

Modeling and Driving Human Body Soundfields through Acoustic Primitives

Chao Huang¹, Dejan Marković², Chenliang Xu¹, and Alexander Richard²

¹ University of Rochester, Rochester, NY, USA

² Codec Avatars Lab, Meta, Pittsburgh, PA, USA

{chao Huang, chenliang.xu}@rochester.edu, {dejanmarkovic, richardalex}@meta.com

Abstract. While rendering and animation of photorealistic 3D human body models have matured and reached an impressive quality over the past years, modeling the spatial audio associated with such full body models has been largely ignored so far. In this work, we present a framework that allows for high-quality spatial audio generation, capable of rendering the full 3D soundfield generated by a human body, including speech, footsteps, hand-body interactions, and others. Given a basic audio-visual representation of the body in form of 3D body pose and audio from a head-mounted microphone, we demonstrate that we can render the full acoustic scene at any point in 3D space efficiently and accurately. To enable near-field and realtime rendering of sound, we borrow the idea of volumetric primitives from graphical neural rendering and transfer them into the acoustic domain. Our acoustic primitives result in an order of magnitude smaller soundfield representations and overcome deficiencies in near-field rendering compared to previous approaches. Our project page: <https://wikichao.github.io/Acoustic-Primitives/>.

Keywords: Human Body Pose · Acoustic Primitives · AR/VR

1 Introduction

Learning, rendering, and animating 3D human body representations has been a long standing research area with applications in gaming, movies, and more recently also AR/VR. MetaHumans [1] and Codec Avatars [3] provide highly realistic models and advances in neural rendering have pushed the visual quality to new frontiers [31, 37, 49]. Animating full-body models has seen significant progress with the availability of generative models, ranging from pose-based animation [3] to audio- and text-driven animation [29, 40, 46]. Overall, *visual* representations of 3D humans these days are of excellent quality and drivable from pose, audio, and text inputs.

However, on the *acoustic* side of the problem, *i.e.*, rendering spatial sound in 3D for these full-body representations, the research landscape looks dire. It has been shown that accurate audio-visual modeling is important for an immersive 3D experience [14] but still almost no research exists that would allow to render spatial audio of virtual humans. Analogous to visual full-body models, acoustic

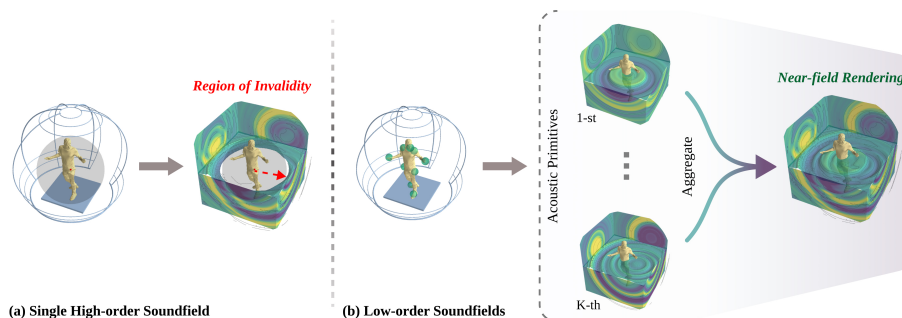


Fig. 1: Single high-order soundfield (a) vs. acoustic primitives (b). Existing approach [44] predicts a high-order ambisonic soundfield around the human body, preventing sound from being rendered in the near-field; our proposed acoustic primitives, represented as small spheres attached to the body, successfully model a complete and accurate 3D body soundfield.

full-body models have similar requirements: first, it must be possible to render spatial sounds produced by a virtual human at any position in 3D space, and second, the soundfield needs to be drivable. In this work, we focus on generating and driving full-body soundfields from 3D body pose and head-mounted microphones.

This problem has recently been addressed in pioneering work by Xu *et al.* [44], who developed a neural soundfield rendering system for full body avatars, driven by body pose and headset microphone input. However, [44] has several major limitations: The approach relies on a single high order ambisonic (spherical harmonics) representation that models the sound emitted from the surface of a sphere around the human body, with a diameter of about 2m. Sound can only be modeled outside of this sphere, such that near-field modeling of signals closer to the body is not possible, see Fig. 1a. Moreover, accurate sound reproduction in [44] relies on extremely high-order ambisonic coefficients which are expensive to compute and instable to estimate. To get around this instability, [44] does not predict the ambisonic coefficients directly, but instead predicts the raw audio signal on 345 positions surrounding the body, and then uses traditional signal processing to compute a 17-th order ambisonic representation from these 345 raw waveforms. This mechanism is computationally inefficient and prevents realtime sound rendering.

In this work, we propose a novel sound rendering method based on acoustic primitives which solves the problems of [44]:

Near-field Rendering. We take inspiration from recent methods in visual neural rendering that rely on *volumetric primitives* like cuboids [25] or Gaussians [20]. Instead of modeling the body soundfield by an ambisonic representation on a single sphere around the full human body as in [44], we attach multiple acoustic primitives (small spheres each representing low-order ambisonics) to the 3D human skeleton and model the sound radiating from each of these acoustic prim-

itives separately, see Fig. 1b. The full soundfield produced by all primitives together is given by the sum of the individual rendered sound from each primitive. This way, sound can be modeled arbitrarily close to the body.

Efficient Soundfield Representation. Instead of a single 17-th order ambisonic representation, we model body sounds by multiple low-order (typically second order) ambisonic primitives. This reduces the number of parameters characterizing the acoustic scene by an order of magnitude and allows for a more compact and efficient soundfield representation.

Efficient Rendering. Instead of predicting 345 raw audio signals and relying on traditional, costly high-order ambisonic encoders and decoders, we predict the low-order ambisonic coefficients of each primitive directly. Efficient sound rendering can then be achieved using spherical wave functions as described in Sec. 3.2.

Drivability. Same as [44], our method can be driven from body pose and a head mounted microphone, *i.e.* 3D soundfields can be generated for novel acoustic input and body motion. Note that this is in stark contrast to its visual counterparts [20, 26] which are designed to synthesize novel views of fixed scenes, but are typically not drivable from user input.

