Signal-Dependent Encoding for First-Order Ambisonic Microphones

Christian Schörkhuber, Franz Zotter, Robert Höldrich
Institute of Electronic Music and Acoustics, University of Music and Performing Arts, Graz
schoerkhuber@iem.at

Introduction

Encoding spatial audio recordings in the Ambisonic format is a popular means to decouple the recording setup from the targeted rendering setup. This is achieved by a frequency-dependent linear transformation of the raw microphone signals into a set of virtual microphone signals, where the characteristics of the virtual microphones correspond to spherical harmonics up to a given order. The advantage of this set of orthogonal virtual microphones is that sound field rotations can be easily implemented by simple linear operations [1]. This property is especially appealing for virtual reality and 360° video applications as dynamic binaural rendering can be implemented by combining dynamic sound field rotations in the spherical harmonics domain and static binaural reproduction using a fixed set of HRTFs corresponding to a set of virtual loudspeakers.

To record 3-dimensional sound fields in Ambisonics, spherical microphone arrays are used and the number of microphones that need to be employed is determined by the desired Ambisonic order. For *First-Order Ambisonics* (FOA) only 4 microphones are required and affordable spherical arrays with high individual transducer quality are available; however, the usable frequency range is limited by the spatial aliasing frequency, which is determined by the array radius. When the recorded sound field is reproduced, errors that are introduced above the spatial aliasing frequency cause signal colourations as well as erroneous spatial cues that would also cause artefacts of direction enhancers such as Harpex [2] or DirAC [3]. To enable full-band FOA encoding, we propose a signal-dependent method where the encoding matrix dynamically adapts to estimated sound field parameters.

First-Order Ambisonic Encoding

Spherical harmonics (SHs) are a set of orthogonal basis functions for square integrable functions on the unit sphere, given by

$$Y_n^m(\Omega) = \begin{cases} N_n^m \cos(m\phi) P_n^m(\cos\theta) & \text{if } m \geq 0, \\ N_n^{|m|} \sin(|m|\phi) P_n^{|m|}(\cos\theta) & \text{if } m < 0, \end{cases}$$

where $\Omega = (\theta, \phi)$ defines a point on the unit sphere in spherical coordinates and $0 < m < M$, $-m \leq n \leq m$ is the SH order and degree, respectively, and P_n^m is the associated Legendre function. The normalization term N_n^m depends on the choice of convention [4].

The so-called B-format of First-Order Ambisonics contains spherical harmonic directivities up to order $M = 1$, resulting in 4 virtual microphone signals corresponding to an omnidirectional receiver (*W channel*) and three orthogonal dipoles (*X, Y, Z channels*) (see Figure 1). The 4 channels of tetrahedral microphone arrays, often called A-format, offer a technically compelling solution to record acoustic scenes. The microphone signals are converted into B-format signals applying the transformation

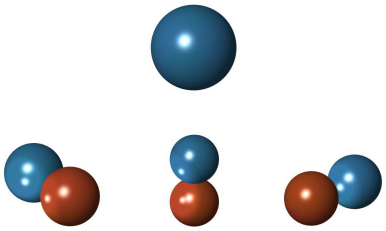


Figure 1: Real-valued spherical harmonics up to order 1. Magnitude is encoded in the radius, the phase (sign) is colour-coded, where blue refers to 0° and red refers to $\pm\pi$.

$$\boldsymbol{z}(\omega, t) = \boldsymbol{W}(\omega)\boldsymbol{x}(\omega, t), \tag{1}$$

where $\boldsymbol{z}(\omega,t) = [z_0^0(\omega,t)\ \ z_1^{-1}(\omega,t)\ \ z_1^0(\omega,t)\ \ z_1^1(\omega,t)]^{\mathrm{T}}$ is the B-format signal vector, $\boldsymbol{x}(\omega,t) = [x_1(\omega,t)\ \ x_2(\omega,t)\ \ x_3(\omega,t)\ \ x_4(\omega,t)]^{\mathrm{T}}$ is the microphone signal vector, $\boldsymbol{W}(\omega)$ is the conversion matrix, which is referred to as the encoder in the realm of Ambisonics, and ω, t index frequency and time, respectively. A standard encoder according to [5] is given by

