

Industrial-Academic Joint Workshop on Emerging Problems and Methods in Audio, Speech and Language Processing

Date

Monday, 1st September, 2025, Istanbul, Turkey

Agenda

Talks (16:00-17:10)

Each talk: 11 minutes presentation + 3 minutes Q&A

1. **Mengyao Zhu** (Huawei, China): Challenges and Requirements in the field of Audio from Huawei
2. **Zheng-Hua Tan** (Aalborg University, Denmark): Emerging Sequence Models for Audio Representation Learning and Speech Enhancement
3. **Jinhua Liang** (Queen Mary University of London, UK):
4. **Wenwu Wang** (University of Surrey, UK): Text-Queried Audio Source Separation
5. **Cem Subakan** (Laval University/Mila-Quebec AI Institute, Canada): Producing Listenable Explanations for Audio Models

Panel Discussion (17:10-18:00)

Panel Members:

Zheng-Hua Tan, Aalborg University, Denmark

Paris Smaragdis, MIT, USA

Cem Subakan, Laval University/Mila-Quebec AI Institute, Canada

Mengyao Zhu, Huawei, China

Wenwu Wang, University of Surrey, UK

Talk Details

Talk 1:

Title:

Challenges and Requirements in the field of Audio from Huawei

Abstract:

Introduction of the Consumer Business Group in Huawei and also the Audio Dept., then the challenges and our requirements in the field of Audio from Huawei CBG. Finally, some student technology competitions co-organized by Huawei in China will be showcased.

Speaker Bio:

Mengyao Zhu received the B.S. and Ph.D. degrees in communication and information system from Zhejiang University, Hangzhou, China, in 2004 and 2009, respectively. Since 2019, he has been a Technical Expert with Audio Department, Huawei CBG on sabbatical leave from Shanghai University, Shanghai, China. He is currently in charge the Spatial audio in Huawei. His research interests include sound field capture and reproduction, audio and speech signal processing, and circuits and system design of multimedia systems. In 2020 and 2021, he was the TPC Co-Chair of CSMT (Conference on Sound and Music Technology). In 2024, he was Vice-Director of Committee on Sound and Music Technology in China Audio Industry Association, and In 2025, he was Vice-Chair of Audio Standard of UHD World Association.

Photo:



Talk 2:

Title:

Emerging Sequence Models for Audio Representation Learning and Speech Enhancement

Abstract:

While Transformer architectures have played a central role in audio and speech modeling, their quadratic complexity and limited scalability have driven the development for more efficient alternatives. Among these, Mamba and xLSTM stand out for their linear scalability and ability to

model long-range dependencies effectively. In this talk, we present our recent work leveraging these architectures to learn general-purpose audio representations from masked spectrogram patches in a self-supervised manner. Both models consistently outperform Transformer-based baselines across ten diverse downstream tasks. Additionally, we explore their applications to speech enhancement, introducing a hybrid architecture that combines Mamba with multi-head attention mechanisms. This approach achieves superior generalization performance on challenging out-of-domain datasets. Our findings demonstrate the potential of these emerging sequence models to advance the state of the art in audio representation learning and speech enhancement.

Speaker Bio:

Zheng-Hua Tan is currently a Professor in the Department of Electronic Systems and a Co-Head of the Centre for Acoustic Signal Processing Research at Aalborg University, Aalborg, Denmark. He is also a Co-Lead of the Pioneer Centre for AI, Denmark. He was a Visiting Scientist at the Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, USA, an Associate Professor at the Department of Electronic Engineering, SJTU, Shanghai, China, and a postdoctoral fellow at the AI Laboratory, KAIST, Daejeon, Korea. His research interests include machine learning, deep learning, noise-robust speech processing, and multimodal signal processing. He has (co)-authored over 280 refereed publications. His works have been recognized by the prestigious IEEE Signal Processing Society 2022 Best Paper Award and International Speech Communication Association 2022 Best Research Paper Award. He was the elected Chair of the IEEE Signal Processing Society Machine Learning for Signal Processing Technical Committee (MLSP TC) from 2021-2022. He is a Member of Speech and Language Processing TC. He is the Lead Editor for IEEE Journal of Selected Topics in Signal Processing Inaugural Special Series on AI in Signal and Data Science. He served as an Associate Editor for IEEE/ACM Transactions on Audio, Speech and Language Processing, Computer Speech and Language, Digital Signal Processing, and Computers and Electrical Engineering. He is the General Chair for ICASSP 2029 and a TPC Co-Chair for ICASSP 2028. He was a TPC Vice-Chair for ICASSP 2024, the General Chair for IEEE MLSP 2018 and a TPC Co-Chair for IEEE SLT 2016.

Photo:



Talk 3:

Title:

LLMs for Audio Intelligence: From Understanding to Generation

Abstract:

Recent advances in large language models (LLMs) have shown their potential beyond text, enabling new paradigms for reasoning and content creation across modalities. This talk will present our efforts in extending LLMs to audio understanding and generation. It will first introduce our work on Acoustic Prompt Tuning (APT), which adapts LLMs for audio perception tasks. This talk will then discuss WavCraft, an open-source agent for controllable and expressive audio editing and synthesis. Together, these works highlight a unified perspective on how LLMs can be leveraged for audio intelligence, paving the way toward foundational models that can understand, reason about, and generate audio content by following user instructions.

Speaker Bio:

Jinhua Liang is a PhD researcher at Queen Mary University of London, advised by Dr. Emmanouil Benetos, Dr. Huy Phan, and Prof. Mark Sandler. His research focuses on multimodal learning for audio intelligence, with the mission of enabling machines to “hear” real-world sounds by integrating audio signals with knowledge from other modalities, and to “create” audio in a controllable and expressive way. He is an active member of the Detection and Classification of Acoustic Scenes and Events (DCASE) community and co-organized DCASE Task 5, Few-shot Bioacoustic Event Detection, in 2024.

Photo:



Talk 4:

Title:

Language Queried Audio Source Separation

Abstract:

Language-queried audio source separation (LASS) is a paradigm that we proposed recently for separating sound sources of interest from an audio mixture using a natural language query. The development of LASS systems offers intuitive and scalable interface tools that are potentially useful for digital audio applications, such as automated audio editing, remixing, and rendering. In this talk, we will introduce present our two newly developed LASS algorithms, AudioSep and FlowSep. AudioSep is a foundational model for open-domain audio source separation driven by natural language queries. It employs a query network and a separation network to predict time-frequency masks, enabling the extraction of target sounds based on text prompts. The model was trained on large-scale multimodal datasets and evaluated extensively on numerous tasks including audio event separation, musical instrument separation, and speech enhancement. FlowSep is a new generative model for LASS based on rectified flow matching (RFM), which models linear flow trajectories from noise to target source features within the latent space of a variational autoencoder (VAE). During inference, the RFM-generated latent features are used to reconstruct a mel-spectrogram through the

pre-trained VAE decoder, which is then passed to a pre-trained vocoder to synthesize the waveform. After this, we will discuss the datasets and performance metrics we developed for evaluating the LASS systems, and the organisation of Task 8 of DCASE 2024 international challenge, building on the AudioSep model. Finally, we conclude the talk by outlining potential future research directions in this area.

Speaker Bio:

Wenwu Wang is a Professor in Signal Processing and Machine Learning, Associate Head of External Engagement, School of Computer Science and Electronic Engineering, University of Surrey, UK. He is also an AI Fellow at the Surrey Institute for People Centred Artificial Intelligence. His current research interests include signal processing, machine learning and perception, artificial intelligence, machine audition (listening), and statistical anomaly detection. He has (co)-authored over 300 papers in these areas. His work has been recognized with more than 15 accolades, including the 2022 IEEE Signal Processing Society Young Author Best Paper Award, ICAUS 2021 Best Paper Award, DCASE 2020 and 2023 Judge's Award, DCASE 2019 and 2020 Reproducible System Award, and LVA/ICA 2018 Best Student Paper Award. He is a Senior Area Editor (2025-2027) of IEEE Open Journal of Signal Processing and an Associate Editor (2024-2026) for IEEE Transactions on Multimedia. He was a Senior Area Editor (2019-2023) and Associate Editor (2014-2018) for IEEE Transactions on Signal Processing, and an Associate Editor (2020-2025) for IEEE/ACM Transactions on Audio Speech and Language Processing. He is Chair (2025-2027) of the EURASIP Technical Area Committee on Acoustic Speech and Music Signal Processing, an elected Member (2021-2026) of the IEEE SPS Signal Processing Theory and Methods Technical Committee. He was the elected Chair (2023-2024) of IEEE Signal Processing Society (SPS) Machine Learning for Signal Processing Technical Committee, and a Board Member (2023-2024) of IEEE SPS Technical Directions Board. He has been on the organising committee of INTERSPEECH 2022, IEEE ICASSP 2019 & 2024, IEEE MLSP 2013 & 2024, and SSP 2009. He is Technical Program Co-Chair of IEEE MLSP 2025. He has been an invited Keynote or Plenary Speaker on more than 20 international conferences and workshops.

Photo:



Talk 5:

Title:

Producing Listenable Explanations for Audio Models

Abstract:

I will talk about our recent works on producing explanations for Audio Models. Deep Learning Models are good when it comes to getting good performance out of them, but they are typically black-box models. Our goal in this line of work is to develop listenable explanation methods for

black-box audio models, without compromising any performance from our original black-box. We show through several metrics that the produced explanations through our methods remain faithful to the original model and we also show that they are indeed listenable and understandable.

Speaker Bio:

Cem Subakan is an assistant professor in Laval University, Computer Science and Software Engineering Department, an affiliate assistant Professor in Concordia University and an associate academic member in Mila-Québec AI Institute. He completed his PhD (in University of Illinois at Urbana-Champaign (UIUC)), and later did a postdoc in Mila. He has extensive research experience in speech and audio and is the leader of source separation part of the highly popular (>9k stars on GitHub) Speech toolkit SpeechBrain. He is an associate member of IEEE Machine Learning for Signal Processing Technical Committee, and he is general chair of 35th IEEE Machine Learning for Signal Processing conference in 2025. He has published papers in venues such as ICML, NeurIPS, ICASSP, Interspeech, TASL, WASPAA, and MLSP. He won the best student paper in the 2017 version of MLSP conference, and was nominated for a best paper award in 2023 in Interspeech.

Photo:



Challenges in the Audio Field of Huawei

Zhu Mengyao, Huawei

向上捅破天，向下扎到根



Some days **you bloom**, some
days you **grow roots**. Both matter.

- 用户体验 Experience
- 技术 Technology

小红书

小红书号：26766964958

Talk details are omitted due to commercial confidentiality.

Emerging Sequence Models for Audio Representation Learning and Speech Enhancement

@Industrial-Academic Joint Workshop on Emerging Problems and
Methods in Audio, Speech and Language Processing at MLSP 2025

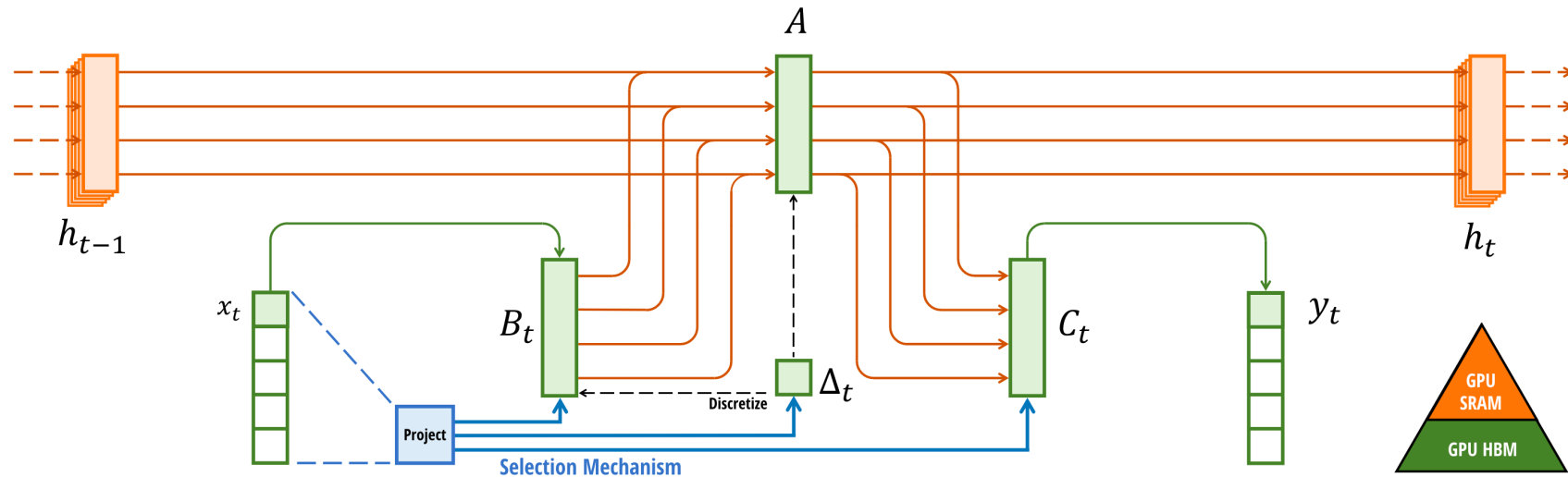
31.08.2025

Prof. Zheng-Hua Tan
Department of Electronic Systems
Aalborg University, Denmark

with Sarthak Yadav and Nikolai Lund Kühne

Mamba (a.k.a. S6: : Selective Structured State Space Sequence models with a Scan)

