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/

Keynote talk on IWAENC 2026, Cremona, Italy

10/09/2026

Acknowledgement

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

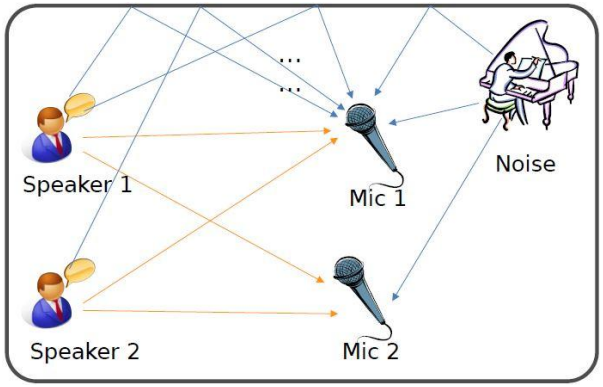
Funding from:

UK Engineering and Physical Sciences Research Council (EPSRC) & University Defence Research Collaboration (UDRC), British Council, Research England, industry including BBC, Samsung, Huawei, etc.

Outline

- **A brief history about audio source separation**
  - Cocktail party problem & speech source separation
  - Example problems and methods
  - Recent developments
- **A brief history about (large) audio-language models**
  - Large audio models
  - Large language models
  - Large audio-language models (examples, datasets, applications)
- **Language queried audio source separation**
  - Early models
  - AudioSep
  - FlowSep & FlowSep2
  - A related problem: language guided audio editing
- **Conclusions and future works**

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)

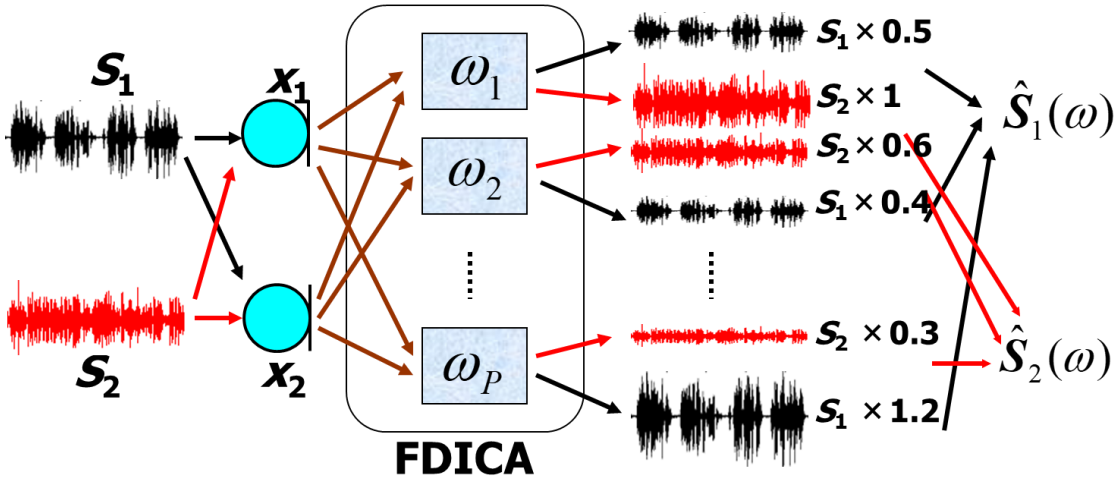
Cocktail party problem may involve a few tasks:

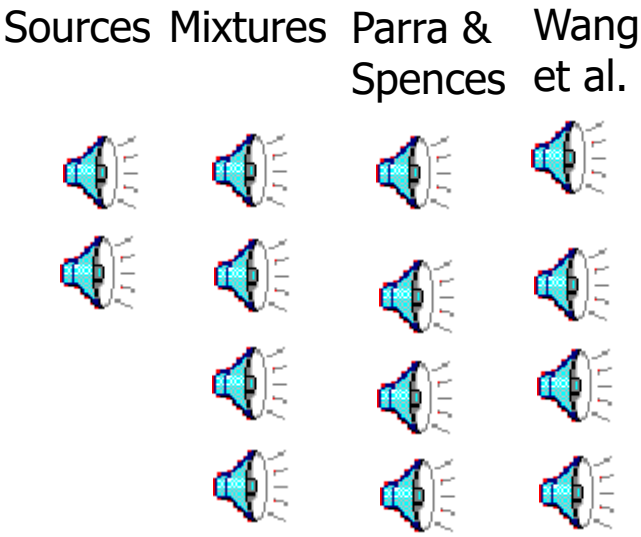
How many speakers?
(Source counting)
Where are they?
(Localization and tracking)
Who speaks and when?
(Diarization)
What are the individual speech sources?
(**Speech source separation**)
What has been said?
(Automatic speech recognition)
What kind of environment?
(Acoustic scene recognition, event detection, room acoustics)

Speech Source Separation

- Various methods explored for the speech source separation problem
  - Beamforming
  - Adaptive filtering
  - Blind source separation and independent component analysis
  - Sparse representation and dictionary learning
  - Matrix/tensor factorization techniques
  - Time-frequency masking
  - Learning based techniques
  - Computational auditory scene analysis
  - Exploiting additional modalities or conditions (e.g. visual signals)
  - …

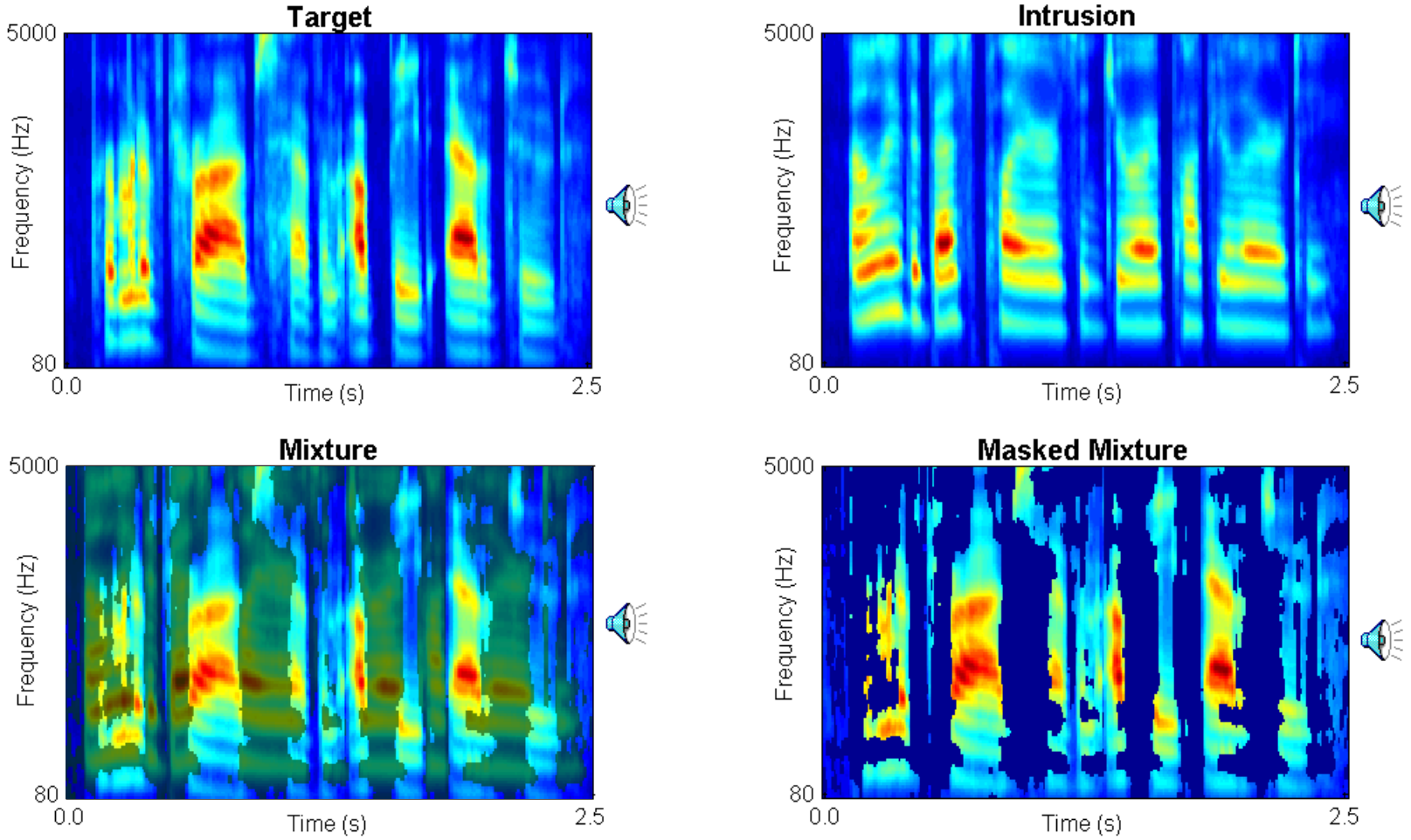
Frequency Domain ICA: Demo





P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," Neurocomput., vol. 22, pp. 21–34, 1998.
L. Parra and C. Spence, "Convolutive blind source separation of nonstationary sources," IEEE Trans. Speech Audio Process., vol. 8, no. 3, pp. 320–327, May 2000.
H. Sawada, R. Mukai, S. Araki, S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," IEEE Trans. on Speech and Audio Processing, vol. 12, no. 5, pp. 530-538, 2004.
W. Wang, S. Sanei, and J. A. Chambers, "Penalty function based joint diagonalization approach for convolutive blind separation of nonstationary sources," IEEE Trans. Signal Processing, vol. 53, no. 5, pp. 1654-1669, May 2005.
H. Buchner, R. Aichner, W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," IEEE Trans. on Speech and Audio Processing, vol. 13, no. 1, pp. 120-134, 2005.

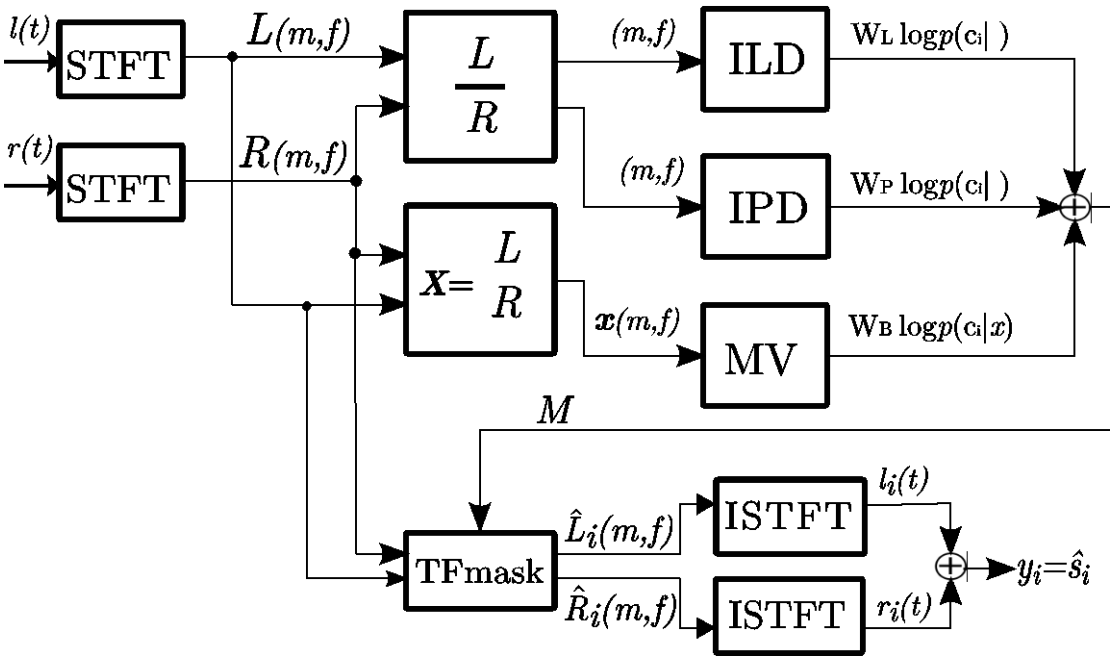
Time-Frequency Masking (TFM)



Psychophysical tests show that the ideal binary mask results in dramatic speech intelligibility improvements (Brungart et al., 2006; Li & Loizou, 2008).

Acknowledgement to Prof DeLiang Wang for this demo.

TFM: An Example



left Source 1

right Source 2

	Sawada et al.	Mandel et al.	Alinaghi et al.
es1			
es2			

O. Ylmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” IEEE Trans. Signal Process., vol. 52, no. 7, pp. 1830–1847, Jul. 2003.

