Supplemental Material for Sounding Bodies: Modeling 3D Spatial Sound of Humans Using Body Pose and Audio

Xudong Xu[*]
Shanghai AI Laboratory
xudongxu9710@gmail.com

Dejan Marković
Meta Reality Labs Research
dejanmarkovic@meta.com

Jacob Sandakly
Meta Reality Labs Research
jasandakly@meta.com

Todd Keebler
Meta Reality Labs Research
toddkeebler@meta.com

Steven Krenn
Meta Reality Labs Research
stevenkrenn@meta.com

Alexander Richard
Meta Reality Labs Research
richardalex@meta.com

A Time Warping

As described in Section 4.2, we employ time warping to tackle the time delay caused by sound propagation. While this delay in principle can be learned by the network, it has been shown that neural spatial audio modeling benefits from explicit time warping (*i.e.*, modeling the delay geometrically by shifting the input signals), as without it the model capacity is spent on learning the delay, causing slow convergence and less accurate results Richard et al. [2021]. While in tasks with a single and known sound source location, geometric time warping is trivial, in our case the origin of the sound is unknown, in fact it is even possible that there are multiple sound sources like left hand and right hand snapping simultaneously. To overcome the issue, we identify six body joints (hands, feet, nose, and hip) that are the most frequent sound emitters, and warp the sound from each of these positions towards the target microphones.

The pipeline of time warping is illustrated in Figure 1. For simplicity, we describe the warping process for a single input microphone (1-channel input). In practice, we apply this process to each of the seven input microphones and concatenate all resulting warped audio signals along the channel dimension.

Similar to Richard et al. [2021], we aim to estimate a geometric warpfield $\rho_{1:T}$ based on the speed of sound v_{sound} and the propagation distance difference between sound sources, input, and target microphones. We consider the aforementioned six body joints to be the potential sound sources $\mathbf{p}_{1:S}^{(\text{src})}$ and simultaneously refer to the nose keypoint as the position of any input microphone $\mathbf{p}_{1:S}^{(\text{in})}$. Besides, the target microphone position (θ, ϕ) will be mapped into the Cartesian space $\mathbf{x} = (x, y, z)$. For a specific time stamp $t \in \{1, ..., T\}$, we first identify the corresponding visual frame $s = \lfloor t/1600 \rfloor$. Accordingly, the time delay can be described as

$$\Delta t = (d_2 - d_1)/v_{\text{sound}} \quad \text{with} \quad d_1 = \left\|\mathbf{x} - \mathbf{p}_s^{(\text{src})}\right\|_2, d_2 = \left\|\mathbf{p}_s^{(\text{in})} - \mathbf{p}_s^{(\text{src})}\right\|_2. \tag{1}$$

In other words, we compute the distance d_1 between the potential sound source and target microphone and the distance d_2 between the potential sound source and input microphone. The time delay to

[*]Work done during internship at Meta Reality Labs Research, Pittsburgh, PA, USA.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

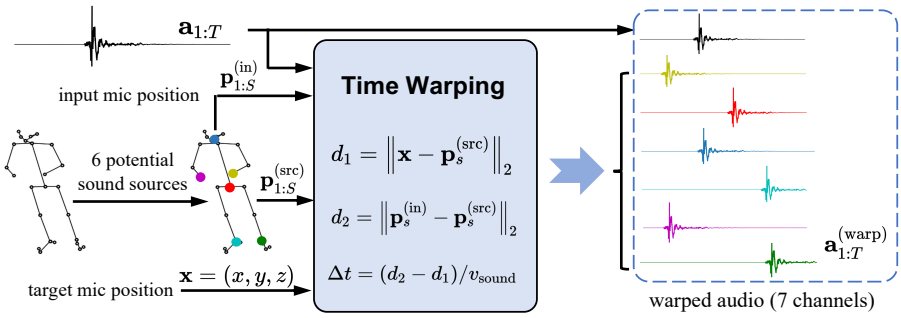

Figure 1: Overview of time warping. We manually select 6 keypoints as the locations of potential sound sources $\mathbf{p}^{(\text{src})}_{1:S}$ and regard the nose keypoint as the input microphone position $\mathbf{p}^{(\text{in})}_{1:S}$. For a specific audio sample a_t, the corresponding visual frame s is first located, and the time warping is conducted according to the sound arriving time difference Δt between input and target microphones. Keeping the original input audio additionally to the warped signals, we obtain *a set of audio* $\mathbf{a}^{(\text{warp})}_{1:T}$ containing seven channels (one original plus six-time warps) for each input microphone.

compensate for is determined by the distance between source and target minus the distance between source and input microphone.

