CLIPSep: Learning Text-queried Sound Separation with Noisy Unlabeled Videos

Hao-Wen Dong[1,2]* Naoya Takahashi[1]† Yuki Mitsufuji[1] Julian McAuley[2] Taylor Berg-Kirkpatrick[2]

[1]Sony Group Corporation [2]University of California San Diego

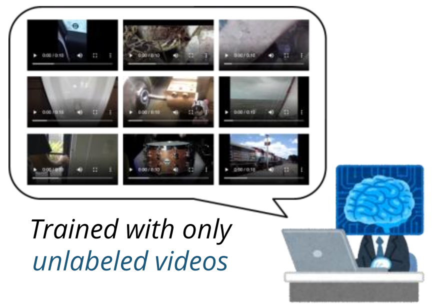
* Work done during an internship at Sony † Corresponding author

Overview

We explore training a sound separation system under a **self-supervised learning** setting. We aim to achieve text-queried universal sound separation by *using only unlabeled data*.

To learn the desired audio-textual correspondence from unlabeled videos, we leverage the visual modality as a bridge using the contrastive image-language pretraining (CLIP) model.

Training



Scalable to larger dataset

Inference

Natural text query-based interface

Contributions

- We propose the *first text-queried universal sound separation model that can be trained on unlabeled videos*.
- We propose a new approach called *noise invariant training* for *training a query-based sound separation model on noisy data*.

Data

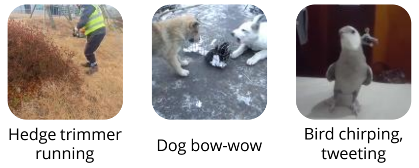
MUSIC

(Zhao et al., 2018)

Music instrument playing videos

VGGSound

(Chen et al., 2020)



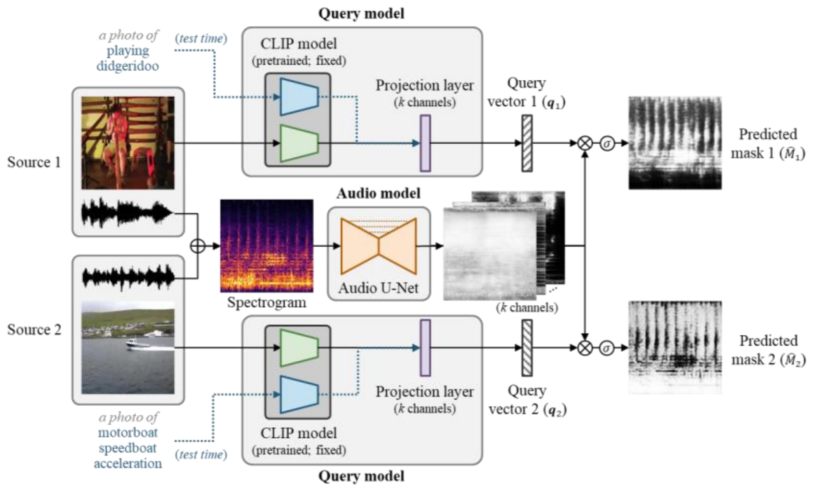
Noisy videos with diverse sounds

CLIPSep

Training

We mix audio from two videos and train the model to separate each audio source given the corresponding video frame as the query:

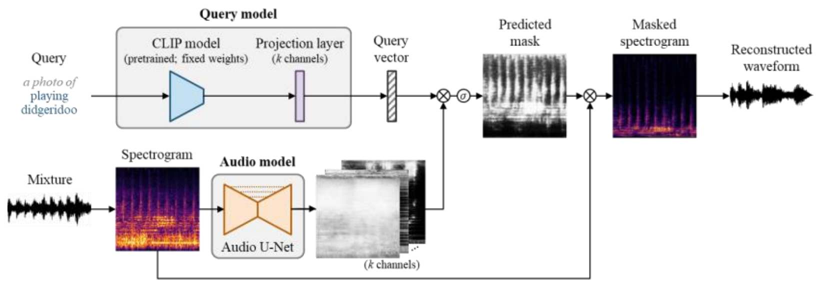
1. An Audio U-Net predicts k intermediate masks $\tilde{M}_1, \dots, \tilde{M}_k$ from the mixture spectrogram.
2. A pretrained CLIP model encodes the input query into a query vector $\boldsymbol{q}_i$.
3. Construct the predicted masks with $\hat{M}_i = \sum_{j=1}^{k} \sigma\left(w_{ij} q_{ij} \tilde{M}_j + b_i\right)$.


