Active Audio-Visual Separation of Dynamic Sound Sources

Sagnik Majumder[1] and Kristen Grauman[1,2]

[1] UT Austin, Austin, TX, USA
[2] Facebook AI Research, Austin, TX, USA
{sagnik, grauman}@cs.utexas.edu

Abstract. We explore active audio-visual separation for dynamic sound sources, where an embodied agent moves intelligently in a 3D environment to *continuously* isolate the *time-varying* audio stream being emitted by an object of interest. The agent hears a mixed stream of multiple audio sources (e.g., multiple people conversing and a band playing music at a noisy party). Given a limited time budget, it needs to extract the target sound accurately at *every step* using egocentric audio-visual observations. We propose a reinforcement learning agent equipped with a novel transformer memory that learns motion policies to control its camera and microphone to recover the dynamic target audio, using self-attention to make high-quality estimates for current timesteps and also simultaneously improve its past estimates. Using highly realistic acoustic SoundSpaces [4] simulations in real-world scanned Matterport3D [3] environments, we show that our model is able to learn efficient behavior to carry out continuous separation of a dynamic audio target. Full paper appearing in the main conference: `https://arxiv.org/abs/2202.00850`. Project: `https://vision.cs.utexas.edu/projects/active-av-dynamic-separation/`.

1 Introduction

Our daily lives are full of *dynamic audio-visual events*, and the activity and physical space around us affect how well we can perceive them. For example, an assistant responding to his boss's call in a noisy office could visually spot other noisy actors nearby and move to a quieter corner to hear better.

These examples show how *smart sensor motion* is necessary for accurate audio-visual understanding of dynamic events. For audio sensing in an environment full of distractor sounds, variations in acoustic attributes like volume, pitch, etc.— in concert with a listener's proximity and direction from it—determine the listener's ability to hear a target sound. However, just getting close or far isn't always enough: effective hearing may require the listener to move around by *visually sensing* competing sound sources and surrounding obstacles, and discovering locations favorable to listening (e.g., an intersection of two walls reflects audio and boosts hearing, a tall cabinet can dull acoustic interferences, etc.).

In this work, we investigate how to induce such intelligent behaviors in autonomous agents through audio-visual learning. Specifically, we propose the

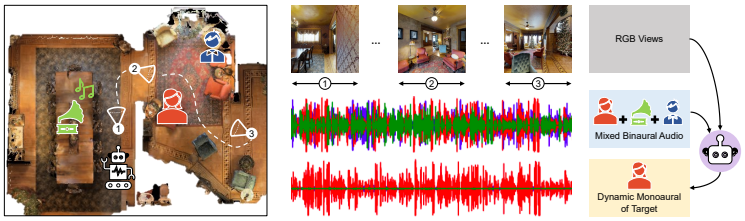

Fig. 1: Active audio-visual separation of dynamic sources. Given multiple *dynamic* (time-varying, non-periodic) audio sources S, all mixed together, the proposed agent separates the audio signal by actively moving in the 3D environment on the basis of its egocentric audio-visual input, so that it is able to accurately retrieve the target signal at every step of its motion.

new task of *active audio-visual separation of dynamic sources*: given a stream of egocentric audio-visual observations, an agent must determine how to move in an environment with multiple sounding objects in order to *continuously* retrieve the *dynamic* (temporally changing) sounds being emitted by some object of interest, in a limited time budget. See Figure 1. This task is relevant for augmented reality (AR) and mobile robotics applications, where a hearing device or service robot needs to aid a user in understanding target sound sources in a busy environment.

Whereas traditional audio-visual separation models extract sounds passively from pre-recorded videos [5,1,10,7,8,12,11], our task requires active placement of the agent's camera and microphones over time. Whereas embodied audio-visual navigation [4,6] entails moving towards a sound source, our task requires recovering the sounds of a target object. The recent Move2Hear model performs active source separation [9] but is limited to *static* (*i.e.*, periodic or constant) sound sources, such as a ringing phone or fire alarm—where recovering one timestep of the sound is sufficient. In contrast, our task involves *dynamic* audio sources and calls for extracting the target audio at *every step* of the agent's movement. Variations in the observed audio arise not only from the room acoustics, but also from the temporally-changing and fleeting nature of the target and distractor sounds. This means the agent must recover a new audio segment at every step of its motion, which it hears only once. The proposed task is thus both more realistic and more difficult than existing active audio-visual separation settings.

