Visual Acoustic Matching

Changan Chen[1,4] Ruohan Gao[2] Paul Calamia[3] Kristen Grauman[1,4]
[1]University of Texas at Austin [2]Stanford University [3]Reality Labs Research at Meta [4]Meta AI

Abstract

We introduce the visual acoustic matching *task, in which an audio clip is transformed to sound like it was recorded in a target environment. Given an image of the target environment and a waveform for the source audio, the goal is to re-synthesize the audio to match the target room acoustics as suggested by its visible geometry and materials. To address this novel task, we propose a cross-modal transformer model that uses audio-visual attention to inject visual properties into the audio and generate realistic audio output. In addition, we devise a self-supervised training objective that can learn acoustic matching from in-the-wild Web videos, despite their lack of acoustically mismatched audio. We demonstrate that our approach successfully translates human speech to a variety of real-world environments depicted in images, outperforming both traditional acoustic matching and more heavily supervised baselines.*

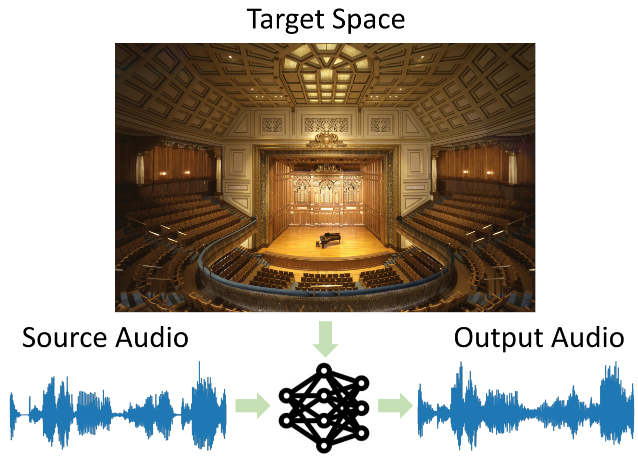

Figure 1. Goal of visual acoustic matching: transform the sound recorded in one space to another space depicted in the target visual scene. For example, given source audio recorded in a studio, re-synthesize that audio to match the room acoustics of a concert hall.

1. Introduction

The audio we hear is always transformed by the space we are in, as a function of the physical environment's geometry, the materials of surfaces and objects in it, and the locations of sound sources around us. This means that we perceive the same sound differently depending on where we hear it. For example, imagine a person singing a song while standing on the hardwood stage in a spacious auditorium versus in a cozy living room with shaggy carpet. The underlying song content would be identical, but we would experience it in two very different ways.

For this reason, it is important to model room acoustics to deliver a realistic and immersive experience for many applications in augmented reality (AR) and virtual reality (VR). Hearing sounds with acoustics *inconsistent* with the scene is disruptive for human perception. In AR/VR, when the real space and virtually reproduced space have different acoustic properties, it causes a cognitive mismatch and the "room divergence effect" damages the user experience [63].

Creating audio signals that are consistent with an environment has a long history in the audio community. If the geometry (often in the form of a 3D mesh) and material properties of the space are known, simulation techniques can be applied to generate a room impulse response (RIR), a transfer function between the sound source and the microphone that describes how the sound gets transformed by the space. RIRs can then be convolved with an arbitrary source audio signal to generate the audio signals received by the microphone [8, 9, 17, 50, 51]. In the absence of geometry and material information, the acoustical properties can be estimated blindly from audio captured in that room (e.g., reverberant speech), then used to auralize a signal [29,42,56]. However, both approaches have practical limitations: the former requires access to the full mesh and material properties of the target space, while the latter gets only limited acoustic information about the target space from the reverberation in the audio sample. Neither uses imagery of the target scene to perform acoustic matching.

We propose a novel task: *visual acoustic matching*. Given an image of the target environment and a source audio clip, the goal is to re-synthesize the audio as if it were recorded in the target environment (see Figure 1). The idea is to transform sounds from one space to another space by altering their scene-driven acoustic signatures. Visual acoustic matching has many potential applications, including smart video editing where a user can inject sounding objects into new backgrounds, film dubbing to make a different actor's voice sound appropriate for the movie scene, audio enhancement for video conference calls, and audio synthesis for AR/VR to make users feel immersed in the visual space displayed to them.

To address visual acoustic matching, we introduce a cross-modal transformer model together with a novel self-supervised training objective that accommodates in-the-wild Web videos having unknown room acoustics.

Our approach accounts for two key challenges: how to faithfully model the complex cross-modal interactions, and how to achieve scalable training data. Regarding the first challenge, different regions of a room affect the acoustics in different ways. For example, reflective glass leads to longer reverberation in high frequencies while absorptive ceilings reduce the reverberation more quickly. Our model provides fine-grained audio-visual reasoning by attending to regions of the image and how they affect the acoustics. Furthermore, to capture the fine details of reverberation effects—which are typically much smaller in magnitude than the direct signal—we use 1D convolutions to generate time-domain signals directly and apply a multi-resolution generative adversarial audio loss.

Regarding the second key challenge, one would ideally have *paired* training data consisting of a sound sample not recorded in the target space plus its proper acoustic rendering for the scene shown in the target image, i.e., a source and target audio for each visual scene in the training set. However, such a strategy requires either physical access to the pictured environments, or knowledge of their room impulse response functions—either of which severely limits the source of viable training data. Meanwhile, though a Web video does exhibit strong correspondence between its visual scene and the scene acoustics, it offers only the audio recorded in the target space. Accounting for these tradeoffs, we propose a self-supervised objective that automatically creates acoustically mismatched audio for training with Web videos. The key insight is to use dereverberation and acoustic randomization to alter the original audio's acoustics while preserving its content.

We demonstrate our approach on challenging real-world sounds and environments, as well as controlled experiments with realistic acoustic simulations in scanned scenes. Our quantitative results and subjective evaluations via human studies show that our model generates audio that matches the target environment with high perceptual quality, outperforming a state-of-the-art model that has heavier supervision requirements [52] as well as traditional acoustic matching models.

