Using Etiquette Strategies to Mitigate Frustration in Computer Automated Job Interviews

Student Name: Ali'sa-Falaq Hussain

Supervisor Name: Suncica Hadzidedic

Submitted as part of the degree of BSc Computer Science to the

Board of Examiners in the Department of Computer Sciences, Durham University

Abstract —

Context/Background - Previous work in human computer interaction has demonstrated that recognising a user's affective state, particularly frustration, and adapting the interface in response, can optimise their task performance. Limited research has been conducted on how automated job interview systems can adapt to improve candidate performance.

Aims - We look to extend existing research on the link between user frustration and task performance to a computer automated interview context. We aim to investigate if adapting a system's communication style with different etiquette comments can mitigate a user's frustration and consequently improve interview performance. To achieve this, we look to use convolutional neural network models to monitor changes in affect as exhibited by facial expressions and speech.

Method - A user experiment is conducted on an automated interview platform which outputs particular etiquette comments when frustration is detected in a user's speech and facial expressions. The user's resulting affective states and interview performance is evaluated to determine the impact of these etiquette comments.

Results - There was a strong negative correlation coefficient of -0.85 between frustration and task performance. In addition to this, there was a statistically significant reduction in user frustration when the system used positive politeness and bald style etiquette comments, although these became less effective as the interview progressed.

Conclusions - Frustration negatively impacts task performance in computer automated interviews. To mitigate this frustration, and in turn improve performance, the outputting of positive politeness etiquette comments, and to a lesser extent, bald etiquette comments, is an effective adaptation.

Keywords — Human Computer Interaction, Affective Computing, Convolutional Neural Networks, Etiquette Strategies, Frustration, Interviews.

I INTRODUCTION

Human computer interaction (HCI) is a multidisciplinary field of study which focuses on the design of computer technology and the interaction between humans and computers (Pantic & Rothkrantz 2003). Affective computing is a specific area of HCI referring to computing that relates to, arises from and deliberately influences emotion (Picard 2000). A user's affect, also known as their emotion or mood, is recognised and the system adapts in response to improve the user's experience (Picard 2000). A common affect of interest in research is frustration, one of the most common experiences in HCI (Waterhouse & Child 1953). Defined as an emotional state appearing when obstacles block the possibility of achieving a goal, frustration impacts productivity, learning and creativity (Waterhouse & Child 1953). The majority of existing research focuses on the recognition of emotional states and their impacts on tasks (Giannopoulos et al. 2018). Few have worked on adaptations to mitigate emotions that may impede task performance (Yang & Dorneich 2015). This is a critical part of research since systems which aim to mitigate emotions, particularly frustration, demonstrate an improvement in responsiveness, task success and motivation(Yang & Dorneich 2018) .

Affective computing has been studied rigorously in fields such as education to improve engagement in students (Woolf et al. 2009). However affect is underexplored in other areas, such as computer automated job interviews. This domain is becoming increasingly important with the rise of automated job interviews, accelerated by the Covid-19 pandemic (*HireVue* 2021). We take the opportunity to exploit a user interface to mitigate a candidate's frustration that could impede task performance and consequently confound judgement on their suitability for a job. We aim to achieve this through objectives outlined in subsection ***C***.

A Affect Recognition

Central to affective computing is the recognition of emotional states so that the system can adapt appropriately to optimise user experience. Emotion can be identified through means such as facial features (Barrón-Estrada et al. 2012) and skin conductivity (Woolf et al. 2009). Equipment necessary can vary from expensive specialist equipment (Qu et al. 2005) to a webcam and microphone (Barrón-Estrada et al. 2012). Facial expression analysis is the most popular (Giannopoulos et al. 2018), a large factor being the availability of datasets. A common model of emotions used across research is Ekman's six basic emotions: fear, anger, disgust, sadness, joy and surprise (Ekman 1992). Using this model is advantageous because of the extensive work done to develop methods for recognising these emotions (Giannopoulos et al. 2018). Additionally, Ekman theorises that these six emotions can be combined to form any complex emotions, including frustration (Ekman 1992). This model forms the basis of many emotion datasets validated by research such as the open-source FER-2013 facial expression dataset. An alternate model is the Facial Action Coding System (FACS), used to code movements of specific groups of muscles also known as action units. Although more precise due to its region-based approach (Ekman 2002), publicly available labelled datasets are difficult to source. For this reason we focus on the former approach.

Fewer developments have been made with speech emotion recognition. Existing work tends to focus on methods for feature extraction and analysis. A recent approach, called multi-task learning uses song to supplement speech data. However, it is difficult to best group these two types of data for stable, optimal outputs. (Zhang et al. 2016). The older RELIEF-F algorithm is commonly used to extract prosodic information including pitch, duration and intensity of utterances (Kononenko 1994). The simplicity and stability of this approach makes it appropriate for a basic interview platform. Again, little research has been done in the application of speech emotion recognition to mitigate negative emotions.

B System Adaptations

Affective computing also encompasses system adaptation. An adaptive system is one that changes its behaviour to optimise user experience and task performance (Yang & Dorneich 2018). These changes can be made in response to a range of variables, both environmental, such as temperature, light and screen orientation, and human, such as a user's preferences, goals and emotional state (Picard 2000). Adaptations are taxonomised into four key categories: function allocation, such as the offloading of tasks; task scheduling which entails timings and duration of tasks; content including quality and quantity; and interaction such as style and interface features (Yang & Dorneich 2018). Research, such as in healthcare, is heavy on function allocation and content adaptations (Yang & Dorneich 2018). This often means selecting information that engages users most, and presenting it in a particular way (Bickmore et al. 2010).

Our work focusses on the influence of etiquette strategies, a form of interaction style adaptation, on a user's emotional state. Etiquette strategies define different styles of communication between humans (Brown et al. 1987). There exist four types of etiquette strategies: bald, negative politeness, positive politeness and off-record. Bald is direct, taking no account into the level of imposition to the hearer. An example is, 'answer the question'. Positive politeness minimises imposition between the speaker and hearer with statements of solidarity and compliments. For example, 'you look sad, can I do anything?' (Brown et al. 1987). These human conversational guidelines have been applied to HCI, to make human-computer communications more natural and polite. Bald and positive politeness strategies have combatted frustration and improved user satisfaction in healthcare and education (Yang & Dorneich 2018), (Bickmore et al. 2010).

Frustration is very common in HCI, particularly affecting task performance, which is an issue when computers arguably exist to optimise work (Waterhouse & Child 1953). The recognition and consequent system adaptation to a user's affect to mitigate frustration could improve the interaction between computers and humans, consequently improving task performance. Whether the benefits of etiquette strategies exist within computer automated interviews remains unexplored. From this emerges our aim: to evaluate the impact of bald and positive politeness etiquette strategies prevalent in human interaction on mitigating user frustration and improving user task performance within a computer automated interview. To achieve this we aim to measure the change in a user's affective state in response to etiquette comments.

C Objectives

To address our aim, we split the study into three sets of objectives: minimum, intermediate and advanced.

