Towards Human-Centric Multimodal Generative Audio AI

You Zhang, PhD

Senior Researcher, Dolby Laboratories

Neil.Zhang@Dolby.com



Télécom Paris
Audio ADASP – Machine Learning S2A
24 April 2026

- Education

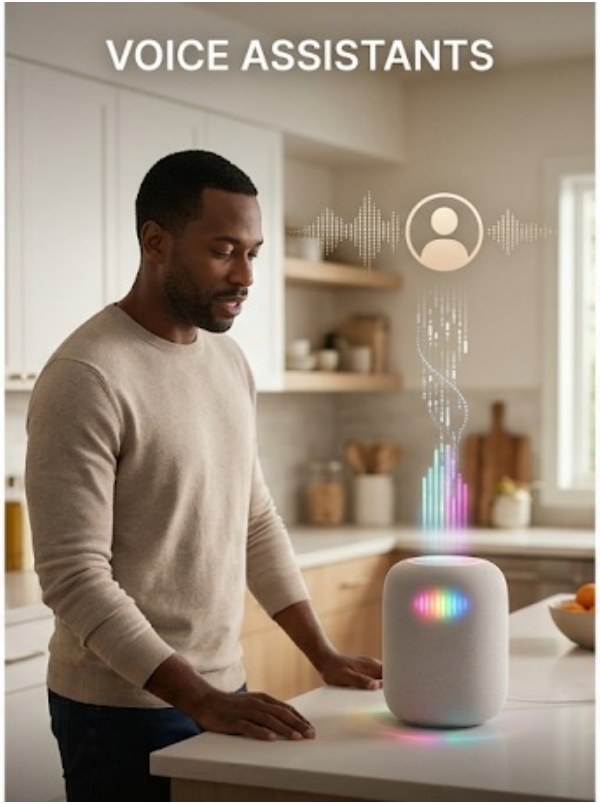
2019	2021	2026
BE	MS	PhD

- Research Interests
  - Spatial Audio AI
  - Multimodal Understanding and Generation
  - Deepfake Detection

- Selected Awards
  - IEEE WASPAA 2025 — Best Student Paper
  - IEEE Signal Processing Society — SPS Scholarship
  - NIJ National Institute of Justice — Graduate Research Fellowship
  - ICASSP 2023 — Rising Star in Signal Processing

Generative audio AI is transforming modern experiences …







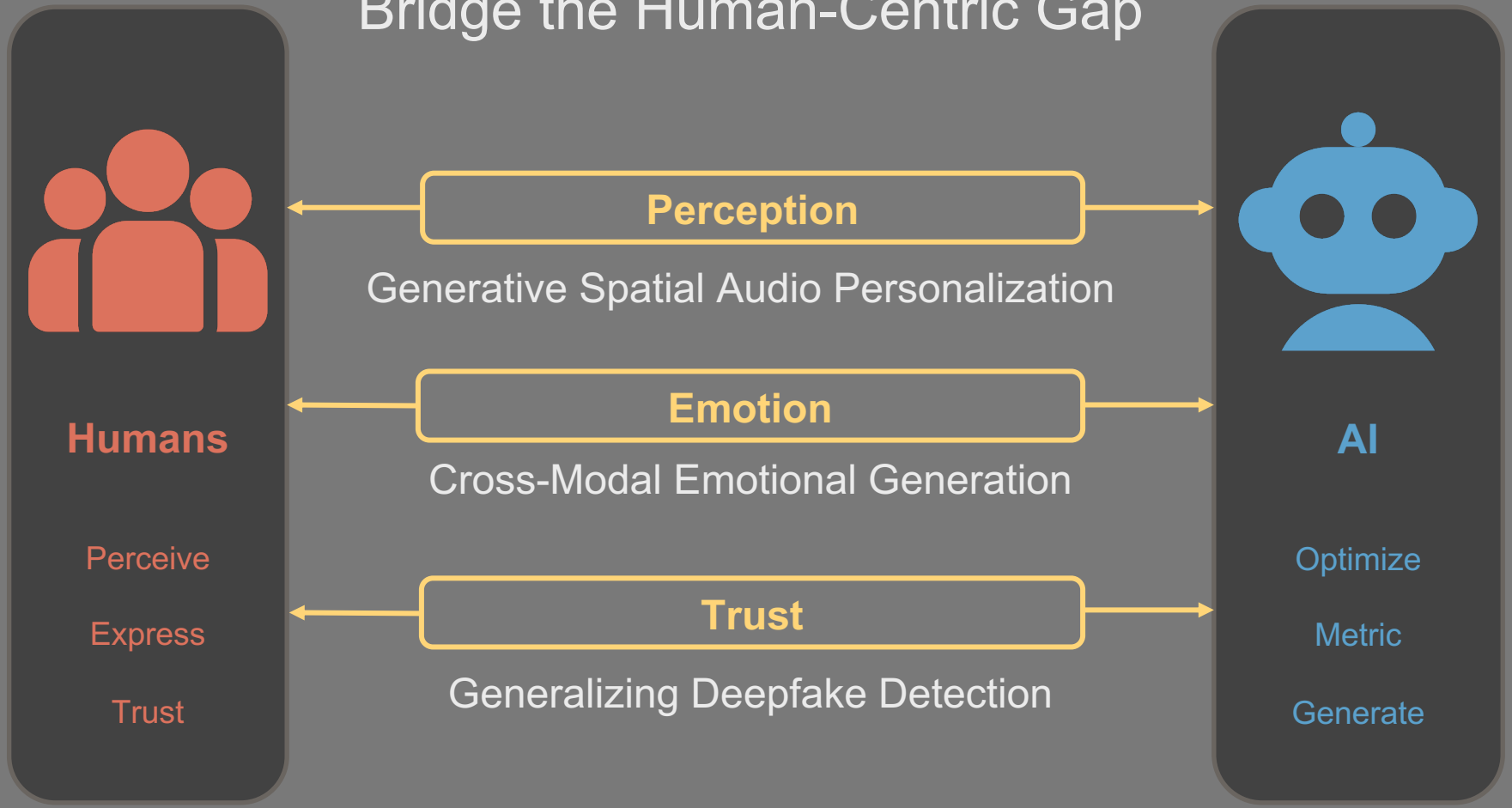
Generative audio AI is NOT fully human-centric …







Bridge the Human-Centric Gap



Bridge the Human-Centric Gap



Humans

Perceive

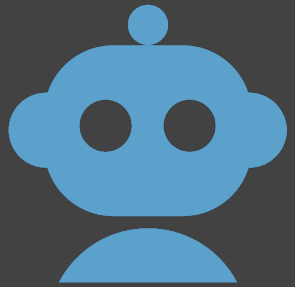
Express

Trust

Perception

Generative Spatial Audio Personalization

How should audio be represented so that models remain structured and perceptually meaningful?



AI

Optimize

Metric

Generate

University of Rochester

ICASSP 2023
4 - 10 JUNE,
RHODES ISLAND, GREECE

HRTF Field: Unifying Measured HRTF Magnitude Representation with Neural Fields

You Zhang, *Yuxiang Wang, Zhiyao Duan*

University of Rochester



Top 3% Accepted Papers

Background: Spatial Effects and Sound Localization

Humans localize sound sources by processing the differences between sounds received by their two ears.

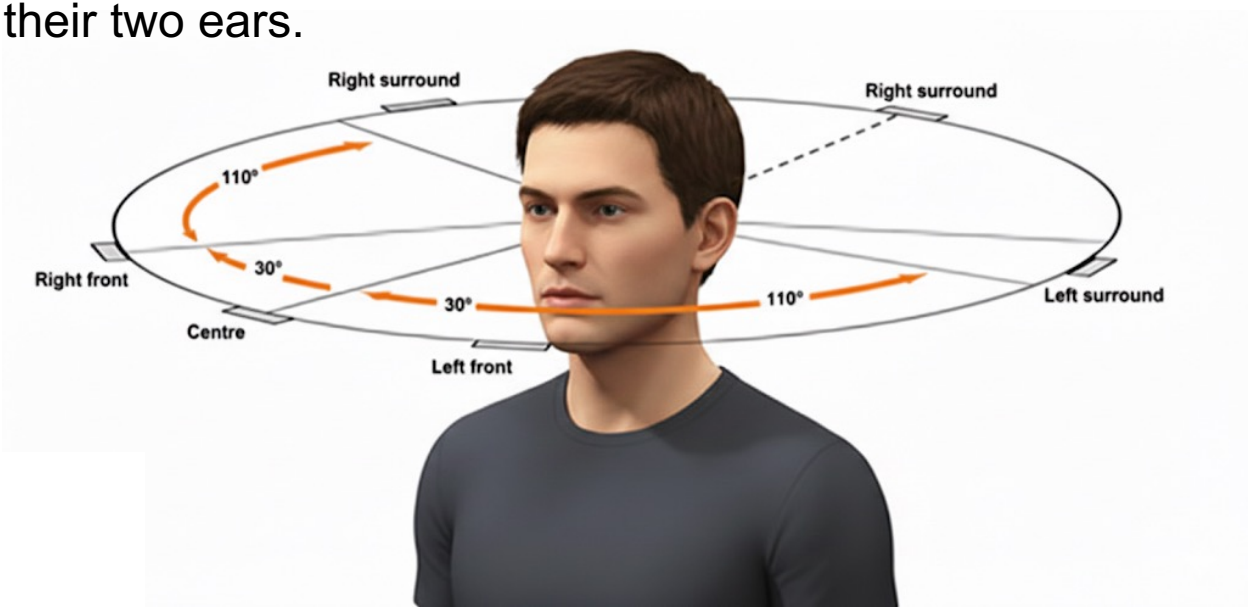
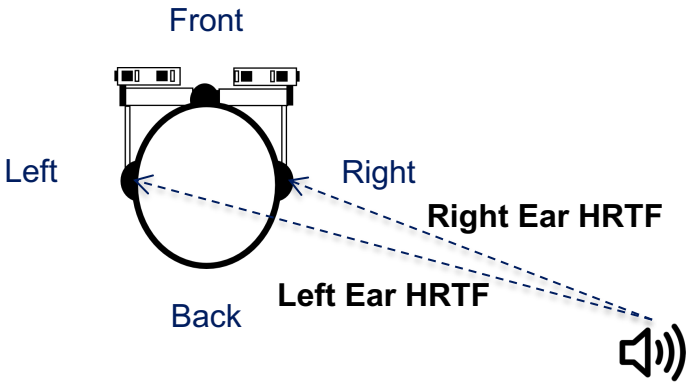


Figure adapted from https://www.soundonsound.com/reviews/mp3-surround

Background: Head-Related Transfer Function (HRTF)

HRTF models the **acoustic filtering** effect of a listener's **head, ears, and torso** to enable 3D sound localization.



Left ear HRTF magnitudes (dB) of the midsagittal plane of one subject

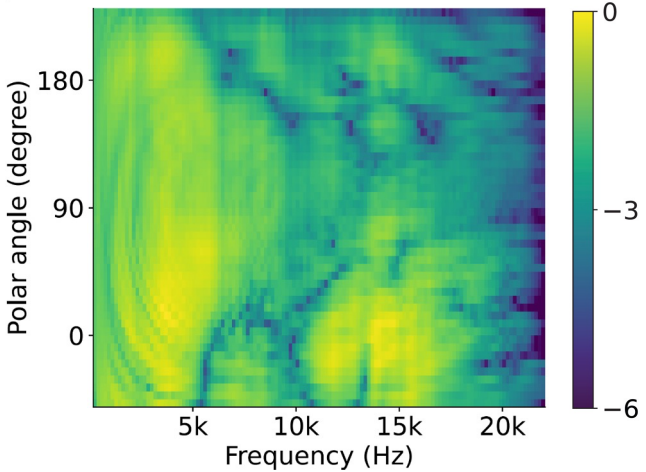


Figure from [Zhang+2023]

HRTF is unique to each person due to differences in ear, head, and torso shape.

You Zhang, Yuxiang Wang, and Zhiyao Duan. "HRTF field: Unifying measured HRTF magnitude representation with neural fields." *ICASSP* 2023.

Background: Measure HRTFs

- An anechoic room
- Multiple loudspeakers on motorized arc
- Two microphones
- Head motion control

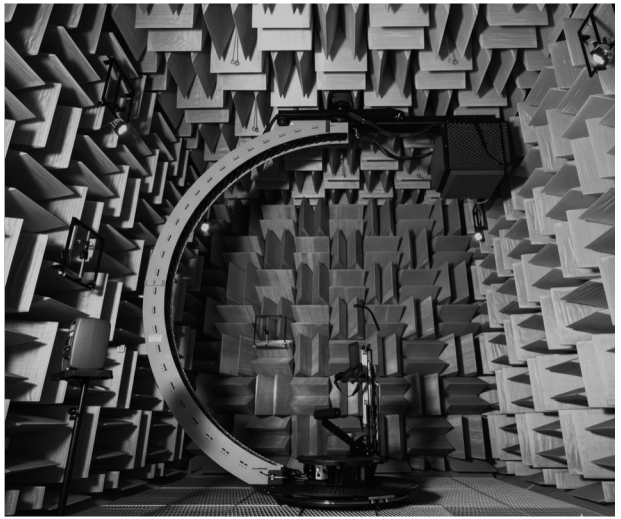
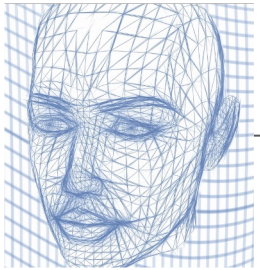


Figure from https://facebookresearch.github.io/SS2_HRTF/

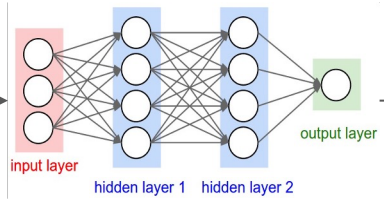
Time-consuming & Resource-intensive!

Motivation: Personalizing HRTF with Machine Learning

Leverage measured data for personalized HRTF prediction

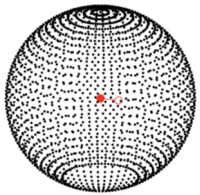


Human physical geometry

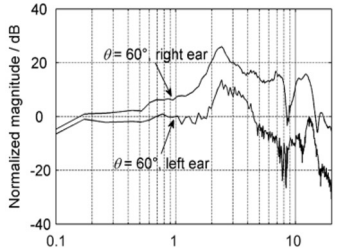


HRTF prediction network

Representation of HRTFs



HRTFs at various spatial locations (of arbitrary spatial sampling schemes)

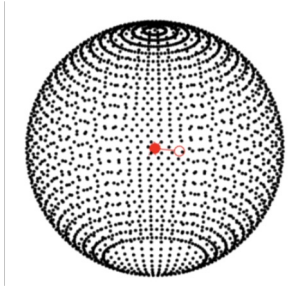


HRTFs at a particular position

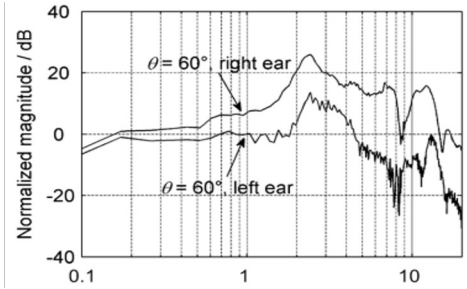
Assumption: Many HRTF characteristics are shared across individuals.