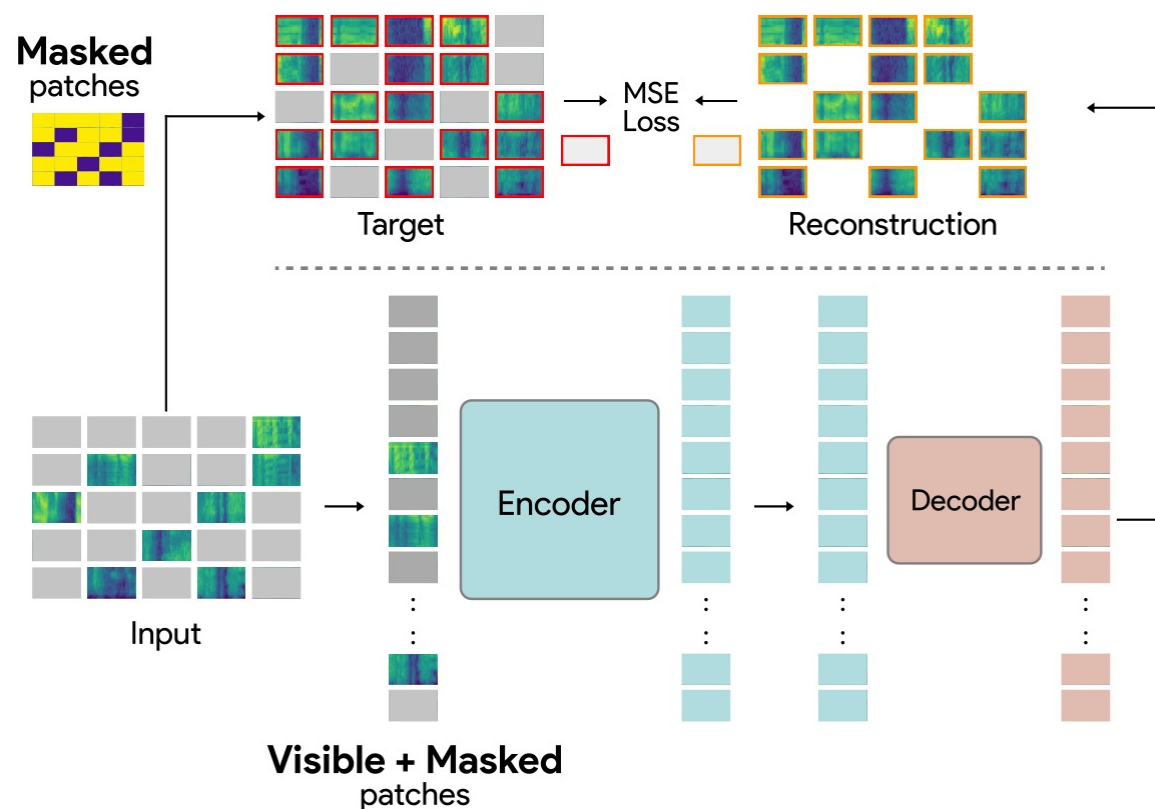
- Mamba dynamically adjusts its internal parameters, Δ, A, B, C , based on the input at each time step t , i.e., the model doesn't use fixed weights for all inputs but selectively adapts them.
- It uses a fully recurrent architecture, processing sequences step-by-step but still retaining the ability to model long-term dependencies.
- Achieves high performance through a hardware aware parallel scan.



Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Self-supervised audio spectrogram transformer (SSAST)

- Autoencoder with masked input.
- Encoder: Get contextual representations using a Vision Transformer (ViT).
- Decoder: Reconstruct the masked patches usually using MLP (can be a transformer).
- Loss: MSE between patches that were masked and corresponding reconstructions from the decoder.



Gong, Y., Lai, C. I., Chung, Y. A., & Glass, J. (2022, June). Ssast: Self-supervised audio spectrogram transformer. In Proceedings of the AAAI Conference on Artificial Intelligence.

Audio Mamba: selective state spaces for self-supervised audio representations

- Compare with the well-established self-supervised audio spectrogram transformer (SSAST)

Model	Data	# Params	BO	CD	ESC-50	LC	Mri-S	Mri-T	NS-5h	SC-5h	F50K	VL	$s(m)$ ↑
SSAST Based													
SSAST [10]	AS+LS	89 M	93.4±0.9	56.5±0.2	68.4±0.4	60.7±0.3	96.7±0.1	96.3±0.1	66.8±0.7	53.5±1.3	38.2±0.1	28.5±0.9	73.1±0.2
SSAST-Tiny	AS	5.4 M	90.4±0.7	46.9±0.2	42.4±0.6	42.7±0.2	95.7±0.1	94.3±0.1	61.2±0.5	50.6±1.6	24.6±0.1	13.8±1.0	56.0±0.2
SSAST-Small	AS	21.5 M	93.2±0.5	51.6±0.2	50.1±0.6	50.0±0.3	96.2±0.1	95.0±0.1	63.8±0.4	58.3±1.2	31.6±0.1	15.6±0.7	63.4±0.3
→ SSAST-Base	AS	85.7 M	93.1±0.7	56.0±0.4	59.6±0.7	52.9±0.3	96.6±0.1	96.2±0.2	64.6±0.8	66.1±1.0	37.5±0.1	19.2±0.9	<u>69.2±0.3</u>
Proposed													
SSAM-Tiny	AS	4.8 M	93.7±0.8	61.8±0.3	70.6±0.2	59.2±0.4	97.1±0.1	94.9±0.1	62.0±0.7	74.8±0.4	41.3±0.2	27.8±1.0	76.3±0.2
SSAM-Small	AS	17.9 M	94.0±0.7	67.5±0.2	78.7±0.6	60.5±0.3	97.5±0.1	96.7±0.1	66.3±0.8	83.7±0.3	48.5±0.1	39.6±0.7	84.4±0.3
→ SSAM-Base	AS	69.3 M	93.2±1.1	70.3±0.2	81.0±0.3	63.5±0.2	97.7±0.1	96.9±0.1	70.5±0.5	87.9±0.3	52.2±0.1	50.4±0.7	<u>89.7±0.3</u>

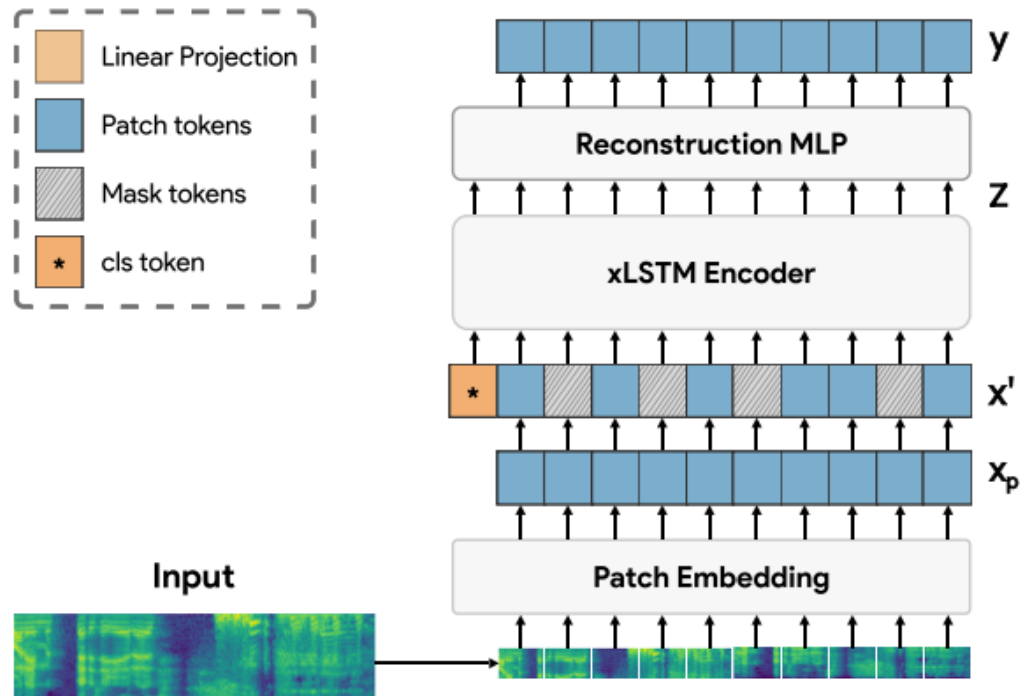
Yadav, S., & Tan, Z. H. (2024). Audio Mamba: Selective State Spaces for Self-Supervised Audio Representations. Interspeech 2024.

xLSTM: Extended long short-term memory

- Introduces exponential gating (i.e., exponential activations on input and forget gates, instead of sigmoid) to better revise storage decisions.
- A normalizer state for better stability.
- 2 building blocks
 - sLSTM:
 - improved memory mixing -> mixing within "heads" but not across them
 - Not parallelizable
 - mLSTM:
 - enhances storage capacity by using a matrix cell state $C \in \mathbb{R}^{(d \times d)}$ instead of a scalar
 - No memory mixing -> parallelizable cell update

Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., ... & Hochreiter, S. (2024). xlstm: Extended long short-term memory. *Advances in Neural Information Processing Systems*, 37.

AxLSTMs: Model architecture



Yadav, S., Theodoridis, S., & Tan, Z. H. (2025). AxLSTMs: learning self-supervised audio representations with xLSTMs. In *Proc. Interspeech 2025*.

AxLSTMs: experimental results

- Pretrained on AudioSet, AxLSTMs outperform comparable self-supervised audio spectrogram transformer (SSAST) baselines
 - by up to 25% in relative performance across ten diverse downstream tasks while
 - having up to 45% fewer parameters.
- Overall score: higher is better:

Model	Data	#M Params	$s(m) \uparrow$
SSAST Based			
SSAST [6]	Mix	89.0	72.5 \pm 0.2
SSAST-Tiny	AS	5.4	55.6 \pm 0.2
<u>SSAST-Small</u>	AS	21.5	63.0 \pm 0.2
<u>SSAST-Base</u>	AS	85.7	68.7 \pm 0.3
Proposed			
AxLSTM-Tiny	AS	4.3	70.7 \pm 0.2
AxLSTM-Small	AS	16.7	81.1 \pm 0.3
<u>AxLSTM-Base</u>	AS	65.6	86.6 \pm 0.2

xLSTM-SENet: xLSTM for Single-Channel Speech Enhancement

- We propose the 1st xLSTM-based speech enhancement system [1], following MP-SENet [2].
- Architecture:

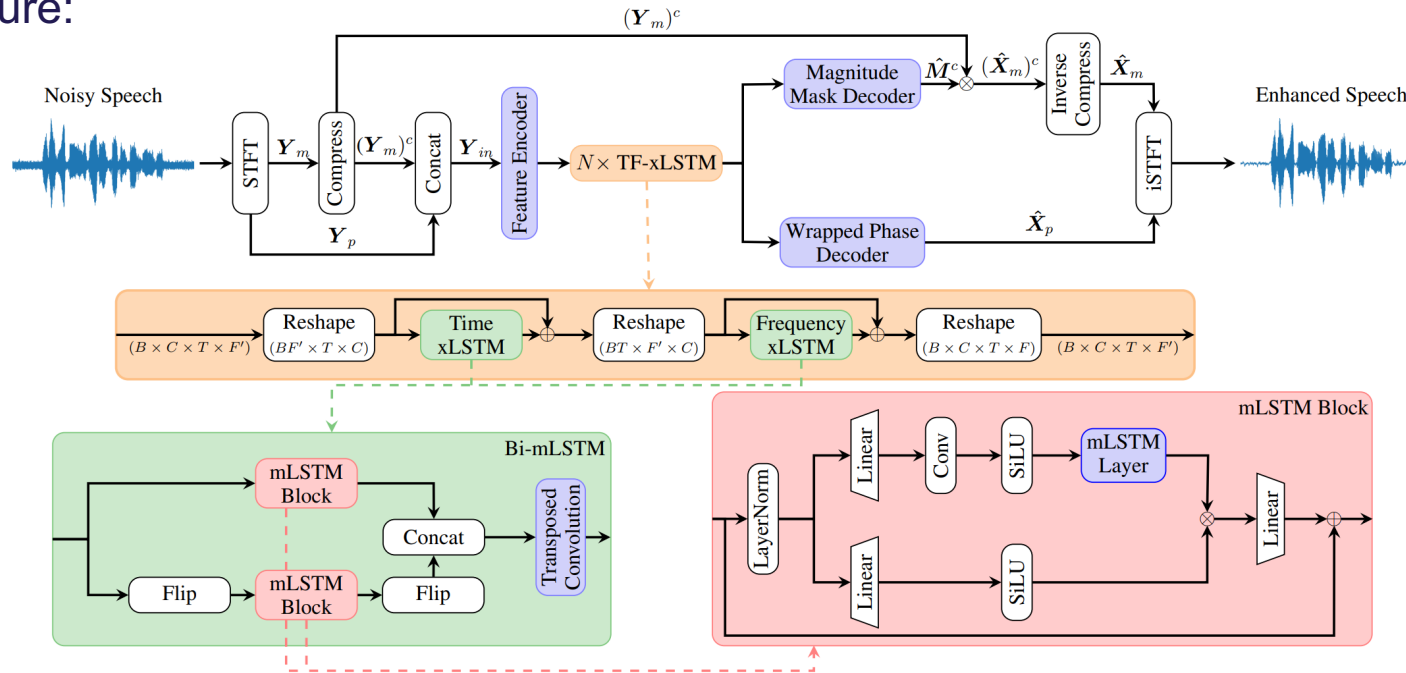


Figure 1: Overall structure of our proposed xLSTM-SENet with parallel magnitude and phase spectra denoising.

[1] **N. L. Kühne**, J. Østergaard, J. Jensen, and Z.-H. Tan, “xlstm-senet: xlstm for single-channel speech enhancement,” In *Proc. Interspeech 2025*.