Considering the audio is sampled at 48kHz, we represent the warpfield as

$$\rho_t = \max(\rho_{t-1}, t - \Delta t \cdot 48000), \tag{2}$$

where monotonicity is ensured by $\max(\rho_{t-1}, \cdot)$ following Richard et al. [2021]. It's noteworthy that the causality property introduced in Richard et al. [2021] doesn't hold here since the target microphone on the spherical array may receive the emitted sound before the head-mounted microphones. For example, a participant might stretch out their hand and make a snapping sound near the target microphone, but an arm-length away from the input microphone. With the predicted warpfield ρ_t, the warped signal can be computed following Richard et al. [2021] as

$$a^{(\text{warp})}_t = (\lceil \rho_t \rceil - \rho_t) \cdot a_{\lfloor \rho_t \rfloor} + (\rho_t - \lfloor \rho_t \rfloor) \cdot a_{\lceil \rho_t \rceil}. \tag{3}$$

The time warping process is implemented on six selected sound sources, thus leading to six channels of warped audio. Additionally, we empirically find the original input audio $\mathbf{a}_{1:T}$ can still facilitate the training and therefore keep it as an additional channel of the warped audio. Finally, the original input audio $\mathbf{a}_{1:T}$ of a single input microphone is warped to 7-channel audio $\mathbf{a}^{(\text{warp})}_{1:T}$ after time warping. We apply this time warping procedure to all input microphones and concatenate the results along the channel dimension.

B Sound Field Encoding

In Section 4.4 we described how to decode the harmonic sound field coefficients into time domain signals at arbitrary locations in space. In this section we address the inverse problem: how to obtain the harmonic coefficients using signals captured by the N microphones arranged on a sphere around the center of the capture stage. We note that, due to spatial aliasing, the maximum harmonic order K of this encoding is limited by the number of microphones to $K = \lfloor \sqrt{N} \rfloor - 1$.

The signal $s_i(t)$ captured by the i-th microphone located at (r_i, θ_i, ϕ_i) can be expressed in terms of sound field coefficients $\beta_{nm}(\tau, f)$ as

$$\text{STFT}\left(s_i(t)\right) = \sum_{n=0}^{K} \sum_{m=-n}^{n} \beta_{nm}(\tau, f) h_n(kr_i) Y_{nm}(\theta_i, \phi_i) \ , \tag{4}$$

where STFT denotes the short-time Fourier transform, τ and f are, respectively, time and frequency bins, $k = 2\pi f / v_{\text{sound}}$ is the wave number, v_{sound} is the speed of sound; $Y_{nm}(\theta, \phi)$ represents the spherical harmonic of order n and degree m (angular spatial component), and $h_n(kr)$ is nth-order spherical Hankel function (radial spatial component).

Given the STFT of all N microphone signals $\mathbf{S}(\tau, f) = \text{STFT}\left([s_1(t), \ldots, s_N(t)]\right)^T$, we can write Equation (4) as a linear system

$$\begin{aligned}\mathbf{S}(\tau, f) &= \begin{bmatrix} h_0(kr_1)Y_{00}(\theta_1, \phi_1) & \ldots & h_K(kr_1)Y_{KK}(\theta_1, \phi_1) \\ \vdots & \ddots & \vdots \\ h_0(kr_N)Y_{00}(\theta_N, \phi_N) & \ldots & h_K(kr_N)Y_{KK}(\theta_N, \phi_N) \end{bmatrix} \begin{bmatrix} \beta_{00}(\tau, f) \\ \vdots \\ \beta_{KK}(\tau, f) \end{bmatrix} \\ &= \mathbf{T}(f)\boldsymbol{\beta}(\tau, f)\ ,\end{aligned}$$

where $\boldsymbol{\beta}(\tau, f)$ is a vector containing $(K+1)^2$ sound field coefficients, and $\mathbf{T}(f)$ is a $N \times (K+1)^2$ matrix that links the coefficients to observations $\mathbf{S}(\tau, f)$. The sound field coefficients can than be obtained as

$$\boldsymbol{\beta}(\tau, f) = \mathbf{T}(f)^\dagger \mathbf{S}(\tau, f)\ , \tag{5}$$

with $(\cdot)^\dagger$ denoting the pseudo-inverse. For more details about spherical harmonic representation of sound fields and 3D recording of large areas see Williams [1999], Samarasinghe and Abhayapala [2012].

