Supplementary Materials for "Modeling and Driving Human Body Soundfields through Acoustic Primitives"

Chao Huang[1], Dejan Marković[2], Chenliang Xu[1], and Alexander Richard[2]

[1] University of Rochester, Rochester, NY, USA
[2] Codec Avatars Lab, Meta, Pittsburgh, PA, USA
{chaohuang,chenliang.xu}@rochester.edu,{dejanmarkovic,richardalex}@meta.com

1 Rendering Audio with Learned Soundfield

As illustrated in Sec 3.2, the sound pressure, *i.e.* the audio signal, produced by a learned soundfield of order N is given by

$$\begin{aligned}\mathbf{w}(r,\theta,\varphi) &= \sum_{n=0}^{N}\sum_{m=-n}^{n}\left(c_{nm}\cdot h_n(kr_{ref})\right)\cdot\frac{h_n(kr)}{h_n(kr_{ref})}\cdot Y_{nm}(\theta,\varphi)\\ &= \sum_{n=0}^{N}\sum_{m=-n}^{n}\tilde{c}_{nm}\cdot\frac{h_n(kr)}{h_n(kr_{ref})}\cdot Y_{nm}(\theta,\varphi),\end{aligned} \tag{1}$$

here, $k = 2\pi f/v_{sound}$ is the corresponding wavenumber; $Y_{nm}(\theta,\varphi)$ represents the spherical harmonic of order n and degree m, which is

$$Y_{nm}(\theta,\varphi) \equiv \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}}P_{nm}(\cos\theta)e^{im\varphi}, \tag{2}$$

$P_{nm}(z)$ is an associated Legendre polynomial, and $h_n(kr)$ is nth-order spherical Hankel functions. All the functions are implemented with PyTorch and, therefore are fully differentiable. In this paper, we choose spherical harmonics up to second order and showcase them in Fig. 1.

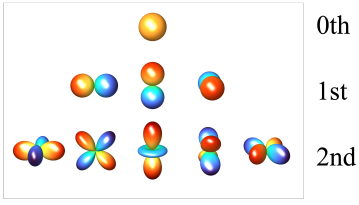

Fig. 1: Spherical harmonics up to second order.

Similarly to $Y_{nm}(\theta,\varphi)$, the learned soundfield, *i.e.* acoustic primitive, can also be decomposed into a series of spherical harmonics functions, each representing a

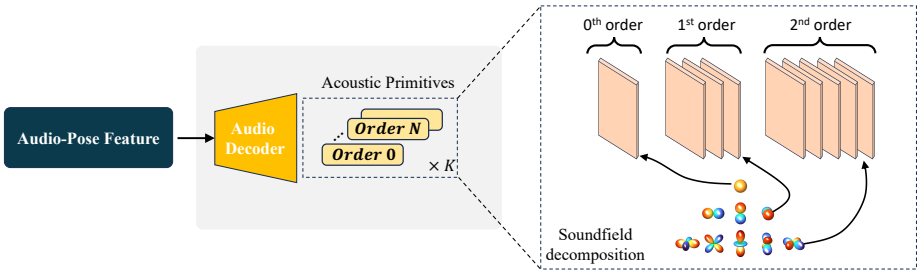

Fig. 2: Decomposition of learned soundfield representation and its connection to second order spherical harmonics.

different spatial component of the soundfield. We demonstrate the decomposition process in Fig. 2. Our learned soundfield representation is enforced to express the same spatial information as spherical harmonics because each predicted acoustic primitive has $(N+1)^2$ channels, which is equivalent to the number of spherical harmonics.

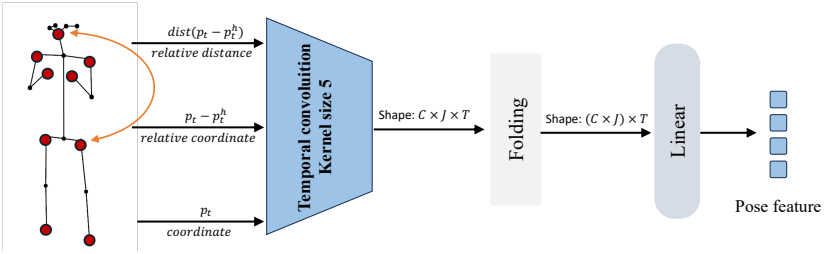

Fig. 3: Illsutration on pose feature encoding, we consider the absolute coordinates and the relative coordinates/distances (to the headset) of body joints.

