SANE2023

Learning Text-queried Sound Separation and Synthesis using Unlabeled Videos and Pretrained Language-Vision Models

UC San Diego

Hao-Wen Dong (University of California San Diego)

Introduction

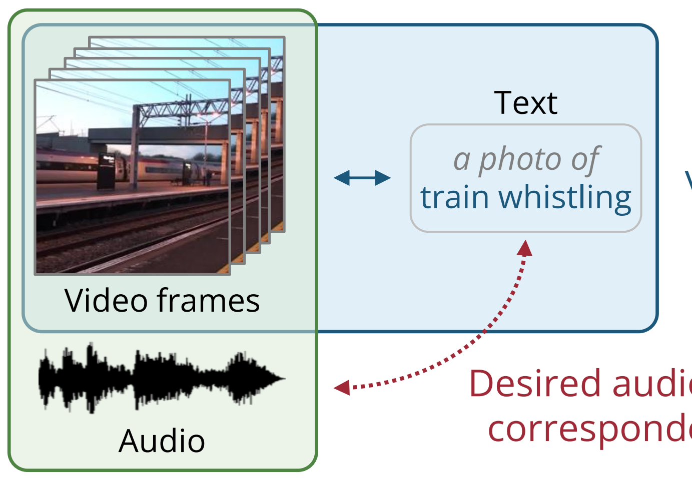
Lately, contrastive language-image pretraining (CLIP) has revolutionized multimodal learning and showed remarkable generalizability in many downstream tasks. While similar attempts have been made to build a counterpart model for language and audio, it remains unclear whether we can scale up text-audio datasets to a size comparable to large-scale text-image datasets.

We explore text-audio data free training **for text-queried sound separation and text-to-audio synthesis**. The proposed models learn the desired text-audio correspondence by combining

- naturally-occurring audio-visual correspondence in videos
- multimodal representation learned by contrastive language-image pretraining (CLIP)

This study offers a new direction of approaching bimodal learning for text and audio through leveraging the visual modality as a bridge.

Audio-visual correspondence in videos



Pretrained vision-language models

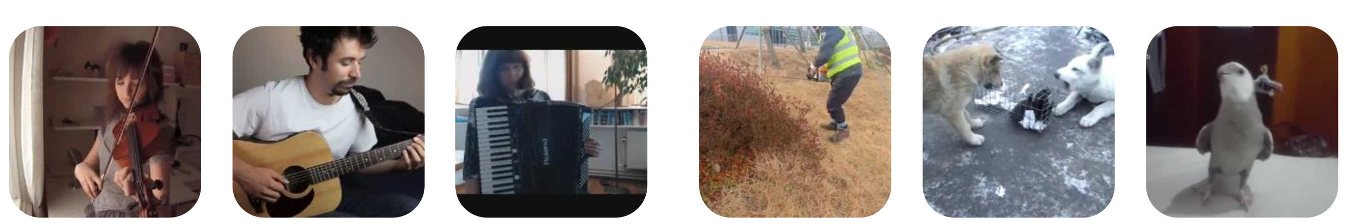
Data

MUSIC

(Zhao et al., 2018)

VGGSound

(Chen et al., 2020)



Violin | Acoustic guitar | Accordion | Hedge trimmer running | Dog bow-wow | Bird chirping, tweeting

