

Learning to Separate Object Sounds by Watching Unlabeled Video

Ruohan Gao

UT Austin

rhgao@cs.utexas.edu

Rogerio Feris

IBM Research

rsferis@us.ibm.com

Kristen Grauman

UT Austin

grauman@cs.utexas.edu

1. Introduction

Understanding scenes and events is inherently a multi-modal experience. We perceive the world by both looking and listening (and touching, smelling, and tasting). Objects generate unique sounds due to their physical properties and interactions with other objects and the environment. For example, perception of a coffee shop scene may include seeing cups, saucers, people, and tables, but also hearing the dishes clatter, the espresso machine grind, and the barista shouting an order. Human developmental learning is also inherently multi-modal, with young children quickly amassing a repertoire of objects and their sounds: dogs bark, cats mew, phones ring.

However, while recognition has made significant progress by “looking”—detecting objects, actions, or people based on their appearance—it often does not listen. Objects in video are often analyzed as if they were silent entities in silent environments. A key challenge is that in a realistic video, object sounds are observed not as separate entities, but as a single audio channel that mixes all their frequencies together. Audio source separation remains a difficult problem with natural data outside of lab settings. Existing methods perform best by capturing the input with multiple microphones, or else assume a clean set of single source audio examples is available for supervision (e.g., a recording of only a violin, another recording containing only a drum, etc.), both of which are very limiting prerequisites. The blind audio separation task evokes challenges similar to image segmentation—and perhaps more, since all sounds overlap in the input signal.

Our goal is to learn how different objects sound by both looking at and listening to unlabeled video containing multiple sounding objects. We propose an unsupervised approach to disentangle mixed audio into its component sound sources. The key insight is that observing sounds in a variety of visual contexts reveals the cues needed to isolate individual audio sources; the different visual contexts lend weak supervision for discovering the associations. For example, having experienced various instruments playing in various combinations before, then given a video with a guitar and saxophone (Fig. 1), one can naturally anticipate what sounds could be present in the accompanying audio, and therefore better separate them. Indeed, neuroscientists report that the mismatch negativity of event-related brain

Figure 1. Goal: audio-visual object source separation in videos.

potentials, which is generated bilaterally within auditory cortices, is elicited only when the visual pattern promotes the segregation of the sounds [6]. This suggests that synchronous presentation of visual stimuli should help to resolve sound ambiguity due to multiple sources, and promote either an integrated or segregated perception of the sounds.

We introduce a novel audio-visual source separation approach that realizes this intuition. Our method first leverages a large collection of unannotated videos to discover a latent sound representation for each visible object. Specifically, we use state-of-the-art image recognition tools to infer the objects present in each video clip, and we perform non-negative matrix factorization (NMF) on each video’s audio channel to recover its set of frequency basis vectors. At this point it is unknown which audio bases go with which visible object(s). To recover the association, we construct a neural network for multi-instance multi-label learning (MIML) that maps audio bases to the distribution of detected visual objects. From this audio basis-object association network, we extract the audio bases linked to each visual object, yielding its prototypical spectral patterns. Finally, given a novel video, we use the learned per-object audio bases to steer audio source separation.

2. Overview of Proposed Approach

Single-channel audio source separation is the problem of obtaining an estimate for each of the J sources s_j from the observed linear mixture $x(t)$: $x(t) = \sum_{j=1}^J s_j(t)$, where $s_j(t)$ are time-discrete signals. The mixture signal can be transformed into a magnitude or power spectrogram, which encode the change of a signal’s frequency and phase content over time. We operate on the frequency domain, and use the inverse short-time Fourier transform (ISTFT) to reconstruct the sources.

The training pipeline is illustrated in Fig. 2. Given an unlabeled video, we extract its visual frames and the corresponding audio track. Then, we perform NMF independently on its audio magnitude spectrogram to obtain its

Figure 2. Unsupervised training pipeline. For each video, we perform NMF on its audio magnitude spectrogram to get M basis vectors. An ImageNet-trained ResNet-152 network is used to make visual predictions to find the potential objects present in the video. Finally, we perform multi-instance multi-label learning to disentangle which extracted audio basis vectors go with which detected visible object(s).

Figure 3. Our deep multi-instance multi-label network takes a bag of M audio basis vectors for each video as input, and gives a bag-level prediction of the objects present in the audio. The visual predictions from an ImageNet-trained CNN are used as weak “labels” to train the network with unlabeled video.

spectral patterns. M audio basis vectors are extracted from each video. For the visual frames, we use an ImageNet pre-trained ResNet-152 network [2] to make object category predictions, and we max-pool over predictions of all frames to obtain a video-level prediction. The top labels are used as weak “labels” for the unlabeled video. The extracted basis vectors and the visual predictions are then fed into our MIML learning framework to discover associations.

