Higher-Order Ambisonics Upscaling Using Gated Recurrent Units

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

***Abstract*—Higher-order Ambisonics (HOA) offer a flexible way to represent 3D sound field information, which makes them suitable for many applications, e.g., virtual reality (VR) and teleconferencing. However, the HOA order which dictates spatial accuracy is practically constrained by the number of microphones and loudspeakers. This work aims to increase the accuracy of the sound field representation by predicting missing HOA coefficients for higher orders. To achieve this, an existing deep learning-based upscaling method utilizes fully connected feedforward neural networks. Our novel approach replaces these fully-connected structures with gated recurrent units (GRUs), which allow to better leverage spatio-temporal dependencies inherent in HOA coefficients. Simulation experiments show that when trained under similar conditions, the proposed model outperforms the previous one by achieving lower mean squared error (MSE) between target and predicted HOA coefficients across various upscaling orders. In further experiments, we train the proposed model on synthetic sinusoidal data and evaluate the performance on test sets of complex real-world recordings. The superior performance of the proposed model in these experiments indicates its value in scenarios where obtaining real acoustic scene data with high orders is impractical.**

***Index Terms*—Audio signal processing, higher-order Ambisonics, recurrent neural networks, upscaling**

I. Introduction

Spatial audio creates realistic acoustic environments by placing sound sources in three-dimensional space. This enables an immersive listening experience, particularly in virtual reality (VR), where these environments are synchronized with the visual effects of the virtual world. The pioneering work of Gerzon [1] laid the foundation for spatial audio technology based on Ambisonics. A notable example is the Soundfield SPS200 microphone, which can capture Ambisonics signals for various applications [2]. Recent advances in multi-channel audio recording and playback systems have made the transition from Ambisonics to higher-order Ambisonics (HOA) feasible, which enhances spatial resolution and enables more accurate sound field representation [3]–[5].

The HOA order is directly related to the accuracy of the sound field representation, but it is fundamentally limited by the number of microphones or loudspeakers in the recording or reproduction setup, respectively [6], [7]. A low HOA order offers only low spatial selectivity, whereas high orders require a large number of microphones to capture that may be impractical in many applications. Consequently, sound fields can only be reproduced with little error within a specific reproduction area known as the *physical sweet spot*, whose size depends on the HOA order and frequency [8], [9].

To enhance sound field representation, parametric approaches such as directional audio coding (DirAC) [10] model sound fields based on the direction of arrival (DOA) and diffuseness. Many studies focus on artificially increasing HOA orders by leveraging sparsity and assuming few incident sound directions [11]–[14]. Other methods specifically concentrate on performing upscaling on sampled spherical grids [15], [16].

Routray et al. [17] proposed a deep learning model that uses fully connected feedforward neural networks for HOA upscaling, which eliminates the need for prior source direction estimation. In this work, we adopt their framework and introduce a novel approach by replacing the fully connected structures with gated recurrent units (GRUs). This modification allows us to capture spatio-temporal dependencies inherent in HOA coefficients more effectively. We demonstrate the proposed model's generalization capacity by training it on an artificially generated set of sinusoidal and harmonic samples, and evaluate the performance on test sets of complex real-world recordings.

We compare the proposed model to the one presented in [17] based on performance. First, we calculate the average mean squared error (MSE) for both models by using given test sets. Second, we illustrate sound field reproduction performance through an example that consists of recorded musical instruments. Finally, to evaluate accuracy in detecting sound source directions, we consider the steered response power (SRP) map by using a sample from the test set as another example.

