Personalized Immersive and Secure Audio for Metaverse

You (Neil) Zhang
University of Rochester

Audio Information Research Laboratory

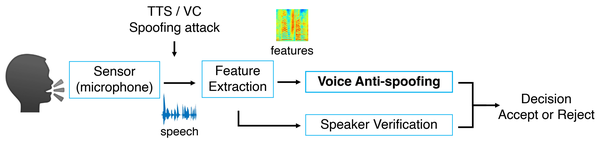
OVERVIEW

Metaverse is a virtual universe comprised of interconnected shared spaces that enable unique experiences through augmented reality (AR) and virtual reality (VR).
My research goal is to design algorithms and systems to support AR/VR with **immersive, personalized, and secure** audio technology.
My work has been focusing on **audio-visual rendering and analysis, personalized spatial audio, and speech anti-spoofing**.

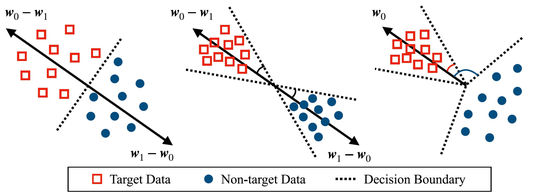
SECURITY

Speech Anti-Spoofing

Synthetic Speech Detection

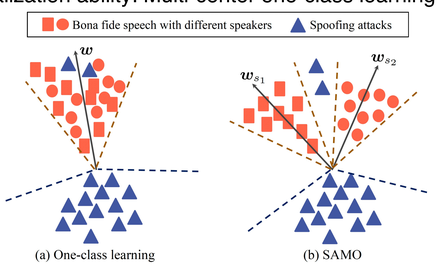


Generalization ability: One-class learning -- OC-Softmax [1, 2]



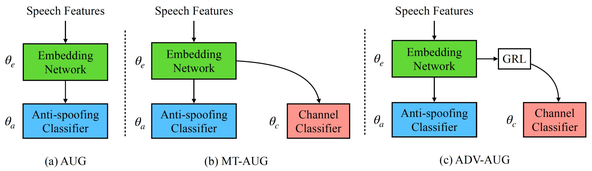
(a) Original Softmax (b) AM-Softmax (c) OC-Softmax (Proposed)

Generalization ability: Multi-center one-class learning [3]



(a) One-class learning (b) SAMO

Channel robustness [2, 4, 5, 6]



(a) AUG (b) MT-AUG (c) ADV-AUG

Joint optimization with speaker verification [7]

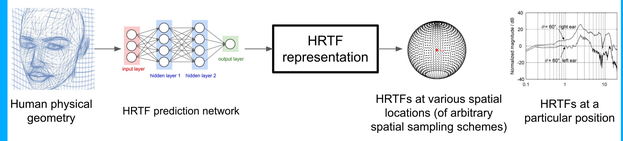
$$P(y^t = 1|x^e_{ASV}, x^t_{ASV}, x^t_{CM}) = P(y^t_{ASV} = 1|x^e_{ASV}, x^t_{ASV}) P(y^t_{CM} = 1|y^t_{ASV}, x^t_{CM}).$$

Future work:
Generalized audio deepfake detection

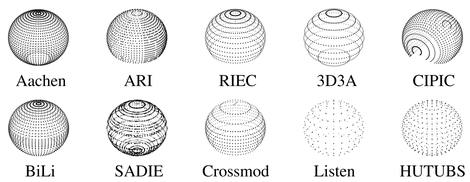
PERSONALIZATION

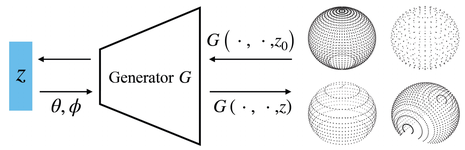
Personalized Spatial Audio

Head-Related Transfer Function (HRTF) Personalization

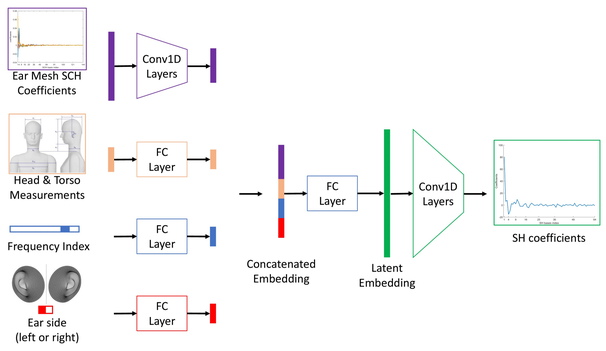


Unified HRTF representation across databases [8, 9]





HRTF personalization across all directions [10, 11]

