Hiding Left from Right: Examining the Robustness of Multi-modal Upmixing Models

Tai Wan Kim
Dartmouth College
Hanover NH
Tai.Wan.Kim.21@dartmouth.edu

Phuc Dai Tran
Dartmouth College
Hanover NH
Phuc.D.Tran.24@dartmouth.edu

Mark Lekina Rorat
Dartmouth College
Hanover NH
Mark.L.Rorat.24@dartmouth.edu

Abstract

Multi-modal audio spatialization task (upmixing) aims to convert a single channel audio from a video to a multi-channel audio by extracting and aligning visual features. In recent work, state of the art performance in upmixing has been achieved by aligning visual features learned from a model pre-trained on a novel pretext task where the model has to correctly classify whether the left and right channels of the audio have been flipped. Our work examines the robustness of this model to noisy visual content. We create noisy video samples via visual masking and RGB perturbation strategies. We observe that while RGB perturbation has no significant effect, loss increases with higher ratio of masking. Qualitative listening of select samples of upmixed audio suggests that the model cannot fully leverage temporal correlation to reconstruct missing information[1].

1. Introduction

Audio spatialization, or upmixing, is the task of converting a single channel, mono audio to a spatial, stereo audio. Upmixed audio enhances the listener's experience of surround sound systems or virtual reality environments and has multiple use cases such as restoring damaged audio or recreating older recordings.

With recent advancements in the field, there have been several attempts to improve upmixing by leveraging deep learning models. In particular, many previous works have focused on multimodal upmixing where a mono audio from a video is upmixed by injecting visual information. A notable example is Gao *et al*. 's Mono2Binaural framework where the ImageNet pre-trained ResNet-18 network extracts visual features to be concatenated to audio features [1].

A recent study further advances the Mono2Binaural framework by introducing a novel self-supervised pre-training scheme. Yang *et al*. [2] proposes a pretext task where the left and right audio channels of each video sample are reversed at a given rate, after which the model has to correctly classify whether the audio has been flipped by leveraging visual cues. The pretext task effectively trains the model to discern whether the auditory and visual content of the video are aligned. [2] demonstrates that visual features pretrained with this strategy improve performance on several downstream tasks including upmixing.

While [2] 's approach is promising, it is unclear whether the model will be effective in real life applications. [2] 's model is pre-trained on Youtube-ASMR-300K dataset, a dataset consisting of autonomous sensory meridian response (ASMR) videos available on Youtube. Considering the scarcity of videos with stereo sound, ASMR videos are in many ways advantageous for upmixing as they provide multi-channel audio and have strong correlation between sight and sound. ASMR videos usually feature a single artist where it is relatively simple for the model to locate and track the source of sound. However, whether the model can be applied for more challenging visual information e.g., a scene from a concert, is open to question.

We evaluate the robustness of [2] 's upmixing model on a noisy test set where we add noise to the visual content. We take three approaches. First, we mask out a number of randomly selected video frames. Second, we mask out a number of sequential, consecutive frames beginning from a

[1]Code available at https://github.com/taytwkim/hiding-left-from-right

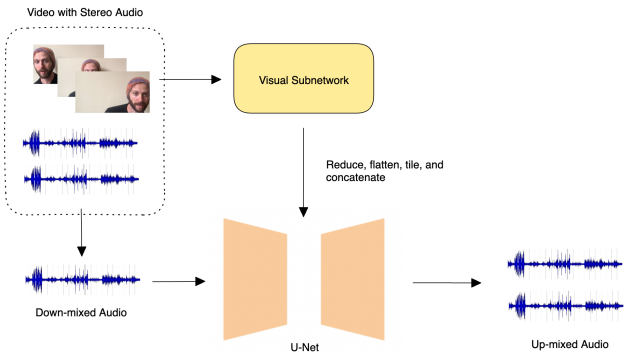

Figure 1. Model Overview. Visual features learned from the base model are concatenated to the U-Net.

randomly selected starting point. Lastly, we perturb a select number of randomly selected video frames' RGB values. We examine if the model's performance degrades under these conditions.