Music instrument playing videos

Noisy videos with diverse sounds

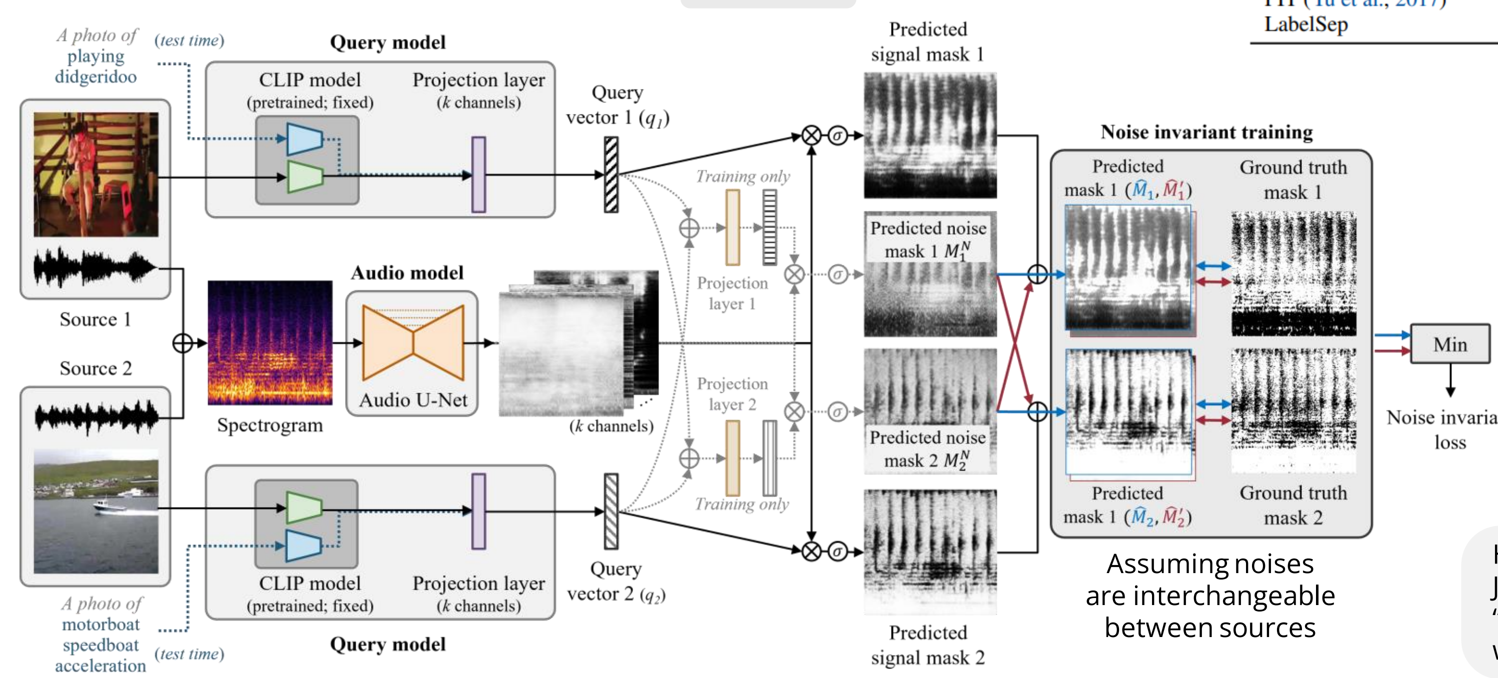
Work done during internships at Sony and Dolby. Joint work with Taylor Berg-Kirkpatrick, Julian McAuley (UC San Diego), Naoya Takahashi, Yuki Mitsufuji (Sony), Xiaoyu Liu, Jordi Pons, Gautam Bhattacharya, Santiago Pascual and Joan Serrà (Dolby).

CLIPSep: Text-queried Sound Separation

Training: We mix the audio track from two videos and train the model to separate each audio source given the corresponding video frame (encoded by the pretrained CLIP-image encoder) as the query.

Inference: We take text queries as inputs by using the pretrained CLIP-text encoder to encode the text.

Model



Quantitative Results

Model	Unlabeled data	MUSIC+ Mean SDR	MUSIC+ Median SDR	VGGSound-Clean+ Mean SDR	VGGSound-Clean+ 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
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

Example Result

Query: "playing bagpipes"

