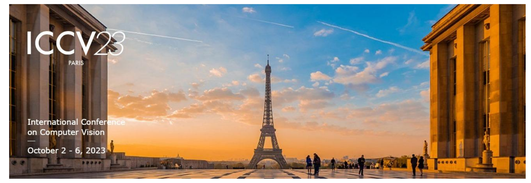
AdVerb: Visually Guided Audio Dereverberation

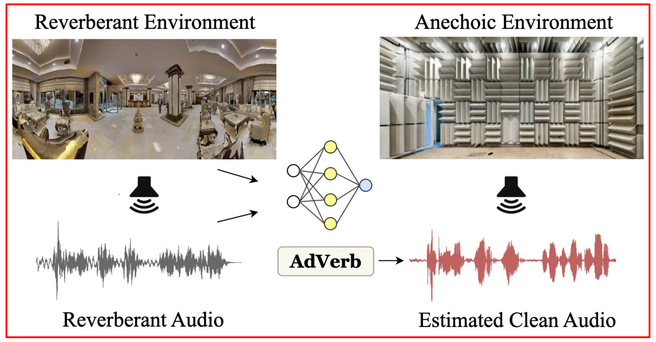
Sanjoy Chowdhury[1], Sreyan Ghosh[1*], Subhrajyoti Dasgupta[2], Anton Ratnarajah[1], Utkarsh Tyagi[1], Dinesh Manocha[1]*

[1] University of Maryland, College Park, [2] Mila and Université de Montréal



Introduction

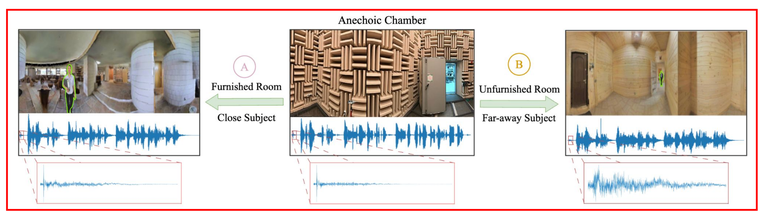
We present AdVerb, a novel audio-visual dereverberation framework that leverages visual cues of the environment to estimate clean audio from reverberant audio. E.g, given a reverberant sound produced in a large hall, our model attempts to remove the reverb effect to predict the anechoic or clean audio



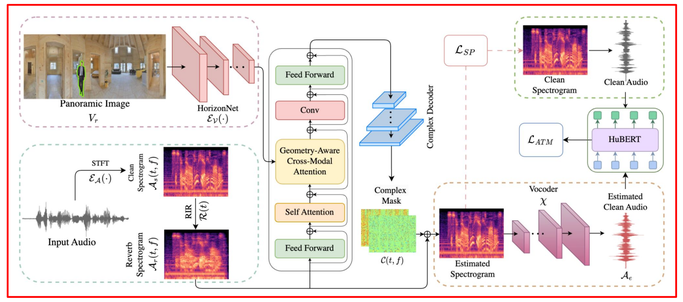
Main Contributions:

- We propose a novel cross-modal framework for dereverberating audio by exploiting complementary low-level visual cues and specially designed relative position embedding
- AdVerb employs a novel geometry-aware conformer network to capture 3D spatial semantic information to equip the network with salient vision cues through (Shifted) Window Blocks and Panoptic Blocks.
- Our architecture involves the prediction of complex ideal ratio mask and simultaneous optimization of two objective functions to estimate the dereverbed speech..
- On objective evaluation our approach significantly outperforms prior methods with a relative improvement in the range 18% - 82% on three downstream tasks: speech enhancement, speech recognition, and speaker verification. It also achieves highly satisfactory RT60 error scores.
- User study analysis reveals our method outperforms prior approaches on perceptual audio quality assessment.

Reverberation is a function of the speaker's relative position and the surrounding environment. The visual signals present critical details that determine the nature of the distortion. E.g,Ⓐin a relatively small furnished room when the speaker is nearby, reverb is less evident, whereas forⒷin a large hallway (especially when the speaker is far away) the reverb effect is very strong. The audio waveform illustrates the nature of reverberation, with the magnified section clearly depicting a stronger reverberation effect in case Ⓑ over Ⓐ.



