Language Queried Audio Source Separation

Wenwu Wang

Centre for Vision, Speech and Signal Processing (CVSSP)
& Surrey Institute for People Centred Artificial Intelligence

University of Surrey

United Kingdom

Email: w.wang@surrey.ac.uk
Web: https://personalpages.surrey.ac.uk/w.wang/

06/12/2024

Outline

- **Introduction**
- **A bit history about my work on source separation**
  - Speech source separation
  - Audio-visual speech source separation
  - Singing voice separation
  - Universal audio source separation
- **Language queried audio source separation**
  - AudioSep
  - FlowSep
  - DCASE challenge
- **Conclusions and future works**

Many thanks to…

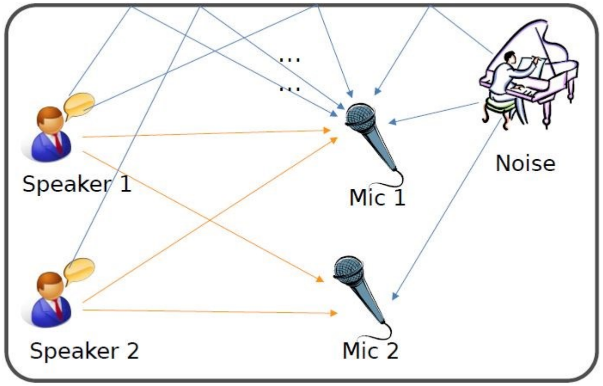
- **Xubo Liu**
- **Yi Yuan**
- **Junqi Zhao**
- **Qingju Liu**,
- Qiuqiang Kong
- Jian Guan
- Feiyang Xiao
- Yong Xu
- Yin Cao
- Qiaoxi Zhu
- Jonathon Chambers
- Philip Jackson
- Mark Barnard,
- Atiyeh Alinaghi,
- Swati Chandna,
- Jian Guan,
- Jing Dong,
- Alfredo Zermini,
- Yang Yu,
- Tariq Jan,
- Tao Xu,
- Josef Kittler,
- Mark Plumbley,
- Saeid Sanei,
- DeLiang Wang
- …

Funding from:

UK Engineering and Physical Sciences Research Council (EPSRC) & University Defence Research Collaboration (UDRC).

UKRI Engineering and Physical Sciences Research Council

Cocktail party problem



Cocktail-party problem (Cherry 1953) or *ball-room problem* (Helmholtz, 1863)

"No machine has yet been constructed to do just that [solving the cocktail party problem]." (Cherry, 1957)

Cocktail party problem may involve a few tasks:

How many speakers?
(Source counting)

Where are they?
(Localization and tracking)

Who speaks and when?
(Diarization)

What are the individual speech sources?
(**Speech separation**)

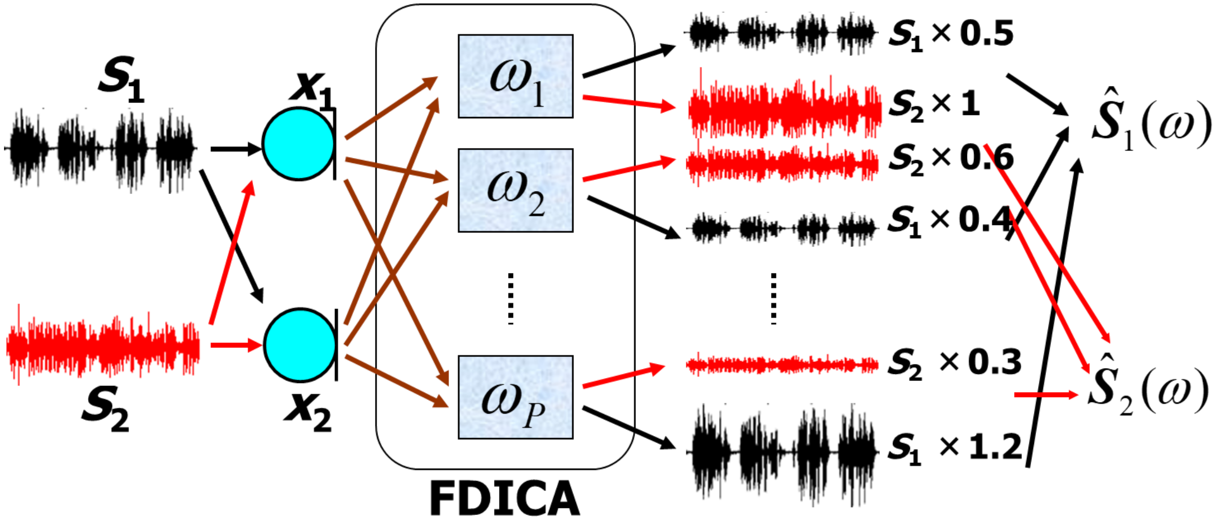
What has been said?
(Automatic speech recognition)

What kind environment?
(Acoustic scene recognition, event detection, room acoustics)

Speech source separation

- Potential techniques for the speech separation problem
  - Beamforming
  - Adaptive filtering
  - Blind source separation and independent component analysis
  - Sparse representation and matrix factorization
  - Matrix/tensor factorization techniques
  - Computational auditory scene analysis (e.g. time-frequency masking)
  - Learning based techniques
  - Exploiting multimodality (e.g. audio-visual coherence)
  - …