Minimum:

- A graphical user interface through which the user interacts with the system. This system should present the user with a set of predetermined questions.
- The collection of a user's facial expression data in real-time.
- The training of a CNN to develop a facial expression recognition model.
- A system that uses hard-coded rules to output etiquette comments in response to facial emotional states.

Intermediate:

- The training of a CNN to develop a speech emotion recognition model.
- The collection of a user's speech data in real-time.
- The use of hard coded rules to output etiquette comments in response to speech emotional states.

Advanced:

- A system that can train the facial expression model on a particular user.
- The analysis of how a user's affective state changes, as indicated by both their facial expression and voice after computer feedback.
- The analysis of the impact of etiquette strategies on a user's overall task performance.

II RELATED WORK

Significant work has been done within affective computing to improve a user's system experience. Work generally focuses either on recognition or adaptation. The former investigates methods to identify a user's emotional state, and their underlying machine learning models. The latter evaluates the impact of system adaptation on a user, with some strategies proving more successful than others.

A Emotions in HCI

Frustration when using a system often interferes with a user's completion of tasks. (Waterhouse & Child 1953). An example of such was an operator's task performance in a robot vehicle teleoperating task (Yang & Dorneich 2015). Face tracking and electrodermal activity indicated a decrease in user satisfaction when there was a delay in system response. Studies have looked into how to account for negative affect in the development of systems. In a driving context, a study looked to mitigate frustration with the use of Ambient Light Patterns (Löcken et al. 2017). Woolf et al worked on a tutor system that would encourage the completion of tasks despite students exhibiting negative affective states such as confusion, sadness and frustration (Woolf et al. 2009). This was achieved by mirroring student actions to show empathy and changing the voice and gestures of an avatar. In a gaming application, a scenario would change in response to a user's level of excitation as determined by their heart rate and EEG, producing a more immersive, realistic game experience (Reuderink et al. 2013). We consider if responsive actions such as these can be used to mitigate frustration and consequently improve task performance in interviews.

B Determining the user's affect

A user's affect is often determined using multiple data sources. Barron et al developed an intelligent tutoring system (ITS) (Barrón-Estrada et al. 2012) which adapted to affect recognised from both a student's voice and facial expressions, improving student performance in 80% of cases. An alternate common technique is self-report. When looking into in-vehicle frustration, Locken et al asked participants to feed back their frustration levels (Löcken et al. 2017). Although allowing them to quickly and cheaply assess adaptations, they recognise other modalities should be used to further validate their findings because of the subjectivity associated with the method. In the development of an affect analyser (Pantic & Rothkrantz 2003), visual, audio and tactile information were combined to understand a user's emotional state. Although they emphasise the objective accurate nature of physiological signals, they are described as invasive. The consensus seems that using a combination of methods is the most reliable approach.

Deep learning developments are key to progress in affect recognition. In 2003, Cohen et al demonstrated the use of Bayesian network classifiers in facial expression recognition (Cohen et al. 2003). Work has progressed since then to a more recent focus on CNNs (O'Shea & Nash

2015). This is a result of increased data availability for training neural networks, and advances in GPU technology (Lopes et al. 2017). Lopes et al looked at combining CNNs with specific pre-processing steps such as artificial rotations and translations, producing accuracies of 96.76% on the Cohn Kanade+ dataset (Lopes et al. 2017). This approach also produced speeds which allowed for real time recognition on standard computers, important for the application in this work.

In a tutoring application (Barrón-Estrada et al. 2012), floating search methods are used to extract vocal features, which are then fed into Kohonen Neural Networks, an approach which achieves up to 74.3% accuracy. Alternatively, the RELIEF-F algorithm (Kononenko 1994) was used in the study of vocal emotion expression in call centres to optimise how calls were prioritised (Petrushin 1999). 700 utterances were analysed, with features such as pitch, speaking rate, formants and energy rate extracted and analysed. A 77% recognition accuracy was achieved with an ensemble of neural network classifiers, each trained on a subset of the training set, making an overall decision using the majority voting principle. Although results were impressive, more recent work supports the use of CNNs (Issa et al. 2020), whose structures such as local connectivity and weight sharing make it better suited to dealing with environment and speaker variations, reducing the error rate.

C Adaptating to user affect

Adaptation styles depend on the goal of a system. Barron et al developed an ITS (Barrón-Estrada et al. 2012) which used content and function-allocation style adaptation, specifically presenting exercises to a user depending on affect. This technique resulted in students improving by 13.4% on average. A dynamic decision network was used to determine the tutor's best course of action, depending on learning and affect utility measures. A hospital animated conversational agent for people with depression was developed with interaction-style adaptation (Bickmore et al. 2010), an approach less common in research. This system used different verbal and non-verbal behaviours indicative of various empathy levels, some proving more effective than others. Our study also outputs affect-dependent comments in a similar manner to this hospital system, but with the specifc aim of minimising negative emotion.

D Etiquette Strategies

Human etiquette strategies have been applied to computers to make interactions with humans more natural and polite. A hospital information system was evaluated in terms of politeness and appropriateness (Bickmore et al. 2010). Ratings were highest in bald, positive and negative politeness conditions. Off-record was reported as less subtle and considerate. In terms of learning efficiency, positive and negative politeness strategies have both shown benefits when using a factory training system (Johnson & Wang 2010). It is clear that particular contexts suit different politeness strategies. Extending this, a study aiming to mitigate student frustation in a tutoring system showed that the most effective strategy varied dependent on whether the student was frustrated (Yang & Dorneich 2018). From this, we look at positive politeness and bald strategies to evaluate the effect on user frustration and task performance.

E Summary

Although many studies focus on developing affect recognition methods (Kononenko 1994), investigating the best ways to extract facial or vocal features for recognition (Pantic & Rothkrantz 2003), and the effect of factors on affect (Löcken et al. 2017), fewer investigate adaptations to mitigate negative affective states. Recently, Yang and Dorneich showed that etiquette strategies provide a way for computers to address affect in a natural way resembling human interaction (Yang & Dorneich 2018). However, specific applications of this, particularly automated interviews, remain unexplored. If conclusions are consistent with previous studies (Barrón-Estrada et al. 2012), (Qu et al. 2005), these adaptations should optimise task performance, which would consequently allow accurate judgements to be made on a candidate's job suitability.

III SOLUTION

In this section we present a solution to investigate the extent to which etiquette strategies mitigate frustration in a computer-automated job interview. To begin, a system architecture outline is presented before an in-depth discussion of the solution, tools used and the underlying algorithms in the subsequent subsections.

A System Architecture Overview

An overview of the system we use to address our aim is illustrated in Figure 1. We divide this into the front-end system with which the user interacts (**a**), and the models that feed into this (created with components **b, c** and **d**). The backbone of the basic system is a simple GUI. Through this the questions can be accessed, which users respond to using their microphone and webcam. A set of bald and positive politeness etiquette comments are established and outputted in response to detected frustration. The resulting affective states are recorded.

