AV-RIR: Audio-Visual Room Impulse Response Estimation

Anton Ratnarajah Sreyan Ghosh Sonal Kumar Purva Chiniya Dinesh Manocha
University of Maryland, College Park
{jeran,sreyang,sonalkum,pchiniya,dmanocha}@umd.edu

1. Table of Contents:

In the Supplementary Material, we provide additional information about:

- The qualitative results of our AV-RIR via a supplementary video[1].
- Quantitative comparison of AV-RIR on far-field Automatic Speech Recognition with other baselines.
- Additional details on our datasets and baselines used for evaluation.
- Additional details of our user study.
- Information about Societal Impact of AV-RIR.

1.1. Supplementary Video

We provide a supplementary video showing the qualitative results of RIR estimation with AV-RIR when applied to three different tasks. In addition, we compare the enhanced speech from our AV-RIR with ground truth clean speech. We also demonstrate our approach's failure cases in the supplementary video.

The RIRs estimated from our approach are evaluated in three practical tasks, that are:

Novel View Acoustic Synthesis : In the novel view acoustic synthesis task, given the audio-visual input from the source viewpoint, we modify the reverberant speech from the source viewpoint to sound as if it is recorded from the target viewpoint. We use reverberant speech as audio input, and the panoramic image and our proposed Geo-Mat feature as our visual input. We use SoundSpaces [4] dataset to perform this task.

To perform this task, we estimate the enhanced speech using audio-visual input from the source viewpoint. We estimate the RIR corresponding to the target viewpoint from audio-visual input. We convolve the enhanced speech from the source viewpoint with RIR from the target viewpoint to make the speech from the source viewpoint sound as if it is recorded from the target viewpoint.

Visual-Acoustic Matching : In the visual-acoustic matching task, we resynthesize the speech from the source environment to match the target environment. We combine the enhanced source environment speech from AV-RIR and the estimated RIR from the target environment to perform this task. Convolving the estimated RIR with clean speech leads to synthesizing speech from the source environment to match the target environment. All our experiments on this task are performed on the SoundSpaces [4] dataset.

Voice Dubbing : Voice dubbing is replacing dialogue in one language with another in a video. Voice dubbing is commonly used to dub movies from one language to another. To test the robustness of RIR estimation from our AV-RIR, we estimated RIR using our AV-RIR on recorded video clips on YouTube. We chose two English video clips in the AVSpeech dataset [10]. We dubbed the video clips with French clean speech from Audiocite [1]. We convolved the French clean speech with the estimated RIR from the YouTube clip to match the room acoustics of French dialogue with the original English dialogue. We replaced the English dialogue with modified French dialogue using our approach.

1.2. Far-field Automatic Speech Recognition

In order to evaluate the effectiveness of RIR estimated from our AV-RIR, we performed a Kaldi Far-field Automatic Speech Recognition (ASR) experiment using a modified KALDI ASR recipe[2]. For our experiment, we use the AMI corpus [2]. The AMI corpus has 100 hours of meeting recording. The meeting is recorded using both an individual headset microphone (IHM) and a single distant microphone (SDM). The IHM data has a high signal-to-distortion ratio when compared to the SDM data. Therefore, IHM data can be considered as clean speech.

To evaluate the benefit of RIRs estimated from our AV-RIR, we take a subset of SDM data with 300 speech samples and estimate the RIRs of the subset of SDM data. We create synthetic reverberant speech data by convolving clean speech from IHM data with the estimated RIR. We train the KALDI ASR recipe with and without modifying the IHM using our audio-only AV-RIR. We evaluated the audio-only version of our AV-RIR because there are no corresponding

[1] https://www.youtube.com/watch?v=tTsKhviukAE

[2] https://github.com/RoyJames/kaldi-reverb/tree/ami/

Table 1. Far-field ASR results. We train the Kaldi ASR recipe with and without modified IHM data and test on SDM data. We modify the IHM data by convolving RIR estimated using our audio-only AV-RIR.

Training Dataset	Word Error Rate ↓ [%]
IHM without Modification	64.2
IHM ⊛ AV-RIR (ours)	**52.1**

visual inputs in the AMI corpus. We test the trained model on far-field SDM data. We use word error rate as our metric to evaluate the performance of the speech recognition system. A lower word error rate indicates improved performance.

Modifying IHM data using our audio-only AV-RIR will bridge the domain gap between the training and test data. From Table 1 we can see that modifying the IHM data with our audio-only AV-RIR improves the word error rate by 12%.

1.3. Additional details on our datasets and baselines

1.3.1 SoundSpaces dataset:

We trained and test our network on SoundSpaces dataset [4]. The SoundSpaces dataset comes with a non-overlapping clean speech from the LibriSpeech dataset [16] and synthetic reverberant speech. The synthetic reverberant speech is simulated using the geometric acoustic simulator in the SoundSpaces platform [4, 6] for 82 Matterport [3] 3D environments. SoundSpaces can simulate highly realistic RIR for any arbitrary camera views and microphone positions by considering direct sounds, early reflections, late reverberations, material and air absorption properties, etc. The panoramic images in the SoundSpaces dataset contain 3D humanoids of the same gender as the speaker in each data. In some data, the speaker is out of view and will not be visible in the panoramic image. The sound spaces dataset has 49,430/2700/2,600 train/validation/test samples respectively.