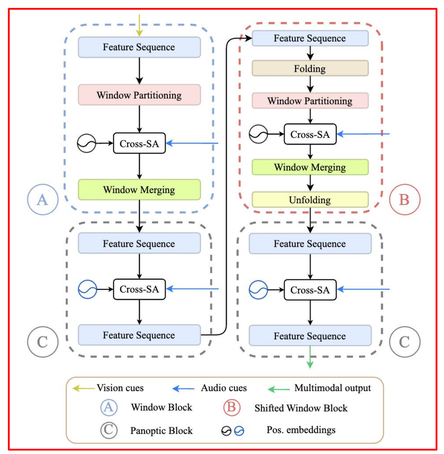
Method



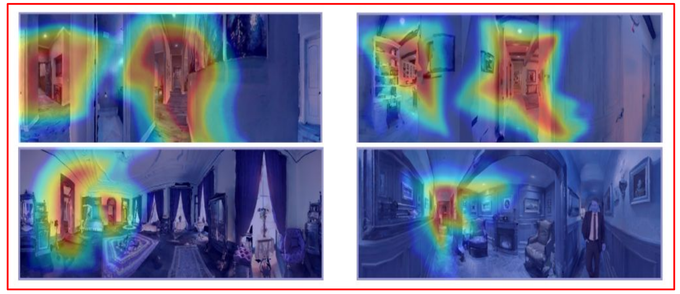
Overview of AdVerb. AdVerb estimates clean source audio from a reverberant speech signal leveraging two primary components: 1: The visual stream processing path comprises a HorizonNet-based backbone to obtain 1D feature sequences, which are subsequently passed to the cross-modal geometry-aware attention subnetwork. 2: The audio processing module applies STFT to get 2D spectrograms which are fed to the cross-modal encoder. The cross-attention subnetwork powered by geometry-aware (Shifted) Window Blocks, Panoptic Blocks, and Relative Position Embedding generates a complex ideal ratio mask.

Geometry-Aware Cross-Modal Attention

Overview of the Geometry-Aware Cross-Modal Attention block. Window and Panoptic Relative Position Embedding (RPE) are fused into Cross-modal Self-Attention (CSA) blocks. In Window Block Ⓐ, partitioning and merging of windows before and after CSA. In Ⓑ, Folding and Unfolding of sequence features before and after CSA, respectively.Ⓒ integrates another RPE to CSA.



Evaluation Results



Grad-CAM visualization of activated regions. Our model attends to regions that cause heavy reverberation effects.

Method	Speech Enhancement (SE)† PESQ ↑	Speech Recognition (SR)† WER (%) ↓	Speech Recognition (SR)† WER-FT (%) ↓	Speaker Verification (SV)† EER (%) ↓	Speaker Verification (SV)† EER-FT (%) ↓	RTE* ↓ (in sec)
Anechoic (Upper bound)	4.72	2.89	2.33	1.53	1.57	-
Reverberant	1.49	8.20	4.44	4.51	4.88	0.382
MetricGAN+ [18]†	2.45 (+64%)	7.48 (+9%)	4.86 (-9%)	4.67 (-4%)	2.85 (+42%)	0.187
HiFi-GAN [38]†	1.83 (+23%)	9.31 (-14%)	5.59 (-26%)	4.32 (+4%)	2.49 (+49%)	0.196
WPE [52]†	1.63 (+9%)	8.43 (-3%)	4.30 (+3%)	5.90 (-31%)	4.11 (+16%)	0.173
SkipConvGAN [40]†	2.10 (+41%)	7.22 (+12%)	4.17 (+6%)	4.86 (-8%)	3.98 (+18%)	0.119
VIDA [12]	2.37 (+59%)	4.44 (+46%)	3.66 (+18%)	3.97 (+12%)	2.40 (+51%)	0.155
ABLATION: AdVerb w/o Image	2.31 (+55%)	3.92 (+52%)	3.41 (+23%)	3.67 (+19%)	2.19 (+55%)	0.119
ABLATION: AdVerb w/ Random Image	2.54 (+70%)	4.12 (+50%)	3.62 (+18%)	3.76 (+17%)	2.26 (+54%)	0.110
ABLATION: AdVerb w/o ATM Loss	2.89 (+94%)	4.67 (+43%)	3.66 (+18%)	3.17 (+30%)	2.07 (+58%)	0.117
ABLATION: AdVerb w/o Complex SA	2.91 (+95%)	3.63 (+56%)	2.98 (+33%)	3.21 (+29%)	2.10 (+57%)	0.117
ABLATION: AdVerb w/o Geometry Aware Block	2.30 (+54%)	4.01 (+51%)	3.12 (+30%)	3.68 (+18%)	2.12 (+57%)	0.113
ABLATION: AdVerb w/o RPE	2.79 (+87%)	3.54 (+57%)	3.01 (+32%)	3.17 (+30%)	2.11 (+57%)	0.107
ABLATION: AdVerb w/o Window Block	2.81 (+89%)	3.61 (+56%)	2.99 (+33%)	3.14 (+30%)	2.12 (+57%)	0.108
ABLATION: AdVerb w/o Panoptic Block	2.92 (+96%)	3.59 (+56%)	2.92 (+34%)	3.29 (+27%)	2.01 (+59%)	0.107
AdVerb (ours)	**2.96 (+98%)**	**3.54 (+57%)**	**2.91 (+34%)**	**3.11 (+31%)**	**1.98 (+59%)**	**0.101**

*Comparison of AdVerb with various baselines on multiple spoken language processing tasks based on the LibriSpeech test-clean set (marked with †) and on sim-to-real transfer based on the AVSpeech dataset (marked with *).*

Baseline Method	SoundSpaces(in %) (A% / B% / C%)	AVSpeech(in %) (A% / B% / C%)
Clean Speech	**61.3** / 8.1 / 30.6	– / – / –
VIDA[12]	16.5 / 6.5 / **77.0**	13.5 / 0.0 / **86.5**
WPE[52]	8.8 / 3.5 / **87.7**	3.7 / 7.4 / **88.9**
SCGAN[40]	9.2 / 0.0 / **90.8**	0.0 / 8.0 / **92.0**

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**

User study results. A% of participants find the baseline audio samples better, B% have no preference, and C% prefer AdVerb. As evident, Users find samples from AdVerb to be perceptually better and cleaner when compared against prior methods

For more details

Paper

Webpage