DIRECTIONAL AUDIO CODING USING PLANAR MICROPHONE ARRAYS

Fabian Kuech, Markus Kallinger*, Richard Schultz-Amling*, Giovanni Del Galdo*, Jukka Ahonen** and Ville Pulkki***

*Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany.
**Laboratory of Acoustics and Audio Sig. Proc., TKK, Helsinki, Finland.
Email: fabian.kuech@iis.fraunhofer.de

ABSTRACT

Multichannel sound systems become more and more established in modern audio applications. Consequently, the recording and the reproduction of spatial audio gains increasing attention. Directional Audio Coding (DirAC) represents an efficient approach to analyze spatial sound and to reproduce it using arbitrary loudspeaker configurations. In DirAC, the direction-of-arrival and the diffuseness of sound within frequency subbands is used to encode the spatial properties of the observed sound field. The estimation of these parameters is based on an energetic sound field analysis using three-dimensional microphone arrays. In practice, however, physical design constraints make three-dimensional microphone configurations often not acceptable. In this paper, we consider a new approach to microphone array processing that allows for an estimation of both direction-of-arrival of sound and diffuseness based on planar microphone configurations. The performance of the proposed method is evaluated via simulations and real measured data.

Index Terms— Spatial audio, microphone arrays

1. INTRODUCTION

Spatial audio processing is becoming more important as the variety of possible applications for multichannel audio is constantly increasing. A common scenario is the use of home entertainment systems to listen to multichannel music or to watch movies featuring surround sound. Furthermore, there is a growing interest in making spatial audio available also for high quality teleconferencing systems. In the latter case it becomes obvious that not only a realistic reproduction of spatial audio is required, but also appropriate methods for the recording of spatial sound have to be provided.

For reasons of flexibility, the recording technique should not put any constraints on the loudspeaker configuration used for reproduction. Thus, common spaced-microphone techniques [1] are not suitable, as they assume a priori knowledge of the loudspeaker systems. Although coincident-microphone approaches, such as Ambisonics [2], are independent of the loudspeaker set-up, they do not yield the desired low level of coherence in the playback channels.

An efficient approach to the analysis and reproduction of spatial sound is the Directional Audio Coding (DirAC) technique [3, 4]. DirAC circumvents the drawbacks of the methods mentioned above by using a parametric representation of sound fields based on the features which are relevant for the perception of spatial sound, namely the direction-of-arrival (DOA) and diffuseness of the sound field in frequency subbands. In fact, DirAC assumes that interaural time differences (ITD) and interaural level differences (ILD) are perceived correctly when the DOA of a sound field is correctly reproduced, while interaural coherence (IC) is perceived correctly, if the diffuseness is reproduced accurately. On the reproduction side, the actual signals of the loudspeaker channels are determined as a function of these parameters so that an accurate spatial rendering can be achieved at a desired listening position.

Note that there are substantial differences between DirAC and parametric multichannel audio coding, such as MPEG Surround [5], although they share very similar processing structures. While MPEG Surround is based on a time/frequency analysis of the different *loudspeaker channels*, DirAC takes *microphone channels* as input. Thus, DirAC also represents an efficient recording technique for spatial audio.

In DirAC, the desired parameters are estimated via an energetic analysis of the sound field. This can be achieved by using B-format microphone signals [2], i.e., an omnidirectional signal and the signals of three figure-of-eight microphones aligned with the axes of a Cartesian coordinate system. These signals can be directly measured using SoundField microphones [2] which are, however, far too expensive for commercial consumer applications. Alternatively, three-dimensional (3D) microphone arrays can be used to generate the required B-format signals [6], where, e.g., six omnidirectional microphones are placed at the vertices of an octahedron.

In many application scenarios, three-dimensional microphone configurations are not feasible due to physical design constraints. Thus, not all of the required B-format signals are available. In order to still apply DirAC, the missing signals have to be replaced by corresponding approximations. In this contribution we consider a planar microphone array composed of five omnidirectional microphones for its application to the 3D analysis of sound fields using DirAC.