2. Related Work

Acoustic matching. The goal of *acoustic matching* is to transform an audio recording made in one environment to sound as if it were recorded in a target environment. The audio community deals with this task with various approaches depending on what information about the target environment is accessible. If audio recorded in the target environment is provided, blind estimation of two acoustic parameters, direct-to-reverberant ratio (DRR), which describes the energy ratio of direct arrival sound and reflected sound, and reverberation time (RT60), the time it takes for a sound to decay 60dB, is sufficient to create simple RIRs that yield plausibly matched audio [15, 18, 29, 38, 42, 65]. Blind estimation of the room impulse response from reverberant speech has also been explored [54, 62]. In music production, acoustic matching is applied to change the reverberation to emulate that of a target space or processing algorithm [33,49]. Recent work conditions the target-audio generation on a low-dimensional audio embedding [56]. Unlike any of the above, we introduce and tackle the *visual* acoustic matching problem, where the target environment is expressed via an input image.

Visual understanding of room acoustics. The room impulse response (RIR) is the (time-domain) transfer function capturing the room acoustics for arbitrary source stimuli given specific source and receiver/listener positions in an environment. Convolving an RIR with a sound waveform yields the sound of that source in the context of the particular physical space. RIRs are traditionally measured with special equipment in the room itself [26, 53] or simulated with sound propagation models [5, 11, 43]. Recent work explores estimating an RIR from an input image [31, 52], which requires access to paired image and impulse response training data. While video recordings provide a natural source for learning the correspondence between space (captured by the visual stream) and acoustics (captured by the audio stream), they have not been explored in the literature. We show how to leverage Web video data for understanding room acoustics in a self-supervised fashion, obviating the need for expensive paired RIR-image training data. Our results demonstrate the advantages.

Audio-visual learning. Recent advances in multi-modal video understanding enable new forms of self-supervised cross-modal feature learning from video [6, 34, 41], object localization [28], and audio-visual speech enhancement and source separation [1, 2, 12, 16, 27, 40, 44, 48, 67, 69]. Work in embodied AI explores acoustic simulations with real visual scans to study audio-visual navigation tasks [11,14,19], where an agent moves intelligently based on the visual and auditory observations. However, no prior work investigates the visual acoustic matching task as we propose.

Multimodal fusion. One standard solution for audio-visual feature fusion is to represent audio as spectrograms, a matrix representation of the spectrum of frequencies of a signal as it varies with time, process them with a CNN, and concatenate with visual features from another CNN [11, 16, 20, 21, 44]. This fusion strategy is limited by using one global feature to represent the scene and thus supports only coarse-grained reasoning. The transformer [60] has proven to be a power tool in vision [22, 30]. Its self-attention operation provides a natural mechanism to fuse high-dimensional signals of different sensory modalities, and it has been used in various tasks such as action recognition [7], self-supervised learning [4, 6, 46], and language modeling [24]. Audio-visual attention [36, 57, 58] has been recently studied to capture the correlation between visual features and audio features. We use cross-modal attention for learning how different regions of the image contribute to reverberation. We show that compared with the conventional concatenation-based fusion, the proposed model predicts acoustics from images more accurately.

3. The Visual Acoustic Matching Task

We introduce a novel task, *visual acoustic matching*. In this task, an audio recording A_S recorded in space S and an image I_T of a different target space T are provided as input. The goal is to predict A_T, which has the same audio content as A_S but sounds as if it were recorded in space T with a microphone co-located with I_T's camera. Our goal is thus to learn a function f such that $f(A_S, I_T) = A_T$. The microphone co-location is important because acoustic properties vary as the listener location changes; inconsistent camera locations would lead to a perceived mismatch between the visuals and acoustics. The space S can have arbitrary acoustic characteristics, from an anechoic recording studio to a concert hall with significant reverberation. We assume there is one sounding object, leaving the handling of background sounds or interference as future work.

Importantly, our task formulation does *not* assume access to the impulse response, nor does it require the input audio to be anechoic. In comparison, the Image2Reverb [52] task requires access to both the impulse response and clean input audio, and does not account for the co-location of the camera and microphone.

4. Datasets

We consider two datasets: simulated audio in scanned real-world environments (Sec. 4.1), and in-the-wild Web videos with their recorded audio (Sec. 4.2). The former has the advantage of clean paired training data for A_T and A_S as well as precise ground truth for evaluating the output audio, but necessarily has a realism gap. The latter has the advantage of total realism, but makes quantitative evaluation more complex.

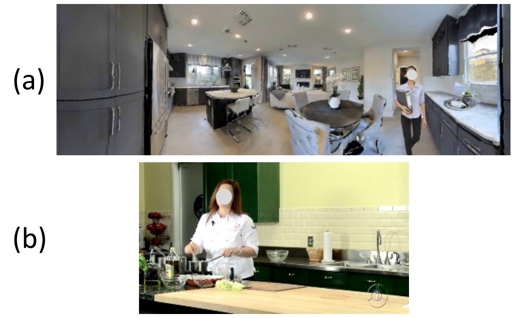

Figure 2. Example images in (a) SoundSpaces and (b) AVSpeech.

For both, we focus on human speech in indoor settings given its relevance to many of the applications cited above, and due to the fact that human listeners have strong prior knowledge about how reverberation should affect speech. However, our model design is not specific to speech. See Supp. for its applicability on non-speech sounds.

4.1. SoundSpaces-Speech Dataset

With the SoundSpaces platform [11], acoustics can be accurately simulated based on 3D scans of real-world environments [10, 55, 64]. This allows highly realistic rendering of arbitrary camera views and arbitrary microphone placements for waveforms of the user's choosing, accounting for all major real-world audio factors: direct sounds, early specular/diffuse reflections, reverberation, binaural spatialization, and effects from materials and air absorption.

We adopt a SoundSpaces-Speech dataset created in [12] consisting of paired clean (anechoic) and reverberant audio samples together with camera views.[1] The RIRs for 82 Matterport3D [10] environments are convolved with non-overlapping speech clips from LibriSpeech [45]. A 3D humanoid of the same gender as the real speaker is inserted at the speaker location and panorama RGB-D images are rendered at the listener location. See Figure 2a. Excluding those samples where the speaker is very distant or out-of-view (for which the visual input does not capture the geometry of the source location), there are 28,853/1,441/1,489 samples for the train/val/test splits.

