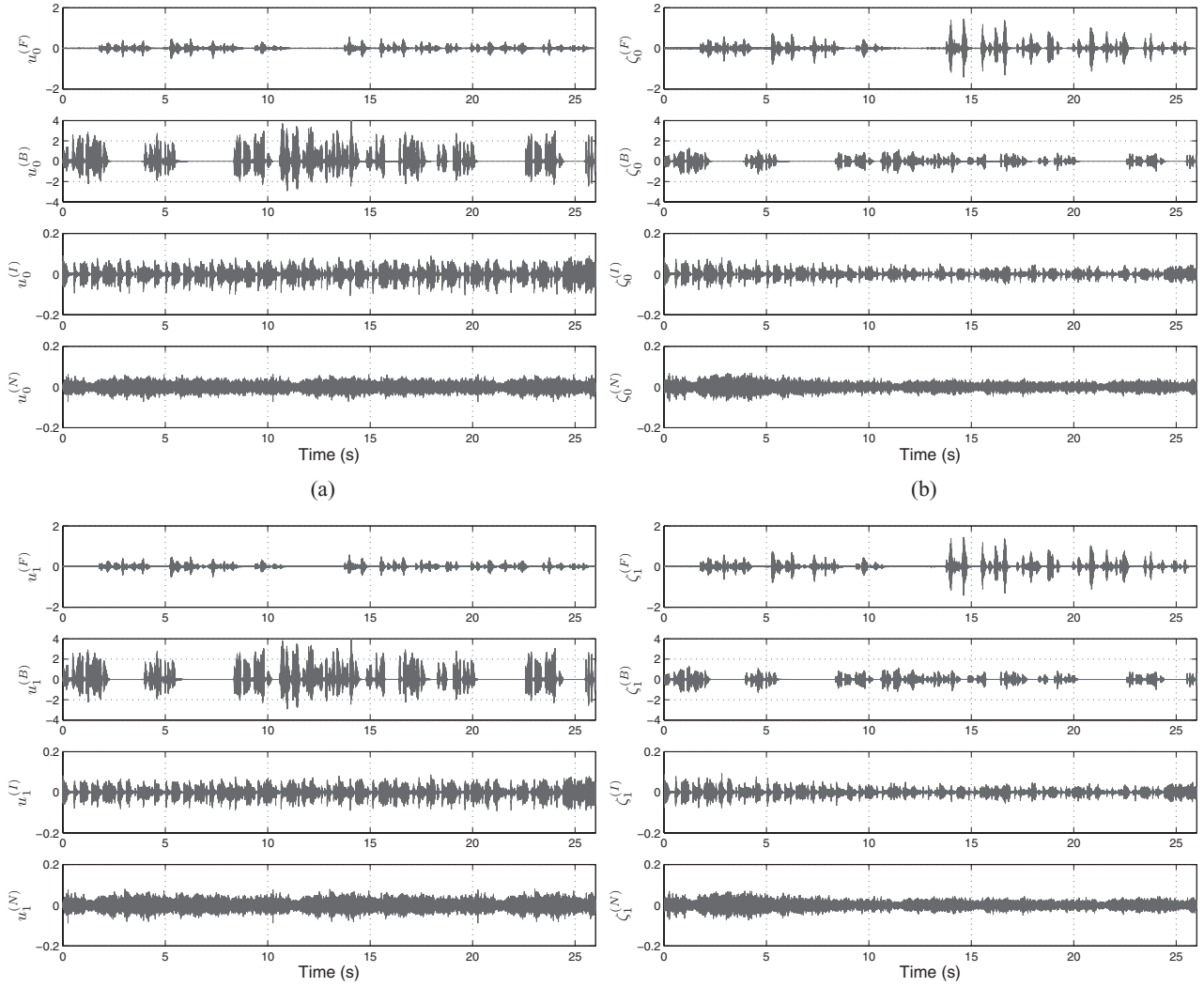
(c) (d)

Fig. 11. AZ simulation scenario 1: (a) left input signal components, (b) left output signal components, (c) right input signal components, and (d) right output signal components.



(a) (b)

Fig. 12. AZ simulation scenario 2: (a) specified gain factors and (b) calculated weight trajectories.