To address active dynamic audio-visual source separation, we introduce a reinforcement learning (RL) framework that trains an agent how to move to continuously listen to the dynamic target sound. Our agent receives a stream of egocentric audio-visual observations in the form of RGB images and mixed binaural audio, along with the target category of interest (human voice, musical instrument, etc.) and decides its next action (translation or rotation of its camera and microphones) at every time step. We test our framework on audio-visual simulations from SoundSpaces [4] together with Matterpor3D [3] environment scans and non-periodic sounds from multiple human speakers, music and common background sources. Our model outpeforms the state-of-the-art Move2Hear [9] model, generalizing better to unheard sounds in unseen environments.

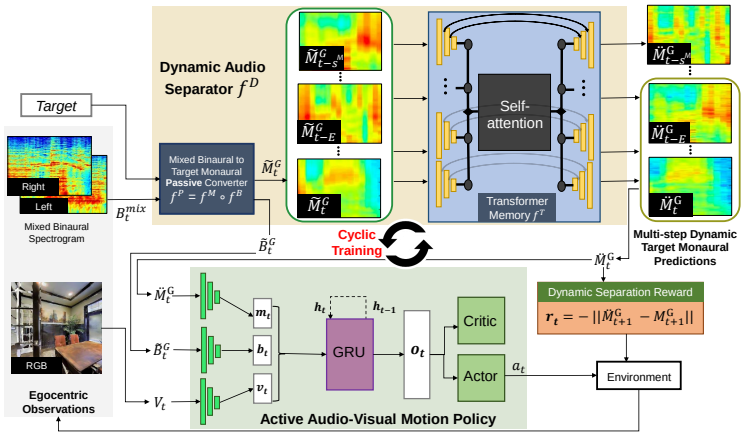

Fig. 2: Our model addresses active audio-visual separation of dynamic sources by leveraging a synergy between a *dynamic audio separator* and an *active audio-visual motion policy*. At timestep t, the dynamic separator f^D uses self-attention to continuously isolate the signal M^G_t for a target sound category G from its received mixed binaural B^{mix}_t on the basis of its past and current initial estimates $\{\tilde{M}^G_{t-s^{\mathcal{M}}}, \ldots, \tilde{M}^G_t\}$, while also using its current initial separation $\tilde{M}^G_t$ to refine past final separations $\{\ddot{M}^G_{t-E}, \ldots, \ddot{M}^G_{t-1}\}$. Here, $s^{\mathcal{M}}$ is size of external memory for storing past estimates. The motion policy uses the separator's outputs, $\ddot{M}^G_t$ and $\tilde{B}^G_t$ and egocentric RGB images V_t to guide the agent to areas suitable for separating future dynamic targets.

2 Approach

We train a motion policy to make sequential movement decisions on the basis of egocentric audio-visual observations, guided by dynamic audio separation quality. Our model has two main components (see Fig. 2): 1) an audio separator network and 2) an active audio-visual (AV) motion policy.

The separator network serves three functions at every step: 1) it passively separates the current target audio segment from its heard mixture, 2) it improves its past separations by exploiting their correlations in acoustic attributes like volume, pitch, quality, content semantics, etc. with the current separation, and 3) it uses the current separation to guide the motion policy towards locations favorable for accurate separation of future audio segments. The motion policy is trained to repeatedly maximize dynamic separation quality. It moves the agent in the 3D scene to get the best possible estimate of the *complete* target signal.

These two components share a symbiotic relationship, each feeding off of useful learning cues from the other while training. This allows the agent to learn the complex links between separation quality and the dynamic acoustic attributes of the target source, its location relative to the agent, the scene's material layout (walls, floors, furniture, etc.), and the inferred spatial arrangement of distractor sources in the 3D environment. Refer to the full paper for more details on our approach.

Table 1: Active audio-visual dynamic source separation.