II. Theoretical Background

A. Plane-Wave Decomposition

A complex-valued sound pressure field in the frequency domain can be expressed as the sum of individual unit-amplitude, single-frequency plane-wave terms of the form $e^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}}$. Assuming that the sound field is composed of infinitely many plane waves with directional amplitude density $x(k, \theta_q, \phi_q)$, the overall sound pressure can be formulated as the surface integral [9], [18]

$$p(k, r, \theta, \phi) = \int_0^{2\pi} \int_0^{\pi} x(k, \theta_q, \phi_q) e^{-\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{r}} \sin\theta_q d\theta_q d\phi_q, \quad (1)$$

where the wave vector is given by $\boldsymbol{k} = -(k, \theta_q, \phi_q)$ and the position $\boldsymbol{r} = (r, \theta, \phi)$ is described in spherical coordinates with radius r, inclination angle θ, and azimuth angle ϕ. Here, $k = \frac{2\pi f}{c}$ denotes the wavenumber with f representing the physical frequency and c the speed of sound. The symbol i denotes the imaginary unit.

ISBN: 978-9-46-459362-4

Although it is defined for plane waves, (1) is also applicable to point sources in the far field, where their distribution can be assumed to be equivalent to that of plane waves [6]. Since the directional amplitude density is defined on the unit sphere, it is appropriate to express it in terms of a spherical harmonics expansion according to [9]

$$x(k,\theta_q,\phi_q) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} x_{nm}(k) Y_n^m(\theta_q,\phi_q), \tag{2}$$

where $Y_n^m(\theta_q,\phi_q)$ are the spherical harmonics and $x_{nm}(k)$ are the corresponding weights with order $n = (0,1,2,\dots)$ and degree $m = (-n,\dots,n)$, which adheres to the Ambisonics channel number (ACN) format [19]. Inserting (2) into (1) and expressing individual plane-wave terms as a summation of spherical harmonics yields an expression of the sound pressure field as a series [9]

$$p(k,r,\theta,\phi) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} 4\pi \mathrm{i}^n x_{nm}(k) j_n(kr) Y_n^m(\theta,\phi) \tag{3}$$

with the spherical Bessel function $j_n(kr)$. The time-domain representation of $x_{nm}(k)$ is referred to as the HOA signal, which provides a more abstract representation of the directional amplitude density, as it does not depend on the source direction (θ_q,ϕ_q) [8]. Note that both spherical harmonics and HOA representations are defined as complex-valued functions. Thus, the pressure field is given as the real part of (3). For simplicity, we omit the explicit notation and consider all quantities as real-valued signals in the following discussion.

B. Band Limitation of Higher-Order Ambisonics Signals

As mentioned in Sec. I, the challenge of representing a sound field through HOA lies in the fact that the coefficients are only available up to a finite truncation order N. Consequently, (3) is merely an approximation in practice, i.e.,

$$p(k,r,\theta,\phi) \approx \sum_{n=0}^{N}\sum_{m=-n}^{n} 4\pi \mathrm{i}^n x_{nm}(k) j_n(kr) Y_n^m(\theta,\phi). \tag{4}$$

For orders $n > kr$, the magnitude of the normalized spherical Bessel function $|4\pi \mathrm{i}^n j_n(kr)|$ in (3) decreases significantly, which causes the individual summation terms to vanish. As a result, the approximation (4) is valid with minimal error within the physical sweet spot given by a radius of approximately $r < \frac{N}{k}$ [9].

Using matrix notation and considering discrete sound sources, the relationship between band-limited HOA coefficients and direction-specific amplitudes can be expressed as [11], [13]

$$\boldsymbol{x_{nm}}(t) = \boldsymbol{Y}\boldsymbol{x}(t), \tag{5}$$

where t denotes the discrete time index. The spherical harmonics evaluated at all sound source directions are stored in an $(N+1)^2 \times Q$ matrix given by

$$\boldsymbol{Y} = \begin{pmatrix} Y_0^0(\theta_1,\phi_1) & Y_0^0(\theta_2,\phi_2) & \dots & Y_0^0(\theta_Q,\phi_Q) \\ Y_1^{-1}(\theta_1,\phi_1) & Y_1^{-1}(\theta_2,\phi_2) & \dots & Y_1^{-1}(\theta_Q,\phi_Q) \\ \vdots & \vdots & \ddots & \vdots \\ Y_N^N(\theta_1,\phi_1) & Y_N^N(\theta_2,\phi_2) & \dots & Y_N^N(\theta_Q,\phi_Q) \end{pmatrix} \tag{6}$$