In summary, we propose an efficient and drivable 3D sound rendering system with

1. **audio-visual driving:** given body pose and an audio signal from head-mounted microphones, we can accurately render the soundfield produced by the body (speech, snapping, clapping, footsteps, etc) in 3D;
2. **real-time rendering:** the introduction of acoustic primitives allows for efficient real-time rendering of 3D sound scenes;
3. **high quality:** although relying only on low-order ambisonic representations, we achieve comparable quality to [44] but avoid the high computational cost.

2 Related Works

Spatial Audio Modeling. Existing works on spatial audio rendering are either based on traditional signal processing and linear filters [4, 6, 39] or on more recent neural binaural renderers [12, 33, 35]. While these approaches can typically produce spatial audio in an efficient way, they come at strong restrictions, particularly, they need to know the exact location of each sound source to render as well as the clean sound signal for each sound location. Such information is available in fully synthetic, artist-created scenes but is usually unknown in real environments and real acoustic scenes. Our approach, in contrast, does not rely on such knowledge and implicitly learns to separate an aggregated audio signal into its distinct sources (the acoustic primitives) through inverse acoustic rendering. More recently, data-driven methods aim to produce binaural audio from audio-visual input information. Chen *et al.* [5] propose a system to render a pre-recorded 3D acoustic scene from novel viewpoints, however, this method can not handle new acoustic scenes. Liang *et al.* [23] propose to reconstruct the 3D audio-visual scene from videos, but the scene is static. Gao *et al.* [10] analyze

a 2D image of a visual scene to generate binaural audio for a given monaural sound source. Note that this approach can only deal with single sound sources and can't handle complex 3D sound scenes. In [28], the authors propose a method to generate spatial audio from mono acoustic input and 360 degree camera inputs. They demonstrate that their approach can correctly localize sound sources in the scene and generate correct spatial audio at a coarse resolution.

Primitives in Volumetric Rendering. Neural rendering has been revolutionized by volumetric rendering methods like neural volumes [24] and neural radiance fields [26]. Follow-up works build on volumetric primitives such as cuboids [25] or Gaussians [20] and render via ray-marching or splatting. These technologies have unlocked real-time rendering for animatable avatars [7, 25, 32, 37, 49]. We borrow the idea of volumetric primitives and transfer them from the visual domain into the acoustic domain to build an efficient and low-parameter characterization of soundfields.

Audio-Visual Learning. Audio-visual learning has been widely applied to find connections between acoustic and 2D visual signals, *e.g.* in audio-visual localization [15, 17, 19, 27, 30, 41], for source-separation [8, 9, 11, 16, 48], or to learn associations from 360-degree videos [18, 22, 28]. Application of audio-visual learning to 3D settings is mostly limited to 3D visual scenarios, such as audio-visual driving of avatars [29, 34, 36, 43, 46] or audio-driven gesture synthesis [2, 13, 21, 47]. While these works use audio-visual input to learn information about a scene, they operate on visual outputs and do not model acoustic scenes.

Most closely related to our work is [44], who address the same task we address in this work. However, as outlined above, [44] has some significant drawbacks such as the inability to render near-field audio, or the lack of real-time rendering capabilities.

3 Pose-Guided Soundfield Generation using Acoustic Primitives

3.1 Problem Definition

Let $\mathbf{a}_{1:T_a}(a_1, \dots, a_{T_a})$ be the input audio signal from one or multiple head-mounted microphones, and $\mathbf{p}_{1:T_p} = (p_1, \dots, p_{T_p})$ be the corresponding sequence of 3D body pose, where each $p_t \in \mathbb{R}^{J \times 3}$ is a vector containing 3D body joint coordinates. We aim to predict a sound signal \mathbf{s} at an arbitrary 3D position (r, θ, φ) . Note that we use spherical coordinates. In other words, we learn to aim a mapping from 3D body pose, headset audio signal, and a 3D position in space to the audio signal at that 3D spatial position,

$$\mathbf{a}_{1:T_a}, \mathbf{p}_{1:T_p}, (r, \theta, \varphi) \mapsto \mathbf{s}_{1:T_a}. \quad (1)$$

The core challenge is how to get training data to learn such a model. It is impossible to place microphones at all positions in 3D space to get dense sampling of the space. Instead, we follow the strategy of [44] (and actually use the same public dataset for our work) and sample soundfield signals $\mathbf{s}_{1:T_a}$ only

on a sphere around the human body. This poses the challenge of rendering the soundfield at positions that are not on the surface of the sphere on which data has been captured. Note the analogy to graphics neural rendering: approaches like NeRF [26] also don't have spatially dense samples of 2D images taken from a scene, yet through inductive biases such as the rendering equation, they succeed in synthesizing the 3D scene from any novel viewpoint. We apply the same strategy for audio, and learn the mapping in Eq. (1) by differentiation through the wave propagation function, which we explain in the next section.

3.2 Sound Radiation using Spherical Wave Functions

The general solution of the homogeneous, time-dependent wave equation in the spherical coordinate system is given by [42, 50]:

$$\mathbf{w}(t, f, r, \theta, \varphi) = \sum_{n=0}^{\infty} \sum_{m=-n}^n [b_{nm}(t, f) \cdot j_n(kr) + c_{nm}(t, f) \cdot h_n(kr)] \cdot Y_{nm}(\theta, \varphi), \quad (2)$$

where (r, θ, φ) are arbitrary coordinates inside a **source-free region**, t and f denote the time and frequency (we will omit them in the following sections for clarity), and $k = 2\pi f/v_{sound}$ is the corresponding wavenumber; $Y_{nm}(\theta, \varphi)$ represents the spherical harmonic of order n and degree m , and $j_n(kr)$ and $h_n(kr)$ are, respectively, n th-order spherical Bessel and Hankel functions. Coefficients $b_{nm}(t, f)$ and $c_{nm}(t, f)$ describe, respectively, incoming and outgoing waves. In particular, considering the scenario depicted in Fig. 1a, only the radiating field component is present, *i.e.* $b_{nm} = 0$, which in literature is known as exterior domain problem [38, 42].