The facial expression dataset, FER-2013 (Goodfellow et al. 2013), and speech audio dataset, RAVDESS, (Livingstone & Russo 2018) used in component (**b**) and (**d**) respectively are acquired and pre-processed as discussed in the following subsections. Two CNN architectures are developed using Keras, a high-level API for Tensorflow, an open source software library for machine learning (Chollet et al. 2015). Training and testing of the models are important to ensure optimal parameter configuration. The application uses both models simultaneously to recognise the real-time affective state of the user.

The personalised facial expression module, (**c**), involves training the model on an individual participant's facial expressions, who may express emotions in a specific way. This not only increases the dataset size but adds relevant images specific to a participant to the existing dataset - two approaches that should theoretically improve the model (Lopes et al. 2017).

The application was implemented in Python because of the extensive Tensorflow and OpenCV documentation. Additionally the PySimpleGUI library for designing graphical interfaces in Python is robust and well-documented.

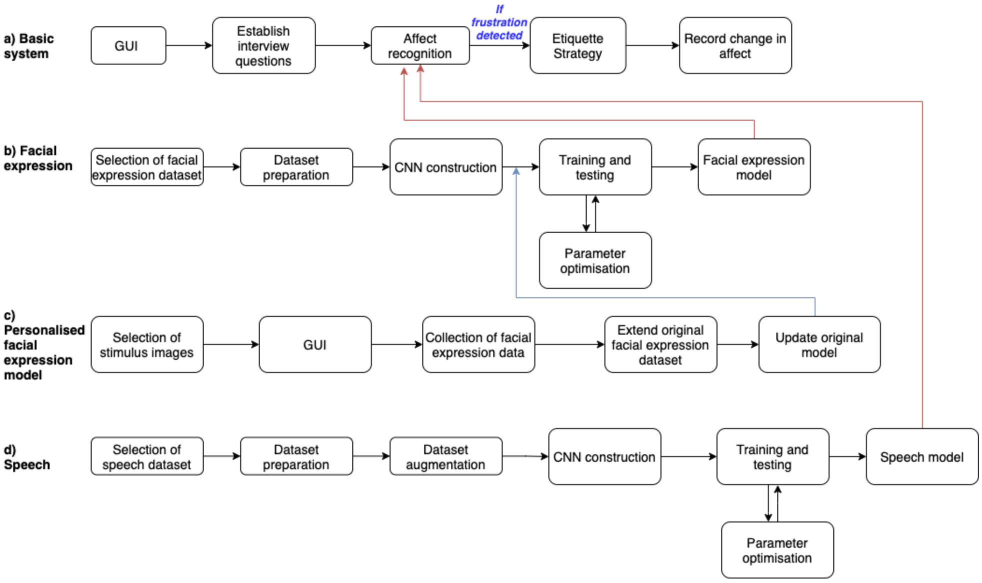


Figure 1: System overview

B Basic System

Structure of the automated interview

The interview structure should resemble a computer automated job interview normally faced by university graduates in terms of style, length and difficulty for reliable, representative results. To establish this level of difficulty, questions were sourced from the Graduate Records Examinations (GRE) (ETS 2021), a standardised test used for graduate school entry in the United States and Canada. There exists a vast amount of available data on the GRE, including the rates of correct answers. It was also used by Yang and Dorneich in their study on mitigating frustration in education (Yang & Dorneich 2018).

The questions are multiple choice verbal reasoning tasks, each for which users have to select two correct answers. As expected in an interview, they must discuss their thought process out loud before selecting their final answers on the interface. Since the GRE leaves an average of 90 seconds per question, this was the standard time allocated per question in the interview. HireVue, an automated computer interview platform that have hosted over 20 million interviews, reported that the average computer-automated interview has between 5 and 8 questions (*HireVue* 2021). As a result, we decided on 6 questions, with a 50% split between easy and hard, similar to the GRE.

We further adjusted the structure of the interview in order to induce frustration. We first described the historically worst answered question as 'easy'. Secondly, we reduced the time allocation for two questions to 50 seconds. These techniques were lifted from Yang and Dorneich's 2018 study (Yang & Dorneich 2018).

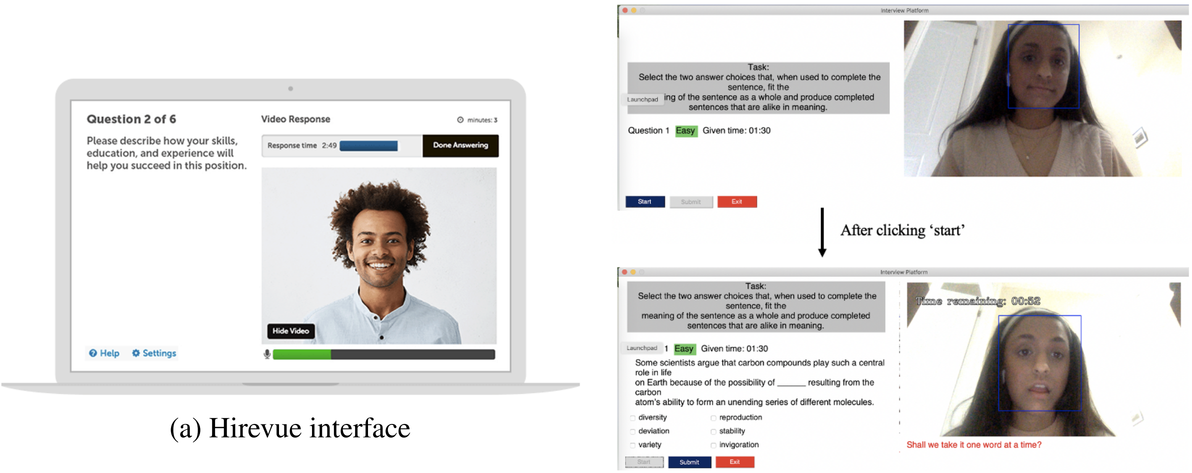
Building a Graphical User Interface

The principal tools used in the design of the interface were PySimpleGUI (Driscoll 2018) and OpenCV (Bradski 2000). Hirevue's interview software interface was analysed and the main features extracted (*HireVue* 2021) . See Figure 2a and 2b for an example of HireVue's interface and our system respectively.

The interface is comprised of PySimpleGUI components. We describe and justify key features below:

- **Webcam Output** OpenCV collects webcam frames in real-time. The window is refreshed every millisecond with the updated frame so the user can view a live stream of themselves.
- **Task Information** The overall task instructions are provided before the timer begins so the user can begin answering the question from the start of the countdown.
- **Allocated time label** The time allocation is displayed before the timer begins to ensure the user is aware of particularly short time constraints, which may contribute towards frustration.
- **Question** After clicking start, and the countdown begins, the question is displayed.
- **Checkboxes** These hold the answer choices and are shown once the user has clicked start. After discussing their thought process, the user checks the final answers they believe to be correct. These are conveniently shown directly below the question. They are disabled once out of time or if the user has clicked 'Submit'.
- **Countdown** A countdown is shown in bold in the top left corner of the webcam output, for easy visibility whenever the user looks at themselves. This is displayed in red when 10 seconds remain to encourage the user to finish off.
- **Etiquette comments** These are outputted in red and bold directly below the webcam output to ensure clear visibility.
- **Final results** These are presented on a separate screen after finishing the entire automated interview so to not confound with the user's affect during the task.