[2] Y.-X. Lu, Y. Ai, and Z.-H. Ling, “Mp-senet: A speech enhancement model with parallel denoising of magnitude and phase spectra,” in *INTERSPEECH*, 2023, pp. 3834–3838.

xLSTM-SENet: experimental results

- Performance on VoiceBank+Demand

Model	Params (M)	PESQ	CSIG	CBAK	COVL	STOI
Noisy	-	1.97	3.35	2.44	2.63	0.91
MetricGAN+ [9]	-	3.15	4.14	3.16	3.64	-
CMGAN [10]	1.83	3.41	4.63	3.94	4.12	0.96
DPT-FSNet [35]	0.88	3.33	4.58	3.72	4.00	0.96
Spiking-S4 [36]	0.53	3.39	4.92	2.64	4.31	-
TridentSE [37]	3.03	3.47	4.70	3.81	4.10	0.96
MP-SENet [14]	2.05	3.50	4.73	3.95	4.22	0.96
SEMamba [21]	2.25	3.55	4.77	3.95	4.26	0.96
MP-SENet*	2.05	3.49±0.02	4.72±0.02	3.92±0.04	4.22±0.02	0.96±0.00
SEMamba*	2.25	3.49±0.01	4.75±0.01	3.94±0.02	4.24±0.01	0.96±0.00
xLSTM-SENet	2.20	3.48±0.00	4.74±0.01	3.93±0.01	4.22±0.01	0.96±0.00

- Match the performance of SOTA Conformer and Mamba models

N. L. Kühne, J. Østergaard, J. Jensen, and Z.-H. Tan, “xlstm-senet: xlstm for single-channel speech enhancement,” In *Proc. Interspeech 2025*.

MambAttention: Mamba with multi-head attention for generalizable single-channel speech enhancement

- Motivation: Self-attending RNNs have shown improved generalization performance over pure recurrent models.
- Combine Mamba with a shared multi-head attention module across time and frequency dimensions.

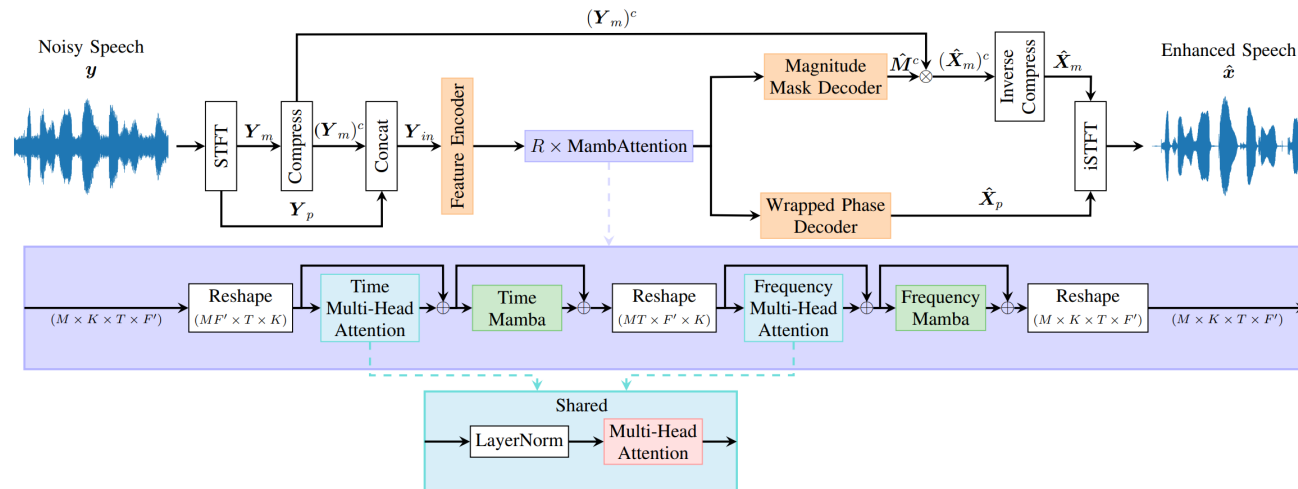


Fig. 1: Overall structure of our proposed MambAttention model. M , K , T , and F' represent the batch size, the number of channels, the number of time frames, and the number of frequency bins, respectively.

Kühne, N.L., Jensen, J., Østergaard, J. and Tan, Z.H., 2025. MambAttention: Mamba with Multi-Head Attention for Generalizable Single-Channel Speech Enhancement. arXiv preprint arXiv:2507.00966.

MambAttention – cont'd

- All models trained exactly in the same framework using *the VB-DemandEx dataset*, only neural architectures differ.
- MambAttention significantly outperforms SOTA systems on the out-of-domain test sets across all reported evaluation metrics.

In-Domain

Out-Of-Domain

Model	Params (M)	VB-DMDEx				DNS 2020			
		PESQ	SSNR	ESTOI	SI-SDR	PESQ	SSNR	ESTOI	SI-SDR
Noisy	-	1.625	-1.068	0.630	4.976	1.582	6.218	0.810	9.071
xLSTM-Attention	2.27	3.019 \pm 0.010	7.689 \pm 0.186	0.800 \pm 0.002	16.653 \pm 0.107	2.801 \pm 0.167	7.187 \pm 0.931	0.886 \pm 0.025	13.913 \pm 1.889
LSTM-Attention	2.48	3.023 \pm 0.037	7.645 \pm 0.339	0.803 \pm 0.008	16.596 \pm 0.279	2.546 \pm 0.183	5.792 \pm 0.878	0.847 \pm 0.032	10.961 \pm 1.622
Conformer [27]	2.05	2.935 \pm 0.065	7.641 \pm 0.283	0.787 \pm 0.010	16.202 \pm 0.318	2.666 \pm 0.010	7.369 \pm 0.382	0.875 \pm 0.009	13.665 \pm 0.892
➔ MambAttention	2.33	3.026 \pm 0.007	7.674 \pm 0.411	0.801 \pm 0.002	16.684 \pm 0.095	2.919 \pm 0.118	8.133 \pm 0.733	0.911 \pm 0.009	15.169 \pm 1.363

Noisy

xLSTM

Conformer

MambAttention

Clean



Kühne, N.L., Jensen, J., Østergaard, J. and Tan, Z.H., 2025. MambAttention: Mamba with Multi-Head Attention for Generalizable Single-Channel Speech Enhancement. arXiv preprint arXiv:2507.00966.

Conclusions

- For audio representation learning
 - both SSAM and AxLSTM outperform comparable transformer baselines, while having fewer parameters
 - SSAM > AxLSTM >> SSAST
 - For speech enhancement
 - xLSTM and Mamba match or outperform in-domain performance of attention-based models (Conformer).
 - Combining Mamba and xLSTM with multi-head attention yields significant improvements in generalization/out-of-domain performance.
 - xLSTM is memory hungry and slower compared to Mamba and Transformer, in our settings
-

Thank you for your attention.

Thank my co-authors as cited in slides.



Engineering and
Physical Sciences
Research Council

LLMs for Audio Intelligence: From Understanding to Generation



centre for digital music

By Jinhua Liang

September 1, 2025

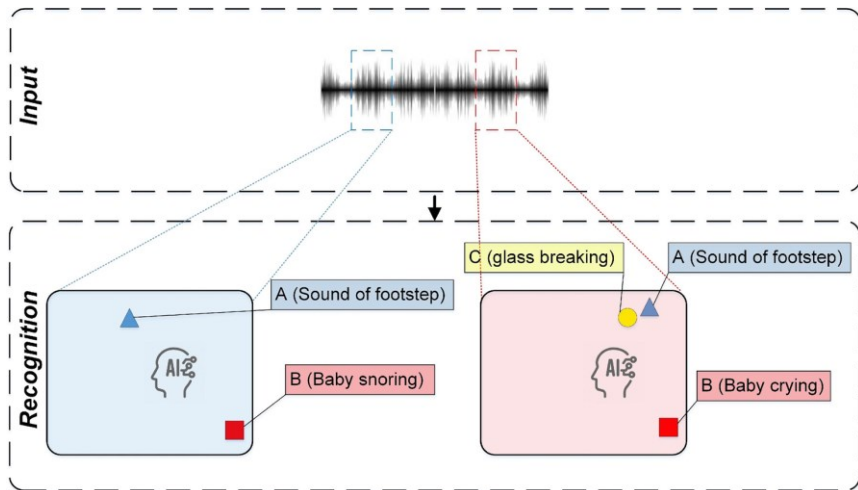
Before start...

I am a final-year Ph.D. candidate advised by Dr. Emmanouil Benetos, Dr. Huy Phan, and Prof. Mark Sandler, at Centre for Digital Music (C4DM), Queen Mary University of London. My research interests are spanning from audio/speech/music understanding, controllable audio generation, multimodal representation learning.

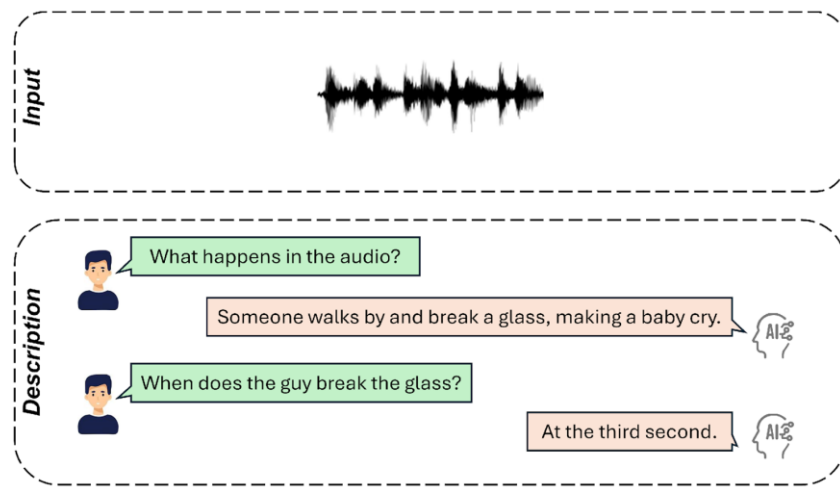
Great thanks to my collaborators:

Emmanouil Benetos (Queen Mary University of London)
Huan Zhang (Queen Mary University of London)
Xubo Liu (Meta)
Huy Phan (Meta)
Haohe Liu (Meta)
Mark D. Plumbley (University of Surrey)
Wenwu Wang (University of Surrey)
Qiuqiang Kong (Chinese University of Hong Kong)
Yin Cao (Xi'an Jiaotong-Liverpool University)
Dan Stowell (Tilburg University)
Zhuo Chen (ByteDance)
Yuxuan Wang (ByteDance)
Sebastian Braun (Microsoft Research)
Hannes Gamper (Microsoft Research)
Ivan Tashev (Microsoft Research)

What’s “wrong” with closed-end classifiers?



- Trained on the task-specific dataset(s)
- Require a predefined set of sound events
- Less practical in the real-world application



- Trained on a broad range of audio tasks
- Describe sound in human language
- Following complex instructions
- Approachable to the user.

This highlights the study on interacting with sounds using natural language

Background: Revisit language modelling

What is a Large Language Model (LLM)?

- LLM, such as ChatGPT, learns to predict next token with massive textual tokens

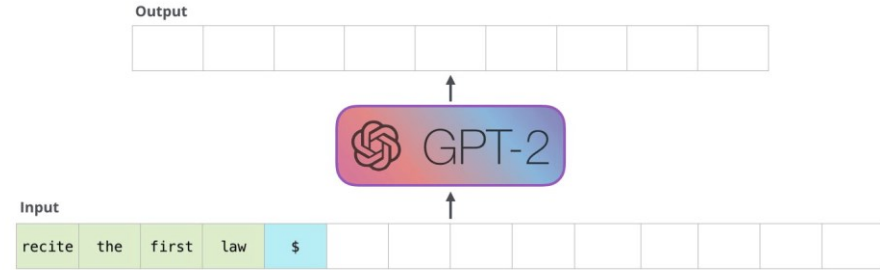


Fig 1. Overall framework of Generative Pre-trained Transformer 2 (GPT-2) [1].
(img borrowed from @JayAlammar[2])

Background: Revisit language modelling

What is a Large Language Model (LLM)?

- LLM, such as ChatGPT, learns to predict next token with massive textual tokens

What is multi-modal LLM?

- A generative network that predicts the next token conditioned on image/video/audio + text.

Why multi-modal LLM?

- Leverage knowledge within LLMs
- Explore Homogeneity across tasks

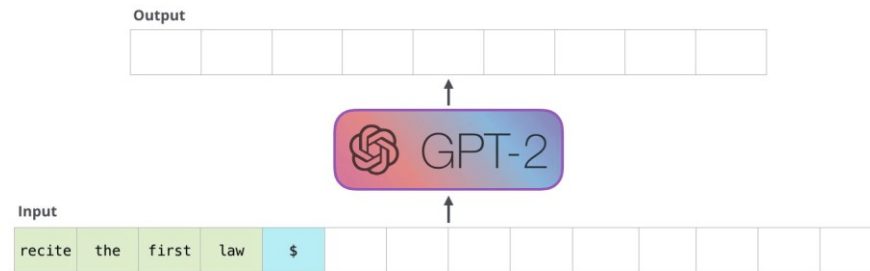


Fig 1. Overall framework of Generative Pre-trained Transformer 2 (GPT-2) [1].
(img borrowed from @JayAlammar[2])