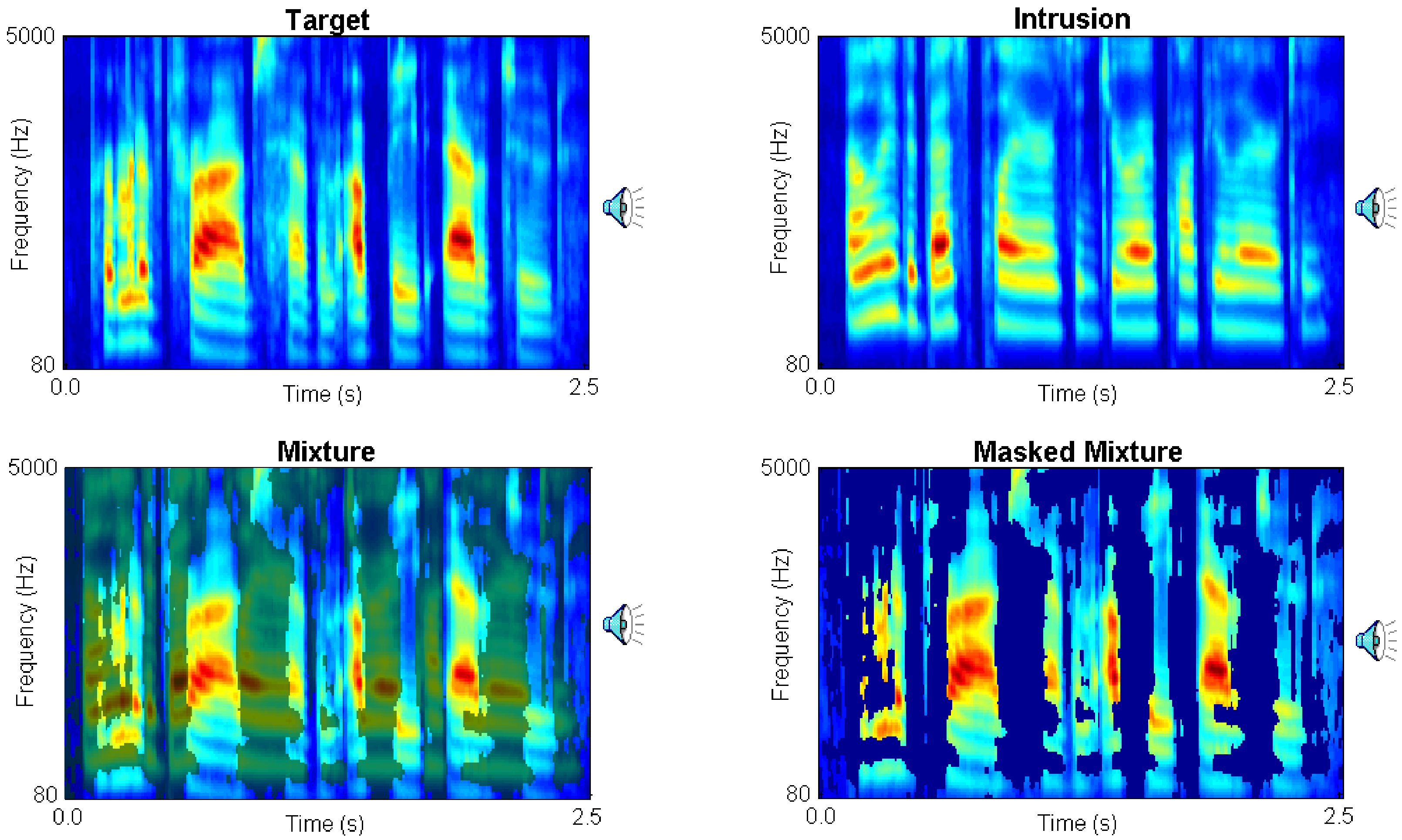
FDICA: Sound Demonstration



	Sources	Mixtures	Parra&Spence	Our approach
Two speaking sentences artificially mixed together	🔊	🔊	🔊	🔊
	🔊	🔊	🔊	🔊
A man speaking with TV on		🔊	🔊	🔊
		🔊	🔊	🔊

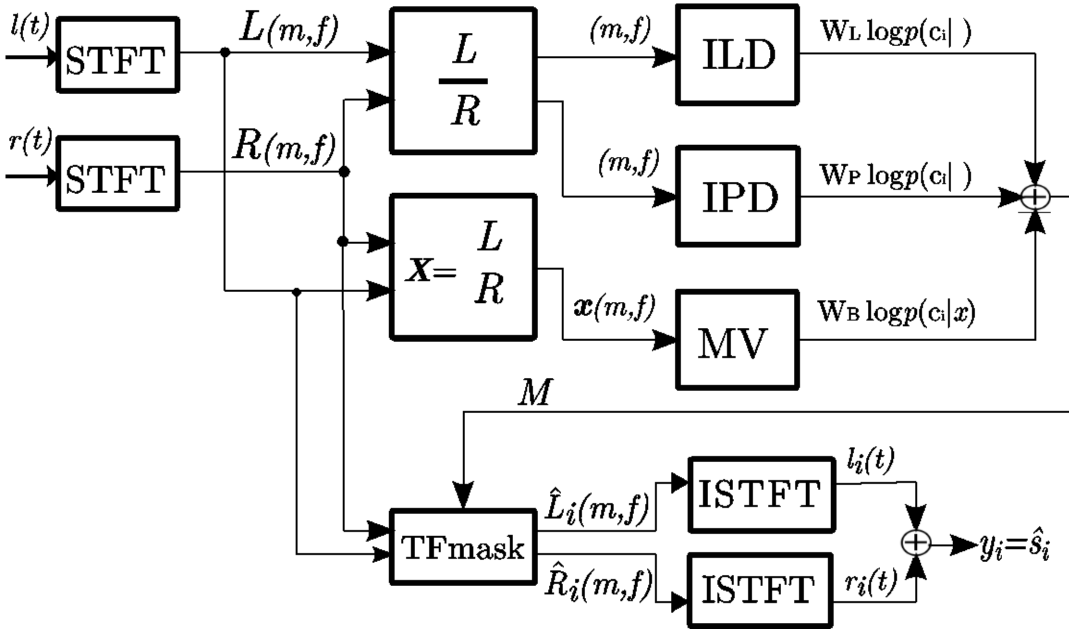
W. Wang, S. Sanei, and J. A. Chambers, Penalty function based joint diagonalization approach for convolutive blind separation of nonstationary sources, in *IEEE Trans. Signal Processing*, vol. 53, no. 5, pp. 1654-1669, May 2005.

Time-Frequency Masking (TFM)



Psychophysical tests show that the ideal binary mask results in dramatic speech intelligibility improvements (Brungart et al.'06; Li & Loizou'08). Acknowledgement to D.L. Wang.

TFM: An Example

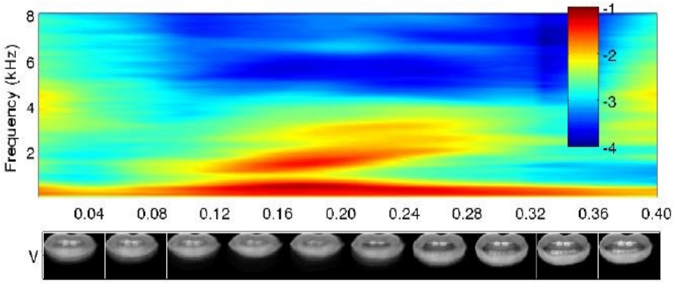


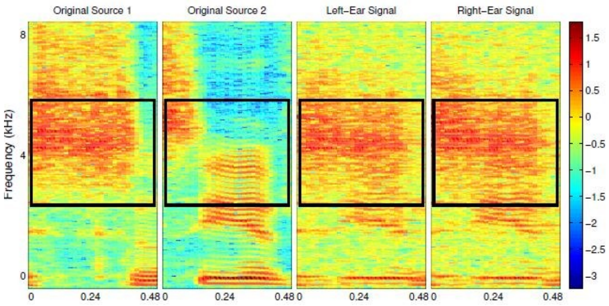
2-source case:

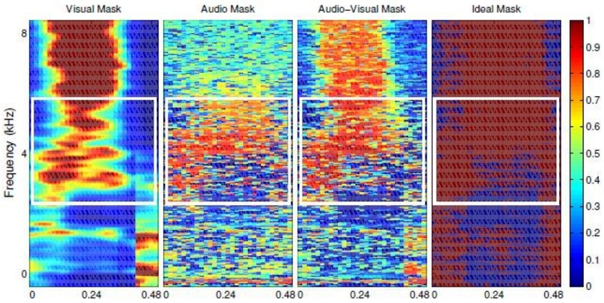
			Mandel et al.	Sawada et al.	Alinaghi et al.	Original
left	🔊	es1	🔊	🔊	🔊	🔊
right	🔊	es2	🔊	🔊	🔊	🔊

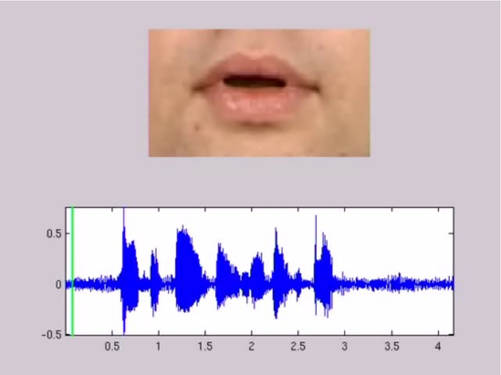
A. Alinaghi, P. Jackson, Q. Liu, and W. Wang, "Joint Mixing Vector and Binaural Model Based Stereo Source Separation", *IEEE Transactions on Audio Speech and Language Processing*, 2014.