Comprehensive instructions describing each interface feature and the overall task are provided to the user before system use.



Question 2 of 6

Please describe how your skills, education, and experience will help you succeed in this position.

Video Response

minutes: 3

Response time 2:49

(a) Hirevue interface

Interview Platform

Task:
Select the two answer choices that, when used to complete the sentence, fit the
ing of the sentence as a whole and produce completed sentences that are alike in meaning.

Question 1 Easy Given time: 01:30

Interview Platform

Task:
Select the two answer choices that, when used to complete the sentence, fit the
meaning of the sentence as a whole and produce completed sentences that are alike in meaning.

1 Easy Given time: 01:30

Some scientists argue that carbon compounds play such a central role in life
on Earth because of the possibility of ______ resulting from the carbon
atom's ability to form an unending series of different molecules.

□ diversity □ reproduction
□ deviation □ stability
□ variety □ invigoration

Time remaining: 00:52

Shall we take it one word at a time?

(b) Our interface

Figure 2: Existing application Hirevue (*HireVue* 2021) alongside our application

System Feedback

Observations consistently show that positive politeness and bald communication strategies encourage struggling students and improves their confidence and satisfaction (Yang & Dorneich 2018). We investigate whether these two strategies effectively mitigate frustration in our interview. To achieve this, when negative affect is flagged by a user's facial expression or speech, we rotate a response between no feedback, a bald comment and a positive politeness comment. The initial is to establish a baseline against which to compare effects of the strategies. We create bald comments with the use of command words such as 'consider'. To reflect positive politeness we make use of the third person with words such as 'we'. The comments used are outlined in Table 1.

Table 1: FEEDBACK COMMENTS

Bald	**Positive Politeness**
Consider each potential word definition carefully.	We can do this, why don't we consider each word's definition?
Ensure you have read the entire text slowly.	Gettting there, let's make sure we've read through the text slowly!
Consider which words make the text flow best.	Shall we see which words seem most similar?
Make sure the whole text is understood.	Shall we take it one word at a time?
Consider which words sound best in the context.	Let's see if we can figure which words fit best.
Consider which words are most similar.	Good concentration! We can try eliminating unlikely words.

C Affect Recognition

Multimodal systems, which detect emotional states in more than one way are standard in research for their increased reliability (Pantic & Rothkrantz 2003). Woolf et al found non-obtrusive measures including facial expression analysis effective and placed users in a natural setting (Woolf et al. 2009), which is not the case when using expensive, intrusive equipment to measure physiological signals. As a result, we analyse speech and facial expression data, which would naturally be collected in an interview and pose no extra cost.

C.1 Methods for facial expression detection

Key tools used in the facial expression recognition module include Keras (Chollet et al. 2015), the FER-2013 dataset (Goodfellow et al. 2013) and the open source computer vision library OpenCV (Bradski 2000). Google Colaboratory was used for model training because of its access to CUDA, providing GPUs for processing.

The valence-arousal scale is a common way to taxonomise emotions (Buechel & Hahn 2016). A valence value describes how positive or negative an emotion is, and an arousal value describes its intensity. Unfortunately, datasets with labelled frustration images were either not publicly accessible or too small for deep learning purposes. Since Ekman's basic emotions make up all other emotions (Ekman 1992), we detect frustration through basic emotions with the most similar valence and arousal values to frustration. These are anger, fear, disgust and sadness (Buechel & Hahn 2016), which we will refer to as 'negative affective states'. We focus on mitigating these in our approach. We will refer to the remaining basic emotions - happy, surprise and neutral - as 'positive affective states'.

The FER-2013 dataset (Figure 3), based on Ekman's emotion model, was chosen for our application (Goodfellow et al. 2013) because of its substantial size, with 28,709 images; the variety of images, being sourced from the internet as opposed to a limited selection of actors; and its extensive use and validation in research (Woolf et al. 2009).

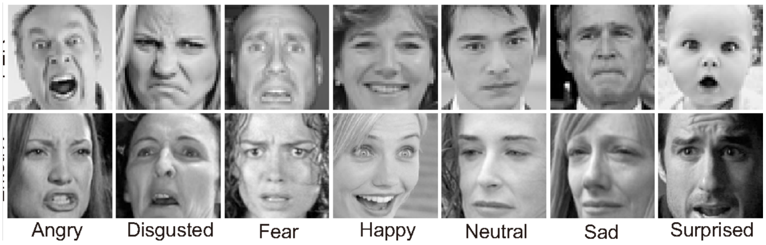


Figure 3: A selection of images from the FER-2013 dataset

Images in the FER-2013 dataset are sized 48x48 and are grayscale. Having a single colour channel, as opposed to three reduces complexity and simplifes the problem so it is less computationally expensive. Data was received as a CSV file, with an integer emotion label (0-6) and 2304 pixel values (0-255) for each image. Pixel values were scaled by 1/255 for more manageable processing. Data was split into training and test sets with proportions 75% and 25% respectively. Finally, we converted the labels to 2-dimensional one-hot encoded arrays.

Using Bayesian networks to classify emotion in facial expressions and speech could be advantageous because of their ability to handle missing data during inference and training (Cohen et al. 2003). However, convolutional neural networks (CNNs) have been used increasingly more with their simpler architecture, faster training and up to 96.7% accuracy (Lopes et al. 2017). For speech emotion analysis, the RAVDESS (Livingstone & Russo 2018) dataset was tested against various models, where CNN proved most accurate, followed by Random Forest (Issa et al. 2020).

A thorough discussion of the underlying theory behind CNNs is beyond the scope of this work, but an overview is given for an understanding of the decisions made in this work. A network is composed of an input, output and hidden layers. These are composed of perceptrons, which each take an input and apply an activation function to produce an output, fed into the next perceptron over a weighted edge. Supervised learning is where data labelled with the expected output (e.g. anger) is inputted into the network. The output is computed, and the network adjusted by changing connection weights to improve classification. Artificial neural networks are used for pattern recognition and classification but are not naturally capable of image feature extraction. Incorporating convolutional and pooling layers produces a CNN, which allows features to be extracted from images (Kalinovskii & Spitsyn 2015). This is because they reduce the number of parameters required to process an image, decreasing computational complexity (Kalinovskii & Spitsyn 2015).

The facial expression recognition CNN architecture is drawn in Figure 4. The blue rectangles represent convolutional layers. Their main objective is to use different weighted kernels to extract high level features from the image. Using more of these layers is appropriate for more complex images (Lopes et al. 2017). Due to the complexity of facial expression features, we use four layers. The kernels are slid over the input image and the dot product is computed on neighbouring pixel values to calculate features for each pixel. (O'Shea & Nash 2015). The CNN learns the weight of the kernels. These features are represented as a two dimensional activation maps which feed into max pooling layers, shown in green.