Query model
a photo of playing didgeridoo
(test time)
CLIP model (pretrained; fixed)
Projection layer (k channels)
Query vector 1 ($\boldsymbol{q}_1$)
Predicted mask 1 ($\hat{M}_1$)
Source 1
Audio model
Spectrogram
Audio U-Net
(k channels)
Source 2
Projection layer (k channels)
Query vector 2 ($\boldsymbol{q}_2$)
Predicted mask 2 ($\hat{M}_2$)
a photo of motorboat speedboat acceleration
(test time)
CLIP model (pretrained; fixed)
Query model

Inference

At test time, we instead use a text query in the form of "*a photo of [user input query]*".

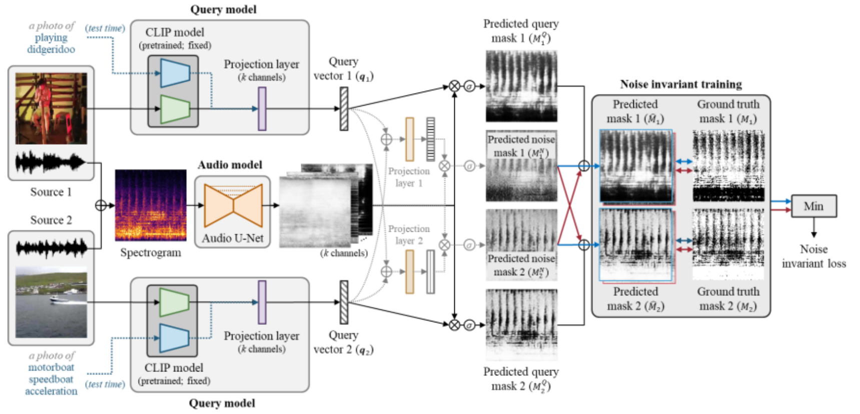
With the pretrained CLIP model, the query vectors we obtain for the image and text queries are expected to be close.



Noise Invariant Training (NIT)

Videos in the wild may contain off-screen sounds and background noise. We introduce two additional noise masks to capture query-irrelevant sounds.

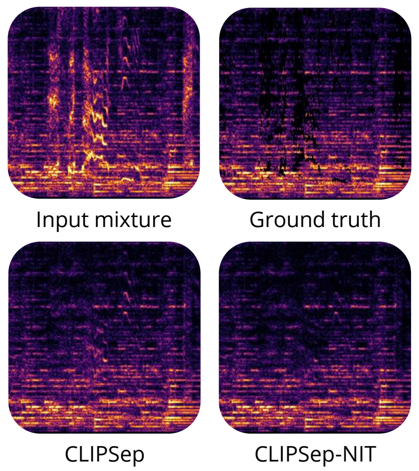
During training, we select the arrangement that has the lowest loss value when combining the noise and query masks. At test time, we discard the noise heads.


Query model
a photo of playing didgeridoo
(test time)
CLIP model (pretrained; fixed)
Projection layer (k channels)
Query vector 1 ($\boldsymbol{q}_1$)
Source 1
Source 2
Audio model
Spectrogram
Audio U-Net
(k channels)
Projection layer 1
Projection layer 2
Predicted query mask 1 (M_1^Q)
Predicted noise mask 1 (M_1^N)
Predicted noise mask 2 (M_2^N)
Predicted query mask 2 (M_2^Q)
Noise invariant training
Predicted mask 1 ($\hat{M}_1$)
Ground truth mask 1 (M_1)
Predicted mask 2 ($\hat{M}_2$)
Ground truth mask 2 (M_2)
Min
Noise invariant loss
Projection layer (k channels)
Query vector 2 ($\boldsymbol{q}_2$)
CLIP model (pretrained; fixed)
a photo of motorboat speedboat acceleration
(test time)
Query model

