Introduction of Problem Context

According to **(Picard),** emotions are a mental state where one experiences pleasure or displeasure with high intensity. They are a very necessary semantic component in human interaction - often, without the context of the speakers emotion, the intentions behind utterances can be ambiguous **(Automatic Emotion Recognition)**. The field of Affective computing aims to understand this phenomenon so that interactions between humans and machines become more naturalistic. Automatic Emotion Recognition has made significant strides in the previous decade but there are still many areas of unexplored territory. Up until the Audio/Visual Emotion Challenge in 2019, there were few works in affective computing recognition literature that supported the common idea that emotions conveyed by facial expressions are mostly universal across cultures **(AVEC 2018, 2019).** There still exist, however, some barriers to universal emotion recognition. It has been found that training machines to recognise emotion from similar language families have shown more accurate results. **(AVEC 2018).**

The dataset we use in this project are audio-visual recordings that have been collected “in the wild”. This phrase simply refers to the idea that standard webcams have been used for recording in a natural setting (home/ work place). The data is not preselected and the behaviour exhibited by the subjects are wholly spontaneous and naturalistic. **(AVEC 2019).** Literature has demonstrated great success with “in the lab” data where the variables are controlled, but this does not account for the noise present in real life situations that “in the wild” data can mimic.

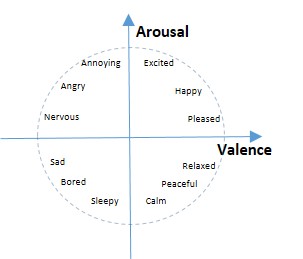
Multiple modalities are exploited in this project meaning that both audio and video data shall be used to extract information about the emotional cues demonstrated. Whilst vocal expressions of emotion have been showed to be less universal across cultures than facial expressions,multi modal systems have achieved relatively high cross cultural accuracies. **(AVEC 2019).**

**Emotional Representation Models**

There are two fundamental approaches when regarding emotion as a measurable concept:

1. Discrete Emotion theory
2. Dimensional models of emotion

Discrete emotion theory describes six basic emotions; anger, disgust, fear, happiness, sadness and surprise **(Ekman, Wikipedia emotion classification).**  Each emotion can be expressed with varying intensities within these discrete categories as opposed to an emotional state. **(Ekman**). There are many algorithms and techniques that realise discrete emotion theory as a classification problem e.g. Support Vector Machines, Naïve Bayes Classifiers and Neural Networks. The focus of this project is on the second approach – treating emotion as a point in a continuous space. This space is described by arousal, valence and dominance which are measures of affective activation, pleasure and control respectively **(Chen, multimodal multitask)**. This plane can be visualised by Thayer’s Arousal-Valence emotion plane **(R. E. Thayer, The Biopsychology of Mood and Arousal. New York: Oxford Univ. Press, 1989.)** Fig 1.

Emotion Recognition based on the emotion plane is a regression problem where observed variables and features are used to predict a real value describing the subject’s emotional state. This approach is free of the ambiguity that is present when treating emotion recognition as a classification problem **(Yang).** Regression is essentially an optimization problem where the value of parameters are set such that the “goodness” of a prediction is maximised in the resulting model. A “good” prediction could be a small prediction error or a large correlation coefficient.**(Continuous Emotion Recognition another look).**

**Machine Learning Methods for Regression**

Automatic Emotion Recognition (AER) is typically tackled through machine learning models and algorithms. The process used by these techniques can be summarized as follows **(Acoustic feature selection)**:

* Feature Extraction Stage:

A feature is an input to a machine learning algorithm. Low Level Descriptor (LLD) features include spectral, prosodic and voice quality information for the audio channel and geometric details for the video channel e.g. face orientation and pixel coordinates for eye points and facial landmarks **(AVEC 2017)**. These features are compiled into feature vectors.

* Data Pre-processing Stage:

Either the most relevant subset of the entire feature set is selected or the number of dimensions in the dataset is reduced.

* Emotion Recognition Stage:

The emotional states of subjects in the recordings are determined by applying machine learning methods to the dataset. The output of these machine learning methods are compared to the labels of the dataset and the parameters of the model are tweaked if necessary. Usually the labels for each recording is merged into a single data series known as the gold standard, which can easily be handled by any machine learning algorithm. **(AVEC 2018)**

The output of the machine learning method is compared to the gold standard using a correlation coefficient. These are measures of how closely the gold standard and the predicted values are correlated. Values of -1.0, 0 and 1.0 indicate perfect negative correlation, no correlation and perfect positive correlation respectively.