In the max pooling layers, the activation maps are split into grids and a downsampling function is applied to each grid (O'Shea & Nash 2015). This is necessary to reduce the volume of data associated with an image, and in turn, the computational complexity of processing this data. Within this layer, the activation function is used is ReLU, $y = max(0, x)$, since it is not computationally expensive to run. In addition it is sparsely activated because of the max function which means a perceptron is only activated for a positive input (Kalinovskii & Spitsyn 2015). Increased sparsity often results in concise models with less overfitting and better predicting power. It is also advantageous as it eliminates the problem of the vanishing gradient problem, where the gradient tends to 0 as x tends to + $\inf$ or -$\inf$ (O'Shea & Nash 2015).

We apply dropout layers, drawn in red, three times throughout the network. These randomly set inputs to 0, preventing overfitting. This ensures the model can produce reliable outputs for unfamiliar faces. This is particularly important since our use of several convolutional layers can encourage overfitting (Lopes et al. 2017). Research supports that using dropout directly after the use of activation functions can improve model performance (O'Shea & Nash 2015).

The brown flatten layer is necessary to convert data into a tabular structure for classification at the end. Finally the dense layers, in yellow, are used to shape the output of the network to the 7 classes in our problem (O'Shea & Nash 2015).

We ran the model over 250 epochs several times, adjusting parameters each time, reaching an optimal accuracy of 76%. Our final parameters were a learning rate of 0.0001 and decay of 1×10^{-6} with the Adam optimizer algorithm. Our first two dropout probabilities were set to 0.25, and the final 0.5. The optimised model was exported and applied to the application.

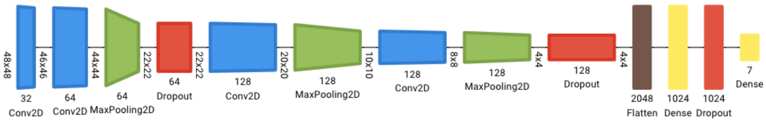


Figure 4: Facial expression recognition CNN architecture diagram

A user's face was detected with a Haar feature-based cascade classifier (Viola & Jones 2001). This is an object detection method that is trained using positive images, which are those with faces, and negative images, which contain no faces. We used a pre-trained model provided by OpenCV which had been validated in research, with better recall values than alternate existing classifiers such as Pico (Kalinovskii & Spitsyn 2015). When a face was detected in a webcam frame, the image was converted to grayscale and resized to 48x48. This was run through the model which outputted a two dimensional array, with each index representing a different emotion. Values in the array summed to one, representing confidence in each affective state. If the overall probability of negative affective states (anger, fear, digust, sadness) outweighed positive (happy, neutral, surprise) we classed the frame as negative, and vice versa.

C.2 A personalised facial expression model

As part of the advanced objectives, we incorporated a module for tailoring the facial expression recognition model towards a specific user. The original dataset is extended with images of a specific user to produce a new model. This requires images of the user naturally displaying each emotional state. Research has used means such as visual stimuli, music and autobigraphical recall to induce emotion. The most effective method has been reported as visual (Siedlecka & Denson 2019). Chosen visual stimuli were taken from the Open Affective Standardized Image Set (OASIS) dataset (Kurdi et al. 2017). These images come labelled with valence and arousal ratings derived from an online study of 822 participants. To determine an appropriate image for each affective state, we refer to literature with empirical evidence of each state's equivalent valence and arousal values (Buechel & Hahn 2016). This is shown graphically in Figure 5. We placed in a GUI images which elicit the same valence and arousal levels as when experiencing each affective state. Surprise was not included since studies have historically struggled to induce the state (Siedlecka & Denson 2019) so images collected would likely confound the model. A sample of images selected are presented in Figure 6. Once a user is exposed to the image, every second frame is recorded, processed and saved to file. Processing the image converts it into the format of FER-2013 images: the frame is converted to grayscale and resized to 48 x 48. Figure 7 shows examples of processed expression images. We combine this data with the FER-2013 images, and retrain the model.

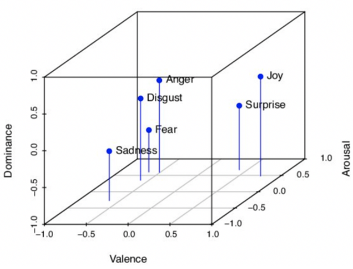


Figure 5: Valence and arousal ratings of the 7 emotions

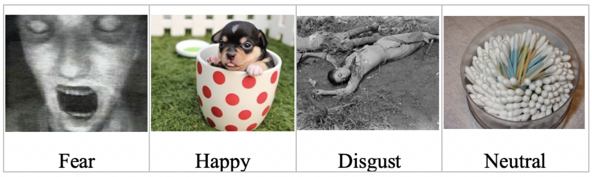


Figure 6: Examples of expressions collected

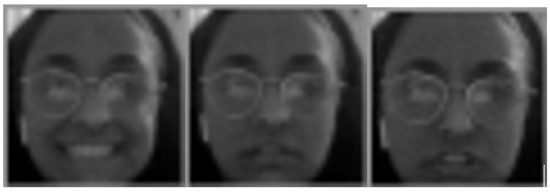


Figure 7: Examples of stimulus images taken from the OASIS dataset

C.3 Methods for speech emotion detection

For speech emotion recognition we utilised the RAVDESS dataset (Livingstone & Russo 2018) which contains 7356 audio (speech and song) and visual recordings of 12 male and 12 female actors pronouncing English sentences with expressions in the categories calm, happy, sad, angry, fearful, surprise, and disgust. The dataset has been used extensively in research, for applications such as the evaluation of machine learning models such as multitask learning (Zhang et al. 2016). This dataset was selected for its availability, research validation and for the emotions available which matched FER-2013. We narrowed data down to speech, of which there were 1440 files. Each clip was around 4 seconds, with the first and final second usually silence.

Data preparation included trimming silence from the clip; relabelling the dataset from 'calm', 'happy' and 'surprise' to 'positive', and 'sad', 'angry', 'disgust' to 'negative'; augmenting prosodic features by stretching the sound, shifting the pitch and adding white noise to improve resistance against noise; and extracting state-of-the-art features called Mel Frequency Cepstral Co-efficients (MFCCs). On the surface, MFCCs are values on the Mel Scale, which is conventional frequency scaled to take into account human perception of frequencies (Issa et al. 2020). We extracted these using the Python audio processing library, Librosa (McFee et al. 2015).

We adapted a CNN from Issa et al's study in 2020 on the use of CNNs for emotion recognition (Issa et al. 2020). Their proposed framework obtained a 71.61% accuracy on the RAVDESS dataset. As shown in the diagram, our model was composed of one-dimensional convolutional layers combined with batch normalisation, dropout and activation layers. Similar to the facial expression CNN atchitecture, the output is activated by ReLU because of its advantages such as faster training. However the speech architecture uses one dimensional convolutional layers, since two dimensions are more suitable for images. After empirically tweaking and testing hyperparameters, we decided on dropout rates of 0.25 and used a particularly large window for our first max-pooling layer, 8, removing a significant volume of data and reducing computational complexity. At the end, the two outputs from the dense layer are normalised using a softmax activation layer, converting them from weighted sum values to probabilities that sum to one. These represent the probability the data belongs to each class, either positive or negative affective states. Our model reached an accuracy of 69.2%.