$$\boldsymbol{W}(\omega) = \mathrm{diag}\left(\boldsymbol{b}(\omega, r)\right)^{-1} \boldsymbol{Y}_e^{\mathrm{H}}, \tag{2}$$

where $\boldsymbol{Y}_e = [\boldsymbol{y}(\Omega_1)\ \boldsymbol{y}(\Omega_2)\ \boldsymbol{y}(\Omega_3)\ \boldsymbol{y}(\Omega_4)]$, with $\boldsymbol{y}(\Omega_l) = [Y_0^0(\Omega_l)\ Y_1^{-1}(\Omega_l)\ Y_1^0(\Omega_l)\ Y_1^1(\Omega_l)]^{\mathrm{T}}$, is a frequency independent matrix containing the spherical harmonics up to order 1 evaluated at the microphone positions Ω_l. The vector $\boldsymbol{b}(\omega, r)$ contains frequency-dependent radial filters depending on the array radius r.

Spatial Aliasing

The A- to B-format conversion obtained by the standard encoder in Eq. (2) is only valid up to the spatial aliasing frequency ω_a [6], and for A-format microphones, this frequency is approximated by $kr = 1$, where $k = \omega/c$ and c is the speed of sound; hence

$$\omega_a \approx \frac{c}{r}. \tag{3}$$

For frequencies above the spatial aliasing frequency, higher order SHs are aliased to lower orders, hence the pick-up patterns of the virtual microphones W, X, Y, and Z start to deviate from the desired patterns as shown in Figure 1

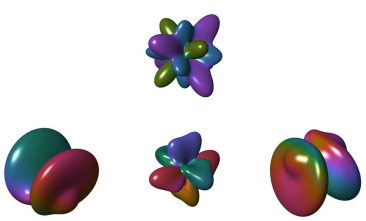


Figure 2: Pick-up patterns of the virtual microphones W (top), Y, Z, and X (bottom: left to right) for $kr = 3.5$. Magnitude is encoded in the radius, the phase is colour-coded.

and Figure 2. The distorted pick-up patterns in Figure 2 are simulated for $kr = 3.5$. Equation (3) signifies that the usable frequency range of A-format microphones can only be extended by reducing the array radius; however, there is a lower bound to the array radius when it comes to signal-to-noise ratio and transducer directivity.

Signal-Dependent Encoder

To enable FOA encoding above the spatial aliasing frequency, we employ a parametric sound field model, where the array signals are modelled as

$$\boldsymbol{x}(\omega,t) \approx \sum_{q}^{Q \ll S} s_q(\omega,t)\boldsymbol{v}(\Omega_q,\omega) + \boldsymbol{d}_x(\omega,t), \tag{4}$$

where s_q is the q-th source signal at the centre of the array, $\boldsymbol{v}(\Omega,\omega)$ is the far-field array response vector in direction Ω, $\boldsymbol{d}_x(\omega,t)$ is a diffuse signal vector, Q is the number of sources active in the time-frequency tile indexed by (ω,t), and S is the total number of sources present in the recorded scene. The assumption that $Q \ll S$ relates to the spectral disjointness between source signals [7], i.e. we assume that in each time-frequency tile very few sources are active.

Similarly, we model the targeted ideal FOA signals as

$$\boldsymbol{z}(\omega,t) \approx \sum_{q}^{Q \ll S} s_q(\omega,t)\boldsymbol{y}(\Omega_q) + \boldsymbol{d}_z(\omega,t). \tag{5}$$

