Triplet-based variable-perspective (6DoF) audio rendering from simultaneous surround recordings taken at multiple perspectives

Christian Schörkhuber[1], Robert Höldrich[2], Franz Zotter[2]

[1] *sonible GmbH, Graz, Austria, Email: christian.schoerkhuber@sonible.com*

[2] *Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Austria,*

Introduction

Recently, Six-Degrees-of-Freedom (6DoF) audio recording and rendering approaches have been proposed to enable a variable-perspective playback for a virtual listener. We assume that a distributed sound scene is simultaneously recorded by various main microphone arrays positioned at multiple perspectives. Variable-perspective rendering enables the virtual subject to freely move within and listen to the (virtual)sound scene. Main representatives of such approaches use single-perspective recordings and map them onto an outer convex hull of the room [1, 2, 3], and moreover, there are works and patents about the interpolation from perspective recordings synchronously taken at multiple perspectives in the room, with parametric concepts to extract and render the sources detected therein and the diffuse or unlocalized parts [4, 5, 6, 7, 8]. Some works explicitly avoid short-term time-frequency-filtering for artifact-free baseline rendering [9, 10, 11, 12, 13, 6, 14], which, however, may stay limited in spatial precision.

This contribution establishes a theory in order to ensure directionally consistent variable-perspective rendering from multi-perspective recordings. While the theory is based on physical metrics such as the intensity and energy density, as, e.g. [15], it actually is meant to work for any kind of distributed-perspective spatial recordings, such using resolution-enhanced first-order Ambisonic microphone recordings (HARPEX [16] or DirAC [17]), higher-order Ambisonic microphone recordings (Eigenmike EM32 or Zylia ZM-1), or ESMA [18], or distributions out of different arrays, for instance. For a moment, let us assume that each of these recordings is able to consistently relate to energy densities and intensities reproduced at the listener to describe perceived sound levels and directions.

The goal of our theory is to obtain a nearly signal-independent strategy avoiding annoying artifacts that signal and position extraction of a complex recorded scene can entail. We assume the presence of a single source and a diffuse soundfield seen by any local triplet of recording perspectives. Related to triplet-based panning approaches such as [19, 20], our contribution shows a model involving linear mixing weights related to the aerial coordinates corrected by distance and diffuseness ratios. This linear combination rule to mix neighboring recording perspectives provides a consistent intensity vector direction at the listener. Moreover, the rendering yields reasonable and robust energy densities.

Descriptors of the recorded sound field

Omitting the constants $\frac{1}{\rho_0 c^2}$ or $\frac{1}{\rho_0 c}$, the acoustic energy density w and the intensity vector $\boldsymbol{i}$ at the position $\boldsymbol{r}$ are given by the short-term expectations, cf. [21]

$$w(\boldsymbol{r}) = \mathrm{E}\left\{p^2(\boldsymbol{r},t)\right\}, \quad \boldsymbol{i}(\boldsymbol{r}) = \mathrm{E}\left\{p(\boldsymbol{r},t)\boldsymbol{v}(\boldsymbol{r},t)\right\}, \quad (1)$$

where $p(\boldsymbol{r},t)$ and $\boldsymbol{v}(\boldsymbol{r},t)$ are the sound pressure and velocity observed, respectively.

Simple fields can be considered to contain a direct component $x(t)$ of a source at $\boldsymbol{s}$, observed at the distance $d = \|\boldsymbol{r}-\boldsymbol{s}\|$, and an uncorrelated diffuse part modeled as normally distributed noise

$$p(\boldsymbol{r},t) = \frac{x\left(t-c^{-1}d\right)}{d} + \sqrt{w_{\mathrm{diff}}}\,\mathcal{N}_p(\boldsymbol{r},t) \quad (2)$$

$$\boldsymbol{v}(\boldsymbol{r},t) = \frac{x\left(t-c^{-1}d\right)}{d}\frac{\boldsymbol{r}-\boldsymbol{s}}{d} + \sqrt{w_{\mathrm{diff}}}\,\boldsymbol{\mathcal{N}}_{\boldsymbol{v}}(\boldsymbol{r},t), \quad (3)$$