Given recorded or predicted microphone signals on the surface of the sphere surrounding the body, SoundingBodies [44] approaches the sound field modeling task as a traditional exterior domain problem and estimates the sound field coefficients $c_{nm}(t, f)$. While the general solution requires an infinite number of harmonic orders, the practical estimates are limited by the available number of microphones M as $N = \sqrt{M} - 1$,

$$\hat{\mathbf{w}}(r, \theta, \varphi) = \sum_{n=0}^N \sum_{m=-n}^n \hat{c}_{nm} \cdot h_n(kr) \cdot Y_{nm}(\theta, \varphi). \quad (3)$$

This reliance on a generic solution of the exterior domain problem limits the practicability of [44]. While the network uses pose information to predict microphone signals, the successive DSP processing of these signals required for spatial sound rendering (*e.g.* binauralization) does not leverage pose information at all. This brings two main issues:

1. To model sound sources located further away from the center of the representation, see R_0 in Fig. 1a, higher harmonic orders are needed, which in turn requires the network to predict a high number of microphones before any spatial rendering can be performed; predicting a smaller number of signals

would limit the harmonic order and, consequently, the rendered scene would collapse towards the center.

2. The wave equation solution is valid only outside the boundary surface encompassing all sources of sound, as shown in Fig. 1a. Inside this region, the high harmonic orders produce chaotic results, limiting the minimum distance at which the scene can be rendered and experienced.

To address the above issues we take a different approach. Instead of using pose-conditioned network to predict microphone signals, we use the network to predict sound field coefficients directly. Furthermore, instead of trying to estimate a generic high-order sound field representation, we leverage the knowledge of possible positions of sound, given by the body pose, and model the sound radiation as a superposition of several small-order elementary sound fields originating from different positions of the body as depicted in Fig. 1b. Similarly to Eq. (3), the sound pressure produced by a single elementary field of order N is given by

$$\begin{aligned} \mathbf{w}(r, \theta, \varphi) &= \sum_{n=0}^N \sum_{m=-n}^n (c_{nm} \cdot h_n(kr_{ref})) \cdot \frac{h_n(kr)}{h_n(kr_{ref})} \cdot Y_{nm}(\theta, \varphi) \\ &= \sum_{n=0}^N \sum_{m=-n}^n \tilde{c}_{nm} \cdot \frac{h_n(kr)}{h_n(kr_{ref})} \cdot Y_{nm}(\theta, \varphi), \end{aligned} \quad (4)$$

where we use $h_n(kr_{ref})$ with $r_{ref} = 0.5\text{m}$ for numerical stability of the learning process, and refer to harmonic coefficients $\mathcal{S} = [\tilde{c}_{00}, \dots, \tilde{c}_{NN}]$ as an **acoustic primitive**.

Given Eq. (4), we can translate the task of modeling 3D spatial sound for the visual body to learning a set of small acoustic primitives $\{\mathcal{S}_i\}_{i=1}^K$, which we choose N up to the second order for harmonic coefficients and set the number of acoustic primitives as K . In practice, capturing ground truth sound field coefficient is infeasible, while the microphone signals received on the surface of the dome are available with prior efforts by [44]. Since the produced sound pressure $\mathbf{w}(r, \theta, \varphi)$ indeed represents the audio signal produced at spherical position (r, θ, φ) , we can therefore decompose the entire learning process into two sub-steps:

- **Learning Acoustic Primitives.** The main objective of this step is to design a neural network \mathcal{F} that consumes audio and pose data as input, and output the sound field representation

$$\{\mathcal{S}_i\}_{i=1}^K = \mathcal{F}(\mathbf{a}_{1:T_a}, \mathbf{p}_{1:T_p}). \quad (5)$$

- **Rendering Audio with Learned Acoustic Primitives.** With the learned acoustic primitives $\{\mathcal{S}_i\}_{i=1}^K$, we leverage Eq. (4) as a differentiable rendering function, denoted as \mathcal{R} , to generate the audio waveform received at the target position

$$\hat{\mathbf{s}}_{1:T_a}(r, \theta, \varphi) = \mathcal{R}(\{\mathcal{S}_i\}_{i=1}^K, r, \theta, \varphi). \quad (6)$$

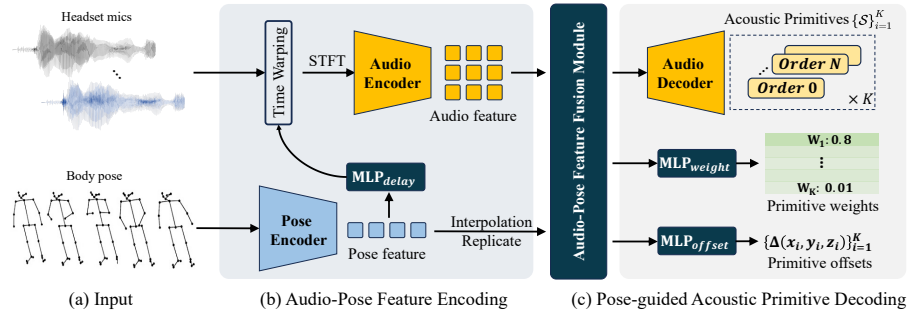


Fig. 2: Our pose-guided acoustic primitive learning framework takes headset microphone signals and body pose information as inputs. It outputs the acoustic primitive representations, weights, and offsets in one pass. The framework consists of two main stages. In the first stage, we employ separate encoders to process the audio and pose signals into feature spaces. An Audio-Pose Feature Fusion Module is then utilized to combine these features. In the second stage, the fused features are fed into an audio decoder network to generate the acoustic primitive coefficients. Additionally, two separate MLP heads are used to predict the weights and offsets for each acoustic primitive.