The paper is organized as follows. Section 2 recalls the DirAC analysis steps, while Section 3 presents and evaluates a method for approximating the missing dipole signal. Finally, we assess the performance of the proposed approach via measurements in Section 4. Section 5 draws the conclusions.

2. DIRECTIONAL AUDIO CODING: ANALYSIS

In the analysis step of DirAC, the parameters are estimated using specific energetic quantities of the observed sound field: The DOA of sound is determined using the active sound intensity vector, whereas the diffuseness of the sound field is estimated by relating the sound intensity to the overall energy density. In the following, we briefly recall the estimation of the desired parameters based on B-format microphone signals, as already presented in [3].

The B-format microphone signals [2] are composed of an omnidirectional signal $w(t)$, and three signals $x(t)$, $y(t)$, and $z(t)$ which

correspond to the output of three dipoles aligned with the x-, y-, and z-direction of a Cartesian coordinate system. Note that the omnidirectional signal is proportional to the sound pressure, whereas the dipole signals correspond to the components of the particle velocity vector.

Let $W(k,n)$ denote the short-time Fourier transform (STFT) of the omnidirectional signal at frequency index n and block time index k. Furthermore, we define the STFT-domain vector $\mathbf{V}(k,n)$ as

$$\mathbf{V}(k,n) = [X(k,n),\, Y(k,n),\, Z(k,n)]^{\mathrm{T}}, \tag{1}$$

where its elements represent the STFT of the corresponding time domain dipole signals. According to [3, 7], the STFT of the active intensity vector $\mathbf{I}(k,n)$ can be computed as

$$\mathbf{I}(k,n) = \frac{1}{\sqrt{2}\rho_o c}\mathrm{Re}\{W^*(k,n)\mathbf{V}(k,n)\}, \tag{2}$$

where ρ_o is the mean density of air and c is the speed of sound. Here, $(\cdot)^*$ denotes the conjugate of a complex number and $\mathrm{Re}\{\cdot\}$ corresponds to its real part. The DOA of sound is obtained from (1) as the opposite direction of $\mathbf{I}(k,n)$.

The frequency representation of the energy density is given by

$$E(k,n) = \frac{1}{2\rho_o c^2}\left[|W(k,n)|^2 + \frac{1}{2}\|\mathbf{V}(k,n)\|^2\right], \tag{3}$$

where it is assumed that the dipole signals are scaled by a factor of $\sqrt{2}$ compared to the magnitude of the omnidirectional signal. The diffuseness $\Psi(k,n)$ is defined as

$$\Psi(k,n) = 1 - \frac{\|\mathrm{E}\{\mathbf{I}(k,n)\}\|}{c\,\mathrm{E}\{E(k,n)\}}, \tag{4}$$

where $\mathrm{E}\{\cdot\}$ denotes the expectation operator. Regarding (1) and (3), the diffuseness can also be expressed using the B-format signals, i.e.,

$$\Psi(k,n) = 1 - \frac{\sqrt{2}\|\mathrm{E}\{\mathrm{Re}\{W^*(k,n)\mathbf{V}(k,n)\}\}\|}{\mathrm{E}\{|W(k,n)|^2 + \|\mathbf{V}(k,n)\|^2/2\}}. \tag{5}$$

In practice, $\mathrm{E}\{\cdot\}$ is replaced by temporal smoothing.

It should be pointed out that the estimation of the DOA is performed for each frequency and block time index independently, based on the instantaneous active intensity vector $\mathbf{I}(k,n)$. As a consequence, the estimates have a relatively high variance. Note that this variance is implicitly reduced by smoothing the directional information in the synthesis step of DirAC, i.e., when computing the loudspeaker signals. As discussed in [4], this approach leads to an improved listening experience compared to directly averaging directions, especially in case of the reproduction of transients.