4.2. Acoustic AVSpeech Web Videos

Web videos offer rich and natural supervision for the association between visuals and acoustics. We adopt a subset of the AVSpeech [16] dataset, which contains 3-10 second YouTube clips from 290k videos of single (visible) human speakers without interfering background noises. We automatically filter the full dataset down to those clips likely to meet our problem formulation criteria: 1) microphone and camera should be co-located and at a position different than the sound source (so that the audio contains not only

[1] Note that [12] uses the data for dereverberation, not acoustic matching.

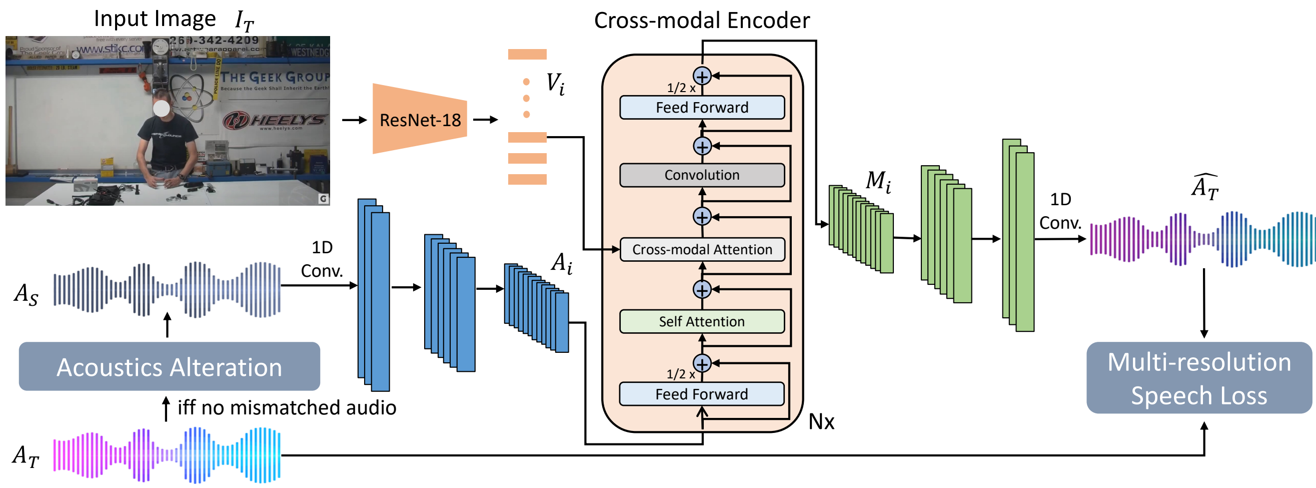

Figure 3. AViTAR model illustration. We extract visual feature sequence V_i from input image I_T with a ResNet-18 [25], and audio feature sequence A_i from input audio A_S with 1D convolutions. V_i and A_i are passed into cross-modal encoders for cross-modal reasoning. The output feature sequence M_i is processed and upsampled with 1D convolutions to recover the output of the same temporal length. Finally, we use a multi-resolution speech GAN loss to guide the audio synthesis to be high fidelity. The acoustics alteration process is applied to the target audio during training if and only if there is no mismatched audio, e.g., on the Acoustic AVSpeech dataset.

the source speech but also the reverberation caused by the environment), and 2) audio recording should be reverberant (so that the physical space has influenced the audio). Cameras in this dataset are typically static, and thus we use single frames and their corresponding audio for this task. See Supp. for details. This yields 113k/3k/3k video clips for train/val/test splits. We refer to this filtered dataset as Acoustic AVSpeech. See Figure 2b.

5. Approach

We present the **A**udio-**Vi**sual **T**ransformer for **A**udio Gene**r**ation model (**AViTAR**) (Figure 3). AViTAR learns to perform cross-modal attention based on sequences of convolutional features of audio and images and then synthesizes the desired waveform $\hat{A}_T$. We first define the audio-visual features (Sec. 5.1) and their cross-modal attention (Sec. 5.2), followed by our approach to waveform generation (Sec. 5.3). Finally, we present our acoustics alteration idea to enable learning from in-the-wild video (Sec. 5.4).

5.1. Audio-Visual Feature Sequence Generation

To apply cross-modal attention, we first need to generate sequences of audio and visual features, where each element in the sequence represents features of a part of the input space. For visual sequence generation from image I_T, we use ResNet18 [25] and flatten the last feature map before the pooling layer, yielding the visual feature sequence V_i.

For audio feature sequence generation from source audio A_S, we generate audio features A_i from the waveform directly with stacked 1D convolutions. We first use one 1D conv layer to embed the input waveform into a latent space. We then apply a sequence of strided 1D convolutions, each doubling the channel size while downsampling the input sequence. The output audio features are a sequence of vectors of size S, with length downsampled D times from the input. Weight normalization is applied to 1D conv layers. We employ 1D convolutions rather than STFT spectrograms so that the audio features are not limited to one resolution and can be optimized end-to-end to learn the most important features for the visual acoustic matching task.

5.2. Cross-Modal Encoder

Prior work often models audio-visual inputs in a simplistic manner by representing the image feature with one single vector and concatenating it with the audio feature [11, 12, 16, 20, 21, 44, 67]. However, for visual acoustic matching, it is important to reason how different regions of the space contribute to the acoustics differently. For example, a highly reflective glass door leads to longer reverberation time for high frequencies, while absorptive ceilings diminish that quickly. Thus, we propose to attend to image regions to reason how different image patches contribute to the acoustics, leveraging recent advances on the transformer architecture [24, 30, 60].

