1 Supplementary

In this supplementary material we provide the following:

1. A video for qualitative evaluation of our model's performance (1.1).
2. Details regarding AVSpeech-Rooms curation (1.2) (referenced in Sec. 4 of main paper)
3. Details on our ablation study with different metric training objectives (1.3) (referenced in Sec. 5 — "Results on AVSpeech-Rooms" of main paper)
4. A sample survey slide from our human perception study (1)
5. Model/training details for our RT60 estimator, de-biaser, discriminator, and reverberator (1.4) (referenced in Sec. 5 — "Implementation Details" of main paper)
6. A more detailed version of Figure 4. in the main paper with both baseline models (2)
7. Details on our data augmentation strategy (1.5) (referenced in Sec. 5 — "Baselines" of main paper)
8. A brief discussion of our work's limitations and broader impact (1.6 1.7)

1.1 Supplementary Video

Our video contains several illustrative examples generated by LeMARA on both SoundSpaces-Speech and AVSpeech-Rooms. We provide audio generated by the current state-of-the-art (AViTAR) for reference on each example. We recommend wearing headphones for a better listening experience.

1.2 AVSpeech-Rooms

Acoustic AVSpeech consists of audio clips from YouTube videos along with an RGB image frame selected randomly from the corresponding video clip. To create AVSpeech-Rooms, we design a set of criteria which we use to filter out samples in which the image contains uninformative, non-natural, or misleading acoustic information about the space. We focus on cases in which the room is not visible, a microphone is being used, or a virtual background/screen is present — any of which will disturb the natural room acoustics for the speaker's voice. We query each sample with our criteria using a Visual Question Answering (VQA) model [6], which we found more reliable than manual annotations we originally obtained on MTurk. 1 contains information about our criteria.

Table 1: Filtering criteria and % of Acoustic AVSpeech samples removed.

Question	Answer	dataset %
Is a microphone or headset visible in the image?	yes	7.2
Is there a whiteboard/blackboard in the background?	yes	3.4
Is the entire background one solid color and material?	yes	23.4
Is there a large projector screen covering most of the background?	yes	2.2
Is part or all of the background virtual?	yes	1.3
Are there multiple screens in the image?	yes	3.5
Is the wider room clearly visible?	no	3.0

1.3 Ablations

Table 2 displays our experiments with different self-supervised training objectives. We report performance on the LibriSpeech evaluation setting. The first three rows correspond to experiments in which we do not utilize the shorcut training strategy (referenced in Sec. 3 — "Training" of main paper). Using SRMR alone (row 1) produces the largest (worst) RTE. Training with the acoustic residue metric instead (row 2) leads to a large improvement in RTE, providing empirical support for our metric as an effective training objective. Using our combined metric and the shortcut training strategy (both described in Sec. 3 — "Training" of our main paper) further improves the performance by a small margin.

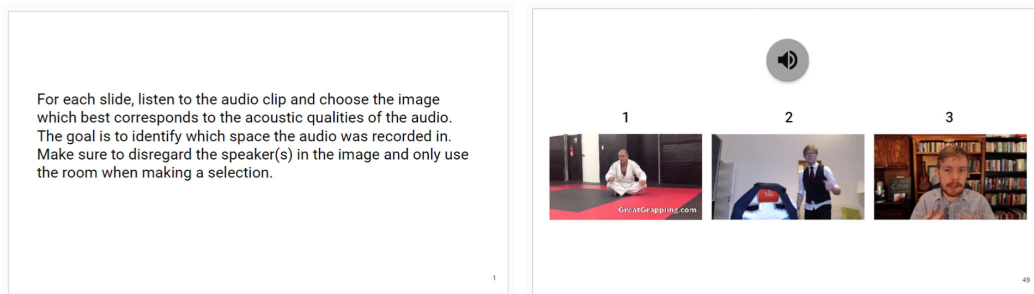

Figure 1: **Human perception study.** The instructions given to the user at the start of the survey (left), and a sample slide from the survey (right). The user is asked to listen to the audio clip, and identify which room image most closely matches its acoustics.

Table 2: Ablation study using different metric training objectives. AR denotes the proposed acoustic residue metric.

Metric	*LibriSpeech* RTE
SRMR [4]	0.2308
AR	0.2156
AR (combined)	0.2123
AR (combined) w/ shortcut	**0.2100**

1.4 Model/Training details

RT60 estimator We adopt the RT60 estimator from [1]. The estimator takes a spectrogram as input, encodes it with a ResNet18 [5], and outputs a scalar RT60 estimate. The model is trained on 2.56s clips of reverberant speech simulated on the SoundSpaces platform [2] paired with the ground truth RT60 computed from the RIR used to generate the reverberant speech. The model trains using MSE loss between predicted and ground truth RT60 values. Ground truth RT60 is computed using the Schroeder method [7].