3. PLANAR MICROPHONE ARRAY

As already mentioned in the introduction, the B-format microphone signals required for computing the DOA and diffuseness can be obtained from 3D microphone configurations composed of omnidirectional sensors [6]. In practice, however, planar microphone configurations are more preferable, since the mounting of the microphones and their integration into any housing is much easier to handle. However, with planar microphone arrays, not all of the required B-format signals are directly accessible. In this section we present a method to approximate a missing dipole signal in order to apply the efficient DirAC approach also for 2D arrays.

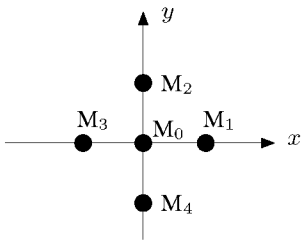

Fig. 1. Planar microphone configuration using five omnidirectional sensors.

In the following, we consider a planar microphone array composed of five omnidirectional microphones as illustrated in Fig. 1. The microphone configuration consists of four outer microphones placed on the corners of square. Furthermore, a fifth microphone is placed in the center of the array. Without loss of generality, we assume that the array lies on the xy-plane of a Cartesian coordinate system, where the outer microphones are aligned with the coordinate axes, and the center microphone is placed at the origin.

The omnidirectional signal and the dipole signals corresponding to the xy-plane can be computed from the planar microphone array according to

$$W(k,n) = P_0(k,n) \tag{6}$$
$$X(k,n) = K(n)\sqrt{2}\,[P_1(k,n) - P_3(k,n)] \tag{7}$$
$$Y(k,n) = K(n)\sqrt{2}\,[P_2(k,n) - P_4(k,n)], \tag{8}$$

where P_i denotes the sound pressure at the i-th microphone M_i. The frequency-dependent normalization factor $K(n)$ is given by

$$K(n) = -j\frac{cN}{d2\pi f_{\mathrm{s}} n}, \tag{9}$$

where j denotes the imaginary unit, N is the number of frequency bins of the STFT, f_{s} is the sampling rate, and d denotes the distance between two opposite outer microphones. Thus, $X(k,n)$ and $Y(k,n)$ are computed as the outputs of first-order differential microphone arrays [8].

3.1. Three-dimensional sound field analysis

Next, we look at approximating the missing dipole signal $Z(k,n)$ using measurable microphone signals only. Note that so far we did not state any assumption on the properties of the observed sound field. We now assume that the sound field consists of a single plane wave, impinging on the array with an azimuth angle φ and elevation ϑ. Then, in the ideal case, the dipole signals can be expressed by

$$X(k,n) = \sqrt{2}W(k,n)\cos\varphi\cos\vartheta \tag{10}$$
$$Y(k,n) = \sqrt{2}W(k,n)\sin\varphi\cos\vartheta \tag{11}$$
$$Z(k,n) = \sqrt{2}W(k,n)\sin\vartheta \tag{12}$$

Note that we omit the dependency of the angles on the frequency and time index for simplicity. Starting from (10)-(12) and making simple trigonometric considerations, we propose to approximate $Z(k,n)$ with $\tilde{Z}(k,n)$ defined as follows

$$\tilde{Z}(k,n) = \sqrt{2|W(k,n)|^2 - |T(k,n)|^2}\,\mathrm{e}^{j\phi_W(k,n)}, \tag{13}$$

where $\phi_W(k,n)$ denotes the phase of the omnidirectional signal $W(k,n)$. The term $|T(k,n)|$ corresponds to the frequency-domain magnitude of an auxiliary signal defined as

$$|T(k,n)| = \sqrt{|X(k,n)|^2 + |Y(k,n)|^2}. \quad (14)$$

Since the directivity pattern of $|T(k,n)|$ corresponds to a torus, it is referred to as *torus signal*. The rotation axis of the torus is aligned with the z-axis.