M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, “Model-based expectation-maximization source separation and localization,” IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 2, pp. 382–394, Feb. 2010

H. Sawada, S. Araki, and S. Makino, “Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment,” IEEE Trans. Audio, Speech, Lang. Process., vol. 19, no. 3, pp. 516–527, Mar. 2011.

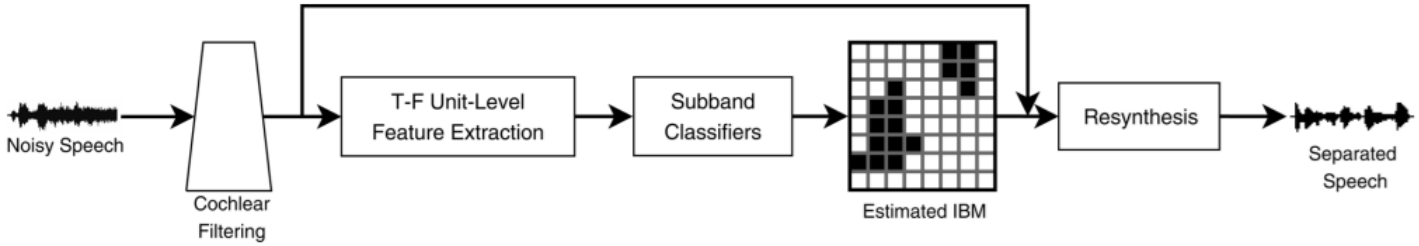
A. Alinaghi, P. Jackson, Q. Liu, and W. Wang, "Joint mixing vector and binaural model based stereo source separation", IEEE Transactions on Audio Speech and Language Processing, vol. 22, no. 9, pp. 1434-1448, 2014.

More Recent Developments

Neural T-F masking

Formulate separation problem as T-F unit classification problem:

- Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," IEEE Transactions on Audio, Speech, and Language Processing, 2013.
- Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," IEEE/ACM TASLP, 22(10):1849–1858, 2014.



Embedding based separation

Deep Clustering: learns the structure of T-F mask:

- J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: discriminative embeddings for segmentation and separation," ICASSP 2016, pp. 31–35.

Deep Attractor Network: learns an embedding space in which each speaker is represented by an attractor point:

- Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," ICASSP 2017, pp. 246–250.

$$\text{mixture} \rightarrow \text{T–F embeddings} \rightarrow \text{clustering} \rightarrow \text{masks} \rightarrow \text{sources}$$

More Recent Developments

UNIVERSITY OF SURREY

Permutation free end to end separation

Permutation Invariant Training (PIT):

- D. Yu, M. Kolbæk, Z.-H. Tan, and J. H. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," ICASSP 2017, pp. 241–245.

$$y_1 = \text{Speaker A}, \quad y_2 = \text{Speaker B}$$

$$\mathcal{L}_{\text{PIT}} = \min_{\pi \in S_K} \mathcal{L}(\hat{\mathbf{s}}_\pi, \mathbf{s})$$

Conv-TasNet: learn the representation of mixture and masks in the time domain

- Y. Luo and N. Mesgarani, "TasNet: surpassing ideal time-frequency masking for speech separation," arXiv:1805.12347, 2018.
- Y. Luo and N. Mesgarani, "Conv-TasNet: surpassing ideal time-frequency magnitude masking for speech separation," IEEE/ACM TASLP, 2019.

$$x(t) \xrightarrow{\text{STFT}} X(f,t) \xrightarrow{\text{mask}} \hat{S}(f,t) \xrightarrow{\text{iSTFT}} \hat{s}(t)$$

$$x(t) \rightarrow \mathbf{w} \rightarrow \mathbf{M} \rightarrow \hat{\mathbf{w}} \rightarrow \hat{s}(t)$$

More Recent Developments

Long context modelling & transformer architecture:

DPRNN:

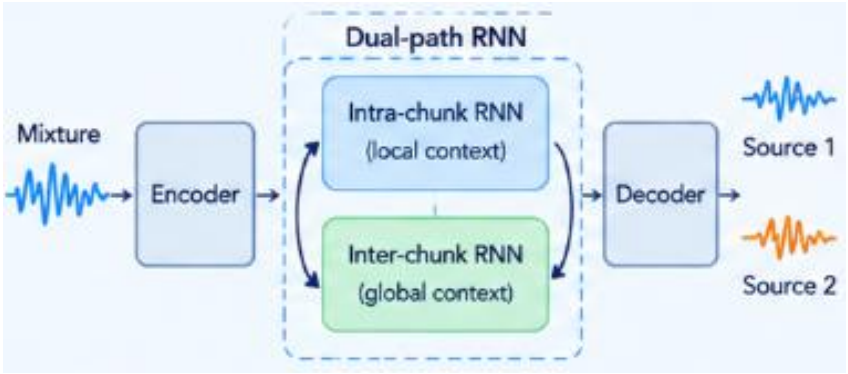
- Y. Luo, Z. Chen, and T. Yoshioka, "Dual-Path RNN: efficient long sequence modeling for time-domain single-channel speech separation," ICASSP, 2020.

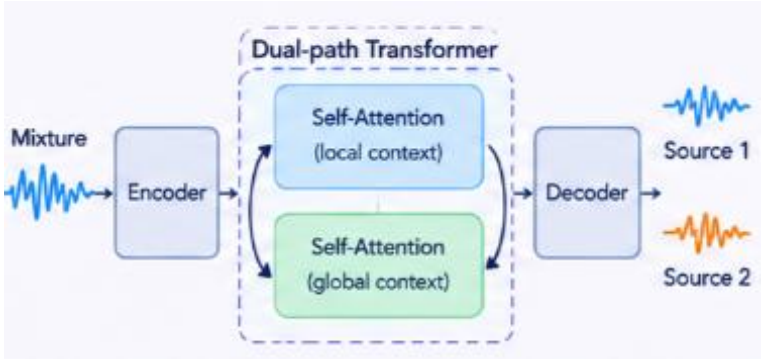
DPTNet:

- J. Chen et al., "Dual-Path transformer network: direct context-aware modeling for end-to-end monaural speech separation," ICASSP, 2020.

SepFormer:

- C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," ICASSP, 2021.



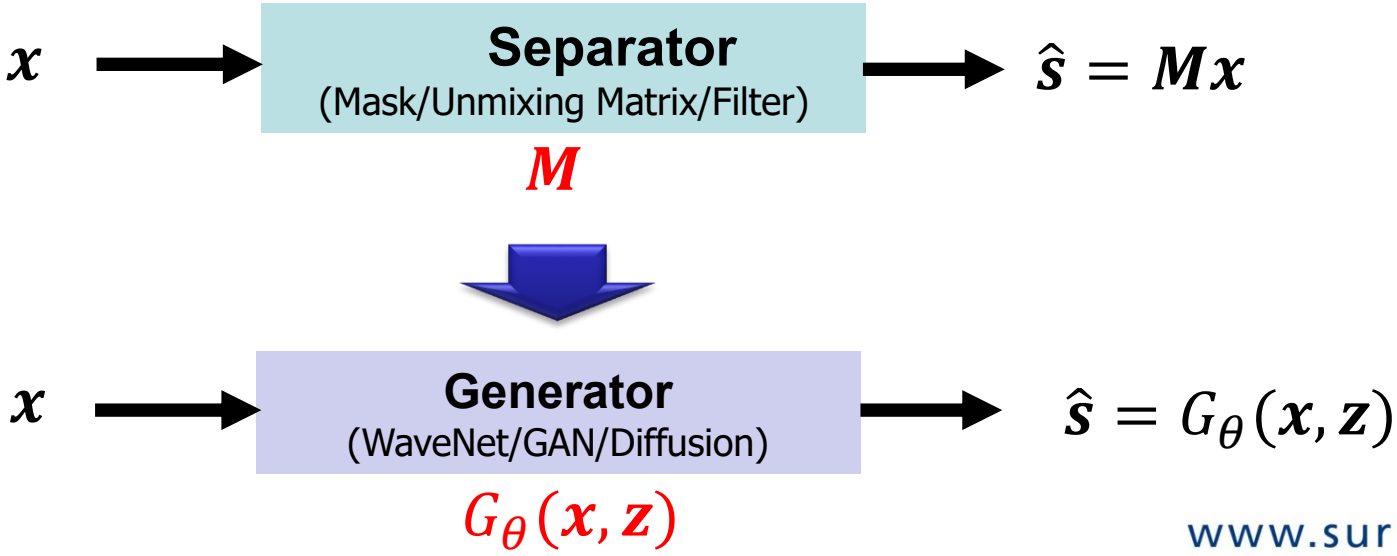


More Recent Developments

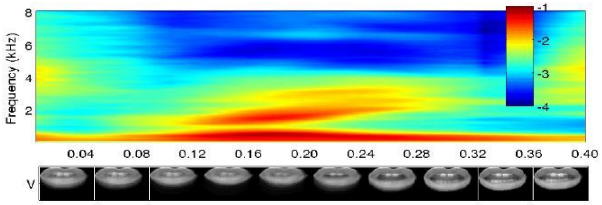
Generative models based separation systems:

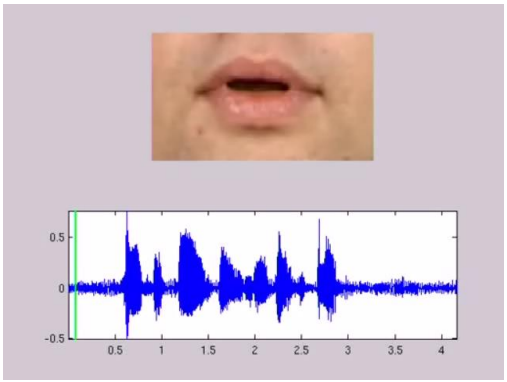
WaveNet/GAN/Diffusion model: $p(\boldsymbol{s}|\boldsymbol{x})$ $\hat{\boldsymbol{s}} = p_\theta(\boldsymbol{s}|\boldsymbol{x})$

- Q. Liu, P. Jackson, and W. Wang, "A speech synthesis approach for high quality speech separation and generation", IEEE Signal Processing Letters, vol. 26, no. 12, pp. 1872-1876, December 2019.
- L. Chen, M. Yu, Y. Qian, D. Su, and D. Yu, "Permutation invariant training of generative adversarial network for monaural speech separation," INTERSPEECH 2018.
- R. Scheibler, Y. Ji, S.-W. Chung, J. Byun, S. Choe, and M.-S. Choi, "Diffusion-based generative speech source separation," ICASSP 2023.
- C. Zhao, G. Sun, K. Deng, C. Li, and P. C. Woodland, "STR-DIFFSEP: Streamable diffusion model for speech separation," ICASSP 2026.



AV Speech or General Sound Separation





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

AV speech separation

D. Sodoyer, J.L. Schwartz, L. Girin, J. Klinkisch, and C. Jutten, "Separation of audio-visual speech sources," EURASIP Journal on Applied Signal Processing, pp. 1165-1173, Nov., 2002.

W. Wang, D. Cosker, Y. Hicks, S. Sanei, and J. A. Chambers, "Video assisted speech source separation," Proc. ICASSP, vol. V, pp.425-428, Philadelphia, USA, March 18-23, 2005.

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.

A. Ephrat et al., “Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation,” ACM TOG, 2018.

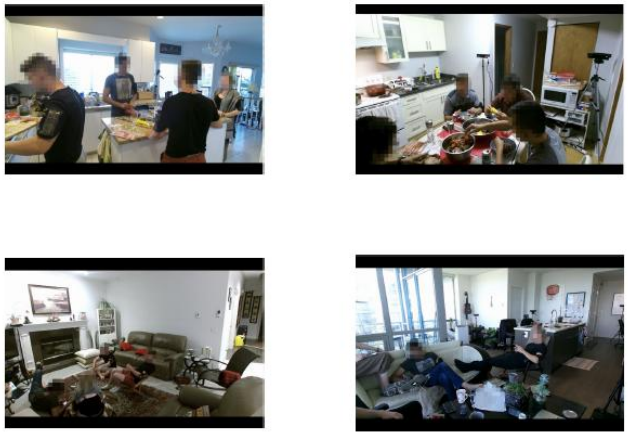
J. Wu et al., “Time domain audio visual speech separation,” IEEE ASRU, 2019.

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.