AV Speech Source Separation- Demo





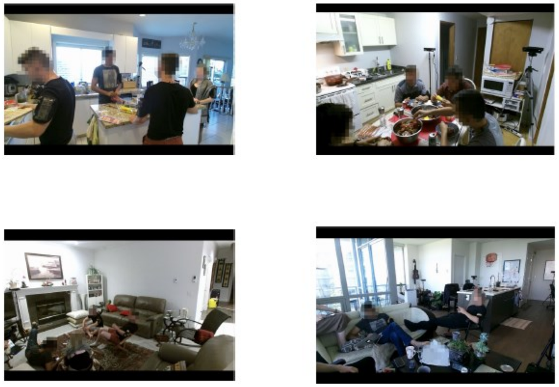




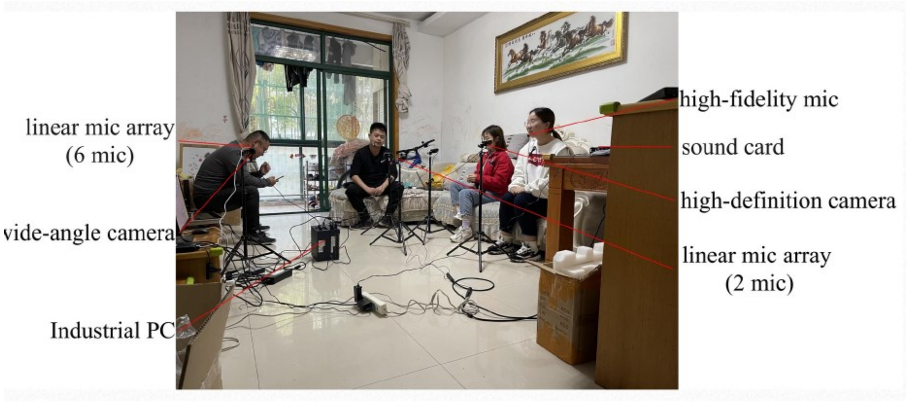
	Mixture	Ideal	Mandel	AV-LIU	AVDL-BSS	Rivet	AVMP-BSS
A	🔊	🔊	🔊	🔊	🔊	🔊	🔊
B	🔊	🔊	🔊	🔊	🔊	🔊	🔊
C	🔊	🔊	🔊	🔊	🔊	🔊	🔊
D	🔊	🔊	🔊	🔊	🔊	🔊	🔊

Q. Liu, W. Wang, et al., "Source separation of convolutive and noisy mixtures using audio-visual dictionary learning and probabilistic time-frequency masking", *IEEE Transactions on Signal Processing*, vol. 61, no. 22, pp. 5520-5535, 2013.

Data Challenges in (AV) Speech Separation



https://chimechallenge.github.io/chime6/

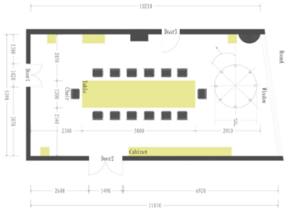


https://mispchallenge.github.io/mispchallenge2022

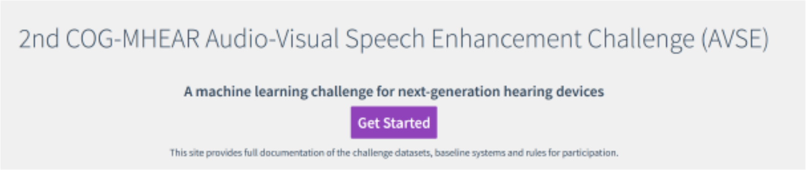
AMI meeting

http://corpus.amiproject.org/

M2MeT challenge -- AliMeeting



https://www.alibabacloud.com/zh/m2met-alimeeting



2nd COG-MHEAR Audio-Visual Speech Enhancement Challenge (AVSE)

A machine learning challenge for next-generation hearing devices

This site provides full documentation of the challenge datasets, baseline systems and rules for participation.

https://challenge.cogmhear.org/

MMCSG (Multi-Modal Conversations in Smart Glasses) datasethttps://ai.meta.com/datasets/mmcsg-dataset/

Other Recent Developments

AV speech separation

R. Gao and K. Grauman, “VISUALVOICE: Audio-Visual Speech Separation with Cross-Modal Consistency,” Proc. CVPR, 2021.

A. Nagrani, et al., “Seeing Voices and Hearing Faces: Cross-modal biometric matching,” in Proc. CVPR, 2018.

AV general sound separation

C. Gan, et al. “Music Gesture for Visual Sound Separation,” in Proc. CVPR, 2020.

E. Tzinis, S. Wisdom, T. Remez, and J.R. Hershey, “AudioScopeV2: Audio-Visual Attention Architectures for Calibrated Open-Domain On-Screen Sound Separation", in Proc. ECCV, 2022.

Universal sound separation

I. Kavalerov, et al., “Universal Sound Separation,” in Proc. IEEE WASPAA, 2019.

Q. Kong et al., “Universal Source Separation with Weakly Labelled Data,” arXiv:2305.07447, 2023.

Queried sound source separation

X. Liu, et al. “Separate What You Describe: Language-Queried Audio Source Separation,” in Proc. Interspeech 2022.

X. Liu, et al. “AudioSep : Separate Anything You Describe”, arXiv:2308.05037, 2023.

Language guided AV source separation

Dong, et al., “CLIPSep: Learning Text-queried Sound Separation with Noisy Unlabeled Videos,” in ICLR 2022.

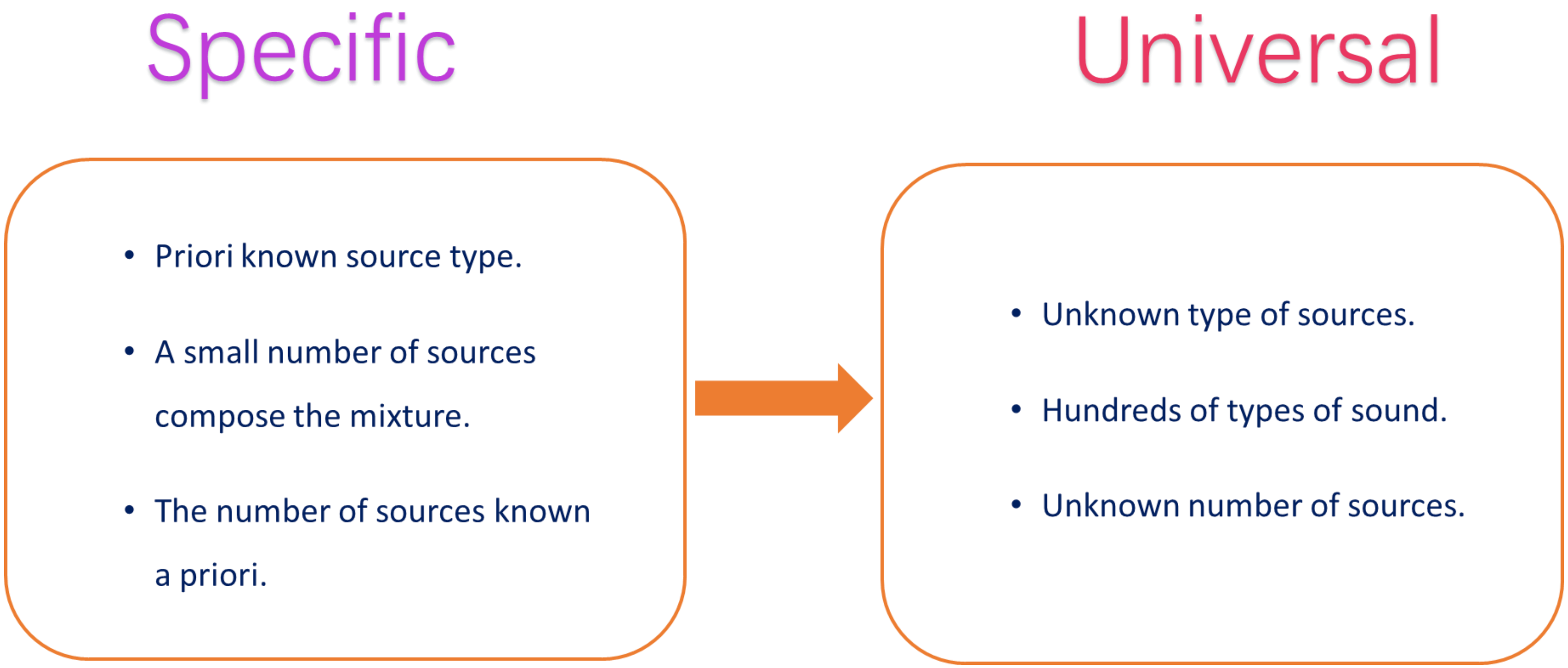
R. Tan, et al., “Language-Guided Audio-Visual Source Separation via Trimodal Consistency,” in Proc. CVPR, 2023.

Singing voice separation

W. Yuan, et al, "Unsupervised Deep Unfolded Representation Learning for Singing Voice Separation," IEEE/ACM Transactions on Audio Speech and Language Processing, vol. 31, pp. 3206 - 3220, 2023.

W. Yuan, et al, "Evolving multi-resolution pooling CNN for monaural singing voice separation", IEEE/ACM Transactions on Audio Speech and Language Processing, vol. 29, pp. 807-821, 2021.