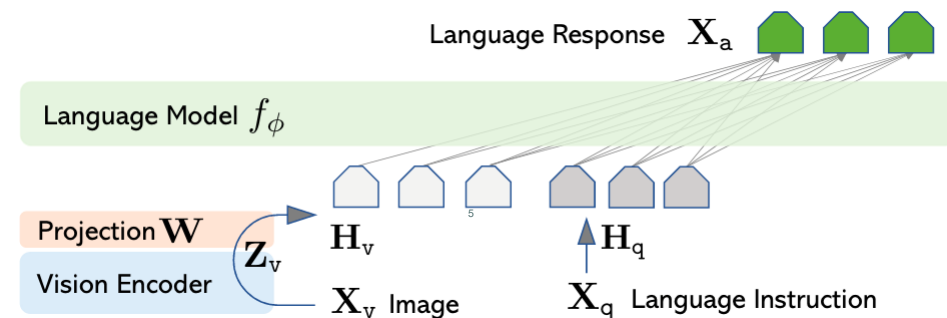
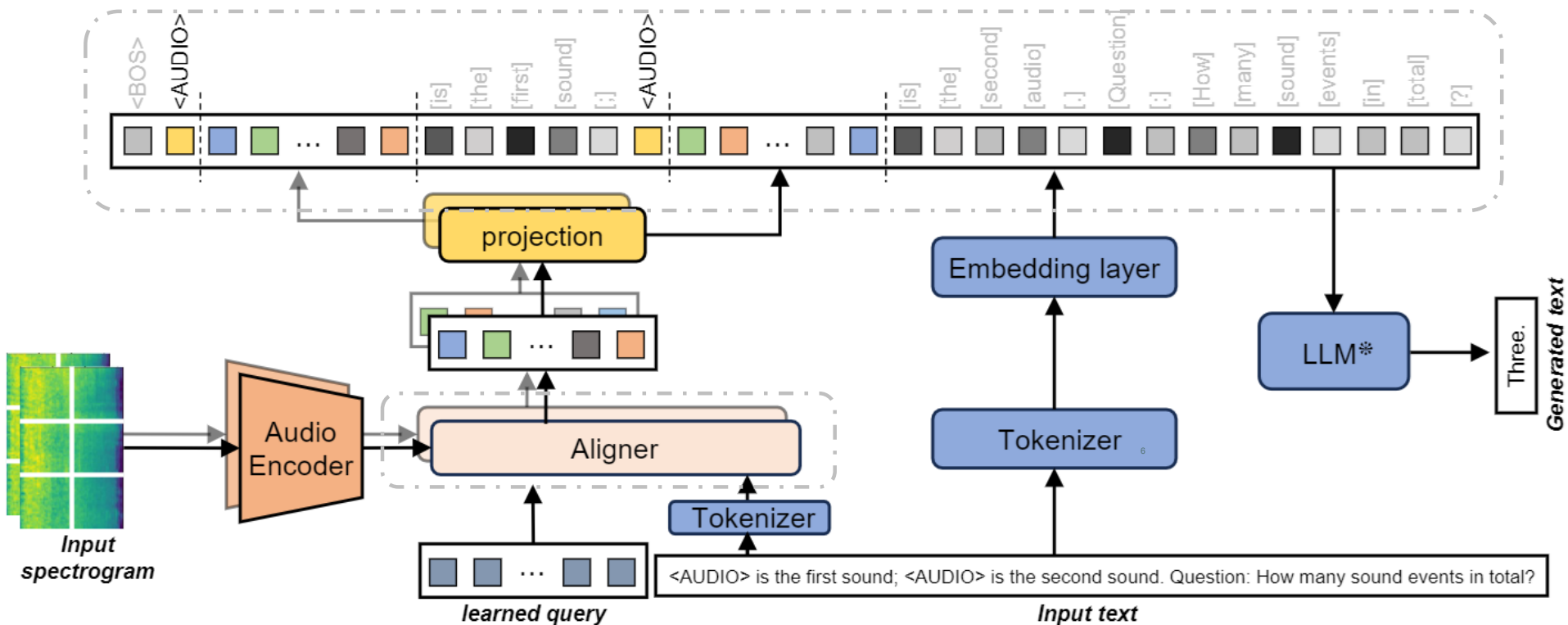


Fig 2. Overall structure of LLAVA [2]

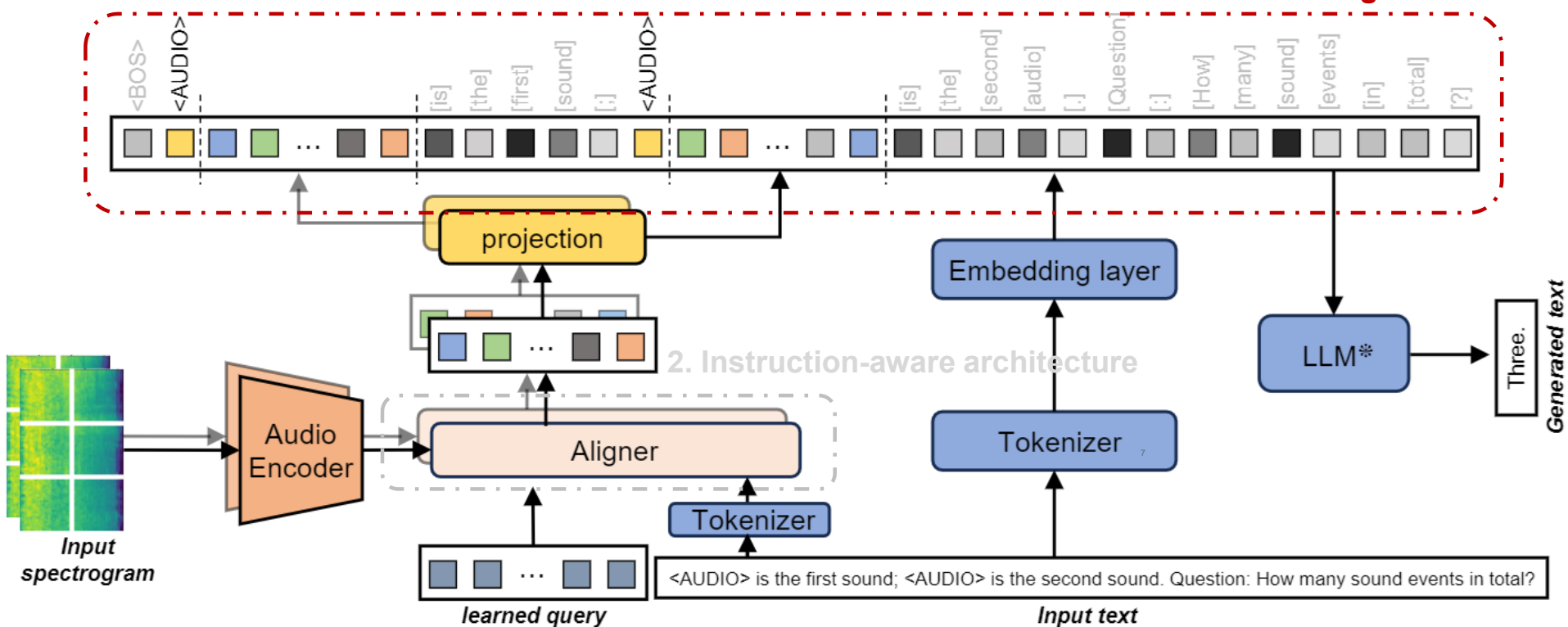
Method: Overview of APT

Acoustic Prompt Tuning (APT): an adapter extending LLMs/VLMs to the audio domain using an improved soft-prompting approach



APT is highlighted three core designs:

1. Interleaved acoustic and text embeddings



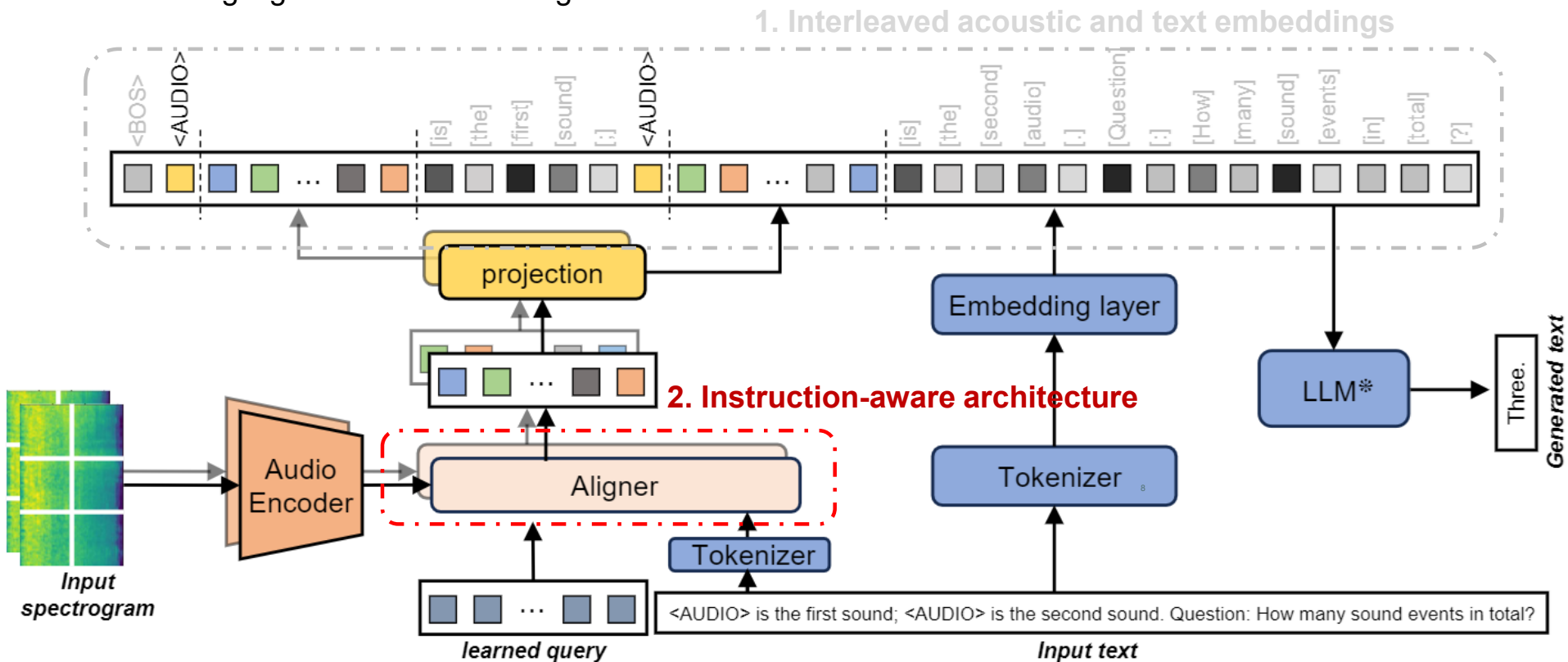
2. Instruction-aware architecture

3. Learning in curriculum

Learning large language models to “hear” the sound

Method: Overview of APT

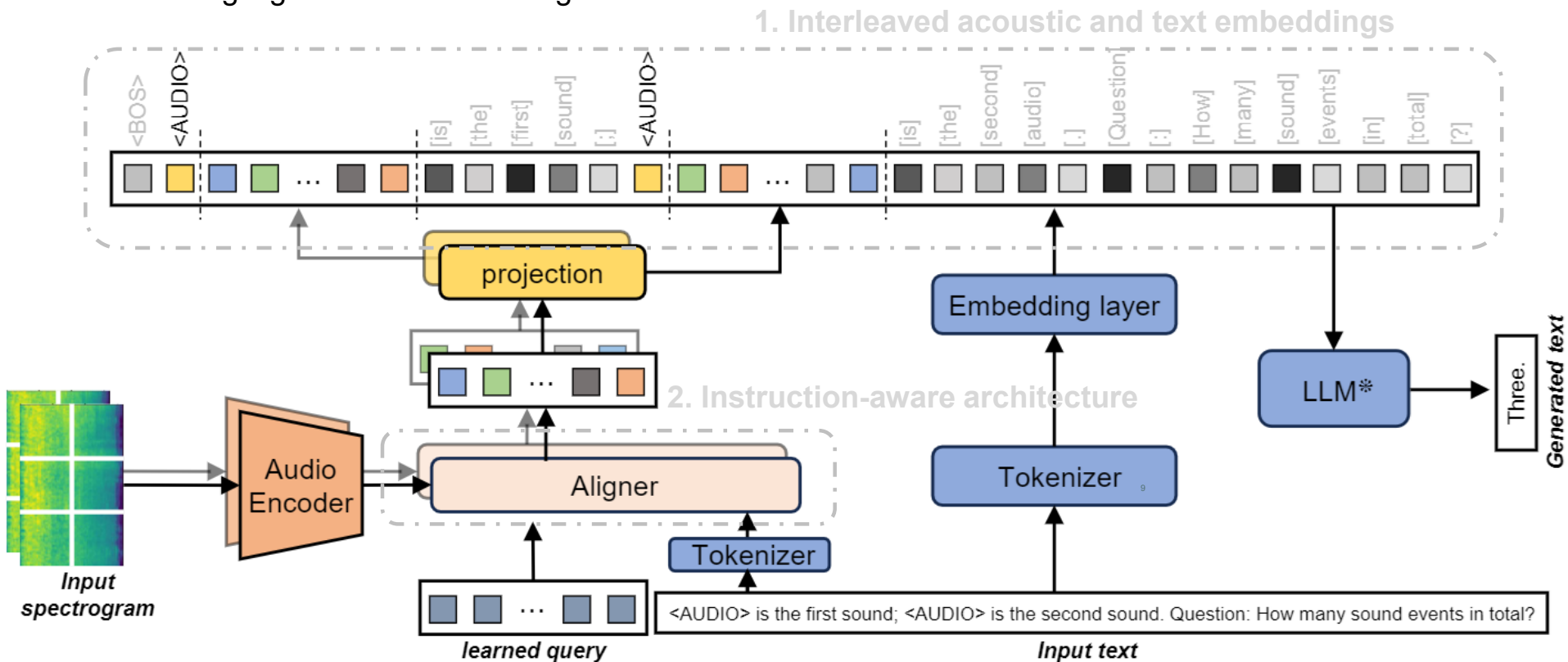
APT is highlighted three core designs:



Learning large language models to “hear” the sound

Method: Overview of APT

APT is highlighted three core designs:



Method: NLAR task setup

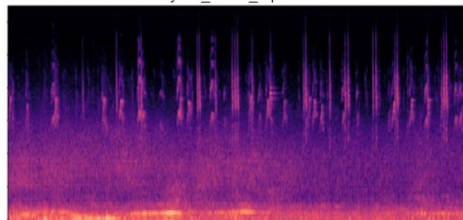
We propose the **natural language audio reasoning** task, to evaluate model’s ability to analyse audio clips by comparison and summarisation.

Table 2: An example demonstrating APT-LLM’s capacity of audio reasoning. It requires audio networks to comprehend recordings and reasoning across multiple recordings.

