LEARNING AUDIO-VISUAL DEREVERBERATION

Changan Chen\textsuperscript{1,2}  Wei Sun\textsuperscript{1}  David Harwath\textsuperscript{1}  Kristen Grauman\textsuperscript{1,2}  
\textsuperscript{1}UT Austin  \textsuperscript{2}FAIR, Meta AI

ABSTRACT

Reverberation not only degrades the quality of speech for human perception, but also severely impacts the accuracy of automatic speech recognition. Prior work attempts to remove reverberation based on the audio modality only. Our idea is to learn to dereverberate speech from audio-visual observations. The visual environment surrounding a human speaker reveals important cues about the room geometry, materials, and speaker location, all of which influence the precise reverberation effects. We introduce Visually-Informed Dereverberation of Audio (VIDA), an end-to-end approach that learns to remove reverberation based on both the observed monaural sound and visual scene. In support of this new task, we develop a large-scale dataset SoundSpaces-Speech that uses realistic acoustic renderings of speech in real-world 3D scans of homes offering a variety of room acoustics. Demonstrating our approach on both simulated and real imagery for speech enhancement, speech recognition, and speaker identification, we show it achieves state-of-the-art performance and substantially improves over audio-only methods.

1. INTRODUCTION AND RELATED WORK

Audio reverberation occurs when multiple reflections from surfaces and objects in the environment build up then decay, altering the original audio signal. While reverberation bestows a realistic sense of spatial context, it also can degrade a listener’s experience. In particular, the quality of human speech is greatly affected by reverberant environments—as illustrated by how difficult it can be to parse the words of a family member speaking loudly from another room in the house, a tour guide describing the artwork down the hall of a magnificent cavernous cathedral, or a colleague participating in a Zoom call from a café. Similarly, automatic speech recognition (ASR) suffers when given reverberant speech input [1, 2]. Thus there is great need for dereverberation algorithms that can strip away reverberant effects for speech enhancement, recognition, and other downstream tasks, which could in turn benefit many applications such as teleconferencing, assistive hearing devices, and augmented reality.

The audio community has made steady progress on speech dereverberation [3, 4, 5]. While past approaches have tackled the problem with signal processing and statistical techniques [6, 7], many modern approaches use neural networks to learn a mapping from reverberant to clean spectrograms [3, 8]. To our knowledge, all existing models for dereverberation rely purely on audio. This underconstrains the dereverberation task since the latent parameters of the recording space are not discernible from the audio alone.\textsuperscript{1}

However, we observe that in many practical settings of interest—video conferencing, augmented reality, Web video indexing—reverberant audio is naturally accompanied by a visual (video) stream. Importantly, the visual stream offers valuable cues about the room acoustics affecting reverberation: where are the walls, how are they shaped, where is the human speaker, what is the layout of major furniture, what are the room’s dominant materials (which affect absorption, reflection, and refraction), and even what is the facial appearance and/or body shape of the person speaking (which affects the acoustic properties of a person’s speech, and reverberation time is frequency dependent). For example, reverberation is typically stronger when the speaker is further away; speech is more reverberant in a large church or hallway; heavy carpet absorbs more sound. See Fig. 1. While some recent works explore acoustic modeling using images [10, 11, 12, 13], no prior work has investigated how to leverage visual-acoustic cues for dereverberation.

Our idea is to learn to dereverberate speech from audio-visual observations (Fig. 1). In this task, the input is reverberant speech and visual observations of the environment surrounding the human speaker, and the output is a prediction of the clean source audio. To tackle this problem, there are two key technical challenges. The first is how to model the multi-modal dereverberation process in order to infer the latent clean audio. The second is how to secure appropriate training data spanning a variety of physical environments for which we can sample speech with known ground truth (non-reverberant, anechoic) audio. The latter is non-trivial because ordinary audio/video recordings are themselves corrupted by reverberation but lack the ground truth source signal we wish to recover.