K. Li et al., “An audio-visual speech separation model inspired by Cortico-Thalamo-Cortical Circuits,” IEEE TPAMI, 2024.

Data Challenges in (AV) Speech Separation



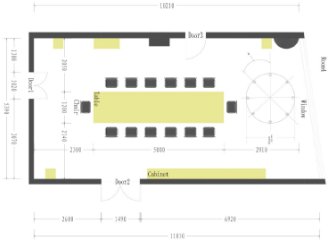
https://chimechallenge.github.io/chime6/

https://mispchallenge.github.io/mispchallenge2022

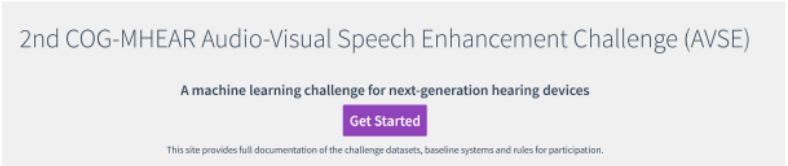
AMI meeting

http://corpus.amiproject.org/

M2MeT challenge -- AliMeeting



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



https://challenge.cogmhear.org/

MMCSG (Multi-Modal Conversations in Smart Glasses) dataset

https://ai.meta.com/datasets/mmcsg-dataset/

Universal Audio Source Separation

UNIVERSITY OF SURREY

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.

I. Kavalerov, et al, "Universal sound separation," WASPAA 2019.
E. Tzinis, et al, "Improving universal sound separation using sound classification", ICASSP 2020.
S. Wisdom, et al, "What's all the fuss about free universal sound separation data?," ICASSP 2021.
Q. Kong, et al, "Universal Source Separation with Weakly Labelled Data," arXiv 2023.

Outline

- **A brief history about audio source separation**
  - Cocktail party problem & speech source separation
  - Example problems and methods
  - Recent developments
- **A brief history about (large) audio-language models**
  - Large audio models
  - Large language models
  - Large audio-language models (examples, datasets, applications)
- **Language queried audio source separation**
  - Early models
  - AudioSep
  - FlowSep & FlowSep2
  - A related problem: language guided audio editing
- **Conclusions and future works**

(Large) Audio Models

Learning **general/universal audio representations** from large scale audio data shows promising performance in downstream tasks (classification, separation, retrieval, etc):

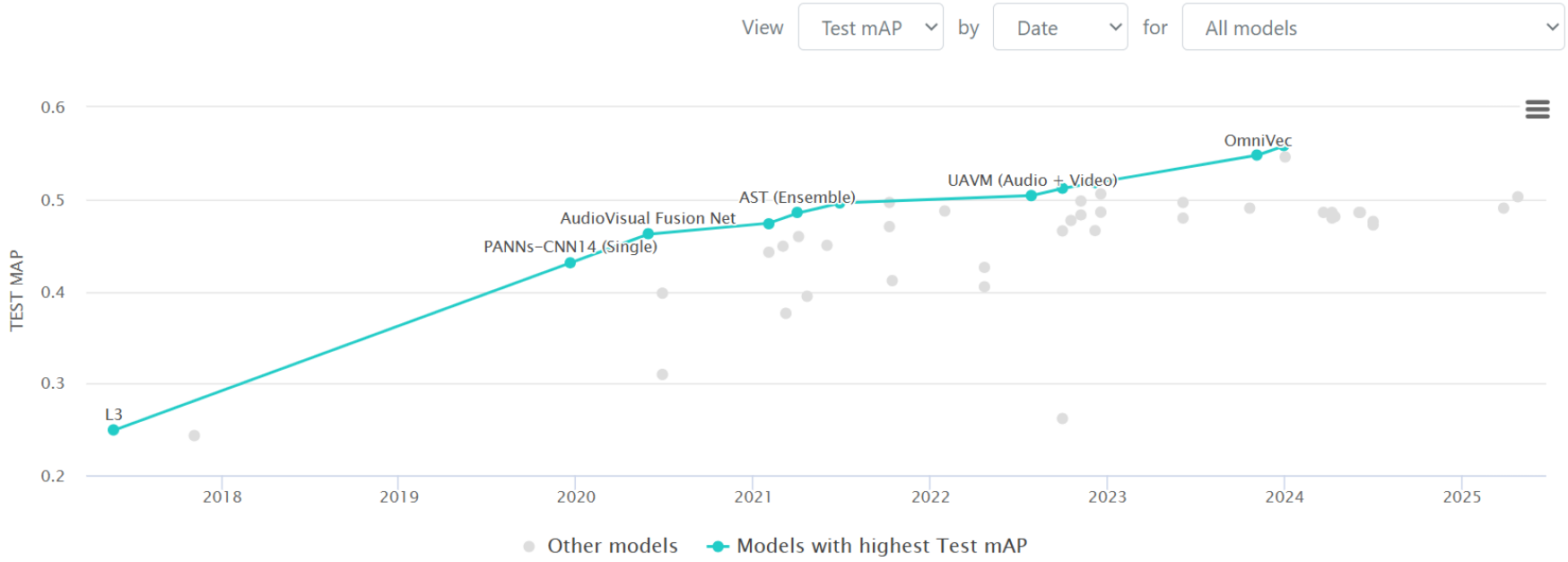
- **PANNs** (Kong, et al, 2020): large scale CNN-based audio model
- Audio2Vec (Tagliasacchi et al, 2020): sequence to sequence unsupervised model
- CLAR (Al-Tahan and Mohsenzadeh, 2021): self-supervised model
- COLA (Saeed et al, 2021): self supervised model
- BOYL-A (Niizumi et al, 2021): self supervised model
- AST (Gong et al, 2021): large scale transformer-based audio model
- ATST (Li and Li, 2022): transformer-based model
- MAE-AST (Baade et al, 2022): self-supervised model
- SSAST (Gong et al, 2022): self-supervised AST model
- Audio-MAE (Huang et al, 2022): self-supervised model
- BEATs (Chen et al, 2023): audio pre-training with acoustic tokenizers
- **ASiT** (Atito et al, 2024): self-supervised models for general audio representation
- SSLAM (Alex et al, 2025): self-supervised learning from audio mixtures

mean Average Precision (mAP) on AudioSet: **31.4** (2017) -> **55.8** (2025)

(Large) Audio Models

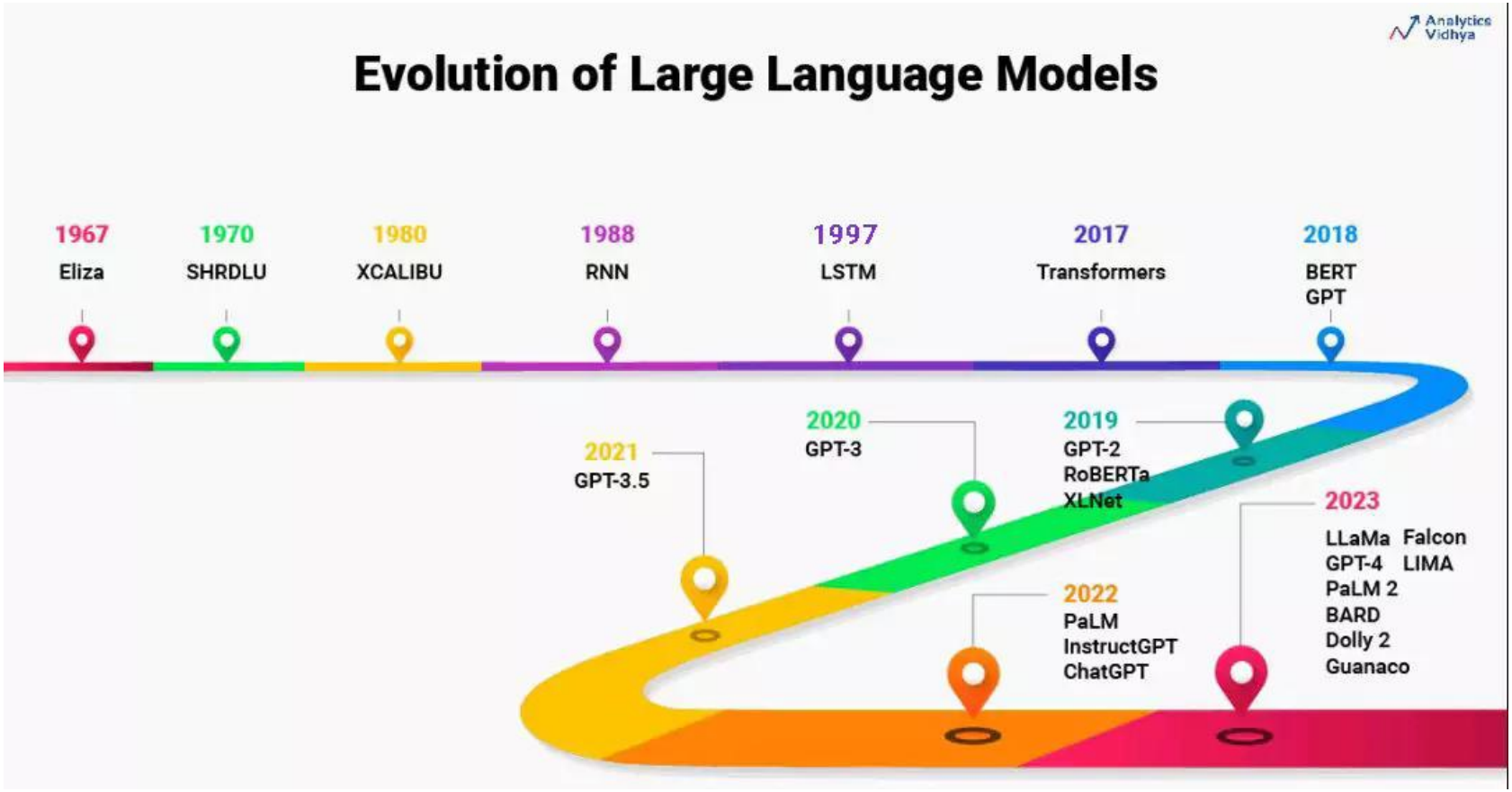
Audio Classification on AudioSet

Leaderboard | Dataset



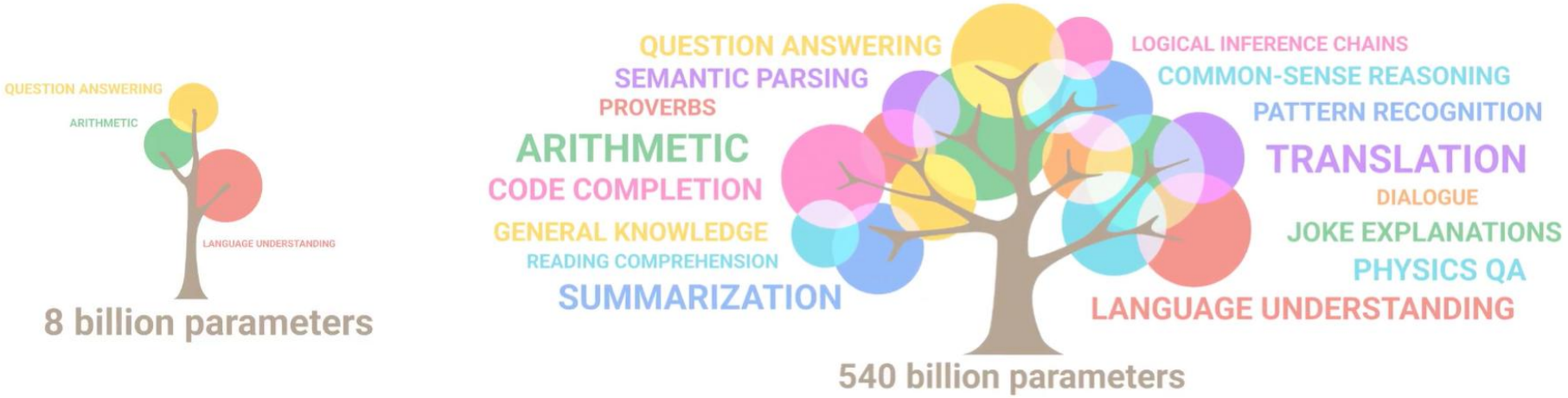
Leaderboard:
https://paperswithcode.com/sota/audio-classification-on-audioset
https://www.codesota.com/audio/classification

Large Language Models (LLMs)



Courtesy to Aravind Pai: https://www.analyticsvidhya.com/blog/2023/07/beginners-guide-to-build-large-language-models-from-scratch/

Large Language Models (LLMs)



Figures from Ryan O'Connor: https://www.assemblyai.com/blog/emergent-abilities-of-large-language-models/

Wei et al, “Emergent abilities of large language models,” Transactions on Machine Learning Research, 2022.

Large Audio-Language Models: Why?

