You (Neil) Zhang | Toward Immersive and Secure Audio Experiences

My research develops intelligent algorithms to **enhance human auditory experiences by making them more immersive and secure**. By integrating innovations in audio signal processing, acoustics, machine learning, and auditory perception, I aim to **create human-centric audio technologies**.

Audio is ubiquitous in daily life, yet often underappreciated. It enhances immersion by conveying rich information—music embodies art and emotion, while speech encodes meaning and sentiment. Historically, sound revolutionized cinema, transforming silent films into immersive audio-visual storytelling. Today, audio AI technologies drive similar advancements, enriching digital and physical spaces.

Immersive audio technologies are critical as they transform how humans interact with the environment, bridging the gap between digital and physical worlds. Today's AI-driven audio techniques thrive in applications such as virtual assistants, virtual and augmented reality (AR/VR), and hearing aids. However, these immersive advancements bring new challenges, as most mainstream audio AI models prioritize what content to deliver (e.g., speech, music, sound events) but neglect how to deliver the content to fit the listener's environment. This lack of adaptability results in systems that fail to deliver seamless and immersive experiences, similar to placing a playful children's game interface in a formal legal discussion.

Furthermore, with the rise of generative AI, our auditory experiences are evolving at an unprecedented pace. Generative audio technologies are poised to blur the boundaries between authentic and synthetic sound. This creates opportunities but also raises critical questions about trust and security in auditory media. For example, recent AI-generated songs are becoming a source of creative expression, while they have also raised concerns about copyright infringement and singer identity theft.

Ensuring immersive and secure audio interactions requires addressing the key challenges, which raise three fundamental questions: 1) How can AI techniques be applied to **create more immersive auditory experiences**? 2) How can we **mitigate the privacy and security risks** posed by generative audio technologies? 3) How can **external information** enhance the immersiveness and security of audio systems?

Back then, I contributed to three key areas for immersive and secure audio technologies— ***spatial audio, audio deepfake detection, and multimodal analysis***.

Immersive Audio through Personalized Spatial Rendering

Creating immersive audio experiences requires **precise modeling of sound propagation** from the source to the human listener's ears through space. This includes models for room acoustics, spatial audio, and individual auditory perception.

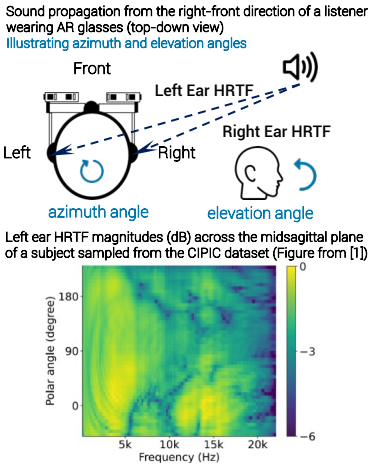
Spatial audio plays a pivotal role in creating immersive experiences through headphones or VR headsets by enabling users to perceive sound direction and distance. By incorporating individual uniqueness in auditory perception, spatial audio rendering can significantly enhance the sense of immersion. A key aspect of this task is to predict personalized head-related transfer functions (HRTFs), which describe spatial filtering effects of human geometry for accurate sound localization.

A. Research Overview

- **Spatial Audio**
  - HRTF Representation [1, 2]
  - HRTF Personalization [3, 4, 5]
- **Audio Deepfake Detection**
  - Generalize to Unseen Attacks [6, 7, 8]
  - Robustness to Channel & Speaker Variation [9, 10, 11, 12]
  - Singing Voice Deepfake Detection [13, 14, 15]
- **Multimodal Analysis**
  - Emotion Rendering [16, 17, 18]
  - Audio-Visual Conversation [19, 20, 21]