For the modeling challenge, we introduce an end-to-end approach called Visually-Informed Dereverberation of Audio (VIDA). VIDA consists of a Visual Acoustics Network that learns reverberation properties of the room geometry, object locations, and speaker position. Coupled with a multi-modal UNet dereverberation module, it learns to remove the reverberations from a single-channel audio stream. In addition, we propose an audio-visual (AV) matching loss to enforce consistency between the visually-inferred reverberation features and those inferred from the audio signal. We leverage the outputs of our model for three downstream tasks: speech enhancement, speech recognition, and speaker verification.

To address the training data challenge, we develop SoundSpaces-Speech, a new large-scale dataset based on SoundSpaces [14], a 3D simulator for real-world scanned environments that allows both visual and acoustic rendering. We insert “clean” audio voices together with

\textsuperscript{1}Our extended manuscript [9] gives more details about related work.

---

Fig. 1: The goal of audio-visual dereverberation is to leverage the visual observation of the environment to improve speech enhancement.
We anticipate that these cues can inform an estimate of the room audio-visual dereverberation and depth image acoustics, and thus the clean source waveform. Given the RGB information that is implicit in the visual stream—as well as the location of the environment’s geometry and material composition—our model, and it has the added benefit of allowing controlled studies that vary the parameters of the capture setting. As we will show, the data also supports sim2real transfer.

We first train and evaluate our model on 82 large-scale environments—each a multi-room home containing a variety of objects—coupled with speech samples from the LibriSpeech dataset [15], then further test with real-world data. We consider both near-field and far-field settings where the human speaker is in-view or quite far from the camera, respectively. The proposed model outperforms methods restricted to the audio stream, improves the state of the art for multiple speech tasks, and transfers successfully to real-world data.

Our main contributions are to 1) present the task of audio-visual dereverberation, 2) address it with a new multi-modal modeling approach and a novel reverber-visual matching loss, 3) provide a benchmark built on both SoundSpaces-Speech and real data, and 4) demonstrate the utility of AV dereverberation for multiple practical tasks. Overall, our work shows the potential for speech dereverberation to benefit from seeing the 3D environment.

2. THE AUDIO-VISUAL DEREVERBERATION TASK

We introduce the novel task of audio-visual dereverberation. In this task, a speaker (or other sound source) and a listener are situated in a 3D environment, such as the interior of a house. The speaker—whose location is unknown to the listener—produces a speech waveform $A_s$. A superposition of the direct sound and the reverber is captured by the listener, denoted $A_r$. The reverberant speech $A_r$ can be modeled as the convolution of the anechoic source waveform $A_s$ with the room impulse response (RIR) $R$, i.e. $A_r(t) = A_s(t) * R(t)$ [16]. $R$ is a function of the environment’s geometry, its materials, and the positioning of the speaker and the listener. It is possible in principle to measure the RIR $R$ for a real-world environment, but doing so can be impractical when the source and listener are able to move around or we must cope with different environments. Furthermore, in the common scenario where we want to process video captured in environments to which we have no physical access, measuring the RIR is simply impossible.

We consider an alternative source of information about the environment: vision. We assume the listener observes its surroundings from an RGB-D camera or an RGB camera coupled with single-image depth estimation [17]. Intuitively, we should be able to leverage the information about the environment’s geometry and material composition that is implicit in the visual stream—as well as the location of the speaker (if visible)—to estimate its reverberant characteristics. We anticipate that these cues can inform an estimate of the room acoustics, and thus the clean source waveform. Given the RGB $I_r$ and depth image $I_d$ captured by the listener from its current vantage point, the task is to predict the source waveform $A_s$ from the images and reverberant audio: $A_s(t) = f_{	heta}(I_r, I_d, A_r(t))$. This setting represents common real-world scenarios previously discussed, and poses new challenges for speech enhancement and recognition.

3. DATASET CURATION

For the proposed task, obtaining the right training data is itself a challenge. Existing video data contains reverberant audio but lacks the ground truth anechoic audio signal, and existing RIR datasets [18, 19] do not have images paired with the microphone position. We introduce both real and simulated datasets to enable reproducible research on audio-visual de- reverberation.

