Visual Acoustic Matching

Changan Chen[1,4] Ruohan Gao[2] Paul Calamia[3] Kristen Grauman[1,4]
[1]University of Texas at Austin [2]Stanford University [3]Reality Labs Research at Meta [4]Meta AI

8. Supplementary Material

In this supplementary material, we provide additional details about:

1. Supplementary video for qualitative assessment of our model's performance.
2. Acoustic AVSpeech filtering process (referenced in Sec. 4 of the main paper).
3. Acoustic changes after each alteration step (referenced in Sec. 5).
4. Implementation and training details (referenced in Sec. 6).
5. Evaluation and baseline details (referenced in Sec. 6).
6. Ablations on acoustics alteration (referenced in Sec. 6.2).
7. Ablations on GAN losses (referenced in Sec. 6).
8. Sim2real generalization (referenced in Sec. 6).
9. Applicability on non-speech sounds (referenced in Sec. 4).
10. Interpretation of the neural network results.
11. Does the model capture room size?
12. User study interface.
13. Societal impact (referenced in Sec. 7).

8.1. Supplementary Video

This video includes examples generated by AViTAR and baselines for SoundSpaces-Speech and Acoustic AVSpeech. We also demonstrate application scenarios for augmented reality and video conferencing. Wear your headphones for a better listening experience.

8.2. Acoustic AVSpeech Filtering Process

As noted in the main paper, we apply a series of automatic filters to the AVSpeech dataset [2] in order to select those clips relevant for our task. Here we detail those steps.

AVSpeech is a large-scale audio-visual dataset comprising speech video clips with no interfering background noises. The segments are 3-10 seconds long, and in each clip the audible sound in the soundtrack belongs to a single person speaking who is visible in the video. In total, the dataset contains roughly 4700 hours of video segments, from a total of 290k YouTube videos, spanning a wide variety of people, languages and face poses.

Since our dereverberation model used during acoustics alteration is trained on an English corpus, we first run a language classification algorithm over all the AVSpeech audio clips and remove clips where the spoken language is not English. After this step, there are still many videos which are almost anechoic, sometimes due to the audio being recorded post video recording, or to the speaker using a microphone very close to his/her mouth. To remove such examples, we train an RT60 predictor on the SoundSpaces-Speech (details in Sec. 8.5), run it on all AVSpeech clips and remove examples where the predicted RT60 is less than 0.1s. Lastly, we balance the distribution of RT60 such that it is not heavily skewed toward the anechoic side.

8.3. Acoustic Changes After Each Alteration Step

In Table 5, we show how the acoustics change after performing each step in the acoustics-alteration process by evaluating RT60 and MOS of the processed speech on the test split. What we expect to see is that the original audio gets cleaner via dereverberation, then becomes increasingly reverberant and noisy as we perform the subsequent steps that are designed to disguise the audio with other room acoustics from the sampled IR. This is indeed what we observe. The original audio input has a high RT60 value on average, but after dereverberation the RT60 drastically goes down to 0.088s and the speech quality becomes better. After reverberating, the average RT60 goes up again, with a lower MOS score. Adding noise slightly improves the RT60 value and reduces the speech quality. For clean speech, its average RT60 is much lower and the MOS score is also high. Note that here we show the MOS scores, not the MOS errors; higher values indicate higher quality speech.

8.4. Implementation and Training Details

The 1D convolutions for encoding and decoding the waveform have kernel sizes of $16, 8, 4, 4$ and strides $8, 4, 2, 2$ respectively. The total downsampling/upsampling

Acoustic Changes	RT60 (s)	MOS
Original audio	0.436	2.778
Dereverb.	0.088	2.970
Dereverb. + Randomization	0.424	2.620
Dereverb. + Randomization + Noise	0.462	2.513
Clean	0.049	3.285

Table 5. Acoustic changes after each alteration step.

rate D is 128. The latent feature size for A_i, V_i and M_i is 512. The number of cross-modal encoders N is 4. There are 8 attention heads in each attention layer. The number of sub-discriminators K is 3 and λ_1 and λ_2 are 1 and 45, respectively. The learning rate for the generator and discriminators are 0.005 and 0.002.