B. Spatial Audio

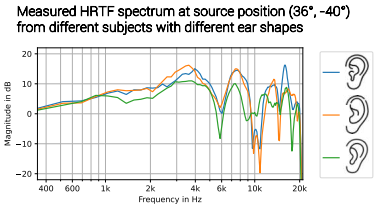
B1. Sound propagation and HRTF



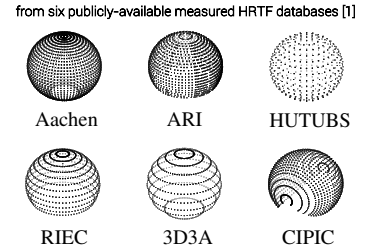
Due to the resource-intensive nature of HRTF measurements, **existing databases often have a limited amount of subjects**, posing challenges for data-intensive machine learning models. HRTFs are inherently high-dimensional, encompassing numerous spatial locations and frequency bins per subject. Prior HRTF modeling methods, trained on a single dataset with limited subjects, often face overfitting and struggle to **generalize across datasets due to their dependence on spatial sampling grids. To address this, I introduced *HRTF Field*** [1], a generative model using neural fields (implicit neural representations) to model HRTFs. Neural fields, neural networks trained on discrete location-magnitude pairs, provide a differentiable representation of the HRTF across all azimuth and elevation angles, allowing HRTF queries in any direction, independent of the original sampling grid. This approach enables training on any dataset or combination of datasets, expanding the available training data and improving generalization. This work, supported by seed funding from our university's Data Science Institute, was **recognized at ICASSP 2023 as one of the top 3% of all accepted papers**. Furthermore, we investigated cross-database differences beyond spatial sampling schemes and introduced a novel normalization strategy to mitigate variations in the HRTF measurement setup during the mix-database training of the HRTF Field [2]. Building on the neural field representation, **I developed a personalization framework that uses anthropometric measurements** to predict the latent representation and integrate it with the HRTF Field model [3].

In an earlier effort to create a generalizable HRTF representation for personalization, I co-developed a machine learning and signal processing approach that compacts HRTFs using spherical harmonics to capture their overall spatial pattern. I then designed a neural network model to **predict spherical harmonic coefficients** from either anthropometric measurements [4] or 3D-scanned head meshes [5]. Unlike prior methods using separate models for different directions, this approach enables prediction for arbitrary spatial directions, **achieving global HRTF personalization**.

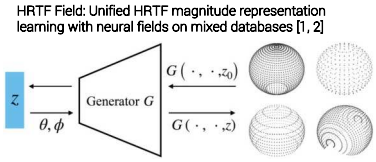
B2. Personalized HRTF



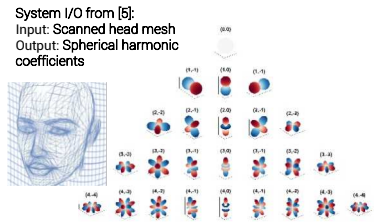
B3. Spatial Sampling Schemes



B4. Neural Field Representation



B5. Personalization System I/O



Safeguarding Audio Experiences with Deepfake Detection

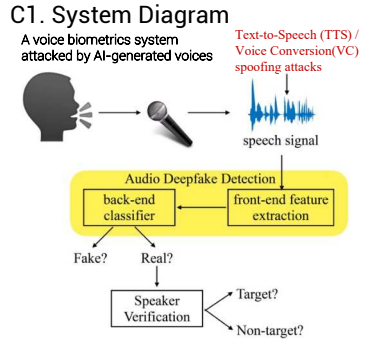
As AI-generated text and voices from systems like ChatGPT become increasingly common and integral to daily life, they present new security challenges. To prevent misinformation and protect biometric systems, it is crucial to distinguish real human communication from synthetic or deepfake audio. **My PhD thesis focuses on developing robust, generalizable algorithms to detect audio deepfake attacks**, ensuring the integrity and trustworthiness of audio technologies.

