COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

DISSERTATION DEFENSE

# ZAKARIA ALDENEH

## Automatic Quantification and Prediction of Affect in Spoken Interactions

Thursday, September 24, 2020
10:00 am
Virtual

**ABSTRACT:** Emotional expression plays a key role in interactions as it communicates the necessary context needed for understanding the behaviors and intentions of individuals.

Therefore, a speech-based Artificial Intelligence (AI) system that can recognize and interpret emotional expression has many potential applications with measurable impact to a variety of areas, including human-computer interaction (HCI) and healthcare. However, there are several factors that make speech emotion recognition (SER) a difficult task; these factors include: variability in speech data, variability in emotion annotations, and data sparsity. This dissertation explores methodologies for improving the robustness of the automatic recognition of emotional expression from speech by addressing the impacts of these factors on various aspects of the SER system pipeline.

The first part of the dissertation focuses on addressing speech data variability in SER. Specifically, we propose modeling techniques that improve SER performance by leveraging short-term dynamical properties of speech. Furthermore, we demonstrate how data augmentation improves SER robustness to speaker variations. Lastly, we discover that we can make more accurate predictions of emotion by considering the fine-grained interactions between the acoustic and lexical components of speech.

The second part focuses on addressing continuous emotion annotation variability in SER. Specifically, we propose SER modeling techniques that account for the behaviors of annotators (i.e., annotators' reaction delay) to improve SER robustness.

Finally, the third part focuses on addressing data sparsity in SER. We investigate two methods that enable us to learn robust embeddings, which highlight the differences that exist between neutral speech and emotionally expressive speech, without requiring emotion annotations. In the first method, we demonstrate how emotionally charged vocal expressions change speaker characteristics as captured by embeddings extracted from a speaker identification model, and we propose the use of these embeddings in SER applications.  In the second method, we propose a framework for learning emotion embeddings using audio-textual data that is not annotated for emotion.

The unification of the methods and results presented in this thesis helps enable the development of more robust SER systems, making key advancements toward an interactive speech-based AI system that is capable of recognizing and interpreting human behaviors.

**CHAIR:** Prof. Emily Mower Provost