We find that i) RGB perturbation has no significant impact on the results; and ii) the model's performance suffers as we increase the masking ratio for random and tube masking. The results also suggest that despite strong temporal correlation of videos, the model cannot effectively predict and reconstruct audio for masked portions.

2. Related Work

Audio spatialization: Audio spatialization refers to the process of up-mixing mono audio to stereo audio while using a concurrent visual stream to introduce spatial cues to the audio. Previous work has used spatial audio as a self-supervisory signal for audio spatialization.

For example, in [3], the authors train a self-supervised audio spatialization network that takes in video and monoaural audio as input and generates spatial audio. It is also the adversarial application for audio spatialization that we have encountered so far that is most relevant to our approach. Here the authors apply a spatial correspondence classifier as an adversarial loss to enhance the model's performance. Similar to [2], the classifier is used to distinguish ground-truth videos from videos whose right and left channels have been swapped. [4] implements a similar approach to generate spatial audio for 360° videos. The authors demonstrate that it is possible to infer the spatial location of sound sources given 360° video and a mono audio track only as input.

Specifically, in [1], a visual stream is used to convert a down-mixed mono audio input to a stereo output. The original stereo audio is used as the training target. The features obtained from the visual stream provide complementary spatial information that is missing from the mono audio to produce stereo audio. Gao *et al*. [1] observe that using the original stereo audio as the training target could result in the model learning to just copy and paste the input audio. To fix this, they propose a new training task; instead of training the model to predict the left and right audio channels, they instead train it to predict the difference signal between the left and right channels. The reasoning here is that predicting this difference signal gives the model a challenging task, which forces the model to reason out the visual features and learn the minute differences between the left and right channels. [1–4] all make use of the U-Net architecture to extract audio features and other CNN-based frameworks (usually the ResNet-18 [5]) to encode video. The U-Net [6], while originally developed for the segmentation of medical images, has been demonstrated to be useful for audio spatialization applications.

Key Prior Work: The most significant paper related to our proposed approach is the paper by Yang *et al*. [2]. The approach presented in this paper exploits spatial cues for self-supervision by taking video with stereo audio as input and learning a model to match spatial audio cues to positions of sound sources in the visual stream. The model achieves this by leveraging the spatial correspondence between the audio and visual signals, and using this to generate a difference signal for the right and left audio channels. In other words, the model learns spatial cues directly from stereo audio to match the perceived localization of a sound with its position in the video. Unlike the approach in [7], the representation is learned without explicitly modeling the target locations of a teacher model. [2] adopts the Mono2Binaural framework of Gao *et al*. [1], using the U-Net to up-mix the audio and concatenating the pre-trained visual features to the innermost layer of the U-Net. One key difference is that [2] uses Tanh activation to produce the difference mask, as opposed to Gao *et al*. [1], who use Sigmoid activation. Yang *et al*. observe that the asymmetry of the Sigmoid layer biases the upmixing in one direction. This is only detected due to the presence of strong binaural cues in the dataset used in [2]. The authors demonstrate that this strategy improves the model's performance on down-

stream tasks such as sound localization, source separation and, more relevant to our approach, audio up-mixing. Our approach aims to build on the work of [2] by borrowing from [3] and [8], which employ adversarial strategies to enhance the robustness of model performance. To our knowledge, this is no prior work that specifically examines the robustness of [2].

3. Proposed Approach

Our research examines the model in [2] (see Figure 1) with black box attacks to suggest future possibilities on adversarial visual-auditory pre-training. We propose three methods: i) masking out a number of randomly selected frames; ii) masking out a number of sequential frames from a random starting point; and iii) adding RGB perturbation to a randomly selected number of frames. We present these as extreme cases of video degradation in real-life scenarios e.g., video quality is low or a hand is momentarily covering the video.

3.1. Dataset

Youtube-ASMR-300K [2] is a dataset consisting of 300K ASMR clips collected from Youtube, each 10 seconds long. Using only the subset of the original dataset, we download 293 videos. Following the approach in [2], each clip is split into shorter clips of 2-3 seconds, resulting in 1465 clips, approximating 49 minutes altogether.

