COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

DISSERTATION DEFENSE

Amrit Romana

Transforming Disfluency Detection: Integrating Large Language and Acoustic Models

Tuesday, June 4, 2024
2:30pm – 4:30pm
3941 Beyster
Hybrid – Zoom

ABSTRACT: Speech disfluencies, such as filled pauses or revisions, are disruptions in the typical flow of speech. While all speakers experience some disfluencies, the frequency of disfluencies increases with certain speaker and environmental characteristics, such as a speech disorder or heightened cognitive load. Modeling disfluent events has been shown to be helpful for a range of downstream tasks, and as a result, disfluency detection and categorization has gained traction as a research area. However, variability across speakers and disfluency types make it difficult to develop scalable and generalizable methods for this task.

In this thesis, I begin by exploring disfluencies as a predictor of cognitive impairment. I use these findings to motivate my investigation into models for automatic disfluency detection and categorization. I find that a fine-tuned transformer-based large language model, namely BERT, outperforms previously used neural networks that rely on hand-crafted features. With scalability in mind, I then address the challenge of performing this task without manually transcribed text. I evaluate the potential of detecting disfluencies from automatic speech recognition (ASR) transcripts, including those generated by Whisper, but I find that ASR errors limit performance of the downstream task.

As an alternative approach, I fine-tune the acoustic ASR models directly for disfluency detection from audio, eliminating the intermediate transcription step. I then propose a multimodal model that combines language and acoustic representations. This multimodal approach effectively compensates for ASR errors and outperforms the unimodal models. Lastly, I consider a multi-task training objective, with disfluency detection as the primary task and ASR as an auxiliary task. I find that this multi-task training results in a model that performs similarly to the multimodal model when evaluated in-domain. Importantly, I also find that multi-task training results in a model that generalizes best out-of-domain.

The overarching goal of this thesis is to develop robust and scalable methods for automatically detecting and categorizing disfluencies. These advancements will lay the groundwork for future research to explore disfluencies as potential signals of speaker or environmental characteristics.

CHAIR: Prof. Emily Mower Provost