Challenge 1: HRTF is High-Dimensional

For each subject, ear, and spatial location, HRTF is a function of frequency.



HRTFs at various spatial locations (of arbitrary spatial sampling schemes)



HRTFs at a particular position

$$\boldsymbol{x} \in \mathbb{R}^{L \times F \times 2}$$

L: number of locations (~1000)

F: number of frequency bins (~128)

2: left and right ear

1000 x 128 x 2 = 256,000. *A huge number!*

Challenge 2: Small Datasets and Different Setups

Existing measured HRTF databases each only contain dozens of subjects.

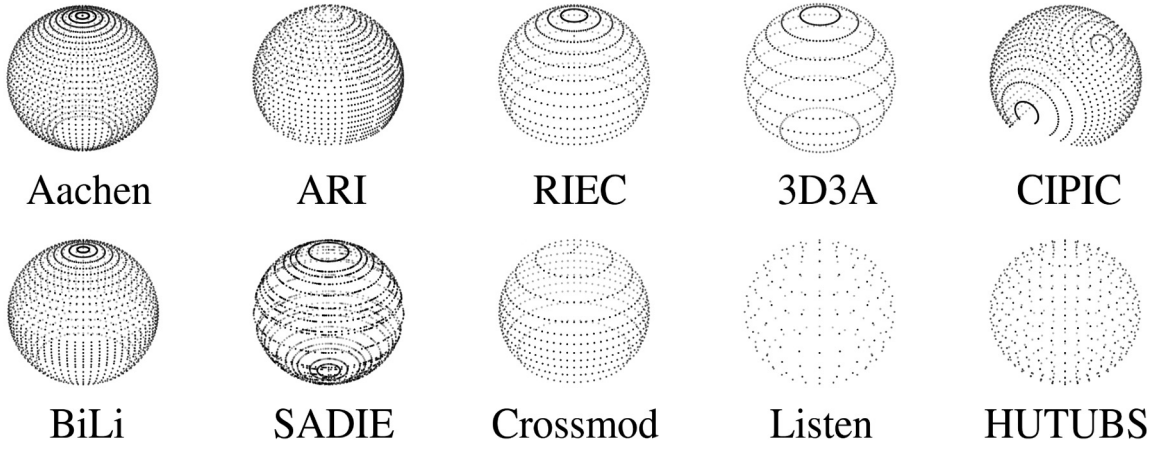
Name	# Subjects	# Locations
3D3A [29]	38	648
Aachen [30]	48	2304
ARI	97	1550
BiLi [31]	52	1680
CIPIC [4]	45	1250
Crossmod	24	651
HUTUBS [17]	96	440
Listen	50	187
RIEC [32]	105	865
SADIE II [2]	18	2818

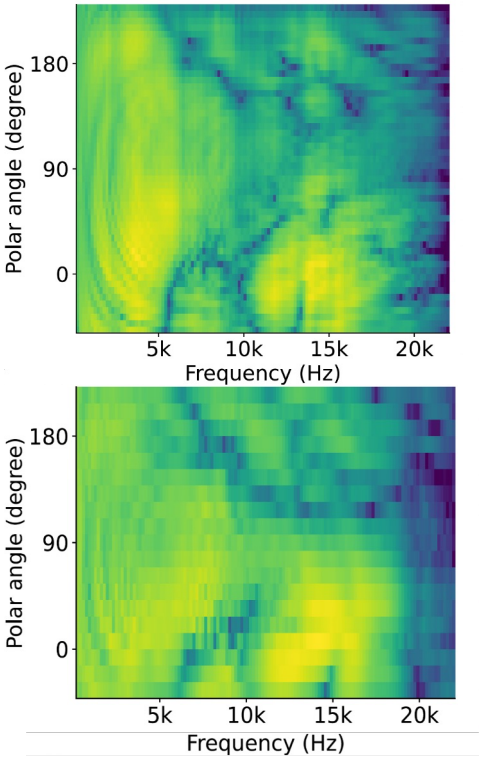
SONICOM [Poole+2025] (Expanded to 300 subjects in 2025)

You Zhang, Yuxiang Wang, and Zhiyao Duan. "HRTF field: Unifying measured HRTF magnitude representation with neural fields." *ICASSP* 2023.
Katarina C. Poole, Julie Meyer, Vincent Martin, Rapolas Daugintis, Nils Marggraf-Turley, Jack Webb, Ludovic Pirard, Nicola La Magna, Oliver Turvey, and Lorenzo Picinali. "The extended SONICOM HRTF dataset and spatial audio metrics toolbox." *Forum Acusticum 2025*.

Challenge 3: Different Spatial Sampling Schemes

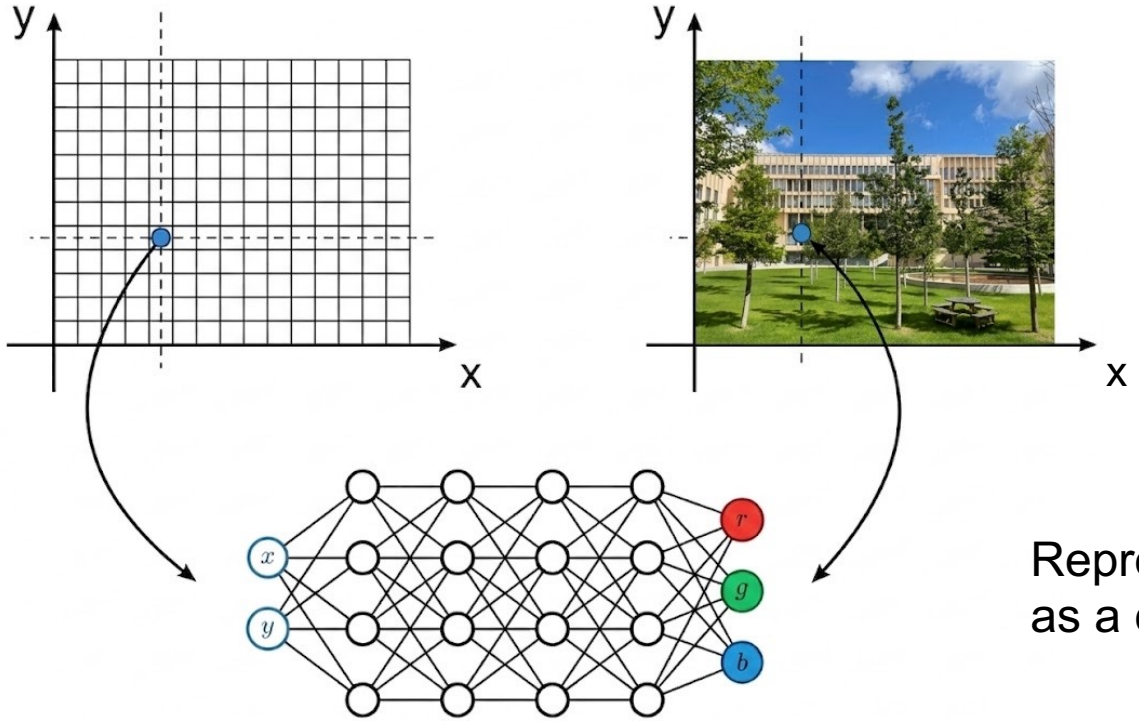
The source location grids used in HRTF databases differ from one to another, making cross-dataset learning difficult.





You Zhang, Yuxiang Wang, and Zhiyao Duan. "HRTF field: Unifying measured HRTF magnitude representation with neural fields." *ICASSP* 2023.

Preliminary: Neural Fields



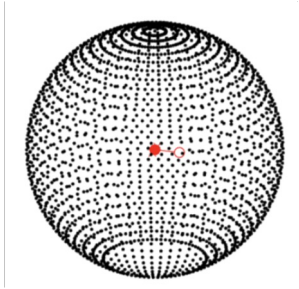
Representing discrete data as a continuous function

Figure adapted from [Skorokhodov+2021] and Telecom Paris wikipedia

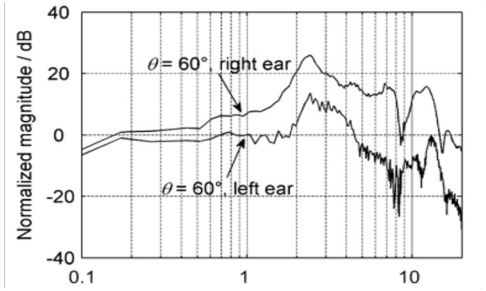
Ivan Skorokhodov, Savva Ignatyev, and Mohamed Elhoseiny. “Adversarial generation of continuous images.” *CVPR* 2021.

Method: HRTF Field

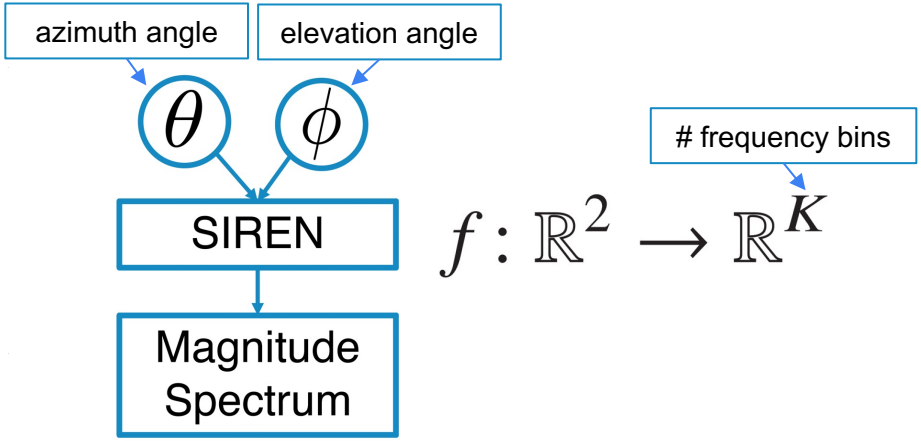
Represent a single subject's HRTFs with a neural field



HRTFs at various spatial locations (of arbitrary spatial sampling schemes)



HRTFs at a particular position



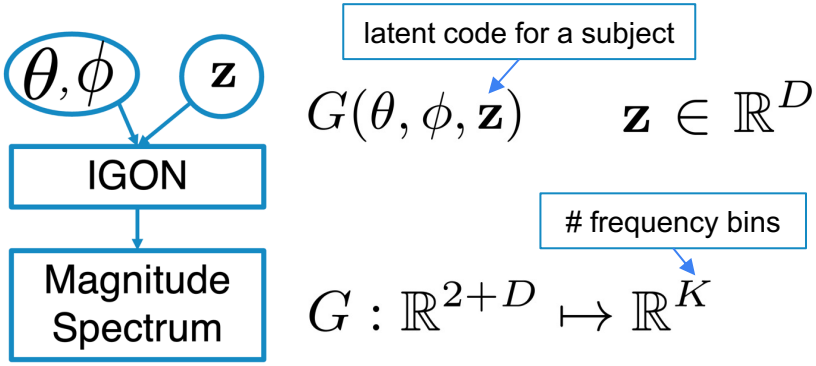
SIREN: a multi-layer perceptron with sine activation functions

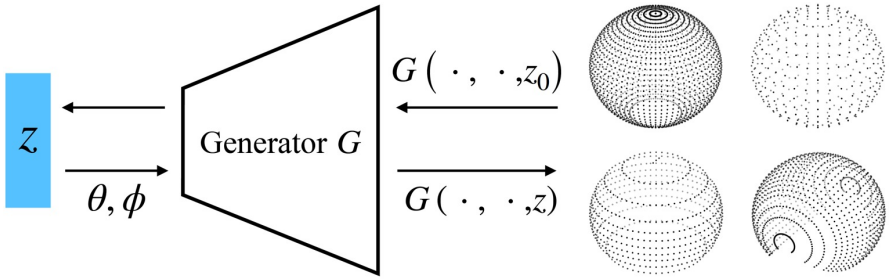
HRTF Field decouples representation from sampling schemes.

Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. "Implicit neural representations with periodic activation functions." *NeurIPS* 2020.

Method: HRTF Field (Cont'd)

Learning HRTF representations across subjects





IGON: implicit gradient origin network that uses SIREN architecture

Enables grid-agnostic mix-database training and cross-database evaluation

You Zhang, Yuxiang Wang, and Zhiyao Duan. "HRTF field: Unifying measured HRTF magnitude representation with neural fields." *ICASSP* 2023.
Sam Bond-Taylor, and Chris G. Willcocks. "Gradient origin networks." *ICLR* 2021.

Impact

10.1109/ICASSPW62465.2024.10626510

NEURAL STEERER: NOVEL STEERING VECTOR SYNTHESIS WITH A CAUSAL NEURAL FIELD OVER FREQUENCY AND DIRECTION

Diego Di Carlo[1,2] *Aditya Arie Nugraha*[1,2] *Mathieu Fontaine*[3,1] *Yoshiaki Bando*[4,1] *Kazuyoshi Yoshii*[2,1]

[1] Center for Advanced Intellig
[2] Graduate School of Infor
[3] LTCI, Télécom Paris, Instit
[4] National Institute of Advanced Indust

IEEE Signal Processing Society IEEE Open Journal of Signal Processing

Received 25 February 2025; revised 13 June 2025; accepted 2 October 2025. Date of publication 31 October 2025; date of current version 25 November 2025. The review of this article was arranged by Associate Editor M. Geronazzo.

Digital Object Identifier 10.1109/OJSP.2025.3627073

A process for calibrating HRTFs based on differentiable implicit representations and domain adversarial learning

1st Thiago Lobato
HEAD acoustics GmbH
Herzogenrath, Germany
thiago.lobato@head-acoustics.com

2nd Roland Sottek
HEAD acoustics GmbH
Herzogenrath, Germany
roland.sottek@head-acoustics.com

SuDaField: Subject- and Dataset-Aware Neural Field for HRTF Modeling

YOSHIKI MASUYAMA [1] (Member, IEEE), GORDON WICHERN [1] (Member, IEEE), FRANÇOIS G. GERMAIN [1] (Member, IEEE), CHRISTOPHER ICK[1,2], AND JONATHAN LE ROUX [1] (Fellow, IEE

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139 USA
[2]Music and Audio Research Laboratory, New York University, Brooklyn, NY 10012 USA

CORRESPONDING AUTHOR: YOSHIKI MASUYAMA (e-mail: masuyama@merl.com).

RANF: Neural Field-Based HRTF Spatial Upsampling With Retrieval Augmentation and Parameter Efficient Fine-Tuning

YOSHIKI MASUYAMA [1] (Member, IEEE), GORDON WICHERN [1] (Member, IEEE), FRANÇOIS G. GERMAIN [1] (Member, IEEE), CHRISTOPHER ICK[1,2], AND JONATHAN LE ROUX [1] (Fellow, IEEE)

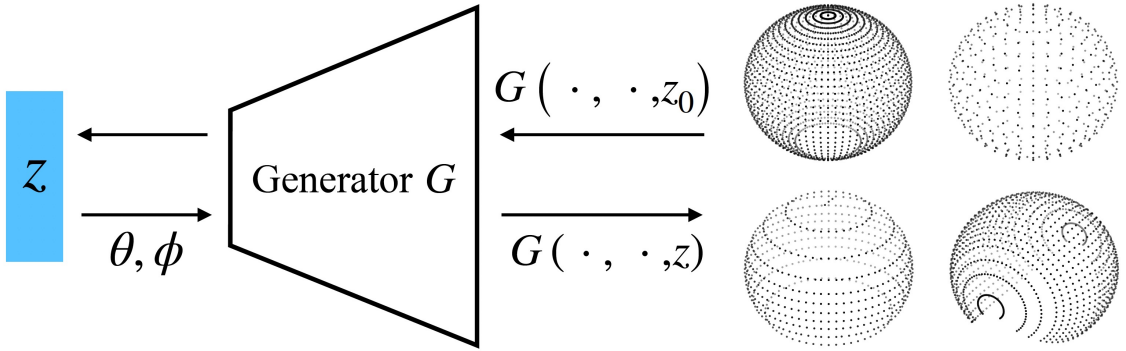
[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139 USA
[2]Music and Audio Research Laboratory, New York University, Brooklyn, NY 10012 USA

CORRESPONDING AUTHOR: YOSHIKI MASUYAMA (email: masuyama@merl.com).