It is straightforward to show that in case of a single planar wave equation (13) simplifies to

$$\tilde{Z}(k,n) = \sqrt{2}W(k,n)\sin|\vartheta|. \quad (15)$$

As the equation above shows, by computing $\tilde{Z}(k,n)$ we can estimate the elevation angle ϑ, however with an ambiguity between the upper and lower hemispheres. This, however, does not represent a problem as common operation scenarios imply that the planar array is placed on a rigid surface, i.e., only positive elevation angles occur.

Obviously, the assumption of a single planar wave does not strictly hold in practice. In fact, even if only one sound source was active, the array would pick up multiple reflections due to reverberation. Thus, an improved model of the sound field would include the superposition of a planar wave and diffuse sound. However, to consider such an enhanced model goes beyond the scope of this paper.

The performance of the proposed method is now exemplified using a simulated room with moderate reverberation. The distance d is set to 2.5 cm. The set-up is as follows: Two speech sources are located at one meter from the microphone array, at azimuth and elevation angles (φ,ϑ) of $(60^\circ, 30^\circ)$ for the first speaker and $(10^\circ, 70^\circ)$ for the second one, respectively. The first speaker is active in the time frame $[0-4]$ seconds, whereas the second speaker in the range $[2-6]$ seconds. To simulate non-ideal microphones, white Gaussian noise with an SNR of 35 dB has been added to each microphone channel. In Fig. 2, the estimates of the instantaneous DOA are shown for each time/frequency tile, where a color coding of the angles has been used for illustration. Note that the DOA of time/frequency tiles

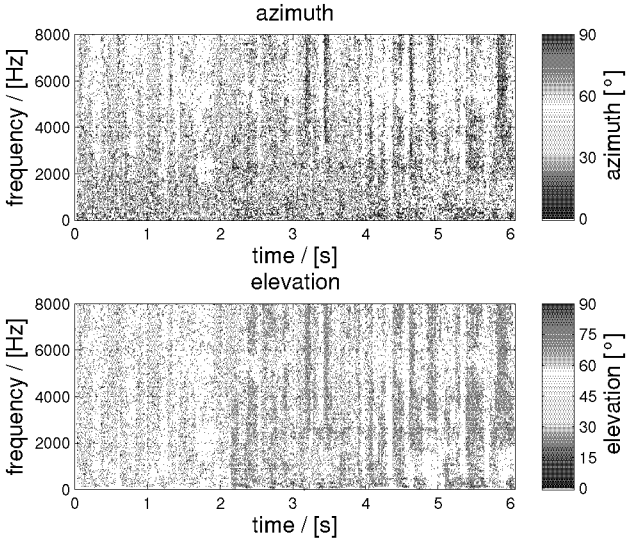

Fig. 2. Estimated instantaneous DOA of sound for a simulated double talk scenario.

have been discarded if the corresponding signal energy is negligible and, thus, would not have any effect on the loudspeaker signals after DirAC synthesis. As it can be seen, not only during single talk, but also during the double talk situation, the DOA of sound, including elevation, could be well estimated. Simulation results with respect to the estimation of diffuseness are presented in the next section.

3.2. Two-dimensional sound field analysis

In applications such as audio communication, the loudspeakers used for reproduction are commonly placed within one plane. In this case, only the azimuth angle has to be estimated in order to sufficiently describe the DOA of sound.

Since the azimuth angle can already be computed based on the x and y component of the intensity vector, only the $X(k,n)$ and $Y(k,n)$ dipole signals are required. A straightforward implementation of 2D DirAC is to simply discard the $Z(k,n)$ dipole signal also for the computation of the diffuseness according to (5). However, it is important to note that the omnidirectional signal $W(k,n)$ still includes all contributions of the 3D sound field arriving from the z direction. Thus, the approach of discarding the z component of the intensity vector leads to a significant overestimation of the diffuseness. This implies that due to an incorrect synthesis of the loudspeaker signals, the rendered sound field will also be far too diffuse. As easily verifiable, this effect is pronounced especially if a sound source is present at high elevation values.