UNIVERSITY OF SURREY

- Leverage knowledge within LLMs to address the limitations of audio models
  - Zero-shot or few-shot classification
- Explore homogeneity across tasks with LLMs
  - Use LLMs as an agent to solve multi-task problems
- Extend the capabilities of audio models for new tasks
  - Extending from audio classification to audio captioning/question answering & reasoning
  - Extending from audio generation to storytelling & controllable editing
  - Extending from audio source separation to language queried audio source separation

Large Audio Language Models - Examples

UNIVERSITY OF SURREY

Whisper: automatic speech recognition (ASR)
Wav2Vec 2.0: speech to text (STT)
DeepSpeech: open-source ASR
Coqui AI: speech synthesis and TTS
Jasper and QuartzNet: ASR
GPT-3 with Whisper: ASR + LLMs
Sonix.ai: speech transcription and analysis
SpeechBrain: platform for speech models
OpenSTT: ASR
Gemini: text/image/speech
WavLLM: speech LLMs

MuseNet: music instrumental composition & style transfer
MusicVAE: melody generation, remixing, and style interpolation
JukeBox: raw music with vocals and lyrics
Riffusion: text to music generation
MusicLM: text to music generation
REMI: symbolic music generation (e.g. MIDI)
DeepBach: classic music composition
AIVA: music generation assistant

Wav2CLIP: audio and language mapping
AudioCLIP: audio-text-image alignment
CLAP: audio-text alignment
CLAP-LAION: audio-text alignment
Pengi: audio classification & AQA
Qwen-Audio: speech, music, general audio
AudioLM: Speech/music generation
LTU: audio QA and reasoning
SALMONN: speech-audio-music LLMs
ImageBind: image, text, audio, depth
ONE-PEACE: audio-text-image
AudioGPT: Speech, Music, Talking Head
UniAudio: speech/sound/music/singing

AudioLDM/AudioLDM2/WavJourney/ DreamAudio: TTA generation
AudioSep/FlowSep/FlowSep2: language guided audio source separation
WavCraft/RFM-Editing/SFM-Adapter/RFM-Editing2: text prompted audio/speech editing
APT-LLM: LLM based AQA and reasoning
GCDance/TeMuDance: Language guided music to dance generation

Audio Language Datasets

- AudioCaps (Kim et al, 2019)
- Clotho (Drossos et al, 2020)
- SoundDescs (Koepke et al, 2021)
- LAION-Audio-630K (Wu et al, 2023)
- Auto-ACD (Xu et al, 2024)
- Audio-FLAN (Xue et al, 2025)
- SeaBench-Audio (Liu et al, 2025)
- MusicSem (Salganik et al, 2026)

- **WavCaps (Mei et al, 2023)**
- **Sound-VECaps (Yuan et al, 2024)**
- **AudioSetCaps (Bai et al, 2025)**

- CLEAR (Abdelnour et al, 2018)
- DAQA (Fayek and Johnson, 2019)
- Clotho-AQA (Lipping et al, 2022)
- MUSIC-AVQA (Li et al, 2022)
- mClothoAQA (Behera et al, 2023)
- OpenAQA-5M (Gong et al, 2023)
- AudioMCQ (He et al, 2025)
- Audio Flamingo 3 (Goel et al, 2025)
- Jamendo-MT-QA (Koh et al, 2026)

Large Audio-Language Models – Recent Progress

Leaderboard: MMAU-v05.15.25

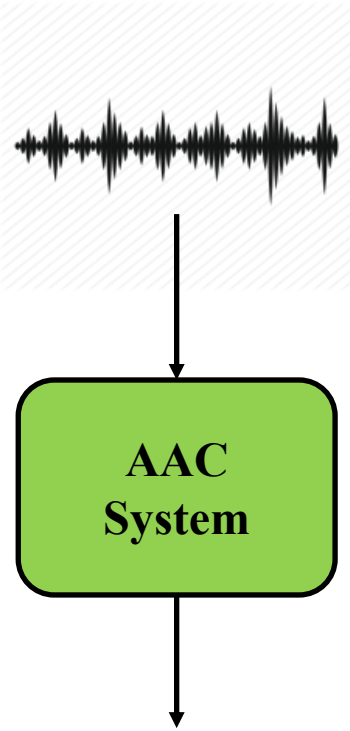
Open-Source | Open-Access | Proprietary | Fine-tuned

		Sound		Music		Speech		Avg	
Name	Size	Test-mini	Test	Test-mini	Test	Test-mini	Test	Test-mini	Test
Audio-Thinker	8.4B	81.98	78.8	74.25	73.8	76.88	75.16	77.7	75.98
Nova 2 Omni	-	81.08	77.87	70.36	66.37	81.98	81.82	77.8	75.28
Step-Audio-2	-	84.04	80.60	73.56	68.23	75.15	72.75	77.58	73.86
MiMo-Audio	7B	81.68	77.2	74.25	69.73	68.17	70.77	74.7	72.59
Audio Flamingo 3	8.2B	79.58	75.83	73.95	74.47	66.37	66.97	73.30	72.42
Qwen2.5-Omni	8.2B	78.10	76.77	65.90	67.33	70.60	68.90	71.50	71.00
Step-Audio-2-mini	8.3B	79.30	75.57	68.44	66.85	68.16	66.49	72.73	70.23
Gemini 2.5 Pro	-	75.08	70.63	68.26	64.77	71.47	72.67	71.60	69.36
Gemini 2.5 Flash	-	73.27	69.50	65.57	69.40	76.58	68.27	71.80	67.39
Gemini 2.0 Flash	-	71.17	68.93	65.27	59.30	75.08	72.87	70.50	67.03
DeSTA2.5-Audio	8B	70.27	66.83	56.29	57.10	71.47	71.94	66.00	65.21
Kimi-Audio	8.2B	75.68	70.70	66.77	65.93	62.16	56.57	68.20	64.40
Audio Reasoner	8.2B	67.87	67.27	69.16	61.53	66.07	62.53	67.70	63.78

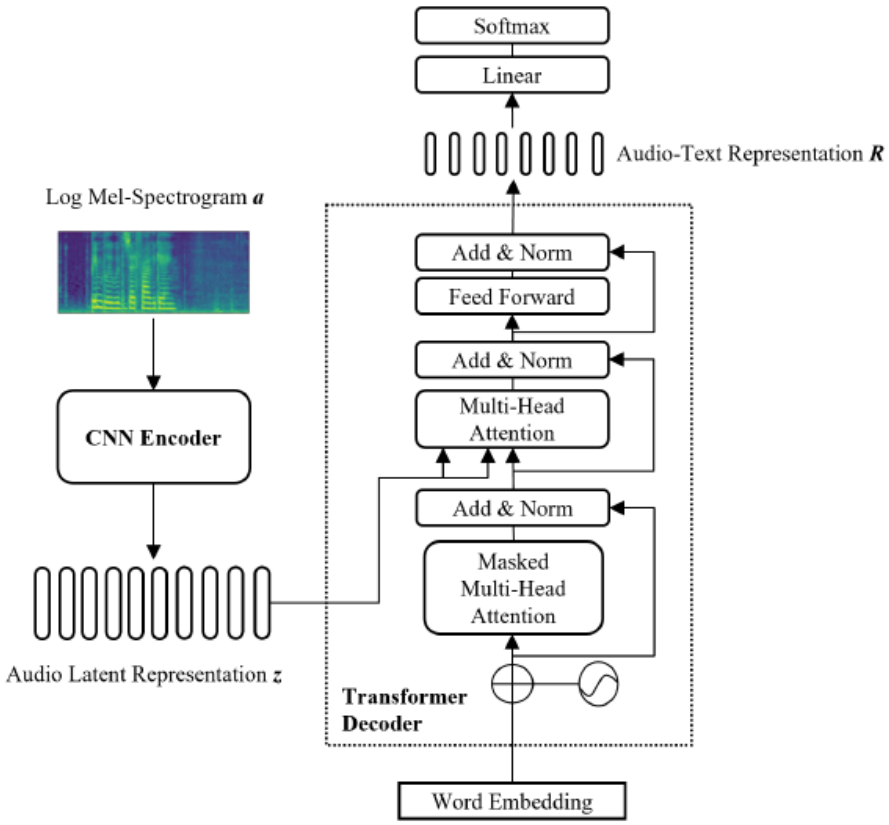
https://sakshi113.github.io/mmau_homepage/#leaderboard-v15-parsed

Application 1 - Audio to Text Generation

- **Automated audio captioning** (AAC) is a cross-modal translation task which aims at generating a natural language description for a given audio clip.
- This task requires understanding the audio content, such the spatial-temporal relationships between events and scenes, and describing such information using natural language.
- Applications
  - Audio retrieval
  - Assist hearing-impared to understand environmental sounds
  - Subtitle for sounds in TV programs



A Typical Model



Common challenges in automated audio captioning:

- Data scarcity
- Representations of audio, text and audio-text
- Diversity of captions
- Multi-lingual captioning
- Interactions with other modalities (e.g. vision)
- ….

Y. Hou, Q. Ren, A. Mitchell, W. Wang, J. Kang, T. Belpaeme, and D. Botteldooren, "Soundscape captioning using sound affective quality network and large language model," IEEE Transactions on Multimedia, 2026.

Y. Zhu, Y. Zhang, L. Xiao, W. Wang, and A. Men, "Zero-shot diverse audio captioning with diffusion models", Knowledge Based Systems, 2026.

Y. Zhang, X. Xu, R. Du, H. Liu, Y. Dong, Z.-H. Tan, W. Wang, and Z. Ma, "Zero-shot audio captioning using soft and hard prompts," IEEE Transactions on Audio Speech and Language Processing, vol. 33, pp. 2045 - 2058, May 2025.

Y. Zhang, H. Yu, R. Du, Z.-H. Tan, W. Wang, Z. Ma, Y. Dong, "ACTUAL: audio captioning with caption feature space regularization," IEEE/ACM Transactions on Audio Speech and Language Processing, vol. 31, pp. 2643 - 2657, 2023.

X. Mei, X. Liu, M. Plumbley, and W. Wang, "Automated audio captioning: an overview of recent progress and new challenges", EURASIP Journal on Audio Speech and Music Processing, 2022.

Audio Captioning Demo

Application 2 - Text to Audio Generation

Potential Applications:

- **Computational "foley artist"**
  - Game developer: e.g., A ghost is haunting a house.
  - Audio producer: e.g., high heels hitting metal ground.
  - Movie producer: e.g., the laser sound from a laser gun.
  - …
- **Automatic content creation**
  - Endless music
  - Audiobook with ambient noises
  - White noise for meditation
  - …
- **Data Augmentations**

Related Works:

Label-to-Audio Generation

- Acoustic Scene (Kong et al., 2019), Sound event (Liu et al., 2019), FootStep (Comunit et al. 2019), …

Text-to-Audio Generation

- DiffSound (Yang et al., 2022), AudioGen (Kreuk et al., 2022), Make-an-Audio (Huang et al., 2023)

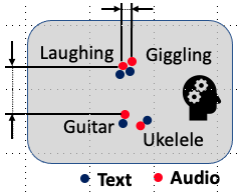
Text-to-Music Generation

- MusicLM (Andrea et al., 2023), Moûsai (Flavio et al., 2023), Noise2Music (Huang et al., 2023)

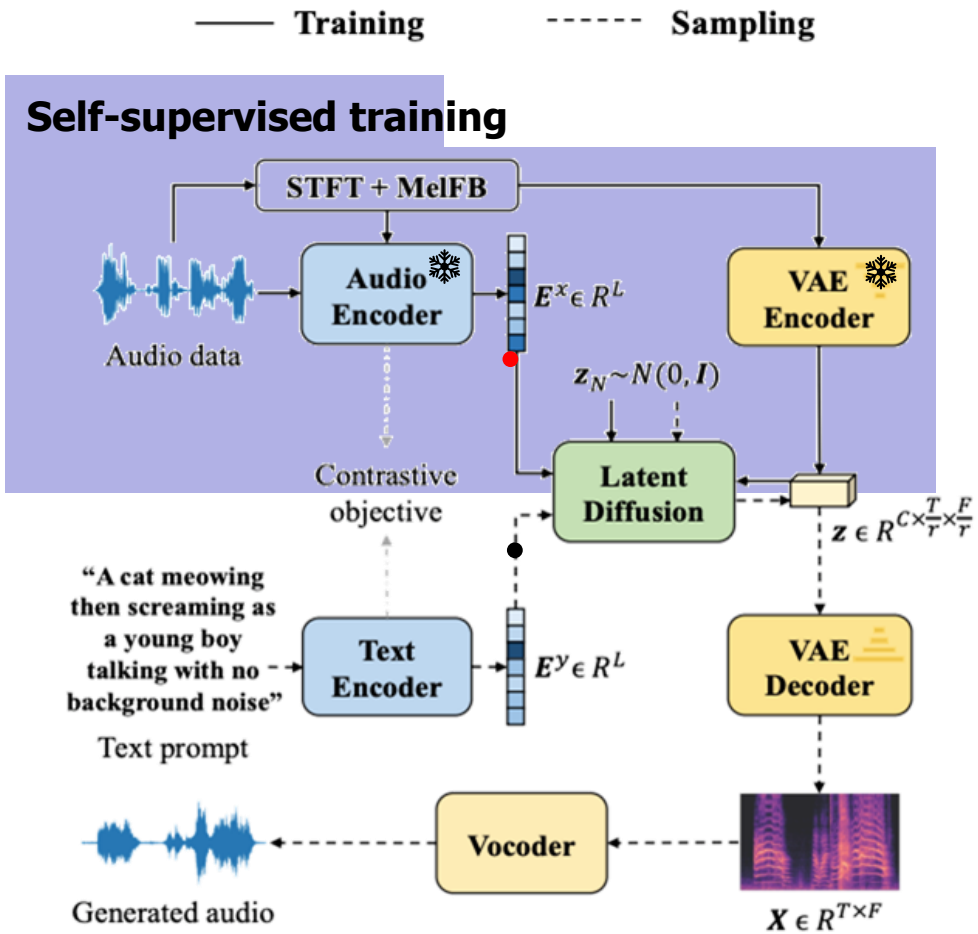
Others

- JukeBox (Dhariwal et al., 2020), AudioLM (Borsos et al., 2022), SingSong (Donahue et al., 2023),…