Inspired the winner of 2024 IEEE LAP Challenge

Summary: Generative Framework for HRTF Representation

- **Goal**: Merge databases with different spatial sampling schemes
- **Approach**: Introduce continuous generative neural fields to model HRTF
- **Contributions**: For the **first time**, it enables mix-database training and cross-database evaluation



Towards Perception-Informed Latent HRTF Representations

You Zhang [1,2], *Andrew Francl* [2], *Ruohan Gao* [3], *Paul Calamia* [2], *Zhiyao Duan* [1], *Ishwarya Ananthabhotla* [2]

[1] University of Rochester
[2] Meta Reality Labs Research
[3] University of Maryland

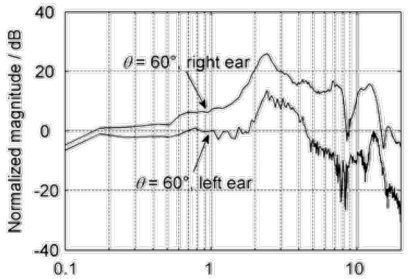
IEEE WASPAA 2025



Best Student Paper

Motivation: Representation not Perceptually Aligned

Most **existing** models are trained and evaluated with **spectral reconstruction**.



Spectral reconstruction





Perceptually plausible HRTF

If two HRTFs are close in spectral distance, are they close perceptually?

Preliminary: HRTF Perception

Perceptual benefits of individual HRTF:

- Reduced **Coloration** (less unwanted spectral distortion)
- Improved **Externalization** (sound appears outside the head)
- Enhanced **Localization** (accurately placing sounds in 3D space)

Majdak, Piotr, Bruno Masiero, and Janina Fels. “Sound localization in individualized and non-individualized crosstalk cancellation systems.” *JASA* 2013.
Brinkmann, Fabian, Alexander Lindau, and Stefan Weinzierl. “On the authenticity of individual dynamic binaural synthesis.” *JASA* 2017.
Jenny, Claudia, and Christoph Reuter. "Usability of individualized head-related transfer functions in virtual reality: Empirical study with perceptual attributes in sagittal plane sound localization." *JMIR Serious Games* 2020.

Preliminary: Computational Auditory Modeling

Coloration: **PBC** [McKenzie+2022] ✓ Differentiable

Externalization: **AEP** [Baumgartner&Majdak2021] ✗ Not differentiable

Localization: **DRMSP** [Barumerli+2023] ✗ Not differentiable



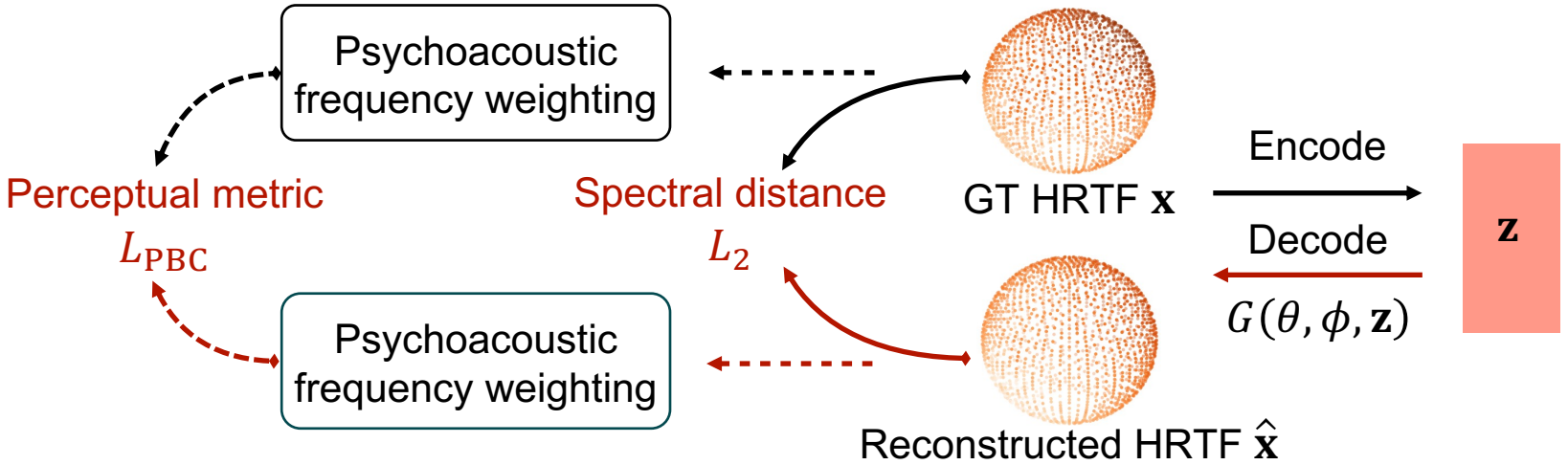
Figure from amtoolbox.org

Thomas McKenzie, Cal Armstrong, Lauren Ward, Damian T. Murphy, and Gavin Kearney. "Predicting the colouration between binaural signals." *Applied Sciences* 2022.
Robert Baumgartner, and Piotr Majdak. "Decision making in auditory externalization perception:model predictions for static conditions." *Acta Acustica* 2021.
Roberto Barumerli, Piotr Majdak, Michele Geronazzo, David Meijer, Federico Avanzini, and Robert Baumgartner. "A Bayesian model for human directional localization of broadband static sound sources." *Acta Acustica* 2023.

Method: Aligning with Perception-Informed Space

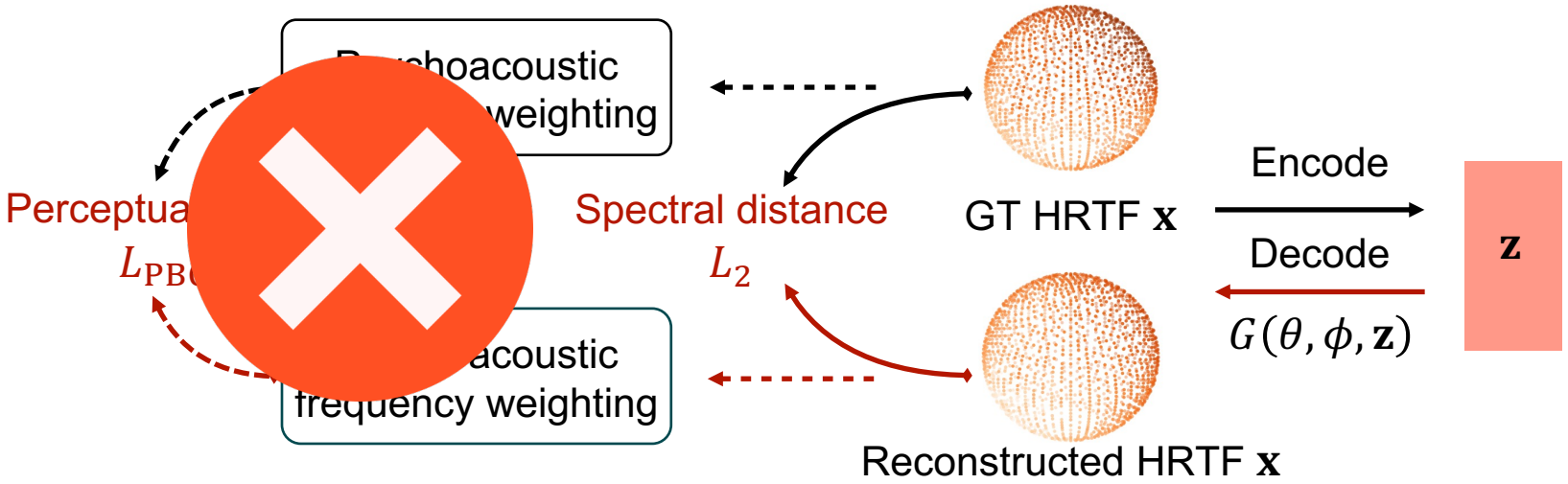
If the perceptual metric is differentiable, a direct perceptual loss can be incorporated.

- This approach applies only to PBC, not to AEP or DRMSP.



Method: Aligning with Perception-Informed Space (Cont'd)

If the perceptual metric is NOT differentiable (AEP, DRMSP)



Preliminary: Metric multi-dimensional scaling (MMDS)

Optimize a new space that preserves the pairwise distances.

Pairwise distance matrix

0	0.4	0.5	0.6	0.9
0.4	0	0.1	0.1	0.4
0.5	0.1	0	0.1	0.5
0.6	0.1	0.1	0	0.6
0.9	0.4	0.5	0.6	0

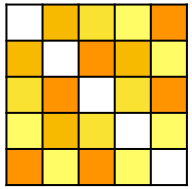
New space



Figure adapted from https://youtu.be/VKSJayDi_IQ

Method: Aligning with Perception-Informed Space (Cont'd)

If the perceptual metric is NOT differentiable (AEP, DRMSP)



Pairwise perceptual distance matrix $\mathbf{M}$

Metric multi-dimensional scaling (MMDS)

$\mathbf{z}_{\mathrm{MDS}}$

MMDS supervision

L_{Align}

This also applies to **differentiable metrics** (PBC).



GT HRTF $\mathbf{x}$

Spectral distance

L_2

Encode

Decode

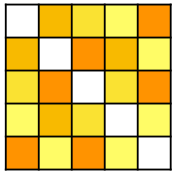
$G(\theta, \phi, \mathbf{z})$

$\mathbf{z}$



Reconstructed HRTF $\hat{\mathbf{x}}$

Method: Overall Pipeline



Pairwise perceptual distance matrix $\mathbf{M}$

Metric multi-dimensional scaling (MMDS)

$\mathbf{z}_{\text{MDS}}$

MMDS supervision

L_{Align}

Encode

$\mathbf{z}$

Decode

$G(\theta, \phi, \mathbf{z})$

GT HRTF $\mathbf{x}$

Reconstructed HRTF $\hat{\mathbf{x}}$

Spectral distance

L_2

Psychoacoustic frequency weighting

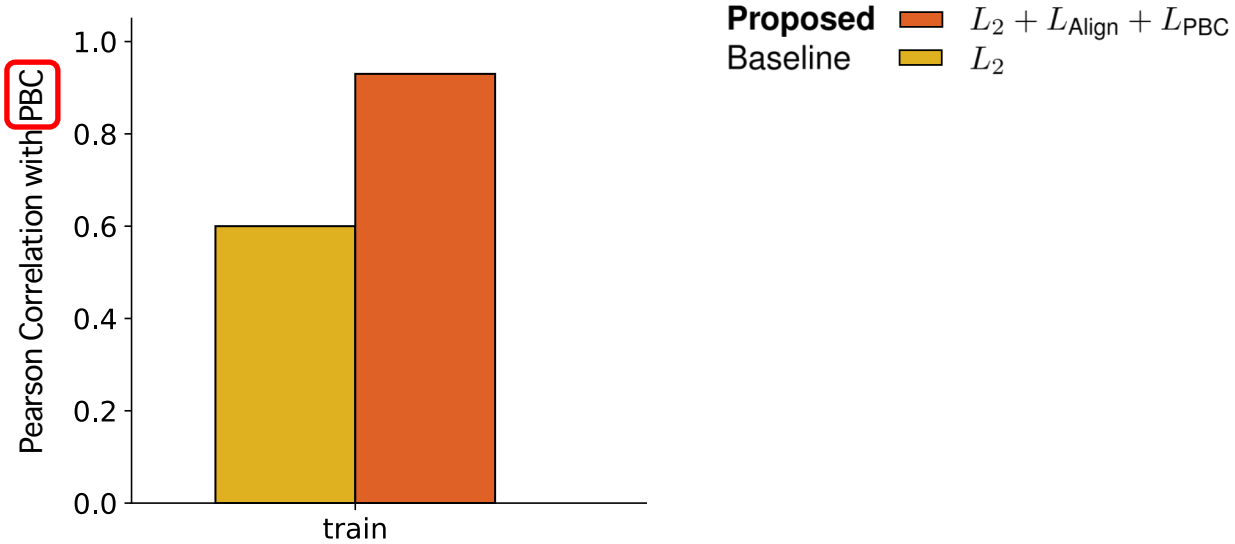
Perceptual metric

L_{PBC}

Psychoacoustic frequency weighting

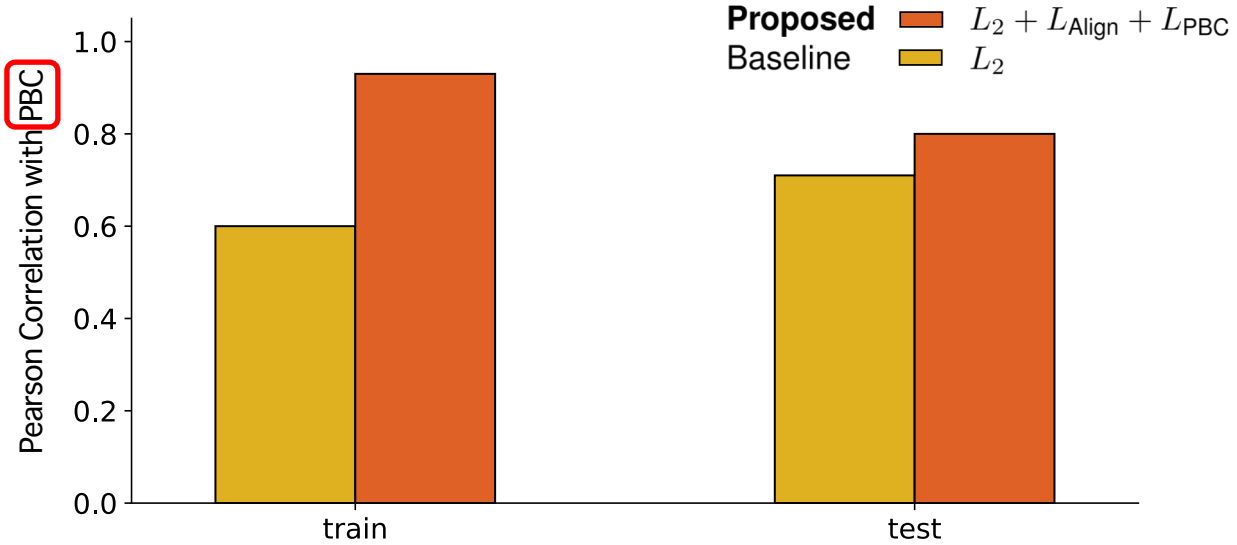
Results: Objective Perceptual Correlation Evaluation on PBC

- Our proposed method **achieves better alignment** with perception-informed space.
- The perceptual correlation learned in training transfer to test subjects (unseen).