C Audio-Visual Time Synchronization

The audio system is a distributed array of MADI-based analog-to-digital converters. Each analog-to-digital converter is phase-locked to each other via the same WordClock signal. The video system is a distributed array of Kinect Azure cameras. Each Kinect is set to "subordinate mode" where the unit will only take an image when they get a digital high signal from the sync-in port. The Kinect array is in a "star" pattern where one master unit sends a digital high signal out to a distribution amplifier. The master Kinect digital high signal is then propagated to all of the subordinate Kinects. We also record the digital high signal into our analog-to-digital audio converters so we have correspondence between a visual frame and an audio sample.

D Additional examples

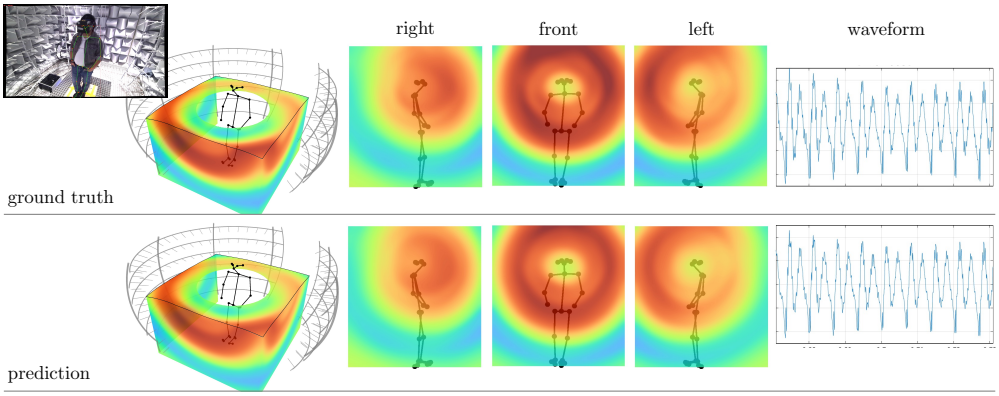

Figure 2: Example of speech directivity.

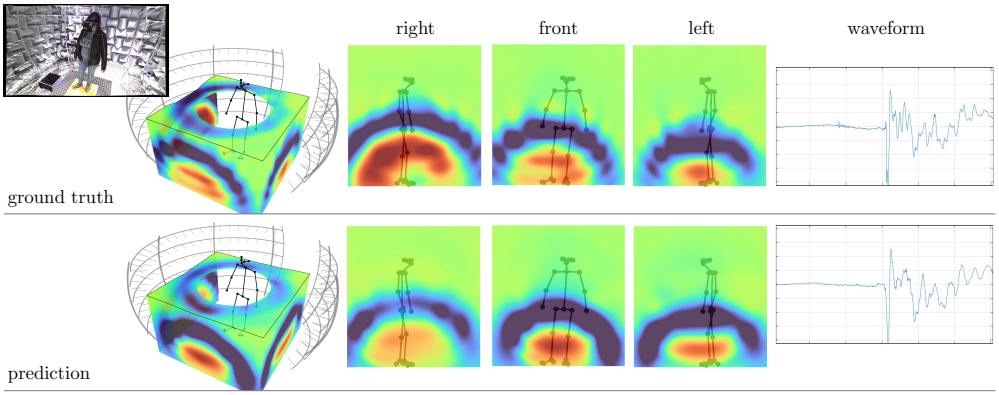

Figure 3: Footstep example.