and so we obtain

$$w(\boldsymbol{r}) = \underbrace{\frac{\sigma^2}{d^2}}_{w_{\mathrm{dir}}} + w_{\mathrm{diff}}, \qquad \boldsymbol{i}(\boldsymbol{r}) = \sigma^2\frac{\boldsymbol{r}-\boldsymbol{s}}{d^3} \quad (4)$$

with $\sigma^2 = \mathrm{E}\left\{x(t)^2\right\}$. And the diffuseness ψ is defined as the ratio of diffuse and total energy density

$$\psi = \frac{w_{\mathrm{diff}}}{w}, \quad \text{hence } w_{\mathrm{diff}} = \psi\, w, \quad w_{\mathrm{dir}} = (1-\psi)\, w. \quad (5)$$

Model of superimposed perspectives

Given the recordings at three positions $\boldsymbol{r}_j$, for $j \in \{1,2,3\}$ in two space dimensions, our goal is to synthesize a recording at an arbitrary location $\boldsymbol{s}$ such that the resulting intensity vector is correct by applying a single interpolation weight g_j for each recording spot. Our hypothesis for interpolation is that pressure and velocity signals

$$p_j(t) = p(\boldsymbol{r}_j,t), \qquad \boldsymbol{v}_j(t) = \boldsymbol{v}(\boldsymbol{r}_j,t) \quad (6)$$

observed at different perspectives $i \neq j$ are uncorrelated $\mathrm{E}\left\{p_i(t)p_j(t)\right\} = 0$, $\mathrm{E}\left\{p_i(t)\boldsymbol{v}_j(t)\right\} = \boldsymbol{0}$. While for diffuse fields or complex audio scenes this assumption is most likely accurate, it implies that we assume direct sounds to be uncorrelated in the different recording perspectives, merely because of the acoustic flight time differences, regardless of the coherence to the source. To compile an estimated signal for a virtual listener, we superimpose the signals of the recorded perspectives by the weights g_j

$$\hat{p}_0(t) = \sum_j g_j\, p_j(t), \qquad \hat{\boldsymbol{v}}_0(t) = \sum_j g_j \boldsymbol{v}_j(t). \quad (7)$$

The resulting estimated energy density and intensity become under the uncorrelatedness assumptions

$$\hat{w}_0 = \sum_{i,j} g_i g_j \mathrm{E}\left\{p_i(t) p_j(t)\right\} = \sum_j w_j\, g_j^2 \qquad (8)$$

$$\hat{\boldsymbol{i}}(\boldsymbol{r}_0) = \sum_{i,j} g_i g_j \mathrm{E}\left\{p_i(t) \boldsymbol{v}_j(t)\right\} = \sum_j \boldsymbol{i}_j\, g_j^2. \qquad (9)$$

Note that such a stochastic superposition can also be enforced by special means when mixing soundfield playback from the different perspectives. For instance, if the perspectives are improved by parametric audio or whenever they do not use coinciding virtual playback systems, a more-or-less stochastic superposition can be assumed.

Source-position invariant interpolation

For a direct sound field of a source at the distance $d_j = \|\boldsymbol{r}_j - \boldsymbol{s}\|$, the interpolation should reconstruct $\hat{\boldsymbol{i}}_0 = \boldsymbol{i}_0$

$$\sum_j \boldsymbol{i}_j g_j^2 = \boldsymbol{i}_0$$

$$\sum_j \frac{\boldsymbol{r}_j - \boldsymbol{s}}{d_j^3} g_j^2 = \frac{\boldsymbol{r}_0 - \boldsymbol{s}}{d_0^3}$$

$$\sum_j (\boldsymbol{r}_j - \boldsymbol{s}) \frac{d_0^3}{d_j^3} g_j^2 = \boldsymbol{r}_0 - \boldsymbol{s}, \qquad (10)$$

which yields an underdetermined system of two linear equations for the three weights g_j^2, and thus has an infinite number of solutions. We may reshape the problem formulation into one that is shift-invariant with regard to the source location $\boldsymbol{s}$ by bringing it to the right-hand side