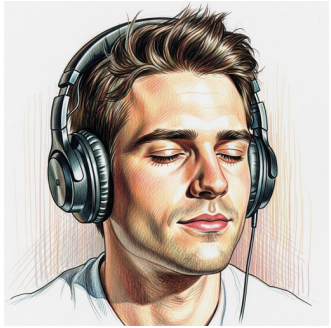
Results: Objective Perceptual Correlation Evaluation on PBC

- Our proposed method **achieves better alignment** with perception-informed space.
- The perceptual correlation learned in training transfer to test subjects (unseen).



Results: Personalized HRTF Selection

For each of the 13 test (unseen) subjects, we select the nearest HRTFs from the 65 training subjects, based on the learned latent representations.



	Method	Metric	Spectral Distance (dB)↓
		PBC↓	
Baseline	L_2	1.30	**2.07**
Ours	$L_2 + L_{\text{Align}} + L_{\text{PBC}}$	**1.21**	2.11
		DRMSP↓	
Baseline	L_2	4.21	**2.07**
Ours	$L_2 + L_{\text{Align}}$	**3.20**	2.12

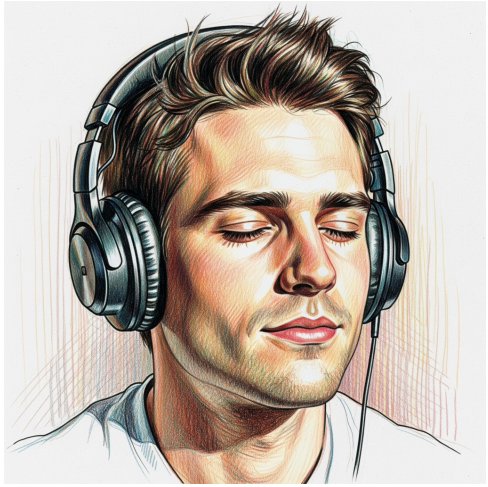
HRTFs selected by our proposed method yield lower perceptual distances.

Perceptually better!

You Zhang, Andrew Francl, Ruohan Gao, Paul Calamia, Zhiyao Duan, and Ishwarya Ananthabhotla. "Towards Perception-Informed Latent HRTF Representations." *WASPAA* 2025.

Application: Virtual Spatial Audio Generation

Personalized HRTFs have wide applications to deliver immersive 3D sound.



Headphones



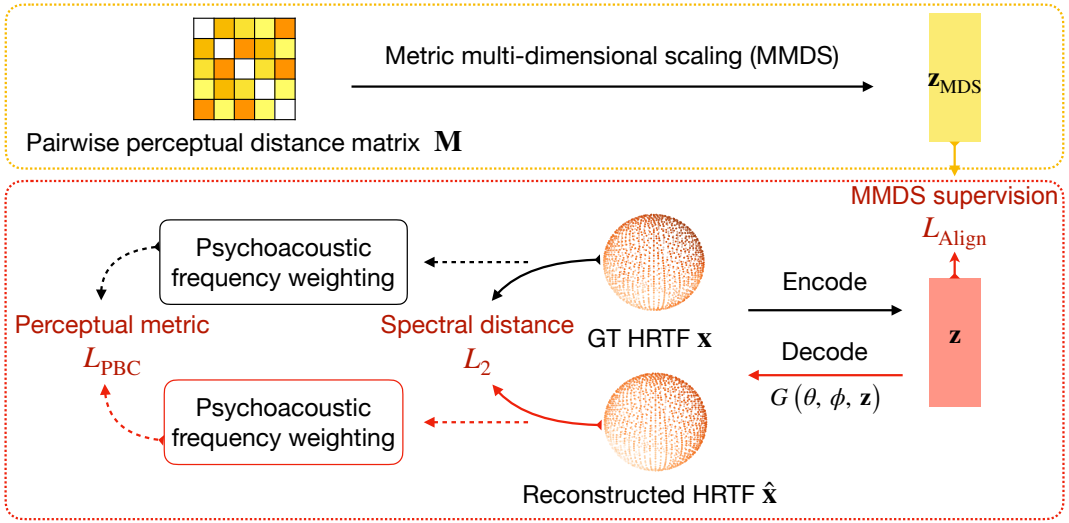
AR smart glasses



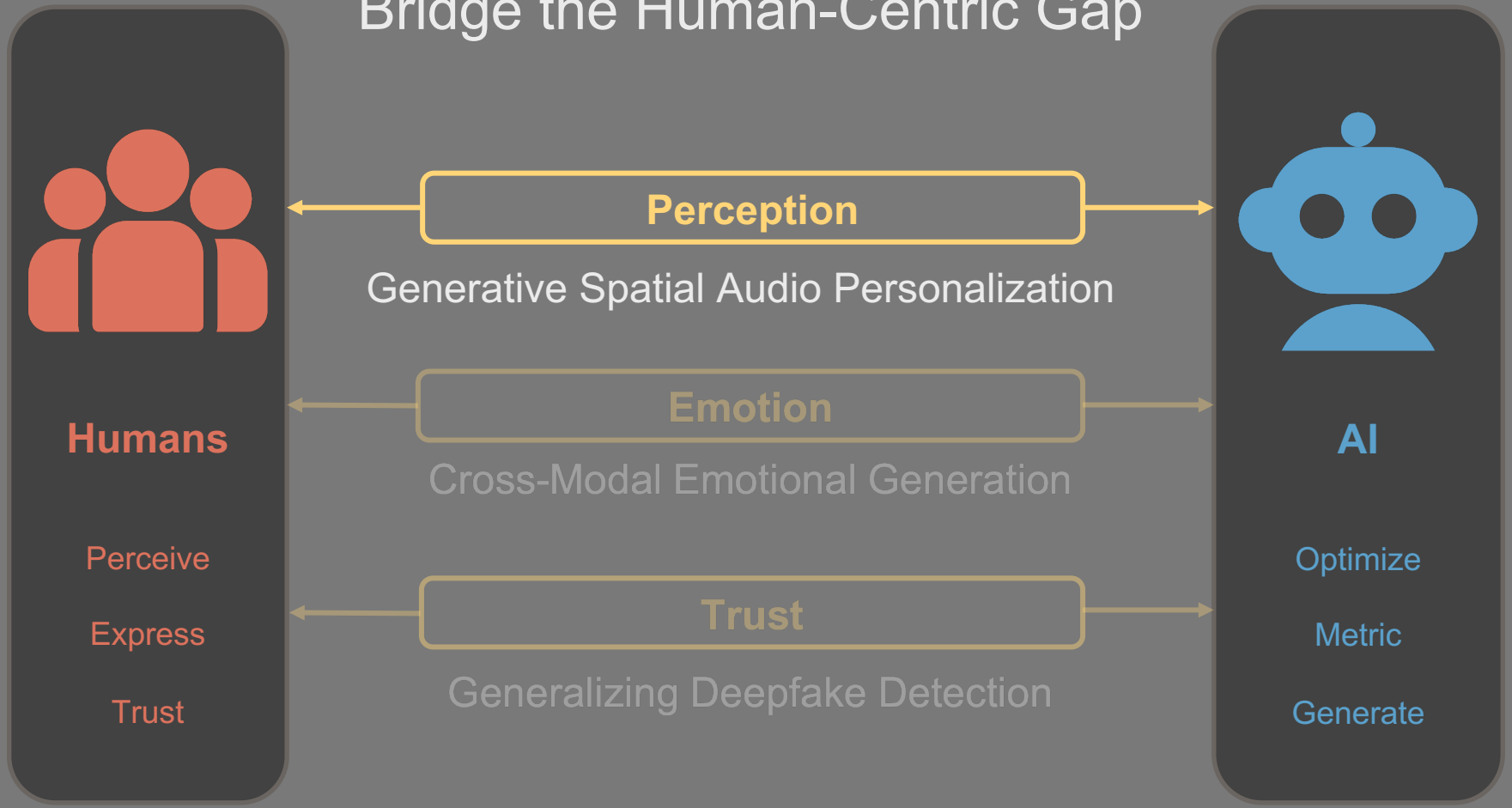
VR headsets

Summary: Perception-Informed Generative Model for HRTF

- **Goal**: Perceptually grounded HRTF personalization
- **Approach**: Generative neural fields for HRTF modeling with constraints on perceptual alignment using perceptual loss and MMDS
- **Contributions**: Better alignment with human perception in HRTF personalization



Bridge the Human-Centric Gap



Bridge the Human-Centric Gap



Humans

Perceive

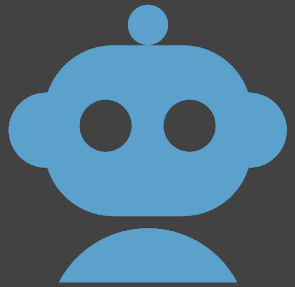
Express

Trust

Emotion

Cross-Modal Emotional Generation

How can emotion be represented and controlled across audio and visual modalities for expressive generation?



AI

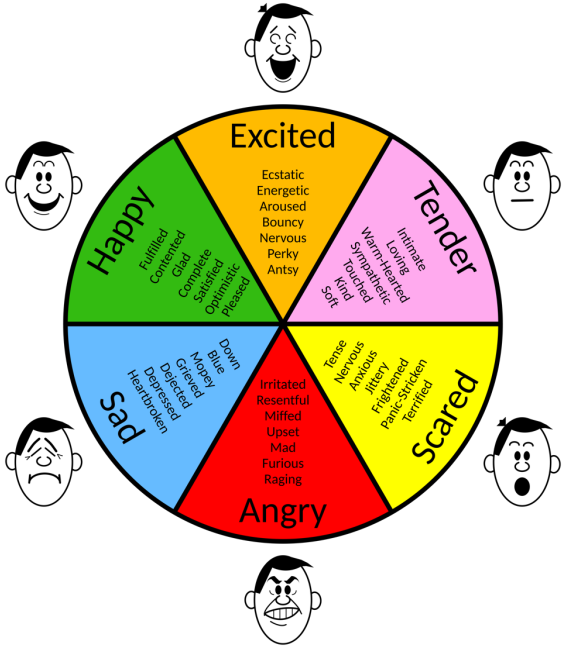
Optimize

Metric

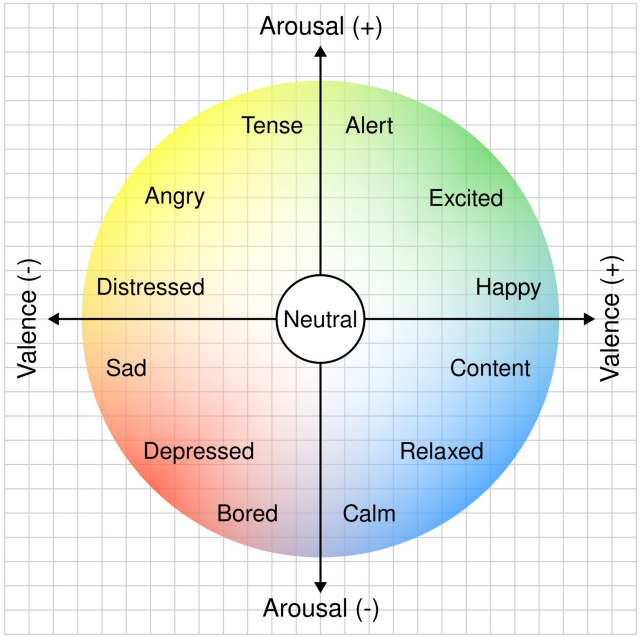
Generate

Background: Emotion Representation

- Categorical

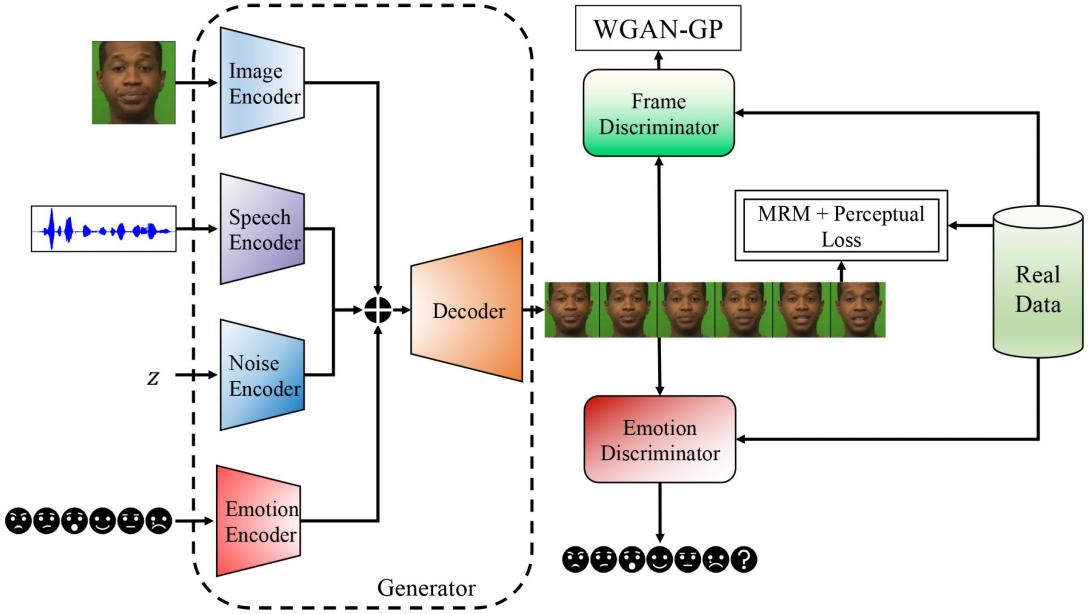


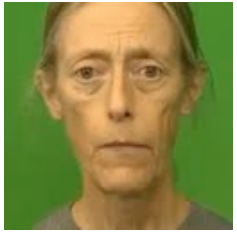
- Dimensional (Continuous)



Figures from Wikipedia https://en.wikipedia.org/wiki/Emotion

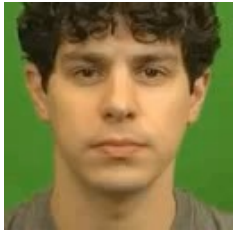
Emotional talking face generation



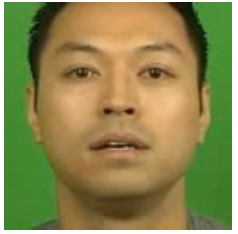




A: happy, V: disgusted



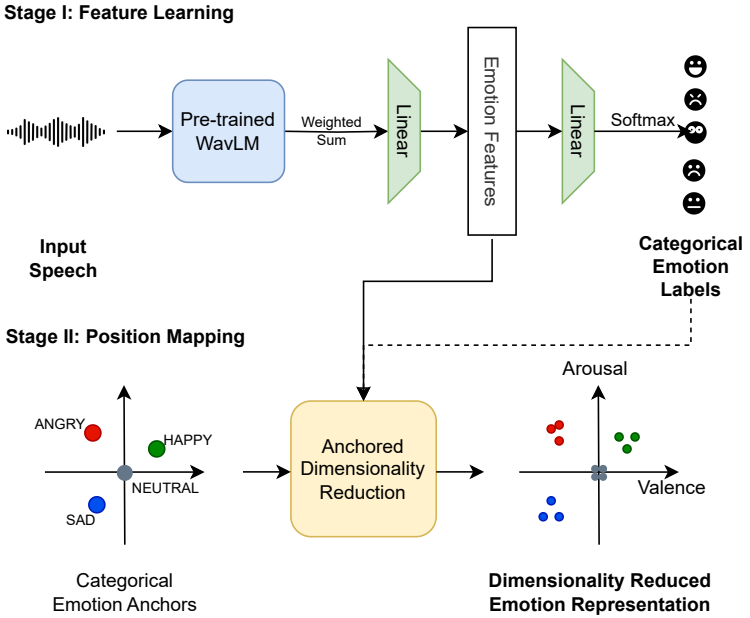
A: neutral, V: happy



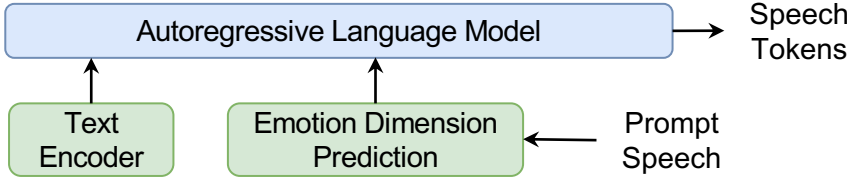
This is **the first** research that introduces explicit emotion control into talking face generation.
It enables the study of incongruency in human emotion perception.