Here we assume that $Q = 1$, hence skip the subscript for source signals and directions, and we moreover assume that the source signal is uncorrelated with the diffuse signals. Under these assumptions, the array and FOA signal covariance matrices become

$$\begin{aligned}\boldsymbol{R}_x(\omega) &= E\left[\boldsymbol{x}(\omega,t)\boldsymbol{x}(\omega,t)^{\mathrm{H}}\right] \qquad (6)\\ &= \sigma_s^2\boldsymbol{v}(\Omega,\omega)\boldsymbol{v}(\Omega,\omega)^{\mathrm{H}} + \sigma_d^2\boldsymbol{\Lambda}(\omega),\\ \boldsymbol{R}_z(\omega) &= E\left[\boldsymbol{z}(\omega,t)\boldsymbol{z}(\omega,t)^{\mathrm{H}}\right] \qquad (7)\\ &= \sigma_s^2\boldsymbol{y}(\Omega)\boldsymbol{y}(\Omega)^{\mathrm{H}} + \sigma_d^2\boldsymbol{I},\end{aligned}$$

where σ_s^2 and σ_d^2 is the power of the direct and diffuse signal, respectively, $\boldsymbol{\Lambda}(\omega)$ is the microphone signal covariance matrix of a uniform diffuse field determined by the array geometry and the microphone characteristics, and $\boldsymbol{I}$ is the identity matrix.

For the encoder $\boldsymbol{W}(\omega)$,

$$\boldsymbol{W}(\omega)\boldsymbol{R}_x(\omega)\boldsymbol{W}(\omega)^{\mathrm{H}} = \boldsymbol{R}_z(\omega) \tag{8}$$

must hold. Inserting Eq. (6) and Eq. (7) into Eq. (8) yields

$$\begin{aligned}\boldsymbol{W}\left(\sigma_s^2\boldsymbol{v}(\Omega)\boldsymbol{v}(\Omega)^{\mathrm{H}} + \sigma_d^2\boldsymbol{\Lambda}\right)\boldsymbol{W}^{\mathrm{H}} &=\\ = \sigma_s^2\boldsymbol{y}(\Omega)\boldsymbol{y}(\Omega)^{\mathrm{H}} + \sigma_d^2\boldsymbol{I}; &\qquad (9)\end{aligned}$$

for the sake of readability, the dependency on ω is not notated, here and below. To avoid the need to estimate the signal-to-diffuse ratio $\Gamma = \sigma_s^2/\sigma_d^2$ and to enforce a distortionless response for the source direction, we split Eq. (9) into two constraints imposed on the encoder. The *directional constraint* is given by

$$\boldsymbol{W}\boldsymbol{v}(\Omega) = \boldsymbol{y}(\Omega), \tag{10}$$

and the *orthogonality constraint* is given by

$$\boldsymbol{W}\boldsymbol{\Lambda}\boldsymbol{W}^{\mathrm{H}} = \boldsymbol{I}. \tag{11}$$

The following paragraphs present a robust closed-form solution for the signal-dependent encoder $\boldsymbol{W}(\Omega,\omega,t)$ that meets both constraints, assuming that the instantaneous direction-of-arrival (DOA) $\Omega(\omega,t)$ is known or can be estimated.

Optimal Encoder

We derive the optimal solution for the encoder in three steps: (i) by defining a parametrized set of solutions fulfilling the orthogonality constraint, (ii) by defining a parametrized subset of these solutions that also fulfil the directional constraint, and (iii) by selecting a particular closed-form solution in this subset that is robust to DOA estimation errors.

Orthogonality constraint. Applying the eigendecomposition $\boldsymbol{\Lambda} = \boldsymbol{U}_x\boldsymbol{S}_x\boldsymbol{U}_x^{\mathrm{H}}$, the set of solutions for Eq. (11) is given by [8, 9]

$$\boldsymbol{W} = \boldsymbol{P}\boldsymbol{S}_x^{-\frac{1}{2}}\boldsymbol{U}_x^{\mathrm{H}}, \tag{12}$$

where $\boldsymbol{P}$ is an arbitrary unitary matrix.