$$\sum_j \frac{d_0^3}{d_j^3} g_j^2 \boldsymbol{r}_j = \boldsymbol{r}_0 - \underbrace{\left[1 - \sum_j \frac{d_0^3}{d_j^3} g_j^2\right]} \boldsymbol{s}, \qquad (11)$$

and by zeroing its (underbraced) coefficient $\sum_j \frac{d_0^3}{d_j^3} g_j^2 = 1$. Substituting $a_j = \frac{d_0^3}{d_j^3} g_j^2$, this step reduces Eq. (11) to

$$\sum_j a_j \boldsymbol{r}_j = \boldsymbol{r}_0 \qquad \text{subject to:} \sum_j a_j = 1. \qquad (12)$$

For a triplet of positions, the solutions a_j are known as normalized barycentric or aerial coordinates. They are positive and limited between $0 \leq a_j \leq 1$ whenever the listener position $\boldsymbol{r}_0$ is located within the triplet of recording positions $\boldsymbol{r}_1$, $\boldsymbol{r}_2$, $\boldsymbol{r}_3$. The values are calculated from the listener position and recording positions by 2×2 matrix inversion

$$[\boldsymbol{r}_2 - \boldsymbol{r}_1,\ \boldsymbol{r}_3 - \boldsymbol{r}_1][a_2,\ a_3]^{\mathrm{T}} = \boldsymbol{r}_0 - \boldsymbol{r}_1$$

$$[\boldsymbol{r}_2 - \boldsymbol{r}_1,\ \boldsymbol{r}_3 - \boldsymbol{r}_1]^{-1}(\boldsymbol{r}_0 - \boldsymbol{r}_1) = [a_2,\ a_3]^{\mathrm{T}} \qquad (13)$$

and $a_1 = 1 - a_2 - a_3$. Together with the distance ratios, the mixing gains become

$$g_j = \sqrt{\frac{d_j^3}{d_0^3} a_j}. \qquad (14)$$

Practical and robust estimation of d_j/d_0

In practice, it might be difficult to robustly estimate the distances between the source and the recording perspectives, as well as the distance of the source to the virtual listener. Without having to rely on time-delay-based estimation or intensity-vector-based estimations that might fail in case of multiple sources, robust estimation of d_j/d_0 is achieved by short-time-averaged levels of $w_j = \mathrm{E}\left\{p_j^2\right\}$. If the diffuseness $\psi_j = 1 - \|\mathrm{E}\{p_j \boldsymbol{v}_j\}\|/\mathrm{E}\{p_j^2\}$ is estimated at the recording perspectives, it enables to isolate the direct energy density from the diffuse one, and we can write:

$$\frac{d_j}{d_0} \approx \sqrt{\frac{\sum_{i=1}^3 w_j\,(1-\psi_j)\,a_j}{w_j(1-\psi_j)}}; \qquad (15)$$

or we would need to assume $\psi_j = 0$ otherwise. The above expression estimates d_j^2 as being proportional to the direct energy density $\sigma^2/w_{\mathrm{dir},j} = \sigma^2/w_j\,(1-\psi_j)$, assuming there being a single direct sound source, only. When estimating d_0, our intention is to avoid loud amplitudes whenever the virtual listener is very close to the recorded source ($d_0 \rightarrow 0$). Therefore, we estimate d_0^2 by the reciprocal of the aerial-coordinate superimposed energy densities over the signal variance, i.e. $\sigma^2/\sum_j w_j\,(1-\psi_j)\,a_j$. In the ratio d_j/d_0 the signal variance cancels, and finally our robust estimator does not require any localization algorithm to actually determine the source position $\boldsymbol{s}$. It is just based on level and diffuseness estimation.

Note that if assuming $\psi_j = 0$ to simplify the estimator, a constant diffuse field would neutralize the estimated ratio $d_j/d_0 \rightarrow 1$, and hereby would actually provide a reasonable rendering of diffuse fields; and yet, the directional rendering of direct fields will be distorted. By contrast, if the diffuseness ψ_j can be estimated at the receivers, directional distortion of direct sounds can be kept small, however, the estimator becomes less robust and less level-preserving in the case of purely diffuse fields with $\psi_j \rightarrow 1$. Any kind of limitation of ψ_j to values smaller than 1 can restore robustness.