1.3.2 AVSpeech Web Video dataset:

To evaluate the robustness of our approach, we evaluate our RIR estimation and Speech enhancement approach using a subset of 1000 speeches in AVSpeech dataset [10] proposed in Visual Acoustic Matching paper [5]. The filtered dataset contains 3-10 seconds YouTube clips with reverberant audio recordings. Also, the filtered dataset microphone and the camera are co-located and placed at a different position than the sound sources. The cameras in the filtered videos are static.

1.3.3 Speech enhancement baselines:

MetricGAN++[11]: MetricGAN++ is an improvised version of the MetricGAN framework where the discriminator network is trained with noisy speech. We use the implementation of MetricGAN in Speechbrain for our comparison [17].

DEMUCS [9]: DEMUCS is the music source separation architecture in the time-domain modified into a time-domain speech enhancer. DEMUCS can work in real-time on consumer-level CPUs.

HiFi-GAN [19]: HiFi-GAN is GAN-based architecture trained on multi-scale adversarial loss in both the time domain and time-frequency domain to enhance real-world speech recording to studio quality.

WPE [15]: WPE is a statistical model-based speech dereverberation approach. WPE can perform dereverberation by removing late reverberation in a reverberant speech signal.

VoiceFixer [14]: Voice-fixer is a two-stage speech dereverberation approach. The analysis stage of the VoiceFixer is modelled using ResUNet and the synthesis stage is modelled using TF-GAN.

SkipConvGAN [13]: SkipConGAN is the GAN-based speech enhancement architecture where the Generator network estimates the complex time-frequency mask and the discriminator network helps to restore the formant structure in the synthesized enhanced speech.

Kotha *et al.* [12]: This speech enhancement network integrates the complex-valued TFA module with the deep complex convolutional recurrent network to improve the overall speech quality of the enhanced speech.

VIDA [7]: VIDA is the audio-visual speech dereverberation network that enhances reverberant speech. Visual input gives valuable information about the room geometry, materials and speaker positions.

AdVerb [8]: Adverb is a geometry-aware cross-modal transformer architecture, that predicts the complex ideal ratio mask. Clean speech is estimated by applying the complex ideal ratio mask to reverberant speech.

1.4. Additional details on User Study

We performed our user study on 50 participants. We only allow participants to perform the survey on a laptop or desktop with headphones to get accurate results. Among the 50 responses, we filtered out noisy responses from our first questions. In the first question, we ask the participants which of the three synthetic reverberant speech matches closely to the ground-truth speech. We place ground truth speech among the synthetic speech and expect the participants with good hearing to choose the ground truth speech. We only counted the responses of 43 participants who chose the ground truth speech.