For cross-modal attention, we first adopt the conformer variant [24] of encoder blocks, which adds one convolution layer inside the block for modeling local interaction for speech features. Based on this block, we insert one cross-modal attention layer $\mathcal{A}_{cm}$ after the first feed-forward layer, described as follows:

$$\mathcal{A}_{cm}(A_i, V_i) = \mathrm{softmax}(\frac{A_i V_i^T}{\sqrt{S}})V_i, \tag{1}$$

where the attention scores between the two sequences of features A_i and V_i are first calculated by dot-product, then normalized by softmax, scaled by $\frac{1}{\sqrt{S}}$, and finally used to weight the visual features V_i. This cross-modal attention allows the model to attend to different image region features and reason about how they affect the reverberation. Absolute positional encoding is added to the visual encoding. After passing V_i and A_i through N encoder blocks, we obtain the fused audio-visual feature sequence M_i, which has the same length as A_i.

5.3. Waveform Generation and Loss

Recent audio-visual work generates audio outputs by inferring spectrograms then using ISTFT reconstruction to obtain a waveform (e.g., [16, 20, 21, 66–68]). While sensible for source separation, where the target signal is a subset of the source signal, ratio mask prediction is inadequate for our task, because reverberation might occupy periods of silence in the input audio and the ratio will be unbounded (as we verify in results). Futhermore, generating audio based on spectrograms is limiting because 1) predicting the coherent phase component remains challenging [3, 13], and 2) the spectrogram has one fixed resolution (one FFT size, hop length, and window size).

Instead, we aim to synthesize time-domain signals directly, skipping the intermediate spectrogram generation step and allowing more flexibility for what losses can be imposed, inspired by recent advances on time-domain speech synthesis [32, 35, 47, 59]. Specifically, with the fused audio-visual feature sequence M_i, we apply a sequence of transposed strided 1D convolutions, each halving the channel size while upsampling the input sequence, which is exactly the reverse operation of the audio encoding. Altogether, we upsample the audio sequence D times and obtain a waveform of the same length as the input.

Next we incorporate a multi-resolution generative loss. We found directly minimizing a Euclidean distance based loss between the target ground truth audio A_T and the inferred audio $\hat{A_T}$ leads to distortion in the generated audio on this task (cf. Figure 5 and Tab. 2). Therefore, to let the model learn how to reverberate the input speech properly, we employ a generative adversarial loss where a set of discriminators operating at different resolutions are trained to identify reverberation patterns and guide the generated audio to sound like real examples. Specifically, we apply an adversarial loss [32] comprised of the generator and discriminator losses:

$$\mathcal{L}_G = \sum_{k=1}^{K}(\mathcal{L}_{Adv}(G; D_k) + \lambda_1 \mathcal{L}_{FM}(G; D_k)) + \lambda_2 \mathcal{L}_{Mel}(G),$$

$$\mathcal{L}_D = \sum_{k=1}^{K} \mathcal{L}_{Adv}(D_k; G),$$

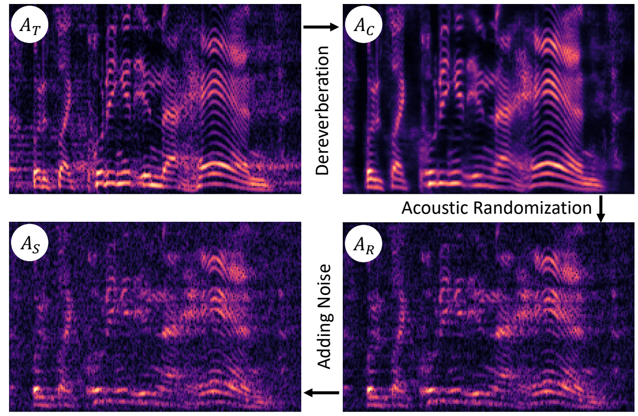

Figure 4. Acoustics alteration process. Spectrograms of the resulting audio after each step are shown. We first dereverberate the target audio A_T to obtain cleaner audio A_C, randomize its acoustics by applying an impulse response of another environment to obtain A_R, and finally, add Gaussian noise to A_R to create A_S. Notice how the spectral pattern changes in this process.

where each D_k is a sub-discriminator that operates at one of K different scales and periods for distinguishing the fake and real examples. $\mathcal{L}_{Adv}$ is the LS-GAN [39] training objective, which trains the generator to fake the discriminator and trains the discriminator to distinguish real examples from fake ones. For the generator G, a feature matching loss [35] $\mathcal{L}_{FM}$ is used, which is a learned similarity metric measured by the difference in features of the discriminator between a ground truth sample and a generated sample. An additional mel-spectrogram loss $\mathcal{L}_{Mel}$ is imposed on the generator for improving the training efficiency and fidelity of the generated audio. λ_1 and λ_2 are two weighting factors for these two losses. The generator loss $\mathcal{L}_G$ and discriminator loss $\mathcal{L}_D$ are trained alternatively competing against each other. For more details, refer to [32].

5.4. Acoustics Alteration for Self-Supervision

The training paradigm differs in one important way depending on the source of training data (cf. Sec. 4). For the simulated SoundSpaces data, we have access to an anechoic audio sample A_S as well as the ground truth reverberated sample A_T as it should be rendered in the target environment for a camera seeing view I_T. This means we can train to (implicitly) discover the mapping that takes the target image to an RIR which, when convolved with A_S, yields A_T.

For the in-the-wild video data (AVSpeech), however, we have only A_T and I_T to train, i.e., we only observe sounds that *do* match their respective views. Thus, to leverage unannotated Web video, we need to create an audio clip that preserves the target audio content but has *mismatched* acoustics. Figure 4 illustrates the steps for this process. First we strip away the original acoustics of the target environment by performing dereverberation on the audio A_T alone with the pretrained model from [12]. Since dereverberation is imperfect, there is residual acoustic information in the dereverberated output A_C, meaning that the resulting "clean" audio is still predictive of the target environment.

Thus, we subsequently randomize the acoustics by convolving that audio with an impulse response of another environment, yielding A_R; that IR is randomly chosen from the corresponding train/val/test split of SoundSpaces-Speech. The idea is to transform the semi-clean intermediate sound into another space to create more acoustic confusion, thereby forcing the model to learn from the target image. Finally, to further suppress the residual acoustics from the training environment, we add Gaussian noise with SNR randomly sampled from 2-10 dB to A_R and obtain the training source audio A_S. See more details about how each step alters the acoustics in Supp. In short, with this strategy, we are able to leverage readily available Web videos for our proposed task, despite its lack of ground truth paired audio.