with Q representing the number of sound sources in the sound field. The direction-specific amplitudes are represented as a vector with length Q,

$$\boldsymbol{x}(t) = (x(t,\theta_1,\phi_1), x(t,\theta_2,\phi_2), \dots, x(t,\theta_Q,\phi_Q))^{\mathrm{T}}, \tag{7}$$

where $(\cdot)^{\mathrm{T}}$ denotes transposition of a vector. The coefficient vector of length $(N+1)^2$ is expressed as

$$\boldsymbol{x_{nm}}(t) = (x_{0,0}(t), x_{1,-1}(t), x_{1,0}(t), \dots, x_{N,N}(t))^{\mathrm{T}}. \tag{8}$$

III. Higher-Order Ambisonics Upscaling

We consider a processing framework for sequential signals, where each signal consists of a single frame with N_t time samples. The problem addressed in this work is to find a block of multiple HOA coefficients of order $\hat{N}$ from a first-order Ambisonics representation, i.e.,

$$\boldsymbol{x}_{\boldsymbol{nm}}^{(1)}(t) \xrightarrow{\text{Upscaling}} \hat{\boldsymbol{x}}_{\boldsymbol{nm}}^{(\hat{N})}(t) \quad \text{with} \quad \hat{N} > 1. \tag{9}$$

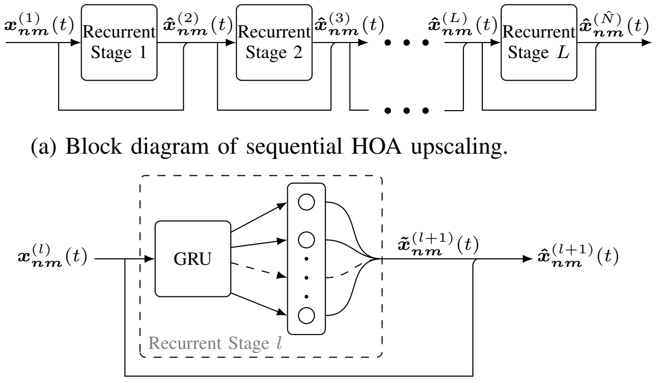
Since the input vector contains 4 coefficients and the target vector $(\hat{N}+1)^2$ coefficients, the problem of HOA upscaling involves predicting $(\hat{N}+1)^2 - 4$ missing coefficients to obtain the $\hat{N}$-th order HOA signal. In the following, we propose a modification to an existing upscaling framework [17] by utilizing recurrent neural networks instead of fully connected layers for this task, and elaborate on the training data.

A. Proposed Framework

It is evident that the number of unknown coefficients relative to the known coefficients increases quadratically as the target order $\hat{N}$ grows. Therefore, it seems reasonable to consider a sequential approach similar to that proposed in [17].

Fig. 1a illustrates the block diagram of the adopted HOA upscaling framework. Our novel approach involves $L = \hat{N} - 1$ independently trained recurrent stages, each of which increments the HOA order by one. Each stage l follows the same training procedure: It takes the l-th order input $\boldsymbol{x}_{\boldsymbol{nm}}^{(l)}(t) \in \mathbb{R}^{(l+1)^2}$, predicts the missing coefficients $\tilde{\boldsymbol{x}}_{\boldsymbol{nm}}^{(l+1)}(t) \in \mathbb{R}^{(2l+3)}$ required for generating the next-order HOA signal, and concatenates these predicted coefficients with its input vector to form the predicted signal $\hat{\boldsymbol{x}}_{\boldsymbol{nm}}^{(l+1)}(t) \in \mathbb{R}^{(l+2)^2}$. The processing is performed iteratively, i.e., the output of stage l becomes the input of stage $l+1$ until the target HOA order $\hat{N}$ is reached.