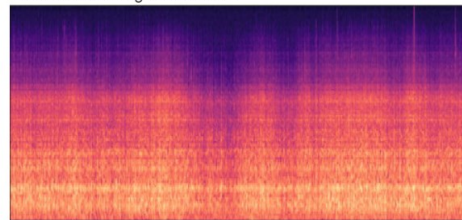
Natural Language Audio Reasoning (NLAR) example: “Where is the sudden sound?”

User

Wav1: “AmbianceBackyard_Quiet_bip.wav”

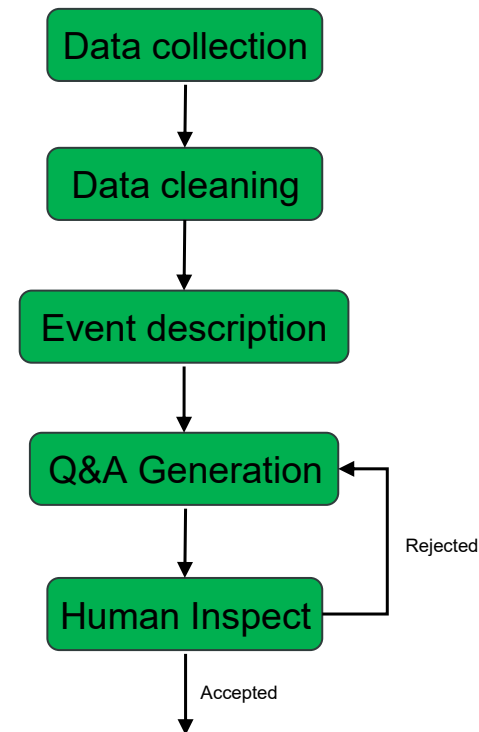


Wav2: “Rain hitting window.wav”



Question: Which recording has a more sudden and startling sound event?

APT-LLM	First.
Ground truth	first



Learning large language models to “hear” the sound

Discussion: Evaluation on the audio understanding tasks

TABLE IV
ZERO-SHOT AUDIO UNDERSTANDING PERFORMANCE (%) COMPARISON WITH AUDIO LANGUAGE MODELS. WE GROUP THE METHODS IN TERMS OF THEIR TRAINING STRATEGY. “#PARAMS.” DENOTES THE NUMBER OF TRAINABLE PARAMETERS AND “#PAIRS” REPRESENTS THE NUMBER OF AUDIO-TEXT PAIRS. ↑ INDICATES THE HIGHER NUMBER, THE BETTER PERFORMANCE.

Model	#Params.	#Pairs	AudioSet (mAP↑)	AudioCaps (SPICE↑)	Clotho (SPICE↑)
<i>Audio-language models trained with the contrastive loss</i>					
AudioCLIP [39]	30M	2M	25.9	-	-
CLAP [40]	190M	128k	5.8	-	-
<i>One-for-all models for various audio tasks</i>					
Qwen-Audio [7]	640M	-	18.5	-	13.6
LTU [9]	96M	5.7M	18.5	17.0	11.9
Pengi [13]	>191M	3.4M	-	18.2	12.6
APT-LLM	101M	2.6M	14.7	17.1	11.6

APT-LLM has a promising result on common audio tasks without fine-tuning on task-specific data. After fine-tuning for two epochs, APT-LLM achieves the best performance on downstream tasks.

TABLE V
PERFORMANCE (%) COMPARISON IN AUTOMATED AUDIO CAPTIONING TASKS. ↑ INDICATES THE HIGHER NUMBER, THE BETTER PERFORMANCE.

Model	AudioCaps		Clotho		Weighted average	
	SPICE↑	SPIDER↑	SPICE↑	SPIDER↑	SPICE↑	SPIDER↑
<i>Specialised systems trained with task-specific examples</i>						
AT+CNN10 [41]	16.8	-	11.5	-	15.1	-
CNN-GPT2 [42]	16.7	43.8	11.1	21.5	14.9	36.7
WSAC+PD [43]	17.3	40.3	12.3	24.7	15.7	35.3
WavCaps [32]	18.2	48.5	13.3	29.7	16.6	42.5
<i>One-for-all models for various audio tasks</i>						
APT-LLM	19.1	40.2	13.2	24.8	17.2	35.3

Learning large language models to “hear” the sound

Discussion: Evaluation on the advanced understanding tasks

TABLE VII
BENCHMARKING APT FOR ACCURACY (%) ON THE NATURAL LANGUAGE
AUDIO REASONING TASK.

Model	Accuracy↑
AAC+ChatGPT	27.9
APT-LLM	62.9
APT-LLM _{1.5}	63.8

TABLE VIII
ACCURACY (%) COMPARISON BETWEEN DIFFERENT MODALITIES IN
AUDIO-VISUAL LEARNING.

Model	Modal	Accuracy↑
BLIP-2 1	Video-only	42.9
APT-LLM	Audio-only	27.7
APT-BLIP-2	Audio-video	59.7

15

APT-LLMs yield the best performance on NLAR and zero-shot audio-visual Q&A tasks.

Audio editing is to change the content of audio by following the instruction precisely.

Existing works treat it as a regeneration task, overlooking the need for high-fidelity and localized editing.

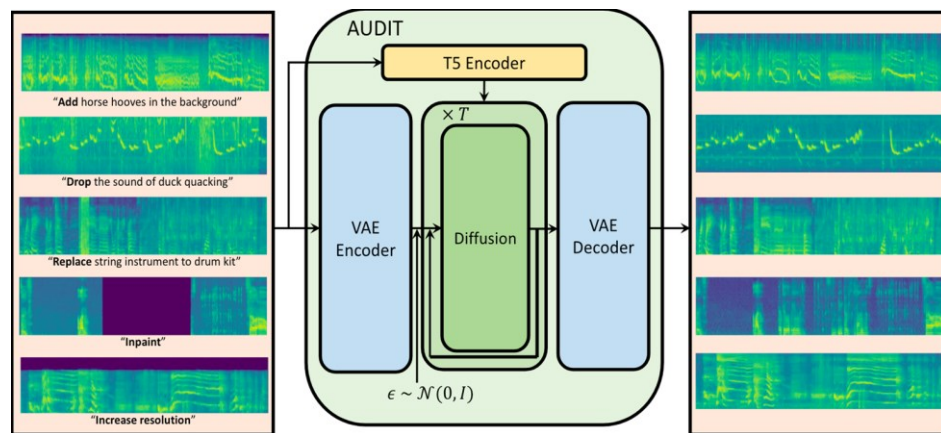
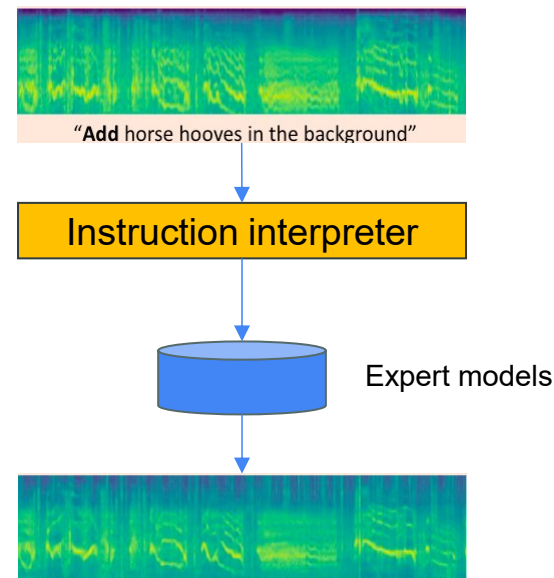
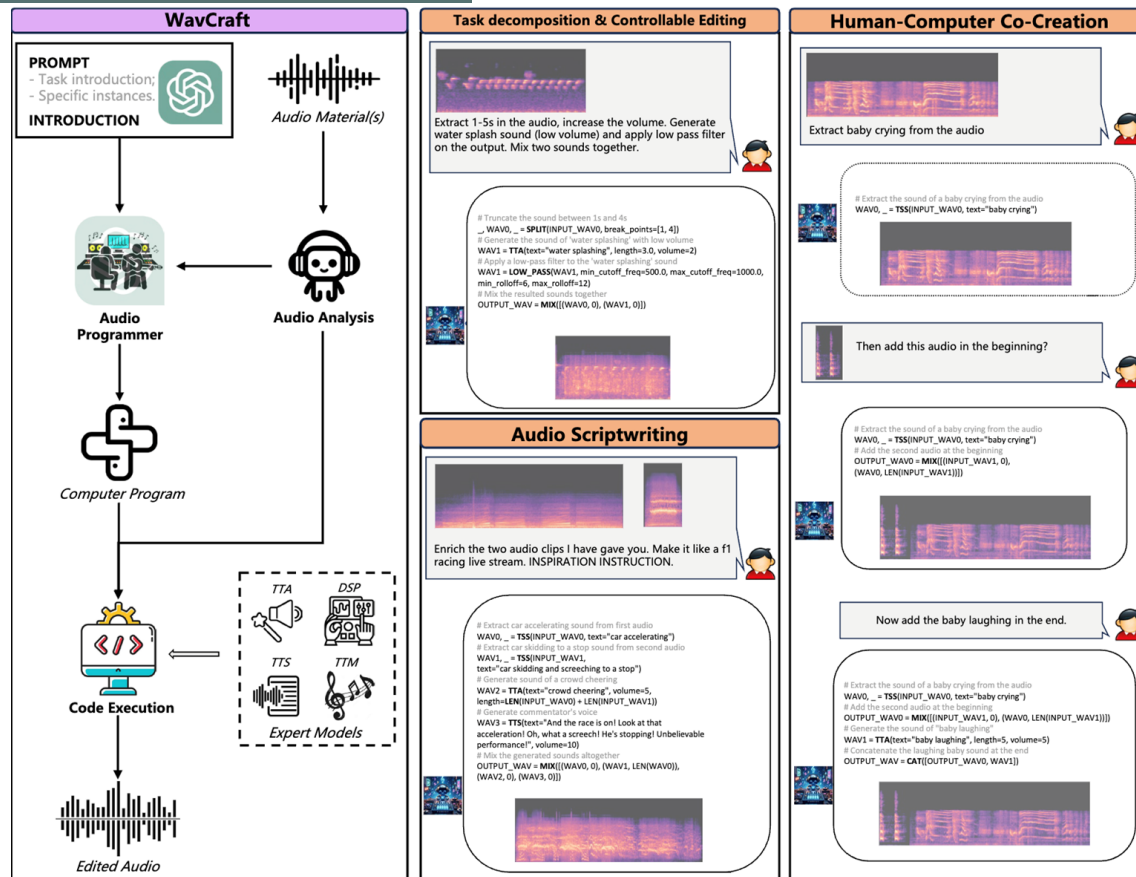


Fig. An example of end-to-end audio editing system [1]



Audio generation and editing with the AI agent

Method: Overview of WavCraft



Audio generation and editing with the AI agent

Discussion: Evaluation on the audio editing tasks



Table 1. Compared to the audio editing model on five editing tasks.
FAD: Fréchet Audio Distance, IS: Inception score, KL: kullback-leibler divergence, LSD: log spectral distance

Task	AUDIT (Wang et al.)				WavCraft			
	FAD ↓	IS ↑	KL ↓	LSD ↓	FAD ↓	IS ↑	KL ↓	LSD ↓
Add	9.27	3.87	3.00	1.95	0.63	6.05	1.45	1.59
Removal	17.57	3.27	4.40	3.46	3.48	6.38	1.72	2.07
Replacement	10.24	2.86	3.10	2.55	0.72	6.09	2.16	1.77
Infilling	12.61	3.88	2.86	3.40	3.31	6.37	1.00	2.10
Super-resolution	13.68	2.62	4.25	2.50	5.98	5.96	1.26	1.93

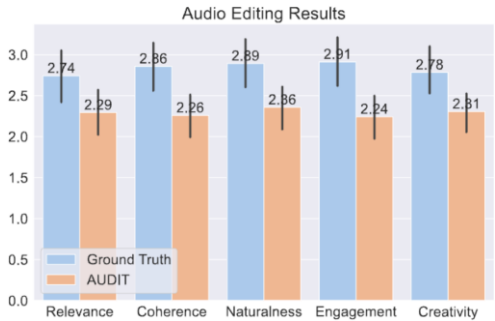


Figure 1. Comparing the ability of audio storytelling

Table 2. Compared to the audio generative models

Model	FAD ↓	KL ↓	IS ↑
AudioLDM (Liu et al., 2023a)	4.65	1.89	7.91
WavJourney(Liu et al., 2023d)	3.38	1.53	7.94
WavCraft	2.95	1.68	8.07



Figure 2. Overall subjective evaluation on audio editing quality

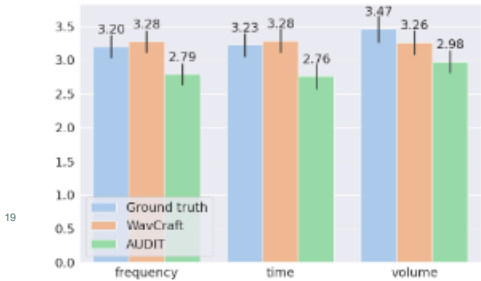


Figure 3. Subjective evaluation on the quality of edited audio

Discussion: Takeaways

- Introduced Acoustic Prompt Tuning (APT) that enables LLMs to process audio by injecting instruction-aware audio embeddings, supporting diverse audio tasks with minimal architectural changes.
- Curriculum learning and interleaved audio-text inputs are performed via various tasks without input format constraints, enabling richer multi-audio reasoning.
- WavCraft is an LLM-driven audio agent that interprets user instructions and decomposes them into subtasks, coordinating specialized models to create or edit audio content.
- It uses natural language prompts and audio descriptions to control generation and editing in a fine-grained, interpretable, and user-friendly way.