Controllable Generation with Fine-Grained Emotion

- Emotional Dimension Prediction



- Emotional Text-to-Speech



No emotion labels required during TTS training

Demo:

Input Text: "After his visit, I told Esprit to take me to the Palais Royal, and I left him at the gates."

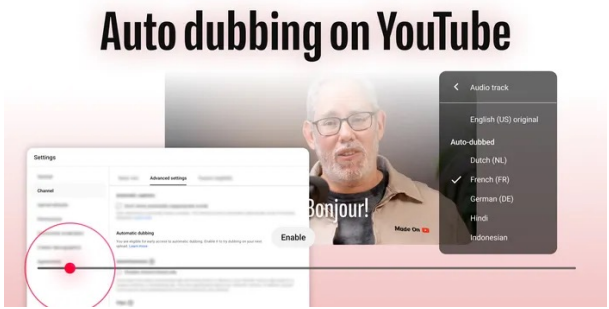
Happy: Excited:

Enables generating a broad spectrum of human emotions

Enting Zhou, **You Zhang**, and Zhiyao Duan. "Learning Arousal-Valence Representation from Categorical Emotion Labels of Speech." *ICASSP* 2024.
Kun Zhou, **You Zhang**, Dianwen Ng, Shengkui Zhao, Hao Wang, and Bin Ma. “Emotional Dimension Control in Language Model-Based Text-to-Speech: Spanning a Broad Spectrum of Human Emotions.” *ICASSP* 2026.

Applications: Multimodal Emotional Rendering

Entertainment



Healthcare

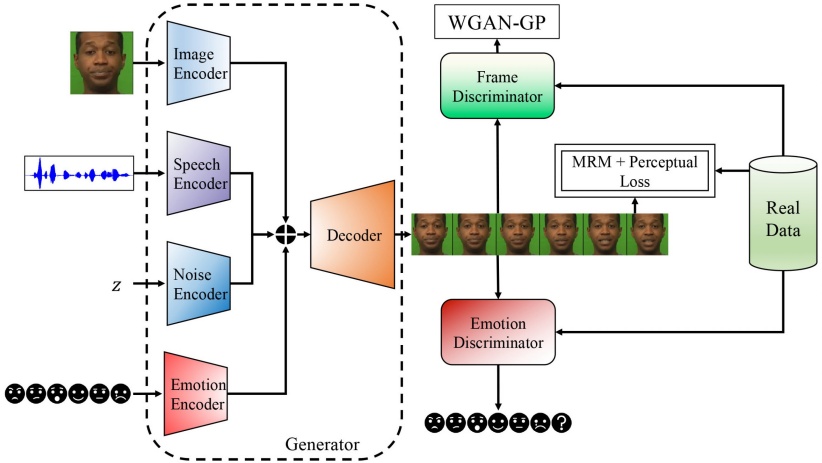


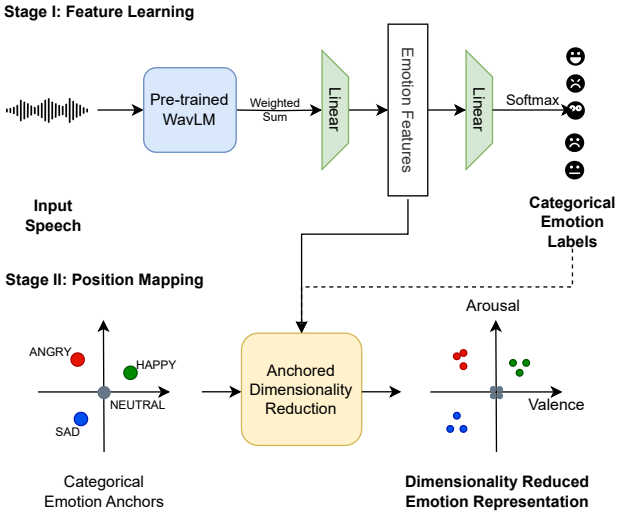
Education



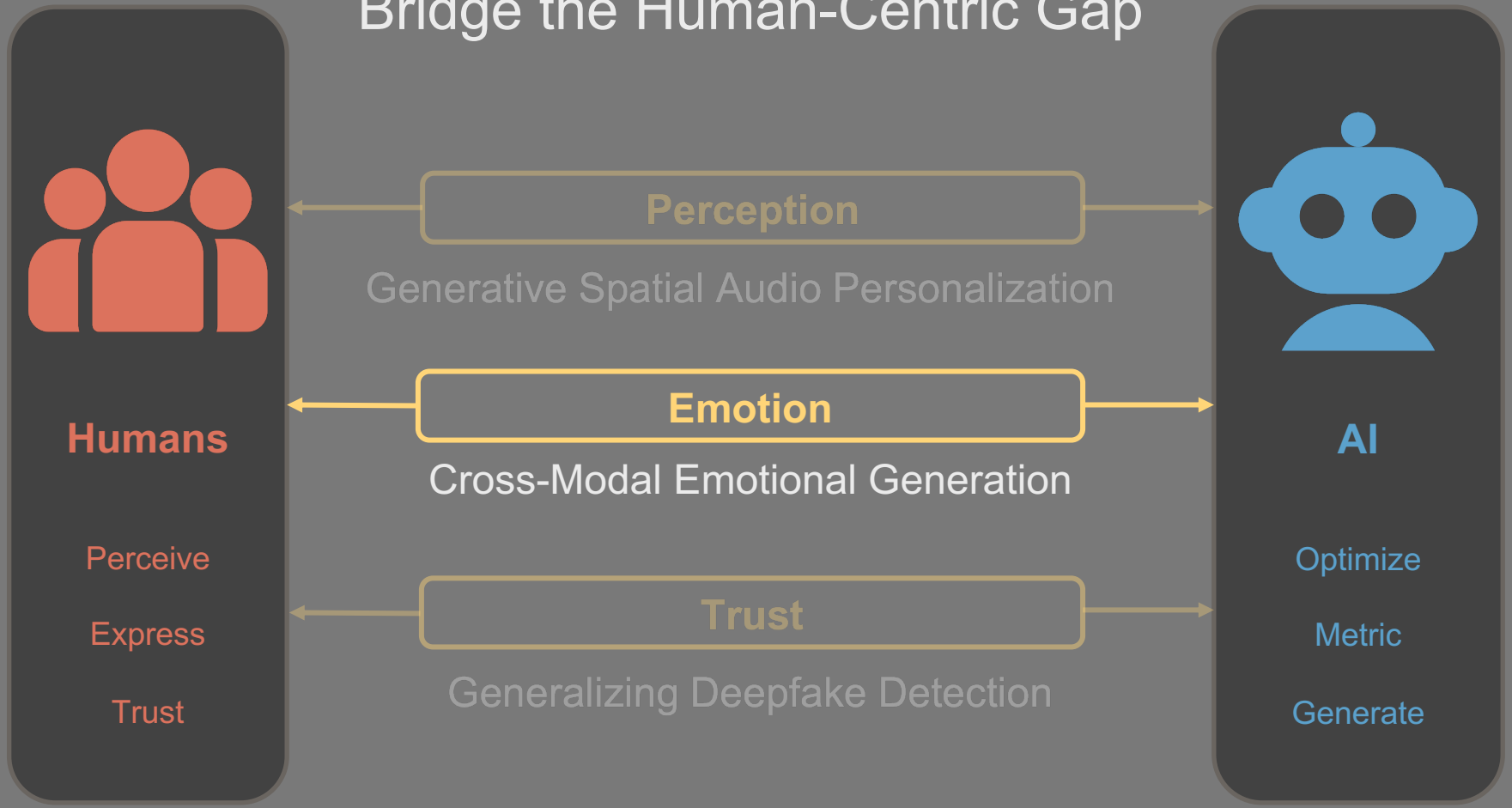
Summary: Cross-Modal Emotional Generation

- **Goal**: Emotional understanding and generation
- **Approach**: Explicit emotion encoder and discriminator, anchored dimensionality reduction from speech features, prefix to a spoken language model
- **Contributions**: Enables generating a broad spectrum of emotions

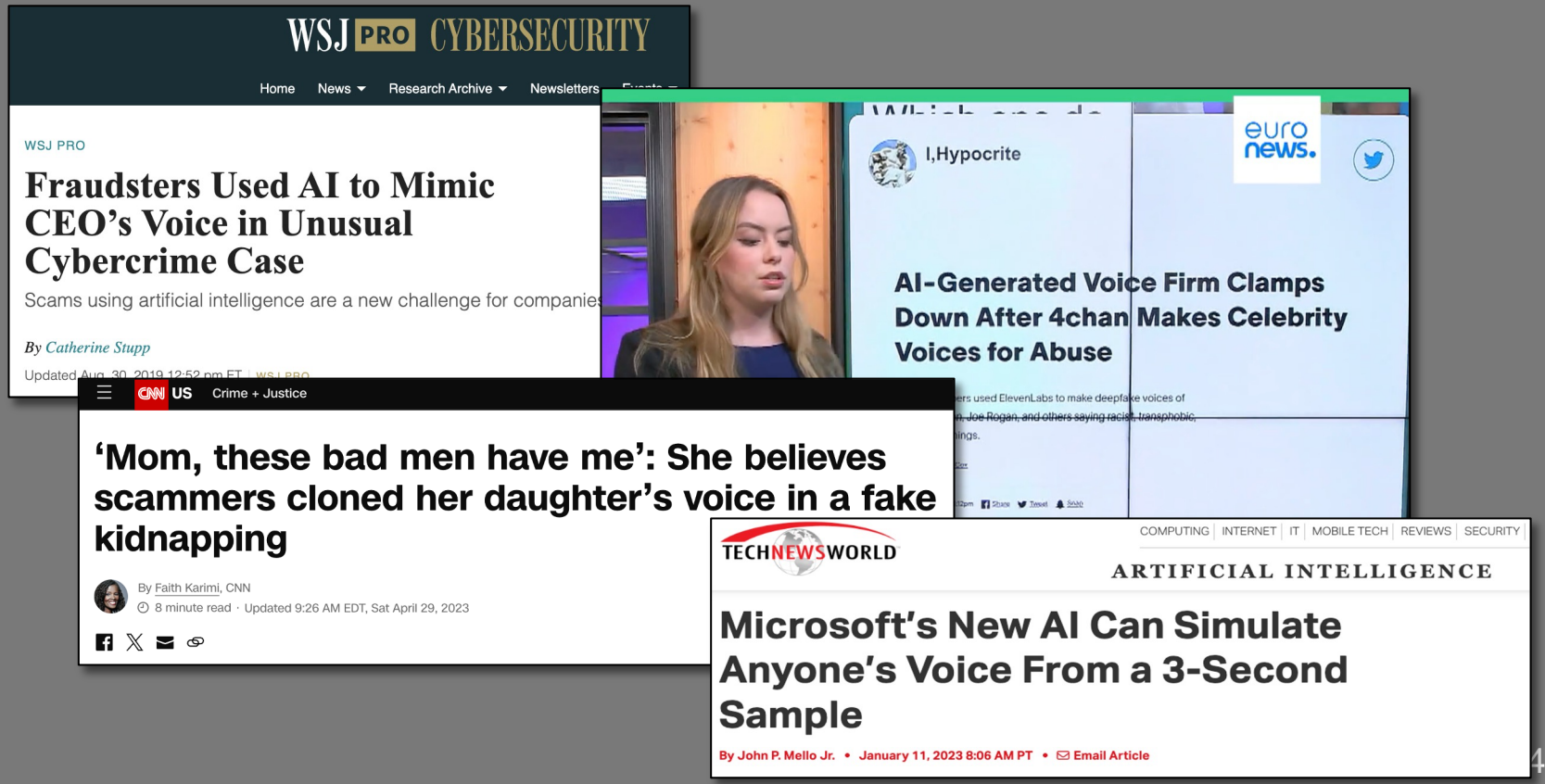




Bridge the Human-Centric Gap



Generative Audio AI raised ethical and societal concern



Bridge the Human-Centric Gap



Humans

Perceive

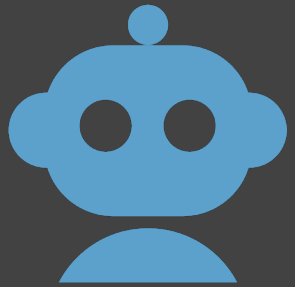
Express

Trust

How do we make generative audio(-visual) systems trustworthy against deepfakes?

Trust

Generalizing Deepfake Detection



AI

Optimize

Metric

Generate

University of Rochester

icassp 2022
Singapore, China, Online

One-class Learning Towards Synthetic Voice Spoofing Detection

You Zhang, *Fei Jiang, Zhiyao Duan*

University of Rochester

Published in *IEEE Signal Processing Letters*

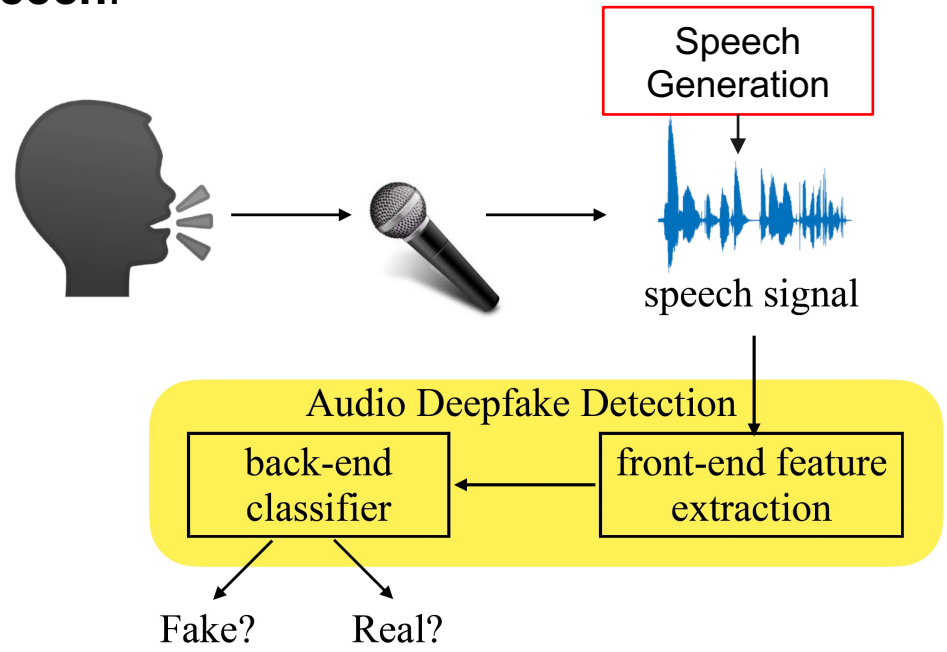


IEEE Signal Processing Letters Best Paper Nomination

Citations: 365

Background: Speech Deepfake Detection

A speech deepfake detection system is desired to distinguish generated **speech** from **bona fide speech**.