(a) (b)

(c) (d)

Fig. 13. AZ simulation scenario 2: (a) left input signal components, (b) left output signal components, (c) right input signal components, and (d) right output signal components.

high-power undesired components in the microphone signals (in this case the back sound source component) to be suppressed more heavily than other undesired components. Moreover, due to the significant spectral attenuation of the back sound source component, the other components will also be suppressed to a larger extent as compared to the first simulation scenario. The resulting gain factors and weight trajectories are shown in Fig. 12, while the waveforms of the individual sound source components in the resulting stereo input and output signals are displayed in Fig. 13. We can observe that the desired AZ effect is indeed achieved, that is, a monotonic level increase of the front sound source during the zoom operation and a fixed level decrease of the back sound source. This comes at the cost of a slight level decrease of the localized interfering sound source and diffuse noise, which is however hardly perceivable.

The multimedia files obtained by combining the stereo input and output signals from both scenarios with the corresponding video scene are available for download[2].

6 CONCLUSION

In this paper, we have proposed an AZ algorithm for audiovisual capture applications, based on a novel multimicrophone sound scene manipulation framework. We have shown that an explicit source separation of the sound scene can be avoided by linearly combining a number of microphone and auxiliary signals, in which the different sound scene components have been independently reweighted relative to each other. In this way, an independent level control of the individual sound sources in a captured microphone signal can be achieved, which can then be exploited to produce the desired AZ effect. The proposed structure for calculating the auxiliary signals is built upon recent results from spatial and spectral noise reduction and speech enhancement literature, and requires a limited amount of computational resources.

An experimental evaluation in two simulated scenarios confirms that the desired AZ effect can indeed be obtained using an array with a small number of low-cost microphones. Nevertheless, a number of issues has been postponed to future work. First of all, the design of the fixed beamformer requires the knowledge of the joint acoustic and microphone responses. In practice, these responses have to be estimated and hence the impact of potential estimation errors should be further investigated. Secondly, since an objective measure for quantifying the perceptual AZ effect is not available at present, subjective listening tests should be executed to verify the perceptual performance of the proposed algorithm.

Finally, we should mention that the proposed multimicrophone sound scene manipulation framework could be valuable in applications other than AZ, such as the synthetic microphone array configuration problem [36] or the design of personalized spatial audio displays for the visually impaired [37].

[22] ftp://ftp.esat.kuleuven.be/pub/sista/vanwaterschoot/abstracts/12-173.html

7 ACKNOWLEDGMENT

T. van Waterschoot is a Postdoctoral Fellow of the Research Foundation Flanders (FWO–Vlaanderen). This research work was carried out at the ESAT Laboratory of Katholieke Universiteit Leuven, in collaboration with and funded by NXP Software, and in the frame of KU Leuven Research Council CoE EF/05/006 “Optimization in Engineering” (OPTEC) and PFV/10/002 (OPTEC), Concerted Research Action GOA-MaNet, the Belgian Programme on Interuniversity Attraction Poles initiated by the Belgian Federal Science Policy Office IUAP P6/04 “Dynamical systems, control and optimization” (DYSCO) 2007–2011, Research Project IBBT, and Research Project FWO nr. G.0600.08 “Signal processing and network design for wireless acoustic sensor networks”. The scientific responsibility is assumed by its authors.

8 REFERENCES

[1] R. B. Welch and D. H. Warren, “Immediate perceptual response to intersensory discrepancy,” *Psychol. Bull.*, vol. 88, no. 3, pp. 638–667, Nov. 1980.

[2] D. Alais and D. Burr, “The ventriloquist effect results from near-optimal bimodal integration,” *Current Biology*, vol. 14, no. 3, pp. 257–262, Feb. 2004.

[3] S. Mateeff, J. Hohnsbein, and T. Noack, “Dynamic visual capture: apparent auditory motion induced by a moving visual target,” *Perception*, vol. 14, no. 6, pp. 721–727, 1985.

[4] S. Soto-Faraco, A. Kingstone, and C. Spence, “Cross-modal dynamic capture,” *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 30, no. 2, pp. 330–345, Apr. 2004.

[5] J. M. Brown, “Visual influences on auditory distance perception,” *J. Acoust. Soc. Amer.*, vol. 104, no. 3, p. 1798, Sept. 1998.

[6] P. Zahorik, “Estimating sound source distance with and without vision,” *Optometry & Vision Science*, vol. 78, no. 5, pp. 270–275, May 2001.

[7] F. B. Kunka and B. Kostek, “ANew Aspects of Virtual Sound Source Localization Research-Impact of Visual Angle and 3-D Video Content on Sound Perception,” *PJ Audio Eng Soc*, vol. V61, no. Issue, pp. F280–L289 (Y2013 May).

[8] C. Spence, “Audiovisual multisensory integration,” *Acoust. Sci. & Tech.*, vol. 28, no. 2, pp. 61–70, 2007.