In this contribution, we propose a modified diffuseness which explicitly takes into account the effect of the missing z-dipole. This is achieved by replacing the omnidirectional signal $W(k,n)$ in (5) by the corresponding torus signal

$$T(k,n) = |T(k,n)|\,\mathrm{e}^{j\phi_W(k,n)}. \quad (16)$$

The magnitude of $T(k,n)$ is given in (14). Then, (5) becomes

$$\Psi_{2\mathrm{D}}(k,n) = 1 - \frac{\sqrt{2}\|\mathrm{E}\{\mathrm{Re}\{T^*(k,n)\mathbf{V}_{2\mathrm{D}}(k,n)\}\}\|}{\mathrm{E}\{|T(k,n)|^2 + \|\mathbf{V}_{2\mathrm{D}}(k,n)\|^2/2\}}, \quad (17)$$

where the 2D vector $\mathbf{V}_{2\mathrm{D}}(k,n) = [X(k,n),\, Y(k,n)]^{\mathrm{T}}$.

From (5) it follows that the influence of $Z(k,n)$ on the diffuseness increases for increasing elevations of sound sources. To illustrate this effect, we consider the same simulation scenario as in the previous section, i.e., the first speaker is located at 30° and the second at 70° elevation, respectively. In Fig. 3, the estimated diffuseness using two different approaches is shown for each time/frequency tile. Again, we only consider estimates for which the signal energy was sufficiently high. Fig. 3(top) shows the case of simply discarding the z-component of the intensity for the computation of the diffuseness. As it can be seen, the estimate of the diffuseness is significantly overestimated as soon as the second talker, located at a high elevation angle, is active. This behavior is obviously not desired, since this abrupt change in diffuseness results in an abrupt change of the perceived acoustic properties of the recording room. From Fig. 3(bottom) we notice that this effect can be avoided, if the diffuseness is computed according to (17). Independently from the source elevation, the diffuseness keeps rather constant values as expected given the constant room characteristics.

4. EXPERIMENTAL RESULTS

In this section, we present DOA estimation results that have been obtained from recordings using a planar microphone array according to Fig. 1. The microphones used are small commercial omnidirectional

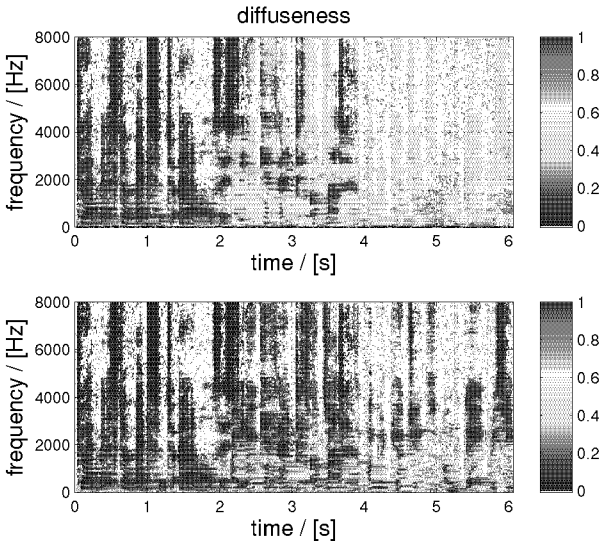

Fig. 3. Estimated diffuseness of a sound field using 2D sound field analysis (top) and the proposed method (bottom) for a simulated double talk scenario.

capsules, where $d = 2.5$ cm. The recording room has a reverberation time of approximately 150 ms. The sound sources have been placed at a distance of 1 m from the microphone array.

First, we look at the performance of the azimuth estimation. The sound field has been generated by playing a sequence of white noise signals from 7 different loudspeakers. They were placed at a constant elevation $\vartheta = 0$ and for azimuths between 0° and 90° with 15° step.

