AudioScopeV2: Audio-Visual Attention Architectures for Calibrated Open-Domain On-Screen Sound Separation

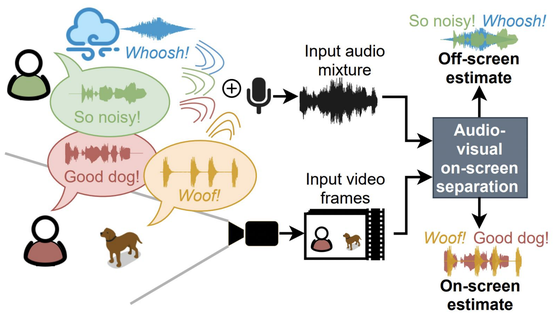
Efthymios Tzinis[1,2], Scott Wisdom[1], Tal Remez[1], John R. Hershey[1]*

[1] Google Research [2]

* Work done during an internship at Google.

Task

- Separate all sounds that originate from on-screen objects, regardless of their class:.



Prior Work

- Most prior audio-visual separation work requires supervised training data, and can only handle specific sound classes (e.g. speech, music).
- AudioScopeV1 [1] proposed a different approach that can learn to separate open-domain on-screen sounds in an unsupervised manner.
  - **(1)** Audio-only separation into sources, trained on raw unsupervised audio with mixture invariant training (MixIT) [2].
  - **(2)** Audio-visual classifier that predicts the probability $\hat{y}_m$ that each source $\hat{s}_m$ is on-screen, trained using MixIT assignments.
  - **(3)** On-screen estimate by using probabilities as mixing weights:

$$\hat{y}_m = \sigma\left(\hat{\ell}_m\right) \in [0,1], \;\; \hat{x}^{\text{on}} = \sum_{m=1}^{M} \hat{y}_m \hat{s}_m$$

Contributions

- Propose more sophisticated cross-modal and self-attention audio-visual architectures, compared to AudioScopeV1 [1].
- Create a new dataset using YFCC100M that is more challenging and unconstrained compared to original AudioScopeV1 [1] dataset.
- Propose a new calibration procedure to precisely tune on-screen reconstruction versus off-screen suppression, that simplifies comparisons across models with different operating points.

References

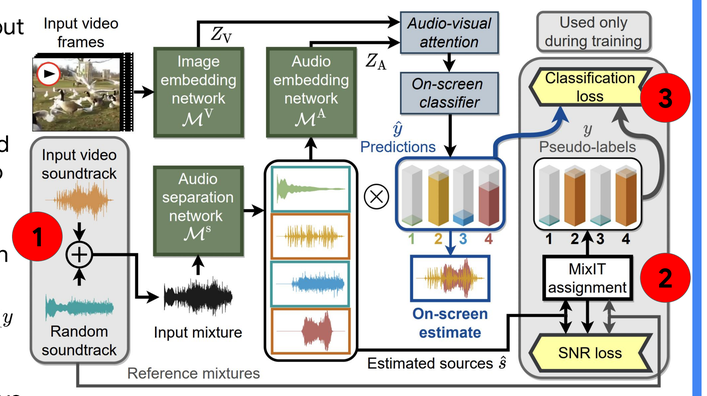
[1] E. Tzinis et al., Into the Wild with AudioScope: Unsupervised Audio-Visual Separation of On-Screen Sounds, ICLR 2021.

[2] S. Wisdom et al., Unsupervised Sound Separation Using Mixture Invariant Training, NeurIPS 2020.

[3] B. Thomee et al., YFCC100M: The New Data in Multimedia Research, Comm. of the ACM 2015.

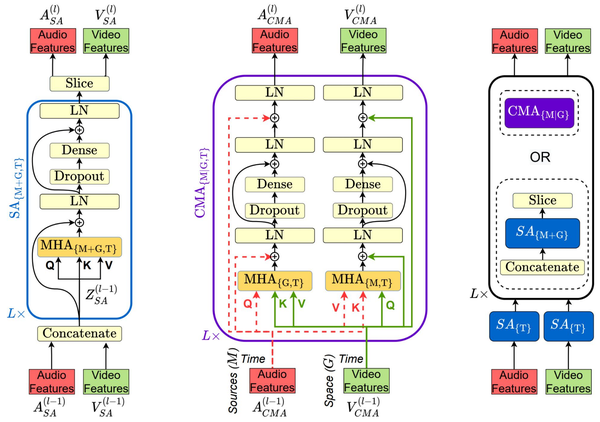
AudioScopeV1 and V2 Model Training

1. Mix a random soundtrack into an input video's soundtrack to make mixture of mixtures (MoM).
2. Separate the MoM and **assign** each source to one of the 2 input soundtracks; use SNR loss to train separation model (MixIT loss).
3. Use the **assignments** y as pseudo-labels to train classifier with multiple-instance active combinations (AC) loss: $\mathcal{L}_{\text{AC}}(y,\hat{y}) = \min_{\ell \in \wp_{\geq 1}(\mathbb{B}^M)} \sum_m \mathcal{L}_{\text{CE}}(\ell_m, \hat{y}_m)$, $\wp_{\geq 1}(\mathbb{B}^M)$ is all subsets of y with at least one positive.



