Synthetic Acoustic Image Generation for Audio-Visual Localization

Valentina Sanguineti[1,2], Pietro Morerio[1], Alessio Del Bue, [3] Vittorio Murino[1,4,5]
[1]Pattern Analysis and Computer Vision, Istituto Italiano di Tecnologia
[2]University of Genova [3]Visual Geometry and Modelling, Istituto Italiano di Tecnologia
[4]University of Verona [5]Huawei Technologies Ltd., Ireland Research Center

Abstract

Acoustic images have the peculiarity to distinguish the spectral signature of sounds coming from different directions in space and they provide richer information than the one derived from mono and binaural microphones. However, they are generated by microphone arrays, which are not as widespread as ordinary microphones mounted on optical cameras. We propose to leverage the generation of synthetic acoustic images from common audio-video data for the task of audio-visual localization[1].

1. Introduction

Humans interpret the world through vision and hearing mostly. More specifically, vision is supported by binaural hearing. In fact, sound signals are received with a certain delay between the left and right ear, as well as a slight difference in intensity, which are critical to perceive spatial clues about the sources of sound. Besides, humans are able to fuse the spatial clues elaborated from their auditory system with those coming from their sight.

To mimic human capabilities, binaural microphone configurations have been lately investigated in audiovisual learning [12, 1]. However, binaural configurations are limited to the estimation of the direction of arrival only along the azimuth direction. In this work, we exploit instead the data gathered by a planar array of microphones, which produce more accurate spatial audio information than stereo audio. In fact, the acoustic signals acquired by an array can be combined via a filter-and-sum beamforming algorithm to produce an *acoustic image* allowing to localize sound sources on a 2-dimensional space, as we can see in Figure 1, rather than along just one single direction [12]. Each pixel of such images contains the spectral signature of the sound coming from the corresponding direction in space. Acoustic images have already been studied by [5], for supervised learning and for distilling their content to audio models, while [8] used them for audio-visual self-supervised learning. Both works showed they are useful to learn good representations. However, they are typically generated by cumbersome microphone arrays. Therefore, we propose to synthesize them from the associated video and its corresponding monaural audio signal, drawing inspiration from methods used for sound *spatialization*, which separate binaural audio from mono audio [12, 1]. In this way we can reconstruct them even without an array of microphones.

[1]This workshop paper is a short version of [7], accepted at AAAI 2021.

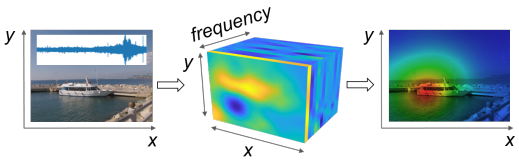

Figure 1: We achieve unsupervised sound source localization through the generation of a spatialized audio, called *acoustic image*: starting from an RGB frame and the corresponding monaural audio (left), we synthesize the spectral signature of the sounds associated with each considered direction, namely each acoustic pixel in the acoustic image (center). Localization is then obtained by extracting the energy of sound (right).

To solve this problem, we propose a novel architecture, which is a hybrid of a Variational Autoencoder (VAE) and a U-Net models. We assess the quality of the generated synthetic acoustic images by common reconstruction metrics and more importantly, on the task of audio-visual localization, by extracting the energy of the spatialized sound. We show that energy is useful to accurately locate the region originating the sound thanks to the precise supervision of the spatial sound distribution provided by the acoustic images. Our model is evaluated by considering both multimodal datasets containing acoustic images and unseen datasets containing just monaural audio signals and RGB frames, showing to transfer well on new domains and reach more accurate localization than previous state-of-the-art models on unseen data. Works for sound localization [4, 3, 9], in fact, usually exploit two-stream deep network architectures to leverage the correlation between a visual object and the corresponding sound, which might be not so reliable as not focusing solely on the region from where the sound was originated, but often from the entire object. In summary, the contributions of our paper are:

1. A novel audio-visual localization, through the generation of acoustic images and by estimating the energy of the synthesized spatialized sound.

2. A new multimodal learning architecture, trained in a self-supervised way, to generate synthetic acoustic images, by jointly processing monaural audio signals and associated RGB images.

3. A set of experiments to evaluate the quality of the reconstruction of in terms of classification and localization. Moreover, ground truth acoustic images allow for a fair evaluation of the sound source localization task, as they are bias-free from human annotations. We also verify that our method generalizes better to datasets never seen in training.