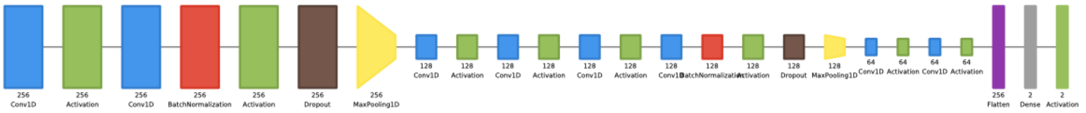


Figure 8: Speech CNN diagram

We refer to a 'clip' as continuous speech input with no pauses. We used the user's microphone to record clips, which were fed into the model in real-time. When the application detected silence the current clip was ended and analysed.

C.4 Combining speech and facial expression data

While the user responds to questions, facial expressions and voice should be processed concurrently. Since Python is naturally synchronous, it was necessary to implement some form of multiprocessing or multithreading. Threads are quicker to spawn than processes, and are in the same memory space, making it easier to share objects, while processes are kept separate. A shared object was necessary to keep track of recognised negative affect. A flag was set either after 8 consecutive frames contained a negative facial expression, or after 3 consecutive negative affect speech clips. When a flag was set, a particular etiquette comment was outputted for 5 seconds, and the detected emotional state from each frame and clip over this time period was recorded and then analysed to determine the impact of this etiquette strategy on affective state.

IV RESULTS

In this section we present and analyse the results of our experiment. We initially detail the experimental settings and our participants before breaking down our analysis into different elements:

we present the average question performance, determine the relationship between a user's negative affective states and performance, and analyse the impact of particular etiquette styles on a user's detected emotional state.

A *Experiment Settings*

The application was compiled as both Windows and Mac executables to maximise the pool of users. Participants were all final year university students from across England, at a stage one would normally complete a computer-based graduate job interview.

Participants were recruited through word of mouth and posts on 5 different university forums across England. Before the experiment, an extensive ethical review (application reference: COMP-2021-01-27T09:44:21-mvlk94) was undertaken because of our use of human participants and their personal data. Each participant read a privacy notice and signed a consent form which outlined the purpose of our data collection and confirmed their understanding of their right to withdraw. They were then each assigned a unique ID - a randomly generated 5 digit number - so the participants could not be identified from their results. A database associating participants with their unique IDs was stored securely and separately from the final results which used the unique IDs.

Due to COVID-19 restrictions, experiments were completed remotely, which made it difficult to directly assist with some technical issues, leading to a selection of participants being unable to complete the experiment. Figure 9 shows an overview of participants recruited, from 48 initial sign-ups to 28 final results for analysis. Issues resulting in this loss of participants included particpants not responding even after further contact attempts, trouble opening the application, corrupted results, and an unforeseen technical issue with question 5. This final issue arose from a lack of testing on Windows laptops of a particular size, which meant part of question 5 and its buttons were off screen, so the experiment could not be completed.

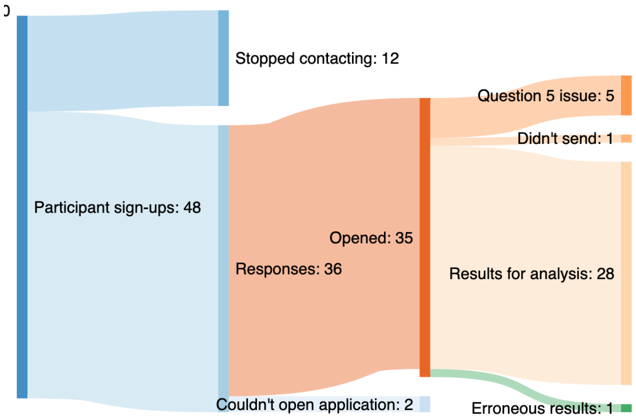


Figure 9: Overview of participants recruited

Participants were provided with comprehensive instructions detailing application setup and the interview. They were asked to carry out the experiment in a quiet environment, as to not interfere with speech recognition, and with nobody else in the webcam's sight, to not interfere with facial expression analysis. The experiment lasted around 10 minutes, and once completed, users sent over anonymised data files, labelled with their unique identifier.

Table 2 shows an overview of questions asked, their labelled difficulty, time allocated and average score across all 28 participants. The rows highlighted are the questions manipulated to induce frustration, through mislabelling or increased time constraints. Question 5, a hard and frustration-inducing task, saw the worst performance. This was followed by another hard task. Surprisingly, question 2, defined as easy by the GRE (ETS 2021), received the third worst performance.

Table 2: QUESTIONS AND THEIR AVERAGE SCORES

Question	Labelled Difficulty	Actual Difficulty	Time Allocated (seconds)	Average Score (%)
1	Easy	Easy	90	68.75
2	Easy	Easy	90	56.25
3	Easy	Easy	50	62.5
4	Hard	Hard	90	43.75
5	Hard	Easy	50	31.25
6	Hard	Hard	90	62.5

B Impact of frustration on task performance

This initial analysis provides indication as to whether the negative affective states underlying frustration impacts a user's task performance when interacting with an automated interview platform, as has been shown in other contexts (Yang & Dorneich 2018). For each question we compare the achieved score with number of times negative affect was detected per 60 seconds. A question was scored at 0%, 50% or 100%, in the case that none, one or both correct answers were selected. Figure 10 shows the relationship between the number of negative emotions detected per minute on a question and the mean question score achieved at that the intensity of frustration.

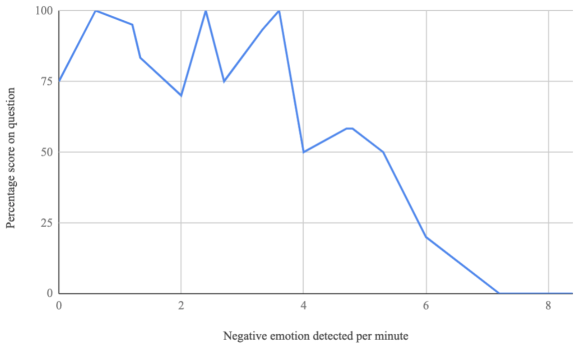


Figure 10: Average question scores at different intensities of negative affective states

Between 0 and 3.6 negative emotions per minute, average scores appear quite erratic, but they all remain above 50%. There is insufficient evidence to confirm why, but this could be indicative of a baseline level of frustration that participants exhibit, even if they perform well. The pearson's correlation coefficient, **r**, as calculated by $\mathbf{r} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}}$ is -0.85, indicating a strong negative correlation, suggesting that performance is detrimentally impacted by negative affective states underlying frustration. This trend is particularly clear in the graph after 3.7 negative emotions per minute.

C Impact of etiquette style on user's affect

To evaluate the impact of each strategy - bald and positive politeness - on a user's affect, we calculate the proportion of affective states detected in the five seconds after a comment was outputted that were positive. We separate the data into question number and feedback type, calculate the mean across participants, and present these results in Figure 11.