We design a deep MIML network (see Fig. 3) for our task. A bag of basis vectors is the input to the network, and within each bag there are M basis vectors extracted from one video. The “labels” are only available at the bag level, and come from noisy visual predictions of the ResNet-152 network trained for ImageNet recognition. The labels for each instance are unknown. We incorporate MIL into the deep network by modeling that there must be at least one audio basis vector from a certain object that constitutes a positive bag, so that the network can output a correct bag-level prediction that agrees with the visual prediction. We use the multi-label hinge loss to train our MIML network.

The MIML network learns from audio-visual associations, but does not itself disentangle them. To collect high quality representative bases for each object category, we use our trained network as a tool. The audio basis-object relation map after the first pooling layer of the MIML network produces matching scores across all basis vectors for all object labels. We perform a dimension-wise softmax over the basis dimension (M) to normalize object matching scores to probabilities along each basis dimension. By examining

the normalized map, we can discover links from bases to objects. We only collect the key bases that trigger the prediction of the correct objects (namely, the visually detected objects). Further, we only collect bases from an unlabeled video if multiple basis vectors strongly activate the correct object(s). See Fig. 5 for examples of typical basis-object relation maps. In short, at the end of this phase, we have a set of audio bases for each visual object, discovered purely from unlabeled video and mixed single-channel audio.

During testing, as shown in Fig. 4, given a novel test video, we obtain its audio magnitude spectrogram through STFT and detect objects using the same ImageNet-trained ResNet-152 network as before. Then, we retrieve the learnt audio basis vectors for each detected object, and use them to “guide” NMF-based audio source separation. Finally, we perform ISTFT on separated spectrogram to reconstruct the audio signals for each detected object.

3. Example Results

We use AudioSet [1] as the source of unlabeled training videos. We use 193k video clips of musical instruments, animals, and vehicles, which span a broad set of unique sound-making objects.

For “in the wild” unlabeled videos, the ground-truth of separated audio sources never exists. To facilitate quantitative evaluation, we construct a dataset of 23 AudioSet videos containing only a single sounding object selected from our val/test set, including 15 musical instruments, 4 animals, and 4 vehicles. We take pairwise video combinations from these videos, and 1) compound their audio tracks

Figure 4. Testing pipeline. Given a novel test video, we detect the objects present in the frames, and retrieve their learnt audio bases. The bases are collected to form a fixed basis dictionary with which to guide NMF of the test video’s audio channel. The basis vectors and the learned activation scores from NMF are finally used to separate the sound for each detected object, respectively.

by normalizing and mixing them and 2) compound their visual channels by max-pooling their respective object predictions. Each compound video is a test video; its reserved source audio tracks are the ground truth for evaluation of separation results. To evaluate source separation quality, we use the widely used BSS-EVAL toolbox [8] and report the Signal to Distortion Ratio (SDR). We perform four sets of experiments: pairwise compound two videos of musical instruments (Instrument Pair), two of animals (Animal Pair), two of vehicles (Vehicle Pair), and two cross-domain videos (Cross-Domain Pair).

Table 1 shows the results. Our method is compared against a series of baselines: 1) Upper-Bound: our performance upper-bound that uses AudioSet ground-truth labels to train the deep MIML network; 2) K-means Clustering: unsupervised NMF approach, where K-means clustering is used to group separated channels; 3) MFCC Unsupervised: a representative off-the-shelf unsupervised audio source separation method [7]; 4) Visual Exemplar: supervised NMF using bases from an exemplar video; 5) Unmatched Bases: supervised NMF using bases of the wrong class; 6) Gaussian Bases: supervised NMF using random bases. The results demonstrate the power of our learned bases. Compared with all baselines, our method achieves large gains, and it also has the capability to match the separated sources to meaningful acoustic objects in the video.

