FREQUENCY-INDEPENDENT AMBISONICS UPSCALING USING DEEP LEARNING

Egke Chatzimoustafa and Peter Jax

Institute of Communication Systems (IKS), RWTH Aachen University, Aachen, Germany
{chatzimoustafa, jax}@iks.rwth-aachen.de

ABSTRACT

Due to its flexibility, the higher-order Ambisonics (HOA) sound format is widely used in spatial audio applications like virtual reality (VR). The spatial accuracy is sensitive to the HOA order, which is limited by the recording and reproduction hardware. To enhance accuracy, the directional audio coding (DirAC) method prioritizes the improvement of perceptual quality over physically accurate spatial representations. We propose a novel neural network architecture that performs HOA upscaling within frequency subbands independently by leveraging the properties of spherical harmonics basis functions that are inherent in the HOA transformations. Although the proposed network is trained on synthetic data, a 63% variance reduction of the spatial similarity is achieved compared to DirAC, while both methods exhibit similar average performance for real-world measurement data. This makes our model beneficial for applications which require reliable spatial representations.

***Index Terms*—** Deep neural networks, frequency domain, higher-order Ambisonics, spatial audio, upscaling

1. INTRODUCTION

Acoustic scenes captured with sound source direction information can create immersive experiences when played back over loudspeakers [1]. A flexible sound format to achieve that is Ambisonics, whose foundations were introduced in the pioneering work of Gerzon [2]. This technology has gained significant interest and found applications in the consumer market [3, 4]. Advancements in recording hardware and loudspeaker systems have made higher-order Ambisonics (HOA) possible, which allows for increased spatial resolution and a larger sweet spot size, and thus a more accurate representation of sound fields [5–8].

Despite enhanced sound field representations through HOA, spatial resolution remains constrained by the number of microphones in recording and loudspeakers in reproduction [9–11]. To overcome this limitation, algorithmic approaches have been proposed which artificially increase HOA order. In [12–14], a small number of sound incidence directions is assumed and a sparse plane-wave decomposition is performed. The directional audio coding (DirAC) introduced by Pulkki [15] is a parametric time-frequency method, which has been standardized in [16]. This algorithm estimates direction of arrival (DOA) and diffuseness parameters from Ambisonics signals in time-frequency blocks. Following this estimation step, vector base amplitude panning (VBAP) [17] gains are applied to steer sound towards desired directions, while decorrelation blocks maintain incoherence in diffuse parts of the sound field. This algorithm prioritizes assumptions about spatial properties and their relation to human auditory perception over physically accurate spatial reproduction. As reported in [15], model parameters must be adjusted based on spatial recording conditions, and the decorrelation filters are designed using complicated heuristics during implementation.

Additionally, deep learning techniques have emerged as viable alternatives to increase the HOA order, which simplify the processing by eliminating the need for DOA and diffuseness estimation. For example, recent studies propose deep learning models that process recorded spherical microphone array signals to enhance sound field reproduction [18, 19]. Other deep learning methods that operate within the HOA domain have also shown potential benefits [20].

In our previous work [21], we proposed to use gated recurrent units (GRUs) [22] in the time domain for the HOA upscaling problem, and presented a data generation method to tackle the challenge of limited training data. This paper builds upon that foundation and introduces a novel deep learning model. As a novelty, our approach leverages the properties of spherical harmonics basis functions inherent in the HOA transformations, which allows us to perform upscaling independently within frequency subbands.

We train GRUs on purely artificial sound scenes characterized by harmonics within frequency subbands. This paper further extends the training dataset from [21] by incorporating time-varying frequencies to reflect non-stationary audio signals in practice [23]. Through simulations, we compare our model against the fullband GRUs from our previous work [21] and DirAC. The comparison is conducted on test datasets that comprise real-world recordings of musical instruments and speech. The evaluation results indicate that our approach consistently achieves higher reliability in reproducing sound source directions in non-diffuse sound fields. We present examples from the test dataset to further substantiate the significance of the performance characteristics of our model.