3D environments and acoustic simulator. First we introduce a large-scale dataset in which we couple real-world visual environments with state-of-the-art acoustic simulations accurately capturing the environments’ spatial effects on real samples of recorded speech. We want our dataset to allow control of a variety of physical environments, the positions of the listener/camera and sources, and the speech content—all while maintaining both the observed reverberant $A_r(t)$ and ground truth anechoic $A_s(t)$ sounds. To this end, we leverage the audio-visual simulator SoundSpaces [14], which provides precomputed RIRs $R(t)$ on a uniform grid of resolution 1 m for the real-world environment scans in Replica [20] and Matterport3D [21]. The SoundSpaces RIRs account for all the major real-world features: direct sounds, early specular/diffuse reflections, reverberations, binaural spatialization, and frequency-dependent effects from materials and air absorption. We use all 82 Matterport environments due to their greater scale and complexity (on average 517 m$^2$ per environment).

SoundSpaces-Speech. We extend SoundSpaces to construct reverberant speech. As the source speech corpus we use LibriSpeech [15], which contains 1,000 hours of 16kHz English speech from audio books. We use the standard train/test splits for both datasets, such that neither the houses nor speaker voices observed at test time are ever observed during training.

For each source utterance, we randomly sample a source-receiver location pair in a random environment, then convolve the speech waveform $A_s(t)$ with the associated SoundSpaces RIR $R(t)$ to obtain the reverberant $A_r(t)$. To augment the visual scene, we insert a 3D humanoid of the same gender as the real speaker at the speaker location and render RGB and depth images at the listener location. We consider two types of image: panorama and normal field of view (FoV), with dimension 192 x 756 and 384 x 256 respectively. See Fig. 3. We generate 49,430/2,700/2,600 such samples for the train/val/test splits, respectively.

Real data collection. To explore whether models trained in simulation can also work in the real world, we also collect a set of real images and speech recordings while preserving the ground truth anechoic audio by playing utterances through a loudspeaker. We collect data from varying environments—auditoriums, meeting rooms, atriums, corridors, and classrooms. For each environment, we vary the speaker location from near-field to mid-field to far-field. During data collection, the microphone also records ambient sounds like people chatting, door opening, AC humming,
We propose the Visually-Informed Dereverberation of Audio (VIDA) model, which leverages visual cues to learn representations of the environmental acoustics and sound source locations to dereverberate audio. VIDA consists of two main components (Figure 4): 1) a Visual Acoustics Network (VAN), which learns to map RGB-D images of the environment to features useful for dereverberation, and 2) the dereverberation module itself, which is based on a UNet encoder-decoder architecture.

**Visual Acoustics Network.** Visual observations of a scene reveal information about room acoustics, including room geometry, materials, object locations, and the speaker position. We devise the VAN to capture all these cues into a latent embedding vector, which is subsequently used to remove reverb. This network takes as its input an RGB image $I_r$ and a depth image $I_d$, capturing material and geometry information respectively. We use two separate ResNet18 [22] networks to extract their features, i.e. $e_r = f_r(I_r)$ and $e_d = f_d(I_d)$. We concatenate $e_r$ and $e_d$ channel-wise and feed the result to a 1x1 convolution layer to reduce the number of total channels to 512 followed by a subsequent pooling layer to reduce the spatial dimension, resulting in the output vector $e_v = f_l(f_c([e_r; e_d]))$.