2. Method: Audiovisual U-VAE

The proposed architecture resembles a VAE with skip connections as in the U-Net model, to exploit the upsides of both: VAEs are very effective generation tools but they show limitation when the size of the output is too large; on the contrary, U-Nets are reconstruction tools which can effectively deal with the details. We name this model U-VAE. VAE can improve reconstruction with respect to using a simple autoencoder since its latent loss acts as a regularizer and by sampling the latent variable VAE can generate data with more variability. We did some ablation studies that show that autoencoders generalize less effectively to different datasets [7]. VAEs are trained to maximize the Evidence Lower Bound (ELBO), which maximizes the log probability of likelihood of generating data similar to real ones $p(\mathbf{x})$:

$$ELBO = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})] - \beta KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})), \quad (1)$$

where $\beta = 1$. The opposite of the first addendum of ELBO in Eq. 1 is often interpreted as a reconstruction loss. The Kullback-Leibler term KL is the latent loss. As proposed by [2], β can be an adjustable hyperparameter that balances the two terms as one regards latent independence constraint and the other one reconstruction accuracy. They propose to consider $\beta > 1$ for good disentangled representations. Instead, we are more interested in obtaining good reconstruction, therefore we weight latent loss using $\beta < 1$, choosing β so that the reconstruction loss and latent loss have the same order of magnitude.

We generate acoustic images starting from monaural audio samples and the corresponding video frame, to provide spatial cues which are missing in omnidirectional microphones.

Ground truth acoustic images are computed from the raw audio signals of the microphones of a planar array combining them with the filter-and-sum beamforming algorithm. They are volumes, with channels corresponding to frequency bins, which were compressed to Mel-Frequency Cepstral Coefficients (MFCC) representation [5], according to audio human perception characteristics, reducing consistently the computational complexity. Thus, acoustic images contain the frequency information for each acoustic pixel represented with MFCC. To simplify the generation task, we feed the MFCC of a single microphone tiled along spatial dimensions, rather than raw waveforms or spectrograms, in order to have homogeneous input-output.

Visual features are extracted using ResNet50, pretrained on ImageNet, modified with the removal of global average pooling and the addition of a 2D convolution layer to match spatial dimensions of the audio encoder. We train the last ResNet50 layer only, to focus on the specific regions producing sound in the considered training datasets. The visual feature map is then concatenated with the last feature map produced by the audio encoder before sampling as shown in Figure 2.

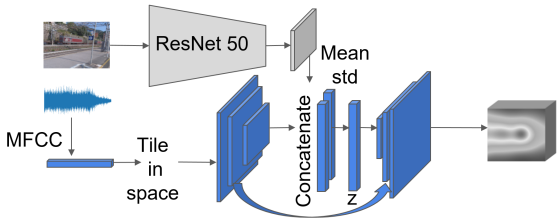

Figure 2: We propose an architecture based on VAE and U-Net (U-VAE) to generate acoustic images. The inputs are monaural audio samples and the corresponding video frame. We compress audio samples to MFCC. ResNet50 visual features are concatenated to audio encoder features.

The network is trained to reconstruct acoustic images for the time interval $1/12\,\mathrm{s}$ as the ground truth acoustic images and RGB images frame rates are 12 frames/s. Therefore, we provide in input MFCC corresponding to $1/12\,\mathrm{s}$ of sound and relative RGB frame. This allows to have almost a real-time estimate of the directional sound, whereas previous works considered from $1\,\mathrm{s}$ [11] up to $20\,\mathrm{s}$ [9] of audio to visually localize the sound. Furthermore, considering one frame only for a long audio track can lead to miss important cues about synchronization.

3. Experiments

In this section we first describe the employed datasets. Subsequently, we assess the reconstruction capability of our U-VAE. Finally, we evaluate audio-visual localization both quantitatively and qualitatively.

3.1. Datasets

We consider the following datasets:

- *ACIVW* [8] is a multimodal dataset including acoustic images containing 5 hours of videos acquired in the wild representing 10 classes.
- *AVIA* [5] is a multimodal dataset including acoustic images with 14 different actions with a characteristic sound performed in 3 scenarios with different noise conditions.
- A random subset of *Flickr-SoundNet* employed by [9], which includes sounds sources positions annotated by three subjects, which facilitates quantitative evaluation. We consider just the test data, which includes 250 pairs of frames and their corresponding sound.
- *VGGSound* is a dataset with over 200k 10s video clips containing an object making sound for 300 audio classes from YouTube videos.

We use the first two datasets for both training and testing. The remaining two are instead used in testing to evaluate the generalization capability of our U-VAE on unseen domains.