Directional constraint. Inserting Eq. (12) into Eq. (10) we obtain

$$\boldsymbol{P}\boldsymbol{q}(\Omega) = \boldsymbol{y}(\Omega), \tag{13}$$

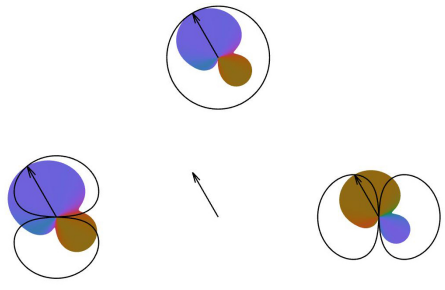
with

$$\boldsymbol{q}(\Omega) = \boldsymbol{S}_x^{-\frac{1}{2}}\boldsymbol{U}_x^{\mathrm{H}}\boldsymbol{v}(\Omega). \tag{14}$$

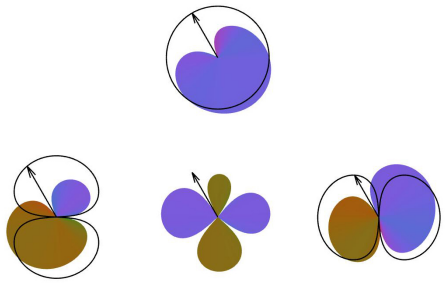
So the remaining task is to find a unitary matrix $\boldsymbol{P}$ that maps $\boldsymbol{q}(\Omega)$ to $\boldsymbol{y}(\Omega)$. It can be shown that $||\boldsymbol{q}(\Omega)||_2 \approx ||\boldsymbol{y}(\Omega)||_2$ for the tetrahedral array configuration; hence, we can write

$$\boldsymbol{P}\bar{\boldsymbol{q}}(\Omega) = \bar{\boldsymbol{y}}(\Omega), \tag{15}$$

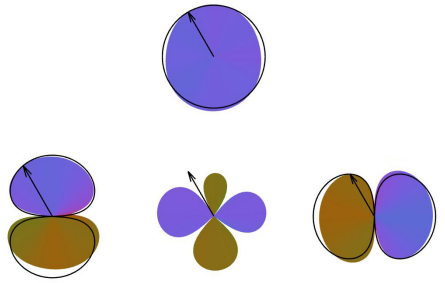
where $\bar{\boldsymbol{q}}(\Omega) = \boldsymbol{q}(\Omega)/||\boldsymbol{q}(\Omega)||_2$ and $\bar{\boldsymbol{y}}(\Omega) = \boldsymbol{y}(\Omega)/||\boldsymbol{y}(\Omega)||_2$.



(a) direct encoder



(b) orthogonal encoder



(c) resulting encoder

Figure 3: Pick-up patterns obtained by the direct encoder $\boldsymbol{W}_{\text{dir}}$ (a), the orthogonal encoder $\boldsymbol{W}_{\text{orth}}$ (b), and the resulting encoder $\boldsymbol{W} = \boldsymbol{W}_{\text{dir}} + \boldsymbol{W}_{\text{orth}}$ (c) for $f = 3000$ Hz and $\Omega = (\pi/2, 2\pi/3)$.

To obtain a parametrized set of solutions, we define the unitary matrices

$$\begin{aligned} \boldsymbol{Q}_y(\Omega) &= \left[\bar{\boldsymbol{y}}(\Omega)\ \boldsymbol{N}_y(\Omega)^{\mathrm{H}}\boldsymbol{\Theta}_y\right] \qquad (16)\\ &= \bar{\boldsymbol{y}}(\Omega)\boldsymbol{e}_1^{\mathrm{H}} + \boldsymbol{N}_y(\Omega)^{\mathrm{H}}\boldsymbol{\Theta}_y\boldsymbol{E}^{\mathrm{H}} \\ \boldsymbol{Q}_q(\Omega) &= \left[\bar{\boldsymbol{q}}(\Omega)\ \boldsymbol{N}_q(\Omega)^{\mathrm{H}}\boldsymbol{\Theta}_q\right] \qquad (17)\\ &= \bar{\boldsymbol{q}}(\Omega)\boldsymbol{e}_1^{\mathrm{H}} + \boldsymbol{N}_q(\Omega)^{\mathrm{H}}\boldsymbol{\Theta}_q\boldsymbol{E}^{\mathrm{H}}, \end{aligned}$$