**Dereverberation Network.** To recover the clean speech audio, we use the UNet [23] architecture, a fully convolutional network often used for image segmentation. We first use the Short-Time Fourier Transform (STFT) to convert the reverberant input audio $A_r$ to a spectrogram $S_r$ with two channels representing the log-magnitude and phase angle, respectively. Our UNet takes spectrograms of a fixed size of $256 \times 256$ as input, but in general the duration of the speech audio we wish to dereverberate will be variable. Therefore, the model processes the full input spectrogram using a series of overlapping, sliding windows. Specifically, we segment the spectrogram along the time dimension into a sequence of fixed-size chunks $S^{seg} = \{S_1^{seg}, S_2^{seg}, ..., S_N^{seg}\}$ using a sliding window of length $s$ frames and 50% overlap between consecutive windows to avoid boundary artifacts. To derive the ground-truth target spectrograms used in training, we perform the exact same segmentation operation on the clean source audio $A_c$, to obtain $S^{seg} = \{S_1, S_2, ..., S_N\}$.

During training, when a particular waveform $S_i$ is selected for inclusion in a data batch, we randomly sample one of its segments $S_{i}^{seg}$ to be the input to the model, and choose the corresponding $S_i$ as the target. We first compute the output of the VAN, $e_{c_i}$, for the environment image associated with $S_i$. Next, $S_{i}^{seg}$ is fed to the UNet’s encoder to extract the intermediate feature map $e_{x_i} = f_{enc}(S_{i}^{seg})$. We then spatially tile and concatenate $e_{x_i}$ channel-wise with $e_{c_i}$, and feed the fused features to the UNet decoder, which predicts the source spectrogram segment $S_{i}^{pred} = f_{dec}(e_{c_i}; e_{x_i})$.

**Spectrogram prediction loss.** Our training objective is to minimize the $L2$ distance between the predicted and ground-truth spectrograms, treating the magnitude and phase separately. For a given predicted spectrogram segment $S_{i}^{pred}$, let $M_{i}^{pred}$ denote the predicted log-magnitude spectrogram, $P_{i}^{pred}$ denote the predicted phase spectrogram, and $M_{i}$ and $P_{i}$ denote the respective ground-truth log-magnitude and phase spectrograms. We define the magnitude loss as:

$$L_{magnitude} = ||M_{i}^{pred} - M_{i}||_2.$$

To address the issue of phase wraparound, we map the phase angle to its corresponding rectangular coordinates on the unit circle and then compute the loss for the phase:

$$L_{phase} = ||\sin(P_{i}^{pred}) - \sin(P_{i})||_2 + ||\cos(P_{i}^{pred}) - \cos(P_{i})||_2.$$

**Reverb-visual matching loss.** To reinforce the consistency between the visually-inferred room acoustics and the reverberation characteristics learned by the UNet encoder, we also employ a contrastive reverb-visual matching loss:

$$L_{matching}(e_{c}, e_{x}, e_{x}^{u}) = \max\left\{d(f_n(e_{c}), f_n(e_{x})) - d(f_n(e_{c}), f_n(e_{x}^{u}) + m, 0]\right\}.

Here, $d(x, y)$ represents L2 distance, $f_n(\cdot)$ applies L2 normalization, $m$ is a margin, and $e_{x}^{u}$ is a different speech embedding sampled from the same data batch. This loss forces the embeddings output by the VAN and the UNet encoder to be consistent, which we empirically show to be beneficial.

**Training.** Our overall training objective is:

$$L_{total} = L_{magnitude} + \lambda_1 L_{phase} + \lambda_2 L_{matching},$$

where $\lambda_1$ and $\lambda_2$ are weighting factors for the phase and matching losses. To augment the data, we further rotate the image view for a random angle for each input during training. This is possible because our audio recording is omni-directional and is independent of camera pose. This data augmentation strategy prevents the model from overfitting; without it our model fails to converge. It creates a one-to-many mapping between reverb and views, forcing the model to learn a viewpoint-invariant representation of room acoustics.

**Testing.** At test time, we wish to re-synthesize the entire clean waveform instead of a single fixed-length segment. In this case, we feed all of the segments for a waveform $S_i$ into the model and temporally concatenate all of the output segments. Finally, to re-synthesize the waveform we use the Griffin-Lim algorithm [24] to iteratively improve the predicted phase for 30 iterations, which we find works better than directly using the predicted phase or using Griffin-Lim with a randomly initialized phase. Please see our extended arXiv manuscript for more details [9].

