Sep-Stereo: Visually Guided Stereophonic Audio Generation by Associating Source Separation - Supplementary Materials -

Hang Zhou*, Xudong Xu*, Dahua Lin, Xiaogang Wang, and Ziwei Liu

CUHK - SenseTime Joint Lab, The Chinese University of Hong Kong
{zhouhang@link,xx018@ie,dhlin@ie,xgwang@ee}.cuhk.edu.hk
zwliu.hust@gmail.com

1 Demo Video Settings

This section describes the settings of the demo video for paper Sep-Stereo [3]. The video can be found at https://hangz-nju-cuhk.github.io/projects/Sep-Stereo.

1.1 Stereophonic Audio Generation

Experiments on FAIR-Play In our video, we first showcase of our stereophonic audio generation results on the standard FAIR-Play [1] dataset. In the first three cases, we show the results of ours compared with that of Mono2Binaural [1], by using directly the videos reported by their paper. We find the validation sets these videos lying in and use the corresponding models for generation. Although the difference is not significant, one can still find that our method can create examples with more precise directional information than theirs.

Experiments on MUSIC We then attempt to generate stereophonic audios for both in-the-wild data and synthetic duet data on MUSIC [2] dataset. Please be noted that all audio samples we use in the MUSIC dataset are converted to mono, thus no stereophonic training is applicable on this dataset. The results of the three models are shown:

- **Mono2Bianural** [1]. We use this baseline model trained on FAIR-Play. Notice that this method cannot be trained with the mono data in MUSIC inherently.
- **Sep-Stereo (Ours)**. Then is our whole Sep-Stereo model that is trained on the mono part of MUSIC and only 50% of the data on split 1 of FAIR-Play. We find that adding more data has subtle effects on human perceptions.
- **Unsupervised Sep-Stereo**. This is our model with only separative and no stereophonic learning. In other words, we train our model only on the mono data of MUSIC without stereophonic supervision. Thus, it is a completely

* Equal Contribution.

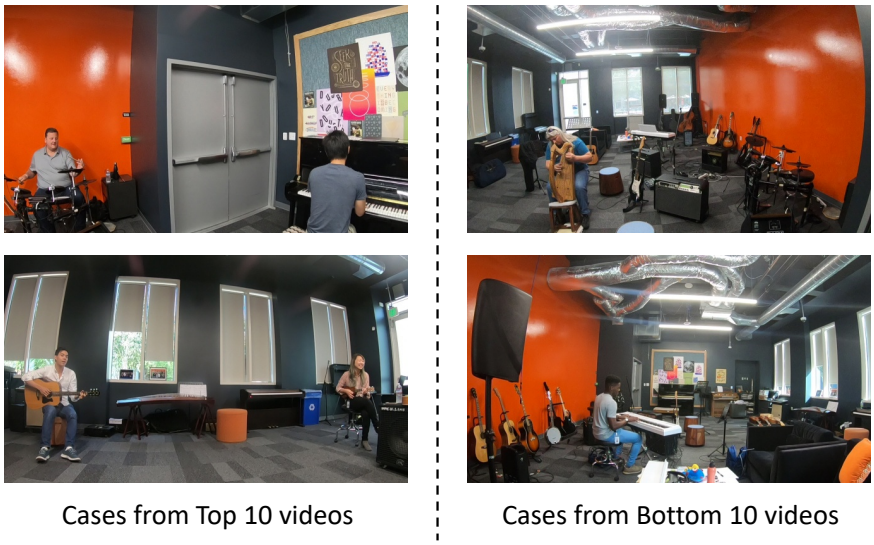

Fig. 1. The cases from the top 10 and bottom 10 videos where our method outperforms Mono2Binaural [1]

unsupervised setting. The idea is to show a rough prediction of separation results. During inference, the whole visual feature map is max-pooled to size $[6, 3, C_v]$ and average-pooled to $[2, 1, C_v]$. We then split them to two $[1, 1, C_V]$ feature maps according to their positions and rearrange them as F_A and F_B in equation (8) of the full paper.

The results are shown in the second part of the video. We can see that while Mono2Binaural has only subtle influence under this setting, our Sep-Stereo can generate more plausible results according to the scenarios. As the distances between players in MUSIC are much smaller than that in FAIR-Play, it is reasonable to create less intense stereophonic audios. The results of our unsupervised model are over-intense as the model is borrowed from source separation. However, it is already a great advantage of our model to achieve stereophonic audio generation in an unsupervised manner.

1.2 Audio Source Separation

Finally, We show our source separation results qualitatively on MUSIC dataset. We mix the mono videos from the validation set as inputs. We explicitly compare our method with PixelPlayer [2], which is designed only for separation. It can be found that our results can outperform theirs for certain channels, which proves that we can achieve at least comparable results as PixelPlayer.

Table 1. More ablation results on FAIR-Play dataset

Metric	w/o L_D	Vertical Extreme	Unet out	Ours
STFT_D	0.976	0.919	0.933	0.879
ENV_D	0.145	0.139	0.140	0.135

2 Detailed Comparison with Mono2Binaural

The differences between our method and that of Mono2Binaural [1] are that 1) we propose separative learning, a completely new training diagram to leverage mono data for enriching stereo learning; and 2) we propose APNet for better visual-audio association in a more explainable way.

We compare our results with Mono2Binaural [1], both trained on the official split1 on FAIR-Play. The 187 results on the validation set are carefully examined. We rank the videos according to the performance gains between ours and [1] w.r.t STFT differences. The larger the value, the better our result is.

1) Among the top 10 videos, we observe that the layouts are mostly with simple backgrounds. Sound sources are located at the far left and right sides of the view. This is reasonable and consistent with our designed framework with feature rearrangement. The results show that our separative learning indeed strengthened the learning of sound source horizontally.

2) Among the bottom 10 videos, the layouts are normally complicated with walls at two sides. The instruments are the rarely-appeared ones in both FAIR-Play and MUSIC. This also reveals the limitation of our work and can be improved in the future.

The arriving time and intensity difference of audios between left and right ears are the base of human sound localization ability. Thus the conclusion that our method can be more distinctive with horizontal cases is of great importance to stereo. Moreover, no previous work has shown results on both the tasks of separation and stereo simultaneously before.

3 More Ablations

In this section, we provide more discussions together with ablation studies to make the design of our network clearer. The experiments are conducted on the FAIR-Play [1] dataset. The results are shown in Table 1

The loss L_D at the output of the U-Net. We found that the loss of L_D is essential for stabling the training process. However, experiments show that the output results from the left and right branch independently from the APNet is better than the direct prediction (UNet out) of the difference map.

The horizontal placement in feature rearrangement. One may argue that we neglect the elevation differences in audios by placing the visual feature maps only at the left and right edges in separative learning, thus an experiment that places the feature maps at vertical extremes could be conducted. The discussions are as follows:

1) We choose only the horizontal extreme points based on the fact that human ears are distributed horizontally. Vertical information is difficult to represent with left-right channels.

2) Experiments with "Vertical Extremes" can be conducted by forming spatially inconsistent pairs (visual top to left channel and bottom to right), but the setting itself is somehow not reasonable. The results in the table 1 show that the network benefits more from our normal layout.

3) However, the elevation issue is indeed a limitation of our work. Also, there is a domain gap between our self-created visual feature map and the directly extracted ones.

References

1. Gao, R., Grauman, K.: 2.5 d visual sound. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 1, 2, 3
2. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018) 1, 2
3. Zhou, H., Xu, X., Lin, D., Wang, X., Liu, Z.: Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 1