The hyperparameter in \mathcal{R} is fixed once the harmonic order N and the number of primitives K are initialized, and all the operations in \mathcal{R} are differentiable, making it feasible to run end-to-end training. With the training data tuple $(\mathbf{a}_{1:T_a}, \mathbf{p}_{1:T_p}, \mathbf{s}_{1:T_a}(r, \theta, \varphi))$ that includes recorded microphone signal at position (r, θ, φ) , we can learn a pose-guided acoustic primitive synthesis system by simply optimizing the loss between $\hat{\mathbf{s}}_{1:T_a}$ and $\mathbf{s}_{1:T_a}$.

3.3 Multimodal Feature Encoding

Pose Encoder. Human body movements offer crucial clues for how sound is distributed in space. To capture these rich spatial cues, we employ a pose encoder that processes the input pose sequence $\mathbf{p}_{1:T_p}$. The pose input is first encoded into a latent feature representation. To capture temporal relationships, we apply two layers of temporal convolutions with a kernel size of 5. Finally, we concatenate the encoded features for all joints and use an MLP to create a compact representation, denoted as $f_p \in \mathbb{R}^{C_p \times T'_p}$. Here C_p is the number of feature channels and T'_p represents the temporal dimension after convolution. More details are provided in the supplementary material.

Audio Encoder. While sound can originate from various points on the body (*e.g.*, hands, feet), it's captured by the headset microphone located at a central position near the head. This difference in location creates a slight time delay between the moment the sound is produced and when it's actually recorded. Previous research has shown that compensating for this time delay can be beneficial [35, 44]. In our approach, we leverage the pose features as guidance and use an MLP (as shown in Fig. 2) to estimate the delay for each acoustic primitive attached to a body joint, and time-warp the audio signal accordingly.

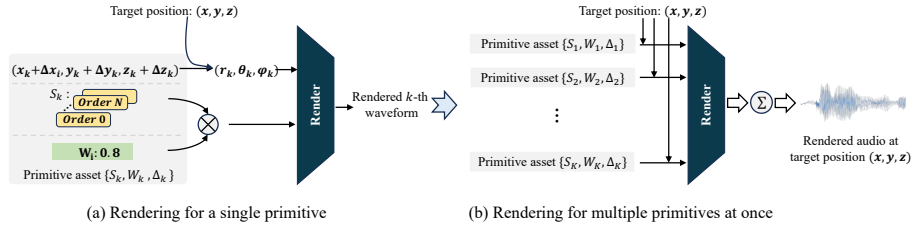


Fig. 3: Illustration on the rendering process with the estimated acoustic primitives. (a) demonstrates how to render a waveform signal given the learned harmonic coefficients S_k , primitive coordinate offset Δ_k , and the weight W_k . Next, we show that for all the primitives, we render audio generated by the primitive at the targeted location and aggregate them to yield the final rendered audio at target position (x, y, z) .

Via STFT, the warped signals are then transformed into complex spectrograms $X_a^c \in \mathbb{R}^{C_h \times F \times T}$ where F and T represent the number of frequency and time bins respectively, and C_h is the number of audio channels. The resulting audio features are then encoded with a network consisting of convolutional- and LSTM layers to capture both local context and long-range dependencies within the audio data. The encoder architecture utilizes four layers, where each layer contains two ResNet blocks, a temporal LSTM block, and a downsampling block with a factor of 2. Eventually, we can obtain the latent audio features $f_a = E_a(X_a) \in \mathbb{R}^{C_a \times \frac{F}{16} \times \frac{T}{16}}$.

Audio-Pose Feature Fusion Module. While headset audio reveals the content of sounds (*e.g.*, finger snapping), it lacks precise spatial information about the source. Conversely, body pose offers strong spatial cues about joint locations, but cannot identify the sound type (*e.g.*, speech) solely from pose data. Therefore, effectively combining audio and pose features is crucial for learning acoustic primitives and determining their contribution to the final sound generation. We first interpolate the pose features f_p to match the temporal size of the audio features f_a . We then employ a lightweight fusion module with two ResNet blocks and one attention block to combine the concatenated audio and pose features, resulting in a new representation denoted as $f_{ap} \in \mathbb{R}^{C_a \times \frac{F}{16} \times \frac{T}{16}}$.

3.4 Acoustic Primitive Decoding

As described in Sec. 3.2, an acoustic primitive determines the audio heard at arbitrary coordinates inside its sound field. It considers factors like the harmonic coefficients, the center coordinate of the primitive, and the target location where the sound is perceived. In this work, we focus on generating sound based on the target location. This translates to learning two key components: the primitive’s coordinates and the harmonic coefficients. However, this is non-trivial as the primitive’s location changes dynamically as the body moves. Additionally, the harmonic coefficients must capture not only the sound content but also the

spatial cues such as sound directivity. In the next section, we will explain how our approach addresses these challenges.

Sound Field Decoder. Leveraging the fused features f_{ap} , our decoder D_a simultaneously generates the sound field representations for all the acoustic primitives. Similar to the input spectrograms X_a , the harmonic coefficients have the same spatial dimensions but differ in the number of channels, which encode the richness of the sound’s spatial information. A higher number of channels allows for more precise control over the perceived location of the sound. For simplicity, we design the decoder to resemble the audio encoder and add skip connections between the encoder and decoder, with the main difference being the number of output channels. The decoder outputs $(N + 1)^2 \times K$ channels. Here, N represents the order of the harmonic coefficients, which controls the level of detail captured in the spatial representation. Finally, we separate the decoder’s output into K distinct harmonic coefficients $\{\mathcal{S}_i\}_{i=1}^K$, one for each acoustic primitive.

Primitive Offsets. We initialize acoustic primitives to be at body joint locations, *e.g.* at the wrists, face and ankles. While the initial 3D coordinates of these body joints provide a reasonable starting point, the actual locations of the sound sources might differ slightly from the body joint positions. For example, the chosen keypoints represent wrists, but the sound of finger snapping originates from the fingers themselves. This discrepancy between the body joint and sound production locations can affect the learning process and the accuracy of the rendered spatial audio. To address this limitation, we learn offsets for the initial coordinates to better represent the actual positions of acoustic primitives. In practice, we employ a three-layer MLP network that operates on the fused features f_{ap} . First, we apply mean pooling along the frequency axis of f_{ap} but keep its time dimension, obtaining \bar{f}_{ap} . Then, we generate the offsets by

$$\Delta(x, y, z) = \sigma \cdot \tanh(\text{MLP}_{\text{offset}}(\bar{f}_{ap})). \quad (7)$$

To constrain the predicted offsets within a reasonable range, we use a tanh activation function and apply a scaling factor of $\sigma = 0.2$ to restrict the offsets to a maximum range of 20 centimeters around the initial locations.