5. EXPERIMENTS

We evaluate our model by dereverberating speech for three downstream tasks: speech enhancement (SE), automatic speech recognition (ASR), and speaker verification (SV). We evaluate using both real scanned Matterport3D environments with simulated audio as well as real-world data collected with a camera and mic.

**Evaluation tasks and metrics.** We report the standard metrics Perceptual Evaluation of Speech Quality (PESQ) [27], Word Error Rate (WER), and Equal Error Rate (EER) for the three tasks, respectively. For ASR and SV, we use pretrained models from SpeechBrain [28]. We evaluate these models off-the-shelf on our (de)reverberated version of the LibriSpeech test-clean set, and also explore finetuning the model on the (de)reverberated LibriSpeech train-clean-360 data to ensure all models have exposure to reverberant speech when training. For speaker verification, we construct a set of 80k sampled utterance pairs consisting of different rooms, mic placements, and genders to account for session variability, similar to [29].
To study how much VIDA leverages visual signals, we evaluated its performance with different visual ablations. Table 1 shows the results for LibriSpeech test-clean set that is reverberated with our environmental simulator (with the exception of the “Anechoic (Upper bound)” setting, which is evaluated on the original audio). FT refers to tests where the models are finetuned with the audio-enhanced data.

|                      | SE PESQ | ASR WER | SV EER |
|----------------------|---------|---------|--------|
| Anechoic (Upper bound) | 4.64    | 2.50    | 1.62   |
| Reverberant           | 1.54    | 8.86    | 4.69   |
| MetricGAN+ [25]       | 2.33    | 7.49    | 4.67   |
| HiFi-GAN [26]         | 1.83    | 9.31    | 5.59   |
| WPE [6]               | 1.63    | 8.18    | 4.30   |
| VIDA w/o VAN           | 2.32    | 4.92    | 3.76   |
| VIDA w/ normal FoV    | 2.33    | 4.85    | 3.73   |
| VIDA w/o matching loss| 2.38    | 4.59    | 3.72   |
| VIDA w/o human mesh   | 2.31    | 4.57    | 3.72   |
| VIDA w/ random image  | 2.34    | 4.94    | 3.82   |
| VIDA                  | 2.37    | 4.44    | 3.66   |

Table 1: Results on LibriSpeech test-clean set that is reverberated with our environmental simulator (with the exception of the “Anechoic (Upper bound)” setting, which is evaluated on the original audio). FT refers to tests where the models are finetuned with the audio-enhanced data.

Baseline models. In addition to evaluating the anechoic and reverberant audio (with no enhancement), we compare against multiple baseline dereverberation models: 1. MetricGAN+ [25]: a state-of-the-art model for speech enhancement, we use the public implementation from SpeechBrain [28], 2. HiFi-GAN [26]: a recent model for denoising and dereverberation, 3. WPE [6]: A widely used statistical speech dereverberation model.

Results on SoundSpaces-Speech. Table 1 shows the results for all models on SE, ASR, and SV. First, since existing methods report results on anechoic audio, we note the pretrained SpeechBrain model applied to anechoic audio (first row) yields errors competitive with the SoTA [30], meaning we have a solid experimental testbed. Comparing the results on anechoic vs. reverberated speech, we see that reverberation significantly degrades performance on all tasks. Our VIDA model outperforms all other models, and by a large margin on the ASR and SV tasks. The results are statistically significant according to a paired t-test. After finetuning the ASR model, the gain is still largely preserved at 0.64% WER (14.88% relative), although it is important to note that finetuning downstream models on enhanced speech is not always feasible, e.g., if using off-the-shelf ASR. Our results demonstrate that learning the acoustic properties from visual signals is very helpful for dereverberating speech, enabling the model to leverage information unavailable in the audio alone.