Simulation Study

Our simulations will now study how well the proposed sound field interpolation works, when regarding the direction of the interpolated intensity $\hat{\boldsymbol{i}}_0$ compared to $\boldsymbol{i}_0$ of the original sound field, the interpolated level $\hat{w}_0$ compared to w_0, and interpolated diffuseness $\hat{\psi}_0 = 1 - \|\hat{\boldsymbol{i}}_0\|/\hat{w}_0$ compared to $\psi_0 = 1 - \|\boldsymbol{i}_0\|/w_0$. The study considers a free-field source at $\boldsymbol{s}$ which is superimposed by a diffuse sound field, related by a signal-to-diffuse ratio $\mathrm{SDR} = \sigma^2/w_{\mathrm{diff}}$. This scene is captured at the recording spots $\boldsymbol{r}_1 = [0, 0]^T$, $\boldsymbol{r}_2 = [2, 0]^T$, $\boldsymbol{r}_3 = [1, 2]^T$. Figure 1 shows the interpolation results of the proposed method for different source positions and SDRs evaluated on a curved path through the triangle using i) the true (unknown) distance ratios assuming $\psi_j = 0$, ii) the estimated distance ratios assuming $\psi_j = 0$, and iii) the estimated distance ratios with accurate diffuseness ψ_j estimated at the recording spots. The upper two rows show the results for a source outside the triangle located at $\boldsymbol{s} = [1.9, 1.3]^T$ and an SDR of 90 dB and 0 dB, respectively. The high SDR case in the first row (a)-(d) is included to show that this condition would enable a distortion-free interpolation purely based on level estimation, however with a tendency of a more diffuse rendering by the interpolation. The total energy density in (d) is flattened by the proposed estimation strategy and hereby more robust than the accurate model. In Fig. 1(e)-(h) and (i)-(l) with the SDR of 0 dB, we see the benefit in the directional mapping (f)(j) for source positions inside and outside the triplet when providing a 3-point diffuseness estimation ψ_j. The interpolation combines intensities that are not aligned and therefore cannot recover the low diffuseness of the target (g)(k). However, the proposed estimation methods are more robust than accurate knowledge of $\frac{d_j}{d_0}$.

Fig. 2 shows the interpolation results for the entire interior of the recording triplet for the same source positions and a weak SDR of 0 dB. The left column shows the target sound field where the background color indicates the total energy density, the direction of the arrows corresponds to the negative direction of the intensity vector, and the length of the arrows represents the directedness (1 - ψ_0); columns two to four display the interpolation results of the proposed method. It becomes obvious from (b)(g) that despite matching directions, the levels will not match those in (a)(f) with a correct estimate of d_0/d_j when $\psi_j = 0$. Moreover, (c)(h) show that d_0/d_j cannot be estimated well enough under diffuse conditions with zeroed diffuseness estimates $\psi_j = 0$; (d)(i) show the proposed robust estimation including ψ_j providing directionally consistent interpolation whose energy densities are well-behaved compared to (b)(g), and yet somewhat free and flatter. None of the interpolation methods is able to restore the vector lengths of the slightly more directional target (a)(f).

Conclusions

We presented a powerful and generic theory enabling a directionally consistent and robust interpolation of distributed 3D audio recordings that we suppose to work regardless of the specific array (FOA/HOA Ambisonics, ESMA) or virtual rendering/upmixing (concentric loudspeaker setups, HARPEX, DirAC) technology used. Our theory assumes perceived direction, diffuseness, loudness of all these solutions to be consistent and reasonably modeled by physical intensities and energy densities. If the linear mix of perspectives is stochastic, the theory permits consistent directional rendering by measuring level and diffuseness in a triplet of recordings. Due to its simplicity, it can serve as high-quality default solution to otherwise fully parametric, jointly object localizing approaches.

References

[1] V. Pihlajamäki, Tapani; Pulkki, "Synthesis of complex sound scenes with transformation of recorded spatial sound in virtual reality," *JAES*, vol. 7/8, no. 63, 2015.