Motivation: Defend Against Unseen Deepfakes

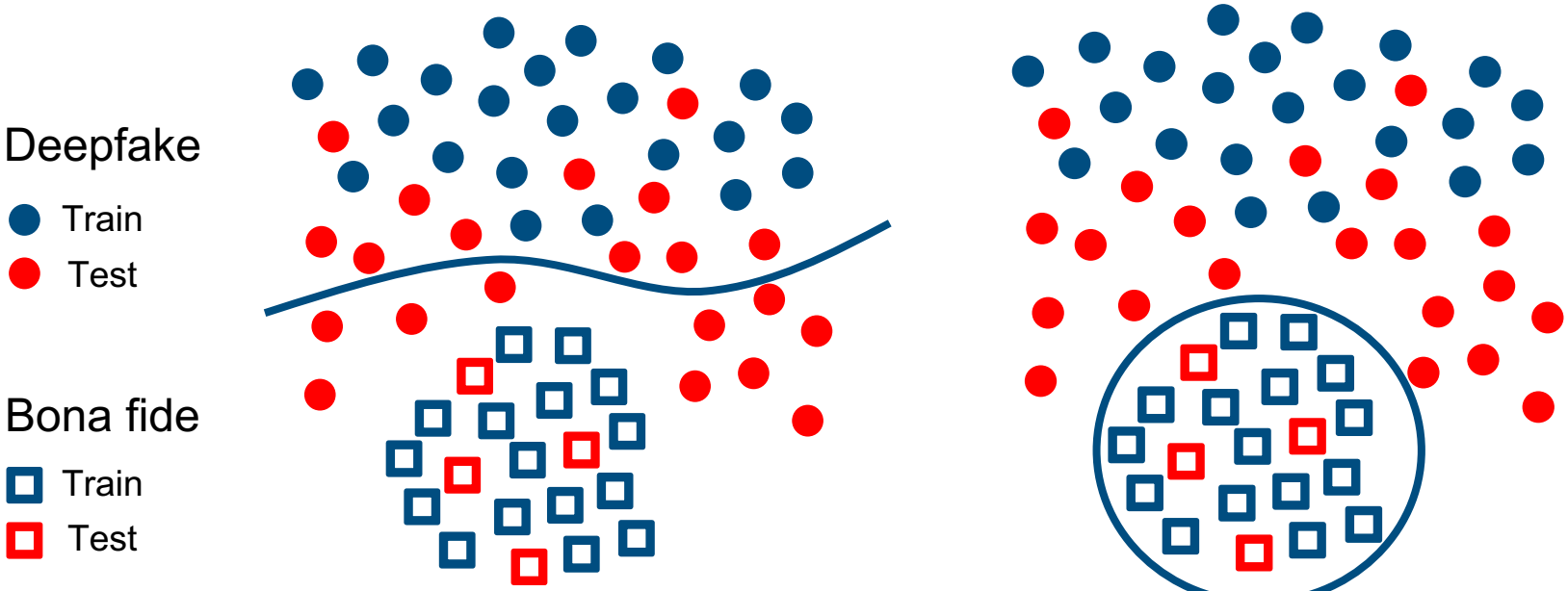
Problem:

- The fast development of speech generation are posing increasingly more threat.
- The **distribution mismatch** between the training set and test set for the deepfake class.

How can the deepfake detection system be better trained to defend against **unseen** deepfakes?

Generalization ability!

Preliminary: Binary classification vs. one-class classification



Bad generalization ability

Good generalization ability

You Zhang, Fei Jiang, Ge Zhu, Xinhui Chen, and Zhiyao Duan. “Generalizing Voice Presentation Attack Detection to Unseen Synthetic Attacks and Channel Variation”, *Handbook of Biometric Anti-spoofing (3rd edition)*, Springer, 2023.

Method: One-class learning

Learn the embedding space to:

- Compact the bona fide speech representation
- Isolate the speech deepfakes

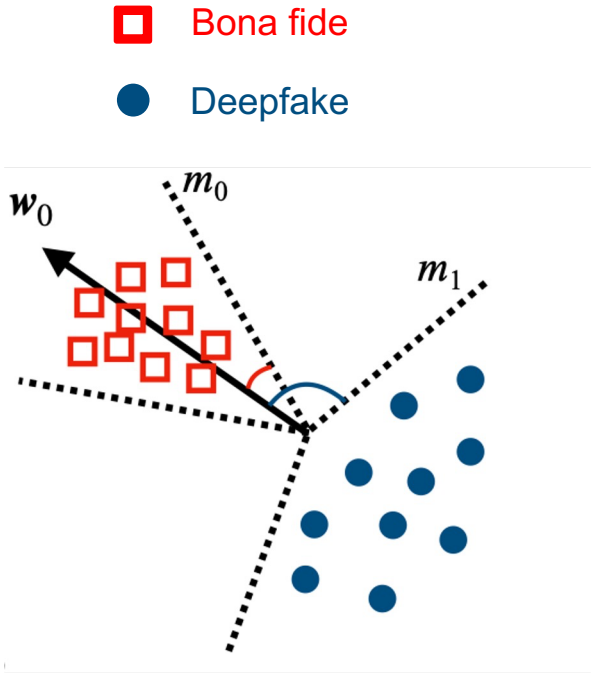
Training: OC-Softmax loss (Proposed)

$$\mathcal{L}_{OCS} = \frac{1}{N}\sum_{i=1}^{N}\log\left(1+e^{\alpha(m_{y_i}-\hat{\boldsymbol{w}}_0\hat{\boldsymbol{x}}_i)(-1)^{y_i}}\right).$$

scale factor, center vector, label, # samples, margin, embedding

Inference: cosine similarity

$$\mathcal{S}_{OCS} = \hat{\boldsymbol{w}}_0\hat{\boldsymbol{x}}_i.$$

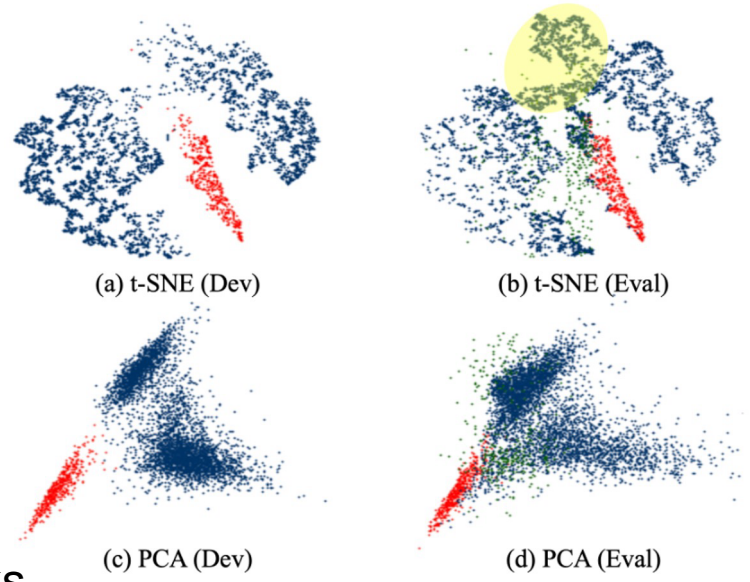


Results: Evaluation of OC-Softmax

Results on the development and evaluation sets of ASVspoof 2019 LA using different losses

Loss	Dev Set		Eval Set	
	EER (%)	min t-DCF	EER (%)	min t-DCF
Softmax	0.35	0.010	4.69	0.125
AM-Softmax	0.43	0.013	3.26	0.082
OC-Softmax	0.20	0.006	**2.19**	**0.059**

- OC-Softmax performs the best on unseen attacks.
- Achieved the state-of-the-art single-system performance.



Feature Embedding Visualization (red: bona fide, green: A17 attack, blue: deepfake)

You Zhang, Fei Jiang, and Zhiyao Duan, "One-class Learning Towards Synthetic Voice Spoofing Detection", *IEEE Signal Processing Letters*, vol. 28, pp. 937-941, 2021.

Impact

Interspeech 2022
18-22 September 2022, Incheon, Korea

A Deep One-Class Learning Method for Replay Att

Yijie Lou, Shiliang Pu, Jianfeng Zhou, Xin Qi, Qinbo Dong,

2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)
2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Improved One-class Learning for Voice Spoofing Detection

ue[§ †], Haipeng Peng[§ †], Yeqing Ren[§ †] and Mengmeng Zhao[§ † ‡]
er, State Key Laboratory of Networking and Switching Technology,
sity of Posts and Telecommunications, Beijing, China
ineering Laboratory for Disaster Backup and Recovery,
sity of Posts and Telecommunications, Beijing, China

DOI: 10.1109/ICASSP48485.2024.10446270

ONE-CLASS KNOWLEDGE DISTILLATION FOR SPOOFING SPEECH DETECTION

Jingze Lu[1,2], *Yuxiang Zhang*[1,2], *Wenchao Wang*[1], *Zengqiang Shang*[1], *Pengyuan Zhang*[1,2,*]

[1]Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing,
[2]University of Chinese Academy of Sciences, I
{lujingze, zhangyuxiang, wangwenchao, shangzengqiang, zh

Lightweight Voice Spoofing Detection Using Improved One-Class Learning and Knowledge Distillation

Yeqing Ren, Haipeng Peng, Lixiang Li, and Yixian Yang

One-Class Neural Network With Directed Statistics Pooling for Spoofing Speech Detection

Guoyuan Lin, Weiqi Luo, *Senior Member, IEEE*, Da Luo, *Member, IEEE*, and Jiwu Huang, *Fellow, IEEE*

Impact

NIJ | NATIONAL INSTITUTE OF JUSTICE
ADVANCING JUSTICE THROUGH SCIENCE
Menu

Home / Funding & Awards

Audio Deepfake Detection for Forensics and Security

Rate This Page

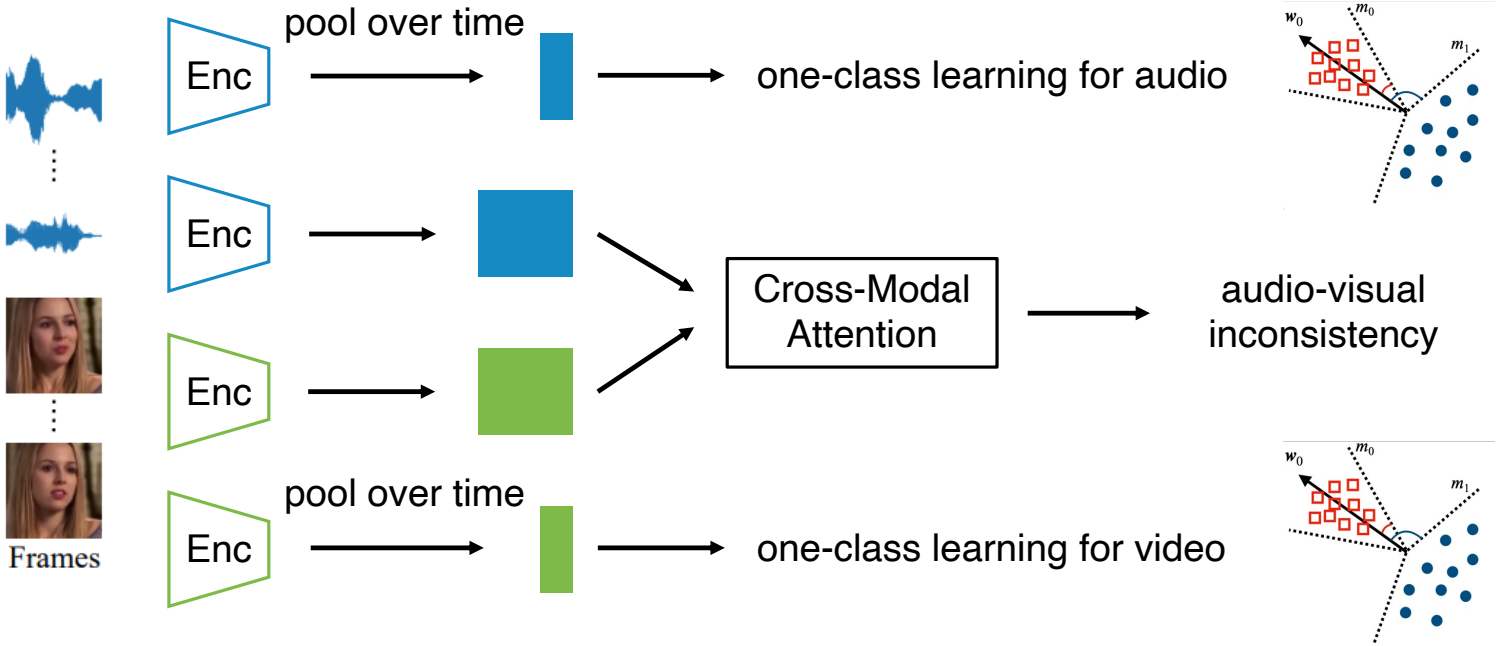
Award Information

Awardee: UNIVERSITY OF ROCHESTER
Award #: 15PNIJ-23-GG-01933-RESS
Funding Category: Competitive Discretionary
Location: ROCHESTER, NY
Congressional District: 25
Status: Open
Funding First Awarded: 2023
Total funding (to date): $64,003
Original Solicitation: NIJ FY23 Graduate Research Fellowship

2023	15PNIJ-23-GG-01933-RESS	You Zhang, UNIVERSITY OF ROCHESTER	Audio Deepfake Detection for Forensics and Security	$64,003

Extension: Audio-Visual Deepfake Detection

- Multi-stream framework with one-class learning against unseen deepfakes



Kyungbok Lee, **You Zhang**, and Zhiyao Duan. “A Multi-Stream Fusion Approach with One-Class Learning for Audio-Visual Deepfake Detection.” *Proc. IEEE 26th International Workshop on Multimedia Signal Processing (MMSP)*, 2024.