2. THEORETICAL BACKGROUND

Considering spatially discrete sound sources, directional information about a sound field which is defined on the surface of a sphere with unit radius $\mathcal{S}^2 = \{\boldsymbol{w} \in \mathbb{R}^3 : \|\boldsymbol{w}\| = 1\}$ can be expressed using the amplitude density vector [12]

$$\boldsymbol{x}(k) = (x(k, \boldsymbol{\Theta}_1), x(k, \boldsymbol{\Theta}_2), \ldots, x(k, \boldsymbol{\Theta}_U))^{\mathrm{T}}, \tag{1}$$

which is characterized by the direction-specific amplitudes among a total of U sources. Here, $(\cdot)^{\mathrm{T}}$ denotes the transpose of a vector, $k = \frac{2\pi f}{c}$ is the wavenumber, f denotes the frequency and $c = 343\ \frac{\mathrm{m}}{\mathrm{s}}$ is the speed of sound. The direction vector $\boldsymbol{\Theta}_u = (\theta_u, \phi_u)^{\mathrm{T}}$ with source index $u = 1, \ldots, U$ consists of inclination $\theta_u \in [0, \pi]$ and azimuth $\phi_u \in [-\pi, \pi)$. As (1) is defined on the unit sphere, it is convenient to express sound fields using the spherical harmonics transform (SHT) [24]

$$\boldsymbol{x}_{\boldsymbol{nm}}(k) = \boldsymbol{Y}\boldsymbol{x}(k) \tag{2}$$

with the $(N+1)^2 \times U$ spherical harmonics matrix

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}(\boldsymbol{\Theta}_1) & \boldsymbol{y}(\boldsymbol{\Theta}_2) & \cdots & \boldsymbol{y}(\boldsymbol{\Theta}_U) \end{pmatrix} \tag{3}$$

and $(N+1)^2$-dimensional spherical harmonics vectors evaluated at the source directions $\boldsymbol{\Theta}_u$

$$\boldsymbol{y}(\boldsymbol{\Theta}_u) = \left(Y_0^0(\boldsymbol{\Theta}_u), Y_1^{-1}(\boldsymbol{\Theta}_u), \ldots, Y_N^N(\boldsymbol{\Theta}_u) \right)^{\mathrm{T}}. \tag{4}$$

Here, N denotes the maximum spherical harmonics order. The vector (4) stores the spherical harmonics basis functions

$$Y_n^m(\boldsymbol{\Theta}_u) = \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}}\,(-1)^m \cdot P_n^m(\cos(\theta_u)) \begin{cases} \sqrt{2}\cos(m\phi_u) & \text{if } m > 0 \\ 1 & \text{if } m = 0 \\ -\sqrt{2}\sin(m\phi_u) & \text{if } m < 0, \end{cases} \tag{5}$$

where $P_n^m(\cdot)$ are the associated Legendre functions and $(\cdot)!$ denotes the factorial function. Although the definition of Y_n^m varies in the literature, the real-valued N3D-normalized definition (5) is used in many applications, e.g., those found in [1, 25]. Furthermore, the spherical harmonics order index $n = 0, 1, 2, \ldots N$ and degree index $m = -n, \ldots, n$ adhere to the Ambisonics channel number (ACN) convention [26]. Thus, the spherical harmonics representations in (2) are expressed by

$$\boldsymbol{x_{nm}}(k) = (x_{0,0}(k), x_{1,-1}(k), \ldots, x_{N,N}(k))^{\mathrm{T}}. \tag{6}$$

This formulation of the sound field is more abstract compared to (1) because the spherical harmonics weights are independent of sound source directions. The time-domain signal $\boldsymbol{x_{nm}}(t)$ with discrete-time index t is referred to as the HOA signal. A perfect representation of a sound field would require $N \to \infty$, which is not feasible. In practice, this limitation can cause sound sources to appear displaced, potentially degrading the virtual reality (VR) experience, as will be further analyzed in Sec. 4.

2.1. Spatial Analysis of Higher-Order Ambisonics Signals

Statistical properties of stationary ergodic processes can be analyzed based on only a single observed process [27]. A powerful metric for assessing the spatial properties of HOA signals is the $(N+1)^2 \times (N+1)^2$ spatial covariance matrix defined as [28]

$$\boldsymbol{\Gamma}_{\boldsymbol{x_{nm}}} = \mathrm{E}_t\{\boldsymbol{x_{nm}}(t)\boldsymbol{x}_{\boldsymbol{nm}}^{\mathrm{T}}(t)\}, \tag{7}$$

