7. Appendix

7.1. Estimated Dereverbed Audio Samples

We provide audio samples of the estimated dereverberant audio generated by AdVerb in the project webpage (use of headphones recommended). Some of these samples were used during the user study. We observe that the perceptual quality of the audio samples generated by our model is significantly better compared to other prior methods (established by the user study).

7.2. Additional Experiments And Results

7.2.1 Performance with 3D Humanoid Removed

To inspect the influence of human speaker cues, we carry out an experiment with the humanoid removed keeping everything else the same. We observe that all the metrics become impacted and achieve inferior scores (Table 6). We conclude the speaker's location-specific information is crucial for better learning ability of the model in order to perform dereverberation.

	SE PESQ ↑	**SR** WER ↓	**SV** EER ↓
AdVerb *w/o human mesh*	2.94	3.67	3.15
AdVerb *w/ human mesh*	2.96	3.54	3.11

Table 6: Result comparison of AdVerb with and without 3D humanoid in the panoramic images.

7.3. Further Ablations on LibriSpeech dataset

7.3.1 Role of Complex Ideal Ratio Mask

As described in Section 4.4, the complex Ideal Ratio Mask (cIRM) is an extension of the conventional ideal ratio mask (IRM) to process the real and imaginary components of the STFT separately. Thus, to measure the influence of the complex aspect of cIRM, we replace it with IRM and make the necessary updates in the pipeline. Table 7 compares the performance of both models. As can be observed, employing cIRM improves performance over using the conventional mask-based pipeline, which is consistent with prior findings [89].

	SE PESQ ↑	**SR** WER ↓	**SV** EER ↓
AdVerb *w/ IRM*	2.52	4.12	3.85
AdVerb *w/ CIRM*	2.96	3.54	3.11

Table 7: Result comparison of AdVerb with and without the Complex Ideal Ratio Mask.

7.4. Evaluation Details

Task Description. As mentioned in Section 5, to compare the performance of AdVerb with the baselines under consideration, we evaluate our models on 3 speech tasks, including automatic Speech Recognition (SR), Speaker Verification (SV), and Speech Enhancement (SE). These 3 tasks can be explained as follows:

- The objective of **SE** is to improve the overall speech quality by suppressing the noise in a noisy speech signal. We evaluate performance on this task using the standard Perceptual Evaluation of Speech Quality (PESQ) metric.
- The goal of **SR** is to automatically transcribe a given speech signal into its corresponding text or contents of the speech utterance. We evaluate performance on this task using the standard Word Error Rate (WER), which calculates the word-level edit distance between the transcribed output and the ground truth text.
- Aim of **SV** is to distinguish if two utterances were from two distinct speakers. We evaluate performance on this task using the standard Equal Error Rate (EER). The EER is the location on a ROC or DET curve where the false acceptance rate and false rejection rate are equal.

7.5. Hyperparameter Tuning Experiments

We train AdVerb for 100 epochs with a batch size of 16 using an Adam optimizer. For model optimization, we find $\lambda = 1$ and $\mu = 0.1$ give the best performance. All hyperparameters were tuned with grid search for the best performance on the dev set. In this sub-section, we show the performance of our model with different values of λ and μ for Spectrogram Prediction and Acoustic Token Matching losses respectively. Table 8 shows the effect of λ on AdVerb performance while μ is kept constant at 0.1. As we clearly see, the performance across all 3 tasks degrades as λ decreases. Table 9 shows the effect of μ on AdVerb performance while λ is kept constant at 0.1. The performance of ASR falls sharply where μ moves to zero, which proves that the Acoustic Token Matching loss helps preserve phonetic information in speech, thereby improving ASR performance. All experiments were done for the non-fine-tuned version of our experimental setup, where a pre-trained model was used from SpeechBrain.

7.6. User Study

7.6.1 More Qualitative Results

Table 10 extends Table 5 to report the subjective evaluation results (30 participants) against the methods in Table 2. Note our approach is complementary to the audio-only

(a) Atrium

(b) Long Corridor

(c) Library

(d) Meeting room

Figure 7: Some examples from the Matterport 3D dataset showing different challenging environments the model was evaluated against.

	SE PESQ ↑	ASR WER ↓	SV EER ↓
$\lambda = 1$	2.96	3.54	3.11
$\lambda = 0.1$	2.11	3.97	3.76
$\lambda = 0.01$	2.02	4.13	4.04
$\lambda = 0.001$	1.99	5.01	4.11
$\lambda = 0$	1.97	5.87	4.21

Table 8: Result comparison of AdVerb for different values of λ across 3 speech tasks on LibriSpeech. All settings are consistent with Section 5 in the paper.