Text to Audio Generation: **AudioLDM**

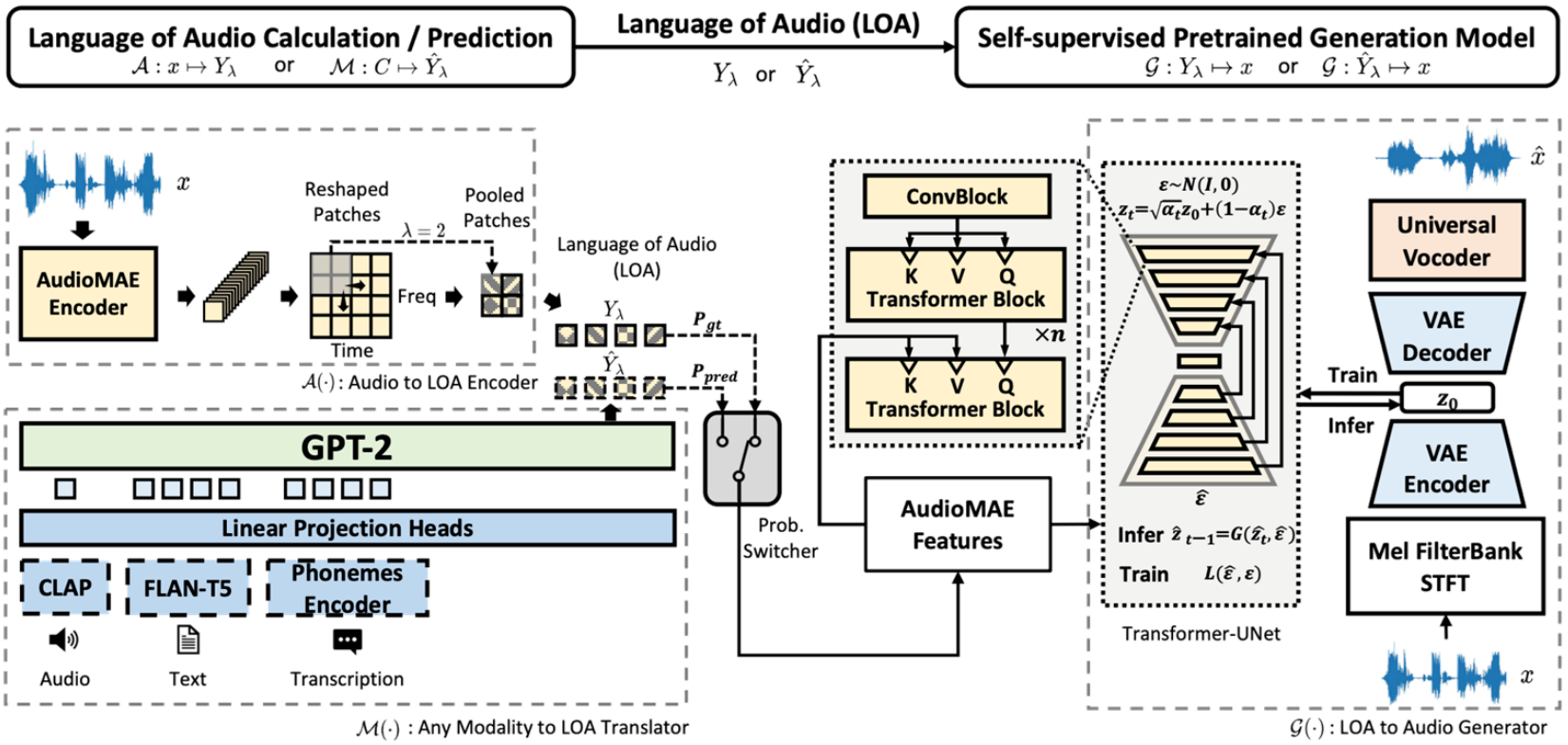


1. **Contrastive Language-Audio Learning (CLAP) Encoders**
   - Align audio and text in one space.
2. **Latent Diffusion Models**
   - Learn to generate VAE latent conditioned on CLAP embedding
3. **Mel-spectrogram Autoencoder**
   - Learn latent representations.
4. **Mel-to-Waveform Vocoder**
   - Reverse Mel back to waveform



H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, M. D. Plumbley, "AudioLDM: text-to-audio generation with latent diffusion models," in Proc. IEEE International Conference on Machine Learning (ICML 2023), Hawaii, USA, 23-29 July, 2023.

Text to Audio Generation: **AudioLDM 2**



H. Liu , Y. Yuan , X. Liu , X. Mei , Q. Kong , Q. Tian , Y. Wang , W. Wang, Y. Wang , M. D. Plumbley, "AudioLDM 2: Learning holistic audio generation with self-supervised pretraining," IEEE/ACM Transactions on Audio Speech and Language Processing, vol. 32, pp. 2871-2883, 2024. [PDF] [code]

AudioLDM 2 - Demo

We generated a total of 350 audio files with prompts (generated by ChatGPT) without cherry-picking.

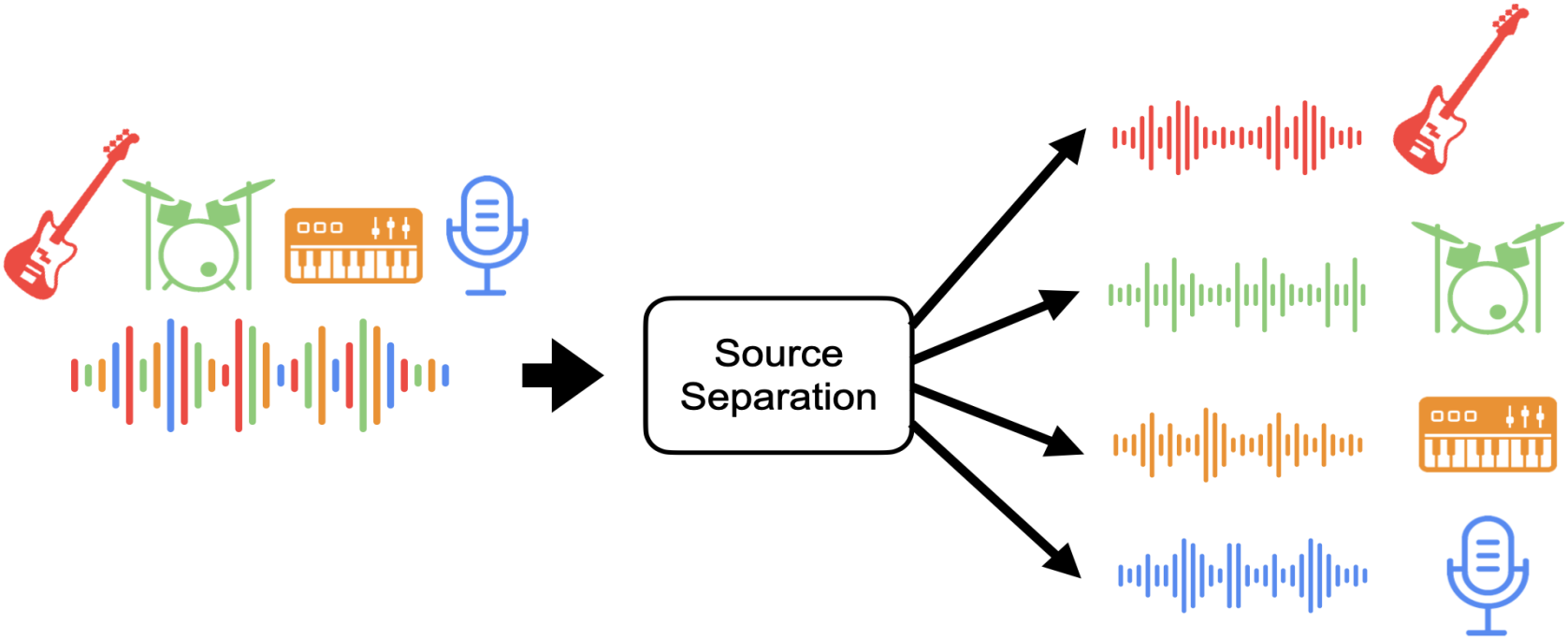
Codes and more demos: https://audioldm.github.io/audioldm2/

Outline

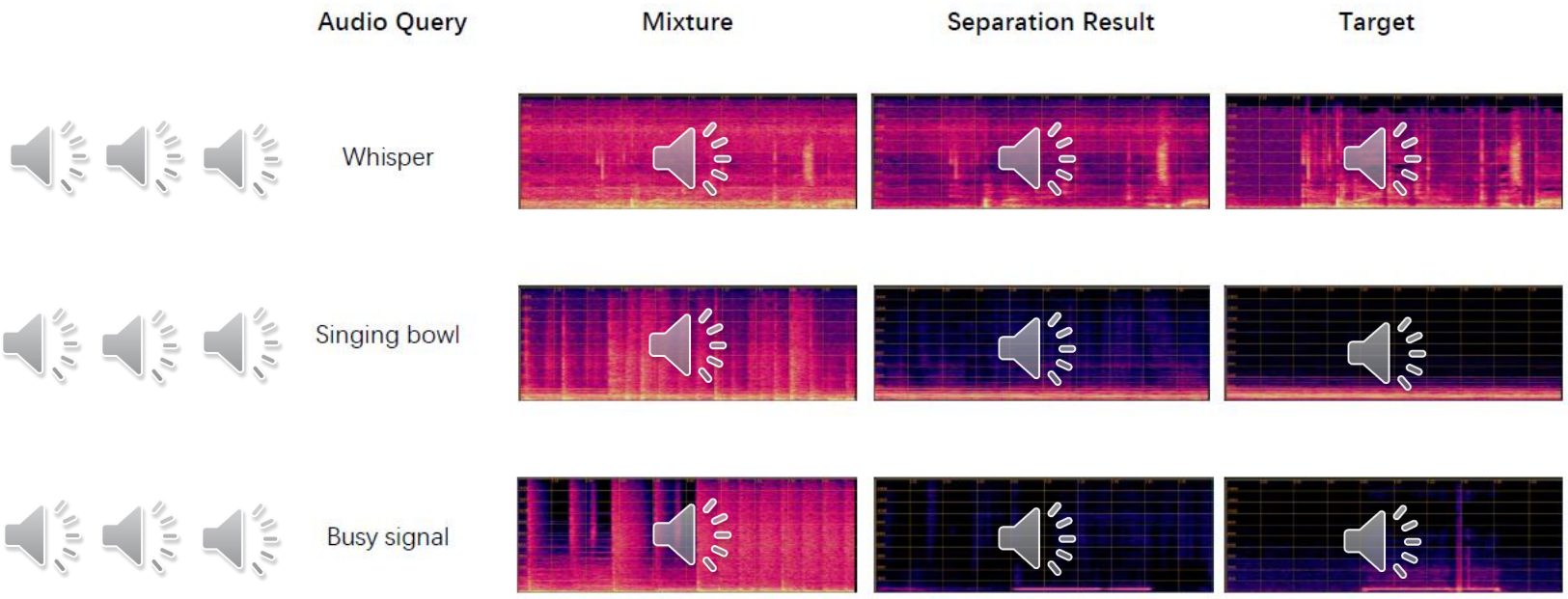
- **A brief history about audio source separation**
  - Cocktail party problem & speech source separation
  - Example problems and methods
  - Recent developments
- **A brief history about (large) audio-language models**
  - Large audio models
  - Large language models
  - Large audio-language models (examples, datasets, applications)
- **Language queried audio source separation**
  - Early models
  - AudioSep
  - FlowSep & FlowSep2
  - A related problem: language guided audio editing
- **Conclusions and future works**

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 the embedding from audio query to extract the sound of interest

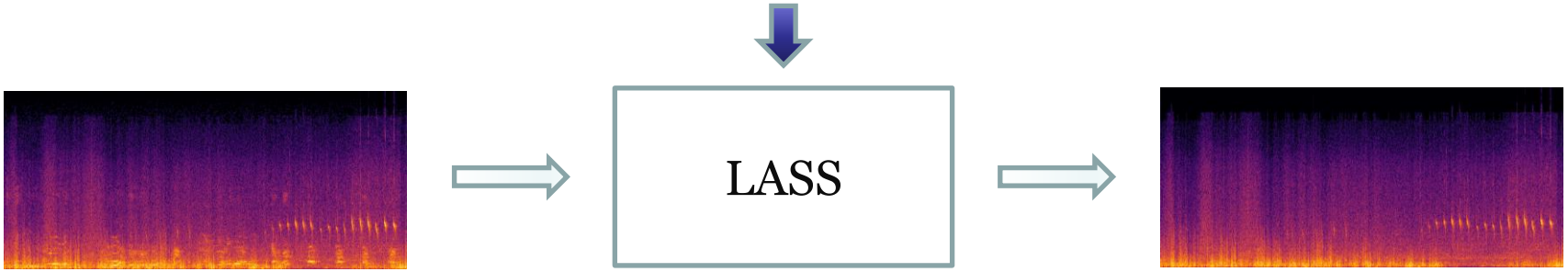
J. Zhao, X. Liu, J. Zhao, Y. Yuan, Q. Kong, M. Plumbley, and W. Wang, "Universal sound separation with self-supervised audio masked autoencoder," in Proceedings of the 32nd European Signal Processing Conference (EUSIPCO 2024), Lyon, France, August 26-30, 2024.