where $\mathrm{E}_t\{\cdot\}$ denotes the expectation operator that is applied over time. In practice, audio signals typically exhibit time-variant statistics and are considered non-stationary. However, stationarity can be assumed in shorter time frames, which is valid, e.g., for speech signals [23]. Consequently, applying (7) to the entire signal yields an average of momentary statistics over shorter, stationary time frames. Using (7), spatial distribution of energy in a sound scene can be calculated at directions $\boldsymbol{\Theta} = (\theta, \phi)^{\mathrm{T}}$ via the steered response power (SRP)

$$P(\boldsymbol{\Theta}) = \frac{4\pi}{(N+1)^2}\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{\Theta})\boldsymbol{\Gamma}_{\boldsymbol{x_{nm}}}\boldsymbol{y}(\boldsymbol{\Theta}), \tag{8}$$

where the term $\frac{4\pi}{(N+1)^2}$ acts as a normalization factor to ensure unit power density when $\boldsymbol{\Gamma}_{\boldsymbol{x_{nm}}} = \boldsymbol{I}$. Based on (7) and (8), a scale-invariant measure that defines the similarity between two sound field power distributions can be calculated as [29]

$$\eta = 1 - \frac{2}{\pi}\cos^{-1}\left(\frac{\iint_{\mathcal{S}} \hat{P}(\boldsymbol{\Theta})\breve{P}(\boldsymbol{\Theta})\,d\mathcal{S}}{\sqrt{\iint_{\mathcal{S}}[\hat{P}(\boldsymbol{\Theta})]^2\,d\mathcal{S}}\sqrt{\iint_{\mathcal{S}}[\breve{P}(\boldsymbol{\Theta})]^2\,d\mathcal{S}}} \right), \tag{9}$$

which is referred to as the spatial similarity in [28]. In practice, the surface integrals in (9) are approximated by evaluating (8) at multiple directions within a defined grid. A widely used sampling scheme is the quasi-uniform Fliege/Maier grid [30], known for its high accuracy in integrating functions on spheres. In Sec. 4, (9) will be used to evaluate the physical accuracy of all models by comparing the target signal SRP $\breve{P}(\boldsymbol{\Theta})$ with the estimated signal SRP $\hat{P}(\boldsymbol{\Theta})$.

3. AMBISONICS UPSCALING

In this work, upscaling HOA signals involves estimating the coefficients needed to obtain the HOA signal at a desired order $\hat{N} > 1$ given a first-order Ambisonics signal, i.e.,

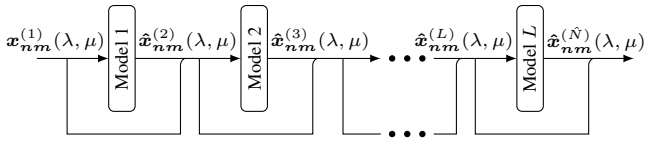
$$\boldsymbol{x}_{\boldsymbol{nm}}^{(1)}(t) \overset{\text{Upscaling}}{\longrightarrow} \hat{\boldsymbol{x}}_{\boldsymbol{nm}}^{(\hat{N})}(t), \tag{10}$$

where $\boldsymbol{x}_{\boldsymbol{nm}}^{(1)}(t) \in \mathbb{R}^4$ represents the first-order Ambisonics signal, and $\hat{\boldsymbol{x}}_{\boldsymbol{nm}}^{(\hat{N})}(t) \in \mathbb{R}^{(\hat{N}+1)^2}$ is the estimated HOA signal. It is evident in (10) that $(\hat{N}+1)^2 - 4$ coefficients need to be estimated, which rises quadratically with $\hat{N}$ and may become infeasible at higher orders. As a remedy, [20] proposed a sequential HOA upscaling framework in which $L = \hat{N} - 1$ individual models perform upscaling from order 1 to $\hat{N}$, where each model l is responsible for upscaling from order l to $l+1$. This approach was later adopted by us in [21], where fully connected structures were replaced by GRUs. In the following, we will take up the idea of this sequential procedure and propose a novel model within this framework based on certain assumptions derived from the signal model presented in Sec. 2.