The input audio clip is 2.56 seconds for both datasets. On SoundSpaces-Speech, the input image size is 192×576, and we randomly shift the panoramic image during training for the model to learn viewpoint-invariant room acoustics features, following the original paper [1]. On Acoustic Speech, the input image is first resized to 270×480, followed by random cropping to size 180×320 and random horizontal flip for data augmentation. We train all models 600 epochs on SoundSpaces-Speech and 300 epochs on Acoustic AVSpeech, and evaluate the checkpoint with the lowest validation loss on the test set. We will share the code and data upon acceptance.

8.5. Evaluation and Baseline Details

RT60 estimator. On SoundSpaces-Speech, we have access to the reverberant speech clip as well as the impulse response. We first encode the 2.56s speech clips as spectrograms, process them with a ResNet18 [3] and predict the RT60 of the speech. The ground truth RT60 is calculated with the Schroeder method [4]. We optimize the MSE loss between the predicted RT60 and the ground truth RT60.

Image2Reverb [5]. We obtained the code from the authors and made some changes to accommodate their model on our dataset. First of all, we replace the depth estimator with the ground truth depth image that we have access to on SoundSpaces-Speech. We also increase the size of the input image to match the size of the panorama. Lastly, we change the sampling rate from 22050 to 16000. The rest of the code stays the same, including the visual encoder pretrained on Places365 and the auxiliary loss on RT60 prediction.

8.6. Ablations on Acoustics Alteration

Table 3 shows ablations on the proposed acoustics-alteration strategy. Removing either the acoustic randomization or noise leads to worse generalization to novel sounds compared to the full process. This is because without these two steps, it is easier for the model to overfit the residual acoustic information in the dereverberated audio rather than use the visual content for recovering correct acoustics. If both are removed ("Dereverb."), the model does not generalize to novel sounds. Similarly, the dereverberation step is also very important. If we simply randomize the acoustics with another IR and add noise to the original audio ("A_T + Randomization + Noise"), there is no training sample that has less reverberation than the target audio, and the model simply learns to perform dereverberation; this leads to poor generalization as well. Altogether, all three steps are necessary to create acoustic mismatch with the image and force the model to recover the correct acoustics based on images.

AViTAR	STFT	RTE (s)	MOSE
Full model	**0.822**	**0.062**	0.195
w/ $\mathcal{L}_{Mel}$	2.907	0.190	0.833
w/ $\mathcal{L}_{FM}$	0.831	0.063	**0.192**

Table 6. Ablations on GAN loss components.

8.7. Ablations on GAN Losses

Here we detail each GAN loss component and how they affect the performance. Mel-spectrogram loss $\mathcal{L}_{Mel}$ is the L1 distance between two mel-spectrograms, which improves the perceptual quality. Feature matching loss $\mathcal{L}_{FM}$ is a learned similarity metric for features of the discriminator between two audio samples. We ablate these two loss terms separately and the results are shown in Table 6. Removing the Mel-spectrogram loss leads to a great reduction on all metrics. Removing the feature matching loss leads to higher STFT distance and RT60 error while marginally lower MOS error. Overall, these two ablations show both components are important for synthesizing realistic audio with matched acoustics.

8.8. Sim2real Generalization

To understand how well the model trained on synthetic dataset generalizes to web videos, we train a new AViTAR model on SoundSpaces-Speech with only RGB input, and then test it on the Acoustic AVSpeech dataset, which yields RTE of 0.278s, MOSE of 0.898, while the model trained and tested on Acoustic AVSpeech gives 0.183s RTE and 0.453 MOSE (Table 1). The newly trained synthetic model tends to generate more reverberation, likely due to the visual discrepancy. This highlights the effectiveness of our self-supervised acoustic alteration strategy.

8.9. Applicability on Non-speech Sounds.

To understand if our models applies to non-speech sounds, we train AViTAR on SoundSpaces by replacing the human speech with non-speech sounds, e.g. ringtone, music, etc., the model has 0.064 RTE on test-unseen, higher

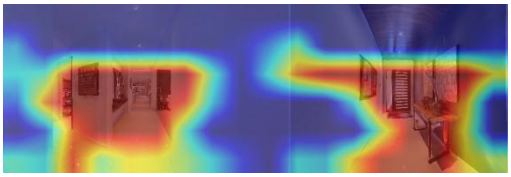

Figure 6. Grad-CAM for corridor scene.