	SE PESQ ↑	ASR WER ↓	SV EER ↓
$\mu = 1$	2.87	3.36	3.27
$\mu = 0.1$	2.96	3.54	3.11
$\mu = 0.01$	2.81	4.16	3.19
$\mu = 0.001$	2.77	4.24	3.01
$\mu = 0$	2.89	4.67	3.17

Table 9: Result comparison of AdVerb for different values of μ across 3 speech tasks on LibriSpeech. All settings are consistent with Section 5 in the paper.

methods (we use audio-visual inputs) and a direct comparison with these methods might not be fair. However, to study the effectiveness of our methods we add the following comparisons.

Interface and Evaluation. We compare the predicted dereverbed audio produced by AdVerb with three other SOTA dereverberation models VIDA [12], SkipConvGAN [40], and WPE [52]. Fig. 8 shows the interface for our user study on Amazon MTurk.

Baseline Method	SoundSpaces(in %) (A% / B% / C%)	AVSpeech(in %) (A% / B% / C%)
Audio-only AdVerb	20.0 / 6.6 / **73.3**	23.3 / 6.6 / **70.0**
DEMUCS [15]	13.3 / 10.0 / **76.6**	16.6 / 10.0 / **73.3**
VoiceFixer [42]	30.0 / 6.6 / **63.3**	23.3 / 10.0 / **66.7**
HiFi-GAN [74]	16.6 / 3.3 / **80.0**	13.3 / 6.7 / **80.0**
Kothapally *et al.* [39]	26.6 / 6.6 / **66.6**	26.6 / 13.3 / **60.0**

Table 10: User study results. Where (***A,B,C***) consistent with Table 5. Users find samples from AdVerb to be perceptually better and cleaner when compared against prior methods

7.7. Visual Environments and Sample Curation

Some examples of different environment settings to which our model was evaluated are presented in Fig.7. The dataset proposed in [12] uses visual environment samples from Matterport 3D [6] with SOTA audio simulations done using SoundSpaces [11] to realistically capture the environments' spatial effects on real samples of recorded speech. The dataset enables flexibility over various physical environments, listener/ source positions as well as speech content of the sources. Matterport offers diverse and complex real-world 3D environments with each environment having multiple rooms spanning an average of 517m^2. LibriSpeech [58] which is widely used for benchmarking in the speech recognition literature, was chosen as the source speech corpus. It contains 1,000 hours of 16kHz read English speech from audiobooks. Following [12] we train our

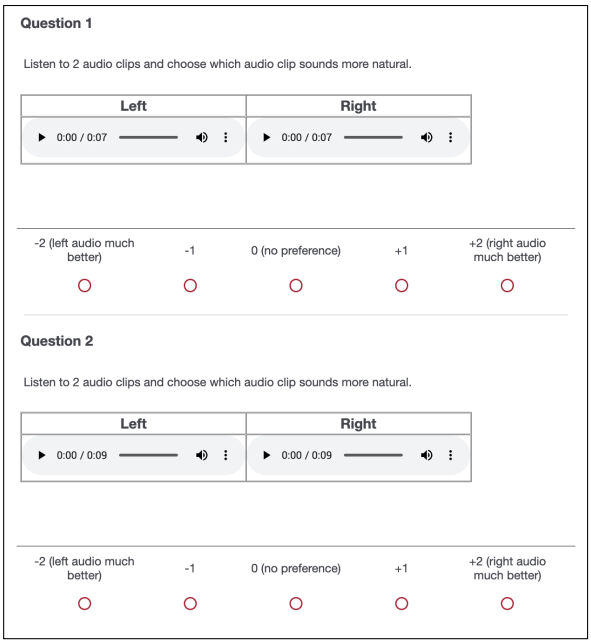

Figure 8: User study interface. Estimated dereverbed audio samples generated by AdVerb are compared against that of 3 other baseline models and the GT clean speech (not present for AVSpeech) VIDA[12], WPE[52] and SCGAN[40]. The audio samples are shuffled so that there is no bias among users while rating the audio samples. Participants are asked to choose the audio sample that sounds cleaner and more realistic.

models with the train-clean-360 split and use the dev-clean and test-clean sets validation and test phases respectively. These splits have non-overlapping speaker identities. Disjoint train/val/test splits for the Matterport 3D visual environments was followed to ensure the house's speaker voices are observed either during training or testing.

7.8. Societal Impact

We believe audio-visual dereverberation can positively influence a myriad of real-world applications involving: teleconferencing systems, speech recognition, hearing aids, and video editing among others. Specifically, dereverberation is very critical for hands-free phones and desktop conferencing terminals because, as the microphones are not close to the sound source in these cases but at a considerable distance.

Lastly, we would like to mention, the user study protocol was approved by Institutional Review Board and we do not collect, share or store any personal information of the participants.