[2] A. Plinge, S. J. Schlecht, O. Thiergart, T. Robotham, O. Rummukainen, and E. A. P. Habets, "Six-degrees-of-freedom binaural audio reproduction of first-order ambisonics with distance information," in *AES Int. Conf. Audio f. Virt. and Aug. Reality*, 2018.

[3] A. Allen and B. Kleijn, "Ambisonic soundfield navigation using directional decomposition and path distance estimation," in *Proc. ICSA*, Graz, 2017.

[4] G. D. Galdo, O. Thiergart, T. Weller, and E. Habets, "Generating virtual microphone signals using geometrical information gathered by distributed arrays," in *Proc. IEEE Workshop HSCMA*, Edinburgh, 2011.

[5] O. Thiergart, G. D. Galdo, M. Taseska, and E. Habets, "Geometry-based spatial sound acquisition using distributed microphone arrays," *IEEE TASLP*, vol. 21, no. 12, 2013.

[6] J. G. Tylka and E. Y. Choueiri, "Domains of practical applicability for parametric interpolation methods of virtual sound field navigation," *JAES*, vol. 67, no. 11, 2019.

[7] ——, "Performance of linear extrapolation methods for virtual sound field navigation," *JAES*, vol. 68, no. 3, 2020.

[8] ——, "Fundamentals of a parametric method for sound field navigation within an array of ambisonics microphones," *JAES*, vol. 68, no. 3, 2020.

[9] P. Grosche, F. Zotter, C. Schörkhuber, M. Frank, and R. Höldrich, "Method and apparatus for acoustic scene playback," *Patent*, WO 2018/077379 A1, 2018.

[10] T. Deppisch and A. Sontacchi, "Browser application for virtual audio walkthrough," in *Forum Media Technology & All Around Audio*, St. Pölten, 2017.

[11] D. Rudrich, M. Frank, and F. Zotter, "Evaluation of interactive localization in virtual acoustic scenes," in *Fortschritte der Akustik - DAGA*, Kiel, 2017.

[12] E. Patricio, A. Ruminński, A. Kuklasiński, Ł. Januszkiewicz, and T. Żernicki, "Toward six degrees of freedom audio recording and playback using multiple ambisonics sound fields," in *prepr. 10141 AES Conv*, Dublin, 2019.

[13] D. R. Méndez, C. Armstrong, J. Stubbs, M. Stiles, and G. Kearney, "Practical recording techniques for music production with six-degrees of freedom virtual reality," in *prepr. 464 AES Conv.*, New York, 2018.

[14] N. Mariette and B. F. Katz, "Sounddelta – large scale, multi-user audio augmented reality," in *EAA Symposium on Auralization*, Espoo, June 2009.

[15] M. A. Gerzon, "General metatheory of auditory localisation," *prepr.3306, 92nd AES Conv*, Vienna 1992.

[16] S. Berge and N. Barrett, "A new method for b-format to binaural transcoding," in *Proc 40th Int AES Conf.*, Tokyo, 2010.

[17] V. Pulkki, "Spatial sound reproduction with directional audio coding," *JAES*, vol. 55, no. 6, 2007.