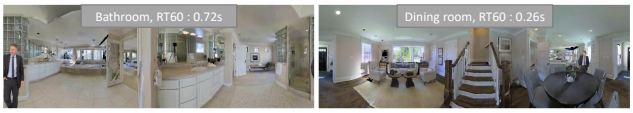

Figure 7. Rooms of similar sizes but different acoustics.

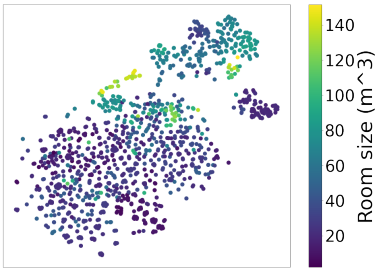

Figure 8. T-SNE projection of visual features colored by room size

than human speech (0.062 RTE), while outperforming AV U-Net (0.074 RTE) and the input (0.176 RTE). So while we focus on speech for application reasons, this positive non-speech result makes sense because our model design is agnostic to the type of audio.

8.10. Interpretation of the Neural Network Results.

To show how the model understands the image, we can use Grad-CAM to visualize the activations. For example, in Fig. 6 Grad-CAM highlights two sides of the corridor because they lead to longer reverberation. Fig. 7 shows two rooms of similar sizes, and our model predicts longer RT60 for the bathroom likely because it has more reflective materials and leads to longer reverberation time.

8.11. Does the Model Capture Room Size?

To understand if our learned model captures room sizes, we check two things: 1) whether the learned visual features manage to pick up on room size (the clustered colors in Fig. 8 suggest yes), and 2) whether we output only a narrow set of acoustics for the same room type (the distribution of RT60s over all kitchens in the test split (Table 7) suggests no). Furthermore, we project visual features on the 2D plane colored by visible room volume with T-SNE (shown in Figure 8). The gradient from small room volumes to large room volumes indicates that room size is captured in visual features. In addition, we show the distribution of RT60s over all kitchen environments in the test-unseen split in Table 7 and it is quite diverse.

RT60 (s)	≤ 0.2	0.2-0.3	0.3-0.4	0.4-0.5	≥ 0.5
Percent (%)	11.9	55.0	27.7	5.5	1.0

Table 7. RT60 distribution over kitchens (other scenes show similar diversity).

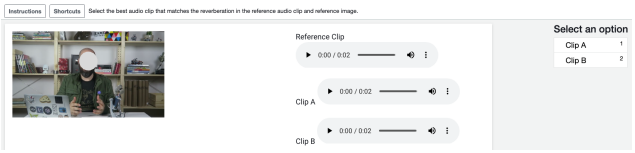

Figure 9. User study interface on MTurk. Given a reference image, a reference audio and two clips generated by AViTAR and a baseline (with shuffled order), participants are asked to pick the best clip that matches the reverberation in the reference image and audio.

8.12. User Study Interface

Figure 9 shows the interface for our user study on MTurk. See details of the instruction in the caption.

8.13. Societal Impact

We believe this work can have a positive impact on many real-world applications, e.g., video editing, film dubbing, and AR/VR, and discussed in the paper. However, future applications built on such technology must also take care to avoid its misuse. The ability to transform a voice to sound like it comes from a new environment could potentially be misused for enhancing deep fake videos, by matching an audio not recorded along with the video to the visual stream.

References

[1] Changan Chen, Wei Sun, David Harwath, and Kristen Grauman. Learning audio-visual dereverberation. In *arXiv*, 2021. 2

[2] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. In *SIGGRAPH*, 2018. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[4] Manfred R. Schroeder. New method of measuring reverberation time. In *The Journal of the Acoustical Society of America 37, 409*, 1965. 2

[5] Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2reverb: Cross-modal reverb impulse response synthesis. In *ICCV*, 2021. 2