20

Related work

- **Jinhua Liang**, Xubo Liu, Wenwu Wang, Mark D. Plumbley, Huy Phan and Emmanouil Benetos, “Acoustic Prompt Tuning: Empowering Large Language Models With Audition Capabilities,” in IEEE Transactions on Audio, Speech and Language Processing, vol. 33, pp. 949-961, 2025.
- **Jinhua Liang**, Huan Zhang, Haohe Liu, Yin Cao, Qiuqiang Kong, Xubo Liu, Wenwu Wang, Mark D. Plumbley, Huy Phan, Emmanouil Benetos. “WavCraft: Audio Editing and Generation with Natural Language Prompts”, International Conference on Learning Representations (ICLR) 2024 Workshop on LLM Agents.
- Huan Zhang, Vincent K.M. Cheung, Hayato Nishioka, Simon Dixon, Shinichi Furuya. (2025) "LLaQo: Towards a Query-Based Coach in Expressive Music Performance Assessment," 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, 2025, pp. 1-5.
- Xubo Liu, Zhongkai Zhu, Haohe Liu, Yi Yuan, Meng Cui, Qiushi Huang, **Jinhua Liang**, Yin Cao, Qiuqiang Kong, Mark D. Plumbley, Wenwu Wang. “Wavjourney: Compositional audio creation with large language models”, IEEE Transactions on Audio, Speech and Language Processing.

E-mail: jinhua.liang@qmul.ac.uk

ORCID: <https://orcid.org/0000-0002-4570-0735>

Homepage: <https://jinhualiang.github.io/>, or:



By Jinhua Liang

September 1, 2025

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/>

01/09/2025

Many thanks to...

- **Xubo Liu**
- **Yi Yuan**
- Haohe Liu
- Xinhao Mei
- Qiuqiang Kong
- Mark Plumbley,
- ...

Funding from:

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

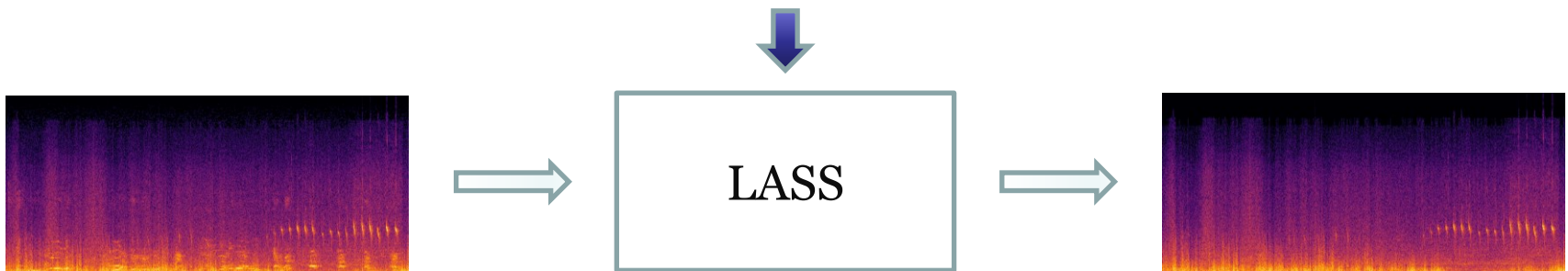


Engineering and
Physical Sciences
Research Council

- **Introduction**
 - What is language queried audio source separation (LASS)?
 - Existing methods and challenges
- **Two new methods**
 - Discriminative model: AudioSep
 - Generative model: FlowSep
- **Conclusions and future works**

- 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.

Existing Methods and Challenges

- **Existing methods**

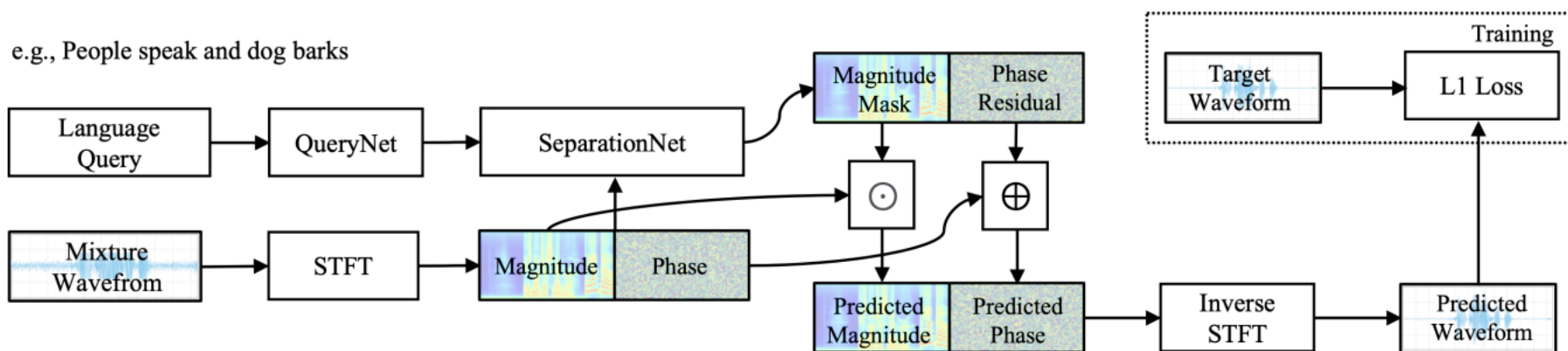
- LASS-Net (Liu et al 2022): first LASS model
- SoundFilter (Kilgour 2022): exploiting audio supervision
- CLIPSep (Dong et al. 2023): exploiting visual supervision

- **Challenges**

- Small scale of data available for training
- Open-domain source separation with texts
- Overlapping sound events
- Processing artefacts and incomplete separation

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



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.

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

BENCHAMRK EVALUATION RESULTS OF AUDIOSEP AND COMPARISON WITH BASELINE SYSTEMS.

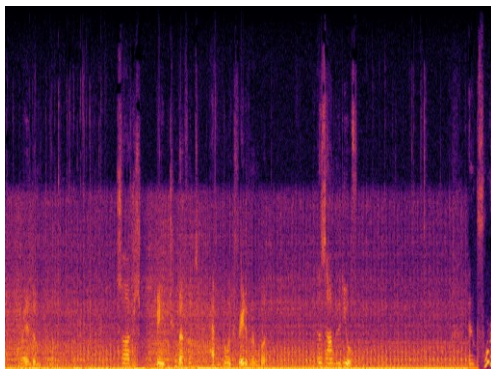
	AudioSet		VGGSound		AudioCaps		Clotho		MUSIC		ESC-50		Voicebank-DEMAND	
	SI-SDR	SDRi	SI-SDR	SDRi	SI-SDR	SDRi	SI-SDR	SDRi	SI-SDR	SDRi	SI-SDR	SDRi	PESQ	SSNR
USS-ResUNet30 [15]	-	5.57	-	-	-	-	-	-	-	-	-	-	2.18	9.00
USS-ResUNet60 [15]	-	5.70	-	-	-	-	-	-	-	-	-	-	2.40	9.35
LASSNet [3]	-3.64	1.47	-4.50	1.17	-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*, vol. 33, 458--471, 2025.

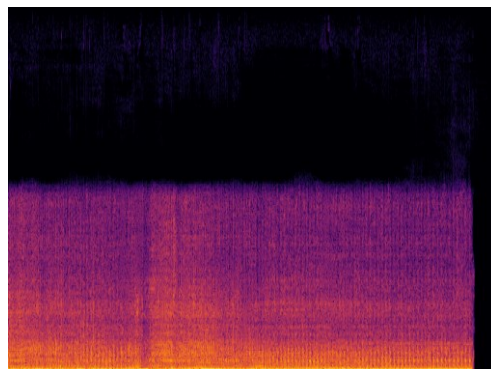
AudioSep: Sound Demos

Human query: “The engine sound of a vehicle”

Mix 

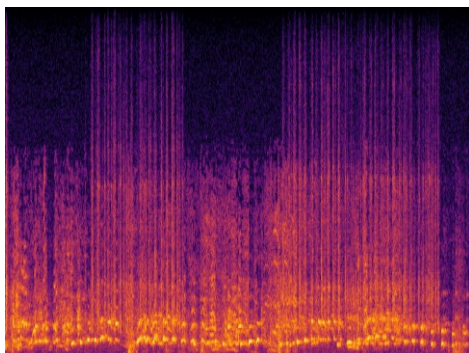



Separated 

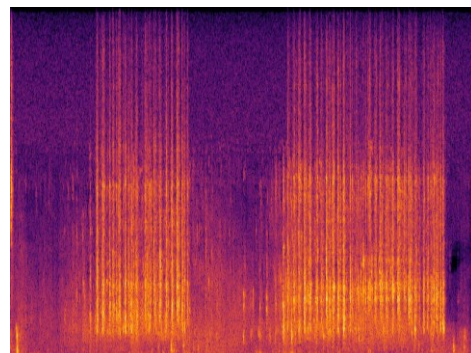


Human query: “The sound of hitting the keyboard”

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:

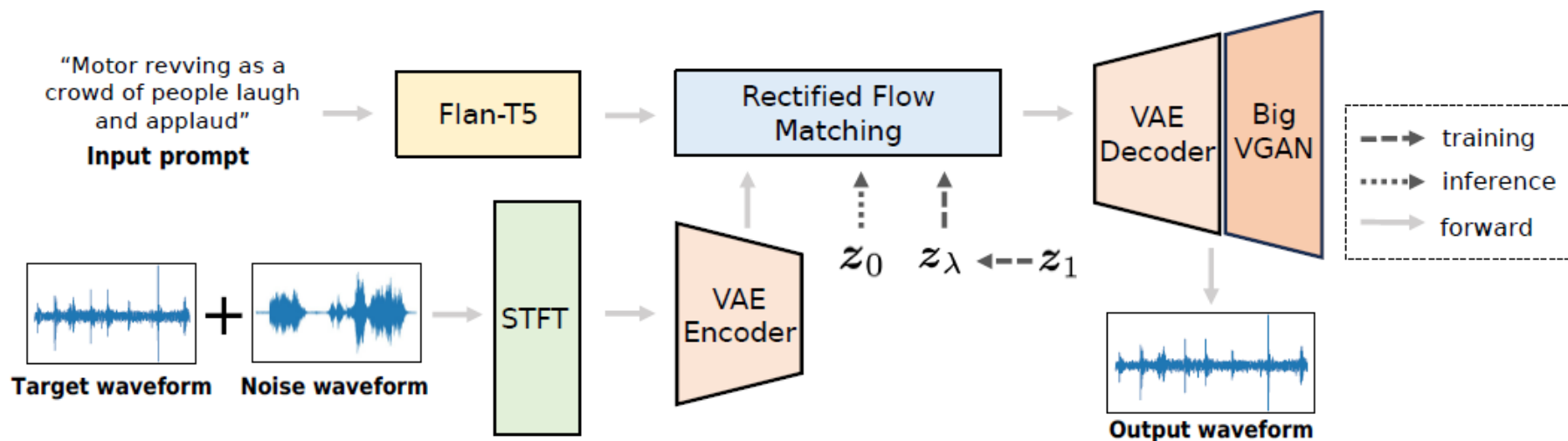
Using the generative approaches.

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. Plumley, 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 not sharing any overlapping sound source classes. All the segments are padded or cropped to 10 seconds with 16kHz sampling rate, and we mix two waveforms 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 between -15 and 15 dB.
- DCASE2024 Task 8:
 - DCASE-Synth includes 3000 mixtures from 1000 selected audio clips under an SNR between -15 and 15 dB.
 - DCASE-Real consists of 100 audio clips from read-world scenarios.

Experimental Results

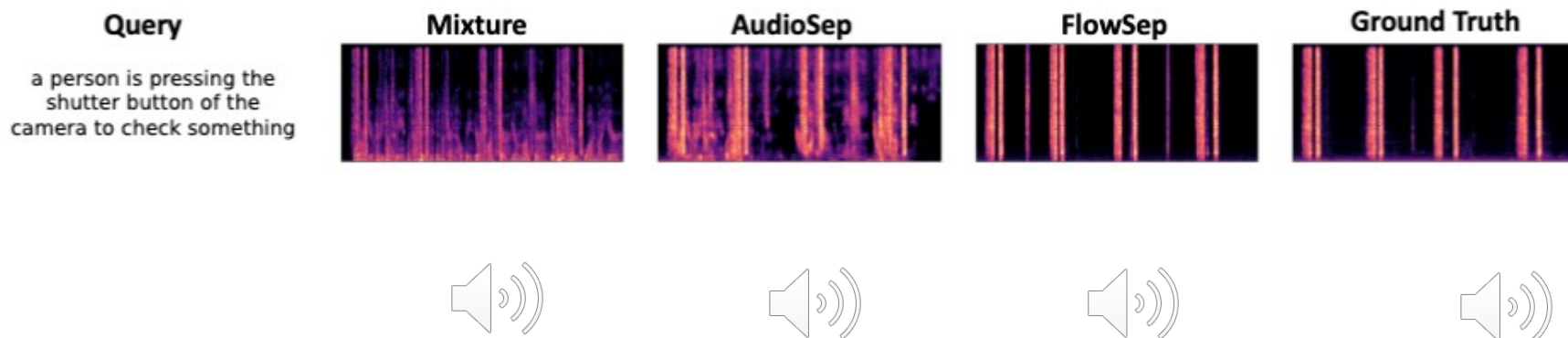
- Unlike discriminative network that modify the original audio clips, generated results do not strictly align with the target audio sample in the temporal dimension.
- Hence, traditional objective metrics are not suitable for evaluating generative models based separation methods.
- We apply FAD, CLAPScore 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

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/.

DCASE Challenge –Task 10

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

Separation
Task 9

Language-Queried Audio Source Separation

DCASE 2024

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

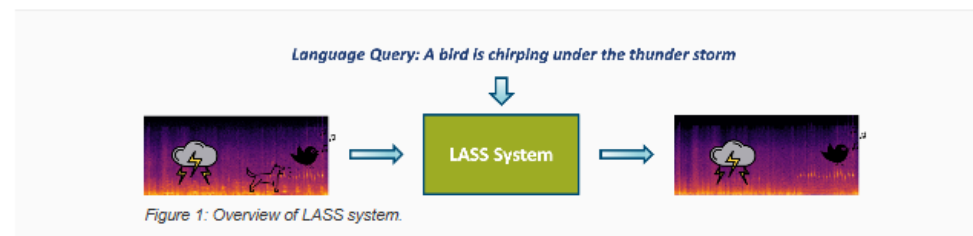
Separate arbitrary audio sources using natural language queries.