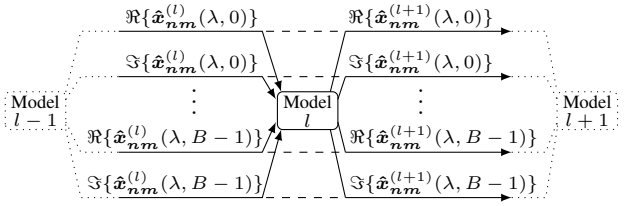
3.1. Proposed Framework

Suppose we have an N-th order spherical harmonics representation denoted as $\boldsymbol{x}_{\boldsymbol{nm}}^{(N)}(k)$. Then, its truncation is based on the $\boldsymbol{Y}$ matrix (3) which is independent of k. This frequency-independency applies also to the "inverse problem" of HOA upscaling, which motivates us to study network architectures with model weights that are shared across different frequency subbands. Moreover, it is reasonable to assume that in an acoustic scene with multiple sound sources, individual frequency bands will be occupied by only a small number of sound sources at any time instance [12, 13]. This allows us for independent upscaling within each subband. By leveraging these properties of the spherical harmonics basis functions and the assumptions about the sound field in our model design, we achieve a low-complexity architecture which is computationally efficient.

Fig. 1a illustrates the sequential framework for HOA upscaling in the time-frequency domain. First, the time-domain Ambisonics signal $\boldsymbol{x}_{\boldsymbol{nm}}^{(1)}(t) \in \mathbb{R}^4$ is divided into Λ overlapping frames and transformed to the time-frequency domain with B frequency bins using the short-term Fourier transform (STFT), which results in the time-frequency representation $\boldsymbol{x}_{\boldsymbol{nm}}^{(1)}(\lambda, \mu) \in \mathbb{C}^4$, where $\lambda = 0, \ldots, \Lambda - 1$ denotes the frame index and $\mu = 0, \ldots, B-1$ the frequency bin index. The first model $l = 1$ is then trained to estimate the next-order HOA frequency bins $\tilde{\boldsymbol{x}}_{\boldsymbol{nm}}^{(l+1)}(\lambda, \mu) \in \mathbb{C}^{(2l+3)}$, which are concatenated with the model's input to construct the next-order HOA frequency bins $\hat{\boldsymbol{x}}_{\boldsymbol{nm}}^{(l+1)}(\lambda, \mu) \in \mathbb{C}^{(l+2)^2}$. Each subsequent model $l > 1$ uses the output from the preceding model $l-1$ and the process continues iteratively. In total, $L = \hat{N} - 1$ independent networks need to be trained for this task. During inference, the input frequency bins pass sequentially through all individual models to obtain the upscaled frequency bins at the desired HOA order, where estimated time-domain signals $\hat{\boldsymbol{x}}_{\boldsymbol{nm}}^{(\hat{N})}(t)$ are reconstructed using the inverse discrete Fourier transform (IDFT), followed by an overlap-add method [23].



(a) Block diagram of sequential architecture of HOA upscaling.



(b) Block diagram showing the models $l > 1$ of the proposed architecture from Fig. 1a. The dashed lines indicate the concatenation of model estimations with the input for the l-th model.

Fig. 1: Architecture of the considered HOA upscaling system.

Fig. 1b provides an in-depth view of model $l > 1$, which is responsible for upscaling its input by one order. Consistent with the previously described simplifications, a single model handles the upscaling of both the real and imaginary parts, $\Re\{\hat{\boldsymbol{x}}_{nm}^{(l)}(\lambda,\mu)\} \in \mathbb{R}^{(l+1)^2}$ and $\Im\{\hat{\boldsymbol{x}}_{nm}^{(l)}(\lambda,\mu)\} \in \mathbb{R}^{(l+1)^2}$. Each model l consists of a bidirectional GRU, which maps $(l+1)^2$ coefficients for each subband to $2 \times N_h$ hidden features per frame λ. The fully connected output layer processes the GRU output using a linear activation. The resulting signal is then concatenated with the input to obtain the next-order signals $\Re\{\hat{\boldsymbol{x}}_{nm}^{(l+1)}(\lambda,\mu)\} \in \mathbb{R}^{(l+2)^2}$ and $\Im\{\hat{\boldsymbol{x}}_{nm}^{(l+1)}(\lambda,\mu)\} \in \mathbb{R}^{(l+2)^2}$.

For this task, alternative architectures such as convolutional neural networks (CNNs) or transformers could also be considered. While CNNs excel at modeling local patterns, they may struggle to capture long-term temporal dependencies, and transformers typically require more data to be effective. We adopt bidirectional GRUs as a computationally efficient model that captures both past and future temporal context and is well suited to limited data and resources.