Prior methods of audio deepfake detection typically frame the problem as a binary classification task (real vs. fake). However, these approaches **struggle to detect unseen attacks**, as speech synthesis evolves rapidly, leading to a distribution that may not be well represented in training data. **To address this, I proposed a series of one-class learning methods [6, 7, 8] that compact natural speech representations** and push away fake speech in the embedding space, thereby improving the generalization ability against unseen attacks. The initial work OC-Softmax [6], published in IEEE Signal Processing Letters (SPL) and presented at ICASSP 2022, has become one of the most cited methods in the field, with 249 citations as of December 2024.

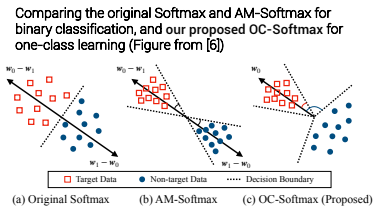
Real speech varies across diverse audio conditions, such as speaker, channel, and codec, which challenge system robustness, particularly when unseen conditions cause **cross-dataset performance degradation**. I was among the first to study this issue in audio deepfake detection, identifying the channel as a critical factor [9].

C. Audio Deepfake Detection

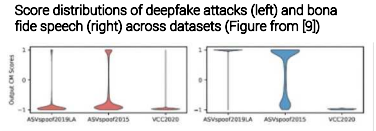
C1. System Diagram



C2. One-Class Learning



C3. Cross-Dataset Performance



I developed training strategies to address channel robustness using data augmentation [9, 10] and phase perturbation [11]. To adapt audio deepfake detection for diverse speakers, I introduced a probabilistic framework for spoofing-aware speaker verification [12], optimizing anti-spoofing by conditioning on speaker verification confidence scores. Additionally, **I refined the one-class learning strategy to support speaker enrollment and incorporated bona fide speech clustering** in the latent space to model speaker diversity while maintaining the generalization ability [8].

With advancements in singing voice generation and the increasing presence of AI singers on social media, **I co-initiated research on singing voice deepfake detection** (SVDD) [13] by curating an in-the-wild dataset from media platforms. **I led a collaboration** among University of Rochester, Carnegie Mellon University, and Nagoya University to **organize the first SVDD challenges** [14]. Held at IEEE SLT and ISMIR, the challenge featured two tracks: CtrSVDD [15], focusing on isolated vocals, and WildSVDD, addressing manipulated vocals with background music. I benchmarked the baselines, summarized submissions, and organized the SVDD special session.

External Information to Enhance Immersiveness and Security

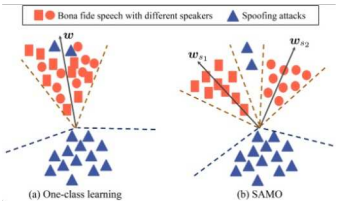
Audio is often presented together with other modalities, such as speech transcripts, music scores, and videos, with external information playing a key role in enhancing immersive and secure audio experiences. Achieving immersiveness requires coherent multimedia rendering across senses, particularly sight and sound. Multimodal approaches can also enhance security, akin to two-factor authentication. My research has explored **multimodal analysis, focusing on emotion modeling and audio-visual tasks**, primarily through internships and student research mentorship.

Emotion, as a paralinguistic feature, is a fundamental aspect of human communication, expressed across modalities such as speech, facial expressions, and body language. **Emotion rendering aims to convey the underlying emotion in synthesized speech or facial expressions.** My interest in emotion modeling began through collaboration with a lab alumnus working on talking face generation. We developed the first end-to-end approach capable of conditioning on categorical emotions (e.g., happy, angry) [16], enabling the generation of talking face videos with even mismatched audio-visual emotions. This innovation opens avenues for psychological studies on audio-visual emotion congruency. To align with psychological research on continuous emotion representation, we pursued a more generalizable approach to predict dimensional emotion representation from self-supervised features [17], reducing the need for labor-intensive annotations and addressing subjective variability across datasets. This framework was later applied to emotional text-to-speech synthesis [18], enabling a broad spectrum of emotion rendering.