6. Experiment

We validate our model on two datasets using comprehensive metrics and baselines. Implementation and training details can be found in Supp.

Evaluation metrics. We measure the quality of the generated audio from three aspects: 1) the closeness to the ground truth (if ground truth audio is available), as measured by **STFT Distance**, i.e., the MSE between the generated and true target audio's magnitude spectrograms; 2) the correctness of the room acoustics, as measured by the **RT60 Error (RTE)** between the true and inferred A_T's RT60 values. RT60 indicates the reverberation time in seconds for the audio signal to decay by 60 dB, a standard metric to characterize room acoustics. We estimate the RT60 directly from magnitude spectrograms of the output audio, using a model trained with disjoint SoundSpaces data (see Supp.), since impulse responses are not available for the target environments; and 3) the speech quality preserved in the synthesized speech, measured by the **Mean Opinion Score Error (MOSE)**, which is the difference in speech quality between the true target audio and generated audio, as assessed by a deep learning based objective model MOSNet [37].[2] Both the RTE and MOSE metrics are content-invariant and thus useful for evaluation when only audio with correct acoustics and mismatched content is available as ground truth, i.e., Web videos. In addition, we conduct user studies to evaluate whether a given audio is perceived as matching the room acoustics of the reference image.

[2]By taking the difference with the true target audio's MOS score (rather than simply the output's score), we account for the fact that properly reverberated speech need not have high speech quality.

Seen and unseen environments. On both datasets, we evaluate by pairing the source audio A_S with a target image I_T coming from either the training set (*Seen*) or test set (*Unseen*). The audio is always unobserved in training. The Seen case is useful to match the audio to scenes where we have video recordings (e.g., the film dubbing case). The Unseen case is important for injecting room acoustics depicted in novel images (e.g., to match sounds for a random Web photo being used as a Zoom call background).

Baselines. We consider the following baselines:

1. **Input audio**. This is the naive baseline that does nothing, simply returning the input A_S as output.

2. **Blind Reverberator**. This is a traditional acoustic matching approach [61] using audio recorded in the target space T as reference with content different from A_T. It first estimates RT60 and DRR from the reference audio (estimators are trained using simulated IRs), and then synthesizes the target IR by shaping an exponentially decaying white noise based on those two parameters. Unlike our model, this method requires reference audio at test time and IRs at training time. It is therefore inapplicable for the Unseen case (no reference audio) and AVSpeech (no training IRs).

3. **Image2Reverb [52]**. This is a recent approach that trains an IR predictor from images, then convolves the predicted IRs with A_S to obtain the target audio. This model requires access to the IR during training and thus is not applicable to the Acoustic AVSpeech dataset. We use the authors' code and convert the SoundSpaces-Speech data into the format of their dataset (see Supp.). We replace their depth prediction model with the ground truth depth image, to improve this baseline's performance.

4. **AV U-Net [20]**. This is an audio-visual model originally proposed for visually guided spatial sound generation based on a U-Net network for processing audio spectrograms. We adapt it for visual acoustic matching by removing the ratio mask prediction (which we find does not work well). Instead, we feed in a magnitude spectrogram, predict the target magnitude spectrograms, and generate the time-domain signals with Griffin Lim [23]. This baseline helps isolate the impact of our proposed cross-modal attention architecture compared to the common U-Net approach [13, 20, 21, 44, 68].

5. **AViTAR w/o visual**. This model is solely audio-based and is the same as our proposed model except that it does not have visual inputs or the cross-modal attention layer.

6.1. Results on SoundSpaces-Speech

For the SoundSpaces data, we have access to clean anechoic speech, which we use as the input A_S. The simulations offer a clean testbed for this task, showing the potential of each model when it is noise-free and the visuals reveal the full geometry via the panoramic RGB-D images.

	SoundSpaces-Speech						*Acoustic AVSpeech*			
	Seen			*Unseen*			*Seen*		*Unseen*	
	STFT	RTE (s)	MOSE	STFT	RTE (s)	MOSE	RTE (s)	MOSE	RTE (s)	MOSE
Input audio	1.192	0.331	0.617	1.206	0.356	0.611	0.387	0.658	0.392	0.634
Blind Reverberator [61]	1.338	0.044	0.312	-	-	-	-	-	-	-
Image2Reverb [52]	2.538	0.293	0.508	2.318	0.317	0.518	-	-	-	-
AV U-Net [20]	**0.638**	0.095	0.353	**0.658**	0.118	0.367	0.156	0.570	0.188	0.540
AViTAR w/o visual	0.862	0.140	0.217	0.902	0.186	0.236	0.194	0.504	0.207	0.478
AViTAR	0.665	**0.034**	**0.161**	0.822	**0.062**	**0.195**	**0.144**	**0.481**	**0.183**	**0.453**

Table 1. Results on the SoundSpaces-Speech and Acoustic AVSpeech datasets for Seen and Unseen environments. All input audio at test time is novel (unheard during training). Note that the STFT metric is applicable only for SoundSpaces, where we can access the ground truth A_T's spectrogram. For all metrics, lower values are better. Standard errors for STFT, RTE and MOSE are all less than 0.04, 0.013s and 0.01 on SoundSpaces-Speech. Standard errors for RTE and MOSE are all less than 0.005s and 0.01 on Acoustic AVSpeech.