Language Queried Audio Source Separation

UNIVERSITY OF SURREY

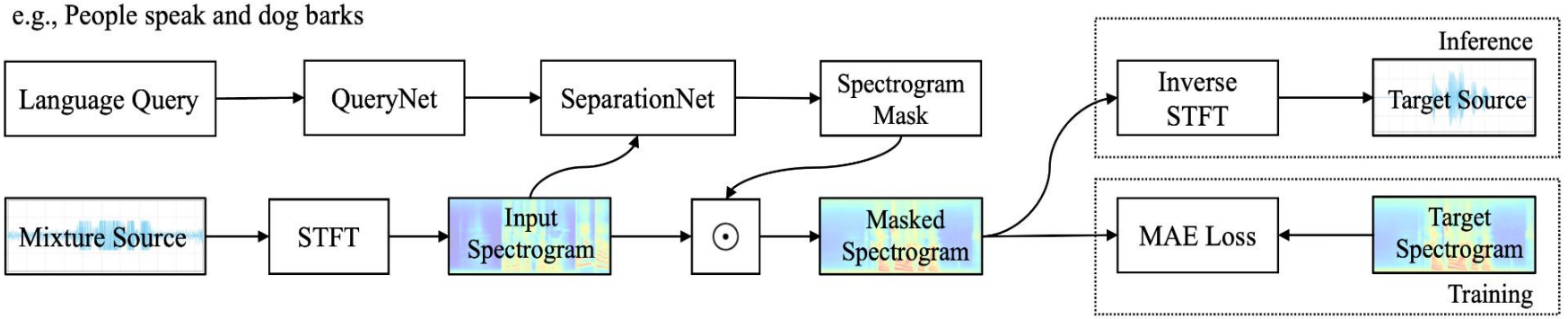
- LASS – Separate a **target source** from an audio mixture based on the **natural language descriptions** of the target source
- First attempt bridging audio source separation and natural language processing
- Support input arbitrary text to separate desired sound sources

Language Query: "A bird is chirping under the thunder storm"



Challenges and Methods

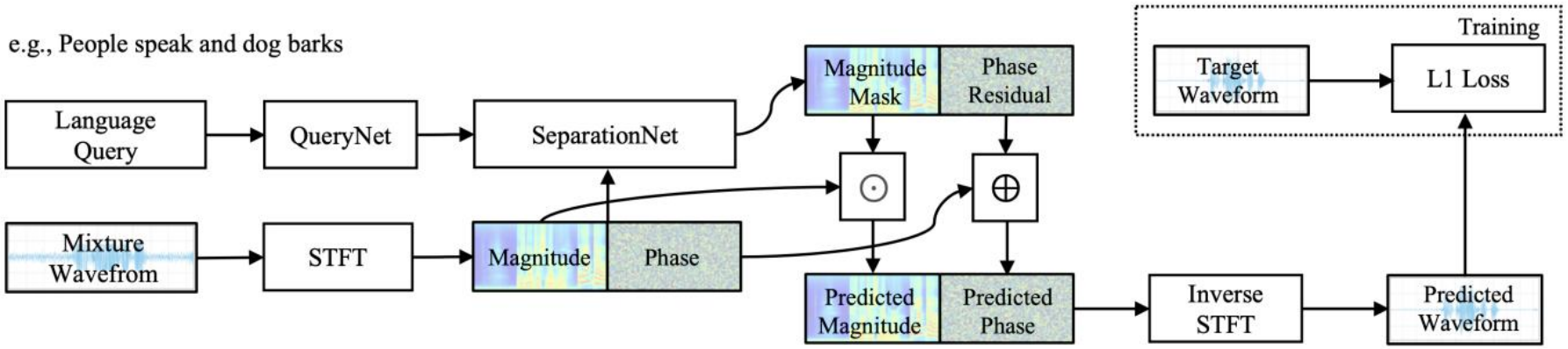
- **How to construct training data?**
  - Training with synthetic mixtures generating from audio-text datasets
- **LASS-Net**
  - QueryNet (BERT) + SeparationNet (ResUNet) + FiLM fusion
  - Trained with 17 hours of data from AudioCaps dataset



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.

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



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, vol. 33, pp. 458-471, Dec 2024.

Training & Testing Data

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	✓	×	×	403 050	7568

THE EVALUATION BENCHMARK FOR OPEN-DOMAIN SOUND SEPARATION WITH NATURAL LANGUAGE QUERIES.

	Num. mixtures	Sr (Hz)	Query Type
AudioSet	5270	32 k	text label
VGGSound	1000	32 k	text label
AudioCaps	4785	32 k	caption
Clotho v2	5225	32 k	caption
ESC-50	2000	32 k	text label
MUSIC	5004	32 k	text label
DCASE 2024 T9	3000	16 k	caption
Voicebank-DEMAND	824	16 k	text label

Experimental Results

- AudioSep achieved excellent results on multiple datasets.
- Impressive zero-shot separation performance on MUSIC and ESC-50.

DIFFERENCES BETWEEN THE AUDIOSEP AND OTHERS IN TERMS OF THE MODEL SIZE, ARCHITECTURE, TRAINING DATA.

	Training Data (hrs)	Parameters	Architecture
LASSNet [3]	17	63.4 M	BERT + ResUNet
CLIPSep [23]	550	181.6 M	CLIP + UNet
USS-ResNet30 [15]	5800	121 M	CNN + ResUNet
AudioSep	14 100	238.6 M	CLAP + ResUNet

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

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

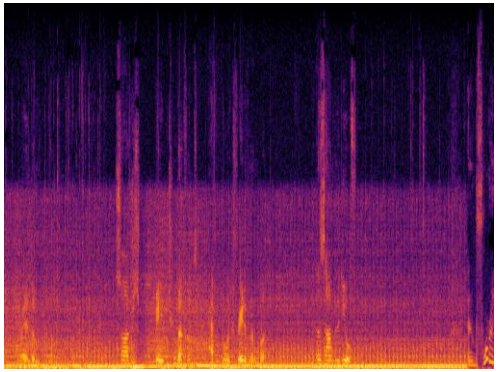
Code: https://github.com/Audio-AGI/AudioSep

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

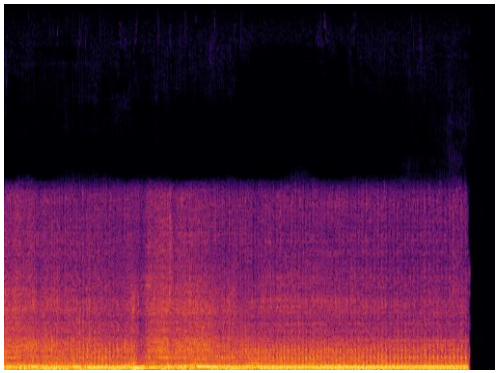
AudioSep: Sound Demos

Human query: “The engine sound of a vehicle”

Mix

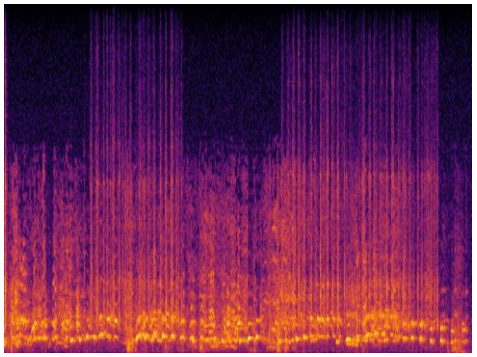


Separated

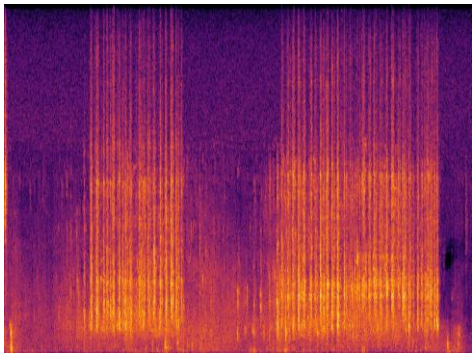


Human query: “The sound of hitting the keyboard”

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:

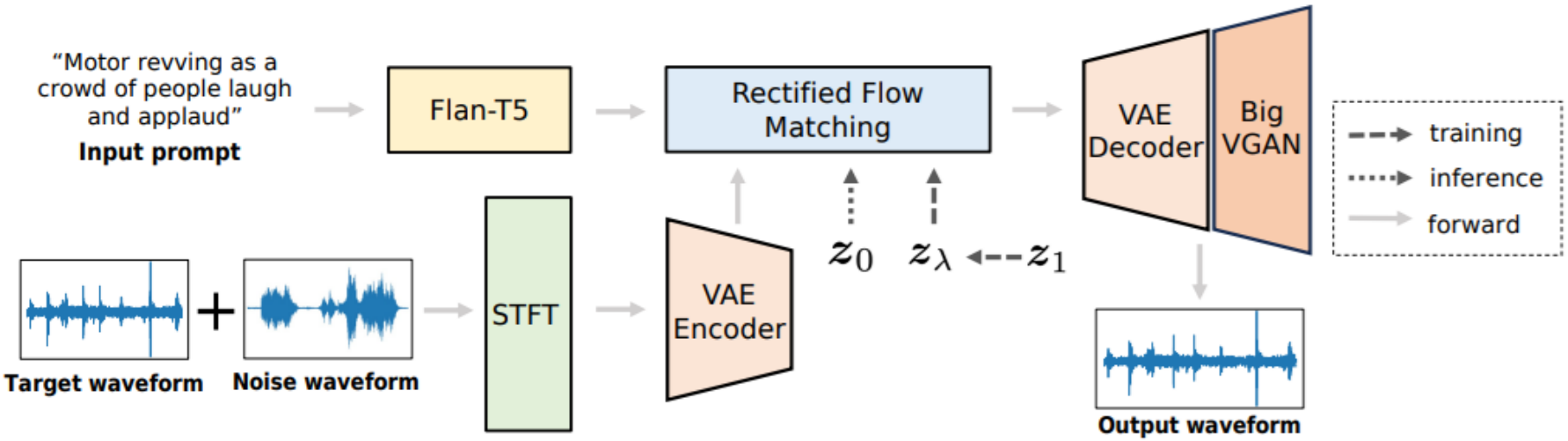
Using the generative approaches.

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.



$$v = \boldsymbol{z}_1 - (1-\sigma)\boldsymbol{z}_0$$

$$L_{\mathrm{RFM}}(\theta) = \mathbb{E}_{\lambda,\boldsymbol{z}_1,\boldsymbol{z}_0} \left\| \mu(\boldsymbol{z}_\lambda, \boldsymbol{E}, \boldsymbol{z}^m) - v \right\|^2$$

Y. Yuan, X. Liu, H. Liu, M.D. Plumbley, and W. Wang,"FlowSep: Language-queried sound separation with rectified flow matching" in ICASSP 2025.

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 CLAPScoreA from generative tasks to evaluate the performance.

