

Comparison of Methods for Plausible Sound Field Translation

Maximilian Kentgens, Peter Jax

RWTH Aachen University, Institute of Communication Systems (IKS), Germany

Email: {kentgens,jax}@iks.rwth-aachen.de

Abstract

Higher-order Ambisonics recordings obtained from a single spherical microphone array inherently allow for immersive reproduction with all three rotational degrees of freedom (3DoF). On the other hand, physically correct implementation of user movement in the additional translational degrees of freedom is possible within a narrow range only. Therefore, a key technology to advance to 3DoF+ or 6DoF are sound field translation methods, which omit physically correct reconstruction in favor of psychoacoustic plausibility. In this work, we review recent advances in the field and juxtapose adaptive and non-adaptive methods. Relevant properties of different methods are compared using a novel visualization scheme. We complete this work with a discussion of limitations and opportunities of the approaches.

Introduction

Higher-order Ambisonics (HOA) have become of large interest for recorded virtual reality applications, not least because of elegant recording possibilities with spherical microphone arrays (SMAs). This spatial audio format is exceptionally well suited for three-(rotational)-degrees-of-freedom (3DoF) spatial audio recording. However, it is subject to ongoing research to extend the format to translational listener movements. With regard to this research, it is usually distinguished between sound field interpolation methods between multiple microphone arrays distributed over the whole space (6DoF) and sound field extrapolation of a single SMA around the recording position (3DoF+) [1]. Here, we focus on the latter.

The fundamental issue is of physical nature and is due to the fact that a HOA signal only represents the sound field in a limited sweet spot around the SMA. Therefore, methods such as the *plane-wave translation* (PWT) [2] exhibit significant signal distortion for larger displacements and higher frequencies, as they directly attempt to evaluate the sound field at the translated position.

In VR applications, it is often sufficient to obtain a signal representation at the translated position which is plausible to a human listener but not necessarily physically meaningful. Benevolent properties of the human auditory system are leveraged in the development of algorithms which allow for translations beyond the signal's sweet spot [3, 4, 5]. Recently, three novel approaches from the authors' lab were presented, namely the space warping (SW) [6], adaptive beamforming (ABF) [7], and adaptive space warping (ASW) [8] methods which will be further investigated and compared in this paper.

The structure of the paper is as follows. The first section

gives the mathematical fundamentals. The section thereafter summarizes the different methods and provides a comparison on the basis of the mathematical formulas. A detailed comparison on the basis of a novel visualization method [9] follows. We finish with a discussion and conclusion.

Mathematical Fundamentals

We consider an N -th order higher-order Ambisonics signal. The coefficients are stored in an $(N + 1)^2$ -dimensional spherical harmonics (SH) coefficient vector \mathbf{x}_{nm} . Any dependency on time and/or frequency is neglected for brevity as only spatial aspects are of interest for most of the methods considered in the following.

The spatial-domain signal $x(\vec{s}_u)$ can be obtained from the SH-domain coefficient column vector \mathbf{x}_{nm} using [10]

$$x(\vec{s}_u) = \mathbf{y}(\vec{s}_u) \mathbf{x}_{nm}. \quad (1)$$

Here, a subscript $(\cdot)_u$ denotes the normalization of a spatial vector to unit length and $\vec{s}_u \in \mathcal{S}^2$ is a direction vector on the surface of the unit sphere \mathcal{S}^2 . The SH row vector $\mathbf{y}(\vec{s}_u)$ consists of the SH functions up to order N evaluated for direction \vec{s}_u . Note that $x(\vec{s}_u)$ is spatially band-limited due to the finite truncation order N .

The inverse operation to Eq. (1) is given by the inner product of $x(\vec{s}_u)$ and the complex conjugate of the SH functions,

$$\mathbf{x}_{nm} = \int_{\mathcal{S}^2} x(\vec{s}_u) \mathbf{y}^H(\vec{s}_u) d\vec{s}_u, \quad (2)$$

where $(\cdot)^H$ denotes the matrix Hermetian and $d\vec{s}_u$ is the scalar surface element on \mathcal{S}^2 for direction \vec{s}_u .

Eqs. (1) and (2) are referred to as the inverse and forward spherical harmonics transforms (ISHT, SHT), respectively.