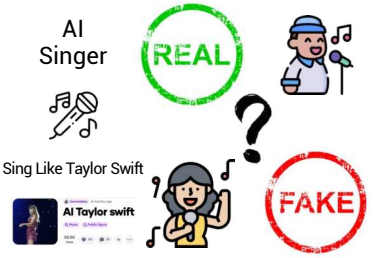
The visual modality aids conversational understanding. I explored audio-visual research during internships and applied this experience to mentoring students interested in **audio-visual conversations and general audio-visual scene analysis**. For audio-visual conversations, we proposed solutions for synchronization issues in active speaker detection [19] and addressed off-screen speaker challenges in audio-visual speaker diarization [20]. For general audio-visual scenes, I designed a cross-modal module to prompt large vision foundation models with audio, showing that leveraging these models improves audio-visual segmentation performance. I also developed a multi-stream approach with one-class learning for audio-visual deepfake detection [21], which offers interpretability and improves generalization ability.

C4. Addressing Speaker Variation

SAMO: Speaker Attractor Multi-Center One-Class Learning [8]



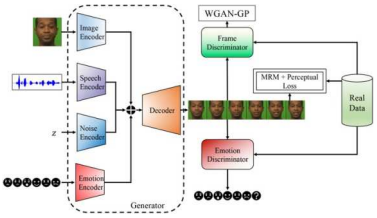
C5. Singing Voice Deepfake Detection



D. Multimodal Analysis

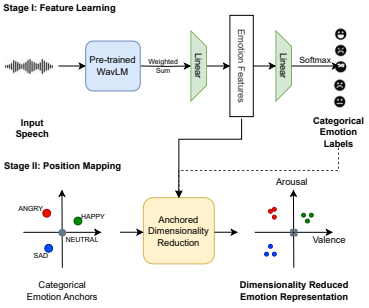
D1. Emotional Talking Face Generation

System overview with emotion-conditional input (from [15])



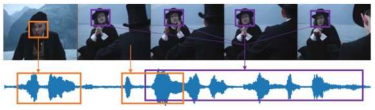
D2. Emotion Representation Learning

Framework of our proposed two-stage arousal-valence learning from categorical emotion labels (from [16])



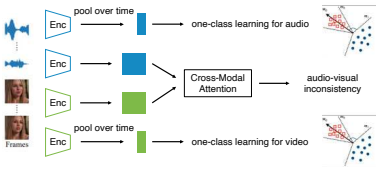
D3. Audio Visual Speaker Diarization

Keyframes and audio of a video segment with two speakers. The second frame shows an off-screen speaker. (from [19])



D4. Audio Visual Deepfake Detection

The multi-stream architecture of our proposed fusion approach with one-class learning strategy (from [20])



Future Research Agenda

My overarching research goal is to **advance immersive and secure multimedia audio content delivery for human-centric technologies**, such as audio scene generation, proactive audio forensics, and embodied audio agents. I plan to establish the **OpenAudio Lab (OpAL)** as a hub for innovation in computer audition. "Open" reflects my commitment to tackling **open challenges** in audio signal processing, promoting **open-science initiatives**, and being **open to interdisciplinary collaboration** to broaden the impact of audio technologies. I foresee three research directions gaining prominence in the coming years:

Human Perceptually Plausible World Understanding and Virtual Generation. I aim to create realistic virtual auditory environments by integrating auditory scene understanding with virtual acoustics modeling and emphasizing human perception. While generative AI excels at simulating virtual spaces, its reliance on limited datasets introduces bias, hindering generalization to rare cases. To achieve human-perceptual plausibility, I propose combining data-driven approaches with perceptual models that capture how humans naturally perceive sound. Auditory neuroscience will be key in developing accurate computational perceptual models, fostering interdisciplinary collaboration. Additionally, I aim to design physics-informed machine learning for acoustics applications to ensure both physical accuracy and perceptual plausibility.