Primitive Weights. At different points in time, primitives have different importance. For instance, when a finger is snapped, the hand primitive emits high energy sound while other primitives emit at most low energy. The relationship is a function of the input audio and body pose. We therefore explicitly model the weight of each primitive as a function of the combined audio and pose encodings \bar{f}_{ap} ,

$$W = \text{softmax}(\text{MLP}_{\text{weight}}(\bar{f}_{ap})). \quad (8)$$

W is the predicted weight for each primitive at each time instance and indicates the relative influence in the final rendered sound.

3.5 Differentiable Acoustic Primitive Renderer

Given the initial primitive locations (*i.e.* the joint locations to which the primitives are attached), the learned offsets, harmonic coefficients, and weights, we

can now render the sound field for each primitive (as shown in Fig. 3). We first compute the primitive’s predicted location by adding the learned offsets to the corresponding body joint location. Now, given a listener position in 3D space at which we want to render the sound, we transform each primitive’s predicted location into spherical coordinates $(r_k, \theta_k, \varphi_k)$ representing the relative position of the listener with respect to each of the K primitives. We now use the differentiable audio renderer from Eq. (6) to render the audio signal $\hat{\mathbf{s}}_{1:T_a}^k$ produced by the k -th primitive at the listener’s position,

$$\hat{\mathbf{s}}_{1:T_a}^k = \mathcal{R}(\mathcal{S}_k \cdot W_k, r_k, \theta_k, \varphi_k), \quad (9)$$

and obtain the full sound field by summation over all acoustic primitives,

$$\hat{\mathbf{s}}_{1:T} = \sum_k^K \hat{\mathbf{s}}_{1:T}^k. \quad (10)$$

3.6 Loss Function

Since our renderer \mathcal{R} is differentiable, it allows us to efficiently train the model using loss functions on the final predicted waveforms. In this work, we employ a multiscale STFT loss [45] between the predicted audio $\hat{\mathbf{s}}_{1:T_a}$ and the ground truth audio $\mathbf{s}_{1:T_a}$ on their amplitude spectrograms, denoted as $\mathcal{L}_{amp}(\hat{\mathbf{s}}_{1:T_a}, \mathbf{s}_{1:T_a})$ and on the real and imaginary parts of spectrograms, denoted as $\mathcal{L}_{ri}(\hat{\mathbf{s}}_{1:T_a}, \mathbf{s}_{1:T_a})$. The window sizes are set as 2048, 1024, 512, 256. As proposed in [44], a shift- $\ell 1$ loss helps reduce the spatial alignment error. We therefore add this loss term as $\mathcal{L}_{sl1}(\hat{\mathbf{s}}_{1:T_a}, \mathbf{s}_{1:T_a})$. Additionally, determining a primitive’s contribution to the final sound (corresponding to the primitive weights W) can be challenging without additional guidance. To overcome this, we leverage clip-level labels (denoted by $y \in \mathbb{R}^K$) that specify which body joint contributes to the received audio. We apply average pooling along the frequency dimension and find the maximum value of W for each primitive across all time steps, resulting in $\tilde{W} \in \mathbb{R}^K$. This essentially summarizes whether an acoustic primitive has contributed to the sound in the audio clip. Finally, a simple cross-entropy loss function $\mathcal{L}_{cts}(\tilde{W}, y)$ is employed to aid in the learning process. Our final loss becomes

$$\mathcal{L}_{total} = \lambda_{amp} \mathcal{L}_{amp} + \lambda_{ri} \mathcal{L}_{ri} + \lambda_{sl1} \mathcal{L}_{sl1} + \lambda_{cts} \mathcal{L}_{cts}. \quad (11)$$

Please refer to the supplementary materials for ablation on the loss terms.

4 Experiments

4.1 Experimental Setting

Dataset. To evaluate our approach, we leverage the publicly available dataset introduced in [44]³. The dataset captures synchronized audio and visual data in

³ <https://github.com/facebookresearch/SoundingBodies>

⁴ Note: data used in the paper and data released publicly differ by 1.5 subjects (8 subjects used in [44] vs 6.5 publicly released). We updated the performance metrics of the baseline [44] to account for this difference.

Table 1: Our method achieves **comparable results** to the baseline SoundingBodies [44] in SDR, amplitude error, and phase error, while significantly outperforming the baseline in inference speed (**15x faster**).

Methods	Speed	non-speech			speech		
		SDR \uparrow	amplitude \downarrow	phase \downarrow	SDR \uparrow	amplitude \downarrow	phase \downarrow
[44]	3.56s	3.052	0.832	0.314	9.635	0.701	0.464
Ours	0.24s	3.597	0.883	0.323	8.448	0.943	0.417

an anechoic chamber, offering multimodal data specifically designed for speech and body sound field modeling research. It utilizes 5 Kinect sensors for body tracking and a large microphone array (345 microphones) arranged in a spherical fashion around the recording area. The data encompasses various participants performing a diverse range of body sounds and speech in different settings (*e.g.*, standing or sitting). The recordings are segmented into non-overlapping one-second clips. We adopt the same train/validation/test splits established by [44], resulting in 10,076/1,469/1,431 clips, respectively.

Implementation Details. In our experimental setup, we employ a sampling rate of 48 kHz for audio signals and a frame rate of 30 fps for body pose data. The audio waveforms are converted into complex spectrograms using a Hann window of size 512 and a hop length of 128 and FFT length of 1022. Within the encoders, both the pose features f_p and audio features f_a are configured to have the same channel size $C_a = C_p = 256$. We set the order of harmonic coefficients to $N = 2$. During training, the batch size is set as 1 per GPU and we randomly select 20 microphones from the available pool of 345 target microphones for each forward pass. The AdamW optimizer with a learning rate of 0.0002 is used, and the network is trained for 100 epochs. To balance different loss terms, we set the weights $\lambda_{amp} = 7$, $\lambda_{ri} = 3$, $\lambda_{s\ell 1} = 0.5$, and $\lambda_{cts} = 1$. The experiments are conducted on 4 NVIDIA Tesla A100 GPUs, with model training for 100 epochs taking approximately 55 hours to complete.