where $\boldsymbol{e}_j$ is a unit vector along the j-th dimension, $\boldsymbol{E} = [\boldsymbol{e}_2\ \boldsymbol{e}_3\ \boldsymbol{e}_4]$, $\boldsymbol{N}_y$ and $\boldsymbol{N}_q$ are 3×4 matrices containing orthonormal basis vectors for the orthogonal complement of $\bar{\boldsymbol{y}}(\Omega)$ and $\bar{\boldsymbol{q}}(\Omega)$, respectively, and $\boldsymbol{\Theta}_y$, $\boldsymbol{\Theta}_q$ are arbitrary unitary 3×3 matrices. As both $\boldsymbol{Q}_y(\Omega)$ and $\boldsymbol{Q}_q(\Omega)$ are unitary, the set of solutions for Eq. (15) is given by

$$\begin{aligned} \boldsymbol{P}(\boldsymbol{\Theta}) &= \boldsymbol{Q}_y(\Omega)\boldsymbol{Q}_q(\Omega)^{\mathrm{H}} \qquad (18)\\ &= \bar{\boldsymbol{y}}(\Omega)\bar{\boldsymbol{q}}(\Omega)^{\mathrm{H}} + \boldsymbol{N}_y(\Omega)^{\mathrm{H}}\boldsymbol{\Theta}\boldsymbol{N}_q(\Omega), \end{aligned}$$

where $\boldsymbol{\Theta}$ is an arbitrary unitary 3×3 matrix.

Robust solution. Since every matrix $\boldsymbol{P}$ obtained by Eq. (18) meets the directional constraint in Eq. (10), the remaining degrees of freedom, parametrized by the unitary matrix $\boldsymbol{\Theta}$, can be used to optimize additional criteria. For example, we can choose a solution that is robust with respect to DOA estimation errors by minimizing the error for a spread of angles around the estimated DOA. By defining the matrices

$$\boldsymbol{V}_c = [\boldsymbol{v}(\Omega_1), \cdots, \boldsymbol{v}(\Omega_K)] \qquad (19)$$

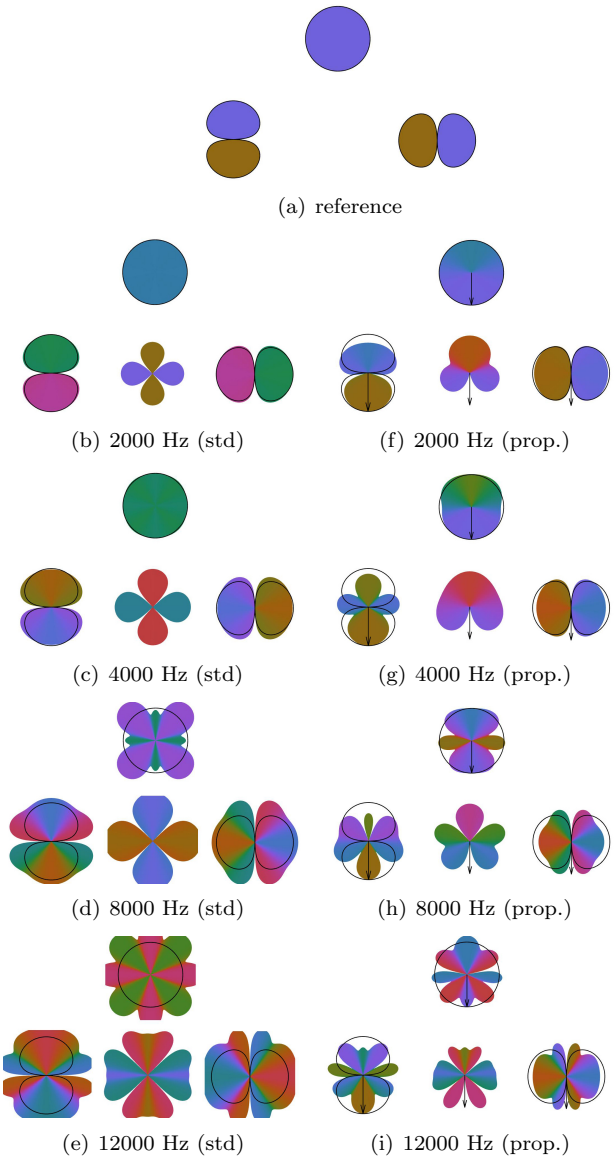
$$\boldsymbol{Y}_c = [\boldsymbol{y}(\Omega_1), \cdots, \boldsymbol{y}(\Omega_K)], \qquad (20)$$