3.2. Data Generation

In the following, we focus on directional sound scenes, which results in a simplified signal model without diffuse components. In our previous work [21], we showed that sinusoidal sources with harmonics can serve as a suitable method for data generation

$$x(t, \boldsymbol{\Theta}_u) = a_u \sum_{\xi=0}^{\Xi} e^{-\xi\beta_u} \sin\left((\xi+1)\, 2\pi \frac{f}{f_s} t + \Delta\alpha_u\right), \quad (11)$$

where we added $\Xi = 4$ harmonics into the signal. Here, a_u denotes the amplitude, β_u the decay factor, f the frequency, f_s the sampling rate, and $\Delta\alpha_u$ the phase shift. In this study, we adopt a similar approach and extend this data generation technique by incorporating time-varying $f(t)$ between subsequent frames to reflect non-stationary behavior of typical audio signals [23]. Hence, the fixed-frequency f in (11) is replaced by a time-varying frequency

$$f(t) = f_0 \cdot \left(1 + \sin\left(2\pi \frac{f_v}{f_s} t\right)\right), \quad (12)$$

where f_0 denotes the fundamental frequency and f_v the frequency of an oscillating term drawn from a random uniform distribution from the interval $[10, 80]$ Hz. The oscillating term has a maximum amplitude of $2f_0$. The following parameters were sampled from random uniform distributions: The amplitude $a_u \in [0.4, 1]$, the decay factor $\beta_u \in [0, 2]$, the phase shift $\Delta\alpha_u \in [0, 2\pi)$, and the fundamental frequency $f_0 \in [20, 1500]$ Hz, which reflects the range of fundamental frequencies of typical musical instruments and speech signals. Using (11) and (2), individual sinusoidal sources were placed in the far field at directions $\boldsymbol{\Theta}_u$ sampled from a random uniform distribution on the whole sphere, i.e., $\theta \in [0, \pi]$ and $\phi \in [-\pi, \pi)$.

In total, 15×10^3 acoustic scenes, each containing between 1 and 5 sinusoidal sources, were generated for training by inserting the randomly chosen parameters into (11) and (12) with $\Xi = 4$ and time-varying frequencies $f(t)$. At a sampling rate of $f_s = 44.1$ kHz, this corresponds to approximately 0.68 hours of training data. To demonstrate the efficacy of the proposed model architecture, we generated an equivalent amount of data using the randomly chosen parameters from the same intervals described above, along with $\Xi = 4$ and time-varying frequencies $f(t)$ to train the fullband GRUs from our previous work [21], which operate in the time domain. A validation dataset with 2×10^3 (approximately 0.09 hours) samples was generated by randomly selecting musical instruments and speech signals as sound stimuli from the EBU-SQAM dataset [31] and using (2).

4. EVALUATION

To assess model performance under different conditions, we considered four scenarios: The first scenario with 2 sound sources per scene, the second with 3 sound sources, the third with 4 sound sources, and the fourth with 5 sound sources. At each scenario, we created a separate test set with 5×10^3 samples (approximately 0.23 hours long) for each upscaling order $l + 1$ by randomly selecting speech stimuli from the HiFi-TTS dataset [32], and musical instrument signals from the musical instrument's sound dataset [33]. The sound sources were placed randomly on Fliege/Maier [30] sampling grids with the numbers of sampling points $Q = 9, 16, 25, 36$ and 49, which correspond to the upscaling orders $l + 1 = 2, 3, 4, 5$ and 6, respectively.

To ensure a fair comparison, we adhered to the signal processing parameters that were configured in the open-source code example of the DirAC method provided by the authors in [15]. We generated HOA signals with a length of 4096 time samples which results in signals that are approximately 93 ms long at $f_s = 44.1$ kHz. In the preprocessing step, these signals were divided into seven overlapping frames, each containing 1024 samples, using Hann windows with 50% overlap. The number of frequency bins for the discrete Fourier transform (DFT) was set to 2048. In accordance with the specified parameters from [21], we generated HOA signals with a length of 512 time samples for training the fullband GRUs in the time domain.

After a hyperparameter search, each model l in Fig. 1 and each fullband GRU from [21] was configured with a hidden size of $N_h = 128$, and trained for 200 epochs to ensure convergence. Adam optimizer [34] was employed with an initial learning rate 10^{-4}, which was reduced by a scheduler if no progress was observed over a patience period of 10 epochs. Mean squared error (MSE) loss was employed to train the models.