Evaluation Metrics. We evaluate the performance of our model using three main metrics: the signal-to-distortion ratio (SDR), the ℓ_2 error on the amplitude spectrogram, and the angular error of the phase spectrogram. The SDR measures the overall quality of the reconstructed sound, with higher values indicating better quality. The amplitude error shows how well the reconstructed sound matches the original in terms of the distribution of sound energy, while the angular error evaluates the timing accuracy of the reconstructed sound waves relative to the original. We report amplitude errors multiplied by a factor of 1000 to remove leading zeros.

4.2 Comparison with Baseline

We compare our method using 12 acoustic primitives of 2nd order with the SoundingBodies [44] baseline. Results are shown in Tab. 1. We can observe that

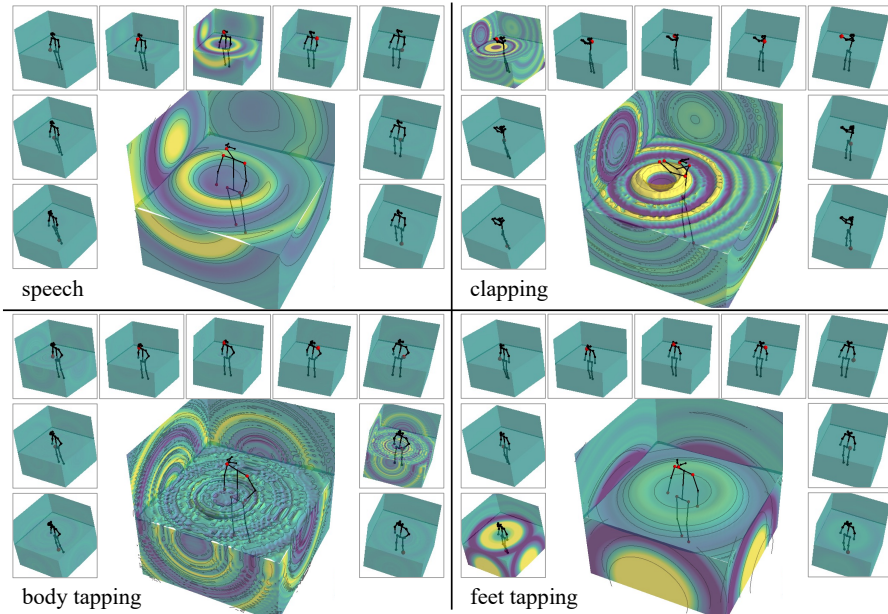


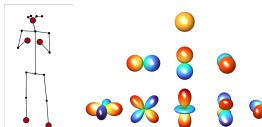
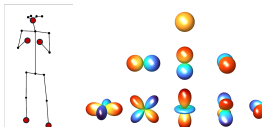
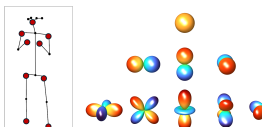
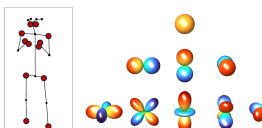
Fig. 4: Sound field visualizations for 4 different kinds of sound. Main sound field is in the center and individual primitive contributions are shown around. We can observe that the method assigns acoustic energy to correct acoustic primitives, e.g. speech comes mostly from the head with only a very small contribution from the shoulder primitives. We can also observe the speech directivity pattern matching the head orientation. For each visualization, the left/right 4 primitives are labeled as follows: *foot*, *hip*, *hand*, and *shoulder* (from bottom to top), and the middle one is the head.

the sound field modeling performance of the proposed method is comparable to [44] while having a much faster inference speed. In particular, proposed method even performs better than the baseline on SDR metric for non-speech sounds and phase metric for speech. Regarding the inference speed, we show average time needed to compute 1 second of audio at 48 kHz. Note that for [44] we only report the time needed for the network to predict the microphone signals. In a practical scenario [44] needs also DSP processing of these microphone signals to obtain the high-order sound field representation, which would further increase the overall processing overhead.

4.3 Ablation Study

In this section we evaluate the impact of the number of acoustic primitives and their harmonic order. Zero-order harmonics are able to model only omnidirectional fields, while higher orders allow for increasingly complex radiation patterns. Note that we limit the maximum order to 2 given that PyTorch implementations of spherical wave functions are available only up to the second order. Intuitively,

Table 2: Ablation study: the number and harmonic order of acoustic primitives.

# K		# N	non-speech			speech		
			SDR \uparrow	amp. \downarrow	phase \downarrow	SDR \uparrow	amp. \downarrow	phase \downarrow
5		0th	2.009	1.013	0.334	4.775	1.225	0.559
		1st	3.534	0.896	0.322	7.261	0.983	0.480
		2nd	3.552	0.911	0.323	7.981	0.977	0.442
9		0th	3.059	0.907	0.327	6.619	1.068	0.496
		1st	3.600	0.895	0.323	7.479	0.983	0.467
		2nd	3.569	0.915	0.323	8.200	0.952	0.434
12		0th	3.020	0.895	0.327	6.893	1.031	0.472
		1st	3.616	0.879	0.325	7.616	0.969	0.466
		2nd	3.597	0.883	0.323	8.448	0.943	0.417
12	no primitive offset	2nd	3.528	0.919	0.321	7.730	0.998	0.456

a higher number of acoustic primitives allows for modeling of more complex overall sound fields. We test three configurations of acoustic primitives: 5 primitives: head, L/R hand, L/R foot; 9 primitives: head, L/R hand, L/R foot, L/R shoulder, L/R hip; and 12 primitives where head and hands have two primitives associated with the same key-point (given location offsets these primitives are not bound to be in the same location allowing approximation of a higher order radiation pattern). Results are shown in Tab. 2. In general, both higher number of primitives and higher primitive order improve the performance as expected. This is especially true for speech. For body sounds on the other hand, increasing from 1st to 2nd order does not seem to be beneficial. We also evaluate the model without the primitive offset adjustment. From Tab. 2 we can observe that removing the offset has similar impact as decreasing the number of primitives from 12 to 9, which intuitively makes sense given that repeated primitives collapse to the same key-point location.

