Supplementary Materials: Be Everywhere - Hear Everything (BEE): Audio Scene Reconstruction by Sparse Audio-Visual Samples

Mingfei Chen [1] Kun Su [1] Eli Shlizerman [1, 2*]
[1] Department of Electrical & Computer Engineering, University of Washington
[2] Department of Applied Mathematics, University of Washington

1. Generated Samples and Overview Video

Please see the **attached video for a short overview and samples of video clips generated by BEE.** Please turn **Audio ON** and please use **headphones** for higher quality audio perception.

2. Immersive VR Demo

To better demonstrate the immersive experience of our proposed audio scene reconstruction task, we built a virtual 'jazz bar' in Horizon Worlds with Meta Quest, according to the layout and structure of a SoundSpaces-Replica scene [2]. Given the emitter actions and locations, we synthesized the audio-visual input using provided data from SoundSpaces, and then deployed BEE to render the target sounds for listener locations. To indicate locations of the emitter and the listener, we inserted the emitter and the listener avatars into the virtual scene and set the motions consistent with the testing settings on SoundSpaces, since the virtual room is similar to the used Replica scene. In the VR demo video in the supplementary video, we show the reconstructed sounds for two listener initial locations and then moving through the scene, with same emitters. Specifically, listener 1 is close to the piano and far from the saxophone such that it can be heard that there is a minor saxophone sound mix with strong piano and vocals. For listener 2, a stronger saxophone sound is heard since the listener is closer to the saxophone. Further, the movement of the vocal is perceived well from the sound especially when headphones are used.

3. Examples of BEE Generated Audio In Comparison to Baselines

In the attached PowerPoint file and in the summary video we include several examples of BEE-generated audio and of two audio-visual baseline methods *Mono2Bi* [3] and *AP-Net* [5] for the same scene and same listener. These can be compared to the ground truth audio. Examples include various settings such as navigation, the same listener location with different emitters and different listener locations with same emitters. Inspection of these results shows qualitatively the overall higher quality of audio generation by BEE. In particular, the spatial aspects of the audio scene are easily noticeable in BEE version vs. other methods, are less noisy, and of higher precision.

4. Model Configurations

In the Joint Audio-Visual Representation (JAVR) module, we use a single ResUNet with a ResNet18 backbone and output dimension of 32 to extract the image features $\mathbf{F}_i$. We utilize a single three-layer hierarchical SparseConvNet to learn a $24 \times 24 \times 5$ visual feature volume $\mathbf{V}'$ with the dimension of 32 for the acoustic propagation space $\mathbf{P}$. Since all locations are sampled from one regular horizontal plane, we squeeze V' to a 24×24 plane with a channel dimension C of 64. For the audio input, we resample the audio with a sampling rate of 16000, then apply a Short-time Fourier transform (STFT) with a window length of 400 and hop length of 152 to transform the audio waveform to a 512×64 spectrogram, which contains 64 frames of 512 frequencies. We concatenate the real and imaginary parts of both the left and right ear spectrogram channels together to form a 4-channel audio input. The audio features are extracted with a 5-layer convolutional network and then integrated with visual and pose features in the feature space $\mathbf{P}$. The channel dimensions d of the audio-visual embedding $\tilde{Q}$ and $\hat{Q}$ are 128. All the sinusoidal encodings for pose embeddings apply 8 frequencies of sin and cos functions.

In Integrated Rendering Head (IRH), we introduce two decoupled branches. For the UpConv branch, we use a 5-layer Transposed Convolutional Network to generate binaural spectrogram weights based on audio features injected with audio-visual embedding $\hat{Q}$ after Cross-Attention. Both the left and right spectrogram weights contain two channels for real and imaginary parts of weights respectively. For the Magnitude branch, a 4-layer MLP is deployed to predict spectrogram magnitude weight for each frequency-time

*Corresponding author: shlizee@uw.edu

pair based on $\hat{Q}$ after Cross-Attention and the corresponding embedding of the target time and frequency index.

5. Role of Vision

To further highlight the role of vision, we conducted an experiment with a model of BEE receiving a fixed image (from one random viewpoint in one scene) as a visual input in training and testing for all scenes (FixImgBEE). We compared it with the full model (on all unseen Replica scenes) with respect to spatial audio reconstruction accuracy for the entire scene, testing all listener locations for accumulated STFT, ENV, and loudness errors. Compared to FixImgBEE, full BEE has acheives better metrics in terms of lower loudness, ENV, and STFT errors by 21%, 12%, and 13%, respectively. These results highlight the importance of visual information for spatial audio reverberation and continuous audio scene reconstruction since visuals can facilitate the effective capture of geometry and material features that are not directly recoverable from audio.

Methods	STFT ↓	ENV ↓	Loudness ↓
FixImgBEE	1.29	8.62	46.30
BEE (Full)	1.12 (-13%)	7.55 (-12%)	36.42 (-21%)

Table 1: Accumulated Error Comparison for All Unseen Scenes.