2 Pose Feature Encoding

To capture these rich spatial cues, we employ a pose encoder that processes the input pose sequence. This sequence, denoted as $\mathbf{p}_{1:T_p}$, contains the 3D coordinates of body joints for each frame $\mathbf{p}_t \in \mathbb{R}^{J\times 3}$. However, since these coordinates are captured from a third-person perspective, they might not fully capture the spatial relationship relevant to the audio, where the sound originates from the body but is received at the headset. To address this, we enhance the pose input by selecting the head joint $\mathbf{p}_t^h$ as an anchor and calculating relative coordinates and Euclidean distances. This extended pose input $[\mathbf{p}_t, \mathbf{p}_t - \mathbf{p}_t^h, dist(\mathbf{p}_t - \mathbf{p}_t^h)] \in \mathbb{R}^{J\times 7}$, consisting of original coordinates, relative coordinates to the head, and distance from the head, provides the pose encoder with a more comprehensive understanding of the body's spatial relationship with the sound. Details are shown in Fig. 3.

3 More Ablations

Visualization of primitive offsets. In Fig. 4, we observe a **time delay** between the predicted and GT audios for the **model without offsets**, likely caused by inaccurate primitive coordinates. In contrast, our model with learned offsets mitigates this issue, resulting in a closer match to the ground truth audio. Also, we visualize the sound fields generated by our framework for different primitives after applying the learned offsets. We observe that the learned offsets generally match location where we would expect the source of sound to be given the particular sound event such as snap, clap, or footstep.

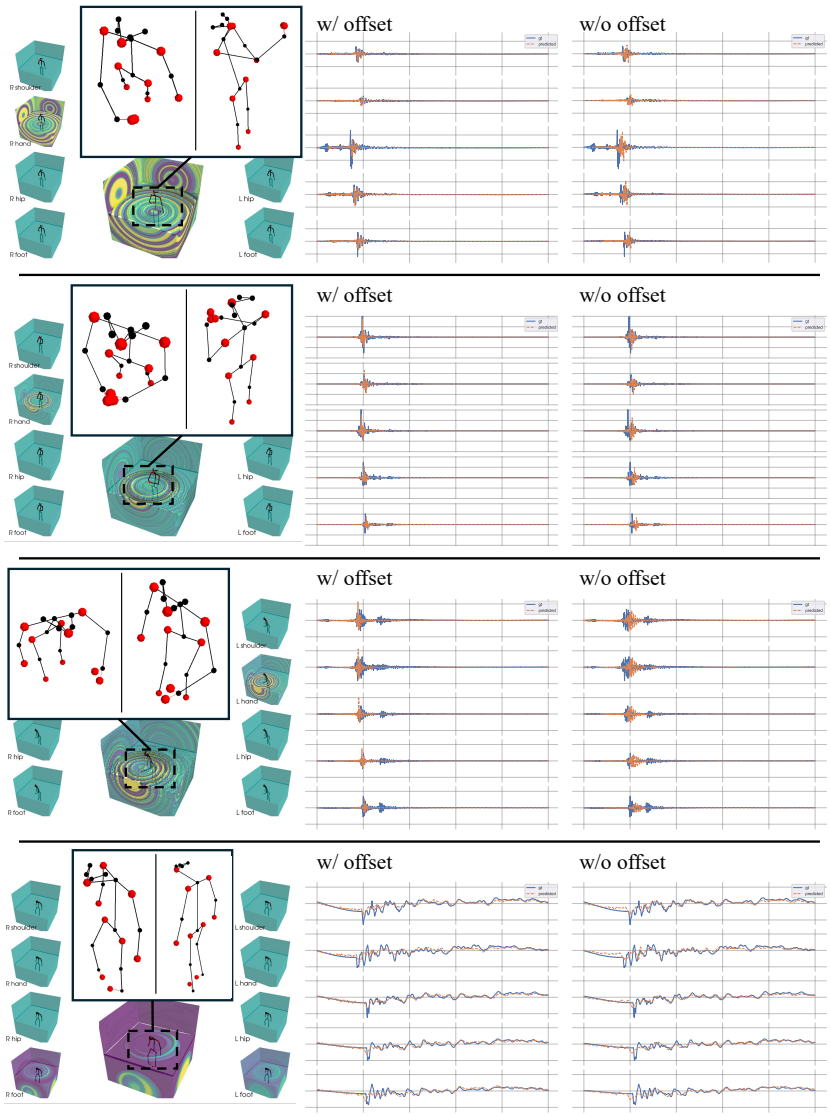

Fig. 4: Ablation on learned acoustic primitive offsets.

Visualization of different harmonic order. Fig. 5 illustrates the impact of sound field order on the accuracy of predicted audio. As shown, the model's prediction using a 2nd-order sound field exhibits **a closer match to the GT audio in terms of amplitude**. This is because higher-order harmonics offer finer spatial rendering capabilities, allowing the model to capture more precise directional details of the sound. In contrast, the predicted 0th-order sound field is omnidirectional, meaning it radiates sound equally in all directions. This limitation hinders its ability to encode specific spatial information, resulting in less accurate audio amplitude prediction.