Universal Audio Source Separation



Specific

- Priori known source type.
- A small number of sources compose the mixture.
- The number of sources known a priori.

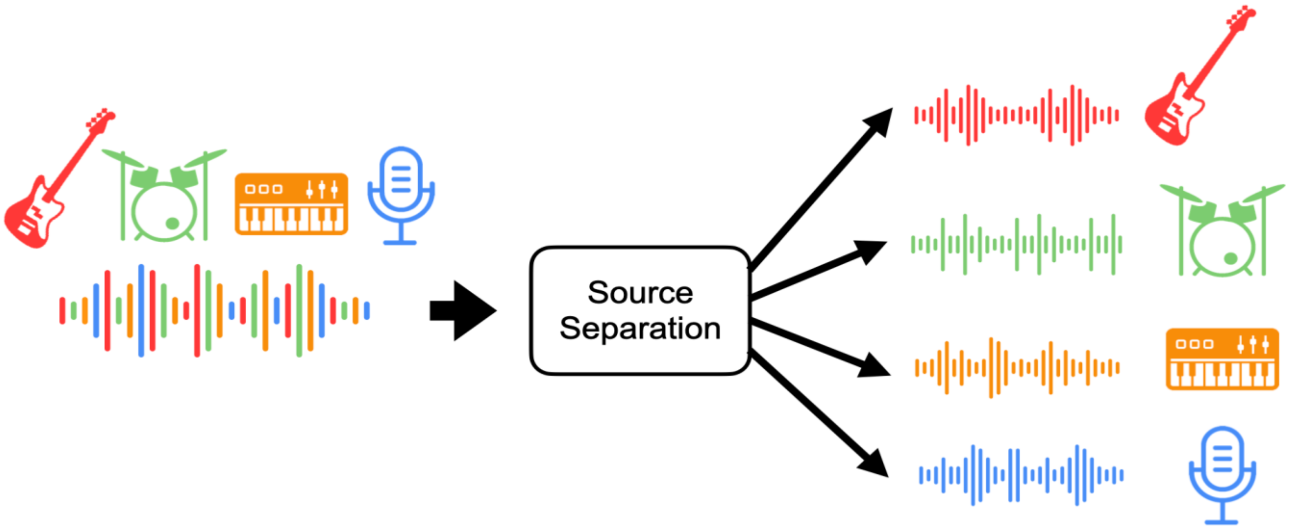
Universal

- Unknown type of sources.
- Hundreds of types of sound.
- Unknown number of sources.

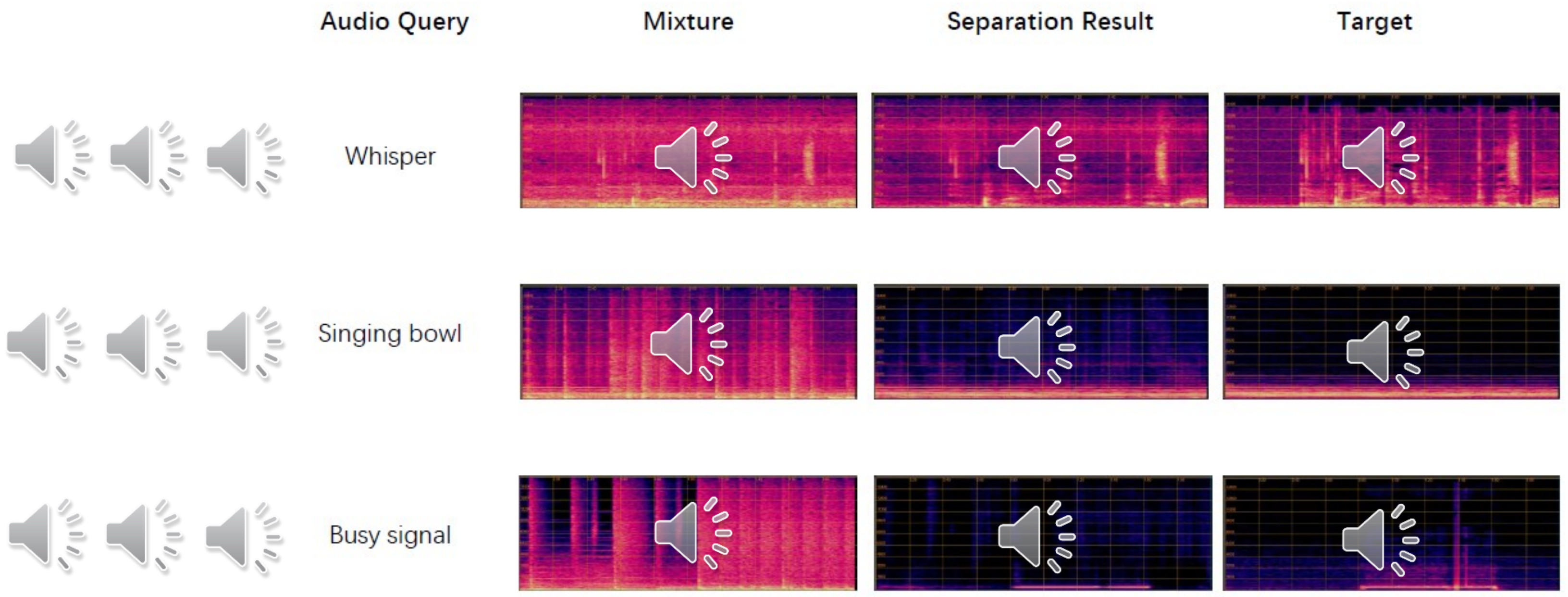
[1] Kavalerov et al, "Universal sound separation," WASPAA 2019.
[2] Tzinis et al, "Improving universal sound separation using sound classification", ICASSP 2020.
[3] Wisdom et al, "What's all the fuss about free universal sound separation data?," ICASSP 2021.
[4] Kong et al, "Universal Source Separation with Weakly Labelled Data," arXiv 2023.

Query-based Audio Source Separation

- Query type: vision, audio, labels, …
- Not flexible and straightforward to separate desired sounds



Universal Audio Source Separation with Query by Embeddings

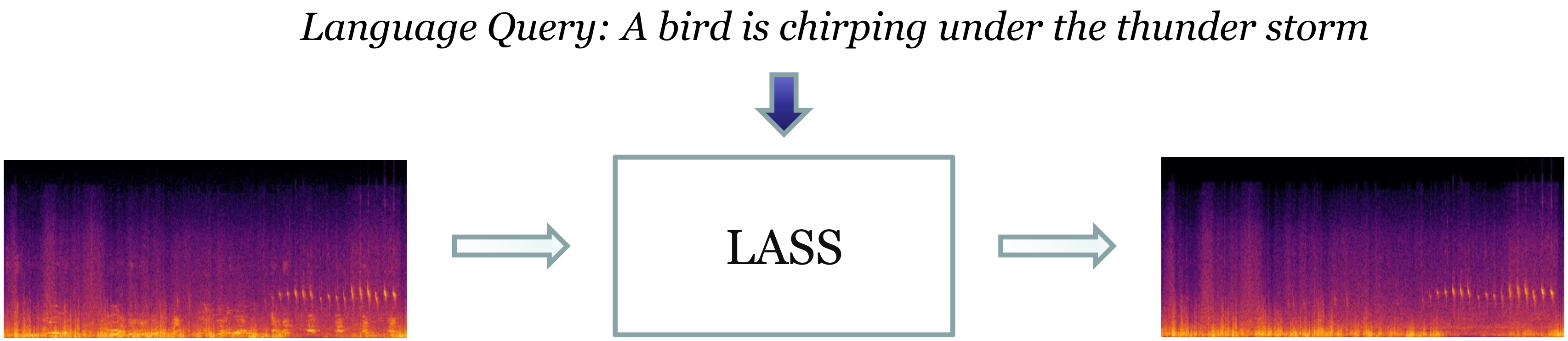


Use embedding from audio query to extract the sound of interest

J. Zhao, X. Liu, J. Zhao, Y. Yuan, Q. Kong, M. Plumbley, and W. Wang, "Universal Sound Separation with Self-Supervised Audio Masked Autoencoder," in *Proceedings of the 32nd European Signal Processing Conference* (EUSIPCO 2024), Lyon, France, August 26-30, 2024.