6. Sampling Strategy of Audio-Visual Samples

In order to capture sufficient audio-visual information of the scene, in BEE implementation we place fixed A/V reference receivers on the midpoints of four edges of the smallest rectangle that contains the room floor plane (*Periphery*), with orientation to the interior of the room.

To investigate the impact of different sampling strategies on the audio-visual samples, we implemented two other sampling strategies: *Random* and *Center*. The *Random* strategy randomly varies the locations and orientation angles of A/V receivers for each scene. While for the *Center* strategy, we place the A/V reference receivers with different orientation angles in the center of each scene. After training, we compare the results of these two sampling strategies to the original results of BEE on Matterport 3D scenes in Table 2, where we place 4 receivers for all strategies. Results in Table 2 illustrate that BEE has improved accuracy and perceptual quality when the A/V receivers are fixed at the edges of the room floor plane.

Indeed, the *Random* strategy lower quality indicates that the location of A/V receivers is important and there could exist A/V receivers of lower quality, e.g., a receiver in the corner of the scene or facing the walls. The performance gap can be larger when testing on scenes seen during training (seen scenes) since the *Periphery* sampling, i.e. BEE, can learn more reliable scene representation by fixing sensors to locations that capture informative A/V samples.

Strategies	Setting	STFT ↓	DPAM ↓	ENV ↓
Random	Seen	0.435	0.326	0.462
Center	Seen	0.436	0.303	0.473
Periphery (BEE)	**Seen**	**0.425**	**0.274**	**0.455**
Random	Unseen	0.457	0.350	0.461
Center	Unseen	0.495	0.370	0.556
Periphery (BEE)	**Unseen**	**0.438**	**0.348**	**0.458**

Table 2: **Comparison for Sampling Strategies on Matterport3D scenes**. The *Periphery* strategy of BEE outperforms other strategies for both seen and unseen settings.

From Table 2, we observe that for *Random* sampling the perceptual quality metric DPAM reduces by 15.95% compared to *Periphery* (BEE) for seen scenes. *Center* strategy faces challenges as well, since also may not capture all needed information to represent the scene and the emitters. While in *Center* sampling the visual sensors are placed in the center and can capture informative visual information about the scene, since the audio sensors are located close to each other, the observed audio input will be similar. In particular, for unseen scenes the input audio samples are critical for effective audio-visual scene representation and in these cases the visual representation is less reliable since the scene have not been seen in training. Therefore, the accuracy of *Center* strategy is expected to be significantly lower for unseen scenes. Compared with *Periphery* (BEE), STFT, DPAM and ENV metrics of BEE with *Center* strategy are less accurate by 11.52%, 5.9% and 17.63% respectively.

7. Robustness of A/V Sensor Density

To verify the robustness of the density of A/V sensors when testing on new scenes, we implement a density analysis experiment. After training with 4 A/V sensors on edge, we test the trained model with selecting a random subset of these sensors as functional sensors to reconstruct the target sound. We set the sensor number in the subset from 1 to 4, where 4 means all sensors in the training process are used for testing. The results are reported in Table 3. STFT, DPAM and ENV metrics remain robust when using 2-4 sensors. After reducing the number of sensors to 1, STFT and ENV errors increase, while the perceptual quality metric DPAM can remains relatively stable.

8. Audio-Visual Feature Plane

JAVR module of BEE constructs an audio-visual representation of feature space for the acoustic propagation of the scene. In the original implementation, we use 24×24 plane as the feature space. To investigate the impact of the plane dimensions and its discretization, we set four differ-

Number	STFT ↓	DPAM ↓	ENV ↓
1 A/V sensor	0.579	0.383	0.594
2 A/V sensors	0.478	0.354	0.503
3 A/V sensors	0.453	0.352	0.481
4 A/V sensors	**0.438**	**0.348**	**0.458**

Table 3: **Robustness Analysis of Different Sensor Density on unseen Matterport3D scenes.** BEE remains relatively robust when a lower sensor density is used with fewer sensors for testing.

ent spatial sizes for the feature plane: 12×12, 24×24, and 48×48. Except for plane dimensions, we keep other configurations fixed and conduct comprative experiments reported in Table 4. In particular, we compare the accuracy of the generated waveforms for different spatial dimensions. For both settings, the reconstruction accuracy of BEE improves when the spatial dimension increases from 12×12 to 24×24. When the spatial dimension is extended beyond, i.e., 48×48, the accuracy for unseen emitters does not improve further and even decreases due to overfitting. Thus for BEE implementation the plane dimensions are chosen as 24×24 .

