

Learning to Separate Object Sounds by Watching Unlabeled Video

Ruohan Gao¹, Rogerio Feris², Kristen Grauman³

¹The University of Texas at Austin, ²IBM Research, ³Facebook AI Research
rhgao@cs.utexas.edu, rsferis@us.ibm.com, grauman@fb.com

Abstract. Perceiving a scene most fully requires all the senses. Yet modeling how objects look and sound is challenging: most natural scenes and events contain multiple objects, and the audio track mixes all the sound sources together. We propose to learn audio-visual object models from unlabeled video, then exploit the visual context to perform audio source separation in novel videos. Our approach relies on a deep multi-instance multi-label learning framework to disentangle the audio frequency bases that map to individual visual objects, even without observing/hearing those objects in isolation. We show how the recovered disentangled bases can be used to guide audio source separation to obtain better-separated, object-level sounds. Our work is the first to learn audio source separation from large-scale “in the wild” videos containing multiple audio sources per video. We obtain state-of-the-art results on visually-aided audio source separation and audio denoising. Our video results: http://vision.cs.utexas.edu/projects/separating_object_sounds/

Fig. 1. Goal: Learn from unlabeled video to separate object sounds

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

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

