Using Diffusion Models to Obfuscate Dementia in Speech

Supervisor(s):
Soteris Demetriou (Department of Computing, s.demetriou@imperial.ac.uk)

Project description:
A fundamental step in natural language understanding is converting speech to text, which allows us to leverage advancements in natural language processing to extract the meaning of a person's utterance. Unfortunately, dementia can be inferred from speech transcripts.
Several works already exist aiming to either obfuscate sensitive attributes or user identity. Some follow a style-transfer approach to target gender, sentiment, age, and other attributes . These (a) are limited by their dependence on the availability of training data for the specific attributes and (b) they require training a model for each target attribute which renders the obfuscation method classifier dependent. Medical datasets of conversations and written text for conditions such as dementia are hard to collect and access. Unlike existing style transfer datasets, parallel corpora between patients and control individuals are not available. This limitation calls for methods which can learn to perform obfuscation even in the absence of training data for the target attribute. Others have used differential privacy which can provide provable guarantees. However, these suffer from poor semantic reconstruction as they increase the level of noise to conceal the authorship. Furthermore, it is difficult to interpret the privacy metric's (ε) value when applied to words in a sentence or a document. More importantly, none of the prior works have been applied to obfuscating sensitive medical attributes such as dementia.
In this project, you will be leveraging diffusion models for text paraphrasing aiming to conceal dementia characteristics while preserving the original semantics of the spoken language.

Timeline (tentative):
Oct 2024: Page plan agreed with the overall approach
Dec 2024: Dementia Classifiers developed and evaluated.
March 2025: Initial results showing the efficacy of the approach in concealing dementia.
June 2025: Rigorous evaluation of the approach and comparison with prior works completes - Poster presentation.
August 2025: Research Paper and Thesis writing completes.

Minimum viable thesis:
We have developed dementia classifiers and an initial obfuscation method based on a transformer model for text paraphrasing. A minimum viable thesis will replicate our results and compare them with previously published papers in the area.

Required background & skills:
Familiarity with machine learning frameworks such as PyTorch. Optionally familiarity with NLP or speech processing. Motivation for pursuing a research career, and resourcefulness.

Representative References:

A. Hlédiková, D. Woszczyk, A. Acman, S. Demetriou, and B. Schuller, "Data Augmentation for Dementia Detection in Spoken Language," 2022, doi: 10.21437/Interspeech.2022-10210
L. Zhang-Kennedy, J. Rocheleau, R. Mohamed, K. Baig, S. Chiasson, and H. Assal, {A4NT}: Author Attribute Anonymity by Adversarial Training of Neural Machine Translation. Usenix Security 2018. Video Available: https://youtu.be/yRhj1FfVNVo
Gröndahl, Tommi, and N. Asokan. "Effective writing style transfer via combinatorial paraphrasing." Proc. Priv. Enhancing Technol. 2020.4 (2020): 175-195.
Mahmood, Asad, et al. "A Girl Has No Name: Automated Authorship Obfuscation using Mutant-X." Proc. Priv. Enhancing Technol. 2019.4 (2019): 54-71.