2.5D Visual Sound
(Supplementary Materials)

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

Kristen Grauman
Facebook AI Research
grauman@fb.com

The supplementary materials consist of:

A. Supplementary video.

B. Details of MONO2BINAURAL network.

C. Details of MIX-AND-SEPARATE network.

D. Implementation Details.

E. User study interface.

A. Supplementary Video

In our supplementary video[1], we show (a) examples of our professional recorded binaural audios, (b) example results of binaural audio prediction, and (c) example results of audio-visual source separation.

B. Details of MONO2BINAURAL Network

Our MONO2BINAURAL network consists of a visual branch and an audio branch. The visual branch takes images of dimension 224x448x3 as input to extract a feature map of dimension 14x7x512 through ImageNet pre-trained ResNet-18 network. The visual feature map is then passed though a 1x1 convolution layer to reduce the channel dimension, producing a feature map of dimension 14x7x8.

The audio branch is of a U-NET style architecture, namely an encoder-decoder network with skip connections. It consists of 5 convolution layers and 5 up-convolution layers. All convolutions and up-convolutions use 4 x 4 spatial filters applied with stride 2, and followed by a BatchNorm layer and a ReLU. After the last layer in the decoder, an up-convolution is followed by a Sigmoid layer to bound the values of the complex mask. The encoder uses leaky ReLUs with a slope of 0.2, while ReLUs in the decoder are not leaky. Skip connections are added between each layer i in the encoder and layer $n-i$ in the decoder, where n is the total number of layers. The skip connections concatenate activations from layer i to layer $n-i$.

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

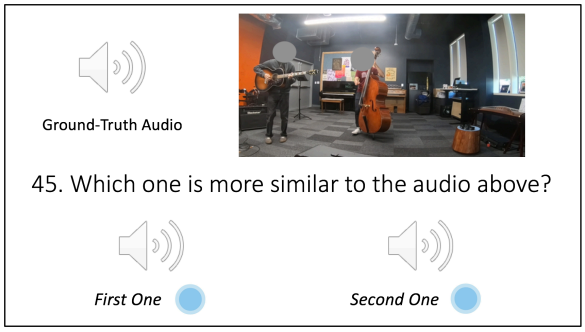
C. Details of MIX-AND-SEPARATE Network

For audio-visual source separation, we use the same base architecture as our MONO2BINAURAL network except that now the input to the network is a pair of training video clips. Two visual branches of shared weights are used and each takes the frame of one video as input to extract visual features. The audio branch takes the mixed audio as input to extract audio features, and is combined with the visual features to predict a mask for each video. Following Zhao *et al*. ECCV'18, we use ratio masks and log magnitude spectrograms. For ratio masks, the ground truth mask of a video is calculated as the ratio of the magnitude spectrogram of the target sound and the mixed sound.

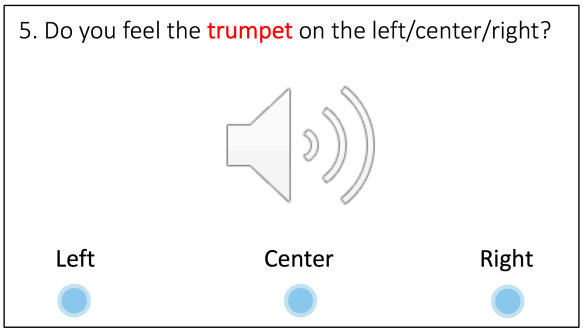
D. Implementation Details

For MONO2BINAURAL training, we randomly sample audio segments of length 0.63s from each 10s audio clip and normalize each segment's RMS level to a constant value. Then we obtain a complex spectrogram of size $257 \times 64 \times 2$ for each channel. For each sampled audio segment, the center video frame is used as the accompanying visual frame and resized to 480×240. We randomly crop 448×224 images and use color and intensity jittering as data augmentation. The network is trained using an Adam optimizer with weight decay 5×10^{-4} and batch size 256. The starting learning rate is set to 0.001, and decreased by 6% every 10 epochs and trained for 1,000 epochs. We use smaller starting learning rate 0.0001 for ResNet-18 because it is pre-trained on ImageNet.

For audio-visual source separation training, we randomly sample pairs of video and take an audio segment of length 2.55s from each video. We mix the two audio segments, and obtain a log magnitude spectrogram of size 257×256 for each channel. A random frame within each audio segment is used as the accompanying visual frame for both videos. The network is trained with batch size 128 and learning rate 0.001, and a smaller learning rate 0.0001 is used for ResNet-18.



(a) Example of user study 1



(b) Example of user study 2

Figure 1: Examples of the interface for the two user studies to test how listeners perceive the predicted binaural audio.

E. User Study Interface

In Fig. 1, we show examples of our user study interface. In the first user study (Fig. 1a), the participants listen to a 10s ground-truth binaural audio and see the accompanying visual frame. Then they listen to two predicted binaural audios generated by our method or a baseline (Ambisonics, Audio-Only, or Mono-Mono). After listening to each pair, participants are asked whether the first one or the second one creates a better 3D sensation that matches the ground-truth binaural audio. In the second user study, we ask every participant to *only listen* to the ground-truth or predicted binaural audio from our method or a baseline, and then choose the direction the sound of a specified instrument is coming from. For example, as shown in Fig. 1b, participants first listen to the audio, and then they are asked to choose whether they feel the trumpet on the left, in the center, or on the right.