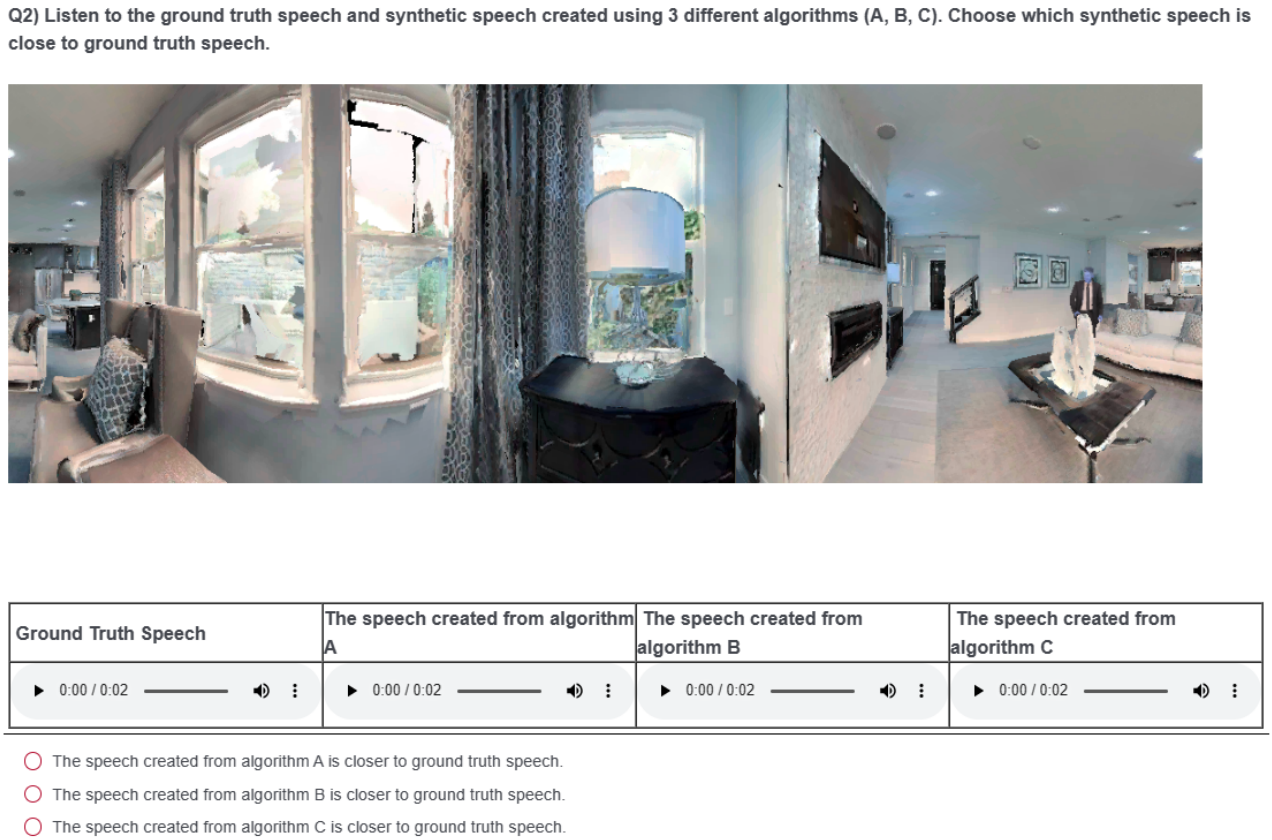

Figure 1. User study interface. We created 3 synthetic speech using our AV-RIR, Image2Reverb [18] and Visual Acoustic Matching [5] and asked the participants, which synthetic reverberant speech matches closely with ground-truth reverberant speech. For each question, we randomly shuffle the order of the synthetic reverberant speech from different approaches.

Out of 50 participants, 33 are male and 17 are female. The six participants are aged between 18-24 years, 28 participants are between 25-34 years and 16 participants are older than 34 years. Figure 1 shows the second question from our user study interface.

1.5. Societal Impact

Our model to estimate the RIR and enhance speech can have positive impacts on real-world applications. For example, the model can give an immersive experience in AR/VR applications and improve voice dubbing in movies. Also, our can be useful for different speech processing applications such as automatic speech recognition systems, telecommunication systems etc. We trained and evaluated our network on open-sourced publicly available datasets.

We got the certification and license to perform user studies from the Institutional Review Board and we followed their protocols. We did not collect any personal information from the participants.

References

[1] Audiocite.net: Livres audio gratuits mp3, 2023. 1

[2] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. The ami meeting corpus: A pre-announcement. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, page 28–39, Berlin, Heidelberg, 2005. Springer-Verlag. 1

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 2

[4] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigaton in 3d environments. In *ECCV*, 2020. 1, 2

[5] 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 (CVPR)*, pages 18858–18868, 2022. 2, 3

[6] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. In *NeurIPS 2022 Datasets and Benchmarks Track*, 2022. 2

[7] Changan Chen, Wei Sun, David Harwath, and Kristen Grauman. Learning audio-visual dereverberation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 2

[8] Sanjoy Chowdhury, Sreyan Ghosh, Subhrajyoti Dasgupta, Anton Ratnarajah, Utkarsh Tyagi, and Dinesh Manocha. Adverb: Visually guided audio dereverberation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7884–7896, 2023. 2

[9] Alexandre Défossez, Gabriel Synnaeve, and Yossi Adi. Real Time Speech Enhancement in the Waveform Domain. In *Proc. Interspeech 2020*, pages 3291–3295, 2020. 2

[10] 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. *ACM Trans. Graph.*, 37(4), 2018. 1, 2

[11] Szu-Wei Fu, Cheng Yu, Tsun-An Hsieh, Peter Plantinga, Mirco Ravanelli, Xugang Lu, and Yu Tsao. Metricgan+: An improved version of metricgan for speech enhancement. *arXiv preprint arXiv:2104.03538*, 2021. 2

[12] Vinay Kothapally and John H.L. Hansen. Complex-Valued Time-Frequency Self-Attention for Speech Dereverberation. In *Proc. Interspeech 2022*, pages 2543–2547, 2022. 2

[13] Vinay Kothapally and John H. L. Hansen. Skipconvgan: Monaural speech dereverberation using generative adversarial networks via complex time-frequency masking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1600–1613, 2022. 2

[14] Haohe Liu, Xubo Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, and Yuxuan Wang. VoiceFixer: A Unified Framework for High-Fidelity Speech Restoration. In *Proc. Interspeech 2022*, pages 4232–4236, 2022. 2

[15] Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang. Speech dereverberation based on variance-normalized delayed linear prediction. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1717–1731, 2010. 2

[16] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015. 2

[17] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A general-purpose speech toolkit, 2021. arXiv:2106.04624. 2

[18] Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2reverb: Cross-modal reverb impulse response synthesis. In *ICCV*, pages 286–295. IEEE, 2021. 3

[19] Jiaqi Su, Zeyu Jin, and Adam Finkelstein. HiFi-GAN: High-Fidelity Denoising and Dereverberation Based on Speech Deep Features in Adversarial Networks. In *Proc. Interspeech 2020*, pages 4506–4510, 2020. 2