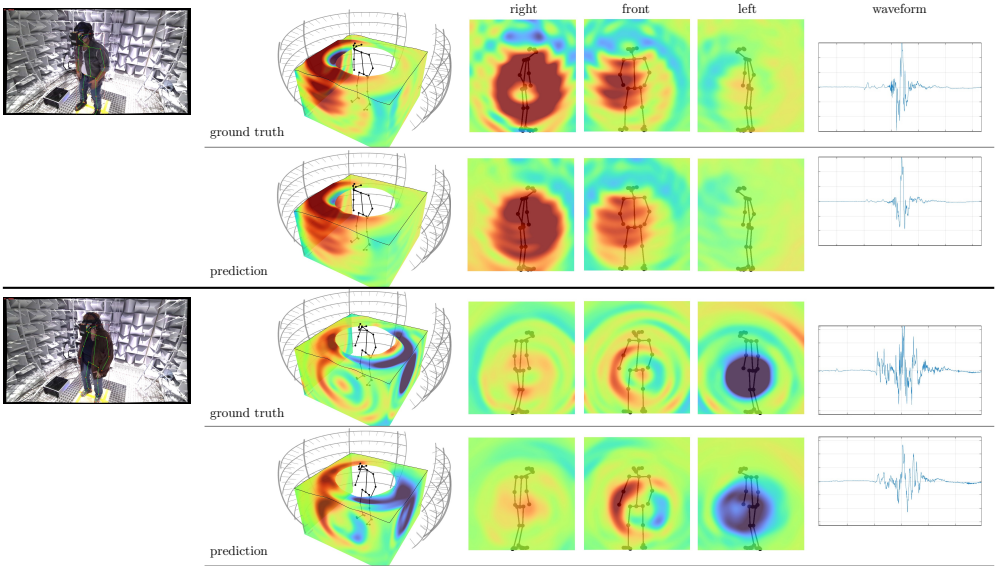

Figure 4: Thigh taps.

E Failure case examples

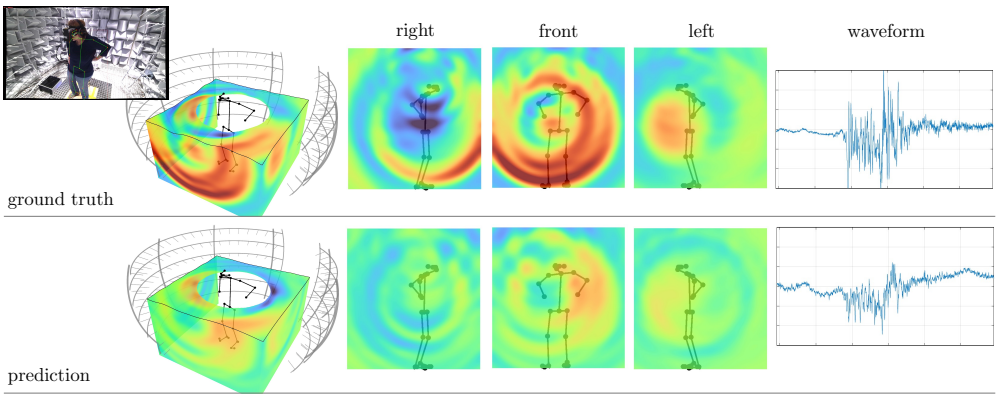

Figure 5: Not always predicted waveforms match the ground truth well. In this example, while sound source is reasonably localized, the signal energy is underestimated.

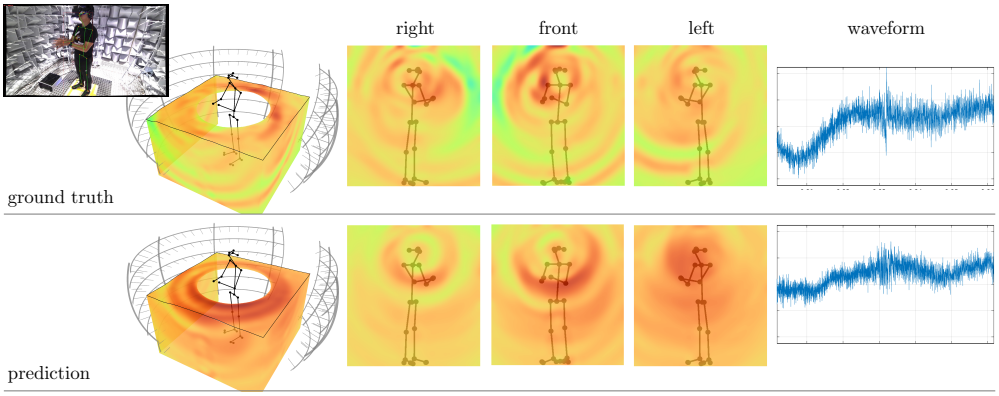

Figure 6: Hands rubbing example. Given noise-like nature and a very low energy of the signal, the model struggles to learn to correctly spatialize it.