Language Queried Audio Source Separation

- LASS – Separate a **target source** from an audio mixture based on the **natural language descriptions** of the target source
- First attempt bridging audio source separation and natural language processing
- Support input arbitrary text to separate desired sound sources

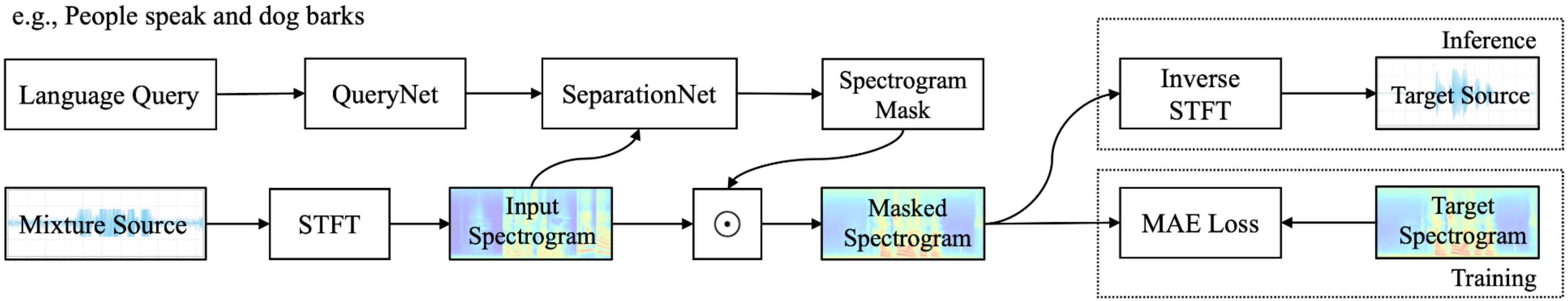


Language Query: A bird is chirping under the thunder storm

X. Liu, H. Liu, Q. Kong, X. Mei, J. Zhao, Q. Huang, M.D. Plumbley, and W. Wang,"Separate What You Describe: Language-Queried Audio Source Separation," in *Proc. 23rd Interspeech Conference* (INTERSPEECH 2022), 18-22 September, 2022, Incheon, Korea.

Challenges and Methods

- **How to construct training data?**
  - Training with synthetic mixtures generating from audio-text datasets
- **LASS-Net**
  - QueryNet (BERT) + SeparationNet (ResUNet) + FiLM fusion
  - Trained with 17 hours of data from AudioCaps dataset



X. Liu, H. Liu, Q. Kong, X. Mei, J. Zhao, Q. Huang, M.D. Plumbley, and W. Wang,"Separate What You Describe: Language-Queried Audio Source Separation," in *Proc. 23rd Interspeech Conference* (INTERSPEECH 2022), 18-22 September, 2022, Incheon, Korea.

Recent LASS

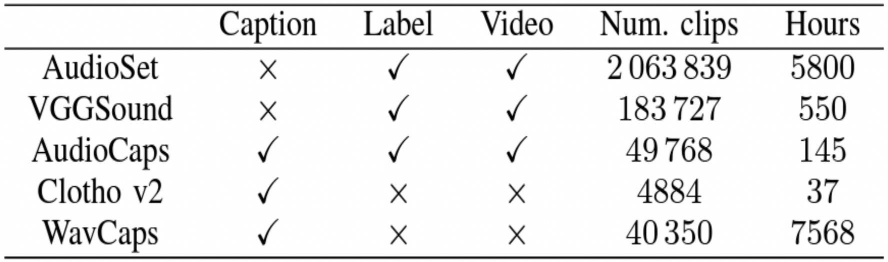
- **Leverage aligned multimodal supervision**
  - Visual supervision (e.g., CLIPSep, *Dong et al. 2023*)
  - Audio supervision (e.g., SoundFilter, *Kilgour 2022*)

- **Great potential to leverage unlabeled rich modality, but**
  - They were trained with small-scale of data (e.g., VGGSound, 500 hours)
  - They may not meet the expectation of open-domain sound separation with texts

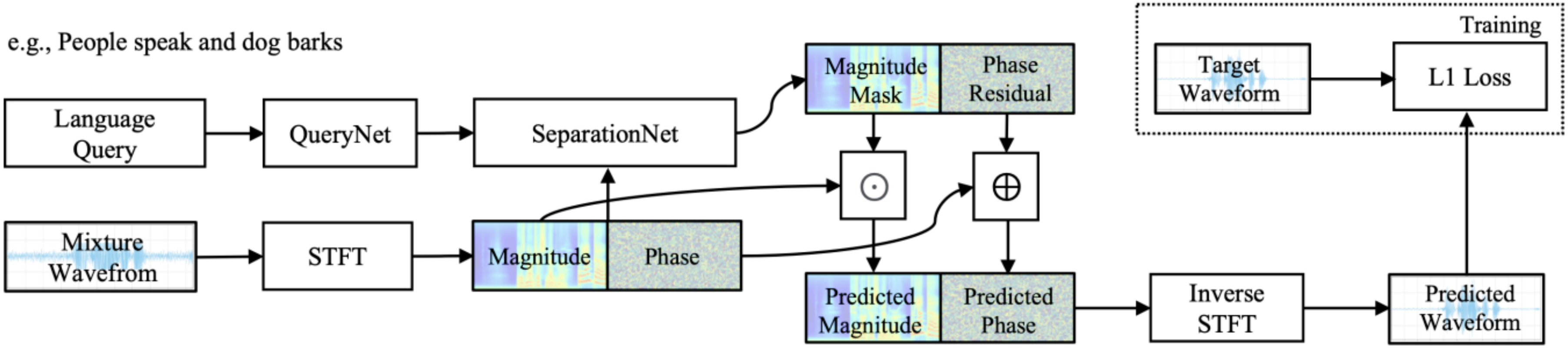
Method 1: AudioSep

- CLAP/CLIP + ResUNet, trained with **14,000** hours of multimodal data
- A foundation model for open-domain sound separation with texts
- Impressive zero-shot performance in separating speech, music, sounds

AUDIOSEP TRAINING DATASETS.



	Caption	Label	Video	Num. clips	Hours
AudioSet	×	✓	✓	2 063 839	5800
VGGSound	×	✓	✓	183 727	550
AudioCaps	✓	✓	✓	49 768	145
Clotho v2	✓	×	×	4884	37
WavCaps	✓	×	×	40 350	7568



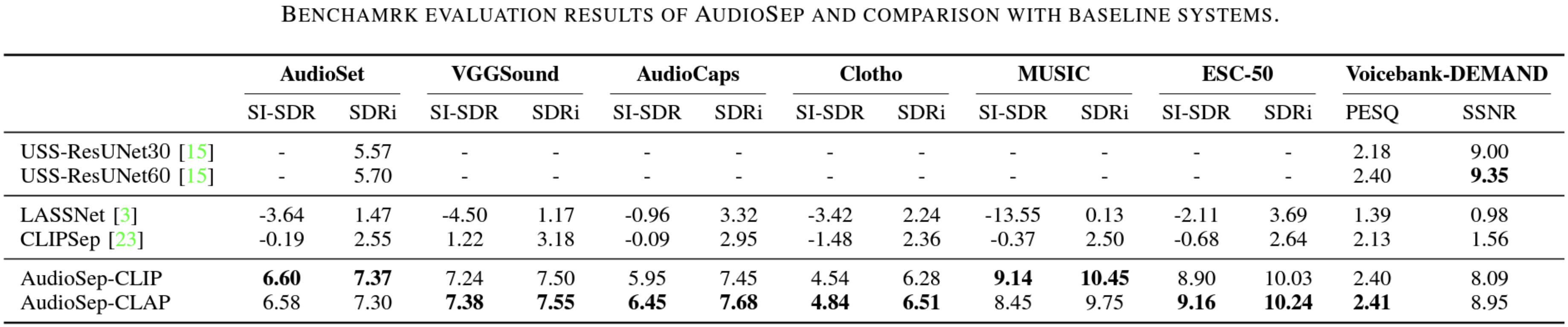
Paper: https://arxiv.org/pdf/2308.05037.pdf

Code: https://github.com/Audio-AGI/AudioSep

Demo: https://huggingface.co/spaces/Audio-AGI/AudioSep

Experimental Results

- AudioSep achieved the **state-of-the-art** results on multiple datasets.
- Impressive zero-shot separation performance on MUSIC and ESC-50.