Spatial Size	Setting	STFT ↓	DPAM ↓	ENV ↓
12×12	Seen	0.441	0.294	0.475
24×24	**Seen**	**0.425**	**0.274**	**0.455**
48×48	Seen	0.448	0.315	0.489
12×12	Unseen	0.467	0.369	0.513
24×24	**Unseen**	**0.438**	**0.348**	**0.458**
48×48	Unseen	0.454	0.383	0.458

Table 4: **Comparison for Different Spatial Dimensions of Audio-Visual Feature Plane on Matterport3D Scenes**. Dimension 24×24 used in the final BEE model performs better than planes of other dimensions for both seen and unseen settings.

9. Spatial Enhancement

We enhance visual features based on the spatial locations of M, L in the JAVR module since they can impact sound propagation paths. To justify this setting, we also implemented a variant that enhances joint audio-visual feature plane with receiver-listener locations for better comparison. However, this variant increases the number of model parameters by 160,000. Furthermore, the perfromance of the variant reduces accuracy, according to DPAM metric by 5.5% and causing approximately 1% higher errors in STFT and ENV, testing on all Matterport scenes. This is since knowledge of locations lack physical significance, and even though implicitly reveal emitter locations do not assist with enhancement of audio features captured by receivers. Therefore we only enhance visual features with the spatial locations.

10. Model Finetuning

To investigate the cost of adapting a pre-trained model to unseen scenes, we randomly select 25% scenes from unseen Matterport3D scenes and finetune the pre-trained models on each scene separately on 3000 emitter-receiver-listener samples with different emitters and emitted sound clips. After finetuning, we test the models in each finetuned scene with novel emitter-receiver-listener samples and emitted sounds. In Table 5, we compare our reconstruction results of finetuned BEE with the finetuning results of other baselines. Besides the three metrics: STFT, DPAM and ENV, we also report the average loudness map error of the scenes. Before finetuning, *BEE_unseen* can achieve higher perceptual quality DPAM than other methods even after finetuning. After a quick finetuning on 3000 samples, BEE exhibits improved accuracy on all metrics significantly and outperform all other baselines by a large margin, according to the results in Table 5.

Method	STFT ↓	DPAM ↓	ENV ↓	Loudness ↓
AViTAR [1]	0.134	0.347	0.283	0.200
Few-ShotRIR [4]	0.194	0.344	0.660	0.307
Mono2Bi [3]	0.158	0.314	0.432	0.193
APNet [5]	0.155	0.311	0.421	0.208
BEE_unseen	0.157	0.263	0.446	0.342
BEE	**0.116**	**0.222**	**0.243**	**0.161**

Table 5: **Finetuning Results on Matterport3D Scenes**. We compare BEE with other baselines after a quick finetuning on 3000 samples in each unseen Matterport3D scene. BEE achieved higher accuracy and outperformed other baselines by a large margin.

11. Multiple Emitters

In dynamic audio scenes there are multiple emitters that move. Furthermore, there could be additional emitters in the target scenes. In the training data, emitter sets contain only 1-2 emitters. In testing, the emitter numbers can be more than 2. In Figure 1, we place 1, 2, 5, 7 emitters in each scene for unseen Matterpor3D scenes and report BEE reconstructed DPAM metric and compare it to audio-visual baselines *APNet* [5] and *Mono2Bi* [3]. For each scene, we randomly test 10 target listener locations for each emitter number and average the DPAM error. The results show that BEE achieves better DPAM metrics indicating better quality and robustness across different numbers of emitters while for other methods DPAM metric increases with increasing number of emitters.

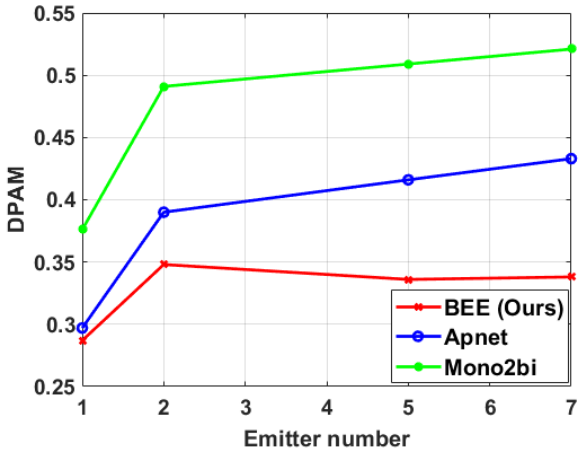

Figure 1: **Perceptual Quality Comparison with Different Emitter Numbers.** We compare the DPAM (**Lower is better**) of BEE (Ours) with *APNet* and *Mono2Bi* on 1, 2, 5 and 7 emitters respectively. BEE achieves better and more stable perceptual quality than the compared two methods even when the number of emitters increases.

12. Dataset License

SoundSpaces [2] is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

References

[1] Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. Visual acoustic matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18858–18868, 2022. 3

[2] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *European Conference on Computer Vision*, pages 17–36. Springer, 2020. 1, 4

[3] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2019. 1, 3

[4] Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics, 2022. 3

[5] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 52–69. Springer, 2020. 1, 3