Proactive Watermarking for Safeguarding Audio AI. Passive, classification-based deepfake detection often falls into an "always-chasing" trap. I am developing proactive defenses, such as watermarking techniques that embed detectable markers into AI-generated content to provide verifiable evidence of its source. As AI-processed speech becomes more prevalent in applications like speech enhancement and anonymization, distinguishing between benign and malicious uses presents new challenges. In such scenarios, proactive methods may surpass traditional deepfake detection in importance. A key challenge is ensuring watermark resilience against distortions caused by compression, re-recording, and model fine-tuning. My goal is to design learning-based methods to effectively enhance watermark robustness under such conditions.

Humanized and Embodied AI Auditory Assistants. I aim to develop auditory assistants capable of responding with emotional nuance, offering comfort (e.g., through music therapy), and generating immersive soundscapes (e.g., simulating a serene beach), building on multimodal analysis work. To enhance personalization and adaptability, I plan to leverage recent advancements in world foundation models beyond the language space, fine-tuning multisensory models to adjust to social and environmental cues dynamically. This involves designing interaction modules and implementing feedback loops that respond to user behaviors and preferences. These innovations can also extend to embodied AI systems, enabling robots to understand and respond to human emotions and environmental contexts through advanced auditory capabilities.

Funding and Industrial Collaborations

My research has been supported by funding from the **IEEE Signal Processing Society (SPS) Scholarship**, the **National Institute of Justice (NIJ) Graduate Research Fellowship Award** (15PNIJ-23-GG-01933-RESS), NSF grants 1741472, 1846184, 2222129, and 1922591, the New York State Center of Excellence in Data Science, and industrial collaborators, including Voice Biometrics Group, IngenID, Microsoft, and Meta. To support my research group in the future, I will pursue funding opportunities from the NSF Division of Information and Intelligent Systems (CISE/IIS) for audio machine learning projects, the NIJ and the Air Force Research Laboratory for deepfake-related projects. I will also seek industrial collaborations with leading tech companies in the fields of speech, music, audio, and biometrics, such as Microsoft, Meta, Apple, Amazon, Adobe, Smule, and Pindrop—many of which are organizations with which my past co-authors are now affiliated.

Broader Impact

I envision my research at the intersection of AI and audio signal processing to advance both scientific understanding and practical applications. In immersive audio, my work could support AI-based evaluation tools for auditory modeling and deepen our understanding of human auditory perception in neuroscience, with spatial rendering designs informed by perceptual findings. In cybersecurity, my research on robust audio forensics and deepfake detection strengthens biometric systems and safeguards sensitive information. Beyond security, it holds promise in healthcare and education—where speech analysis could predict cognitive impairment progression, support emotional well-being, and improve teaching methods and learning outcomes. The AI techniques I develop for audio analysis also have broad applications in other time series domains, such as fintech, autonomous driving, and biosignals, as audio shares key characteristics with other 1D signals.

References

[1] **You Zhang**, Yuxiang Wang, and Zhiyao Duan. HRTF field: Unifying measured HRTF magnitude representation with neural fields. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023 (Recognized as one of the top 3% of all papers accepted at ICASSP 2023).

[2] Yutong Wen, **You Zhang**, and Zhiyao Duan. Mitigating cross-database differences for learning unified HRTF representation. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1--5, 2023.

[3] **You Zhang**, Yuxiang Wang, Mark Bocko, and Zhiyao Duan. Grid-agnostic personalized head-related transfer function modeling with neural fields. *The Journal of the Acoustical Society of America*, 153(3_supplement):A125--A125, 2023 (Recognized by Signal Processing at the ASA Student Paper Award - Second Place).

[4] Yuxiang Wang, **You Zhang**, Zhiyao Duan, and Mark Bocko. Global HRTF personalization using anthropometric measures. In *Audio Engineering Society Convention 150*. Audio Engineering Society, 2021.