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.



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.
- Both discriminative (**AudioSep**) and generative approaches (**FlowSep**) have been developed, offering state-of-the-art performance.

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 "annoying 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/>

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:

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

Google Research

Stars 1.3k



Explainability for Audio Models

Cem Subakan



UNIVERSITÉ
LAVAL

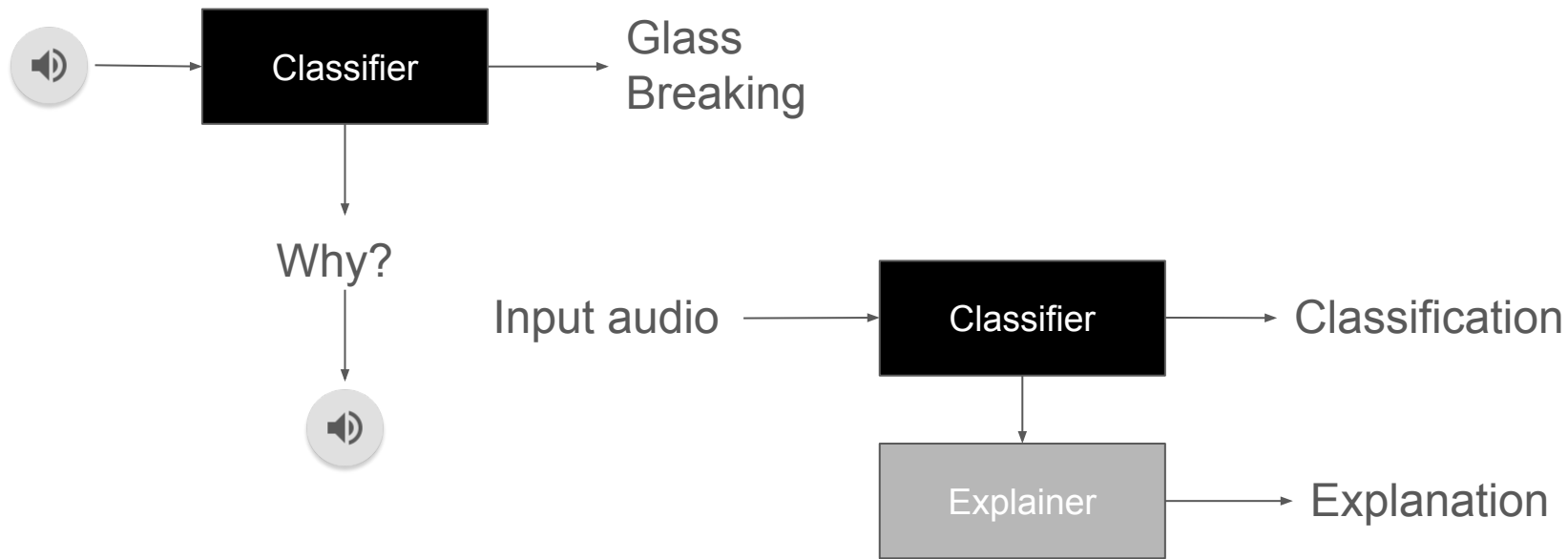


UNIVERSITÉ
Concordia
UNIVERSITY



Mila

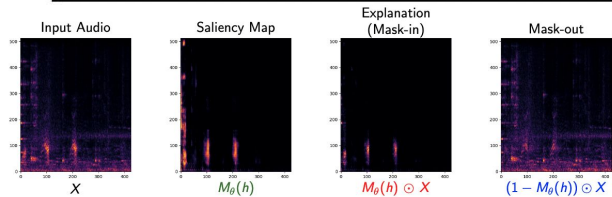
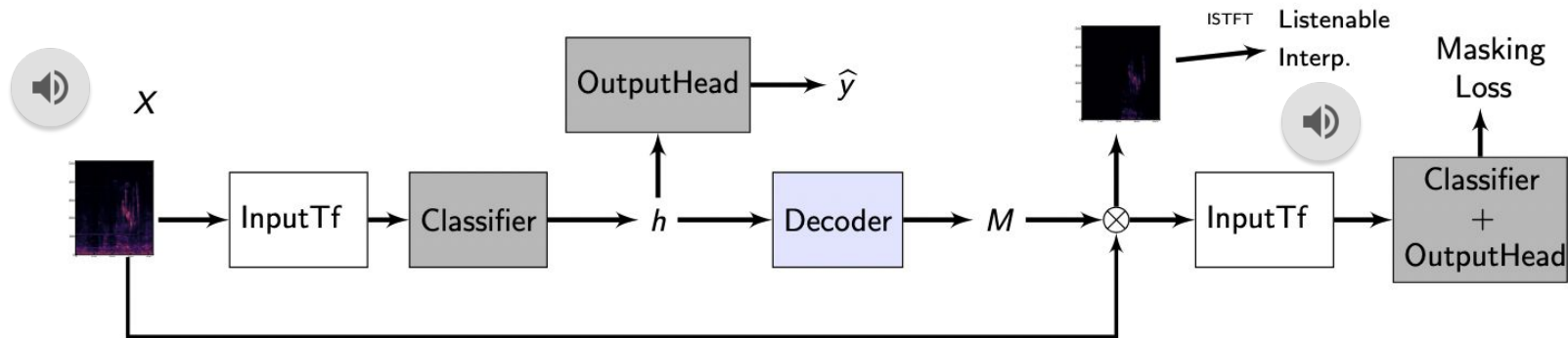
Listenable Posthoc Explanations for Audio Classifiers



Desiderata: Faithful, Listenable, Understandable Explanations

Explainability in Decision Critical Applications (e.g. Healthcare, DeepFake detection)

Listenable Maps for Audio Classifiers (L-MAC)

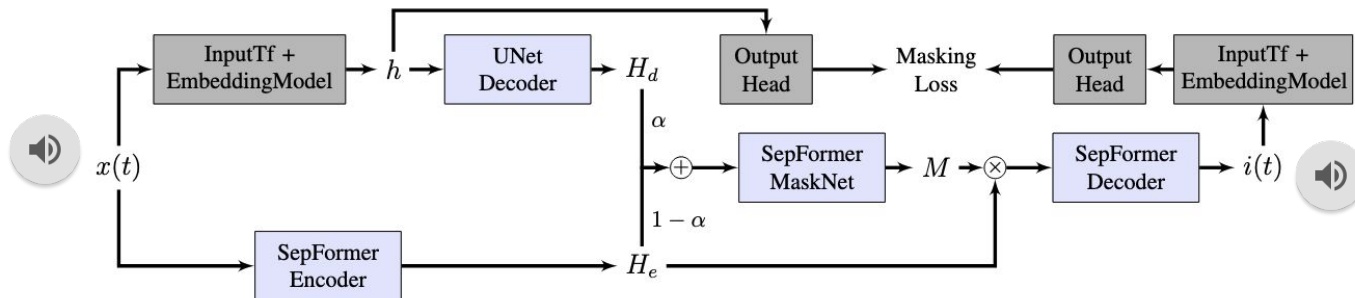


$$\min_{\theta} \underbrace{\lambda_{in} \mathcal{L}_{in}(\log f(M_{\theta}(h) \odot X), \hat{y})}_{\text{Mask-in}} - \underbrace{\lambda_{out} \mathcal{L}_{out}(\log f((1 - M_{\theta}(h)) \odot X), \hat{y})}_{\text{Mask-out}} + \underbrace{|M_{\theta}(h)|}_{\text{Mask Reg}}$$

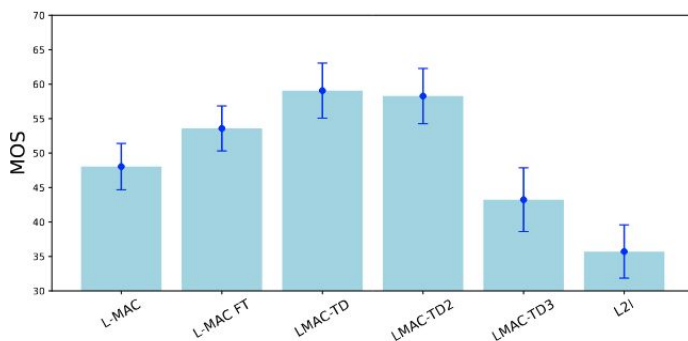
Faithful & Understandable Explanations!

	Metric	AI (↑)	AD (↓)	AG (↑)	FF (↑)	Fid-In (↑)	SPS (↑)	COMP (↓)
Listenable (STFT → Mel)	Saliency	0.00	15.79	0.00	0.05	0.07	0.39	5.48
	Smoothgrad	0.00	15.71	0.00	0.03	0.05	0.42	5.32
	IG	0.25	15.45	0.01	0.07	0.13	0.43	5.11
	GradCAM	8.50	10.11	1.47	0.17	0.33	0.34	5.64
	Guided GradCAM	0.00	15.61	0.00	0.05	0.06	0.44	5.12
	Guided Backprop	0.00	15.66	0.00	0.05	0.06	0.39	5.47
	L2I, RT=0.2	1.63	12.78	0.42	0.11	0.15	0.25	5.50
	SHAP	0.00	15.79	0.00	0.05	0.06	0.43	5.24
	L-MAC (ours)	36.25	1.15	23.50	0.20	0.42	0.47	4.71
	L-MAC, FT, $\lambda_g = 4$ (ours)	32.37	1.98	18.74	0.21	0.41	0.43	5.20
Not Listenable (Mel)	Saliency	0.00	15.81	0.00	0.10	0.07	0.39	4.53
	Smoothgrad	0.00	15.61	0.00	0.07	0.04	0.39	4.54
	IG	0.00	15.55	0.00	0.12	0.08	0.42	4.36
	GradCAM	7.00	10.93	1.04	0.17	0.29	0.34	4.72
	Guided GradCAM	0.125	15.40	6.67	0.08	0.07	0.45	4.17
	Guided Backprop	0.125	15.54	0.00	0.10	0.08	0.39	4.53
	SHAP	0.00	15.57	0.00	0.11	0.08	0.41	4.42
	L-MAC (ours)	35.63	1.59	24.28	0.22	0.42	0.45	4.11
	L-MAC (ours) FT, $\lambda_g = 4$	36.13	1.28	21.15	0.23	0.42	0.32	4.71

L-MAC in Time Domain



Improves the Sound Quality

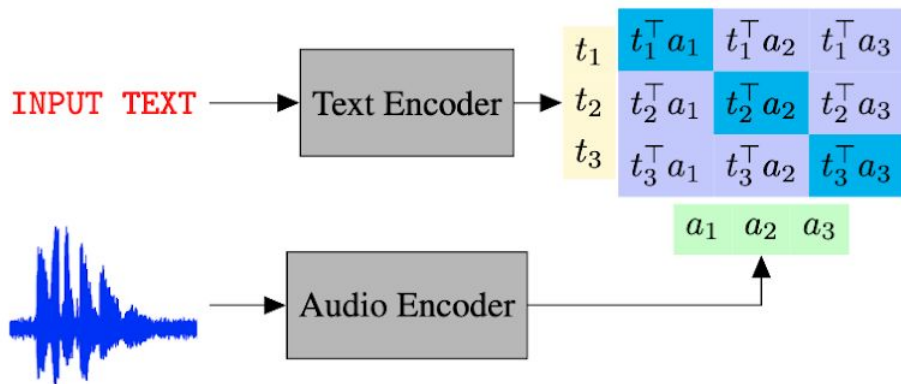


Improves the Faithfulness (for the most part)

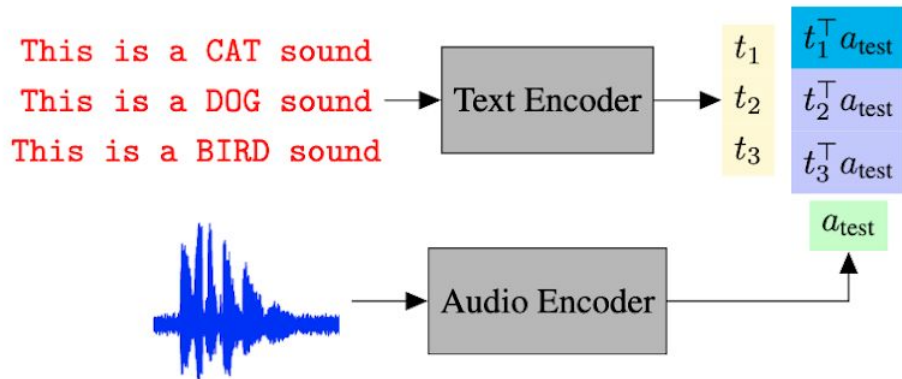
Metric	AI (↑)	AD (↓)	AG (↑)	FF (↑)	Fid-In (↑)	SPS (↑)	COMP (↓)
Saliency	0.00	15.79	0.00	0.05	0.07	0.39	5.48
Smoothgrad	0.00	15.71	0.00	0.03	0.05	0.42	5.32
IG	0.25	15.45	0.01	0.07	0.13	0.43	5.11
GradCAM	8.50	10.11	1.47	0.17	0.33	0.34	5.64
Guided GradCAM	0.00	15.61	0.00	0.05	0.06	0.44	5.12
Guided Backprop	0.00	15.66	0.00	0.05	0.06	0.39	5.47
L2I, RT=0.2	1.63	12.78	0.42	0.11	0.15	0.25	5.50
SHAP	0.00	15.79	0.00	0.05	0.06	0.43	5.24
L-MAC	36.25	1.15	23.50	0.20	0.42	0.47	4.71
L-MAC, FT, $\lambda_g = 4$	32.37	1.98	18.74	0.21	0.41	0.43	5.20
LMAC-TD, $\alpha = 1.00$ (ours)	66.00	2.62	22.39	0.42	0.87	0.86	10.50
LMAC-TD, $\alpha = 0.75$ (ours)	69.75	2.10	28.07	0.42	0.91	0.86	10.53
LMAC-TD, $\alpha = 0.00$ (ours)	46.50	5.55	11.86	0.42	0.86	0.80	10.88