where Ω_k, $k \in \{1, \cdots, K\}$ is a set of directions around the estimated DOA Ω, we can find the optimal choice for $\boldsymbol{\Theta}$ by solving

$$\boldsymbol{\Theta}^* = \arg\min_{\boldsymbol{\Theta}} ||\boldsymbol{P}(\boldsymbol{\Theta})\boldsymbol{S}_x^{-\frac{1}{2}}\boldsymbol{U}_x^{\mathrm{H}}\boldsymbol{V}_c - \boldsymbol{Y}_c||_{\mathrm{F}}^2 \qquad (21)$$

$$\text{subject to } \boldsymbol{\Theta}\boldsymbol{\Theta}^{\mathrm{H}} = \boldsymbol{I},$$

where $||\cdot||_{\mathrm{F}}$ is the Frobenius norm of a matrix. This non-convex constrained problem can be cast as an unconstrained optimization problem on the Stiefel manifold [10] and a local optimizer can be found by iterative methods [11]. However, in order to get a closed-form solution



(a) reference

(b) 2000 Hz (std) (f) 2000 Hz (prop.)

(c) 4000 Hz (std) (g) 4000 Hz (prop.)

(d) 8000 Hz (std) (h) 8000 Hz (prop.)

(e) 12000 Hz (std) (i) 12000 Hz (prop.)

Figure 4: Comparison of pick-up patterns for different frequencies and a horizontal 2D slice ($\theta = \pi/2$, $-\pi < \phi < \pi$). (a) ideal patterns, (b-e) patterns of the standard encoder, (f-i) patterns of the proposed encoder with $\Omega = (\pi/2, -\pi/2)$.

instead, we compute the unconstrained least squares solution

$$\tilde{\boldsymbol{\Theta}}^* = \boldsymbol{N}_y^{\mathrm{H}} \bar{\boldsymbol{Y}}_c \left(\boldsymbol{N}_q \boldsymbol{S}_x^{-\frac{1}{2}} \boldsymbol{U}_x^{\mathrm{H}} \boldsymbol{V}_c \right)^{\dagger} \tag{22}$$

where † denotes the pseudo-inverse of a matrix, and

$$\bar{\boldsymbol{Y}}_c = \boldsymbol{Y}_c - \bar{\boldsymbol{y}}(\Omega) \bar{\boldsymbol{q}}(\Omega)^{\mathrm{H}} \boldsymbol{S}_x^{-\frac{1}{2}} \boldsymbol{U}_x^{\mathrm{H}} \boldsymbol{V}_c, \tag{23}$$

and then select the unitary matrix closest to $\tilde{\boldsymbol{\Theta}}^*$ given by

$$\boldsymbol{\Theta}^* = \boldsymbol{J} \boldsymbol{L}^{\mathrm{H}}, \tag{24}$$

where $\boldsymbol{J}$ and $\boldsymbol{L}$ are the left- and right-singular vectors of $\tilde{\boldsymbol{\Theta}}^*$, respectively.

By inserting Eq. (24) into Eq. (18) and Eq. (12), the closed-form solution for the signal-dependent encoder can be written as

$$\boldsymbol{W}(\Omega) = \boldsymbol{W}_{\mathrm{dir}}(\Omega) + \boldsymbol{W}_{\mathrm{orth}}(\Omega), \tag{25}$$

with

$$\boldsymbol{W}_{\mathrm{dir}}(\Omega) = \boldsymbol{y}(\Omega) \frac{\boldsymbol{v}(\Omega)^{\mathrm{H}} \boldsymbol{\Lambda}^{-1}}{\boldsymbol{v}(\Omega)^{\mathrm{H}} \boldsymbol{\Lambda}^{-1} \boldsymbol{v}(\Omega)} \tag{26}$$

$$\boldsymbol{W}_{\mathrm{orth}}(\Omega) = \boldsymbol{N}_y^{\mathrm{H}} \boldsymbol{J} \boldsymbol{L}^{\mathrm{H}} \boldsymbol{N}_q \boldsymbol{S}_x^{-\frac{1}{2}} \boldsymbol{U}_x^{\mathrm{H}}, \tag{27}$$

where $\boldsymbol{W}_{\mathrm{dir}}(\Omega)$ is a super-directive beamformer scaled by $\boldsymbol{y}(\Omega)$, and $\boldsymbol{W}_{\mathrm{orth}}(\Omega)$ ensures that the orthogonality constraint is met and makes the encoder more robust to DOA estimation errors.