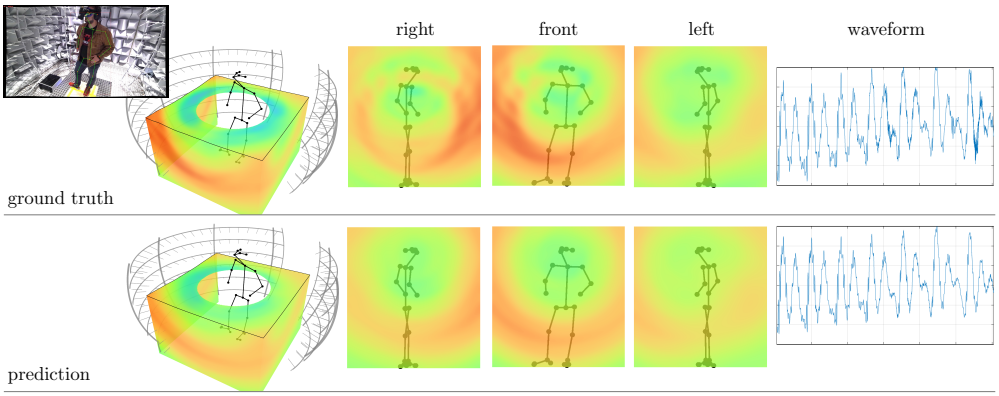

Figure 7: In this example a body tapping sound occurs at the same time as speech. Given disparity between signal energies, the model spatializes speech well but not the taping sound.

Table 1: **Capture Script.** Each participant performed the following activities three times: once standing with light clothing (T-Shirt, light sweater, ...), once standing with heavy clothing that produces more noise (jackets, coats, ...), and once with light clothing while sitting.

activity	description	duration
	speech	
conversation_speech	conversational speech	300s
rainbow_L	reading rainbow passage w/ left hand covering mouth	150s
rainbow_R	reading rainbow passage w/ right hand covering mouth	150s
	non-speech	
head_scratch	scratch head w/ one hand at a time	15s
face_rub	play and rub face w/ one hand at a time	15s
neck_scratch	scratching neck w/ one hand at a time	15s
back_scratch	scratching back w/ one hand at a time	15s
arms_stretching	arms/fingers stretching	15s
hands_rubbing	rubbing hands (*e.g.*, washing hands)	15s
hands_arms_swipe	swipe hands and arms	15s
hands_scratch	scratching arms and hands w/ one hand at a time	15s
hands_punch	punch other hand w/ one hand at a time	15s
fist_bump	fist bump	15s
applause	applause, *i.e.* clap hands in front of body	10s
clapping	clapping, covering as many spatial positions as possible	20s
snapping_L	snapping w/ left hand, covering as many spatial positions as possible	20s
snapping_R	snapping w/ right hand, covering as many spatial positions as possible	20s
snapping_LR	snapping w/ both hands, covering as many spatial positions as possible	15s
trunk_twist	twisting the trunk	10s
body_stretching	stretch body and yawn	10s
chest_tap	tapping chest w/ one hand at a time	20s
belly_rub	rubbing belly w/ one hand at a time	15s
belly_tap	tapping belly w/ one hand at a time	15s
thighs_rub	rubbing/scratching thighs w/ one hand at a time	15s
thighs_tap	tap thighs w/ one hand at a time	20s
thighs_tap_LR	tap thighs w/ both hands simultaneously	15s
knees_rub	rubbing/scratching behind knees w/ one hand at a time	15s
body_itching	itching and scratching the whole body	15s
body_tapping	tapping whole body (searching for phone, keys, wallet, ...)	15s
feet_tap	tap feet, one leg at a time	15s
walk	walk in place	15s
	mixed	
conversation_body	conversational speech w/ body sounds and gestures	300s

F Data Capture

Eight participants perform the script in Table 1 in three different settings: once standing with light clothing (T-Shirt, light sweater, ...), once standing with heavy clothing that produces more noise (jackets, coats, ...), and once with light clothing while sitting. During the capture, all participants wear a VR headset on which we mounted seven microphones. For the reading activities, participants are shown the text to read on the display of the VR headset. For non-speech activities, participants are shown an example video of an instructor performing the respective activity. For conversational tasks, participants are instructed to talk for the given duration about any topic of their choosing.

References

Alexander Richard, Dejan Markovic, Israel D Gebru, Steven Krenn, Gladstone Alexander Butler, Fernando Torre, and Yaser Sheikh. Neural synthesis of binaural speech from mono audio. In *International Conference on Learning Representations*, 2021.

P. N. Samarasinghe and T. D. Abhayapala. 3D spatial soundfield recording over large regions. In *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012.

E. G. Williams. *Fourier Acoustics*. Academic Press, 1999.