For more experiments, such as ablation on the loss terms and visualizations with different harmonic orders, please refer to the supplementary materials.

4.4 Qualitative Results

Some sound field visualization examples are shown in Fig. 4. We can observe that the network is able to correctly associate different kinds of sounds to the appropriate acoustic primitives. We can also observe the speech radiation pattern matching the head orientation. Furthermore, Fig. 5 shows predicted and

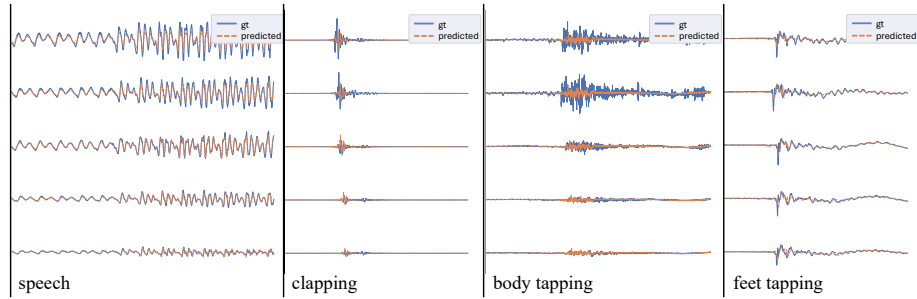


Fig. 5: Predicted and ground truth microphone signals at 5 different locations around the dome. We can observe good temporal alignment and good amplitude match except for the low-energy body tapping sound. We recommend zooming in for better visibility.

ground truth waveforms at different microphone locations. We can observe good temporal alignment and amplitude match for most cases. One exception is the body tapping sound in which the amplitude does not match across different microphones. This may be due to the primitive struggling to match a highly variable radiation pattern.

5 Conclusion

We propose a neural rendering system for sound that allows to generate and render 3D sound fields from sparse user input like body pose and headset audio. We demonstrate that we maintain similar quality to state-of-the-art sound rendering, while improving significantly on speed and soundfield completeness: our approach is an order of magnitude faster than the approach from [44] and is capable of rendering sound in the near-field, *i.e.* close to the transmitter’s body, where the previous approach from [44] failed.

Moreover, we want to highlight the design similarities to successful neural renderers from computer graphics: By leveraging an acoustic rendering equation and acoustic primitives, similar to leveraging volumetric primitives in graphical neural rendering, we design a 3D spatial audio system with a conceptual duality to its visual counterpart. We hope this work will impact sound rendering in 3D settings like computer games and AR/VR.

Limitations. Albeit the promising results in terms of quality and efficiency, our approach is still far from broad availability: model training relies on data collected with a multi-microphone capture stage that is not broadly available. Future directions need to aim at enabling learning such acoustic scenes with simpler setups, ideally with commodity hardware like smartphones. Generalization beyond human bodies is another natural extension that emerges from the availability of broader data sources for spatial sound.

Potential Society Impact. Ethical and societal risks in this work are low since no data is manipulated in a generative fashion - pure spatialization has little potential for harmful actors.

References

1. Metahuman creator. <https://metahuman.unrealengine.com> (2021)
2. Ahuja, C., Lee, D.W., Nakano, Y.I., Morency, L.P.: Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. pp. 248–265. Springer (2020)
3. Bagautdinov, T., Wu, C., Simon, T., Prada, F., Shiratori, T., Wei, S.E., Xu, W., Sheikh, Y., Saragih, J.: Driving-signal aware full-body avatars. *ACM Transactions on Graphics (TOG)* **40**(4), 1–17 (2021)
4. Chen, C., Jain, U., Schissler, C., Gari, S.V.A., Al-Halah, Z., Ithapu, V.K., Robinson, P., Grauman, K.: Soundspaces: Audio-visual navigation in 3d environments. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. pp. 17–36. Springer (2020)
5. Chen, C., Richard, A., Shapovalov, R., Ithapu, V.K., Neverova, N., Grauman, K., Vedaldi, A.: Novel-view acoustic synthesis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023)
6. Chen, C., Schissler, C., Garg, S., Kobernik, P., Clegg, A., Calamia, P., Batra, D., Robinson, P., Grauman, K.: Soundspaces 2.0: A simulation platform for visual-acoustic learning. *Advances in Neural Information Processing Systems* **35**, 8896–8911 (2022)
7. Chen, Z., Hong, F., Mei, H., Wang, G., Yang, L., Liu, Z.: Primdiffusion: Volumetric primitives diffusion for 3d human generation. In: *Thirty-seventh Conference on Neural Information Processing Systems* (2023)
8. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics (TOG)* **37**(4), 1–11 (2018)
9. Gao, R., Feris, R., Grauman, K.: Learning to separate object sounds by watching unlabeled video. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 35–53 (2018)
10. Gao, R., Grauman, K.: 2.5 d visual sound. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 324–333 (2019)
11. Gao, R., Grauman, K.: Visualvoice: Audio-visual speech separation with cross-modal consistency. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 15490–15500. IEEE (2021)
12. Gebru, I.D., Marković, D., Richard, A., Krenn, S., Butler, G.A., De la Torre, F., Sheikh, Y.: Implicit hrtf modeling using temporal convolutional networks. In: *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 3385–3389. IEEE (2021)
13. Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J.: Learning individual styles of conversational gesture. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3497–3506 (2019)
14. Hendrix, C., Barfield, W.: The sense of presence within auditory virtual environments. *Presence: Teleoperators and Virtual Environments* **5**(3), 290–301 (1996)
15. Hu, D., Qian, R., Jiang, M., Tan, X., Wen, S., Ding, E., Lin, W., Dou, D.: Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems* **33**, 10077–10087 (2020)
16. Huang, C., Liang, S., Tian, Y., Kumar, A., Xu, C.: Davis: High-quality audio-visual separation with generative diffusion models. *arXiv preprint arXiv:2308.00122* (2023)