Zero-Shot Audio Classification

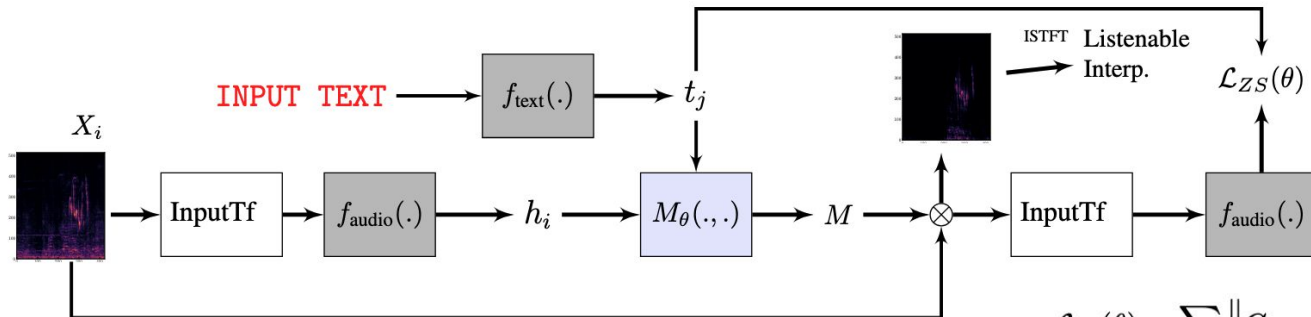
Contrastive Training



Zero Shot Classification

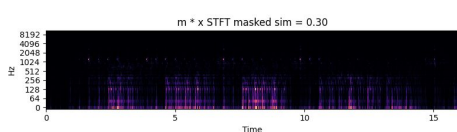
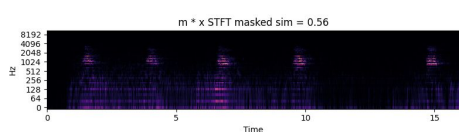
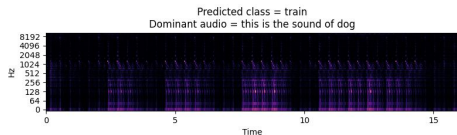
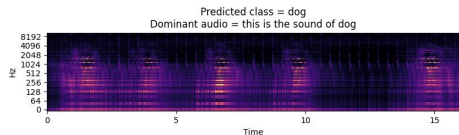
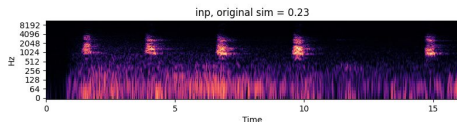
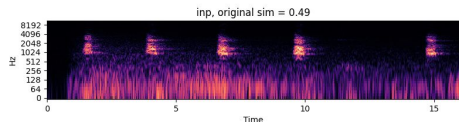


LMAC-ZS: Listenable Maps for Zero-Shot Audio Classifiers



$$\mathcal{L}_{ZS}(\theta) = \sum_{i,j} \left\| C_{i,j} - t_i^{\top} f_{\text{audio}} \left(M_{\theta}(t_i, h_j) \odot X_{\text{audio},j} \right) \right\| -$$

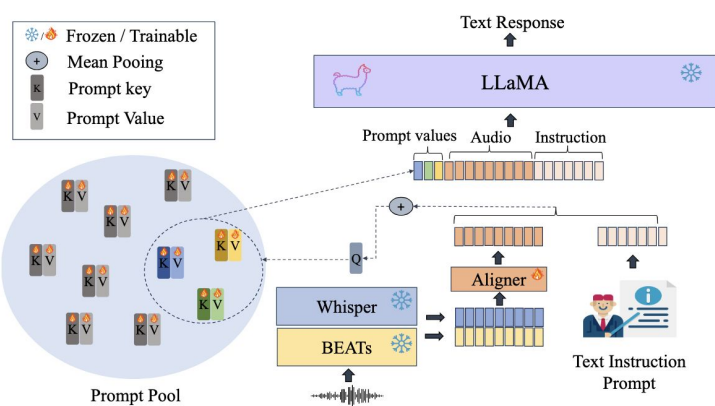
And still faithful..



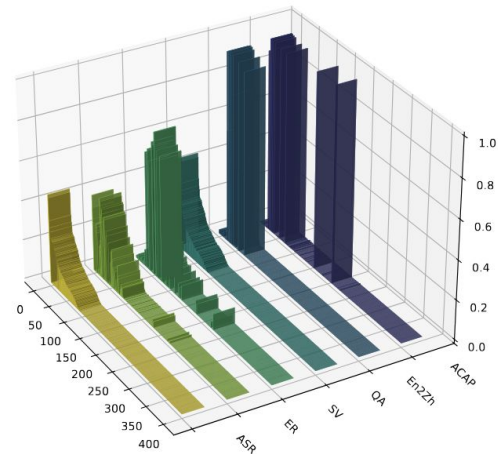
Metric	AI (↑)	AD (↓)	AG (↑)	FF (↑)	Fid-In (↑)	SPS (↑)	COMP (↓)	MM
<i>ZS classification on ESC50, Mel-Masking, 80.7% accuracy</i>								
Gradcam	2.90	45.85	1.01	0.28	0.19	0.71	9.52	0.15
GradCam++	8.45	35.07	3.19	0.50	0.39	0.41	10.32	0.35
SmoothGrad	0.50	52.76	0.12	0.024	0.036	0.301	10.52	0.039
IG	0.25	53.47	0.054	0.064	0.022	0.57	10.09	0.037
LMAC-ZS (CT)	29.00	12.25	12.93	0.49	0.80	0.78	9.40	0.14
LMAC-ZS (Full)	23.45	17.12	10.31	0.51	0.68	0.80	9.12	0.17
<i>ZS classification on ESC50, STFT-Masking, 78.9% accuracy</i>								
GradCam	20.30	23.75	7.77	0.78	0.58	0.72	11.54	0.14
GradCam++	32.50	8.97	7.95	0.79	0.84	0.41	12.41	0.35
SmoothGrad	6.95	32.75	2.85	0.78	0.47	0.53	11.98	0.0001
IG	16.10	21.51	6.05	0.79	0.65	0.74	11.58	0.0095
LMAC-ZS (CT)	37.40	7.43	11.26	0.78	0.86	0.50	12.29	0.11
LMAC-ZS (Full)	43.35	4.29	10.57	0.78	0.9	0.65	11.86	0.1

Ongoing Work: Explainable MLSP

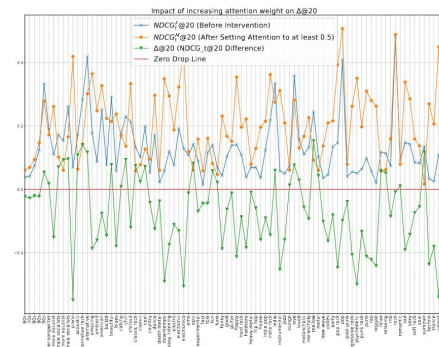
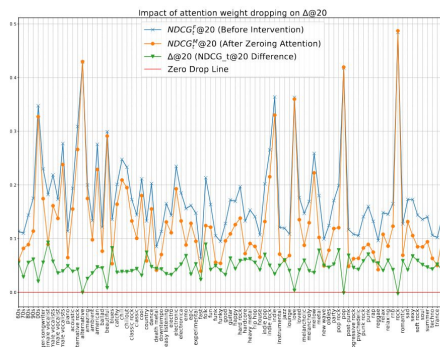
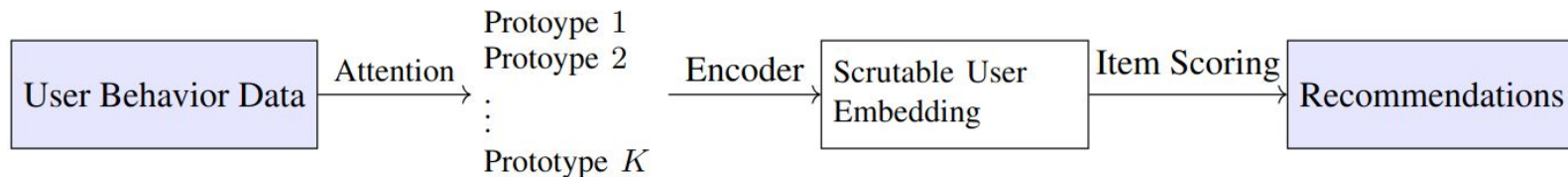
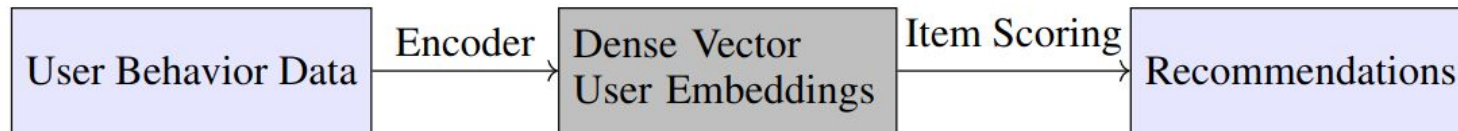
- Posthoc methods are great! Works (faithful / listenable) on unimodal and multimodal models.
 - But can we create explainable models?
 - Explainable LLMs? (Interspeech 2025)



- Posthoc explanation methods for Audio Language Models.
- Controllable Recommendation Systems



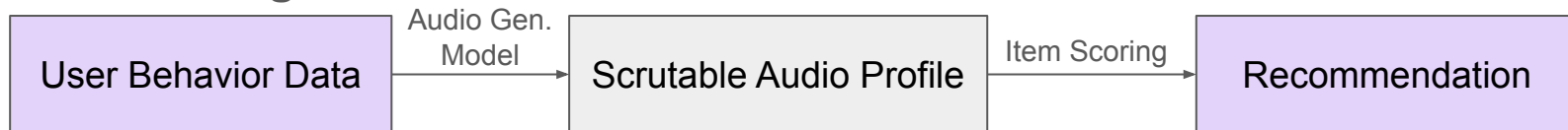
Scrutable Audio / Multimodal RecSys



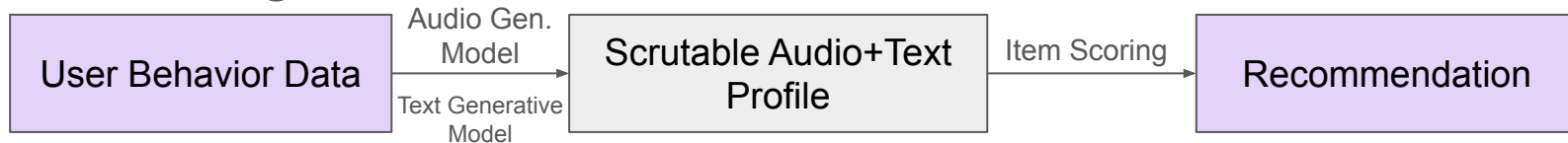
F. Oncel, E. Penalzo, H. Wu, S. Gupta, M. Ravanelli, L. Charlin., C. Subakan; Audio Prototype Network for Controllable Music Recommendation, MLSP 2025

Scrutable Audio / Multimodal RecSys

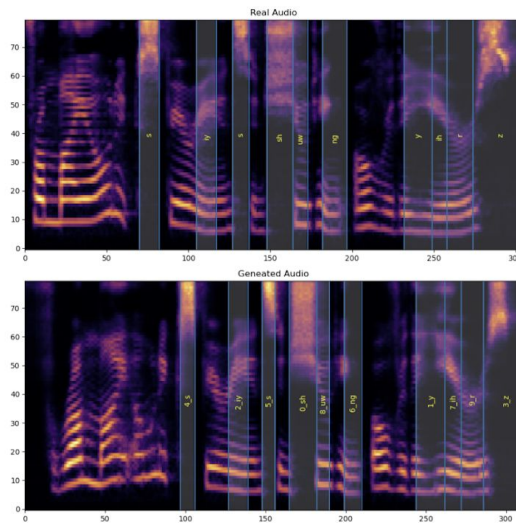
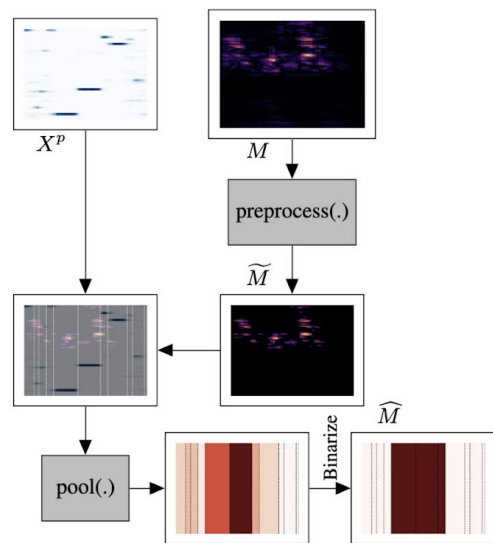
- Generating Personalized Controllable Audio Summaries



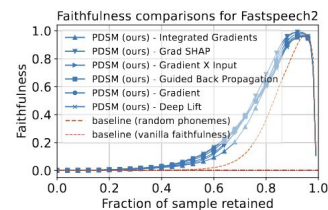
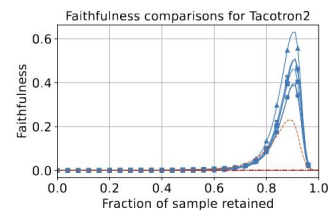
- Generating Personalized Controllable Multi-Modal Summaries



Explainable Detection of AI Generated Voice



Faithful!



Q: More general discretization techniques to improve understandability.