Lukas Dillingham

ECE 447 Computer Audition

Peer Review: Attention Based 3-D Convolutional Neural Network for Speech Emotion (Meiying Chen)

Meiying introduces a 3-D convolutional neural network model with the goal of detecting emotions in speech. In this case those emotions are anger, sadness, happiness, and normal as defined by the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset. Although no initial results are presented, it is clear that Meiying is very knowledgeable about speech emotion detection and discusses the reasons behind the architecture chosen.

Chen et al. is referenced multiple times when explaining the choice for using 3-D CNNs, but it would be helpful for the reader to have a bit more explanation in this paper as to why Chen et al determines 3-D CNNs are superior to 1-D or 2-D, and why the author agrees with Chen et al. More explanation as to why log-Mel, deltas, delta-deltas are used as the 3-D input would help strengthen her argument about the choice of 3-D CNNs as well. Similarly, some more explanation as to why an attention layer is preferred over an LSTM layer would be helpful rather than just listing the reference. It would be great to have a diagram of the architecture as well.

Overall the paper is very well written and the author appears to understand the problem of emotion detection in speech quite well. It appears that the author is expanding on Chen et al's work and modifying some of the architecture of the model. Should Chen et al be a be included in the base line as well because of this? It will be interesting once some results are presented comparing this method to some other methods. During the poster session, if a live demo allowing a user to speak into a mic and detect the emotion was available, it would be quite impressive.