Results: Audio-Visual Deepfake Detection

Accuracy

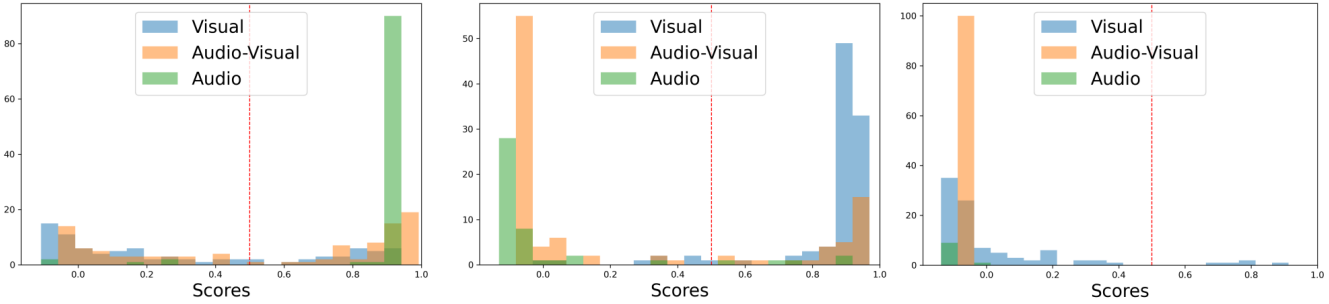
	Model	RAFV	FAFV	FARV
Baselines	AVDF	50.88	86.38	51.38
	MRDF-CE	54.38	88.25	47.25
	MRDF-Margin	55.12	86.88	47.62
Ours	**MSOC**	**60.25**	**89.88**	**74.38**

RAFV:
Real Audio
Fake Video

FAFV:
Fake Audio
Fake Video

FARV:
Fake Audio
Real Video

Score Distribution



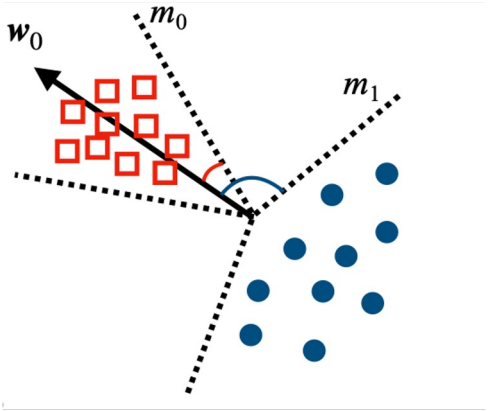
(a) RAFV (b) FARV (c) FAFV

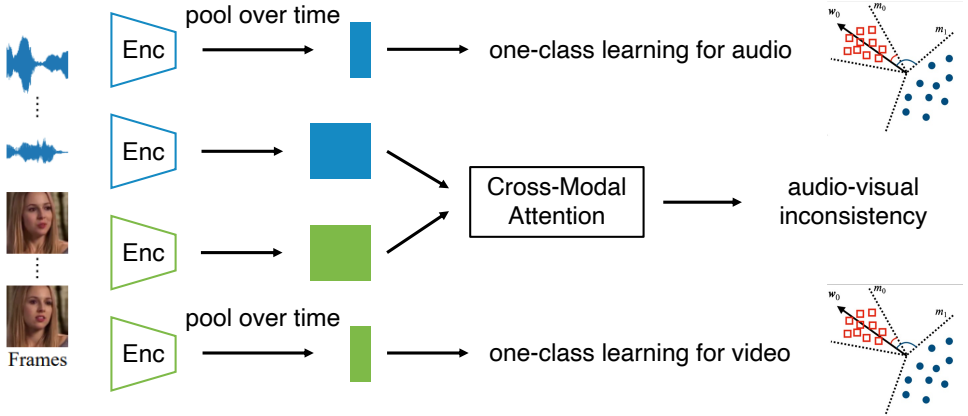
Our proposed method provides better generalization ability and interpretability.

Kyungbok Lee, **You Zhang**, and Zhiyao Duan. “A Multi-Stream Fusion Approach with One-Class Learning for Audio-Visual Deepfake Detection.” *Proc. IEEE 26th International Workshop on Multimedia Signal Processing (MMSP)*, 2024.

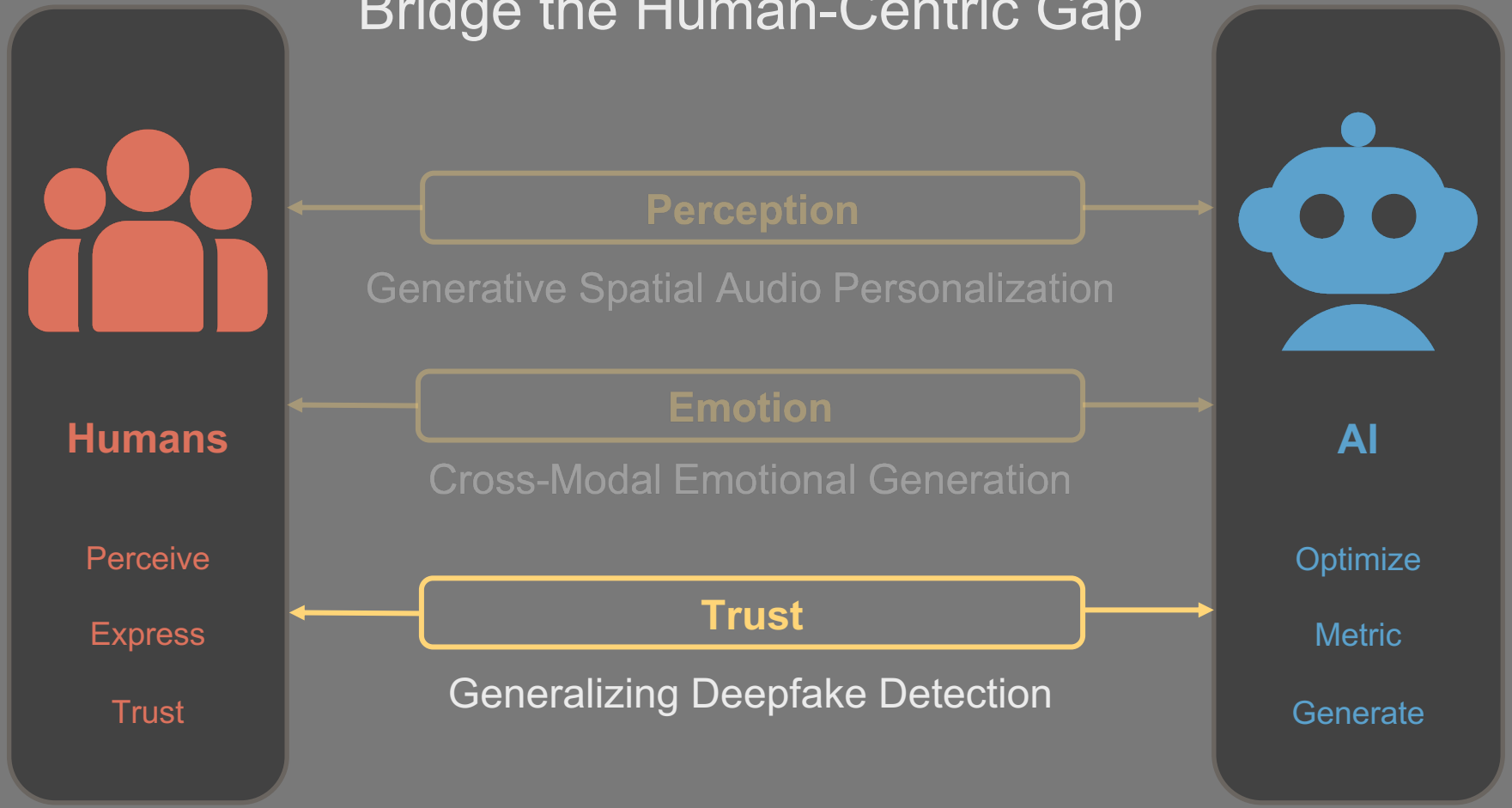
Summary: Generalizing Audio Deepfake Detection

- **Goal**: Defend against unseen deepfakes in audio and audio-visual deepfake detection
- **Approach**: One-class learning, multi-stream fusion
- **Contributions**: Improves generalization ability and interpretability





Bridge the Human-Centric Gap



Towards Human-Centric Multimodal Generative Audio AI

Generative Spatial Audio Personalization	**Perception**	Audio + Spatial Information + Human Perception	Generative Modeling
Cross-Modal Emotional Generation	**Emotion**	Audio + Vision + Text + Emotion	Controllable Generation
Generalizing Deepfake Detection	**Trust**	Audio + Vision	Safeguarding Generation

Other selected works on multimodal generative audio AI

- ## Spatial Audio + Large Audio Language Model

S Sakshi, Vaibhavi Lokegaonkar, **You Zhang,** Ramani Duraiswami, Sreyan Ghosh, Dinesh Manocha, and Lie Lu. "SPUR: A Plug-and-Play Framework for Integrating Spatial Audio Understanding and Reasoning into Large Audio-Language Models." *arXiv* 2026.

- ## Audio-Visual Analysis

Kyungbok Lee, **You Zhang**, and Zhiyao Duan. “Audio Visual Segmentation Through Text Embeddings.” IEEE ICIP 2025.
Abudukelimu Wuerkaixi, Kunda Yan, **You Zhang**, Zhiyao Duan, and Changshui Zhang. "DyViSE: Dynamic Vision-Guided Speaker Embedding for Audio-Visual Speaker Diarization." IEEE MMSP 2022.
Abudukelimu Wuerkaixi, **You Zhang**, Zhiyao Duan, and Changshui Zhang. "Rethinking Audio-Visual Synchronization for Active Speaker Detection." IEEE MLSP 2022.

- ## Singing Voice Deepfake Detection

Co-initiated this research direction and led the collaboration on organizing the first challenge

You Zhang, Yongyi Zang, Jiatong Shi, Ryuichi Yamamoto, Tomoki Toda, and Zhiyao Duan. "SVDD 2024: The Inaugural Singing Voice Deepfake Detection Challenge." IEEE SLT 2024.
Yongyi Zang, Jiatong Shi, **You Zhang**, Ryuichi Yamamoto, Jionghao Han, Yuxun Tang, Shengyuan Xu, Wenxiao Zhao, Jing Guo, Tomoki Toda, and Zhiyao Duan. "CtrSVDD: A Benchmark Dataset and Baseline Analysis for Controlled Singing Voice Deepfake Detection." Interspeech 2024.
Yongyi Zang*, **You Zhang***, Mojtaba Heydari, and Zhiyao Duan. “SingFake: Singing Voice Deepfake Detection.” IEEE ICASSP 2024. (***Equal Contributions**)

Future Vision: Human-Centric Multimodal Generative Audio AI



Humans

Perceive

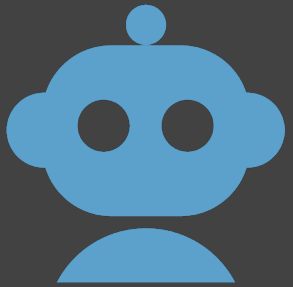
Express

Trust

Scientifically Grounded Generation

Multimodal Aligned Generation

Trustworthy Traceable Generation



AI

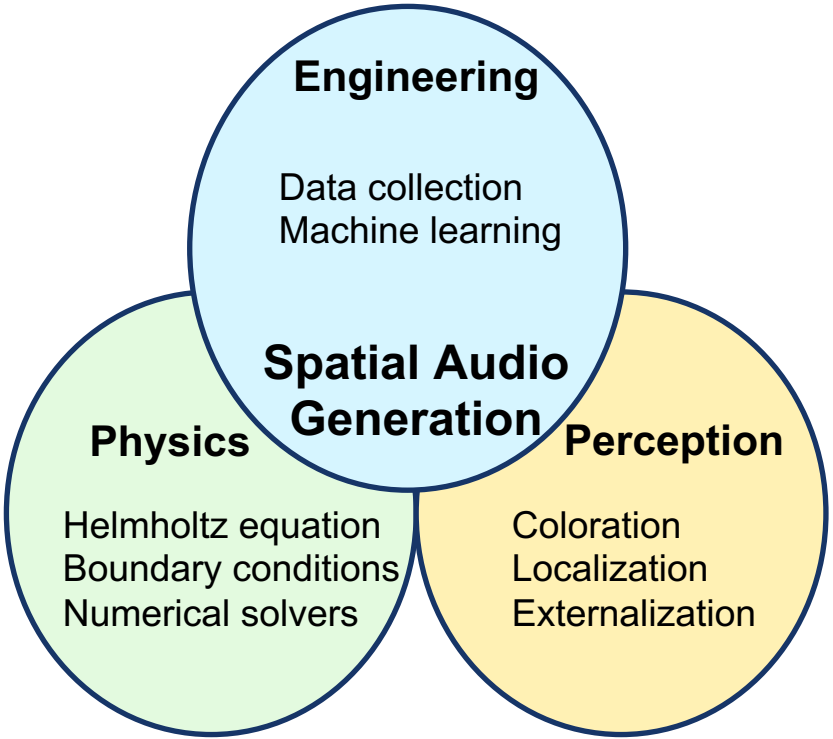
Optimize

Metric

Generate

Scientifically Grounded Spatial Audio Generation

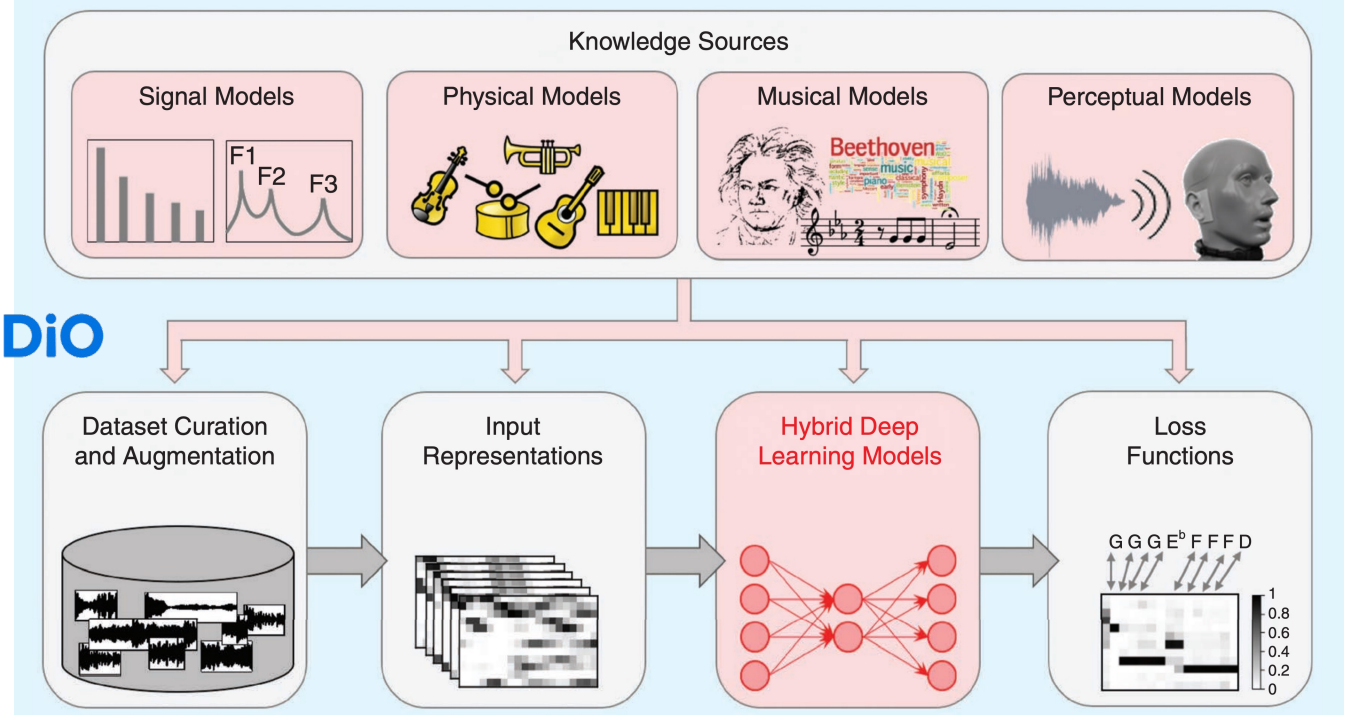
Physics-Informed Neural Network for Acoustic Field