BENCHAMRK EVALUATION RESULTS OF AUDIOSEP AND COMPARISON WITH BASELINE SYSTEMS.

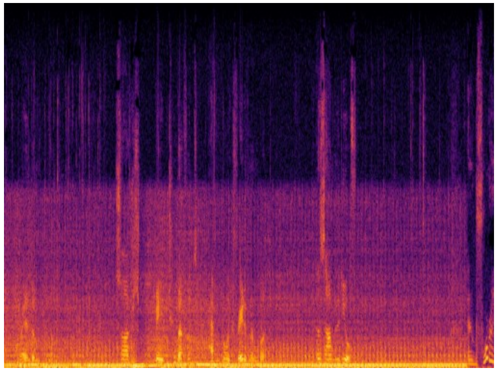
	AudioSet		VGGSound		AudioCaps		Clotho		MUSIC		ESC-50		Voicebank-DEMAND	
	SI-SDR	SDRi	SI-SDR	SDRi	SI-SDR	SDRi	SI-SDR	SDRi	SI-SDR	SDRi	SI-SDR	SDRi	PESQ	SSNR
USS-ResUNet30 [15]	-	5.57	-	-	-	-	-	-	-	-	-	-	2.18	9.00
USS-ResUNet60 [15]	-	5.70	-	-	-	-	-	-	-	-	-	-	2.40	**9.35**
LASSNet [3]	-3.64	1.47	-4.50	1.17	-0.96	3.32	-3.42	2.24	-13.55	0.13	-2.11	3.69	1.39	0.98
CLIPSep [23]	-0.19	2.55	1.22	3.18	-0.09	2.95	-1.48	2.36	-0.37	2.50	-0.68	2.64	2.13	1.56
AudioSep-CLIP	**6.60**	**7.37**	7.24	7.50	5.95	7.45	4.54	6.28	**9.14**	**10.45**	8.90	10.03	2.40	8.09
AudioSep-CLAP	6.58	7.30	**7.38**	**7.55**	**6.45**	**7.68**	**4.84**	**6.51**	8.45	9.75	**9.16**	**10.24**	**2.41**	8.95

X. Liu, Q. Kong, Y. Zhao, H. Liu, Y. Yuan, Y. Liu, R. Xia, Y. Wang, M.D. Plumbley, and W. Wang,"Separate Anything You Describe" in *IEEE/ACM Transactions on Audio Speech and Language Processing,* 2024, accepted.

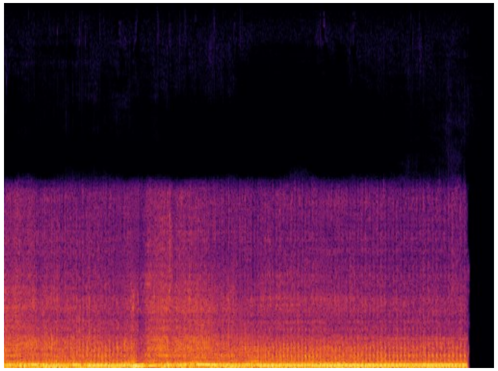
AudioSep: Sound Demos

Human query: “The engine sound of a vehicle”

Mix

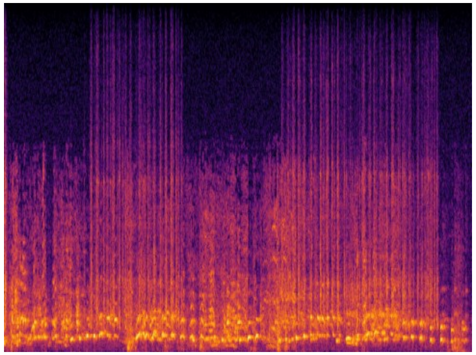


Separated

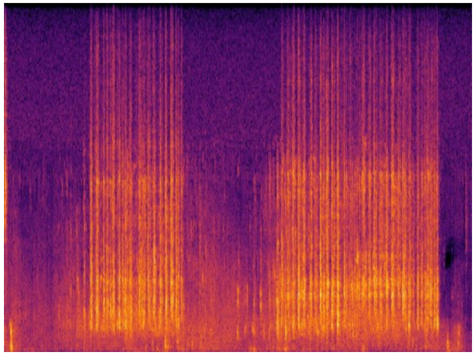


Human query: “The sound of hitting the keyboard”

Mix



Separated



Method 2: FlowSep

Existing Ideas:

- Discriminative approaches.
- Time-frequency masking on spectrogram to remove the noise sound sources.

Challenges:

- Challenging with overlapping sound events.
- Excessive and insufficient masking leads to **artifacts**, including spectral holes and incomplete separation.

A “New” Idea:

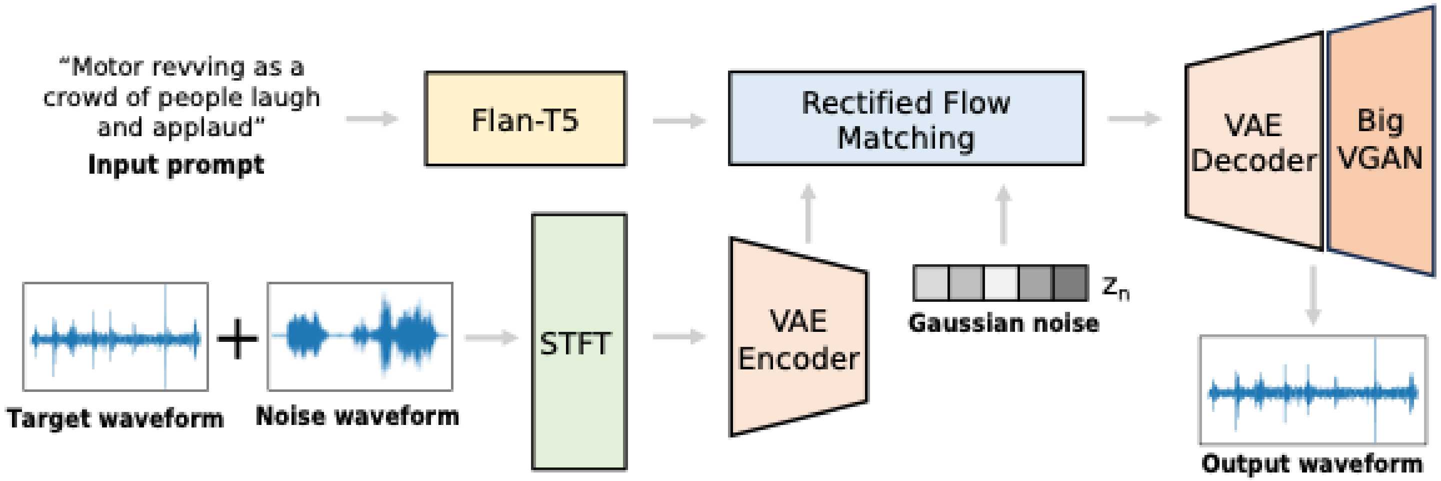
Using the generative approaches.

Diffusion-based generation framework with rectified flow matching.

Separation system by generating new audio samples with the noise clips and text prompts as a condition.

FlowSep: Architecture

- Text-to-audio generation model as the backbone, Rectified Flow Matching for feature generation.
- Extended VAE latent space to integrate the noise audio feature.
- Flan-T5 text encoder, VAE latent decoder and BigVGAN vocoder.



Y. Yuan, X. Liu, H. Liu, M.D. Plumbley, and W. Wang,"FlowSep: Language-Quried Sound Separation with Rectified Flow Matching" in *ICASSP 2025*, submitted.

Training Data

A total of 1,680 hours of audio from various datasets for training. When creating the mixture audio samples, every two audio clips are ensure not sharing any overlapping sound source classes. All the segments are padded or cropped to 10 seconds with 16kHz sampling rate, and we mixture two waveform with a random SNR between -15 and 15 dB.