Since DirAC yields signals at the Fliege/Maier grid sampling points instead of directly in the HOA domain, the resulting signals were converted into the HOA domain using (2). In this process, the spherical harmonics matrix $\boldsymbol{Y}$ was evaluated at the directions of the sampling points. Note that the aim of creating the test set completely on Fliege/Maier grid sampling points is to prevent additional errors that could otherwise arise for DirAC if the grid sampling points do not exactly match the ground truth source directions.

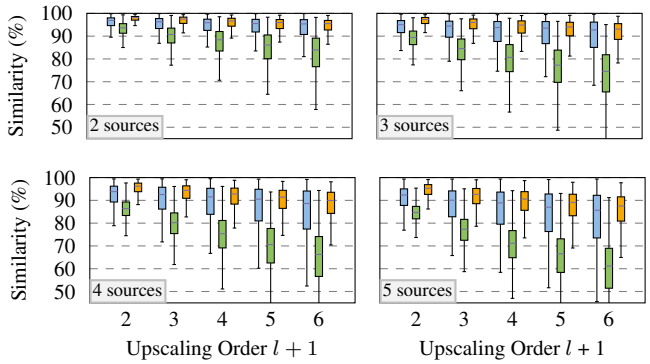


Fig. 2: Box plots representing spatial similarity distributions on test sets for DirAC (), the fullband model from [21] (), and the proposed model (). Both deep learning models were trained on sinusoidal data with harmonics and time-varying frequencies (as defined in (12)).

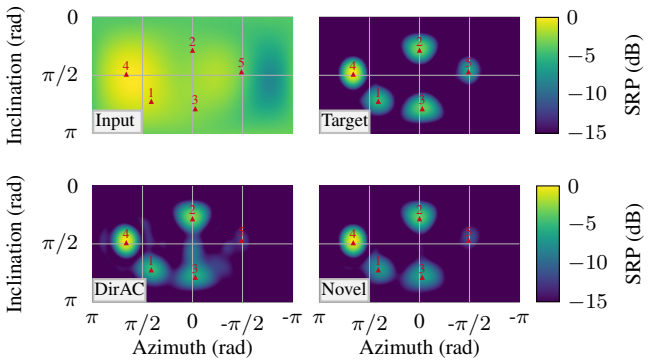


Fig. 3: SRP maps from first-order Ambisonics input, target HOA signal at order 6, and estimated HOA signals from DirAC and proposed models at order 6 for a sound field from the test set. Ground truth sound source directions are marked by triangles (▲).

4.1. Spatial Similarity

We calculated the spatial covariance matrices for the estimated signal $\mathbf{\Gamma}_{\hat{\mathbf{x}}_{nm}}$ and reference signal $\mathbf{\Gamma}_{\check{\mathbf{x}}_{nm}}$ using (7) for each test sample at each upscaling order $l+1$. Subsequently, we computed SRP maps using (8) on a dense Fliege/Maier grid [30] with $Q = 900$ sampling points to ensure high accuracy in approximating the integral in (9).

Fig. 2 displays the spatial similarity from DirAC, fullband and the proposed models across all test datasets, represented as box plots. Higher similarity implies more physically accurate sound field reproduction, and the proposed model trained on sinusoidal data with harmonics and time-varying frequencies outperforms the others at every upscaling order and in all cases. In contrast, the fullband model exhibits the lowest median spatial similarity across all scenarios. For example, in the case of two sources, the average spatial similarity of the proposed model is 1.5% higher than DirAC and 4.5% higher than the fullband model at order 2. As the upscaling order rises, median similarity decreases and the variance increases for all models, which hints at less reliable estimations. However, the advantage of the proposed model becomes evident as it consistently exhibits remarkably the lowest variance in all cases.

Additionally, performance declines across all models as the number of sound sources in each scene increases. Under these conditions, the proposed model shows higher median values and lower variance, whereas the fullband model maintains lower variance but exhibits notably lower median values than DirAC across different upscaling orders. For example, at the upscaling order 6 with five sources, the proposed model achieves a median spatial similarity of 87.5% with variance not exceeding 0.011. In contrast, DirAC has a median of 85.5% with variance slightly above 0.029, while the fullband model has a median of 61% and a variance greater than 0.016. This improvement corresponds to an approximate 63% reduction in spatial similarity variance while maintaining a similar median similarity across the entire dataset by the proposed model, which indicates more physically accurate sound field representation even in more complicated scenarios involving multiple sources.

