Contribution ID: **33** Type: **"Classic" talk**

Towards an interpretable and transferable acoustic emotion recognition for the detection of mental disorders

Thursday, 27 May 2021 11:35 (20 minutes)

Motivation
Automatic speech emotion recognition (ASER) refers to a group of algorithms that deduce the emotional state of an individual from their speech utterances. The methods are deployed in a wide range of tasks, including the detection and intervention of mental disorders. State-of-the art ASER techniques have evolved from the more conventional ML based methods to the current advanced deep neural network based solutions. Despite the long history of research contributions in this domain, state-of-art methods still struggle to generalize across languages, between corpora with different recording conditions, etc. Furthermore, most of the methods lack in interpretation and transparency of the models and their decision making process. These aspects are especially crucial when the methods are deployed in applications with impact on human lives.

Contribution
Autoencoders and latent representation studies are useful tools in the exploration of interpretable and generalizable models. We present results on the benefits of using autoencoders and its variants for ASER, predominantly on emotional states like anger, sadness, happiness and the neutral state. We show that the clusters in the latent space are representative of the desired emotional clusters, although some classes of emotions are more discriminative than others. We take a step further to illustrate the use of SHAP and DeepLIFT to gain insights into the feature subsets that contribute to the discriminative clustering of emotion classes in the latent space. Furthermore, we study the robustness of the methods by investigating the differences that occur in the latent representations when the underlying data conditions are modified. In other words, how the differences in the language of the corpus, recording conditions of the corpus~(acted, 'in the wild') manifest in the latent space. In addition, we explore the discrete and continuous scales for their appropriateness in modelling speech emotions and their correspondence to each other. Lastly, we use the feature subset that provides the most stable representations of emotional states over different corpora, languages and recording conditions to transfer the knowledge to languages with few or no labelled emotion corpus.

Primary author: DAS, Sneha (Technical University of Denmark)

Co-authors: LØNFELDT, Nicole Nadine (Region Hovedstadens Psykiatri); PAGSBERG, Anne Katrine (Region Hovedstadens Psykiatri); CLEMMENSEN, Line H (Technical University of Denmark)

Presenter: DAS, Sneha (Technical University of Denmark)

Session Classification: Morning 1

Track Classification: Models and Inference