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.

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.
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 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 evalua-

![Image](image_url)

Figure 2. Example images in (a) SoundSpaces and (b) A\(^{\text{VSpeech}}\).

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.
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 Audio-Visual Transformer for Audio Generation 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 audiovisual 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 $A_{cm}$ after the first feed-forward layer, described as follows:

$$A_{cm}(A_i, V_i) = \text{softmax}(\frac{A_iV_i^T}{\sqrt{S}})V_i,$$  \hspace{1cm} (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 \( \frac{1}{\sqrt{5}} \), 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 upampling 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 \( A_T \) 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:

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

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

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. \( \lambda_1 \) and \( \lambda_2 \) 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 en-
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].\footnote{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).

3. 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.
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. 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 2. Ablations on model design and data.

<table>
<thead>
<tr>
<th>Ablations</th>
<th>STFT</th>
<th>RTE (s)</th>
<th>MOSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full model</td>
<td>0.822</td>
<td>0.062</td>
<td>0.195</td>
</tr>
<tr>
<td>w/ pooled visual feature</td>
<td>0.850</td>
<td>0.067</td>
<td>0.193</td>
</tr>
<tr>
<td>w/o generative loss</td>
<td>0.777</td>
<td>0.081</td>
<td>0.314</td>
</tr>
<tr>
<td>w/o human</td>
<td>0.884</td>
<td>0.139</td>
<td>0.218</td>
</tr>
<tr>
<td>w/ random image</td>
<td>0.940</td>
<td>0.236</td>
<td>0.250</td>
</tr>
</tbody>
</table>

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.

The synthetic dataset provides access to meta information useful to evaluate whether and how much AViSTAR 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] (S) paired with target images (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 AViSTAR 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.

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

<table>
<thead>
<tr>
<th>Acoustics Alteration</th>
<th>Seen</th>
<th>Unseen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dereverb. + Randomization + Noise</td>
<td>0.144</td>
<td>0.183</td>
</tr>
<tr>
<td>Dereverb. + Randomization</td>
<td>0.178</td>
<td>0.197</td>
</tr>
<tr>
<td>Dereverb. + Noise</td>
<td>0.170</td>
<td>0.208</td>
</tr>
<tr>
<td>Dereverb.</td>
<td>0.230</td>
<td>0.250</td>
</tr>
<tr>
<td>$A_T$ + Randomization + Noise</td>
<td>0.236</td>
<td>0.249</td>
</tr>
</tbody>
</table>

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.

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 sim2real generalization.

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 $A_T'$ 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

[20] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In CVPR, 2019. 3, 4, 5, 6, 7, 8
[40] Daniel Michelanti, 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
[52] Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2reverb: Cross-modal reverb impulse response synthesis. In ICCV, 2021. 3, 6, 7, 8
[57] Thanh-Dat Truong, Chi Nhan Duong, The De Vu, Hoang Anh Pham, Bhioka 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.


[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