Each recurrent stage of the model is selected as a GRU followed by a fully connected output layer, chosen for their simpler architecture and fewer parameters than those of fully connected structures in [17], which reduce the risk of overfitting [20]. Additionally, the use of GRUs allows for effective memory retention over longer observation periods, even with shorter block lengths. This capability is particularly beneficial when dealing with relatively stationary signals, where source positions do not change abruptly. In such scenarios, GRUs can be interpreted as providing smoothing across both time and spatial information in the HOA coefficients, which makes them more suitable than simple fully connected layers for capturing spatio-temporal dependencies. Fig. 1b illustrates the structure of each proposed recurrent stage within the upscaling



(a) Block diagram of sequential HOA upscaling.

(b) Structure of recurrent stage l.

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

framework. Each GRU in these stages is configured as unidirectional, i.e., each time step's output depends solely on the current and previous time steps. This configuration allows the GRU model to be suitable for online applications. The GRU consists of a single layer that takes $(l+1)^2$ inputs and produces N_h outputs per time t, where N_h denotes the number of features in the hidden state. The fully connected output layer performs a linear transformation that maps the GRU's output to $\tilde{\boldsymbol{x}}^{(l+1)}_{\boldsymbol{nm}}(t) \in \mathbb{R}^{(2l+3)}$ with a linear activation. The GRU hidden size was selected experimentally as $N_h = 128$. Inference based on the upscaling framework is performed in the same sequential manner as previously described.

B. Data Generation

The proposed approach will be tested for practical usability, which requires data. Since real-world data are not always available in sufficient amounts, this study explores scenarios where data are generated synthetically. Several plausible variants for data generation are presented in this section and their impact on model performance will be compared in the evaluation.

First, we generated HOA signals by using (5), where individual sound stimuli were randomly selected from a subset of the EBU-SQAM database [21]. This subset consists of recordings, which contain 6 tonal signals, 36 musical instruments, 4 opera pieces, and 6 speech excerpts. From this subset, we created 2×10^5 training samples by using only the speech signals and another 2×10^5 samples that included the remaining types of signals. Each generated acoustic scene consists of 1 to 5 individual sound sources randomly selected from a uniform distribution, with random amplitudes in the range $u_q \in (0.1, 1)$ with source index q. Each source emits plane waves from a single, random direction θ_q and ϕ_q within the angles $\theta_q \in (0, \pi/2)$ and $\phi_q \in (0, \pi)$, respectively. These angle ranges are defined identically to those described in [17].

We generated another 4×10^5 acoustic scenes, each consisting of 1 to 5 artificial sinusoidal sources, without recordings, unlike the previously mentioned dataset. Each sinusoidal source in an acoustic scene has a random amplitude in range $u_q \in (0.1, 1)$, a phase shift $\Delta\alpha_q \in (0, 2\pi)$, and a frequency $f_q \in (200, 1500)\,\mathrm{Hz}$. This frequency range was motivated by the fundamental frequencies of typical (non-bass) musical instruments and speech signals. The plane waves from each source arrive from directions θ_q and ϕ_q within their respective ranges. This simple dataset created a controlled environment that allowed for quick adjustments of various configurations and enabled us to test their impact on model performance.

For the evaluation, we created two test sets: One with 8×10^4 speech signals and the other with 8×10^4 tonal components, musical instruments, and opera samples from the EBU-SQAM subset. As in the training set based on the EBU-SQAM subset, we used the same parameters to generate the acoustic scenes: Each scene consists of 1 to 5 sources with random amplitudes and source directions within the same specified ranges.

Furthermore, considering the inherent tonality of musical instruments and speech signals [22], we introduced harmonics into the sinusoidal training set to improve the generalization of the proposed model. To achieve this, we added four harmonic components to approximately half of the sinusoidal scenes by incorporating the original signal along with integer multiples of the fundamental frequency and an exponential decay factor. This modified the source signal q as follows:

$$x(t, \theta_q, \phi_q) = u_q \sum_{\kappa=0}^{4} e^{-\kappa\beta_q} \sin((\kappa + 1)\Omega_q t + \Delta\alpha_q) \tag{10}$$

with normalized angular frequency $\Omega_q = 2\pi \frac{f_q}{f_s}$ and a decay factor of $\beta_q = 1$. Thus, we created the third training set by using (10). The effect of this set will be discussed in Sec. IV.