4.2. Steered Response Power Maps

To gain insights into the relationship between spatial similarity values and reproduced sound source directions, we consider an example from the test set featuring $U = 5$ sources in the far field: A drum at the direction $\mathbf{\Theta}_1 = (\frac{7}{10}\pi, \frac{2}{5}\pi)^{\mathrm{T}}$, violins at $\mathbf{\Theta}_2 = (\frac{13}{50}\pi, 0)^{\mathrm{T}}$ and $\mathbf{\Theta}_3 = (\frac{39}{50}\pi, \frac{-7}{250}\pi)^{\mathrm{T}}$, a piano at $\mathbf{\Theta}_4 = (\frac{43}{90}\pi, \frac{10}{15}\pi)^{\mathrm{T}}$, and an acoustic guitar at $\mathbf{\Theta}_5 = (\frac{41}{90}\pi, \frac{-22}{45}\pi)^{\mathrm{T}}$. Fig. 3 displays the SRP maps from the input signal of order 1, the target signal of order 6, and the estimated HOA signals from both DirAC and the proposed model. The associated sound source indices are included to indicate the correspondence. Here, DirAC achieves a spatial similarity of $\eta = 83\%$, while the proposed model yields $\eta = 86\%$. These values represent the portions surrounding the medians for both models in Fig. 2 to ensure a fair comparison in typical scenarios.

In all cases, non-infinitesimal main lobe width and surrounding ringings result from the order truncation of the HOA signals [28]. The low order of the input leads to insufficient spatial resolution, resulting in a wide main lobe that makes it difficult to distinguish between the sound source directions. The target displays five clear peaks, some with lower magnitudes due to randomly assigned amplitudes during data generation (cf. Sec. 3.2). Although DirAC accurately reflects most peaks, it shows high-magnitude areas between certain peaks where there is no sound source present. For example, this occurs between the two violins (indices 2-3), and the violin and the acoustic guitar (indices 3-5). In contrast, the proposed model captures all these peaks while effectively suppressing undesired components in the resulting signal, ultimately achieving better spatial resolution. In the given example, the proposed model demonstrates better performance in terms of achieving a more physically accurate sound field representation as also indicated by its higher spatial similarity.

5. CONCLUSION

We propose a low-complexity deep learning approach for Ambisonics upscaling that exploits the frequency-independent spherical harmonics basis functions in the HOA transformations. By processing each subband independently, the model remains computationally efficient. It is trained on synthetic sinusoidal data with harmonics and time-varying frequencies, and evaluated on musical instrument and speech recordings. Simulation results show that, despite being trained only on synthetic data, the proposed method outperforms the parametric time-frequency approach DirAC in reproducing non-diffuse sound fields, using spatial similarity between estimated and target signals as an objective metric. Subjective listening tests are left for future work.

6. REFERENCES

[1] M. Kentgens, S. Kühl, C. Antweiler, and P. Jax, "From spatial recording to immersive reproduction—Design & implementation of a 3DOF audio-visual VR system," in *Audio Engineering Society Convention*, New York, NY, USA, October 2018.

[2] M. A. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Eng. Soc.*, vol. 21, no. 1, pp. 2–10, 1973.

[3] Røde Microphones, "NT-SF1 Ambisonics microphone." [Online]. Available: https://www.rode.com/

[4] Soundfield Ltd., "SPS200 Ambisonics microphone." [Online]. Available: https://www.soundfield.com/#/products/sps200/

[5] M. Frank, "How to make Ambisonics sound good," in *Proceedings Convention of the European Acoustics Association Forum Acusticum*, Krakow, Poland, September 2014.

[6] J. Daniel, S. Moreau, and R. Nicol, "Further investigations of high-order Ambisonics and wavefield synthesis for holophonic sound imaging," in *Audio Engineering Society Convention*, Amsterdam, The Netherlands, March 2003.

[7] A. Solvang, "Spectral impairment of two-dimensional higher order Ambisonics," *J. Audio Eng. Soc.*, vol. 56, no. 4, pp. 267–279, 2008.

[8] M. Frank, F. Zotter, and A. Sontacchi, "Localization experiments using different 2D Ambisonics decoders," in *VDT International Convention*, Leipzig, Germany, November 2008.