De-biaser architecture The de-biaser G takes a magnitude spectogram as input. This is passed to a bi-directional LSTM with input size 257 and two hidden layers each of size 200, which produces an output with the same temporal length as the input spectrogram. This is passed through a linear layer of size 300 and a leakyReLU activation, followed by another linear layer of size 257 and a Sigmoid activation. The final mask is multiplied with the input magnitude spectrogram to create the generated magnitude spectrogram. A resynthesis module computes phase information from the input audio waveform, combines this with the generated magnitude spectrogram, and performs an inverse STFT to produce the generated waveform. The discriminator D consists of 4 2D Convolutional layers with kernel size (5,5) and 15 output channels, followed by a channel averaging operation and two linear layers of sizes 50 and 10. A LeakyReLU activation with negative slope $= 0.3$ is used after each intermediate layer. The final layer outputs a scalar-valued metric score estimate.

De-biaser training In stage (1) (see Sec. 3 — "Training" of our main paper), we train with batch size 32. During stage (3) fine-tuning, we use a batch size of 2. G and D are trained with learning rates of 2e-6 and 5e-4 respectively in both stages. In each epoch, We train on 10k samples randomly selected from the train set without replacement. The reverberator models R_v and R_b are updated with the target networks at a frequency of $E = 8$ epochs. For all models, we clip each audio sample to 2.56s during training and evaluation.

Reverberator training We train the reverberators with batch size 4 and a learning rate of 1e-2 in stage (2). During stage (3) fine=tuning, we use batch size 2 and a learning rate of 1e-6. Both reverberator models and the ViGAS baseline are trained with MSE loss between the log magnitude spectrogram of predicted and ground truth audio.

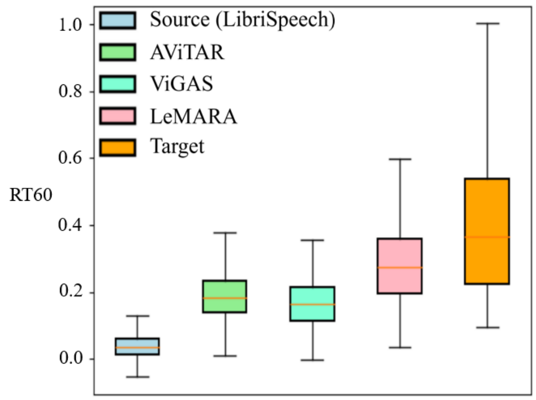

Figure 2: **LibriSpeech evaluation.** A more detailed version of Figure 4 (see main paper) with both baseline models for reference. The distribution of RT60 values for LeMARA reverberated audio (pink) better matches the ground truth distribution (orange) than either baseline (AViTAR and ViGAS).

Baseline training details We use a learning rate of 1e-2 and a batch size of 4 to train ViGAS. We train AViTAR with batch size 4 — all other hyperparameters are set as described in [1].

Compute All models are trained on 8 NVIDIA Quadro RTX 6000 GPUs.

1.5 Augmentation strategy

We follow a data augmentation strategy similar to that proposed in [1] for training the baseline models, which was shown to produce better generalization performance on the LibriSpeech setting than when trained without this augmentation strategy. In particular, to each batch of dereverberated audio we add colored noise, perform a polarity inversion on the waveform with $p = 0.5$, and convolve the waveform with a randomly selected Room Impulse Response (RIR) from a different acoustic environment with $p = 0.9$. At test time, we evaluate without these audio augmentations. This strategy is designed to mask over residual acoustic information in dereverberated audio during training. We do not use this augmentation strategy in our approach as our model directly learns to remove residual acoustic information, obviating the need for a heuristic strategy to mask it out.

1.6 Limitations

Our approach focuses on visual acoustic matching on mono-channel audio exclusively. However, binaural cues in audio play a fundamental role in our perception of reverberation and room acoustics [3]. We leave it to future work to extend our approach to binaural audio.

1.7 Broader impact

While training on in-the-wild web videos allows wider access to a diverse variety of speakers and environments, it also introduces uncontrolled biases, speaker privacy concerns, and potentially harmful content into the model.

1.8 Data examples

Refer to video to view samples from both SoundSpaces-Speech and AVSpeech-Rooms.

References

[1] Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. Visual acoustic matching. In *CVPR*, 2022.

[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 *ECCV*, 2020.

[3] Sasha Devore and Bertrand Delgutte. Effects of Reverberation on the Directional Sensitivity of Auditory Neurons across the Tonotopic Axis: Influences of Interaural Time and Level Differences. *Journal of Neuroscience*, 30(23):7826–7837, 2010. Publisher: Society for Neuroscience _eprint: https://www.jneurosci.org/content/30/23/7826.full.pdf.

[4] Tiago H. Falk, Chenxi Zheng, and Wai-Yip Chan. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1766–1774, 2010.

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

[6] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision, 2021.

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