The sound field translation methods considered in the following are linear operations described by matrices \mathbf{T} to obtain a plausible signal estimate $\tilde{\mathbf{x}}_{nm}$ at the translated position from the input signal \mathbf{x}_{nm} , i.e.,

$$\tilde{\mathbf{x}}_{nm} = \mathbf{T} \mathbf{x}_{nm}. \quad (3)$$

Methods Recap

This section first recaps the physically meaningful PWT method for reference. Subsequently, the recently proposed methods for plausible sound field translation SW, ABF, and ASW are revisited.

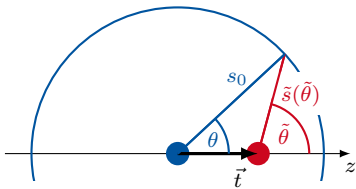


Figure 1: Geometric sound field translation model [8]

Plane-Wave Translation (PWT) [2] is a method for *physically correct* sound field translation, which we consider for reference. This method exploits that any source-free sound field can be decomposed as a superposition of plane waves. Translating along a plane wave yields a phase rotation only. Accordingly, the PWT applies a phase term per direction \vec{s}_u in the spatial domain dependent on the displacement \vec{t} :

$$\mathbf{T}^{(\text{PWT})} = \int_{S^2} \mathbf{y}^H(\vec{s}_u) e^{-ik\vec{s}_u \cdot \vec{t}} \mathbf{y}(\vec{s}_u) d\vec{s}_u. \quad (4)$$

Here, k is the wave number which is proportional to frequency. Due to the finite truncation order, the SHT has limited spatial selectivity, which results in strong signal degradation when $k\|\vec{t}\| \ll N$ is not met. This is usually the case for practically relevant displacements $\|\vec{t}\|$, especially at higher frequencies [6] and necessitates more specialized methods which allow for larger displacements at the expense of physical correctness.

Geometric Model for Plausible Sound Field Translation Unlike the physically correct PWT, the *psychoacoustically plausible* sound field translation methods considered in the following employ a geometric model for the acoustic scene. All sound sources are assumed to be in constant distance s_0 from the origin. A displacement along \vec{t} then results in a change of source directions and distances relative to the position of interest as shown in Fig. 1.

Plain Space Warping (SW) [6] exploits the spatial modification technique from [11] for sound field translation. It is based on the idea of stretching and squeezing the angles of the spatial representation of the signal according to the geometric model from Fig. 1. The SW definition

$$\mathbf{T}^{(\text{SW})} = \int_{S^2} \mathbf{y}^H(\vec{s}_u) g(\vec{s}_u) \mathbf{y}(\vec{s}_u) d\vec{s}_u \quad (5)$$

reveals some structural similarities to the PWT despite being a fundamentally different approach. An ISHT is applied analogous to the PWT in order to transform the signal into the spatial domain. However, the transformation back to the SH-domain is performed on warped spherical coordinates such that the directional components are panned to the target directions according to the geometric model with $\vec{s} = \vec{s} - \vec{t}$ and $\vec{s} = s_0\vec{s}_u$. The second fundamental difference is that no complex phase-rotation term but a real-valued gain $g(\vec{s}_u)$ is applied per direction. It is chosen to adjust the sound level when approaching or moving away from a source dependent on the change in source distance according to a simple point-source gain

model. A real-valued gain is of advantage as no detrimental destructive interference occurs due to the lack of phase shifts [6].

Adaptive Beamforming (ABF) [7] A disadvantage of the SW approach is that reverberation and diffuse noise are spuriously modified. The ABF approach, in contrast, aims to leave such ambient sound unchanged. Only the primary part, which consists mainly of direct sound, is changed, in order to yield psychoacoustically reasonable sound field translation [12, 3, 5]. The operation is defined as

$$\mathbf{T}^{(\text{ABF})} = \underbrace{\left(\sum_{j=1}^J \mathbf{y}^H(\vec{s}_{j,u}) g_j \mathbf{w}_j^H \right)}_{\text{primary term}} + \underbrace{\left(\mathbf{I} - \sum_{j=1}^J \mathbf{y}^H(\vec{s}_{j,u}) \mathbf{w}_j^H \right)}_{\text{residual term}} \quad (6)$$

and works as follows:

1. Based on an eigenvalue decomposition of the SH coefficients' covariance matrix of the input signal \mathbf{x}_{nm} , a direction-of-arrival analysis is performed, which also incorporates an estimation of the number of distinct primary plane waves J .
2. J null-steering beamformers \mathbf{w}_j^H are steered into the J source directions to extract the primary signal components. Similar to the SW approach, a distance-dependent gain g_j is applied to each primary component. These components are then panned to the target directions $\vec{s}_{j,u}$.
3. The residual term adds every other signal component, which is not processed by the beamformers, without any spatial modification. Here, \mathbf{I} denotes the identity matrix.