	Heard		*Unheard*	
Model	SI-SDR ↑	STFT ↓	SI-SDR ↑	STFT ↓
Stand In-Place	2.49	0.328	2.03	0.343
Rotate In-Place	2.50	0.327	2.04	0.343
DoA	2.78	0.313	1.88	0.342
Random	2.81	0.314	1.95	0.343
Proximity Prior	2.92	0.309	2.05	0.338
Novelty [2]	1.68	0.358	1.44	0.366
Move2Hear [9]	2.31	0.331	2.06	0.339
Ours	**3.93**	**0.273**	**2.57**	**0.318**

3 Experiments

We evalaute our model using standard metrics, like the **STFT disance** between the predicted and the ground-truth spectrograms for the target audio, and **SI-SDR**, a scale-invariant measure of the extent of distortion in the separated audio. We compare against different baselines and existing methods, including passive baselines that holds the starting pose of the agent (**Stand In-Place**) or keeps rotating at the start location (**Rotate In-Place**). Other baselines are agents that samples the direct sound from a step away from the target source (**DoA**), or randomly scouts the area around the target (**Random** and **Proximity Prior**). We also compare against a model that maximizes area coverage within the time budget and samples diverse audio cues (**Novelty [2]**) and a state-of-the-art active separation model for static sources (**Move2Hear [9]**).

Table 1 shows the separation quality of all models. The passive baselines Stand and Rotate In-Place fare worse than the ones that move and sample more diverse audio cues, like Random and Proximity Prior. However, despite being able to move, Novelty [2] performs poorly; in its effort to maximize coverage of the environment, it wanders too far from the target and fails to hear it in certain phases of its motion. DoA improves over the stationary baselines, since standing a step away from the target source allows it to sample a cleaner audio cue.

Our model outperforms all baselines and Move2Hear by a statistically significant margin ($p \leq 0.05$). Its performance demonstrates the impact of our active policy and long-term memory. Simply staying at or close to the source to be able to continuously hear the target (as done by Stand or Rotate In-Place, DoA, and Proximity Prior) is not enough. Our improvement over Move2Hear [9] further emphasizes the advantage of our transformer memory f^T in dealing with dynamic audio, both in terms of boosting separation when the agent is able to sample a cleaner signal, and providing robustness to the separator when the agent is passing through zones that are relatively less suitable for separation. Refer to the full paper for more experiments and model analysis.

Acknowledgements: Thank you to Ziad Al-Halah for very valuable discussions. Thanks to Tushar Nagarajan, Kumar Ashutosh, and David Harwarth for feedback on paper drafts. UT Austin is supported in part by DARPA L2M, NSF CCRI, and the IFML NSF AI Institute. K.G. is paid as a research scientist by Meta.

References

1. Afouras, T., Chung, J.S., Zisserman, A.: The conversation: Deep audio-visual speech enhancement. arXiv preprint arXiv:1804.04121 (2018)
2. Bellemare, M.G., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., Munos, R.: Unifying count-based exploration and intrinsic motivation. arXiv preprint arXiv:1606.01868 (2016)
3. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. International Conference on 3D Vision (3DV) (2017), matterPort3D dataset license available at: `http://kaldir.vc.in.tum.de/matterport/MP_TOS.pdf`
4. Chen, C., Jain, U., Schissler, C., Gari, S.V.A., Al-Halah, Z., Ithapu, V.K., Robinson, P., Grauman, K.: SoundSpaces: Audio-visual navigation in 3D environments. In: Proceedings of the European Conference on Computer Vision (ECCV). Springer (2020)
5. Gabbay, A., Shamir, A., Peleg, S.: Visual speech enhancement. arXiv preprint arXiv:1711.08789 (2017)
6. Gan, C., Zhang, Y., Wu, J., Gong, B., Tenenbaum, J.B.: Look, listen, and act: Towards audio-visual embodied navigation. In: ICRA (2020)
7. Gao, R., Grauman, K.: Co-separating sounds of visual objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3879–3888 (2019)
8. Gao, R., Grauman, K.: Visualvoice: Audio-visual speech separation with cross-modal consistency. arXiv preprint arXiv:2101.03149 (2021)
9. Majumder, S., Al-Halah, Z., Grauman, K.: Move2Hear: Active audio-visual source separation. In: ICCV (2021)
10. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 631–648 (2018)
11. Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., Zhong, J.: Attention is all you need in speech separation. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 21–25. IEEE (2021)
12. Zadeh, A., Ma, T., Poria, S., Morency, L.P.: Wildmix dataset and spectro-temporal transformer model for monoaural audio source separation. arXiv preprint arXiv:1911.09783 (2019)