17. Huang, C., Tian, Y., Kumar, A., Xu, C.: Egocentric audio-visual object localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22910–22921 (June 2023)
18. Huang, H., Solah, M., Li, D., Yu, L.F.: Audible panorama: Automatic spatial audio generation for panorama imagery. In: Proceedings of the 2019 CHI conference on human factors in computing systems. pp. 1–11 (2019)
19. Jiang, H., Murdock, C., Ithapu, V.K.: Egocentric deep multi-channel audio-visual active speaker localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10544–10552 (2022)
20. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics* **42**(4) (July 2023)
21. Lee, H.Y., Yang, X., Liu, M.Y., Wang, T.C., Lu, Y.D., Yang, M.H., Kautz, J.: Dancing to music. *Advances in neural information processing systems* **32** (2019)
22. Li, D., Langlois, T.R., Zheng, C.: Scene-aware audio for 360 videos. *ACM Transactions on Graphics (TOG)* **37**(4), 1–12 (2018)
23. Liang, S., Huang, C., Tian, Y., Kumar, A., Xu, C.: Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. In: Conference on Neural Information Processing Systems (NeurIPS) (2023)
24. Lombardi, S., Simon, T., Saragih, J., Schwartz, G., Lehrmann, A., Sheikh, Y.: Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.* **38**(4), 65:1–65:14 (Jul 2019)
25. Lombardi, S., Simon, T., Schwartz, G., Zollhoefer, M., Sheikh, Y., Saragih, J.: Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (ToG)* **40**(4), 1–13 (2021)
26. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 405–421. Springer (2020)
27. Mo, S., Morgado, P.: Localizing visual sounds the easy way. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII. pp. 218–234. Springer (2022)
28. Morgado, P., Nvasconcelos, N., Langlois, T., Wang, O.: Self-supervised generation of spatial audio for 360 video. *Advances in neural information processing systems* **31** (2018)
29. Ng, E., Romero, J., Bagautdinov, T., Bai, S., Darrell, T., Kanazawa, A., Richard, A.: From audio to photoreal embodiment: Synthesizing humans in conversations. In: IEEE Conference on Computer Vision and Pattern Recognition (2024)
30. Qian, R., Hu, D., Dinkel, H., Wu, M., Xu, N., Lin, W.: Multiple sound sources localization from coarse to fine. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. pp. 292–308. Springer (2020)
31. Qian, S., Kirschstein, T., Schoneveld, L., Davoli, D., Giebenhain, S., Nießner, M.: Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians (2023)
32. Remelli, E., Bagautdinov, T., Saito, S., Wu, C., Simon, T., Wei, S.E., Guo, K., Cao, Z., Prada, F., Saragih, J., et al.: Drivable volumetric avatars using texel-aligned features. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–9 (2022)
33. Richard, A., Dodds, P., Ithapu, V.K.: Deep impulse responses: Estimating and parameterizing filters with deep networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (2022)

34. Richard, A., Lea, C., Ma, S., Gall, J., de la Torre, F., Sheikh, Y.: Audio- and gaze-driven facial animation of codec avatars. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 41–50 (2021)
35. Richard, A., Markovic, D., Gebu, I.D., Krenn, S., Butler, G.A., Torre, F., Sheikh, Y.: Neural synthesis of binaural speech from mono audio. In: International Conference on Learning Representations (2021)
36. Richard, A., Zollhoefer, M., Wen, Y., de la Torre, F., Sheikh, Y.: Meshtalk: 3d face animation from speech using cross-modality disentanglement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
37. Saito, S., Schwartz, G., Simon, T., Li, J., Nam, G.: Relightable gaussian codec avatars (2023)
38. Samarasinghe, P.N., Abhayapala, T.D.: 3D spatial soundfield recording over large regions. In: Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC) (2012)
39. Savioja, L., Huopaniemi, J., Lokki, T., Väänänen, R.: Creating interactive virtual acoustic environments. *Journal of the Audio Engineering Society* **47**(9), 675–705 (1999)
40. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-or, D., Bermano, A.H.: Human motion diffusion model. In: The Eleventh International Conference on Learning Representations (2023)
41. Tian, Y., Hu, D., Xu, C.: Cyclic co-learning of sounding object visual grounding and sound separation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2745–2754 (2021)
42. Williams, E.G.: *Fourier Acoustics*. Academic Press (1999)
43. Xing, J., Xia, M., Zhang, Y., Cun, X., Wang, J., Wong, T.T.: Codetalker: Speech-driven 3d facial animation with discrete motion prior. arXiv preprint arXiv:2301.02379 (2023)
44. Xudong, X., Markovic, D., Sandakly, J., Keebler, T., Krenn, S., Richard, A.: Sounding bodies: Modeling 3d spatial sound of humans using body pose and audio. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
45. Yamamoto, R., Song, E., Hwang, M.J., Kim, J.M.: Parallel waveform synthesis based on generative adversarial networks with voicing-aware conditional discriminators. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6039–6043. IEEE (2021)
46. Yi, H., Liang, H., Liu, Y., Cao, Q., Wen, Y., Bolkart, T., Tao, D., Black, M.J.: Generating holistic 3d human motion from speech. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 469–480 (June 2023)
47. Yoon, Y., Cha, B., Lee, J.H., Jang, M., Lee, J., Kim, J., Lee, G.: Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)* **39**(6), 1–16 (2020)
48. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: Proceedings of the European conference on computer vision (ECCV). pp. 570–586 (2018)
49. Zielonka, W., Bagautdinov, T., Saito, S., Zollhöfer, M., Thies, J., Romero, J.: Drivable 3d gaussian avatars (2023)
50. Zotter, F., Frank, M.: *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Springer Topics in Signal Processing, Springer International Publishing (2019)