The primary part term in Eq. (6) resembles the SW definition Eq. (5). However, instead of integrating over all directions on the sphere, we only sum over the J primary directions.

Adaptive Space Warping (ASW) [8] pursues the same goal as the ABF: only primary sound shall be spatially modified. A spatial mask $\mathbf{\Gamma}$ is used to segregate primary and ambient sound. Space warping is then applied on the primary part only. The inverse mask $\mathbf{I} - \mathbf{\Gamma}$ then adds the ambient part without modification:

$$\mathbf{T}^{(\text{ASW})} = \underbrace{\mathbf{T}^{(\text{SW})} \mathbf{\Gamma}}_{\text{primary term}} + \underbrace{(\mathbf{I} - \mathbf{\Gamma})}_{\text{residual term}}. \quad (7)$$

The mask is the result of an optimal filter derivation and is of the form of a multichannel Wiener filter. It can be constructed as $\mathbf{\Gamma} = \mathbf{\Psi}_d \mathbf{\Psi}_x^{-1}$, where $\mathbf{\Psi}_d$ and $\mathbf{\Psi}_x$ are the covariance matrices of the primary and total signal, respectively. Furthermore, $(\cdot)^{-1}$ denotes a matrix inverse.

While the ABF method is parametric and requires estimation of the number of sources and their direction, the ASW approach requires estimation of covariance matrices. In practice, both adaptive methods are usually applied independently per time-frequency bin in a time-frequency domain [7]. Covariances of the total signal are measured

i	a_i	θ_i	ϕ_i
1	0.8	$\frac{1}{5}\pi$	0
2	1	$\frac{1}{2}\pi$	$\frac{1}{3}\pi$
3	0.4	$\frac{4}{5}\pi$	$-\frac{9}{10}\pi$

Table 1: Source properties

by recursive smoothing over time and/or smoothing over adjacent frequency bands.

Filter Analysis

Filter Visualization We compare the previously presented methods for a translation in the vertical z -direction. The SH truncation order is chosen as $N = 4$. The displacement is chosen as $\|\vec{t}\| = 0.8 \cdot s_0$. Moreover, we consider a wave number of $k = 10/s_0$, which is only relevant for the frequency-dependent PWT method. For the evaluation of the signal-adaptive methods, we assume three plane-wave sources with indices $i = 1, 2, 3$. The respective inclination and azimuth angles of incidence θ_i and ϕ_i as well as amplitudes a_i can be found in Table 1. Diffuse noise is added at an SNR of -5 dB. These source properties are provided to the ABF and ASW methods as oracle information in the form of direction-of-arrivals and covariance matrices, respectively. This means that no parameter estimation is considered here.

The resulting operations $\mathbf{T}^{(\text{PWT})}$, $\mathbf{T}^{(\text{SW})}$, $\mathbf{T}^{(\text{ABF})}$, and $\mathbf{T}^{(\text{ASW})}$ are illustrated in Fig. 2 by means of the visualization method proposed in [9].

Fig. 2a shows a non-satisfactory behavior of the PWT for the given displacement which is significantly beyond the sweet spot as $k\|\vec{t}\| = 8 \gg N = 4$. Sources (blue crosses) are not projected to the desired target directions (red crosses) and exhibit an undesired attenuation (bluish background). Note that this behavior is frequency dependent and would result in strong artifacts in the translated output signal.

Fig. 2b visualizes the non-adaptive SW filter. Sound impinging from the top is amplified while sound incidence from the bottom is attenuated. This is as expected as we are investigating a translation along the z -axis. Impinging sound is systematically projected into directions with increased inclination angle. This does not only affect sources but also ambient signal components. Comparing the actual spatial modification (length of the arrows) with the geometric model (distance of the red and blue crosses) reveals a slight angular bias of the SW approach as will be discussed further down.

The ABF and ASW approaches are depicted in Figs. 2c and 2d, respectively. As intended, both approaches show a bypass characteristic with no spatial modification and unit gain in those directions where no source is present. A spatial adaptation only occurs for directions in which sources are present. The exact behavior of the two approaches, however, is slightly different as discussed in the following.

