2.5D Visual Sound

Ruohan Gao*
The University of Texas at Austin
rhgao@cs.utexas.edu

Kristen Grauman
Facebook AI Research
grauman@fb.com†

Abstract

Binaural audio provides a listener with 3D sound sensation, allowing a rich perceptual experience of the scene. However, binaural recordings are scarcely available and require nontrivial expertise and equipment to obtain. We propose to convert common monaural audio into binaural audio by leveraging video. The key idea is that visual frames reveal significant spatial cues that, while explicitly lacking in the accompanying single-channel audio, are strongly linked to it. Our multi-modal approach recovers this link from unlabeled video. We devise a deep convolutional neural network that learns to decode the monaural (single-channel) soundtrack into its binaural counterpart by injecting visual information about object and scene configurations. We call the resulting output 2.5D visual sound*—the visual stream helps "lift" the flat single channel audio into spatialized sound. In addition to sound generation, we show the self-supervised representation learned by our network benefits audio-visual source separation. This paper summarizes our key ideas and results of our recent conference paper*[1] *[1]. Our video results:* http://vision.cs.utexas.edu/projects/2.5D_visual_sound/

1. Introduction

Multi-modal perception is essential to capture the richness of real-world sensory data and environments. People perceive the world by combining a number of simultaneous sensory streams, among which the visual and audio streams often carry paramount information. In particular, both audio and visual data convey significant *spatial* information. We see where objects are and how the room is laid out. We also hear these things: sound-emitting objects indicate their location, and sound reverberations reveal the room's main surfaces, materials, and dimensions. Similarly, as in the famous cocktail party scenario, while having a conversation at a noisy party, one can hear another voice calling out and turn to face it. The two senses naturally work in concert to interpret spatial signals.

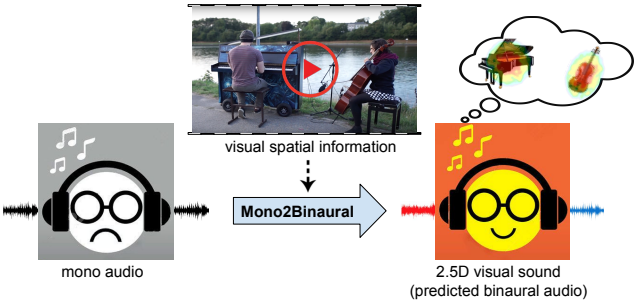

Figure 1: Binaural audio creates a 3D soundscape for listeners, but such recordings remain rare. The proposed approach infers *2.5D visual sound* by injecting the spatial information contained in the video frames accompanying a typical monaural audio stream.

The key insight of this work is that video accompanying monaural audio has the potential to unlock spatial sound, lifting a flat audio signal into what we call "2.5D visual sound". Although a single channel audio track alone does not encode any spatial information, its accompanying visual frames do contain object and scene configurations. For example, as shown in Fig. 1, we observe from the video frame that a man is playing the piano on the left and a man is playing the cello on the right. Although we cannot sense the locations of the sound sources by listening to the mono recording, we can nonetheless anticipate what we *would* hear if we were personally in the scene by inference from the visual frames.

We introduce an approach to realize this intuition. Given unlabeled video as training data, we devise a MONO2BINAURAL deep convolutional neural network to convert monaural audio to binaural audio by injecting the spatial cues embedded in the visual frames. Our encoder-decoder style network takes a mixed single-channel audio and its accompanying visual frames as input to perform joint audio-visual analysis, and attempts to predict a two-channel binaural audio that agrees with the spatial configu-

*Work done during an internship at Facebook AI Research.

†On leave from The University of Texas at Austin (grauman@cs.utexas.edu).

[1]To appear at 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition.

	FAIR-Play		REC-STREET		YT-CLEAN		YT-MUSIC	
	STFT	ENV	STFT	ENV	STFT	ENV	STFT	ENV
Ambisonics [4]	-	-	0.744	0.126	1.435	0.155	1.885	0.183
Audio-Only	0.966	0.141	0.590	0.114	1.065	0.131	1.553	0.167
Flipped-Visual	1.145	0.149	0.658	0.123	1.095	0.132	1.590	0.165
Mono-Mono	1.155	0.153	0.774	0.136	1.369	0.153	1.853	0.184
MONO2BINAURAL (Ours)	**0.836**	**0.132**	**0.565**	**0.109**	**1.027**	**0.130**	**1.451**	**0.156**

Table 1: Quantitative results of binaural audio prediction on four diverse datasets. Lower is better.

rations in the video. When listening to the predicted binaural audio, listeners can then feel the locations of the sound sources as they are displayed in the video. Moreover, we show that the MONO2BINAURAL conversion process also benefits audio-visual source separation, a key challenge in audio-visual analysis.