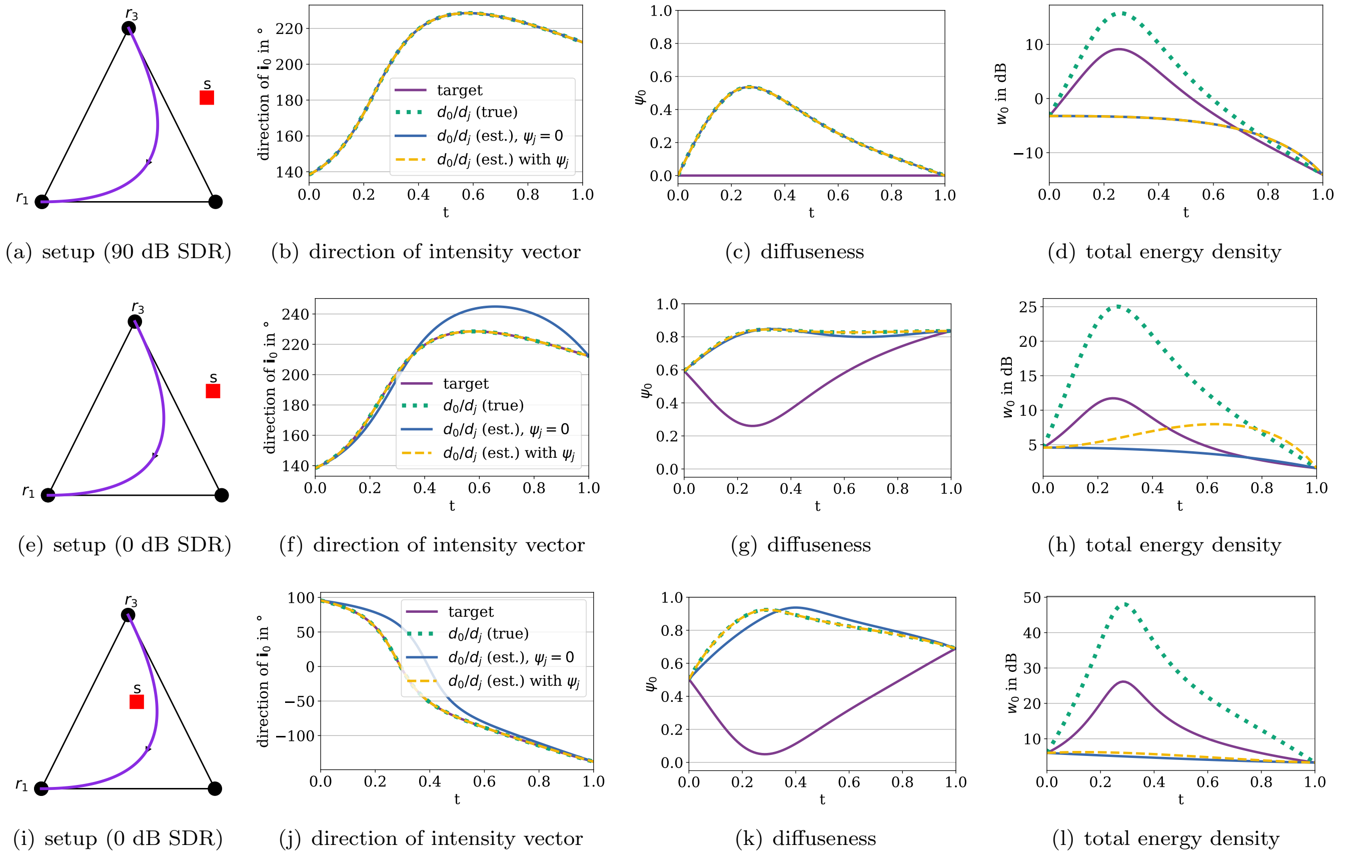
[18] H. Lee, "Capturing 360° audio using an equal segment microphone array (esma)," *JAES*, vol. 67, no. 1/2, 2019.

[19] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *JAES*, vol. 45, no. 6, 1997.

[20] C. Borß, "A polygon-based panning method for 3d loudspeaker setups," in *prepr. 9106, 137th AES Conv.*, Los Angeles, October 2014.

[21] H. Kuttruff, *Room Acoustics*, 6th ed, CRC Press, 2016.

[22] A. Franck, W. Wang, and F. M. Fazi, "Sparse l1-optimal multiloudspeaker panning and its relation to vector base amplitude panning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 996–1010, 2017.