In all generated datasets, each individual training sample represents a unique acoustic scene, with a frame size $N_t = 512$ at a sampling rate of $f_s = 44.1\,\mathrm{kHz}$. All frames are processed by the networks independently.

IV. Evaluation

We consider the model from [17] as the baseline and compare it with the proposed model. Each model was trained sequentially up to order $\hat{N} = 6$. The baseline model was trained only on the EBU-SQAM subset, whereas the proposed model was trained on three different training sets: EBU-SQAM subset, sinusoidal dataset, or sinusoidal dataset with harmonics according to (10). During training of each individual stage for upscaling from order l to $l + 1$, we used the MSE loss function to compare predicted HOA signals $\tilde{\boldsymbol{x}}^{(l+1)}_{\boldsymbol{nm}}(t)$ with their corresponding targets $\breve{\boldsymbol{x}}^{(l+1)}_{\boldsymbol{nm}}(t)$. The batch size was set to 64, and the learning rate was configured at 10^{-4}. We employed Adam optimizer for stochastic optimization.

When training the baseline model, we closely followed the parameters from the original paper [17] to ensure consistency and comparability. Each stage in the baseline model was trained for 300 epochs based on convergence patterns indicated by stabilized training and validation loss. In contrast, the recurrent stages in the proposed model on all datasets achieved convergence within a maximum of 100 epochs under similar conditions. To monitor convergence patterns and modify the learning rate accordingly, we employed a scheduler with a patience of 10 epochs.

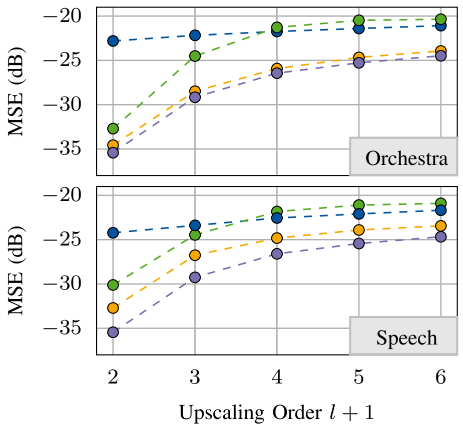


Fig. 2: MSE on orchestra and speech test sets for the baseline model (--●--), and the proposed model trained on EBU-SQAM subset, sinusoidal data, or sinusoidal data with harmonics (as defined in (10)) (--●--/ --●--/ --●--), respectively.

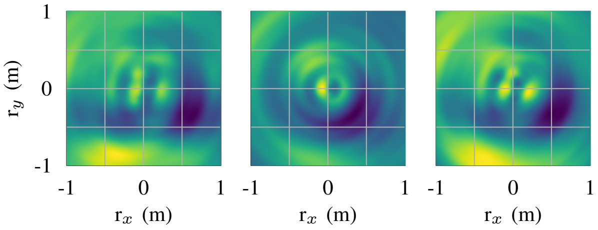
A. Deviation of Higher-Order Ambisonics Coefficients

To compare the overall performance of the models, we first examined HOA prediction accuracy across 8×10^4 orchestra and 8×10^4 speech scenes used as test sets (as detailed in Sec. III-B). We computed the average MSE between the $2l+3$ upscaled coefficients $\tilde{\boldsymbol{x}}_{\boldsymbol{nm}}^{(l+1)}(t)$ and the targets $\breve{\boldsymbol{x}}_{\boldsymbol{nm}}^{(l+1)}(t)$ for each upscaling order $l+1$ until the final order $\hat{N}$, i.e., for $l = 1, 2, \ldots, \hat{N}-1$. The target signal was calculated using (5).