Ablations. To study how much VIDA leverages visual signals, we ablate the visual network VAN (audio-only). Table 1 shows the results. All performance degrades significantly, showing that visual acoustic features are helpful for dereverberation. To understand how well VIDA works with a normal field-of-view (FoV) camera, we replace the panorama image input with a FoV of 80 degrees randomly sampled from the current view. All metrics drop compared to using a panorama, as expected. Compared to the audio-only ablation, however, VIDA still performs better; even a partial view of the environment helps the model understand the scene and dereverberate the audio. Next, we ablate the proposed reverber-visual matching loss (“w/o matching loss”). Without it, VIDA’s performance declines on all metrics. This shows by forcing the visual feature to agree with the reverberation feature, our model learns a better representation of room acoustics. To examine how much the model leverages the human speaker cues and uses the visual scene, we evaluate VIDA on the same test data but with the 3D humanoid removed (“w/o human mesh”) or train VIDA with random images (“w/ random image”) and re-evaluate. All three metrics become worse. This shows our model pays attention to both the presence of the human speaker and the scene geometry to better anticipate reverberation.

Table 3: Breakdown of word error rate (WER) for VIDA without and with VAN on real test data.

| Environment | SE PESQ | ASR FT | SV EER | Corridor FT |
|-------------|---------|--------|--------|-------------|
| Near-field  | 11.2    | 4.24   | 1.54   | 1.44        |
| Mid-field   | 21.8    | 18.39  | 6.5    | 5.9         |
| Far-field   | 52.4    | 50.5   | 6.7    | 5.9         |

Table 2: Results on real data demonstrating sim2real transfer.

6. CONCLUSION

We introduced the novel task of audio-visual dereverberation. The proposed VIDA approach learns to remove reverb by attending to both the audio and visual streams, recovering valuable signals about room geometry, materials, and speaker locations from visual encodings of the environment. In support of this task, we develop a large-scale dataset providing realistic, spatially registered observations of speech and 3D environments. VIDA successfully dereverberates novel voices in novel environments more accurately than an array of baselines, improving multiple downstream tasks. In future work, we will explore temporal models for dereverberation with real-world video.

Acknowledgements: UT Austin is supported in part by the IFML NSF AI Institute and NSF CCRI. KG is paid as a researcher at Meta.
7. REFERENCES

[1] Keisuke Kinoshita, Marc Delcroix, Sharon Gannot, Emanuël Habets, Reinhold Haeb-Umbach, Walter Kellermann, Volker Leutnant, Roland Maas, Tomohiro Nakatani, Bhiksha Raj, Armin Sehr, and Takuya Yoshioka, “A summary of the reverb challenge: State-of-the-art and remaining challenges in reverberant speech processing research,” EURASIP Journal on Advances in Signal Processing, 2016.

[2] Igor Szöke, Miroslav Skácel, Ladislav Mošner, Jakub Paliesek, and Jan Cernocky, “Building and evaluation of a real room impulse response dataset,” IEEE Journal of Selected Topics in Signal Processing, 2019.

[3] Kun Han, Yuxuan Wang, DeLiang Wang, William S. Woods, Ivo Merks, and Tao Zhang, “Learning spectral mapping for speech dereverberation and denoising.” in IEEE Trans Audio Speech Lang Process, 2015.

[4] Bo Wu, Kehuang Li, Minglei Yang, and Chin-Hui Lee, “A dereverberation-time-aware approach to speech dereverberation based on deep neural networks,” in IEEE Trans Audio Speech Lang Process, 2016.

[5] Yan Zhao, Zhong-Qiu Wang, and DeLiang Wang, “Two-stage deep learning for noisy-reverberant speech enhancement,” IEEE Trans Audio Speech Lang Process, 2019.

[6] Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang, “Speech dereverberation based on variance-normalized delayed linear prediction.” in IEEE Trans Audio Speech Lang Process, 2010.

[7] Patrick A. Naylor and Nikolay D. Gaubitch, Speech Dereverberation, Springer, 1st edition, 2010.