- AudioCaps:
  - One of the largest publicly available audio captioning dataset, containing 49837 10-second audio clips with human annotated captions.
- VGGSound:
  - Audio dataset with 200,000 audio clips. Each sample has a duration of 10 seconds and annotated with labels.
- WavCaps
  - Large-scale audio dataset with weakly-labelled captions generated with LLM. We only use the samples less than 10 seconds and collected a total of 400,000 clips.

Evaluation Data

- VGGSound:
  - 2000 mixtures generated from a group of 200 clean and distinct audio samples, mixed with random LUFS loudness between -35 and -25 dB.
- ESC-50:
  - 2000 mixtures with a SNR at 0 dB.
- AudioCaps:
  - 928 samples by mixing the audio from testing set under random SNR rate between -15 and 15 dB.
- DCASE2024 Task 8:
  - DCASE-Synth includes 3000 mixtures from 1000 selected audio clips under an SNR rate between -15 and 15 dB.
  - DCASE-Real consists of 100 audio clips from read-world scenarios.

Experimental Results

- Unlike discriminative network that modify on the original audio clips, generated results do not strictly align with the target audio sample in the temporal dimension.
- Hence, traditional objective metrics are not suitable for generative based models.
- We apply FAD, CLAPScore and CLAPScore$_A$ from generative tasks to evaluate the performance.

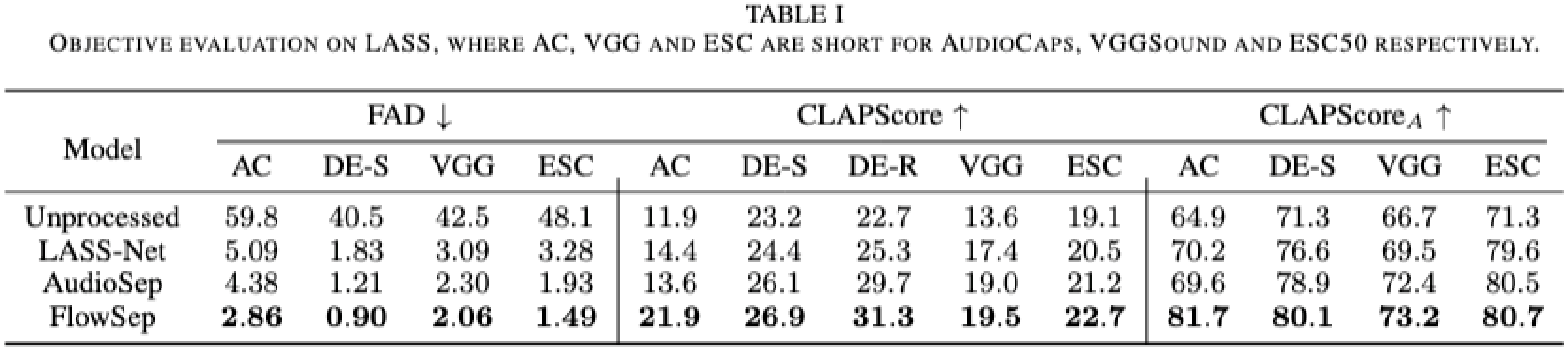


TABLE I
OBJECTIVE EVALUATION ON LASS, WHERE AC, VGG AND ESC ARE SHORT FOR AUDIOCAPS, VGGSOUND AND ESC50 RESPECTIVELY.

Model	FAD ↓				CLAPScore ↑					CLAPScore$_A$ ↑			
	AC	DE-S	VGG	ESC	AC	DE-S	DE-R	VGG	ESC	AC	DE-S	VGG	ESC
Unprocessed	59.8	40.5	42.5	48.1	11.9	23.2	22.7	13.6	19.1	64.9	71.3	66.7	71.3
LASS-Net	5.09	1.83	3.09	3.28	14.4	24.4	25.3	17.4	20.5	70.2	76.6	69.5	79.6
AudioSep	4.38	1.21	2.30	1.93	13.6	26.1	29.7	19.0	21.2	69.6	78.9	72.4	80.5
FlowSep	**2.86**	**0.90**	**2.06**	**1.49**	**21.9**	**26.9**	**31.3**	**19.5**	**22.7**	**81.7**	**80.1**	**73.2**	**80.7**

Experimental Results (Cont.)

- Results on subjective evaluations.

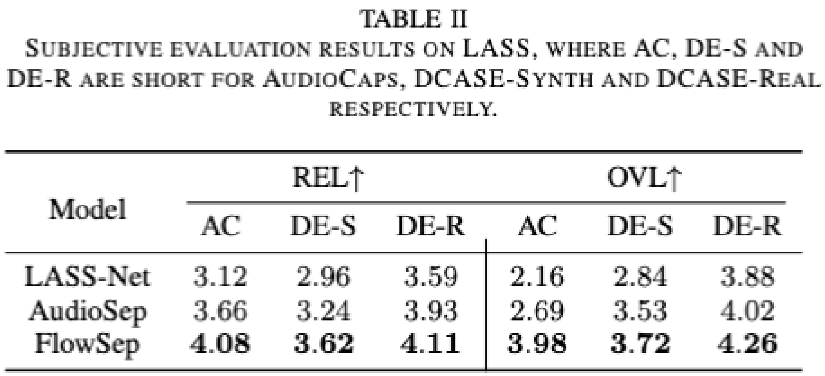


TABLE II
SUBJECTIVE EVALUATION RESULTS ON LASS, WHERE AC, DE-S AND DE-R ARE SHORT FOR AUDIOCAPS, DCASE-SYNTH AND DCASE-REAL RESPECTIVELY.

Model	REL↑			OVL↑		
	AC	DE-S	DE-R	AC	DE-S	DE-R
LASS-Net	3.12	2.96	3.59	2.16	2.84	3.88
AudioSep	3.66	3.24	3.93	2.69	3.53	4.02
FlowSep	**4.08**	**3.62**	**4.11**	**3.98**	**3.72**	**4.26**

- Ablation studies between RFM and traditional diffusion-based models.

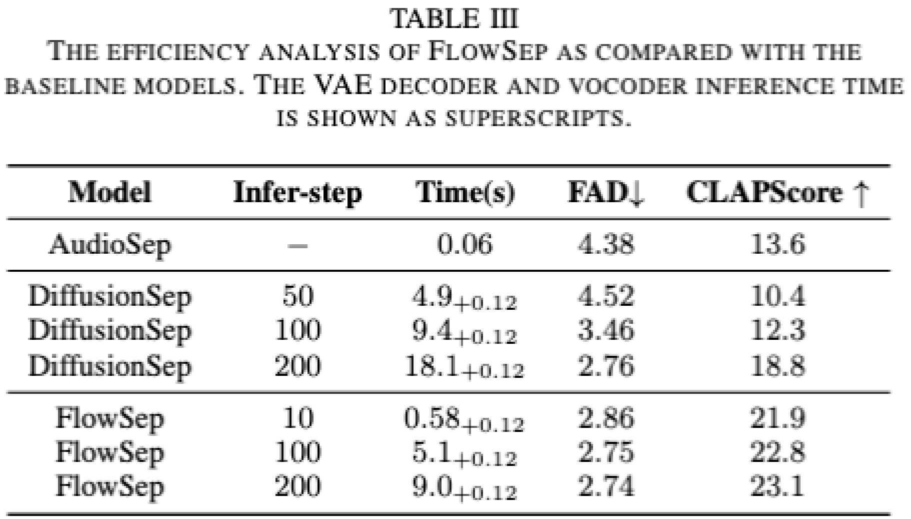
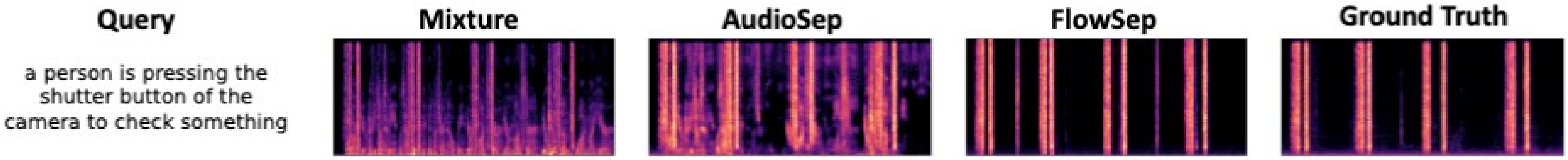


TABLE III
THE EFFICIENCY ANALYSIS OF FLOWSEP AS COMPARED WITH THE BASELINE MODELS. THE VAE DECODER AND VOCODER INFERENCE TIME IS SHOWN AS SUPERSCRIPTS.

