Visually Informed Binaural Audio Generation without Binaural Audios (Supplementary Material)

Xudong Xu[1*] Hang Zhou[1*] Ziwei Liu[2] Bo Dai[2] Xiaogang Wang[1] Dahua Lin[1]
[1]CUHK - SenseTime Joint Lab, The Chinese University of Hong Kong
[2]S-Lab, Nanyang Technological University
{xx018@ie,zhouhang@link,xgwang@ee,dhlin@ie}.cuhk.edu.hk, {ziwei.liu,bo.dai}@ntu.edu.sg

The supplementary materials contain:

1. Demo video.
2. Separation training details.
3. Implementation details.
4. More ablation studies.
5. Experiment on Youtube-ASMR dataset.
6. Limitations.

1. Demo Video

In the demo video, we show our prediction results and the comparisons between our predictions and that of the baseline method [1] on FAIR-Play, MUSIC-Stereo and Youtube-ASMR 5, respectively. The video is encoded in H.264 codec.

2. Separation Training Details

The task of visually guided sound source separation [6, 2] aims at separating a mixed audio into independent ones, according to their sound source's visual appearances. We adopt the setting with a mixture of two sources, which is kept the same as Sep-Stereo [7] for fair comparisons. The backbone audio network Net_a is shared across stereo and separation learning, following [7].

During training, we follow the Mix-and-Separate pipeline [6, 2] to mix two mono audios (s_a and s_b) of two solo videos as input in the form of STFTs. This can be written as $S_{mix} = S_a + S_b$. The network renders complex masks $\mathcal{M}_a$ and $\mathcal{M}_b$ to predict the STFTs of the audios themselves. We use two APNet [7] structures for the prediction of complex masks. The loss function can be written as:

$$\mathcal{L}_{sep} = ||S_a - \mathcal{M}_a * S_{mix}||_2^2 + ||S_b - \mathcal{M}_b * S_{mix}||_2^2 \quad (1)$$

This training is done in parallel with stereo, thus the overall loss function is:

$$\mathcal{L} = \mathcal{L}_{stereo} + \lambda_{sep}\mathcal{L}_{sep}. \quad (2)$$

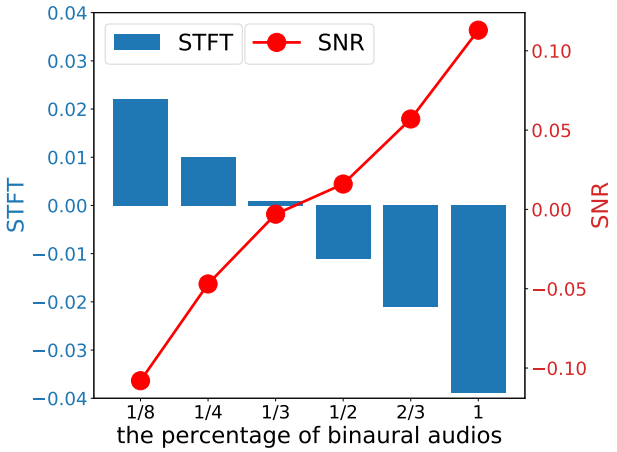

Figure 1: The curve of our relative performances *w.r.t* the percentage of binaural audios we use on FAIR-Play. It can be observed that our approach can reach their full performance using only $1/3$ of the ground-truth binaural audios.

The weight λ_{sep} is empirically set to 1 in our experiments. We observe that it would not affect the results much when $\lambda_{sep} \in [0.5, 1.5]$.

3. Implementation Details

Following [1], we use Python package Audiolab to resample the audio at 16kHz, which significantly accelerates the IO process during training. Besides, our follow the same details for our network structure and the training protocol as in Mono2Binaural [1] and Sep-Stereo [7]. The spectrograms are of size $257 \times 64 \times 2$, and the visual inputs are $224 \times 448 \times 3$ images. During testing, we find an inappropriate normalization operation in the demo-generating script presented in the public code of [1] and [7]. Specifically, since the ground-truth binaural audio is unknown in advance, the normalization step should be implemented in the mono audio instead. After fixing this bug, we discover the hop

Table 1: Quantitative results of binaural audio generation on Youtube-ASMR dataset with five evaluation metrics. Note that, our method PseudoBinaural still outperforms Mono-Mono on the non-musical dataset. Owing to the huge data amount and relatively simpler scenarios in Youtube-ASMR dataset, Augment-PseudoBinaural can only achieve a minor improvement over the supervised counterpart [1].