2. Overview of Proposed Approach

We denote the signal received at the left and right ears by $x^L(t)$ and $x^R(t)$, respectively. If we mix the two channels into a single channel $x^M(t) = x^L(t) + x^R(t)$, then all spatial information collapses. We can formulate a self-supervised task to take the mixed monaural signal $x^M(t)$ as input and split it into two separate channels $\tilde{x}^L(t)$ and $\tilde{x}^R(t)$, using the original $x^L(t)$, $x^R(t)$ as ground-truth during training. However, this is a highly under-constrained problem, as $x^M(t)$ lacks the necessary information to recover both channels. Our key idea is to guide the MONO2BINAURAL process with the accompanying video frames, from which *visual* spatial information can serve as supervision.

Instead of directly predicting the two channels, we predict the difference of the two channels:

$$x^D(t) = x^L(t) - x^R(t). \quad (1)$$

More specifically, we operate on the frequency domain and perform short-time Fourier transform (STFT) [2] on $x^M(t)$ to obtain the complex-valued spectrogram $\mathbf{X}^M$, and the objective is to predict the complex-valued spectrogram $\mathbf{X}^D$ for $x^D(t)$:

$$\mathbf{X}^M = \{\mathbf{X}^M_{t,f}\}^{T,F}_{t=1,f=1}, \quad \mathbf{X}^D = \{\mathbf{X}^D_{t,f}\}^{T,F}_{t=1,f=1}, \quad (2)$$

where t and f are the time frame and frequency bin indices, respectively, and T and F are the numbers of bins. Then we obtain the predicted difference signal $\tilde{x}^D(t)$ by the inverse short-time Fourier transform (ISTFT) [2] of $\mathbf{X}^D$. Finally, we recover both channels—the binaural audio output:

$$\tilde{x}^L(t) = \frac{x^M(t) + \tilde{x}^D(t)}{2}, \quad \tilde{x}^R(t) = \frac{x^M(t) - \tilde{x}^D(t)}{2}. \quad (3)$$

We devise a MONO2BINAURAL deep network that performs audio spatialization. The network takes the mono audio $x^M(t)$ and visual frames as input and predicts $x^D(t)$. We extract visual features from the center frame of the audio segment using ResNet-18 [3], which is pre-trained on ImageNet. On the audio side, we adopt a U-NET [5] style architecture to extract audio feature, combined with the visual features to perform binaural audio prediction. We train our MONO2BINAURAL network using L2 loss to minimize the distance between the ground-truth complex spectrogram and the predicted one. Finally, using ISTFT, we obtain the predicted difference signal $\tilde{x}^D(t)$, through which we recover the two channels $\tilde{x}^L(t)$ and $\tilde{x}^R(t)$ as defined in Eq. 3.

3. Example Results

We validate our approach for generation and separation. We use four challenging datasets: FAIR-Play, REC-STREET, YT-CLEAN and YT-MUSIC [4]. FAIR-Play dataset[2] collected by us is the first dataset of its kind that contains videos of professional recorded binaural audio. REC-STREET is a dataset of outdoor street scenes, and YT-CLEAN and YT-MUSIC contain ∼1,000 "in the wild" videos from YouTube. These videos contain diverse scenes such as meeting rooms, travel, sports, *etc*.

We evaluate the quality of our predicted binaural audio by comparing to the following baselines: 1) Ambisonics [4]: predicting ambisonics using the pre-trained models provided by [4]; 2) Audio-Only: a baseline that removes the visual stream and uses only audio as input; 3) Flipped-Visual: flipping visual frames to perform prediction using the wrong visual information; 4) Mono-Mono: a straightforward baseline that copies the mixed monaural audio onto both channels to create a fake binaural audio.

Table 1 shows the binaural generation results. Our method outperforms all baselines consistently on all four datasets. Our MONO2BINAURAL approach performs better than the Audio-Only baseline, indicating the visual stream is essential to guide conversion. Note that the Audio-Only baseline uses the same network design as our method, so it has reasonably good performance as well. Flipped-Visual performs much worse, demonstrating that the network requires the correct visual spatial information in order to predict binaural audio correctly. The Ambisonics [4] approach also does not do as well. Please see our video results[3].

[2] https://github.com/facebookresearch/FAIR-Play

[3] http://vision.cs.utexas.edu/projects/2.5D_visual_sound/

References

[1] R. Gao and K. Grauman. 2.5d visual sound. In *CVPR*, 2019. 1

[2] D. Griffin and J. Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1984. 2

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[4] P. Morgado, N. Vasconcelos, T. Langlois, and O. Wang. Self-supervised generation of spatial audio for 360° video. *arXiv preprint arXiv:1809.02587*, 2018. 2

[5] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015. 2