The result of the instantaneous azimuth estimation for each time/frequency tile is shown in Fig. 4. The DirAC approach is able

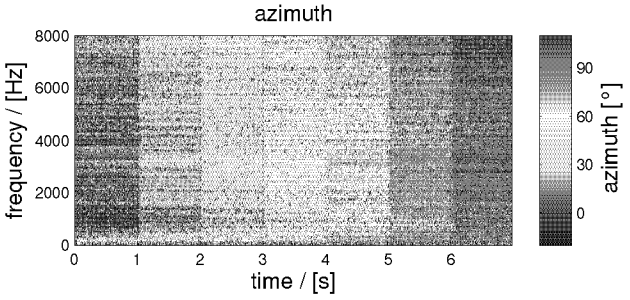

Fig. 4. Estimated azimuth angles using recorded microphone signals of a planar array.

to reliably estimate the DOA of sound within a large range of both, azimuth angles and frequencies.

Next, we consider the elevation estimation using the proposed method for planar arrays. The experimental set-up is similar to the one used for the azimuth. In this case, however, the azimuth is kept constant and the elevation angles varies from 0° to 90°. The resulting instantaneous elevation estimates are depicted in Fig. 5. Note that the computation of $|\tilde{Z}(k,n)|$ in (13) relies on the planar wave assumption. In practice, this assumption does not always sufficiently hold and $2|W(k,n)|^2 < |T(k,n)|^2$ occurs in some situations. This

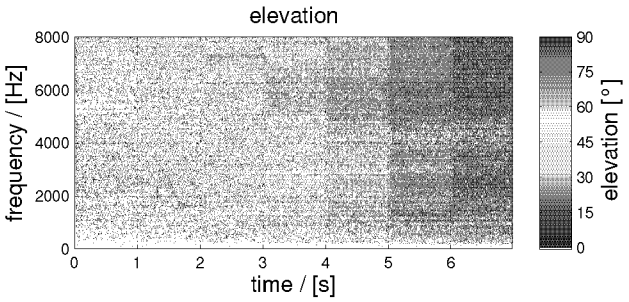

Fig. 5. Estimated elevation angles using recorded microphone signals of a planar array.

corresponds to a physically impossible case according to our model, making the observed signals not interpretable. Thus, the corresponding estimates are discarded in Fig. 5.

It should also be mentioned that for low elevations, a saturation of the estimated angles is observable. This can be explained by the superposition of two effects: On the one hand, there is a systematic bias due to the dipole approximation by a differential array [8]. On the other hand, the assumption of a single planar wave does not hold in the presence of diffuse sound, leading to a biased estimate of $|\tilde{Z}(k,n)|$. Methods for the correction of the corresponding estimation errors are out of the scope of this paper and therefore are not discussed here further.

5. CONCLUSION

In this contribution, we have presented an approach to DirAC parameter estimation using planar microphone arrays. The simulation results show that the azimuth and diffuseness can be reliably estimated for a wide range of sound source configurations. With the proposes method, also the 3D DOA of sound, i.e., including positive elevation angles can be estimated. The experimental results indicate that the accuracy of the estimate increases for increasing elevation angles.

6. REFERENCES

[1] J. Eargle, *The microphone book*, Boston: Focal Press, 2001.

[2] R. K. Furness, "Ambisonics - An overview," in *AES 8th International Conference*, April 1990, pp. 181–189.

[3] V. Pulkki and C. Faller, "Directional audio coding: Filterbank and STFT-based design," in *120th AES Convention,* Paper 6658, Paris, May 2006.

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

[5] L. Villemoes, J. Herre, and J. Breebaart et al., "MPEG Surround: The forthcoming ISO standard for spatial audio coding," in *AES 28th International Conference*, Piteå, June 2006.

[6] J. Merimaa, "Applications of a 3-D microphone array," in *112th AES Convention,* Paper 5501, Munich, May 2002.

[7] F. J. Fahy, *Sound intensity*, Essex: Elsevier Science Publishers Ltd., 1989.

[8] G. W. Elko, "Superdirectional microphone arrays," in *S. G. Gay, J. Benesty (eds.): Acoustic Signal Processing for Telecommunication*, Kluwer Academic Press, 2000.