To facilitate comparison to prior audio-visual methods (none of which report results on AudioSet), we also perform the same experiment as in [5] on visually-assisted audio denoising on three benchmark videos used in previous studies: Violin Yanni, Wooden Horse, and Guitar Solo. Following the same setup as [5], the audio signals in all videos are corrupted with white noise with the signal to noise ratio set to 0 dB. To perform audio denoising, our method retrieves bases of detected object(s) and appends the same number of randomly initialized bases as the weight matrix to supervise NMF. The randomly initialized bases are intended to capture the noise signal. As in [5], we report Normalized SDR

	Instrument	Animal	Vehicle	Cross-Domain
Upper-Bound	2.05	0.35	0.60	2.79
K-means Clustering	-2.85	-3.76	-2.71	-3.32
MFCC Unsupervised [7]	0.47	-0.21	-0.05	1.49
Visual Exemplar	-2.41	-4.75	-2.21	-2.28
Unmatched Bases	-2.12	-2.46	-1.99	-1.93
Gaussian Bases	-8.74	-9.12	-7.39	-8.21
Ours	1.83	0.23	0.49	2.53

Table 1. We pairwise mix the sounds of two single source AudioSet videos and perform audio source separation. Mean Signal to Distortion Ratio (SDR in dB, higher is better) is reported to represent the overall separation performance.

	Wooden Horse	Violin Yanni	Guitar Solo	Average
Kidron et al. [3]	4.36	5.30	5.71	5.12
Lock et al. [4]	4.54	4.43	2.64	3.87
Pu et al. [5]	8.82	5.90	14.1	9.61
Ours	12.3	7.88	11.4	10.5

Table 2. Visually-assisted audio denoising results on three benchmark videos, in terms of NSDR (in dB, higher is better).

(NSDR), which measures the improvement of the SDR between the mixed noisy signal and the denoised sound.

Table 2 shows the results¹. Note that the method of Pu et al. [5] is tailored to separate noise from the foreground sound by exploiting the low-rank nature of background sounds. Still, our method outperforms [5] on 2 out of the 3 videos, and performs much better than the other two prior audio-visual methods [3, 4]. Pu et al. [5] also exploit motion in manually segmented regions. On Guitar Solo, the hand’s motion may strongly correlate with the sound, leading to their better performance.

Next we show some qualitative results to illustrate the effectiveness of MIML training. Fig. 5 shows example unlabeled videos and their discovered audio basis associations. For each example, we show sample video frames, ImageNet CNN visual object predictions, as well as the corresponding audio basis-object relation map predicted by our MIML

¹We take the numbers for existing methods from Pu et al. [5].

Figure 5. In each example, we show the video frames, visual predictions, and the corresponding basis-label relation maps predicted by our MIML network. Please see our supplementary video for more examples and the corresponding audio tracks.

network. We also report the AudioSet audio ground truth labels, but note that they are never seen by our method. The first example (Fig. 5-a) has both piano and violin in the visual frames, which are correctly detected by the CNN. The audio also contains the sounds of both instruments, and our method appropriately activates bases for both the violin and piano. Fig. 5-b shows a man playing the violin in the visual frames, but both piano and violin are strongly activated. Listening to the audio, we can hear that an out-of-view player is indeed playing the piano. This example accentuates the advantage of learning object sounds from thousands of unlabeled videos; our method has learned the correct audio bases for piano, and “hears” it even though it is off-camera in this test video. Fig. 5-c/d shows two examples with inaccurate visual predictions, and our model correctly activates the label of the object in the audio. Fig. 5-e/f show two more examples of an animal and a vehicle, and the results are similar. These examples suggest that our MIML network has successfully learned the prototypical spectral patterns of different sounds, and is capable of associating audio bases with object categories.

Please see our supplementary video for more qualitative results, where we use our system to detect and separate object sounds for novel “in the wild” videos. They lack ground truth, but results can be manually inspected for quality.

Acknowledgements: This research is supported in part by an IBM Faculty Award, IBM Open Collaboration Award, and NSF IIS -1514118.

References

- [1] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In ICASSP, 2017. 2
- [2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016. 2
- [3] E. Kidron, Y. Y. Schechner, and M. Elad. Pixels that sound. In CVPR, 2005. 3
- [4] E. F. Lock, K. A. Hoadley, J. S. Marron, and A. B. Nobel. Joint and individual variation explained (jive) for integrated analysis of multiple data types. The annals of applied statistics, 2013. 3
- [5] J. Pu, Y. Panagakis, S. Petridis, and M. Pantic. Audio-visual object localization and separation using low-rank and sparsity. In ICASSP, 2017. 3
- [6] T. Rahne, M. Böckmann, H. von Specht, and E. S. Sussman. Visual cues can modulate integration and segregation of objects in auditory scene analysis. Brain research, 2007. 1
- [7] M. Spiertz and V. Gnan. Source-filter based clustering for monaural blind source separation. In 12th International Conference on Digital Audio Effects, 2009. 3
- [8] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. IEEE transactions on audio, speech, and language processing, 2006. 3