Results and Discussion

To illustrate the contributions of $\boldsymbol{W}_{\mathrm{dir}}(\Omega)$ and $\boldsymbol{W}_{\mathrm{orth}}(\Omega)$, Figure 3 shows the individual pick-up patterns. The results were obtained by simulating an open-sphere tetrahedral cardioid-microphone array with a radius of 3 cm and a spatial aliasing frequency of approximately 1.8 kHz. The pick-up patterns of the direct encoder are correct for the estimated DOA (indicated by the black arrows) but deteriorate rapidly for other directions. What is more, the output signals of the direct encoder would be perfectly correlated in a diffuse sound field. Adding the orthogonal encoder, which exhibits a deep null in the source direction, resolves these problems so that the resulting encoder is more robust to DOA estimation errors and yields decorrelated signals in a diffuse sound field.

In Figure 4 the pick-up patterns obtained by the standard encoder in Eq. (2) and the proposed encoder in Eq. (26) are depicted for different frequencies. The standard encoder introduces considerable magnitude and phase errors above 2 kHz because the existence of spatial aliasing is ignored. When the signals thus obtained are reproduced or directionally enhanced, these artefacts cause erroneous spatial cues which strongly vary with both frequency and source directions. The proposed encoder, on the other hand, always yields the desired responses in the direction of the source while maintaining orthogonality of the virtual microphones; towards higher frequencies, however, robustness with respect to DOA estimation errors decreases. By design, the signals picked up by the virtual microphones in a diffuse sound field are uncorrelated and have the same energy.

Conclusion

We proposed a signal-dependent encoding scheme that enables the conversion of spherical microphone array recordings to First-Order Ambisonics above the spatial aliasing frequency. The proposed encoder is time-frequency variant and relies on instantaneous estimates of the source direction. Its evaluation in real-world scenarios is subject of ongoing research.

References

[1] M. Kronlachner and F. Zotter, "Spatial transformations for the enhancement of Ambisonic recordings," in *Proceedings of the 2nd International Conference on Spatial Audio*, 2014.

[2] S. Berge and N. Barrett, "High Angular Resolution Planewave Expansion," *Ambisonics Symposium*, 2010.

[3] V. Pulkki, "Spatial sound reproduction with directional audio coding," *Journal of the Audio Engineering Society*, pp. 503–516, 2007.

[4] C. Nachbar, F. Zotter, E. Deleflie, and A. Sontacchi, "AMBIX - A Suggested Ambisonics Format," in *Ambisonics Symposium*, 2011.

[5] I. Balmages and B. Rafaely, "Open-sphere designs for spherical microphone arrays," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 727–732, 2007.

[6] B. Rafaely, B. Weiss, and E. Bachmat, "Spatial aliasing in spherical microphone arrays," *Signal Processing, IEEE Transactions on*, vol. 55, no. 3, pp. 1003–1010, 2007.

[7] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002.

[8] J. Vilkamo, T. Bäckström, and A. Kuntz, "Optimized covariance domain framework for time-frequency processing of spatial audio," *AES: Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 403–411, 2013.

[9] J. Vilkamo and S. Delikaris-Manias, "Perceptual Reproduction of Spatial Sound Using Loudspeaker-Signal-Domain Parametrization," *Audio, Speech, and Language Proc., IEEE Transactions on*, vol. 23, no. 10, pp. 1660–1669, 2015.

[10] J. H. Manton, "Optimization algorithms exploiting unitary constraints," *IEEE Transactions on Signal Processing*, vol. 50, no. 3, pp. 635–650, 2002.

[11] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre, "Manopt, a matlab toolbox for optimization on manifolds," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1455–1459, 2014.