3.2. Model

We replicate the model from [2] with best pre-trained weights. We use the version where the base model has been trained from scratch with ResNet-18 [5] as the visual sub-network and SE Net [9] as the audio sub-network. The pre-trained visual features from the base model are concatenated to the innermost layer of the U-Net from the Mono2Binaural framework [1] for up-mixing.

3.3. Adding Noise

3.3.1 Random Masking

Since up-mixing relies heavily on visual-audio correspondence, masking may potentially misdirect the model in locating the sound source. For our first approach, we choose a number of randomly selected and independent frames to mask out as black. Similarly, we also experiment with masking a number of sequential frames starting at a random start point to test whether the model can predict audio for the masked portion based on long-term dependencies on temporally distant frames[2] (See Figure 2).

[2]Selected samples of upmixed audio are publicly available here.

3.3.2 RGB Perturbation

RGB perturbation is a common black-box attacking strategy for image classification. We propose the same approach to attack our up-mixing model. Instead of retrieving the gradient from the model's losses, we randomly generate RGB noises. We choose our epsilon value to be 0.1 and add a uniform distribution from -0.1 to 0.1 to all pixels of selected frames. We clip all noisy pixel values to the possible RGB range between 0 and 255.

4. Experimental Results

4.1. Implementation Detail

4.1.1 Dataset Download and Preprocessing

We obtain a dataset which provides the URL and the start/end timestamps of each ASMR video to download. The videos are downloaded via the youtube-dl[3] package. Following [2]'s approach, we set the video frame rate to 30 fps and audio bitrate to 192000 bps. The original stereo audio is down-sampled to a mono audio by simply taking the average between the left and right audio channels. *Short-Term Fourier Transform* (STFT) is applied to the difference between the left and right channels to obtain the spectrogram of the audio. The input and output of the model are both spectrogram representations.

4.1.2 Metrics

For evaluation criteria, we use the L1 distance between the STFT spectrograms of the original stereo audio and the up-mixed audio.

4.2. Results

4.2.1 Masking Random (Non-Sequential) Frames

Masking Ratio	Total Loss	Mean Loss
0 (baseline)	144.8925	0.0989
0.1	144.7421	0.0988
0.25	145.1574	0.0990
0.5	146.3045	0.0998
0.75	147.5198	0.1006
1	149.5971	0.1021

Table 1. The results from masking randomly selected frames. Performance starts degrading starting at the masking ratio of 0.25.

We experiment with the masking ratio r of $0.1, 0.25, 0.5$, and 0.75, where $r = 0$ corresponds to no masking and $r = 1$ corresponds to no visual cues available. The results are reported in Table 1. As expected, we observe a

[3]http://youtube-dl.org/

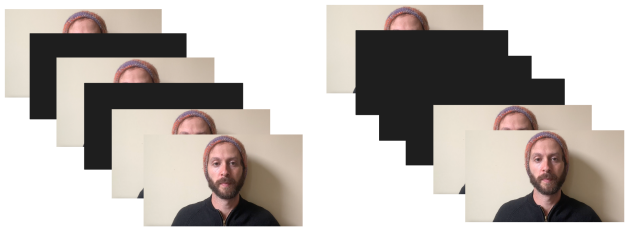

Figure 2. Masking. On the left is random masking and on the right is sequential masking.

consistent increase in loss as the masking ratio increases. Qualitative examination (listening) of selected audio samples demonstrate that the quality of upmixed audio significantly degrades when more than half of the frames are masked ($r > 0.5$). As videos tend to be temporally correlated, videos traditionally require a higher masking ratio to effectively conceal information [10]; even when half of the randomly selected frames are masked, the left-to-right motion of the ASMR artist in the video is apparent to human eyes in most cases. However, our results suggest that the model cannot effectively reconstruct missing pieces by leveraging temporally redundant visual features.

4.2.2 Masking Sequential Frames

Masking Ratio	Total Loss	Mean Loss
0.1	145.5720	0.09936
0.25	146.3417	0.09989
0.5	147.06125	0.1003
0.75	147.8563	0.1009

Table 2. The results from masking consecutive frames. For each masking ratio, we observe that the loss is greater than its counterpart for masking random (non-consecutive) frames.

