Exploration of a Multimodal Emotion Recognition System and the Physical Correlates of Emotion

Clodagh Lynch, Master of Science in Computer Science (Data Science)
University of Dublin, Trinity College, 2021

Supervisors: Professor Khurshid Ahmad and Dr. Carl Vogel

Automatic Emotion Recognition (AER) is an emerging area of research that has most recently focused on the use of multiple modalities within an emotion recognition system. This work puts forward a specification of a modality pipeline for multimodal emotion recognition that uses facial expression emotion, speech emotion, and text sentiment. A novel two-step decision-level data fusion process is proposed that fuses the results of each of these modalities to determine if an emotion is deemed strongly evident across each 2000ms segment of a video. The dataset used is comprised of 50 videos of male and female Irish politicians and is used to examine the physical correlates of emotion from an Irish context. The physical correlates of emotion in facial expression are Action Units (AUs) such as from Ekman's FACS, with one of the most insightful physical correlates of emotion in speech being Fundamental Frequency (F0). Findings in this study show the 'most dominant' AUs to be AU6 and AU12 for joy and AU4 and AU9 for anger, which supports some findings in previous literature. Mean F0 is analysed with respect to emotion, gender, and age, with results for the mean F0 for joy and anger being supportive of findings from previous literature, finding both mean F0s to be high. It's widely established in previous literature that the mean F0 of female speakers is higher than male speakers however, the limitation posed by the small sample size in this study hindered this pattern from being observed in Irish-English speakers. Variations in mean F0 with respect to age observed in previous literature are not reflected in results of this study, except the mean F0 of male Irish-English speakers remaining largely stationary between ages 42 to 55, which is supportive of previous research. Systems implemented in this pipeline include Emotient FACET, OpenSMILE, and RockSteady for facial expression, speech, and text processing, respectively. Automatic Speech Recognition (ASR) services are implemented to transcribe audio, with Microsoft Azure Speech-to-Text achieving the best mean WER rate of 7.2% on a subset of this dataset. The systems implemented in this pipeline are not fully understood systems and therefore, this work provides an insight into the calibration of each system with respect to each other. Future work includes implementing machine learning-based text sentiment analysis in order to experiment with other decision-level data fusion equations proposed in previous literature.