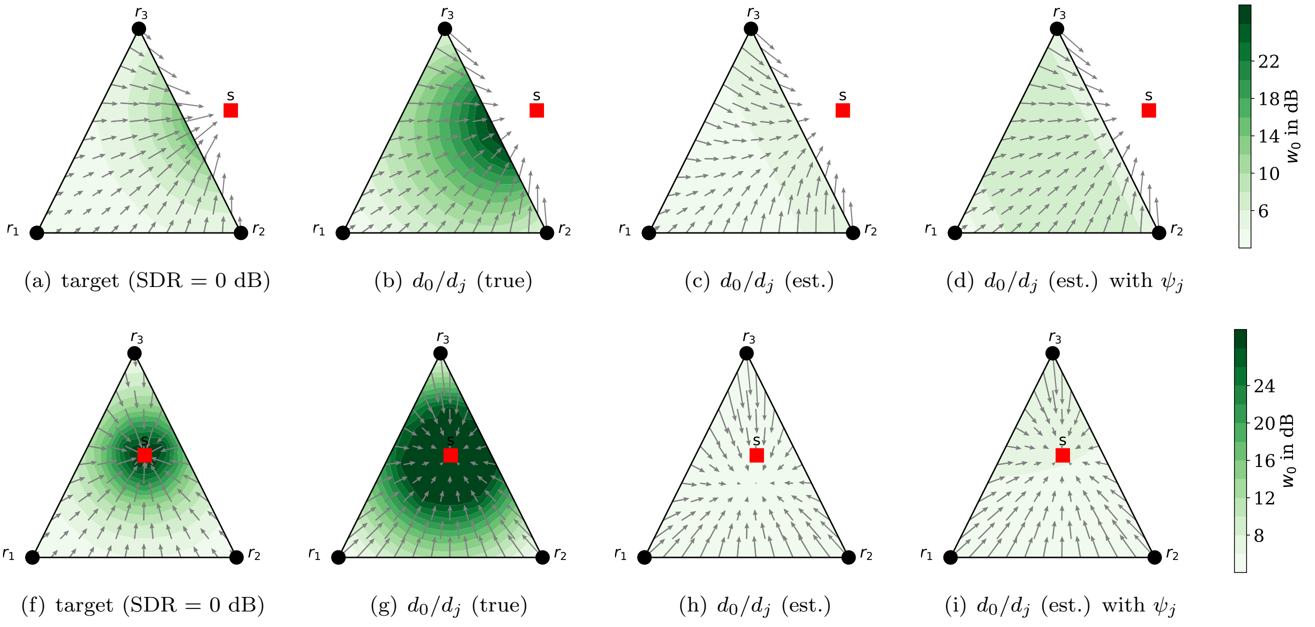


(a) setup (90 dB SDR) (b) direction of intensity vector (c) diffuseness (d) total energy density

(e) setup (0 dB SDR) (f) direction of intensity vector (g) diffuseness (h) total energy density

(i) setup (0 dB SDR) (j) direction of intensity vector (k) diffuseness (l) total energy density

Figure 1: Sound field interpolation results along a curved path. Upper two rows show interpolation results for a source outside the recording triplet $\boldsymbol{s} = [1.9, 1.2]$ and SDR of 90 dB (a)-(d) vs. 0 dB (e)-(h). Bottom row: source located inside triplet at $\boldsymbol{s} = [1.1, 1.0]^T$, SDR is 0 dB. Left column: layout of listener path wrt. recording positions and source location. Columns 2 to 4: interpolation results of the proposed method along the path compared to the target sound field parameters, where interpolation results are reported for using i) the true (unknown) distance ratios, ii) the estimated distance ratios with $\psi_j = 0$, and iii) the estimated distance ratios ${}^{d_0}/{}_{d_j}$ when the diffuseness ψ_j is accurately estimated.



(a) target (SDR = 0 dB) (b) d_0/d_j (true) (c) d_0/d_j (est.) (d) d_0/d_j (est.) with ψ_j

(f) target (SDR = 0 dB) (g) d_0/d_j (true) (h) d_0/d_j (est.) (i) d_0/d_j (est.) with ψ_j

Figure 2: Sound field interpolation results. In the upper two rows, the interpolation results are depicted for a source outside the recording triplet at $\boldsymbol{s} = [2.4, 1.3]$ and an SDR 0 dB. In the bottom row the source is located inside the triplet at $\boldsymbol{s} = [1.1, 1.0]^T$. The left column shows the target sound field, the background color the total energy density, the arrows are the opposite-sign intensity vectors, their length correspond to directedness (1 - ψ_0). Columns 2 and 4 show the interpolation results using the same estimators as in Fig. 1.