Fig. 2 illustrates performance of all models on both test sets for different upscaled HOA orders. A significant difference in performance is observed when comparing the baseline model with the proposed model trained on the same EBU-SQAM subset. For example, the MSE is reduced by over 10 dB at upscaling order 2 for both test sets. This discrepancy is expected because the proposed model leverages the inherent spatio-temporal dependencies in HOA signals more effectively. In all cases, errors increase with higher upscaling orders, while the proposed model's performance remains better than that of the baseline model across all considered upscaling orders.

When evaluating the proposed model trained on sinusoidal training sets, we observe a significant performance improvement by incorporating harmonics (according to (10)), compared to training solely on sinusoidal data. Interestingly, the latter variant performs worse than the baseline starting from upscaling order 4 for both the speech and orchestra test sets. This indicates that training exclusively on sinusoidal data is not sufficient for effective generalization on the EBU-SQAM subset, and incorporating harmonics enhances the model's ability to represent the more complex test sets.

Another advantage of using a sinusoidal training set with harmonics becomes evident when comparing the proposed model trained on harmonics to the one trained on the EBU-SQAM subset. The results for the orchestra test set are on-par, while those for the speech test set are nearly similar, with the difference in the latter case bounded by approximately 3 dB. Overall, this demonstrates that training on sinusoidal data with harmonics achieves similar performance without requiring prior recordings of sound sources. This approach could be particularly beneficial in situations where recording sound scenes may be impractical. In the following, we will focus exclusively on this proposed model variant.



(a) Target (b) Baseline (c) Proposed

Fig. 3: Reproduced sound fields with HOA of order 6 using (4), with target, baseline and proposed model coefficients.

B. Sound Field Reproduction

For better interpretation of the results, we selected a sample from the test set and visualized sound field reproduction performance at HOA of order 6. Fig. 3 illustrates the reproduced sound fields which feature three musical instruments in the far field: A grand piano, an oboe, and a violoncello, with their plane waves arriving at the center from directions $(\theta_1, \phi_1) = (\frac{\pi}{6}, \frac{\pi}{18})$, $(\theta_2, \phi_2) = (\frac{\pi}{4}, \frac{17}{18}\pi)$, and $(\theta_3, \phi_3) = (\frac{7}{18}\pi, \frac{\pi}{2})$, respectively. The reproduced sound fields were simulated by using (4) over an area of 4 m^2, with the target signal, baseline model predictions, and proposed model predictions inserted as HOA signals. We observe the sound field in the horizontal plane, i.e., at $\theta = \frac{\pi}{2}$, and at a fixed time $t = 210$.

This chosen scene creates a complex sound field due to wave interference within the reproduction area. When reproduced with HOA of order 6, many aspects of this complexity can be accurately represented, as evident in the target signal (see Fig. 3a). Comparing the reproduced sound fields from both models reveals that the baseline model captures only a portion of the overall sound field (see Fig. 3b), while the proposed model's results closely align with the target (see Fig. 3c). This indicates that the proposed model successfully reflects the larger physical sweet spot through the higher HOA order.

Remarkably, even with the proposed model, some waves appear slightly rotated clockwise. This phenomenon can be attributed to errors in the resulting coefficients, where certain orders leak into others, which alters spatial information.

C. Steered Response Power Maps

Another important aspect of spatial information about a sound field is the sound source directions, which can be visualized with HOA through signal energy distribution via SRP maps [8]. To compare model performance in detecting source directions, we considered the same exemplary sound field from the test set described in Sec. IV-B, and created four SRP maps based on this sound field: One using initial first-order