OBJECTIVE EVALUATION ON LASS, WHERE AC, VGG AND ESC ARE SHORT FOR AUDIOCAPS [20], VGGSOUND [21] AND ESC50 [30] RESPECTIVELY.FAD DOES NOT APPLY ON UNPROCESSED DATA AS IT CALCULATE THE DIFFERENCE BETWEEN TWO GROUPS OF AUDIO, WHILE THE TARGET AND MIXED AUDIO SHARE THE SAME AUDIO EVENTS.

Model	FAD ↓				CLAPScore ↑					CLAPScore$_A$ ↑			
	AC	DE-S	VGG	ESC	AC	DE-S	DE-R	VGG	ESC	AC	DE-S	VGG	ESC
Unprocessed	–	–	–	–	11.9	23.2	22.7	13.6	19.1	64.9	71.3	66.7	71.3
LASS-Net [7]	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 [13]	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.
- Ablation studies between RFM and traditional diffusion-based models.

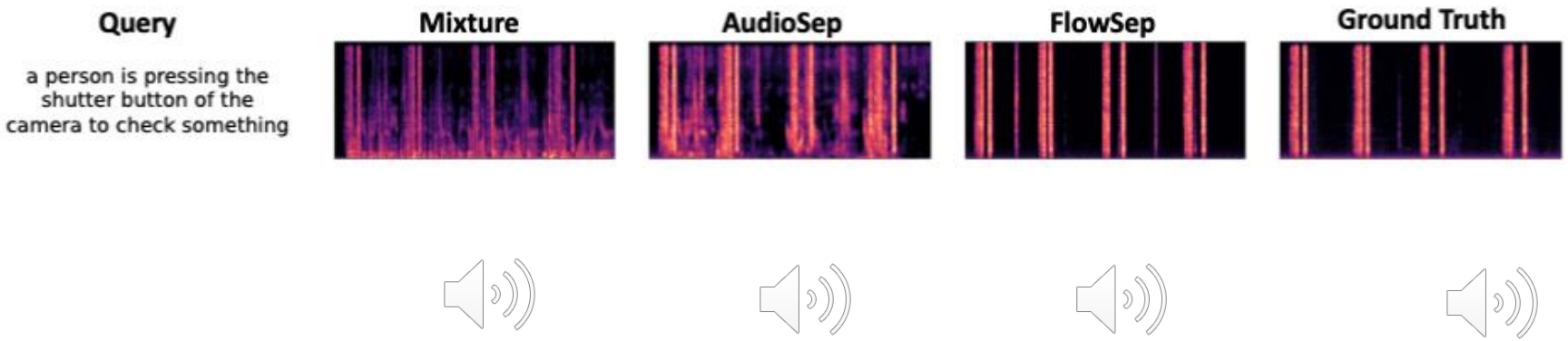
TABLE II
SUBJECTIVE EVALUATION RESULTS ON LASS, WHERE AC, DE-S AND DE-R ARE SHORT FOR AUDIOCAPS, DCASE-SYNTH AND DCASE-REAL RESPECTIVELY.

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**

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

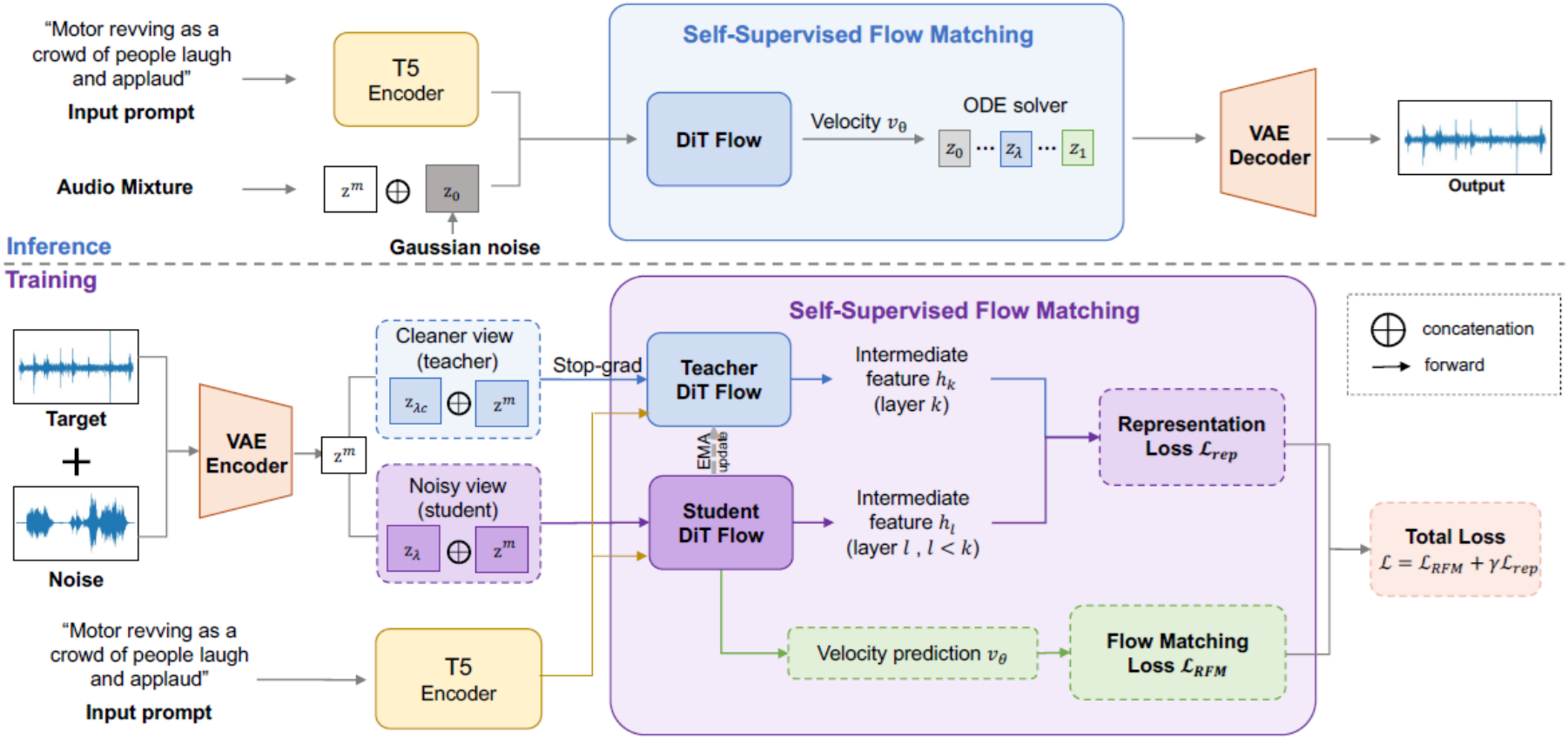
FlowSep – Sound Demos



- 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-queried sound separation with rectified flow matching" in ICASSP 2025.

FlowSep 2



Y. Yuan, X. Liu, H. Liu, X. Kang, M. D. Plumbley, W. Wang, "FlowSep 2: Self-supervised flow matching for language-queried audio source separation", submitted to IEEE Transactions on Audio Speech and Language Processing, August 2026. **https://arxiv.org/pdf/2608.22111**

FlowSep 2 – Training and Testing Datasets

UNIVERSITY OF SURREY

SUMMARY OF DATASETS USED FOR TRAINING AND EVALUATION IN FLOWSEP2.

Split	Dataset	Num.	Hours	Target Type
Train	VGGSound	183,727	510	Label
	AudioCaps	49,837	140	Caption
	AudioSetCaps	1,400,000	3,900	Caption
	WavCaps	400,000	1,100	Caption
Test	VGGSound	2,000	5.6	Label
	ESC-50	2,000	5.6	Label
	MUSDB	192	0.93	Label
	AudioCaps	928	2.6	Caption
	DCASE-Synth	3,000	8.3	Caption
	DCASE-Real	100	0.3	Caption

EXPERIMENTAL SETUPS AND MODEL SIZES, SF DENOTES SELF-FLOW AND RFM IS SHORT FOR RECTIFIED FLOW-MATCHING

Model	Training-Objectives	Encoder	Backbone	Params. (M)
AudioSep	Masking-Based	STFT	UNet	238.6
SAM-Audio	RFM	VAE	DiT	2,934
FlowSep	RFM	VAE	UNet	265.5
FlowSep2-UNet	RFM	VAE	UNet	352.1
FlowSep2-DDPM	DDPM		DiT	559.5
FlowSep2-RFM	RFM	VAE	DiT	558.2
FlowSep2-RAE		RAE		685.8
FlowSep2-REPA		VAE		605.3
FlowSep2	SF	VAE	DiT	598.9

FlowSep 2 – Key Results

UNIVERSITY OF SURREY

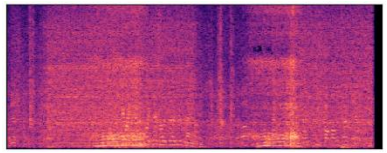
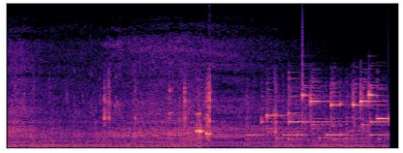
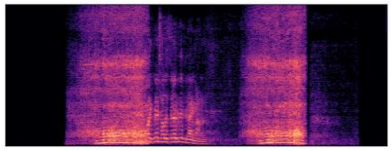
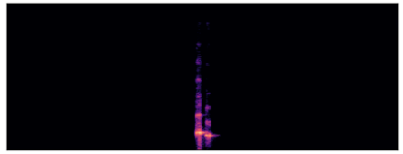
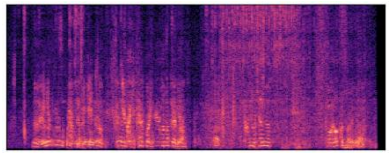
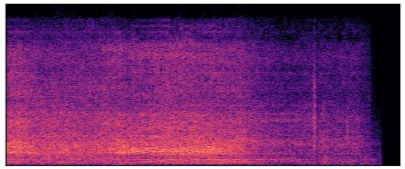
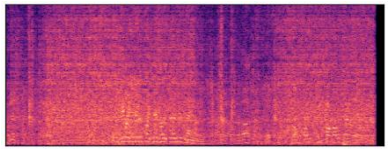
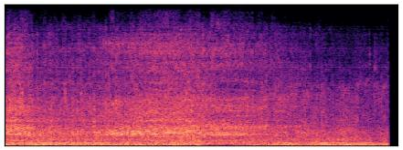
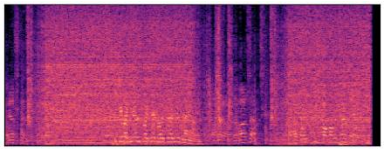
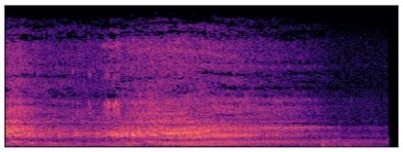
OBJECTIVE EVALUATION ON LASS, WHERE AC, VGG, ESC, MUC ARE SHORT FOR AUDIOCAPS [63], VGGSOUND [62], ESC-50 [65], AND MUSDB18 [66] RESPECTIVELY. FAD CANNOT BE APPLIED TO UNPROCESSED DATA AS IT CALCULATES THE DIFFERENCE BETWEEN TWO GROUPS OF AUDIO, WHILE THE TARGET AND MIXED AUDIO SHARE THE SAME AUDIO EVENTS.

Model	FAD ↓				CLAP Score ↑					CLAP$_A$ Score ↑				
	AC	DE-S	VGG	ESC	AC	DE-S	DE-R	VGG	ESC	AC	DE-S	VGG	ESC	MUC
Unprocessed	–	–	–	–	31.7	42.7	41.7	33.6	38.6	54.4	66.3	62.7	66.8	50.2
LASS-Net [14]	5.09	1.83	3.09	3.28	33.9	44.4	44.8	36.4	40.5	65.2	72.1	64.5	75.6	–
AudioSep [20]	4.38	1.21	2.30	1.93	33.6	45.1	49.7	38.5	40.2	65.1	74.4	67.4	76.0	50.4
FlowSep [32]	2.86	0.90	2.06	1.49	41.9	46.4	50.3	39.5	42.2	76.7	75.6	69.2	75.7	54.1
SAM-Audio [48]	1.58	0.98	1.86	1.20	37.8	43.4	45.8	39.1	40.9	64.1	71.3	70.5	78.6	**62.8**
FlowSep2-M	**0.88**	**0.72**	**1.32**	**1.15**	**44.9**	**47.5**	**51.0**	**42.0**	**44.5**	**80.5**	**78.5**	**72.0**	**80.6**	58.5

SUBJECTIVE AND HUMAN-ALIGNED EVALUATION RESULTS ON LASS, WHERE AC, DE-S AND DE-R DENOTE AUDIOCAPS [63], DCASE-SYNTH AND DCASE-REAL RESPECTIVELY. SAJ (SAM AUDIO JUDGE), AA (AUDIOBOX AESTHETICS), REL AND OVL ARE EVALUATED USING A 1–5 LIKERT SCALE. ↑ INDICATES THAT HIGHER SCORES ARE BETTER.