Results

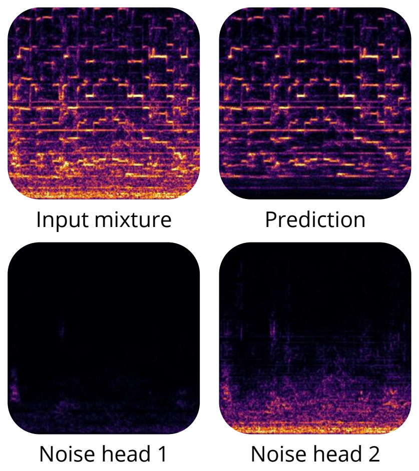
Audio samples available at sony.github.io/CLIPSep/

Sound Separation



Query: "playing harpsichord"
Interference: "people coughing"

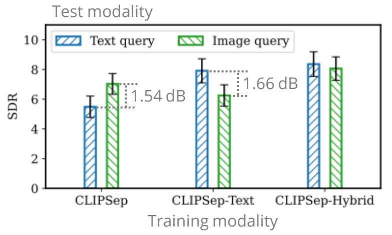
Noise Removal



Query: "playing bagpipes"
Interference: none

Results on MUSIC (clean data)

Model	Unlabeled data	Post-proc. free	Query type		SDR [dB]	
			Training	Test	Mean	Median
Mixture	-	-	-	-	0.00 ± 0.89	0.00
Text-queried models						
CLIPSep	✓	✓	Image	Text	**5.49 ± 0.72**	**4.97**
CLIPSep-Text		✓	Text	Text	7.91 ± 0.81	7.46
CLIPSep-Hybrid		✓	Text + Image	Text	**8.36 ± 0.83**	**8.72**
Image-queried models						
SOP (Zhao et al., 2018)	✓	✓	Image	Image	6.59 ± 0.85	**6.22**
CLIPSep	✓	✓	Image	Image	**7.03 ± 0.70**	5.85
CLIPSep-Text		✓	Text	Image	6.25 ± 0.72	6.19
CLIPSep-Hybrid		✓	Text + Image	Image	**8.06 ± 0.79**	**8.01**
Nonqueried models						
LabelSep		✓	Label	Label	8.18 ± 0.80	**7.82**
PIT (Yu et al., 2017)	✓		×	×	**8.68 ± 0.76**	7.67



Results on VGGSound (noisy data)

Model	Unlabeled data	Post-proc. free	MUSIC+		VGGSound-Clean+	
			Mean SDR	Median SDR	Mean SDR	Median SDR
Mixture	-	-	4.49 ± 1.41	2.04	-0.77 ± 1.31	-0.84
Text-queried models						
CLIPSep	✓	✓	9.71 ± 1.21	8.73	2.76 ± 1.00	**3.95**
CLIPSep-NIT	✓	✓	**10.27 ± 1.04**	**10.02**	**3.05 ± 0.73**	3.26
BERTSep		✓	4.67 ± 0.44	4.41	5.09 ± 0.80	5.49
CLIPSep-Text		✓	10.73 ± 0.99	9.93	5.49 ± 0.82	5.06
Image-queried models						
SOP (Zhao et al., 2018)	✓	✓	11.44 ± 1.18	11.18	2.99 ± 0.84	3.89
CLIPSep	✓	✓	**12.20 ± 1.17**	**12.42**	**5.46 ± 0.79**	**5.35**
CLIPSep-NIT	✓	✓	11.28 ± 1.08	10.83	4.84 ± 0.66	3.57
CLIPSep-Text		✓	9.89 ± 1.04	8.09	2.45 ± 0.70	1.74
Nonqueried models						
PIT (Yu et al., 2017)	✓		**12.24 ± 1.20**	**12.53**	**5.73 ± 0.79**	**4.97**
LabelSep		✓	-	-	5.55 ± 0.81	5.29

CLIPSep successfully learned text-queried sound separation on noisy data.

Noise invariant training improves the mean SDRs.

Paper: arxiv.org/abs/2212.07065
Demo: sony.github.io/CLIPSep/
Code: github.com/sony/CLIPSep

Paper Demo Code