Voice-profiling is the deduction of a speaker’s characteristics from their voice, and has many applications in audio forensics. Characteristics that are determined include the speaker’s gender, age, and ethnicity along with other physical characteristics. Another challenging voice-profiling problem is facial image reconstruction, where an image of the speaker’s face is generated from a voice recording. One of the major challenges in this field is choosing a generalizable representation of voice that encodes all relevant speaker characteristics.

Prior computational voice-profiling techniques modelled the production of voice as a physical system, and defined multiple voice signal features that encode speaker characteristics. Recent advances in artificial neural networks has resulted in an improvement in performance across voice profiling tasks, but such methods are often purely data-driven; the representation and relationships between voice and speaker characteristics are learned from a large dataset, not necessarily leveraging the knowledge-based voice features from prior work.

In this work, we combine domain-specific signal-processing features with state of the art neural network techniques to learn a generalizable audio representation for voice-profiling. The learned representation is evaluated on multiple voice-profiling tasks including speaker identification, gender classification, and age prediction. Additionally, we present a novel framework for face generation tasks and evaluate our voice representation on facial image reconstruction from voice.


Thesis Committee:
Bhiksha Raj and Rita Singh