Model	**Infer-step**	**Time(s)**	**FAD↓**	**CLAPScore ↑**
AudioSep	–	0.06	4.38	13.6
DiffusionSep	50	$4.9_{+0.12}$	4.52	10.4
DiffusionSep	100	$9.4_{+0.12}$	3.46	12.3
DiffusionSep	200	$18.1_{+0.12}$	2.76	18.8
FlowSep	10	$0.58_{+0.12}$	2.86	21.9
FlowSep	100	$5.1_{+0.12}$	2.75	22.8
FlowSep	200	$9.0_{+0.12}$	2.74	23.1

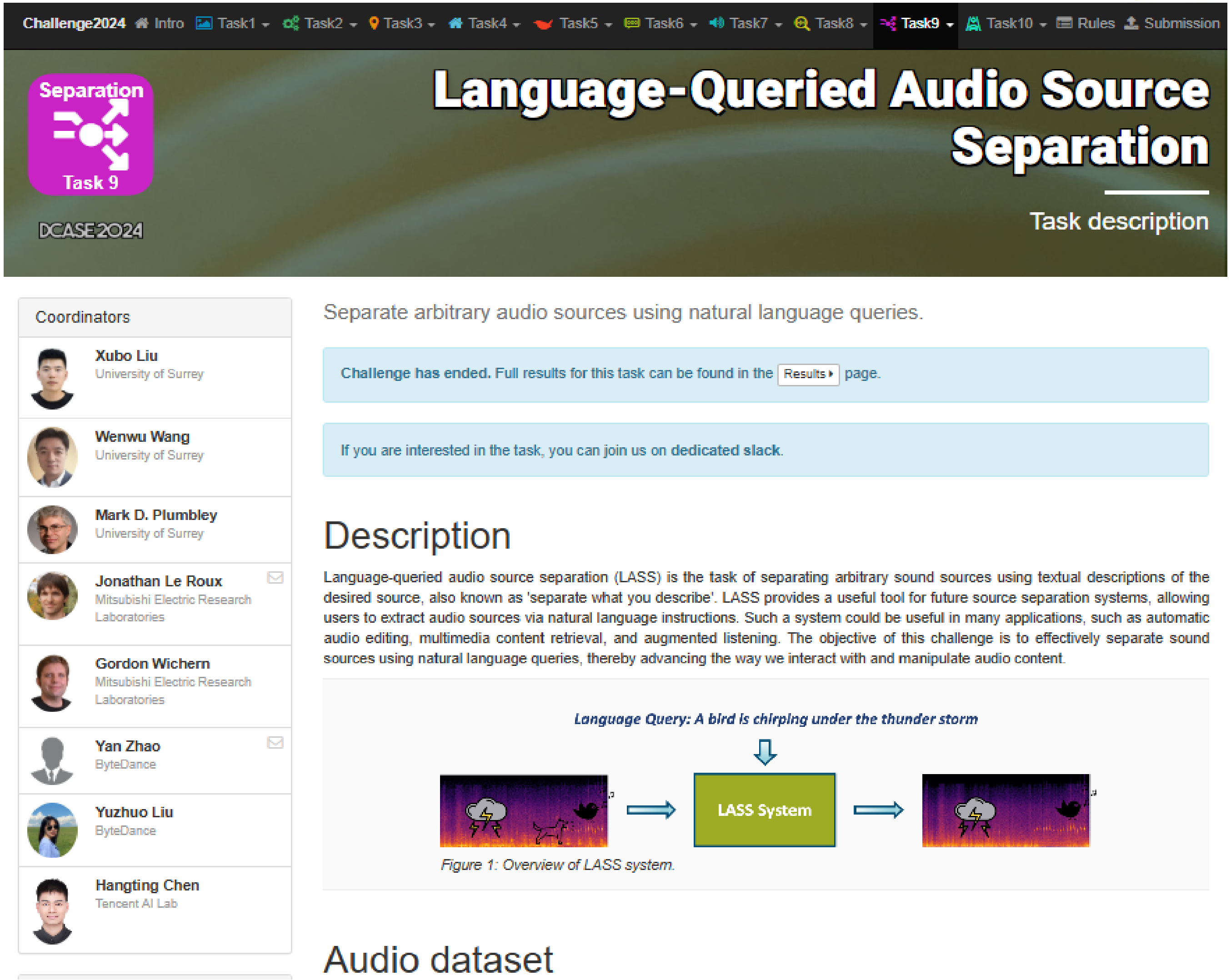
FlowSep - Demos



- Baseline models show incomplete separation with noticeable spectral gaps.
- FlowSep demonstrates promising capabilities in such situations.
- More demos please refer to https://audio-agi.github.io/FlowSep_demo/.

Y. Yuan, X. Liu, H. Liu, M.D. Plumbley, and W. Wang,"FlowSep: Language-Quried Sound Separation with Rectified Flow Matching" in *ICASSP 2025*, submitted.

DCASE Challenge –Task 10



Challenge2024 Intro Task1 Task2 Task3 Task4 Task5 Task6 Task7 Task8 Task9 Task10 Rules Submission

Separation

Task 9

DCASE2024

Language-Queried Audio Source Separation

Task description

Coordinators

Xubo Liu
University of Surrey

Wenwu Wang
University of Surrey

Mark D. Plumbley
University of Surrey

Jonathan Le Roux
Mitsubishi Electric Research Laboratories

Gordon Wichern
Mitsubishi Electric Research Laboratories

Yan Zhao
ByteDance

Yuzhuo Liu
ByteDance

Hangting Chen
Tencent AI Lab

Separate arbitrary audio sources using natural language queries.

Challenge has ended. Full results for this task can be found in the Results ▸ page.

If you are interested in the task, you can join us on dedicated slack.

Description

Language-queried audio source separation (LASS) is the task of separating arbitrary sound sources using textual descriptions of the desired source, also known as 'separate what you describe'. LASS provides a useful tool for future source separation systems, allowing users to extract audio sources via natural language instructions. Such a system could be useful in many applications, such as automatic audio editing, multimedia content retrieval, and augmented listening. The objective of this challenge is to effectively separate sound sources using natural language queries, thereby advancing the way we interact with and manipulate audio content.

Figure 1: Overview of LASS system.

Audio dataset

https://dcase.community/challenge2024/task-language-queried-audio-source-separation

Possible Future Directions

Conclusions:

- Language queried audio source separation offers tools for users to control which sound to be separated from the sound mixtures, using language-based queries.

Future directions:

- Leverage generative models (e.g., diffusion models) to improve the perceptual quality of separated sounds.
- Explore advanced reasoning capabilities of LLMs for LASS (e.g., separating complex queries like "the second sound" or "annoyed sounds").
- Apply self-supervised techniques (e.g., MixIT) for pre-training to enhance separation performance.

Take Away

EMRPCNN:
Code/demos at project page:
https://github.com/tuxzz/emrpcnn_pub
https://tuxzz.org/emrpcnn-ckpt/

My contact:

Email: w.wang@surrey.ac.uk
Web: https://personalpages.surrey.ac.uk/w.wang/

FlowSep:
Paper/Code/Demo:
https://audio-agi.github.io/FlowSep_demo

AudioSep:
Code/paper/demo:

- DCASE 2024 Task 9: “Language-Queried Audio Source Separation”
- GitHub: https://github.com/Audio-AGI/AudioSep
- HuggingFace: https://huggingface.co/spaces/Audio-AGI/AudioSep
- Media coverage:

Google Research

Stars 1.3k

gradio

UNITE.AI

CyberEd.io

ARCHETYP Staffing Magazine

Medium

AIPressRoom