As above, we experiment with the masking ratio r of $0.1, 0.25, 0.5$, and 0.75. The results are reported in Table 2. As expected, we observe that masking consecutive frames leads to higher loss than masking randomly selected frames. By masking consecutive numbers of frames, we make it more difficult for the model to predict the left-to-right motion of the ASMR artist in the video.

4.2.3 RGB Perturbation

As above, we experiment with the ratio r of $0.1, 0.25, 0.5$, and 0.75. The results are reported in Table 3. We observe that RGB perturbation has no significant effect. This seems to be because RGB perturbation does not obstruct the left-to-right motion of the ASMR artist, which is the primary feature the model learns from.

Masking Ratio	Total Loss	Mean Loss
0.1	144.8602	0.09888
0.25	144.8204	0.09885
0.5	144.7015	0.09877
0.75	144.7142	0.09878
1	144.6245	0.09871

Table 3. The results from RGB perturbation with 0.1 epsilon value. Performance is not significantly affected compared to the baseline loss of 0.0989 from Table 1

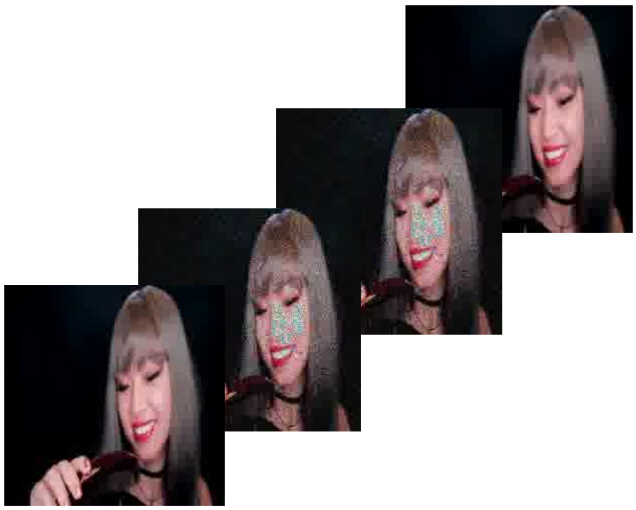

Figure 3. RGB Perturbation. Perturbation is added to each frame in this example with varying epsilon values.

5. Conclusion

We have explored the robustness of an existing model to noisy visual input. We find that masking even a relatively little number of frames (0.5) can lead to a significant decrease in performance. Our work has several limitations. First, our work can benefit from in-depth analysis of the upmixed audio to examine the loss for specific portions of the audio (i.e. the masked portion or the unmasked portion). This will allow us to draw stronger conclusions on the degree to which mask affects the corresponding or neighboring audio bins. Furthermore, our work is but an evaluation

of an existing approach and can be extended to proposing a solution to the problem. Our future plan is to pre-train the model on adversarial samples to examine if the model becomes more robust to these types of noisy input.

References

[1] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2019. 1, 2, 3

[2] Karren Yang, Bryan Russell, and Justin Salamon. Telling left from right: Learning spatial correspondence of sight and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9932–9941, 2020. 1, 2, 3

[3] Yu-Ding Lu, Hsin-Ying Lee, Hung-Yu Tseng, and Ming-Hsuan Yang. Self-supervised audio spatialization with correspondence classifier. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3347–3351. IEEE, 2019. 2, 3

[4] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. *Advances in neural information processing systems*, 31, 2018. 2

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 3

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2

[7] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7053–7062, 2019. 2

[8] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 3

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3

[10] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 4

Task Assignment

Data

- Debugging Youtube-dl python script and setting it up on Colab - Tay
- Downloading test set videos from YouTube - Lekina
- Pre-processing videos and audios (adjust frame rate, bit rate, etc.) - Lekina

Model Set Up & Evaluation

- Debugging clip generator and model, setting it up on Colab - Tay
- Code for adding noise to video (masking) - Tay
- Code for adding noise to video (RGB perturbation) - Phuc
- Evaluation for masked videos - Tay
- Evaluation for RGB perturbed videos - Phuc
- Select samples for qualitative listening - Tay

Final Report

- Figures and tables - Tay, Phuc, Lekina
- Abstract and introduction/conclusion - Tay
- Related works - Lekina
- Proposed approach - Phuc
- Experimental results - Tay
- References - Lekina