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

[10] A. Gupta and T. D. Abhayapala, "Three-dimensional sound field reproduction using multiple circular loudspeaker arrays," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1149–1159, 2010.

[11] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *J. Audio Eng. Soc.*, vol. 53, no. 11, pp. 1004–1025, 2005.

[12] A. Wabnitz, N. Epain, A. McEwan, and C. Jin, "Upscaling Ambisonic sound scenes using compressed sensing techniques," in *Proceedings IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, USA, October 2011, pp. 1–4.

[13] A. Wabnitz, N. Epain, and C. T. Jin, "A frequency-domain algorithm to upscale Ambisonic sound scenes," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Kyoto, Japan, March 2012, pp. 385–388.

[14] G. Routray and R. M. Hegde, "Sparse plane-wave decomposition for upscaling Ambisonic signals," in *Proceedings IEEE International Conference on Signal Processing and Communications*, virtual conference, July 2020, pp. 1–5.

[15] V. Pulkki, S. Delikaris-Manias, and A. Politis, *Parametric time-frequency domain spatial audio*. Wiley Online Library, 2018.

[16] J. Paulus, L. Laaksonen, T. Pihlajakuja, M.-V. Laitinen, J. Vilkamo, and A. Vasilache, "Metadata-assisted spatial audio (MASA)–an overview," in *Proceedings IEEE International Symposium on the Internet of Sounds*, Erlangen, Germany, September 2024, pp. 1–10.

[17] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, 1997.

[18] L. Zhang, X. Wang, R. Hu, D. Li, and W. Tu, "Estimation of spherical harmonic coefficients in sound field recording using feed-forward neural networks," *Multimedia Tools and Applications*, vol. 80, pp. 6187–6202, 2021.

[19] J. Xia and W. Zhang, "Upmix B-format Ambisonic room impulse responses using a generative model," *Applied Sciences*, vol. 13, no. 21, p. 11810, 2023.

[20] G. Routray, S. Basu, P. Baldev, and R. M. Hegde, "Deep-sound field analysis for upscaling Ambisonic signals," in *Proceedings EAA Spatial Audio Signal Processing Symposium*, Paris, France, September 2019, pp. 1–6.

[21] E. Chatzimoustafa and P. Jax, "Higher-order Ambisonics upscaling using gated recurrent units," in *Proceedings IEEE European Signal Processing Conference*, Palermo, Italy, September 2025, pp. 171–175.

[22] K. Cho, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[23] P. Vary and R. Martin, *Digital Speech Transmission and Enhancement*. John Wiley & Sons, 2024.

[24] B. Rafaely, *Fundamentals of Spherical Array Processing*. Springer, 2015, vol. 8.

[25] J. Daniel, "Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new Ambisonic format," in *Proceedings AES International Conference on Signal Processing in Audio Recording and Reproduction*, Helsingør, Denmark, May 2003.

[26] C. Nachbar, F. Zotter, E. Deleflie, and A. Sontacchi, "Ambix-a suggested Ambisonics format," in *Proceedings Ambisonics Symposium*, Lexington, Kentucky, USA, July 2011.

[27] R. M. Gray, *Probability, random processes, and ergodic properties*. Springer Science & Business Media, 2009.

[28] M. Kentgens, *Signal Processing Concepts for User Movement in Scene-Based Spatial Audio*. Ph.D. dissertation. Shaker Verlag, 2023.

[29] A. Herzog and E. A. Habets, "Direction preserving Wiener matrix filtering for Ambisonic input-output systems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Brighton, UK, May 2019, pp. 446–450.

[30] J. Fliege and U. Maier, "The distribution of points on the sphere and corresponding cubature formulae," *IMA Journal of Numerical Analysis*, vol. 19, no. 2, pp. 317–334, 1999.

[31] E. B. Union, "EBU Tech 3253 - Sound quality assessment material recordings for subjective tests," Geneva, Switzerland, September 2008.

[32] E. Bakhturina, V. Lavrukhin, B. Ginsburg, and Y. Zhang, "Hi-Fi Multi-Speaker English TTS Dataset," *arXiv preprint arXiv:2104.01497*, 2021.

[33] S. P. Mohanty, "Musical instrument's sound dataset." [Online]. Available: https://www.kaggle.com/datasets/soumendraprasad/musical-instruments-sound-dataset/data

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings Conference on Learning Representations*, San Diego, California, US, May 2015.