[9] M. Ercan Altinsoy, “The quality of auditory-tactile virtual environments,” *J. Audio Eng. Soc.*, vol. 60, no. 1/2, pp. 38–46, Jan./Feb. 2012.

[10] J. W. Lewis, M. S. Beauchamp, and E. A. DeYoe, “A comparison of visual and auditory motion processing in human cerebral cortex,” *Cerebral Cortex*, vol. 10, no. 9, pp. 873–888, Sept. 2000.

[11] O. Baumann and M. W. Greenlee, “Neural correlates of coherent audiovisual motion perception,” *Cerebral Cortex*, vol. 17, no. 6, pp. 1433–1443, June 2007.

[12] P. W. Battaglia, R. A. Jacobs, and R. N. Aslin, “Bayesian integration of visual and auditory signals for spatial localization,” *J. Opt. Soc. Am. A*, vol. 20, no. 7, pp. 1391–1397, 2003.

[13] F. Rumsey, "Audio in multimodal applications," *J. Audio Eng. Soc.*, vol. 58, no. 3, pp. 191–195, Mar. 2010.

[14] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst, "Auditory distance perception in humans: a summary of past and present research," *Acta Acust. United Ac.*, vol. 91, no. 3, pp. 409–420, May/Jun. 2005.

[15] P. Zahorik, "Assessing auditory distance perception using virtual acoustics," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1832–1846, Apr. 2002.

[16] Y. Ishigaki, M. Yamamoto, K. Totsuka, and N. Miyaji, "Zoom microphone," in *Preprints AES 67th Convention*, New York, NY, USA , Nov. 1980, AES Preprint 1713.

[17] M. Matsumoto, H. Naono, H. Saitoh, K. Fujimura, and Y. Yasuno, "Stereo zoom microphone for consumer video cameras," *IEEE Trans. Consum. Electron.*, vol. 35, no. 4, pp. 759–766, Nov. 1989.

[18] R. M. M. Derkx, "First-order superdirectional acoustic zooming in the presence of directional interferences," in *Proc. 2011 Joint Workshop Hands-free Speech Commun. Microphone Arrays (HSCMA '11)*, Edinburgh, UK , May 2011, pp. 176–179.

[19] J. Meyer and G. W. Elko, "Eigenbeam beamforming for microphone arrays," *J. Acoust. Soc. Amer.*, vol. 120, no. 5, p. 3177, Nov. 2006.

[20] –, "Spherical microphone arrays for 3D sound recording," in *Audio signal processing for next-generation multimedia communication systems*, Y. Huang and J. Benesty, Eds. Dordrecht : Kluwer, 2004, ch. 3.

[21] R. Schultz-Amling, F. Kuech, O. Thiergart, and M. Kallinger, "Acoustical zooming based on a parametric sound field representation," in *Preprints AES 128th Convention*, London, UK , May 2010, AES Preprint 8120.

[22] O. M. Bucci, D. D'Elia, G. Mazzarella, and G. Panatiello, "Antenna pattern synthesis: a new general approach," *Proc. IEEE*, vol. 82, no. 3, pp. 358–371, Mar. 1994.

[23] H. Lebret and S. Boyd, "Antenna array pattern synthesis via convex optimization," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 526–532, Mar. 1997.

[24] G. W. Elko and J. Meyer, "Microphone arrays," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and J. Benesty, Eds. Berlin : Springer, 2008, ch. 50.

[25] S. Doclo and M. Moonen, "Superdirective beamforming robust against microphone mismatch," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 2, pp. 617–631, Feb. 2007.

[26] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[27] S. Doclo and M. Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 9, pp. 2230–2244, Sept. 2002.

[28] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 2, pp. 260–276, Feb. 2010.

[29] S. Doclo, A. Spriet, J. Wouters, and M. Moonen, "Frequency-domain criterion for the speech distortion weighted multichannel wiener filter for robust noise reduction," *Speech Commun.*, vol. 49, no. 7–8, pp. 636–656, Jul.-Aug. 2007, special Issue on Speech Enhancement.

[30] G. Rombouts and M. Moonen, "QRD-based unconstrained optimal filtering for acoustic noise reduction," *Signal Processing*, vol. 83, no. 9, pp. 1889–1904, Sept. 2003.

[31] –, "Stochastic gradient-based implementation of spatially preprocessed speech distortion weighted multichannel wiener filtering for noise reduction in hearing aids," *IEEE Trans. Signal Process.*, vol. 53, no. 3, pp. 911–925, Mar. 2005.

[32] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[33] E. Hänsler and G. Schmidt, *Acoustic echo and noise control: a practical approach*. New York : Wiley-IEEE, 2004.