[8] Szu-Wei Fu, Chien-Feng Liao, Yu Tsao, and Shou-De Lin, “MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement,” in ICML, 2019.

[9] Chengan Chen, Wei Sun, David Harwath, and Kristen Grauman, “Learning audio-visual dereverberation.” arXiv:2106.07732, 2021.

[10] Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman, “Few-shot audio-visual learning of environment acoustics,” in NeurIPS, 2022.

[11] Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman, “Visual acoustic matching,” in CVPR, 2022.

[12] Andrew Luo, Yilun Du, Michael J Tarr, Joshua B Tenenbaum, Antonio Torralba, and Chuang Gan, “Learning neural acoustic fields,” in NeurIPS, 2022.

[13] Hansung Kim, Luca Remaggi, Sam Fowler, Philip JB Jackson, and Adrian Hilton, “Acoustic room modelling using 360 stereo cameras,” IEEE Transactions on Multimedia, 2020.

[14] 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.

[15] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “LibriSpeech: An asr corpus based on public domain audio books,” in ICASSP, 2015.

[16] Stephen T. Neely and Jont B. Allen, “Invertibility of a room impulse response,” in Journal of the Acoustical Society of America, 1979.

[17] Clément Godard, Osin Mac Aodha, Michael Firman, and Gabriel J. Brostow, “Digging into self-supervised monocular depth prediction,” in ICCV, 2019.

[18] Peter Vary Marco Jeub, Magnus Schafer, “A binaural room impulse response database for the evaluation of dereverberation algorithms,” in Proceedings of International Conference on Digital Signal Processing, 2009.

[19] Damian T Murphy and Simon Shelley, “Openair: An interactive auralization web resource and database,” in Audio Engineering Society Convention 129, 2010.

[20] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shibhbit Verma, et al., “The replica dataset: A digital replica of indoor spaces,” arXiv:1906.05797, 2019.

[21] 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,” 3DV, 2017. MatterPort3D dataset license available at: http://kaldir.vc.in.tum.de/matterport/MP_TOS.pdf.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in CVPR, 2016.

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical image computing and computer-assisted intervention, 2015.

[24] Daniel Griffin and Jae Lim, “Signal estimation from modified short-time fourier transform,” IEEE Transactions on acoustics, speech, and signal processing, vol. 32, no. 2, pp. 236–243, 1984.

[25] 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:2104.03538, 2021.

[26] Jiabi Su, Zeyu Jin, and Adam Finkelstein, “HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks,” in INTERSPEECH, 2020.

[27] Antony W. Rix, John G. Beerends, Michael P. Hollier, and Andrés P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephony networks and codecs,” in ICASSP, 2001.

[28] Mirco Ravanelli, Titouan Parcollet, Aku Rouhe, Peter Plantinga, Elena Rastorgueva, Loren Lugosch, Nauman Dawalatabad, Chou Ju-Chieh, Abdel Heba, Francois Grondin, William Ariis, Chien-Feng Liao, Samuele Cornell, Sung-Lin Yeh, Hwidong Na, Yan Gao, Szu-Wei Fu, Cem Subakan, Renato De Mori, and Yoshua Bengio, “Speechbrain.” https://github.com/speechbrain/speechbrain, 2021.

[29] Colleen Richey, Maria A. Barrios, Zeb Armstrong, Chris Bartels, Horacio Franco, Martin Graciarena, Aaron Lawson, Mahesh Kumar Nadwanda, Allen Stauffer, Julian van Hout, Paul Gamble, Jeff Hetherly, Cory Stephenson, and Karl Ni, “Voices obscured in complex environmental settings (voices) corpus,” arXiv:1804.05053, 2018.

[30] Anmol Gulati, Chung-Cheng Chiu, James Qin, Jiuhui Yu, Nikh Parmar, Ruomeng Pang, Shibo Wang, Wei Han, Yonghui Wu, Yu Zhang, and Zhengdong Zhang, “Conformer: Convolution-augmented transformer for speech recognition,” in Interspeech, 2020.