Perception-Informed Binaural Synthesis and Reproduction

Scientifically Grounded Spatial Audio Generation (Cont'd)

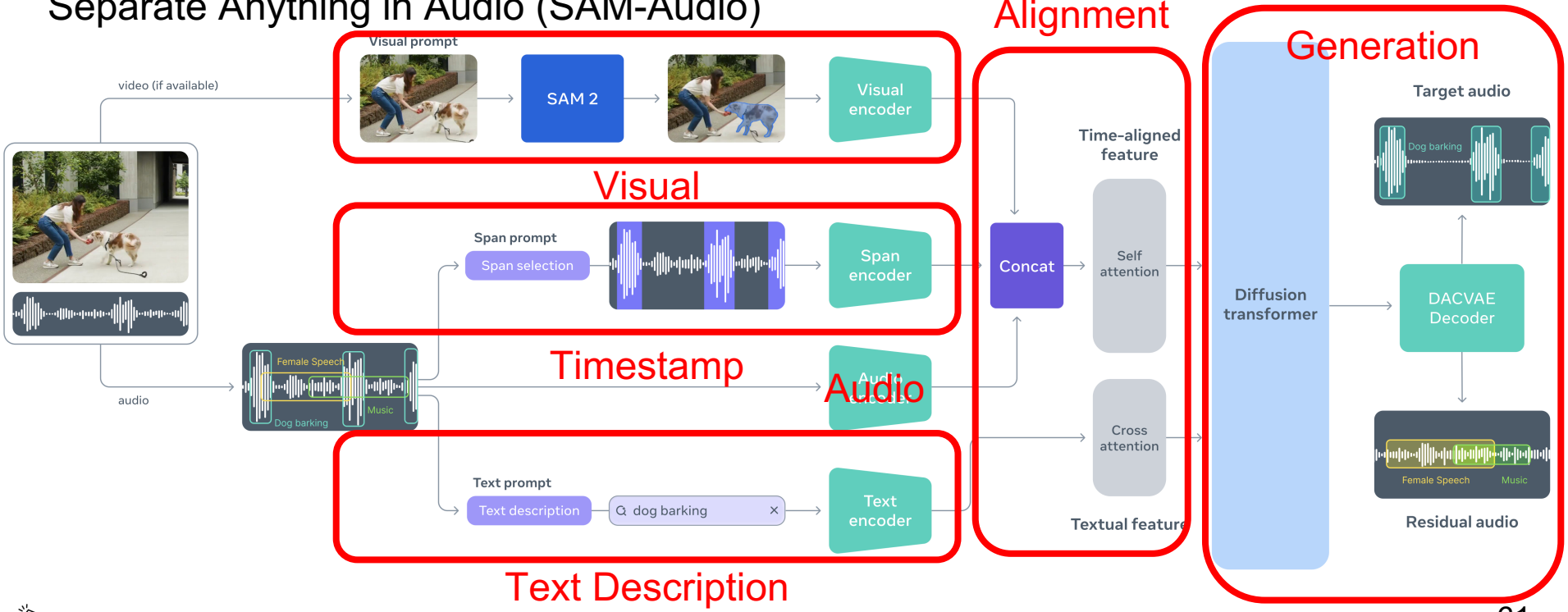
Hi-AUDiO



G. Richard, V. Lostanlen, Y. -H. Yang and M. Müller, "Model-Based Deep Learning for Music Information Research: Leveraging diverse knowledge sources to enhance explainability, controllability, and resource efficiency," *IEEE Signal Processing Magazine* 2024

Controllable Generation with Multimodal Alignment

Separate Anything in Audio (SAM-Audio)



Bowen Shi, Andros Tjandra, John Hoffman, Helin Wang, Yi-Chiao Wu, Luya Gao, Julius Richter et al. "SAM Audio: Segment anything in audio." *arXiv* 2025.

Controllable Generation with Multimodal Alignment (Cont'd)

More modalities:

Spatial direction

Musical scores

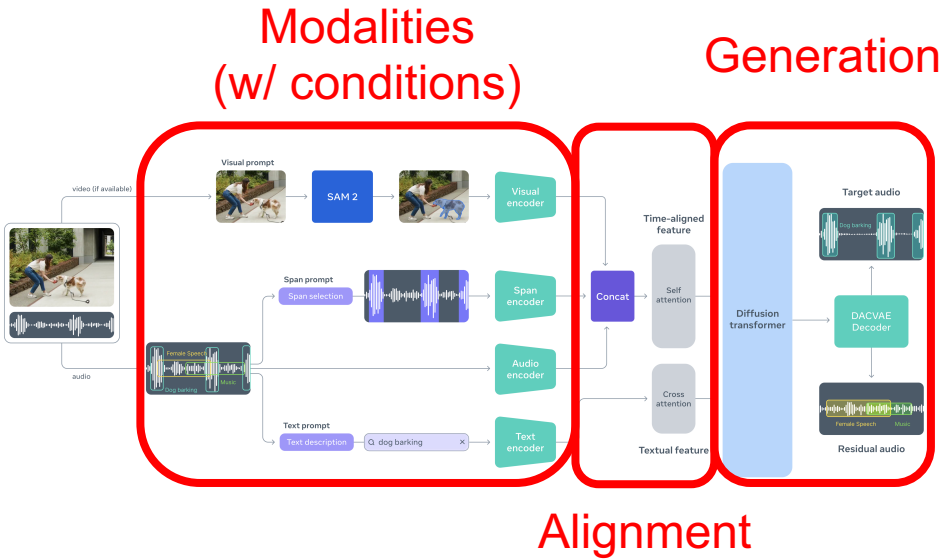
Enrollment voice

Speech transcript

Human vocal imitation

Event roll

…



Generative tasks:

Separation

Editing

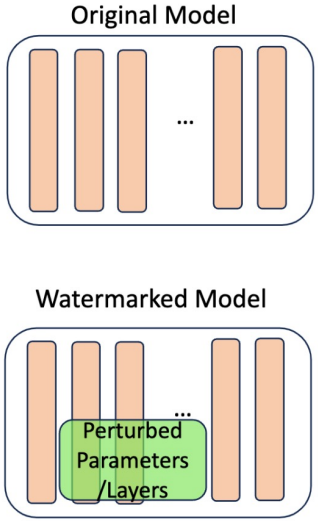
Conversion

Generation

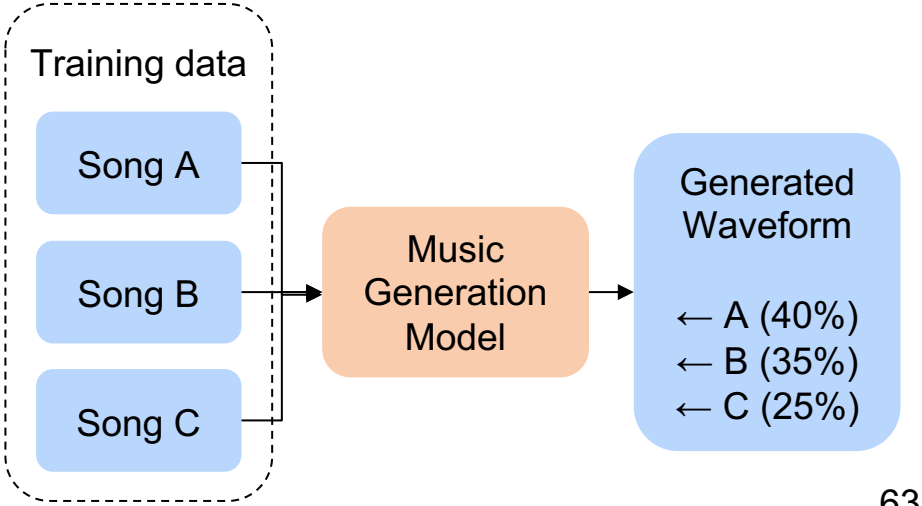
…

Watermarking and Source Tracing for Generative Audio AI

- Inherent Watermarking
  - Distribute watermarks into the model weights
  - Robust against fine-tuning



- Traceable Generation
  - Which training data contribute to the generated samples
  - Machine unlearning for privacy



Future Vision: Human-Centric Multimodal Generative Audio AI



Humans

Perceive

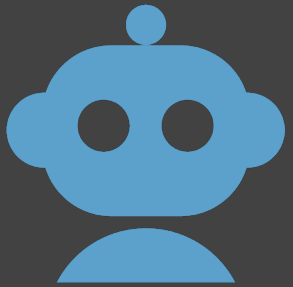
Express

Trust

Scientifically Grounded Generation

Multimodal Aligned Generation

Trustworthy Traceable Generation



AI

Optimize

Metric

Generate

Potential Funding Sources

- Early-stage investigators
  - ANR JCJC – Individual projects
  - ANR PRME – Ambitious single-team projects
- Collaborative
  - ANR PRC – national collaborative projects
  - ANR PRCI – international bilateral collaborations
  - Horizon Europe / Marie Skłodowska-Curie grant
  - NSF-French ANR international collaborations
- Industry-oriented
  - ANR PRCE – collaborative projects including a company partner
  - Cifre (“Industrial Agreements for Training through Research”)

My funding proposal writing experiences:

- NIJ independent fellowship award (sole PI)
- Contributed to 6 grants from public and private sectors, including National AI Research Resource Pilot, Meta, Microsoft, etc.
- Total: $253,992 and GPU credits)



Career Roadmap

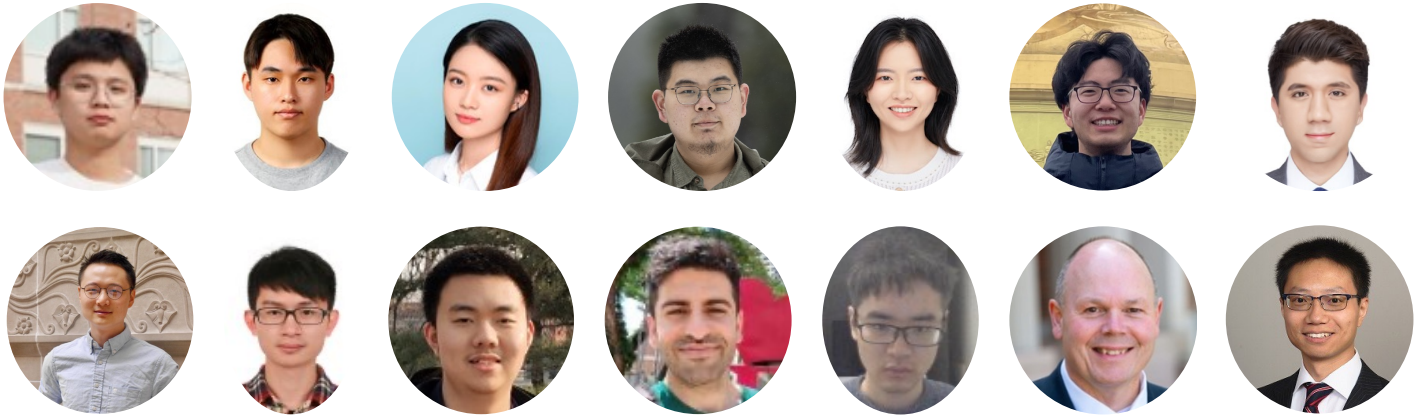
Phase	Timeline	Deliverables
Launch	Year 1–2	First ANR submission (JCJC); recruit 2 PhD students (incl. Cifre); 4 flagship papers; establish ADASP collaborations
Build	Year 3–5	First PhD student graduates, recruit more PhD students; ANR PRC + Horizon Europe awarded; Organize a challenge and special session at ICASSP; ERC Starting Grant application
Consolidate	Year 6+	HDR dossier submission; 3 PhD students graduated; recognized group in multimodal generative audio AI

Grounded in my past mentoring and collaborative experiences:

- 11 papers with students as first/co-first author across 9 mentored students at UR
- 2 research interns co-mentored at Dolby
- SVDD challenge: led multi-institution organization across CMU, Nagoya, Rochester
- 3 conference tutorials organized and presented

Acknowledgment

UR Colleagues



External Collaborators



Funding Agencies

NSF

NIJ NATIONAL INSTITUTE OF JUSTICE ADVANCING JUSTICE THROUGH SCIENCE

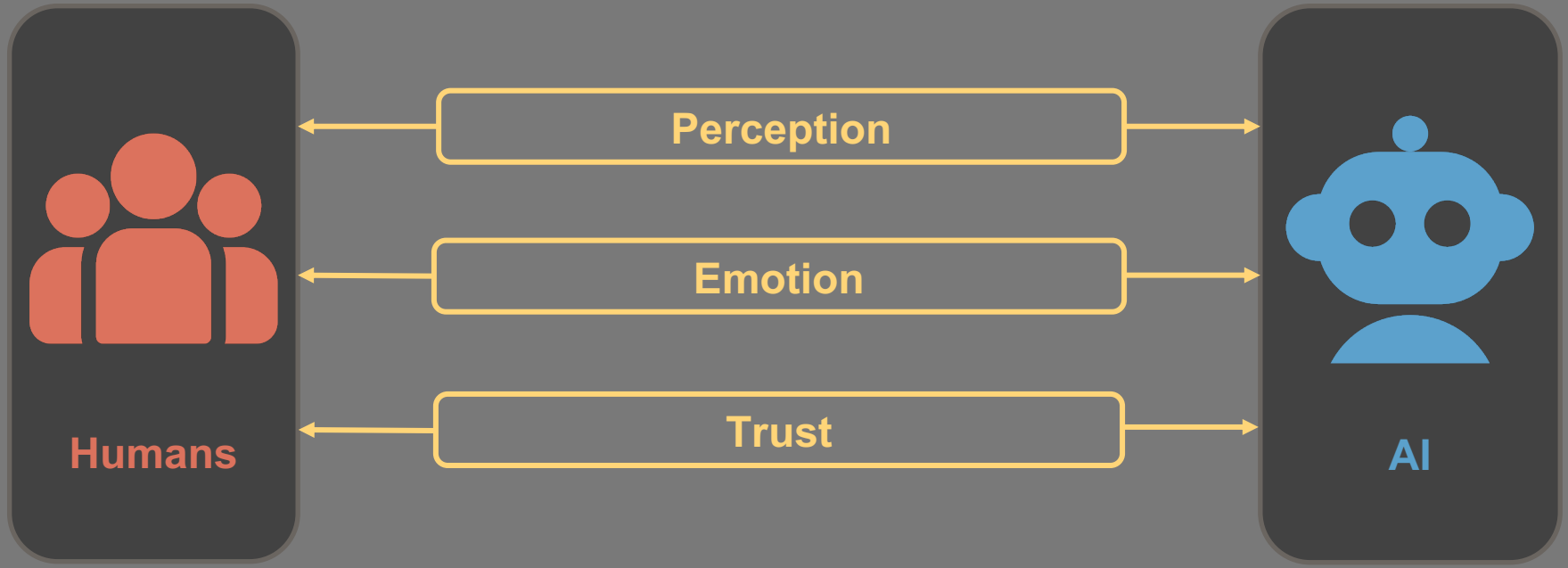
A NEW YORK STATE CENTER OF EXCELLENCE

Meta

Microsoft

ingenID Biometric Innovations

Towards Human-Centric Multimodal Generative Audio AI



My future research builds generative audio AI that is scientifically grounded, multimodal aligned, and trustworthy by design.

Thank You! Questions?