[34] M. R. Schroeder, "Improvement of acoustic-feedback stability by frequency shifting," *J. Acoust. Soc. Amer.*, vol. 36, no. 9, pp. 1718–1724, Sept. 1964.

[35] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, Dec. 2004.

[36] Y. Hur, J. S. Abel, Y.-C. Park, and D. H. Youn, "Techniques for synthetic reconfiguration of microphone arrays," *J. Audio Eng. Soc.*, vol. 59, no. 6, pp. 404–418, June 2011.

[37] M. Bujacz, P. Skulimowski, and P. Strumiłło, "Naviton – a prototype mobility aid for auditory presentation of three-dimensional scenes to visually impaired," *J. Audio Eng. Soc.*, vol. 60, no. 9, pp. 696–708, Sept. 2012.

[38] E. J. Diethorn, "Subband noise reduction methods for speech enhancement," in *Acoustic signal processing for telecommunication*, S. L. Gay and J. Benesty, Eds. Dordrecht : Kluwer, 2000, ch. 9.

THE AUTHORS


Toon van Waterschoot


Wouter Joos Tirry


Marc Moonen

Toon van Waterschoot received the Master's degree (2001) and the PhD degree (2009) in Electrical Engineering, both from KU Leuven, Belgium.

Since 2011, he has been a Postdoctoral Research Fellow of the Research Foundation - Flanders (FWO) at KU Leuven, Belgium. In 2002, he spent a year as a Teaching Assistant with the Antwerp Maritime Academy (Hogere Zeevaartschool Antwerpen), Belgium. From 2002 to 2003, and from 2008 to 2009, he was a Research Assistant with KU Leuven, Belgium, while from 2004 to 2007, he was a Research Assistant with the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT), Belgium. After his PhD graduation, he was a Postdoctoral Research Fellow at KU Leuven, Belgium (2009–2010) and at Delft University of Technology, The Netherlands (2010–2011). Since 2005, he has been a Visiting Lecturer at the Advanced Learning and Research Institute of the University of Lugano (Universit della Svizzera italiana), Switzerland, where he is teaching Digital Signal Processing. His research interests are in adaptive and distributed signal processing and parameter estimation, with application to acoustic signal enhancement, speech and audio processing, wireless communications, and sensor networks.

Dr. van Waterschoot is currently serving as an Associate Editor for the Journal of the Audio Engineering Society and for the EURASIP Journal on Audio, Music, and Speech Processing, and as a Nominated Officer for the European Association for Signal Processing (EURASIP). He has been serving as an Area Chair for Speech Processing at the European Signal Processing Conference (EUSIPCO 2010 and 2013), and has been a Technical Program Committee member for several other international conferences. He is a member of the Audio Engineering Society, the Acoustical Society of America, EURASIP, and IEEE.

●

Wouter Joos Tirry is the Audio Technology Development Leader at NXP Software Leuven, Belgium. He received his PhD degree in Solar Physics at KU Leuven, Belgium, in 1998. Since 1999, he has been building up expertise in speech enhancement for mobile devices at Philips, Philips Software and NXP Software as research engineer and system architect. Currently he is responsible for the supervision of NXP Software's audio technology development activities and corresponding technology roadmap.

●

Marc Moonen is a Full Professor at the Electrical Engineering Department of KU Leuven, where he is heading a research team working in the area of numerical algorithms and signal processing for digital communications, wireless communications, DSL and audio signal processing.

He is a Fellow of the IEEE (2007). He received the 1994 KU Leuven Research Council Award, the 1997 Alcatel Bell (Belgium) Award (with Piet Vandaele), the 2004 Alcatel Bell (Belgium) Award (with Raphael Cendrillon), and was a 1997 Laureate of the Belgium Royal Academy of Science. He received a journal best paper award from the IEEE Transactions on Signal Processing (with Geert Leus) and from Elsevier Signal Processing (with Simon Doclo).

He was chairman of the IEEE Benelux Signal Processing Chapter (1998–2002), and a member of the IEEE Signal Processing Society Technical Committee on Signal Processing for Communications, and is currently Past-President of EURASIP (European Association for Signal Processing).

He has served as Editor-in-Chief for the EURASIP Journal on Applied Signal Processing (2003–2005), and has been a member of the Editorial Board of IEEE Transactions on Circuits and Systems II, IEEE Signal Processing Magazine, Integration-the VLSI Journal, EURASIP Journal on Wireless Communications and Networking, and Signal Processing. He is currently a member of the editorial board of EURASIP Journal on Applied Signal Processing and Area Editor for Feature Articles in IEEE Signal Processing Magazine.