Table 1 (left) shows the results. As expected, the clean input audio baseline does poorly because it does not account for the target environment. Our AViTAR model has the lowest RT60 error and MOS error, indicating that it best predicts the correct acoustics from images, injects them into the speech, and synthesizes high-quality audio. The AV U-Net baseline has slightly lower STFT distance than ours, likely because its training objective is to minimize STFT distance. However it has higher perceptual errors (RTE and MOSE). Image2Reverb's [52] high errors reveal the difficulty of our task and data, and its inapplicability to AVSpeech highlights our model's self-supervised training advantage. Despite having the estimated RT60 as input (and thus having low RT60 error), Blind Reverberator's STFT and MOS errors are much higher than AViTAR's, showing that images are a promising way to characterize room acoustics beyond the traditional RT60. Plus, its inapplicability for the other scenarios highlights fundamental advantages of AViTAR. Without access to visual information ("w/o visual"), AViTAR can only learn to add an average amount of reverberation to the input audio; this confirms that our model successfully learns the acoustics from the visual scene. Although this variant has higher RT60 error than AV U-Net, its MOS error is lower because the audio quality is better. See Supp. video for examples.

Ablations. Table 2 shows results for ablations on unseen images. For the model architecture, to understand if attending to different image regions with cross-modal attention is helpful, we train the full model with the length of visual feature sequence reduced to one by mean pooling the final ResNet feature map ("w/ pooled visual feature"). This model underperforms the full model on both STFT and RT60 metrics, showing that the audio-visual attention leads to a better visual understanding of room acoustics. Next we ablate the generative loss and replace it with the non-generative multi-resolution STFT loss [35] ("w/o generative loss"), which slightly improves the STFT error but leads to a large drop on the acoustics recovery and speech quality. Despite being multi-resolution, without learnable discriminators to learn to model those fine reverberation details, the audio quality gets worse. See Supp. for GAN loss ablations.

AViTAR	STFT	RTE (s)	MOSE
Full model	0.822	**0.062**	0.195
w/ pooled visual feature	0.850	0.067	**0.193**
w/o generative loss	**0.777**	0.081	0.314
w/o human	0.884	0.139	0.218
w/ random image	0.940	0.236	0.250

Table 2. Ablations on model design and data.

The synthetic dataset provides access to meta information useful to evaluate whether and how much AViTAR reasons about different visual properties. The location of the sound source matters for acoustics because it directly influences acoustic characteristics like the direct-to-reverberant ratio (DRR). When we remove the 3D humanoid from the scene ("w/o human") in all test images, all error metrics increase, which indicates that our model reasons about the location of the sound source in the image for accurate acoustic matching. To understand if the model learns meaningful information from the visuals, we replace the target image with a random image ("w/ random image"); this significantly harms our model's performance.

6.2. Results on Acoustic AVSpeech

Next, we train our model on the in-the-wild AVSpeech videos, and test it on novel clean speech clips from LibriSpeech [45] (A_S) paired with target images (I_T) from AVSpeech. Here we do not have ground truth for the target speech, so we evaluate with RTE and MOSE.

Table 1 (right) shows the results. Our proposed AViTAR model achieves the lowest RT60 error compared to all baselines. This shows our model trained in its self-supervised fashion successfully generalizes to novel images and novel audio, and demonstrates we can do acoustic matching even

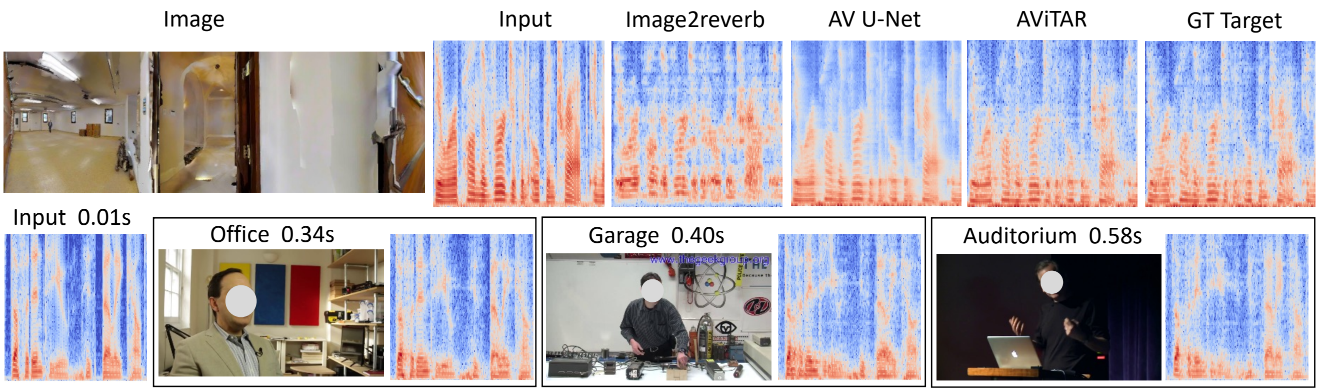

Figure 5. Qualitative predicted audio. For all audio clips, we compute the magnitude spectrogram, convert the magnitude to dB, and plot the spectrogram with x-axis spanning from 0 to 1.28 s (left to right) and y-axis from 0 to 3000 Hz (bottom to top). Row 1: SoundSpaces-Speech example where the target space is a large empty room with a lot of reverberation. Our model predicts the audio closest to the target clip. AV U-Net's spectrogram is too smoothed compared to ours and misses some fine reverb details, which leads to perceptual distortion. Row 2: examples on Acoustic AVSpeech (unseen images). We feed one clean audio clip to match three different scenarios (office, garage, auditorium). From left to right, the audio spectrogram becomes more reverberant as phoneme patterns get extended and blurred on the temporal axis (est. RT60 times shown). NB: AViTAR processes waveforms, not spectrograms; here they are for visualization.

Acoustics Alteration	Seen	Unseen
Dereverb. + Randomization + Noise	**0.144**	**0.183**
Dereverb. + Randomization	0.178	0.197
Dereverb. + Noise	0.170	0.208
Dereverb.	0.230	0.250
A_T + Randomization + Noise	0.236	0.249

Table 3. Ablations on acoustics alteration. RTE is reported.