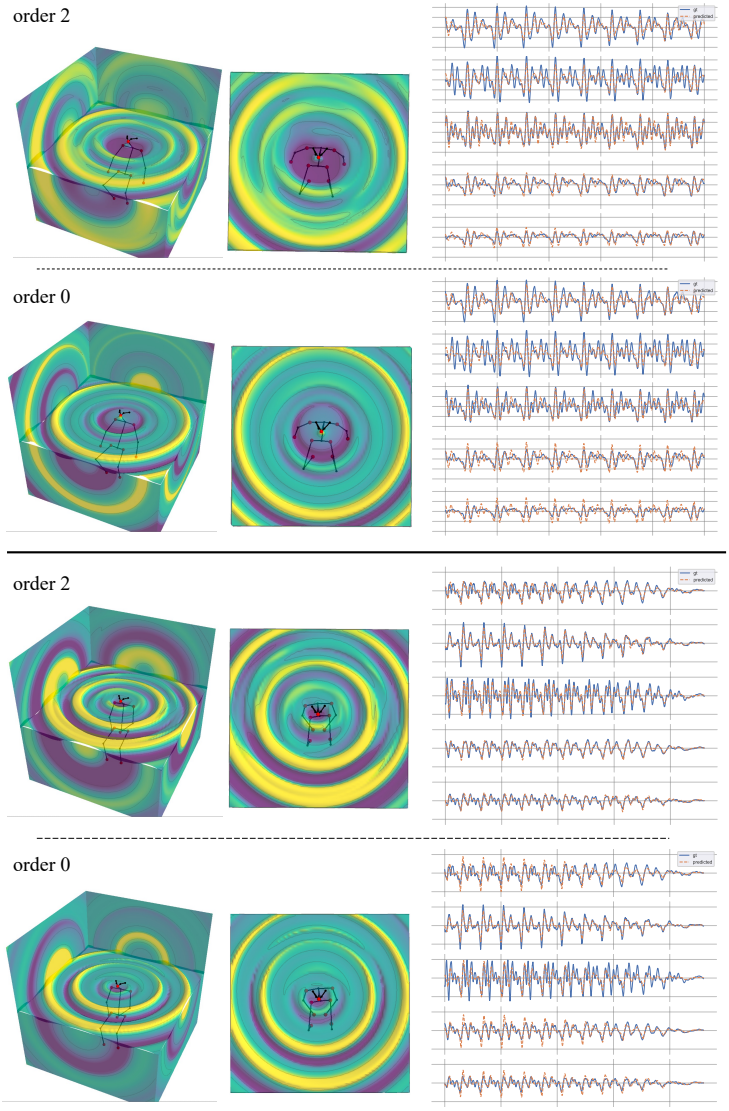

Fig. 5: Ablation on different harmonic orders.

Table 1: Ablation study on loss function. Each loss term is removed from the total loss function one at a time. The best and second-best results are highlighted in green and blue, respectively.

	non-speech			speech		
Loss	SDR ↑	amplitude ↓	phase ↓	SDR ↑	amplitude ↓	phase ↓
$\mathcal{L}_{total}$	3.597	0.883	0.323	8.448	0.943	0.417
w/o $\mathcal{L}_{amp}$	2.099	1.179	0.303	8.370	1.038	0.381
w/o $\mathcal{L}_{ri}$	3.579	0.894	0.325	8.492	0.933	0.418
w/o $\mathcal{L}_{s\ell 1}$	3.338	0.821	0.330	7.722	0.915	0.491
w/o $\mathcal{L}_{cts}$	3.523	0.903	0.324	8.239	0.938	0.422

Ablation on the choices of loss function. In Tab. 1, we conduct an ablation study to investigate the effectiveness of each loss term in the total loss function (Eq. (11) in the main paper). We remove each loss term from the total loss $\mathcal{L}_{total}$ one at a time. The results show that including $\mathcal{L}_{cts}$ improves the overall performance on both speech and non-speech data, and combining all the loss terms yields the best or second-best performance across different metrics for both speech and non-speech data and generally the best performance on average over the metrics.

4 Demo Video

We have prepared a supplementary video to visually demonstrate the capabilities of our method in spatial audio rendering. The video showcases a full-body avatar producing correctly spatialized binaural audio corresponding to various actions and interactions using our trained model. In particular, the input of the audio system is a single channel mono audio that contains the mixture of all sounds being made. Our model can render them with the correct spatial locations using wearer's body pose. This means that the wearer can *clap left, clap right, applaud, snap around,* and the sounds will be appropriately positioned. Additionally, the system works with *objects that the wearer may be using*, such as an egg shaker.

Acknowledgement

The authors would like to thank Frank Yu for engineering work on the VR demo.