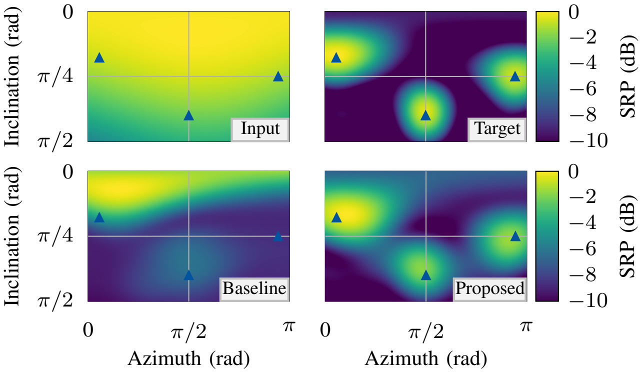


Fig. 4: SRP maps created using first-order Ambisonics input, target HOA signal at order 6, and predicted HOA coefficients by the baseline and proposed models at order 6 for an exemplary sound field from the test set. Ground truth sound source directions are marked by triangles (▲).

Ambisonics coefficients as input of the models, one using the target HOA signal at order 6, and two using predicted HOA coefficients from both the baseline and proposed models at the upscaling order 6.

Fig. 4 displays the resulting SRP maps. Due to strong band limitation from the lower order of the initial first-order Ambisonics, there is low spatial resolution that results in significant spatial blur [8], which makes it difficult to distinguish between sound sources. The map from the target signal at order 6 exhibits an energy distribution with three distinct peaks which correspond to the ground truth sound source directions. The baseline model reflects only one of these peaks and shows lower energy distribution for the remaining sound sources. In contrast, the proposed model accurately captures all three peaks, which demonstrates a high degree of spatial directivity. However, both the target and proposed models still exhibit some spatial blur due to the HOA band limitation. For the proposed model, imperfect HOA estimations slightly exaggerate this blur. Despite this effect, the overall energy distribution with three distinct peaks is preserved.

V. Conclusion

We consider an existing deep learning framework for HOA upscaling and propose using GRUs instead of fully connected networks within this framework. Simulation experiments show that leveraging spatio-temporal dependencies in HOA signals more effectively through GRUs enhances model performance in predicting missing coefficients for higher orders. The proposed model outperforms the previous one when both are trained on realistic data. In further experiments, we trained the proposed model by using artificial sinusoidal data with harmonics, which shows better performance than the previous model trained on realistic data. This improvement indicates that incorporating synthetic training samples could serve as a suitable method for data augmentation. In resource-constrained situations, e.g., when real scene training data are unavailable or costly, the proposed model remains effective by achieving good performance on complex real-world data of musical instruments and speech. In future research, we aim to conduct listening tests and incorporate parametric methods, e.g., DirAC for a comparative analysis with our deep learning approaches.

References

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

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

[3] 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 114*, Amsterdam, The Netherlands, March 2003.

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

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

[6] D. B. Ward and T. D. Abhayapala, “Reproduction of a plane-wave sound field using an array of loudspeakers,” *IEEE Trans. Speech Audio Process.*, vol. 9, no. 6, pp. 697–707, 2001.

[7] 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.

[8] M. Kentgens, “Signal processing concepts for user movement in scene-based spatial audio,” Ph.D. dissertation, RWTH Aachen University, 2023.

[9] B. Rafaely, *Fundamentals of spherical array processing*. Springer, 2015, vol. 8.

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

[11] 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. (ICASSP)*, Kyoto, Japan, March 2012, pp. 385–388.

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

[13] 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 (WASPAA)*, New Paltz, New York, USA, October 2011, pp. 1–4.

[14] M. Kentgens, S. Al Hares, and P. Jax, “On the upscaling of higher-order Ambisonics signals for sound field translation,” in *Proceedings IEEE European Signal Processing Conference (EUSIPCO)*, Dublin, Ireland, August 2021, pp. 81–85.

[15] 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.

[16] T. Lübeck, J. M. Arend, and C. Pörschmann, “Spatial upsampling of sparse spherical microphone array signals,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1163–1174, 2023.

[17] 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.

[18] E. Williams, *Fourier Acoustics: Sound Radiation and Nearfield Acoustical Holography*. Academic Press, 1999.

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

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

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

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