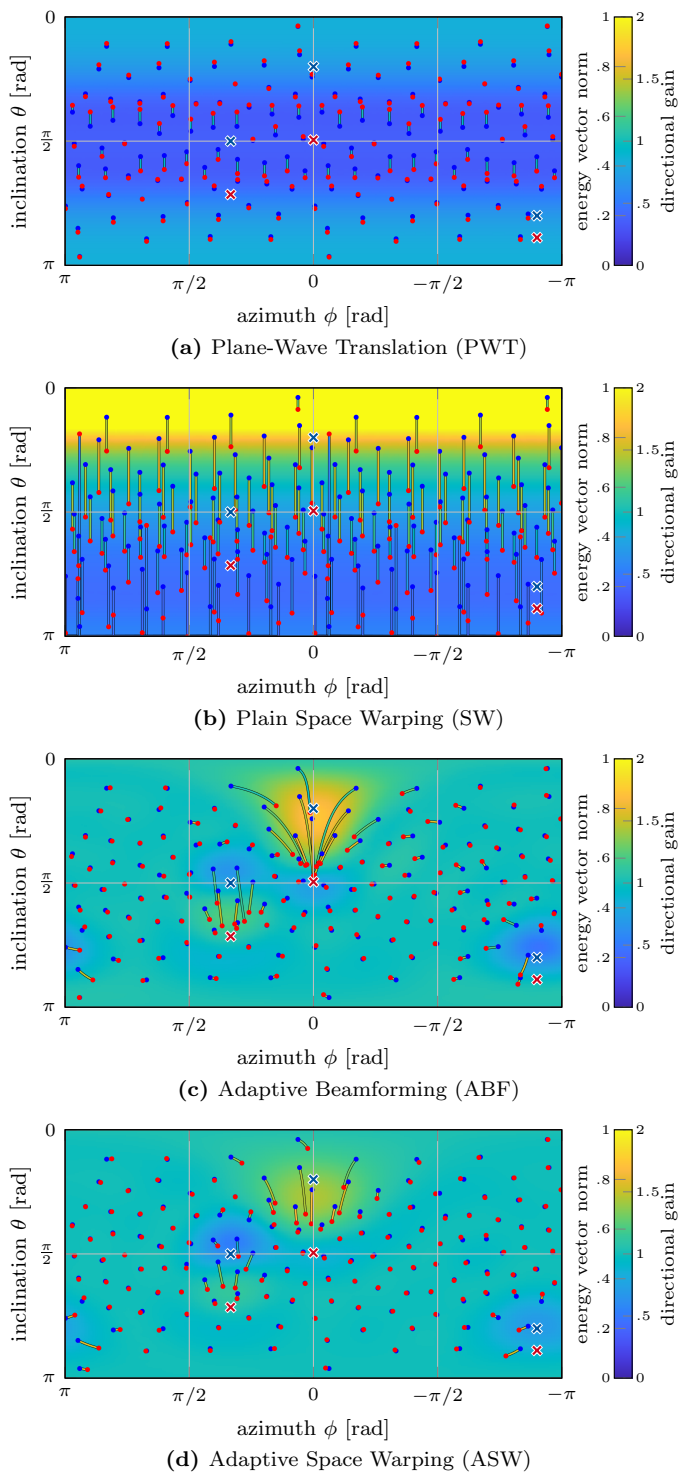


Figure 2: Visualization of different translation methods. The red dots \bullet indicate how a directional excitation from the direction of the blue dots \bullet is spatially modified. The colored background shows the amplification of a directional excitation [9]. The blue crosses \boxtimes indicate the actual source positions before translation, the red crosses \boxtimes indicate desired sources positions after translation.

Close Look on Source #1 As a further investigation, we examine the isolated impact of the different translation operations on source #1. The responses $\tilde{x}_1(\vec{s}_u) = \mathbf{y}(\vec{s}_u) \cdot \mathbf{T} \cdot a_1 \mathbf{y}^H(\theta_1, \phi_1)$ are illustrated in Fig. 3 for $\phi = 0$.