Mixture | Prediction | Noise head 1 | Noise head 2

Paper Demo

ICLR 2023

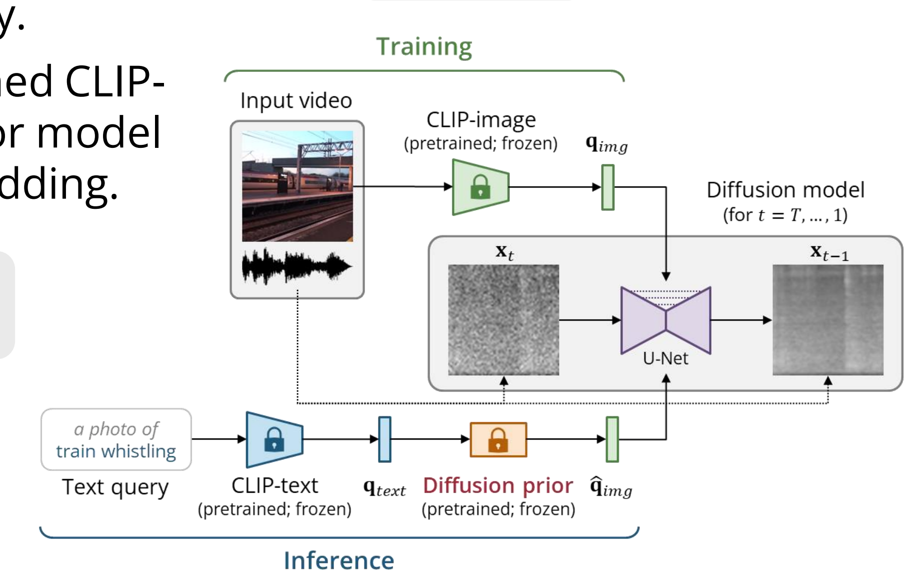
Hao-Wen Dong, Naoya Takahashi, Yuki Mitsufuji, Julian McAuley, and Taylor Berg-Kirkpatrick, "CLIPSep: Learning Text-queried Sound Separation with Noisy Unlabeled Videos," *ICLR*, 2023.

CLIPSonic: Text-to-audio Synthesis

Training: Similarly, we train a diffusion model that generates a mel spectrogram given the corresponding video frame as the query.

Inference: We take text queries as inputs by using the pretrained CLIP-text encoder to encode the text and a pretrained diffusion prior model to generate a CLIP-image embedding from the CLIP-text embedding.

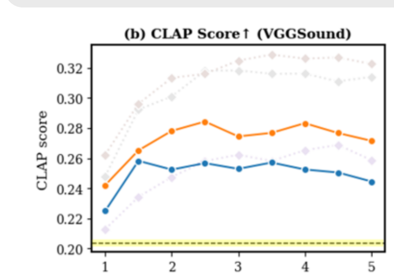
Model



Text-to-audio Synthesis Results

Model	VGGSound Fidelity	VGGSound Relevance	MUSIC Fidelity	MUSIC Relevance
CLIPSonic-ZS	2.55 ± 0.22	2.01 ± 0.27	2.98 ± 0.23	3.87 ± 0.24
CLIPSonic-PD	**3.04 ± 0.20**	2.86 ± 0.25	**3.67 ± 0.18**	3.91 ± 0.24
CLIPSonic-SD	2.96 ± 0.21	**3.49 ± 0.28**	3.36 ± 0.20	**4.07 ± 0.22**
Ground truth	3.78 ± 0.19	3.54 ± 0.29	3.90 ± 0.17	4.34 ± 0.18

Effects of classifier-free guidance



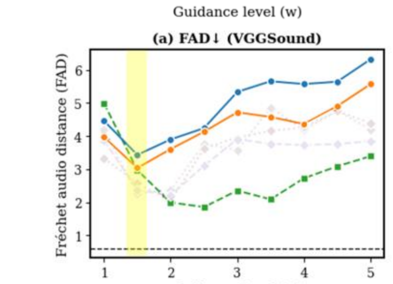


Image-to-audio Synthesis Results

Model	Fidelity	Relevance
CLIPSonic-IQ (image-queried)	**3.29 ± 0.16**	3.80 ± 0.19
SpecVQGAN [20]	2.15 ± 0.17	2.54 ± 0.23
im2wav [21]	2.19 ± 0.15	**3.90 ± 0.22**

Paper Demo

WASPAA 2023

Hao-Wen Dong, Xiaoyu Liu, Jordi Pons, Gautam Bhattacharya, Santiago Pascual, Joan Serrà, Taylor Berg-Kirkpatrick, and Julian McAuley, "CLIPSonic: Text-to-Audio Synthesis with Unlabeled Videos and Pretrained Language-Vision Models," *WASPAA*, 2023.