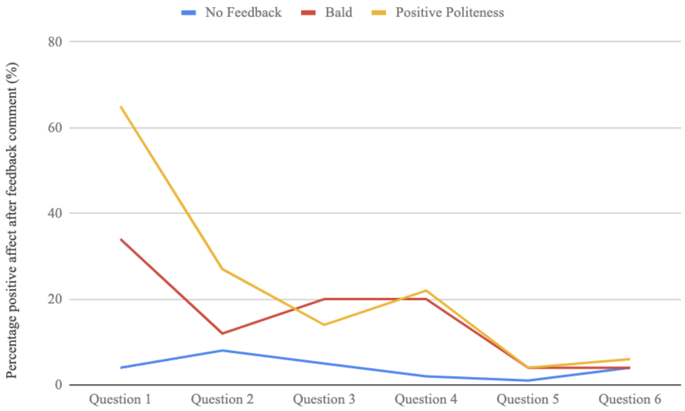


Figure 11: Impact of different strategies on a user's immediate affect

Apart from the bald condition in question 6, the etiquette strategies consistently indicate an increase in the average positive affect detected, when using no feedback as a baseline. Outputting a feedback comment therefore appears beneficial to a user's emotional state. Furthermore, positive politeness comments result in a higher average positive affect in all questions apart from 3, where bald comments performs better, and 5, where both strategies perform the same. These two questions were the shorter, frustration-inducing questions. Data is not sufficient to conclude, but increased frustration could weaken the impact of positive politeness, or make bald the more appropriate strategy. We apply a paired t-test to determine whether the etiquette conditions are statistically significantly different from the control condition. For bald, t = -2.434 and p = 0.017 indicating a 1.7% chance the results are due to chance. Under the positive politeness condition, t = -2.064 and p = 0.032. Therefore the improvement of positive politeness and bald strategies are both statistically significant at $p < 0.05$.

Additionally, another notable trend is that etiquette comments become less effective in later questions. For example there is an 18% difference between no feedback and positive politeness for question one, which decreases to a 2% difference for question 6. More research is required to determine why but this could be attributed to the lack of variety of feedback comments, which users become used to as the exercise continues.

D Custom model

7 of the 28 participants agreed to take part in the preliminary exercise where data was gathered to train tailored facial expression models. For each emotion (except surprise), 20 images were taken from the participant and appended to the existing FER-2013 dataset. Trends detected were very similar between those with custom and standard models as evidenced in Figure 12. Applying a t-test between these two conditions, we calculate t = 1.15 and p = 0.133. There is no significant

difference between the two conditions at $p < 0.05$. In order to evaluate the true accuracy of the custom model over the standard model, a self-report may have been appropriate so detected and self-gauged emotion could be compared.

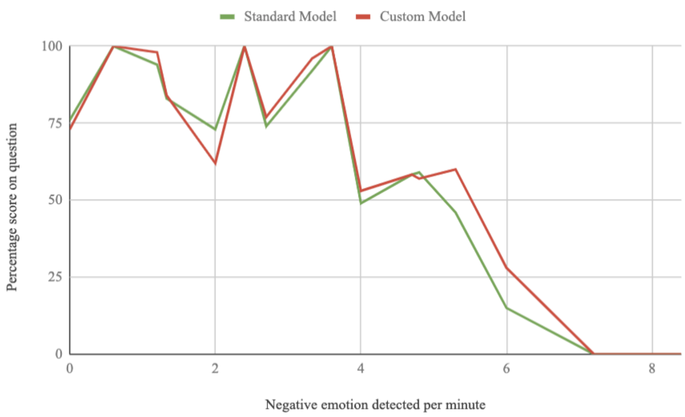


Figure 12: Comparison of trend between frustration and performance using the custom and standard models

V EVALUATION

In this section, we evaluate the strengths and limitations of our system, and consequently determine the suitability of our solution.

A Solution strength and limitations

Our approach carried several strengths, allowing our system to collect ample data that could be analysed to establish reliable trends. Data was collected using webcams and microphones, convenient and inexpensive mediums because of the availability of these in most modern laptops the interview system would be run on. The extensive application of speech and facial expression analysis in research also meant there were multiple validated CNN architectures that could confidently be used and adapted (Issa et al. 2020). An additional strength of our solution is its multimodal nature. Assessing both the user's voice and face in real-time doubles the data available for analysis. As a result, the impact of anomolous results are minimised and trends can be seen more clearly. This increases the reliability of conclusions drawn. Overall, we used a waterfall development approach which was appropriate because there were no stakeholders with changing requirements, and it allowed us to clearly define steps towards a solution. We also introduced an iterative element to development by optimising code in response to regular code reviews. This was advantageous as it brought to light logic errors and code inconsistencies.

On the other hand, we recognise limitations in our solution, providing areas for improvement in future work. The RAVDESS dataset was chosen because of its size and widespread use in research. However, the majority of clips are the same two sentences of less than four seconds. Emotion features in these short clips may not have been complex enough to be effective emotion indicators for the large varied chunks of speech in an interview. Additionally, the clips are all spoken by American actors. Those with different accents or from different cultures may express emotion differently, rendering the model generated as unrepresentative. A further weakness to our solution was the limited data produced. To verify conclusions, our study should be repeated with more questions. In addition, more frustration inducing questions could be included. It appeared that positive politeness was less effective in frustration-inducing conditions, but two instances is not sufficient to draw a conclusion.

In our experiment we evaluated the immediate impact of feedback comments on a user's affect. Although we noticed some trend, it would be interesting to investigate if there are any long-term impacts of these strategies.

B Suitability of the solution

We refer to our objectives in order to determine the suitability of our solution for addressing the investigation of bald and positive politeness etiquette strategies on minimising negative affective states underpinning frustration. We fully satisfied our minimum and intermediate objectives. We created an interview platform with a GUI that provided users with predetermined questions. Multiprocessing was used to simultaneously collect speech and facial expression information in real time. We developed two CNNs: our facial expression model was trained to 76.6% accuracy and speech emotion model to 69.2% accuracy. Etiquette comments were displayed to participants in response to the model outputs. In terms of advanced objectives, collecting facial and speech data enabled us to determine the impact of particular strategies on a user's affect across the interview exercise. It also allowed us to confirm the pattern between task performance and frustration evident in other studies. Additionally, a system was created that elicited and captured particular expressions from participants, which were used to produce a personalised facial expression model. However, this objective was left unsatisfied, since we had insufficient data to evaluate the effectiveness of this model at recognising affect over the standard model. This could be overcome in future by supplementing facial expression and speech data with self-reports to enable a comparison between detected and actual affect. Overall, our application satisfied the majority of our objectives, proving it suitable for addressing our aim. In future, using and evaluating models produced from alternate machine learning methods, such as Bayesian networks could further confirm the suitability of our solution, or improve it, by producing a faster, more accurate way for detecting a user's affect and its changes in response to etiquette strategies.

Although our solution has its shortcomings, our application effectively addresses the problem being investigated. We successfully determined the impact of bald and positive politeness etiquette comments on mitigating user frustration and how this impacts task performance.