Proposed Audio-Visual Attention Architectures

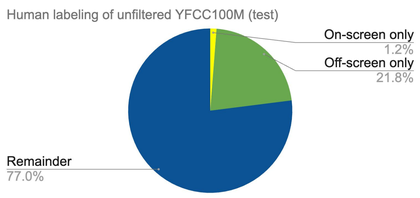
1. Self-attention (SA) attends over audio sources (M), spatial locations (G), and time (T).
2. Cross-modal attention allows shape (G, T) video tensors to attend to shape (M, T) audio tensors, and vice versa.
3. Separable versions split attention over audio sources M/spatial locations G and time T.



(a) Self-Attention (SA). (b) Cross-modal attention (CMA). (c) Separable.

New Unfiltered Dataset

- All data sourced from Creative Commons-licensed YFCC100M [3] videos.
- First AudioScope model [1] trained with YFCC100M data filtered by an unsupervised audio-visual coincidence model.
  - We found this filtering method introduces bias (see results to the right).
- We construct a new unfiltered version of YFCC100M with new human annotations (3 raters per clip) for on-screen and off-screen sounds.
- 1600 hours (4.85M 5s clips) training data
- Total / on-screen-only / off-screen-only:
  - Train: 20000 / 480 / 4664
  - Validation: 6500 / 109 / 1421
  - Test: 3500 / 43 / 762



Proposed Calibration Procedure

- Signal-to-noise ratio (SNR) measures reconstruction of on-screen sounds.

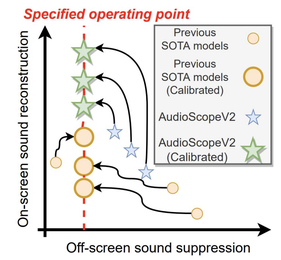
$$\text{SNR}(x^{\text{on}}, \hat{x}^{\text{on}}) = 20\log_{10}\frac{\|x^{\text{on}}\|}{\|x^{\text{on}} - \hat{x}^{\text{on}}\|}$$

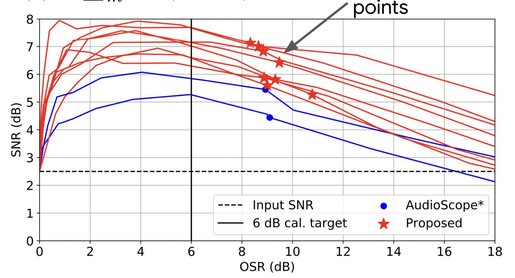
- Off-screen suppression ratio (OSR) measures rejection of off-screen sounds.

$$\text{OSR}(x, \hat{x}^{\text{on}}) = 20\log_{10}\frac{\|x\|}{\|\hat{x}^{\text{on}}\|}$$

- Inherent tradeoff between SNR and OSR, and models achieve different random operating points after training.
- To compare fairly, measure SNR at a calibrated target OSR.
- Can calibrate to target OSR by tuning a scalar bias on logits:

$$\tilde{x}^{\text{on}}(\theta) = \textstyle\sum_m \hat{s}_m \sigma(\hat{\ell}_m + \theta)$$

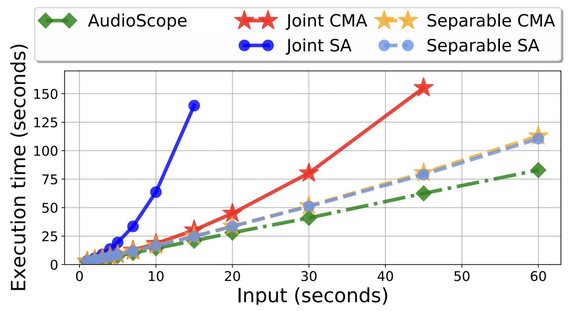




Results

- Measure on-screen SNR calibrated to 6 dB OSR for all models.
- Source power-weighted AUC-ROC to measure classifier performance.
- Evaluate on the original AudioScopeV1 filtered dataset, and the new proposed unfiltered dataset.

Separation model audio-only pre-training		Trained on filtered	Filtered [1]				Unfiltered (new proposed)			
			1 FPS		16 FPS		1 FPS		16 FPS	
AV alignment	Complexity	PT Filt.	**SNR**	**AUC**	**SNR**	**AUC**	**SNR**	**AUC**	**SNR**	**AUC**
No processing ($\hat{x}^{\text{on}} = x$ with 0dB OSR)			4.4	–	4.4	–	2.5	–	2.5	–
No processing ($\hat{x}^{\text{on}} = x/2$ with 6dB OSR)			4.7	–	4.7	–	4.1	–	4.1	–
AudioScope [45]	$\mathcal{O}(TMG)$	✓	6.0	0.79	–	–	2.7	0.69	–	–
AudioScope*	$\mathcal{O}(TMG)$	✓	8.2	0.80	5.9	0.77	5.8	0.78	5.2	0.71
SA Joint	$\mathcal{O}(T^2[M+G]^2)$	✓	**10.0**	0.84	9.9	**0.86**	7.2	0.82	**7.7**	0.83
SA Sep.	$\mathcal{O}(T^2+[M+G]^2)$	✓	9.6	0.84	8.2	0.83	6.6	0.78	6.6	0.80
CMA Joint	$\mathcal{O}(T^2MG)$	✓	**10.0**	**0.88**	**10.0**	0.85	**7.3**	**0.83**	**7.7**	**0.84**
CMA Sep.	$\mathcal{O}(T^2+MG)$	✓	9.5	0.83	9.3	0.82	6.4	0.78	7.1	0.80



- Computational efficiency of audio-visual attention architectures.
- Separable versions are much more efficient for longer inputs.

- Online demos of AudioScopeV2 separation and dataset recipe: google-research.github.io/sound-separation/papers/audioscope-v2