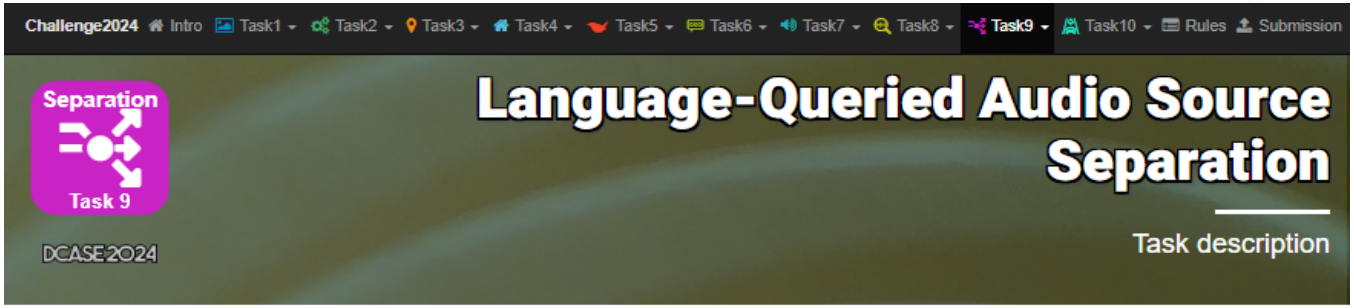
Model	SAJ ↑			AA ↑				REL ↑				OVL ↑			
	AC	MUC	DE-S	AC	MUC	DE-S	DE-R	AC	MUC	DE-S	DE-R	AC	MUC	DE-S	DE-R
LASS-Net [14]	2.57	2.03	2.68	4.86	5.27	5.41	4.89	3.12	2.87	2.96	3.59	2.16	2.45	2.84	3.88
AudioSep [20]	2.84	2.97	2.93	5.18	5.94	6.03	5.57	3.66	3.99	3.24	3.93	2.69	3.85	3.53	4.02
SAM-Audio [48]	2.58	**3.76**	2.91	**5.56**	6.48	**6.14**	5.79	3.54	**4.35**	3.24	3.85	2.69	**4.44**	3.75	4.26
FlowSep [32]	2.73	3.08	3.02	5.36	6.07	5.83	5.28	4.08	3.71	**3.62**	4.11	3.98	3.84	3.72	4.26
FlowSep2-M	**2.96**	3.39	**3.47**	5.49	**6.67**	6.08	**5.86**	**4.13**	3.91	**3.62**	**4.16**	**4.09**	4.02	**3.88**	**4.28**

FlowSep 2 - Demos

Prompt	*"A couple of men speaking as metal clanks and a power tool operates"*	*"An aircraft engine is taking off"*
Mixture		
SAM Audio		
FlowSep		
FlowSep2		
Ground Truth		

https://audio-agi.github.io/Flowsep2-demo/

DCASE Challenge –Task 10



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.

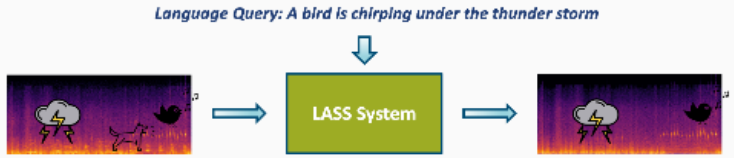
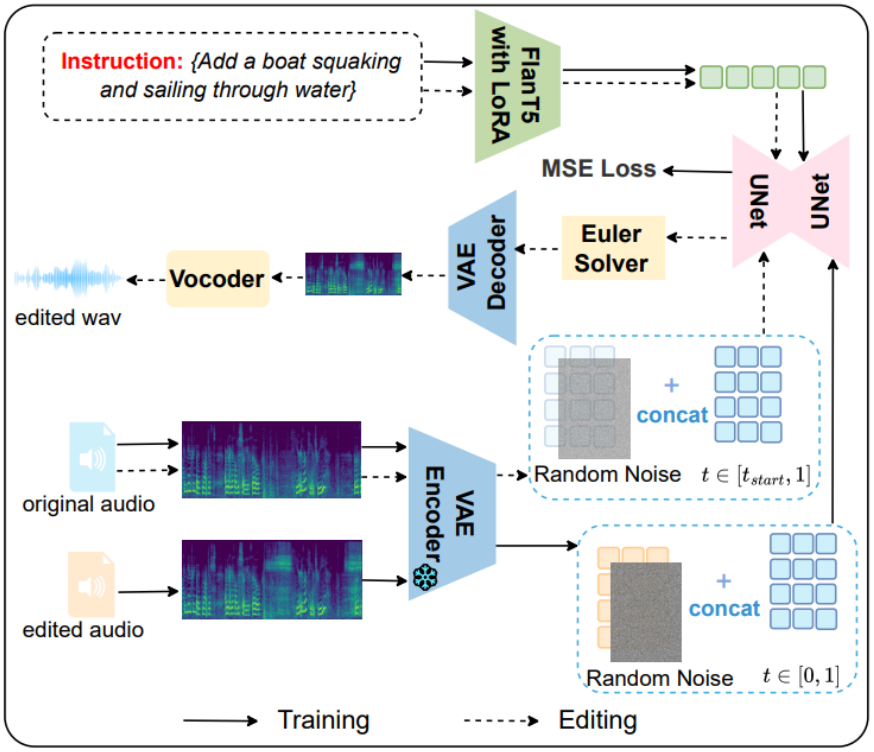


Figure 1: Overview of LASS system.

Audio dataset

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

Controllable Audio Editing:
RFM-EDITING & **RFM-EDITING2**



Demos here:
https://katelin-glt.github.io/RFM-Editing-Demo/

Remove continuous frying noises:

Original:

Edited:

Replace someone suddenly sneezes out loud with several pigeons cooing:

Original:

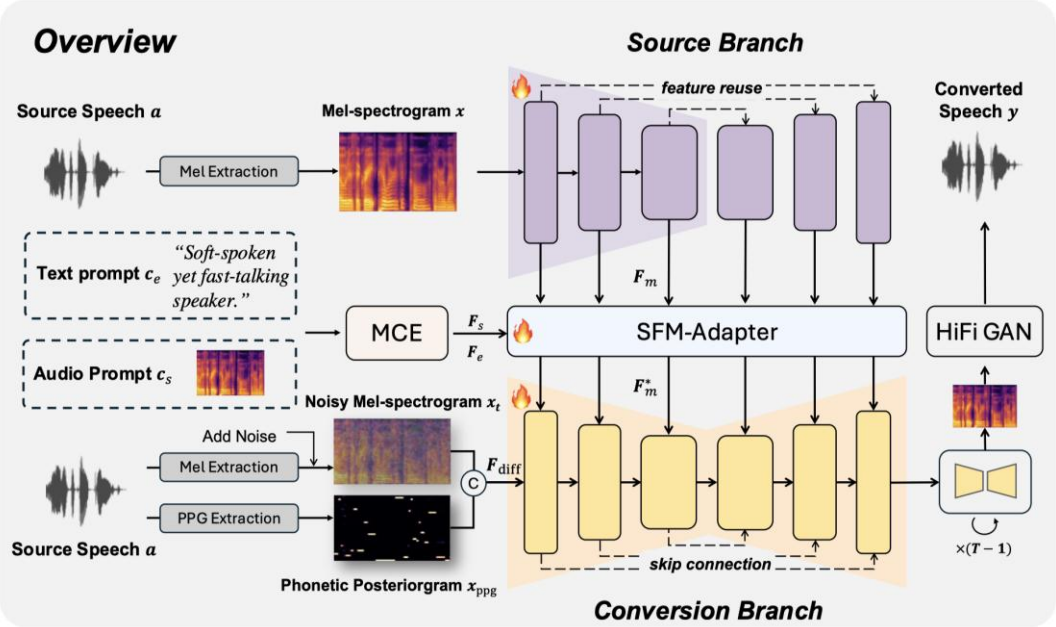
Edited:

- L. Gao, Y. Yuan, Y. Chen, Y. Cheng, Z. Li, J. Wen, S. Zhang, and W. Wang, "RFM-EDITING: Rectified flow matching for text-guided audio editing," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2026), Barcelona, Spain, May 4-8, 2026. [PDF]
- L. Gao, Y. Zhu, Y. Chen, D. Wang, S. Zhang, Z. Li, J.Y. Guillemaut, W. Wang, "RFM-Editing 2: Text-guided audio editing with rectified flow matching and coarse-to-fine diffusion transformers," submitted to IEEE Trans. On Audio Speech and Language Processing, July 2026. https://arxiv.org/abs/2606.20101v3

Controllable Speech Editing: **SFM-Adapter**

The goal of **Speech Style Edit:** ***Change How Speech Sounds, Not What It Says !***

Main framework: a reconstruction-based diffusion architecture that can incorporate external style conditions (*e.g., natural language and audio*).



Text Prompt

Speaking with intention, her unhappy voice held a high pitch and low energy.

A fatigued disconsolate male voice, characterized by a slow speaking speed and a deep, low pitch, emits a subtle energy, resulting in a soothing audio experience that exudes tranquility.

Y. Chen, H. Liu, Q. Chen, A. Singh, J. Zhao, W. Wang, P. J. B. Jackson, M. D. Plumbley, "SFM-Adapter: Style-aware feature manipulation adapter for speech style editing", IEEE Transactions on Audio Speech and Language Processing, 2026. [PDF] [code, data, & demo], vol. 34, pp. 3132-3145, 2026.

Summary & Future Directions

Summary:

- Language queried audio source separation offers **helpful** and **flexible** tools for users to control what sound to be separated from the sound mixtures, using **natural language-based queries**.
- FlowSep & FlowSep2 offers better separation results than AudioSep, meaning generative models perform better as compared with time-frequency masks-based models.
- Rectified flow matching based loss works better than diffusion loss. Self-supervised flow representation further improves the results of flow matching.

Future directions:

- Explore advanced reasoning capabilities of LALMs for LASS (e.g., separating complex queries like "the second sound" or "annoying sounds").
- Further scale the datasets and improve diversities of datasets.
- Improve generalisation to more and unseen sound classes.
- Apply self-supervised techniques (e.g., MixIT) for pre-training to enhance separation performance.
- Include additional modalities (e.g. visual, EEG).

Take Away: Papers, Codes, Data, Demos

UNIVERSITY OF SURREY

AudioLDM:
Paper: https://arxiv.org/abs/2301.12503
Project Page: https://audioldm.github.io/
Pretrained model: https://github.com/haoheliu/AudioLDM
Evaluation tools: https://github.com/haoheliu/audioldm_eval
YouTube: https://www.youtube.com/watch?v=_0VTltNYhao

SemantiCodec:
Paper/code/demos at project page:
https://haoheliu.github.io/SemantiCodec/

AudioSep & FlowSep & FlowSep2:
https://github.com/Audio-AGI/AudioSep
https://audio-agi.github.io/FlowSep_demo/
https://audio-agi.github.io/Flowsep2-demo/

RFM-EDITING & RFM-EDITING2:
https://katelin-glt.github.io/RFM-Editing-Demo/

WavCaps:
Paper: https://arxiv.org/abs/2303.17395
Code: https://github.com/xinhaomei/wavcaps

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

More code about other works avaliable at:
https://github.com/XinhaoMei/DCASE2021_task6_v2
https://github.com/XinhaoMei/ACT
https://github.com/liuxubo717/cl4ac

AudioLDM2:
Project Page: https://audioldm.github.io/audioldm2/

APT:
Code: https://github.com/JinhuaLiang/APT

WavCraft:
Code: https://github.com/JinhuaLiang/WavCraft

WavJourney:
Paper: https://arxiv.org/abs/2307.14335
Code: https://github.com/Audio-AGI/WavJourney
Demo: https://huggingface.co/spaces/Audio-AGI/WavJourney

AudioSetCaps:
Data and code: https://github.com/JishengBai/AudioSetCaps
https://huggingface.co/datasets/baijs/AudioSetCaps

SFM-Adapter: https://ychenn1.github.io/SFM-Adapter/

Sound-VECaps: paper, data & code:
https://yyua8222.github.io/Sound-VECaps-demo

GCDance: paper, data & code:
https://xinranliu7715.github.io/gcdance/