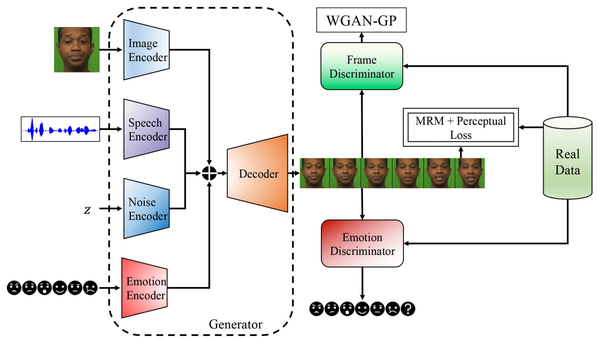


Future work:
Personalized binaural synthesis from mono audio

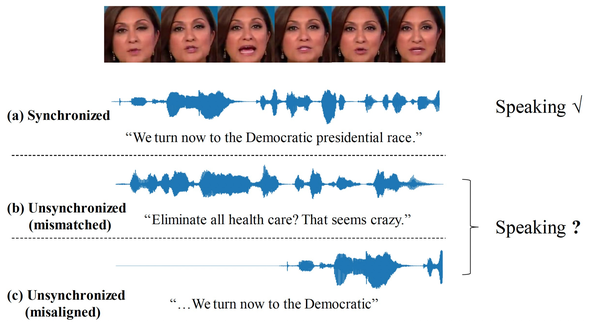
IMMERSIVENESS

Audio-Visual Rendering and Analysis

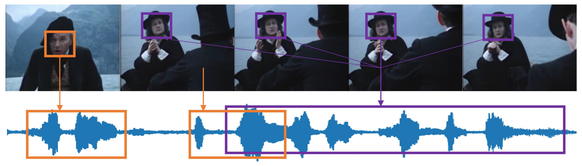
Talking Face Generation: Emotional Rendering [12]



Active Speaker Detection: Audio-Visual Synchronization [13]



Audio-Visual Speaker Diarization: Off-screen Speakers [14]



Future work:
Audio-visual scene understanding

REFERENCES

[1] **Zhang, Y.**, Jiang, F. and Duan, Z., 2021. One-class learning towards synthetic voice spoofing detection. *IEEE Signal Processing Letters*, *28*, pp.937-941.
[2] **Zhang, Y.**, Jiang, F., Zhu, G., Chen, X. and Duan, Z., 2023. Generalizing voice presentation attack detection to unseen synthetic attacks and channel variation. In *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment* (pp. 421-443). Springer Nature Singapore.
[3] Ding, S., **Zhang, Y.** and Duan, Z., 2023. SAMO: Speaker attractor multi-center one-class Learning for voice anti-spoofing. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
[4] **Zhang, Y.**, Zhu, G., Jiang, F. and Duan, Z., 2021. An empirical study on channel effects for synthetic voice spoofing countermeasure systems. *Proc. Interspeech*, pp.4309-4313.
[5] Chen, X.*, **Zhang, Y.***, Zhu, G.* and Duan, Z., 2021. UR channel-robust synthetic speech detection system for ASVspoof 2021. *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pp.75-82. (* equal contribution)
[6] Zang, Y., **Zhang, Y.** and Duan, Z., 2023. Phase perturbation improves channel robustness for speech spoofing countermeasures. *Proc. Interspeech*. Accepted.
[7] **Zhang, Y.**, Zhu, G. and Duan, Z., 2022. A probabilistic fusion framework for spoofing aware speaker verification. *Proc. The Speaker and Language Recognition Workshop (Odyssey)*, pp.77-84.
[8] **Zhang, Y.**, Wang, Y. and Duan, Z., 2023. HRTF field: Unifying measured HRTF magnitude representation with neural fields. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
[9] Wen, Y., **Zhang, Y.** and Duan, Z., 2023. Mitigating cross-database differences for learning unified HRTF representation. Under review.
[10] Wang, Y., **Zhang, Y.**, Duan, Z. and Bocko, M., 2021. Global HRTF personalization using anthropometric measures. In *Audio Engineering Society Convention 150*. Audio Engineering Society.
[11] Wang, Y., **Zhang, Y.**, Duan, Z. and Bocko, M., 2022. Predicting global head-related transfer functions from scanned head geometry using deep learning and compact representations. Under review.
[12] Eskimez, S.E., **Zhang, Y.** and Duan, Z., 2021. Speech driven talking face generation from a single image and an emotion condition. *IEEE Transactions on Multimedia, 24*, pp.3480-3490.
[13] Wuerkaixi, A., **Zhang, Y.**, Duan, Z. and Zhang, C., 2022. Rethinking audio-visual synchronization for active speaker detection. *Proc. IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP)*.
[14] Wuerkaixi, A., Yan, K., **Zhang, Y.**, Duan, Z. and Zhang, C., 2022. DyViSE: Dynamic vision-guided speaker embedding for audio-visual speaker diarization. *Proc. IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*.

Personal Website

https://yzyouzhang.com

ACKNOWLEDGMENTS

My research has been supported by NSF under Grant 1741472, DGE-1922591, and a New York State Center of Excellence in Data Science award.