[5] Yuxiang Wang, **You Zhang**, Zhiyao Duan, and Mark Bocko. Predicting global head-related transfer functions from scanned head geometry using deep learning and compact representations. *arXiv preprint arXiv:2207.14352*, 2022.

[6] **You Zhang**, Fei Jiang, and Zhiyao Duan. One-class learning towards synthetic voice spoofing detection. *IEEE Signal Processing Letters*, 28:937--941, 2021.

[7] **You Zhang**, Fei Jiang, Ge Zhu, Xinhui Chen, and Zhiyao Duan. Generalizing voice presentation attack detection to unseen synthetic attacks and channel variation. In Sébastien Marcel, Julian Fierrez, and Nicholas Evans, editors, *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*, pages 421--443. Springer, 2023.

[8] Siwen Ding, **You Zhang**, and Zhiyao Duan. SAMO: Speaker attractor multi-center one-class learning for voice anti-spoofing. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[9] **You Zhang**, Ge Zhu, Fei Jiang, and Zhiyao Duan. An empirical study on channel effects for synthetic voice spoofing countermeasure systems. In *Proc. Interspeech*, pages 4309--4313, 2021.

[10] Xinhui Chen*, **You Zhang***, Ge Zhu*, and Zhiyao Duan. UR channel-robust synthetic speech detection system for ASVspoof 2021. In *Proc. Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pages 75--82, 2021 (* **equal contribution**).

[11] Yongyi Zang, **You Zhang**, and Zhiyao Duan. Phase perturbation improves channel robustness for speech spoofing countermeasures. In *Proc. Interspeech*, pages 3162--3166, 2023.

[12] **You Zhang**, Ge Zhu, and Zhiyao Duan. A probabilistic fusion framework for spoofing aware speaker verification. In *Proc. The Speaker and Language Recognition Workshop (Odyssey)*, pages 77--84, 2022.

[13] Yongyi Zang*, **You Zhang***, Mojtaba Heydari, and Zhiyao Duan. SingFake: Singing voice deepfake detection. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12156--12160, 2024 (* **equal contribution**).

[14] **You Zhang**, Yongyi Zang, Jiatong Shi, Ryuichi Yamamoto, Tomoki Toda, and Zhiyao Duan. SVDD 2024: The inaugural singing voice deepfake detection challenge. In *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2024.

[15] 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. In *Proc. Interspeech*, pages 4783--4787, 2024.

[16] Sefik Emre Eskimez, **You Zhang**, and Zhiyao Duan. Speech driven talking face generation from a single image and an emotion condition. *IEEE Transactions on Multimedia*, 24:3480--3490, 2022.

[17] Enting Zhou, **You Zhang**, and Zhiyao Duan. Learning arousal-valence representation from categorical emotion labels of speech. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12126--12130, 2024.

[18] Kun Zhou, **You Zhang**, Shengkui Zhao, Hao Wang, Zexu Pan, Dianwen Ng, Chong Zhang, Chongjia Ni, Yukun Ma, Trung Hieu Nguyen, et al. Emotional dimension control in language model-based text-to-speech: Spanning a broad spectrum of human emotions. *arXiv preprint arXiv:2409.16681*, 2024.

[19] Abudukelimu Wuerkaixi, **You Zhang**, Zhiyao Duan, and Changshui Zhang. Rethinking audio-visual synchronization for active speaker detection. In *Proc. IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 01--06, 2022.

[20] Abudukelimu Wuerkaixi, Kunda Yan, **You Zhang**, Zhiyao Duan, and Changshui Zhang. DyViSE: Dynamic vision-guided speaker embedding for audio-visual speaker diarization. In *Proc. IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1--6, 2022.

[21] Kyungbok Lee, **You Zhang**, and Zhiyao Duan. A multi-stream fusion approach with one-class learning for audio-visual deepfake detection. In *Proc. IEEE 26th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1--6, 2024.