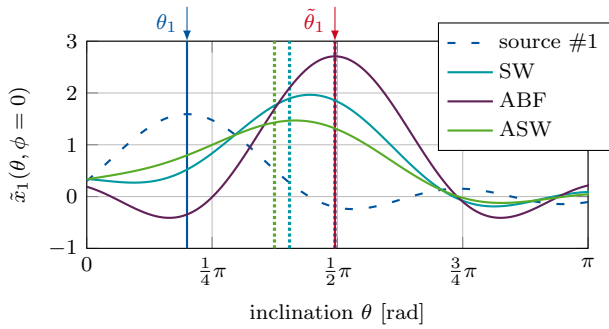


Figure 3: Impact of translation operations on source #1: responses $\tilde{x}_1(\vec{s}_u) = \mathbf{y}(\vec{s}_u) \cdot \mathbf{T} \cdot \mathbf{a}_1 \mathbf{y}^H(\theta_1, \phi_1)$

The output using $\mathbf{T}^{(\text{ABF})}$ results in an order-truncated plane wave exactly in the direction of the desired target direction. This is due to the fact that the simulation setup perfectly matches the parametric model of the ABF approach. In contrast, the response using $\mathbf{T}^{(\text{SW})}$ is slightly skewed and blurred, resulting in an angular bias of the response energy vector (dotted line). This is inherent for the space warping operation due to the limited SH truncation order. The effect is even more pronounced for $\mathbf{T}^{(\text{ASW})}$. The Wiener weighting used within ASW is optimized to keep a certain amount of energy in the original directions due to the presence of strong additive diffuse noise which also explains the lower amplitude of the response. In contrast, the ABF approach systematically overestimates the primary energy due to the absence of such a weighting.

Discussion

The comparison presented reveals many details of the behavior of the various translation operations. However, another important aspect was not considered here: the robustness of estimates of the covariances and directions-of-arrival in the ASW and ABF approaches, respectively. The statistical analysis of the signal is of particular importance in complex acoustic scenes. In the end, only a listening test can give rise to evidence for psychoacoustic performance assessment. First investigations in this direction are presented in [7]. Our findings in the present paper are the basis for further investigations in the future.

Conclusion

In this paper, we reviewed and compared recent advances in plausible sound field translation. An evaluation using a novel visualization scheme provided evidence that the SW, ASW and ABF methods are superior to the physically correct methods for translations beyond the sweet spot of the input HOA signal. We studied various system details and pointed out important similarities and differences. In summary, all three recently proposed methods show great potential despite featuring different behavior in various aspects. This needs to be further investigated from a psychoacoustic point of view in future research.

Acknowledgments

The authors thank Georg Krekel for fruitful discussions.

References

- [1] J. G. Tylka, “Virtual navigation of ambisonics-encoded sound fields containing near-field sources,” Ph.D. dissertation, Princeton University, Princeton, NJ, USA, 2019.
- [2] F. Schultz and S. Spors, “Data-based binaural synthesis including rotational and translatory head-movements,” in *AES 52nd International Conference on Sound Field Control*, Sept. 2013.
- [3] T. Pihlajamaki and V. Pulkki, “Synthesis of complex sound scenes with transformation of recorded spatial sound in virtual reality,” *Journal of the Audio Engineering Society*, vol. 63, no. 7/8, pp. 542–551, Aug. 2015.
- [4] A. Allen and W. B. Kleijn, “Ambisonics soundfield navigation using directional decomposition and path distance estimation,” in *4th International Conference on Spatial Audio (ICSA)*, Graz, Austria, Sept. 2017.
- [5] 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 International Conference on Audio for Virtual and Augmented Reality (AVAR)*, Aug. 2018.
- [6] M. Kentgens and P. Jax, “Translation of a higher-order ambisonics sound scene by space warping,” in *AES International Conference on Audio for Virtual and Augmented Reality (AVAR)*, Aug. 2020.
- [7] M. Kentgens, A. Behler, and P. Jax, “Translation of a higher order ambisonics sound scene based on parametric decomposition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020.
- [8] M. Kentgens and P. Jax, “Ambient-aware sound field translation using optimal spatial filtering,” Accepted at IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), Oct. 2021.
- [9] M. Kentgens and P. Jax, “Visualization of linear operations in the spherical harmonics domain,” in *International Conference on Immersive and 3D Audio (I3DA)*, Sept. 2021, pre-print available arXiv:2104.13069.
- [10] B. Rafaely, *Fundamentals of Spherical Array Processing*. Springer, 2015.
- [11] H. Pomberger and F. Zotter, “Warping of 3D Ambisonic recordings,” in *3rd International Symposium on Ambisonics and Spherical Acoustics*, June 2011.
- [12] A. Neidhardt, A. Ignatious-Tommy, and A. D. Peralpandan, “Plausibility of an interactive approaching motion towards a virtual sound source based on simplified BRIR sets,” in *AES 144th Convention*, May 2018.