Method	STFT	ENV	Mag	$\mathcal{D}_{phase}$	SNR ↑
Mono-Mono	0.286	0.070	0.571	1.570	2.111
Mono2Binaural [1]	0.198	0.055	0.395	1.396	3.855
PseudoBinaural (Ours)	0.252	0.064	0.504	1.517	2.634
Augment-PseudoBinaural	**0.196**	**0.055**	**0.394**	**1.394**	**3.860**

length of sliding window will not affect the inference performance. Hence, the sliding window is set to 0.1s for all the experiments.

While creating the pseudo visual-stereo pairs, we leverage the off-the-shelf human detector and instrument detector of Faster RCNN [4, 3] to crop the visual patches from the videos with mono audios. The center of the cropped patches is placed in an arbitrary position on a blank background. For the generation of pseudo binaurals, the number of speakers in the virtual array is defined as $N = 8$ following the common practice. The speakers are uniformly placed around the frontal part of the head.

As for the activation map in Fig.4, it comes from the feature map of the last convolutional layer in the visual network. For a specific image input, the corresponding feature map is averagely pooled along the channel dimension and normalized to $[0, 1]$. And then, we deploy a bilinear upsampling operation on the feature map, making the size align with the original input image. In the end, we set the transparency ratio of upsampled feature map as 0.4, and combine it with the input image to obtain the final activation map.

4. More Ablation Studies

4.1. Ablation Study on the Amount of Binaural Recordings

Similar to the extensive analysis in [7], we provide an ablation study on the amount of binaural recording audios used in the augmentation experiments. Fig 1 shows the relative performance gains *w.r.t* the percentage of binaural audios used on FAIR-Play. The curve is drawn in a relative setting, where the performance of Mono2Binaural serves as the reference (zero in both metrics). As illustrated in Fig 1, Augment-PseudoBinaural can achieve the comparable performance based on only $1/3$ of the ground-truth binaural audios, which further demonstrates the effectiveness of our proposed method.

Table 2: Results for different mixes of K on FAIR-Play.

Mix ratio	$1:1:1$	$4:5:1$	$4:1:5$	$1:5:4$
STFT ↓	0.880	**0.878**	0.884	0.885
SNR ↑	5.312	**5.316**	5.310	5.306

4.2. Ablation Study on Mix Ratios of K

We conduct an ablation study on the selection of different mix ratios. The portions are the partitions of videos with one, two and three mixed sources used during training. During our implementation, the ratios are selected as $4:5:1$. The ablation results on FAIR-Play are listed in Table 2. It indicates that the influence of different mix ratios is minor.

5. Experiment on Youtube-ASMR Dataset

Youtube-ASMR introduced by Yang *et al*. [5] is a large-scale video dataset collected from YouTube. It consists of approximately 300K 10-second video clips with spatial audio and lasts about 904 hours in total. In an ASMR video, only an individual actor or "ASMRtist" is speaking or making different sounds towards a dummy head or the microphone arrays while facing the cameras. Compared to FAIR-Play or MUSIC-Stereo, the binaural scenarios in Youtube-ASMR dataset are relatively simpler since a typical sample just contains one sound source and the visual scene only involves a face in most cases. Similarly, the supervised method Mono2Binaural [1] and our approach are also implemented on this non-musical dataset. As shown in Tabel 1, PseudoBinaural can still surpass Mono-Mono on all the metrics, while the augmentation version only improves modestly over the supervised paradigm [1].

6. Limitations

There are also certain limitations in our work. The first is the domain gap between self-created images and real images. One direction towards solving this problem is through better blending techniques to the background. The second is that we do not specifically model room reverberations, particu-

larly cannot be adapted to any given environment. Domain adaptation techniques for vision might be useful for it. Moreover, our method does rely on the videos that contains only one visual and auditory sound source. Despite these limitations, our binaural-recording-free solution can still reach a satisfactory performance and be of enough contribution to the community.

References

[1] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2

[2] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 1

[3] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems (NeurIPS)*, 2015. 2

[5] Karren Yang, Bryan Russell, and Justin Salamon. Telling left from right: Learning spatial correspondence of sight and sound. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[6] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1

[7] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. *ECCV*, 2020. 1, 2