	SoundSpaces	AVSpeech
Input Speech	42.1% / **57.9%**	40.1% / **59.9%**
Image2Reverb [52]	25.9% / **74.1%**	- / -
AV U-Net [20]	29.8% / **70.2%**	27.2% / **72.8%**
AViTAR w/o visual	39.6% / **60.4%**	46.3% / **53.9%**

Table 4. User study results. X%/Y% indicates among all paired examples for this baseline and AViTAR, X% of participants prefer this baseline while **Y%** prefer AViTAR.

for non-anechoic inputs. AViTAR's MOS error is also the lowest compared to all baselines, showing that it is able to synthesize high-fidelity audio while injecting the proper amount of reverberation into the speech. The absolute errors on AVSpeech are higher than on SoundSpaces, which makes sense because the YouTube imagery is more variable, and it has a narrower field of view and no depth, making the geometry and materials of the scene only partly visible. See Supp. for sim2real generalization.

Ablations on acoustic alteration. Table 3 shows ablations on the proposed acoustics alteration strategy. In short, all three steps are necessary to create an acoustic mismatch with the image, thereby forcing the model to recover the correct acoustics based on the image and allowing better generalization to novel sounds. See Supp. for details.

User study. To supplement the quantitative metrics and directly capture the perceptual quality of the generated samples, we next conduct a user study. We show participants the image of the target environment I_T, the accompanying ground truth audio clip A_T as reference, and paired audio clips $\hat{A}_T$ generated by AViTAR and each baseline. We ask participants to select the clip that most sounds as if it were recorded in the target environment and best matches the reverberation in the given clip. We select 30 reverberant examples from SoundSpaces-Speech and AVSpeech and ask 30 participants to complete the assignment on MTurk.

Table 4 shows the resulting preference scores. Compared to each baseline, AViTAR is always preferred. Note that no participant has a background in acoustics, and some might simply pick the one that sounds "clean" rather than having the correct room acoustics. This may be the reason even the anechoic input has a higher preference score than the U-Net model. Despite the lack of domain knowledge, participants still consistently favor our model over other baselines.

Qualitative examples. Figure 5 shows example outputs. Please see the Supp. video to gauge the audio quality.

7. Conclusion

We proposed the visual acoustic matching task and introduced the first model to address it. Given an image and audio clip, our method injects realistic room acoustics to match the target environment. Our results validate their realism with both objective and perceptual measures. Importantly, the proposed model is trainable with unannotated, in-the-wild Web videos. In future work we aim to extend our model to leverage the dynamics in target visual scenes in video. We discuss potential societal impact in Supp.

Acknowledgements UT Austin is supported in part by a gift from Google and the IFML NSF AI Institute.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. In *INTERSPEECH*, 2018. 2

[2] Triantafyllos Afouras, Andrew Owens, Joon-Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *ECCV*, 2020. 2

[3] Yang Ai and Zhen-Hua Ling. A neural vocoder with hierarchical generation of amplitude and phase spectra for statistical parametric speech synthesis. In *IEEE Transactions on Audio, Speech and Language Processing*, 2019. 5

[4] Hassan Akbari, Linagzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *arXiv preprint arXiv:2104.11178*, 2021. 3

[5] Jont B. Allen and David A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979. 2

[6] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *NeurIPS*, 2020. 2, 3

[7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021. 3

[8] Stefan Bilbao. Modeling of complex geometries and boundary conditions in finite difference/finite volume time domain room acoustics simulation. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(7):1524–1533, 2013. 1

[9] Chunxiao Cao, Zhong Ren, Carl Schissler, Dinesh Manocha, and Kun Zhou. Interactive sound propagation with bidirectional path tracing. *ACM Transactions on Graphics (TOG)*, 35(6):1–11, 2016. 1

[10] 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`. 3

[11] 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. 2, 3, 4

[12] Changan Chen, Wei Sun, David Harwath, and Kristen Grauman. Learning audio-visual dereverberation. In *arXiv*, 2021. 2, 3, 4, 6

[13] Hyeong-Seok Choi, Jang-Hyun Kim, Jaesung Huh, Adrian Kim, Jung-Woo Ha, and Kyogu Lee. Phase-aware speech enhancement with deep complex u-net. In *ICLR*, 2019. 5, 6

[14] Victoria Dean, Shubham Tulsiani, and Abhinav Gupta. See, hear, explore: Curiosity via audio-visual association. In *NeurIPS*, 2020. 2

[15] James Eaton, Nikolay Gaubitch, Allistair Moore, and Patrick Naylor. Estimation of room acoustic parameters: The ACE challenge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10), 2016. 2

[16] 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. In *SIGGRAPH*, 2018. 2, 3, 4, 5

[17] Thomas Funkhouser, Nicolas Tsingos, Ingrid Carlbom, Gary Elko, Mohan Sondhi, James E West, Gopal Pingali, Patrick Min, and Addy Ngan. A beam tracing method for interactive architectural acoustics. *The Journal of the acoustical society of America*, 115(2):739–756, 2004. 1

[18] Hannes Gamper and Ivan J Tashev. Blind reverberation time estimation using a convolutional neural network. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 136–140, 2018. 2

[19] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *ICRA*, 2020. 2

[20] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *CVPR*, 2019. 3, 4, 5, 6, 7, 8

[21] Ruohan Gao and Kristen Grauman. VisualVoice: Audio-visual speech separation with cross-modal consistency. In *CVPR*, 2021. 3, 4, 5, 6

[22] Rohit Girdhar and Kristen Grauman. Anticipative Video Transformer. In *ICCV*, 2021. 3

[23] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984. 6

[24] Anmol Gulati, Chung-Cheng Chiu, James Qin, Jiahui Yu, Niki Parmar, Ruoming Pang, Shibo Wang, Wei Han, Yonghui Wu, Yu Zhang, and Zhengdong Zhang. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech*, 2020. 3, 4

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[26] Martin Holters, Tobias Corbach, and Udo Zölzer. Impulse response measurement techniques and their applicability in the real world. In *Proceedings of the 12th International Conference on Digital Audio Effects, DAFx 2009*, 2009. 2

[27] Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Yu Tsao, Hsiu-Wen Chang, and Hsin-Min Wang. Audio-visual speech enhancement using multimodal deep convolutional neural networks. In *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2017. 2