3.2. Evaluation of Reconstruction

In Table 1 we evaluate the reconstruction of acoustic images for both the test sets of ACIVW and AVIA datasets. This is done by using the following metrics:

- **Mean square error (MSE)** measures the reconstruction error for each acoustic pixel.
- **GAN-test [10]** measures the accuracy of a classifier trained on real acoustic images but evaluated on generated images (we evaluate also on real ones) to quantify semantic similarity to real samples. To classify acoustic images, we consider the DualCamNet network introduced by [5].
  We see that when training on ACIVW dataset we have only a 1% drop if testing on generated acoustic images. AVIA dataset has a bigger drop, 16%, as its acoustic images were collected in noisy scenarios and contain periodic sounds. We also test on synthetic acoustic images created by replicating single-microphone MFCC along the 2 spatial dimensions. We see that the drop in accuracy is huge: 30% for ACIVW and 63% for AVIA, showing that our architecture is essential to generate different MFCC for each acoustic pixel, namely to modulate sound in space.
- **GAN-train [10]** measures the accuracy of a classifier trained on generated data and evaluated on real test images (we evaluate also on generated ones). GAN-train metric captures the diversity of generated samples.
  The best result is obtained on ACIVW dataset, where we have only a 9% drop when testing on real samples. When testing on AVIA, instead, the drop is 14%. Nevertheless, we notice that on generated data we have good results for both datasets when testing on generated data.
  Last, we train the classifier on uniform acoustic images artificially created by replicating MFCC from a single microphone and testing on real acoustic images. This experiment is designed to show how our U-VAE is actually modulating MFCC for each spatial direction. Furthermore, when both training and testing on replicated single microphone MFCC we get worse performance than when training and testing from acoustic images (GAN-test on real), proving that spatialized audio allows increasing classification accuracy.

	Test	ACIVW	AVIA
MSE	-	1.1426±0.0053	0.9483±0.0026
GAN-test	real	0.8497±0.0014	0.8383±0.0022
	gen.	0.8342±0.0093	0.6700±0.0009
	MFCC	0.5410±0.0175	0.2091±0.0027
GAN-train (on gen.)	gen.	0.8512±0.0089	0.7871±0.0039
	real	0.7661±0.0065	0.6456±0.0100
GAN-train (on MFCC)	MFCC	0.7323±0.0072	0.6614±0.0038
	real	0.4270±0.0186	0.1307±0.0119

Table 1: Reconstruction metrics for AVIA and ACIVW models. MSE values are multiplied by 10^{-2}. We specify considered test modalities: real acoustic images, generated acoustic images, tiled MFCC from a single microphone.

Train	Test	AUC	IoU
ACIVW	ACIVW	59.7±0.2	76.8±0.2
AVIA	AVIA	51.2±0.3	54.4±0.7
Train	**Test**	**AUC**	**cIoU**
Senocak 1 [9]	Flickr-SoundNet (subset)	44.9	43.6
Senocak 2 [9]		51.2	52.4
Senocak 3 [9]		55.8	66.0
ACIVW		50.3±0.5	53.1±1.9
AVIA		37.2±1.8	20.1±3.0
Hu, Nie, and Li 1 [3]		45.2	41.6
Hu et al. [4]		49.2	50.0
Qian et al. [6]		49.6	52.2
Hu, Nie, and Li 2 [3]		56.8	67.1

Table 2: Audio-visual localization of ACIVW and AVIA models compared with other benchmarks. [9] 1: Unsupervised 10k, [9] 2: Unsupervised 144k ReLU, [9] 3: Unsupervised 144k, [3] 1: Unsupervised 20k AudioSet, [3] 2: Unsupervised 400k Flickr-SoundNet.

3.3. Audio-Visual Localization

We evaluate now localization results both quantitatively and qualitatively firstly on ACIVW, AVIA using intersection over union (IoU) and area under the curve (AUC), then on Flickr-SoundNet using consensus IoU and AUC, whereas we have no ground truth for VGGSound dataset for which can only show some qualitative results.

Results for ACIVW and AVIA

1. Quantitative Results

Given synthetic and true energy, we evaluate our results quantitatively using IoU and AUC. The results are at the top of Table 2. We see that training and testing on ACIVW dataset we have a better result than when training and testing on AVIA because we have more data and less noise.

2. Qualitative Results

As regards ACIVW dataset, the energy of our reconstruction in Figure 3b is very similar to the energy of real test samples in Figure 3a. As it can be noticed from the first row of Figure 3b, the reconstructed image is sometimes even less noisy than the ground truth one.

The AVIA dataset is a more challenging benchmark not only because of noise present in some scenarios but also due to the periodicity of considered sounds: in some frames we do not have any sound but only background noise, such as in the second row of Figure 3c, so that it is difficult to match sound and video. On the contrary, ACIVW dataset contains continuous sound and energy is always mapping with visual objects. As we can see in the first row of Figure 3d, in the anechoic chamber the sound localization is very precise because the sound is present and there is little noise (compare to real energy in Figure 3c). In the last row of Figure 3d we see that also in AVIA we can sometimes improve sound localization when there is noise in the original image.