# Research Statement

Shimaa Ahmed

Speech is a natural form of humans communication which makes it a very convenient vehicle for human-computer interactions as well. This technology has become a reality due to the massive growth of smart devices, the advancement of machine learning algorithms, and the abundance of public speech data. Thanks to these advancements, machines can now understand speech (automatic speech recognition) [4, 5], generate close-to-natural speech (speech synthesis) [6], differentiate between different speakers (speaker recognition) [8], and even detect many paralinguistic features about the speakers [7] such as their emotions, health condition, age, gender, and mental health.

Cloud operators offer speech technologies as a machine learning as a service (MLaaS) model. This model enables the integration of such technologies in many applications and devices that we interact with in our daily life such as voice-enabled devices and vehicles, live caption of online meeting and education platforms, and voice authentication into banks and secure facilities. These technologies are developed with performance and user experience as their main driving objectives. However, they are accompanied by unprecedented privacy, security, and integrity threats that have become more prevalent with their wide deployment. These threats include cloud access to private recordings, unauthorized voice activations of smart speakers, unauthorized voice biometrics collection, and speaker impersonation. Recent privacy regulations, such as the GDPR and CCPA, provide guidelines for protecting the users' privacy. However, current technologies and cloud services fall behind in meeting these requirements, especially for the case of speech data. Thus, *there is a need for practical solutions that fill the gap between the current technologies, the privacy regulations, and the user expectations.*

In my research, I analyze the emerging privacy and security threats accompanying speech technologies and I develop practical systems, based on solid theoretical underpinnings, to mitigate the risks while preserving the utility and convenience of the current technology. First, in the context of cloud-based speech recognition, I propose an end-to-end system (Pr$\epsilon\epsilon$ch [1]) that applies voice conversion and differential privacy to protect the speakers' voice biometrics and textual content, while fully utilizing the accurate cloud transcription services. Second, I quantify the privacy and integrity risks of the *accidental* and *malicious* false activations of voice assistants (VAs). I develop a system (EKOS [2]) to enhance the VAs robustness against these risks. EKOS leverages the physical channel diversity, speech semantics, architectures diversity, and integrates them in an ensemble of models to enhances the robustness of the keyword spotting system against false activations. Third, I argue that voice biometrics should not be used for security critical authentication. I show, physically, that the acoustic channel has a direct impact on the speaker recognition models reliability (Mystique [3]). I design an impersonation attack using only physical (analog) objects, such as a tube, that fools both the speaker identification and the spoofing detection systems. Finally, I work on analyzing the fairness of speech technologies towards underrepresented groups, and quantifying the privacy and security risks that stem from the models performance disparity towards these groups.

# References


[1] Shimaa Ahmed, Amrita Roy Chowdhury, Kassem Fawaz, and Parmesh Ramanathan. Preech: A system for privacy-preserving speech transcription. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 2703–2720. USENIX Association, August 2020.

[2] Shimaa Ahmed, Ilia Shumailov, Nicolas Papernot, and Kassem Fawaz. Towards more robust keyword spotting for voice assistants. In *31st USENIX Security Symposium*, 2022.

[3] Shimaa Ahmed, Yash Wani, Ali Shahin Shamsabadi, Mohammad Yaghini, Ilia Shumailov, Nicolas Papernot, and Kassem Fawaz. Pipe overflow: Smashing voice authentication for fun and profit. *arXiv preprint arXiv:2202.02751*, 2022.

[4] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182, 2016.

[5] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.

[6] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in Neural Information Processing Systems 31*, pages 4480–4490. Curran Associates, Inc., 2018.

[7] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian MüLler, and Shrikanth Narayanan. Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, 27(1):4–39, 2013.

[8] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. X-vectors: Robust DNN embeddings for speaker recognition. In *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, April 2018.