[28] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. In *NeurIPS*, 2020. 2

[29] Florian Klein, Annika Neidhardt, and Marius Seipel. Real-time estimation of reverberation time for selection of suitable binaural room impulse responses. In *Audio for Virtual, Augmented and Mixed Realities: Proceedings of 5th International Conference on Spatial Audio (ICSA 2019)*, pages 145–150. 1, 2

[30] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. In *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, 2021. 3, 4

[31] Homare Kon and Hideki Koike. Estimation of late reverberation characteristics from a single two-dimensional environmental image using convolutional neural networks. In *Journal of the Audio Engineering Society*, 2019. 2

[32] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *NeurIPS*, 2020. 5

[33] Junghyun Koo, Seungryeol Paik, and Kyogu Lee. Reverb conversion of mixed vocal tracks using an end-to-end convolutional deep neural network. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 81–85. IEEE, 2021. 2

[34] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *NeurIPS*, 2018. 2

[35] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. In *NeurIPS*, 2019. 5, 7

[36] Yan-Bo Lin and Yu-Chiang Frank Wang. Audiovisual transformer with instance attention for audio-visual event localization. In *ACCV*, 2020. 3

[37] Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. Mosnet: Deep learning based objective assessment for voice conversion. In *Proc. Interspeech 2019*, 2019. 6

[38] Wolfgang Mack, Shuwen Deng, and Emanuël AP Habets. Single-channel blind direct-to-reverberation ratio estimation using masking. In *INTERSPEECH*, pages 5066–5070, 2020. 2

[39] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. *arXiv preprint arXiv:1611.04076*, 2016. 5

[40] Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen. An overview of deep-learning-based audio-visual speech enhancement and separation. In *arXiv*, 2020. 2

[41] Pedro Morgado, Yi Li, and Nuno Vasconcelos. Learning representations from audio-visual spatial alignment. In *NeurIPS*, 2020. 2

[42] Prateek Murgai, Mark Rau, and Jean-Marc Jot. Blind estimation of the reverberation fingerprint of unknown acoustic environments. In *Audio Engineering Society Convention 143*. Audio Engineering Society, 2017. 1, 2

[43] D.T. Murphy, Antti Kelloniemi, Jack Mullen, and Simon Shelley. Acoustic modeling using the digital waveguide mesh. *Signal Processing Magazine, IEEE*, 24:55 – 66, 04 2007. 2

[44] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 2, 3, 4, 6

[45] 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. IEEE, 2015. 3, 7

[46] Mandela Patrick, Yuki M Asano, Bernie Huang, Ishan Misra, Florian Metze, Joao Henriques, and Andrea Vedaldi. Space-time crop & attend: Improving cross-modal video representation learning. *arXiv preprint arXiv:2103.10211*, 2021. 3

[47] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *arXiv*, 2018. 5

[48] Mostafa Sadeghi, Simon Leglaive, Xavier Alameda-PIneda, Laurent Girin, and Radu Horaud. Audio-visual speech enhancement using conditional variational auto-encoders. In *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2020. 2

[49] Andy Sarroff and Roth Michaels. Blind arbitrary reverb matching. In *Proceedings of the 23rd International Conference on Digital Audio Effects (DAFx-2020)*, 2020. 2

[50] Lauri Savioja and U Peter Svensson. Overview of geometrical room acoustic modeling techniques. *The Journal of the Acoustical Society of America*, 138(2):708–730, 2015. 1

[51] Lauri Savioja and Ning Xiang. Simulation-based auralization of room acoustics. *Acoust. Today*, 16(4):48–55, 2020. 1

[52] Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2reverb: Cross-modal reverb impulse response synthesis. In *ICCV*, 2021. 2, 3, 6, 7, 8

[53] Guy-Bart Stan, Jean-Jacques Embrechts, and Dominique Archambeau. Comparison of different impulse response measurement techniques. *Journal of the Audio Engineering Society*, 50(4):249–262, april 2002. 2

[54] Christian Steinmetz, Vamsi Krishna Ithapu, and Paul Calamia. Filtered noise shaping for time domain room impulse response estimation from reverberant speech. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021. 2

[55] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 3

[56] Jiaqi Su, Zeyu Jin, and Adam Finkelstein. Acoustic matching by embedding impulse responses. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 426–430. IEEE, 2020. 1, 2

[57] Thanh-Dat Truong, Chi Nhan Duong, The De Vu, Hoang Anh Pham, Bhiksha Raj, Ngan Le, and Khoa Luu. The right to talk: An audio-visual transformer approach. In *ICCV*, 2021. 3

[58] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov.

Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Florence, Italy, 7 2019. Association for Computational Linguistics. 3

[59] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *arXiv*, 2016. 5

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3, 4

[61] Vesa Välimäki, Julian Parker, Lauri Savioja, Julius O. Smith, and Jonathan Abel. More than 50 years of artificial reverberation. In *60th International Conference: DREAMS*, 2016. 6, 7

[62] Sanna Wager, Keunwoo Choi, and Simon Durand. Dereverberation using joint estimation of dry speech signal and acoustic system. *arXiv preprint arXiv:2007.12581*, 2020. 2

[63] Stephan Werner, Florian Klein, Annika Neidhardt, Ulrike Sloma, Christian Schneiderwind, and Karlheinz Brandenburg. Creation of auditory augmented reality using a position-dynamic binaural synthesis system—technical components, psychoacoustic needs, and perceptual evaluation. *Applied Sciences*, 11(3), 2021. 1

[64] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson Env: real-world perception for embodied agents. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018. 3

[65] Feifei Xiong, Stefan Goetze, Birger Kollmeier, and Bernd T Meyer. Joint estimation of reverberation time and early-to-late reverberation ratio from single-channel speech signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(2):255–267, 2018. 2

[66] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP*, 2020. 5

[67] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *ICCV*, 2019. 2, 4, 5

[68] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *The European Conference on Computer Vision (ECCV)*, September 2018. 5, 6

[69] Hang Zhou, Ziwei Liu, Xudong Xu, Ping Luo, and Xiaogang Wang. Vision-infused deep audio inpainting. In *ICCV*, 2019. 2