VI CONCLUSIONS

In this project, we produced a solution to investigate the mitigation of frustration in computer automated job interviews. We achieved this by focussing on a form of interaction-style adaptation: the use of etiquette strategies prevalent in human interaction. The solution was composed of the development of an interface a user would navigate through as they answer questions, affect recognition and system adaptation. Recognition involved identifying emotional states that underly frustration as defined by the valence-arousal scale of emotion: anger, sadness, disgust and fear. We trained CNN models on facial expression and speech data, using architectures adapted from literature (Lopes et al. 2017) (Issa et al. 2020), which would then be used to analyse user emotion in real time. Finally, adaptation entailed outputting bald and positive politeness style comments when negative affective states were consistently observed. User affect immediately after adaptation was observed to determine the impact of each strategy.

Our work successfully discovered a trend indicating that frustration in a computer automated interview negatively impacts task performance. This extends previous similar findings in the context of human controlled robots and tutoring systems (Yang & Dorneich 2015) (Yang & Dorneich 2018). Additionally, our work provides evidence that both bald and positive politeness comments reduce negative affect in participants. In the majority of cases, positive politeness comments proved most effective.

In the future, work should be carried out to determine the impact of indirect and negative politeness etiquette strategies on the minimisation of frustration in computer automated interviews. It would also be valuable to validate this work by reproducing findings with alternate affect recognition methods such as heart rate or self reports.

The conclusions of this work have several real world applications. Positive politeness and bald etiquette comments could be incorporated in computer automated job interviews to mitigate frustration and in turn improve performance. This would enable candidates to demonstrate their true potential, unconfounded by external causes of frustration. After further context-specific research, these techniques could be incorporated in other computer systems to improve user task performance.

References

Barrón-Estrada, M. L., Zatarain-Cabada, R., Pérez, Y. H. et al. (2012), An intelligent and affective tutoring system within a social network for learning mathematics, *in* 'Ibero-American Conference on Artificial Intelligence', Springer, pp. 651–661.

Bickmore, T. W., Mitchell, S. E., Jack, B. W., Paasche-Orlow, M. K., Pfeifer, L. M. & O'Donnell, J. (2010), 'Response to a relational agent by hospital patients with depressive symptoms', *Interacting with computers* **22**(4), 289–298.

Bradski, G. (2000), 'The OpenCV Library', *Dr. Dobb's Journal of Software Tools* .

Brown, P., Levinson, S. C. & Levinson, S. C. (1987), *Politeness: Some universals in language usage*, Vol. 4, Cambridge university press.

Buechel, S. & Hahn, U. (2016), Emotion analysis as a regression problem—dimensional models and their implications on emotion representation and metrical evaluation, *in* 'Proceedings of the Twenty-second European Conference on Artificial Intelligence', pp. 1114–1122.

Chollet, F. et al. (2015), 'Keras'.
URL: *https://github.com/fchollet/keras*

Cohen, I., Sebe, N., Gozman, F., Cirelo, M. C. & Huang, T. S. (2003), Learning bayesian network classifiers for facial expression recognition both labeled and unlabeled data, *in* '2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.', Vol. 1, IEEE, pp. I–I.

Driscoll, M. (2018), 'Pysimplegui', https://github.com/PySimpleGUI/PySimpleGUI.

Ekman, P. (1992), 'An argument for basic emotions', *Cognition & emotion* **6**(3-4), 169–200.

Ekman, P. (2002), 'Facial action coding system (facs)', *A human face* .

ETS (2021), 'Gre® general test', https://www.ets.org/gre. (Accessed on 04/26/2021).

Giannopoulos, P., Perikos, I. & Hatzilygeroudis, I. (2018), Deep learning approaches for facial emotion recognition: A case study on fer-2013, *in* 'Advances in hybridization of intelligent methods', Springer, pp. 1–16.

Goodfellow, I., Erhan, D., Carrier, P.-L., Courville, A. & Mirza, M. (2013), 'Challenges in representation learning: A report on three machine learning contests'.

HireVue (2021), https://www.hirevue.com/platform/online-video-interviewing-software. (Accessed on 04/26/2021).

Issa, D., Fatih Demirci, M. & Yazici, A. (2020), 'Speech emotion recognition with deep convolutional neural networks', *Biomedical Signal Processing and Control* **59**, 101894.
URL: *https://www.sciencedirect.com/science/article/pii/S1746809420300501*

Johnson, W. L. & Wang, N. (2010), 'The role of politeness in interactive educational software for language tutoring', *Human-Computer Etiquette: Cultural Expectations and the Design Implications They Place on Computers and Technology. Auerbach Publications* pp. 91–113.

Kalinovskii, I. & Spitsyn, V. (2015), 'Compact convolutional neural network cascade for face detection', *arXiv preprint arXiv:1508.01292* .

Kononenko, I. (1994), Estimating attributes: Analysis and extensions of relief, *in* 'European conference on machine learning', Springer, pp. 171–182.

Kurdi, B., Lozano, S. & Banaji, M. R. (2017), 'Introducing the open affective standardized image set (oasis)', *Behavior research methods* **49**(2), 457–470.

Livingstone, S. R. & Russo, F. A. (2018), 'The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english', *PloS one* **13**(5), e0196391.

Löcken, A., Ihme, K. & Unni, A. (2017), Towards designing affect-aware systems for mitigating the effects of in-vehicle frustration, *in* 'Proceedings of the 9th international conference on automotive user interfaces and interactive vehicular applications adjunct', pp. 88–93.

Lopes, A. T., de Aguiar, E., De Souza, A. F. & Oliveira-Santos, T. (2017), 'Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order', *Pattern Recognition* **61**, 610–628.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E. & Nieto, O. (2015), librosa: Audio and music signal analysis in python.

O'Shea, K. & Nash, R. (2015), 'An introduction to convolutional neural networks', *arXiv preprint arXiv:1511.08458* .

Pantic, M. & Rothkrantz, L. J. (2003), 'Toward an affect-sensitive multimodal human-computer interaction', *Proceedings of the IEEE* **91**(9), 1370–1390.

Petrushin, V. (1999), Emotion in speech: Recognition and application to call centers, *in* 'Proceedings of artificial neural networks in engineering', Vol. 710, Citeseer, p. 22.

Picard, R. W. (2000), *Affective computing*, MIT press.

Qu, L., Wang, N. & Johnson, W. L. (2005), Using learner focus of attention to detect learner motivation factors, *in* 'International Conference on User Modeling', Springer, pp. 70–73.

Reuderink, B., Mühl, C. & Poel, M. (2013), 'Valence, arousal and dominance in the eeg during game play', *International journal of autonomous and adaptive communications systems* **6**(1), 45–62.

Siedlecka, E. & Denson, T. F. (2019), 'Experimental methods for inducing basic emotions: A qualitative review', *Emotion Review* **11**(1), 87–97.

Viola, P. & Jones, M. (2001), Rapid object detection using a boosted cascade of simple features, *in* 'Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001', Vol. 1, IEEE, pp. I–I.

Waterhouse, I. K. & Child, I. L. (1953), 'Frustration and the quality of performance.', *Journal of personality* .

Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D. & Picard, R. (2009), 'Affect-aware tutors: recognising and responding to student affect', *International Journal of Learning Technology* **4**(3-4), 129–164.

Yang, E. & Dorneich, M. C. (2015), The effect of time delay on emotion, arousal, and satisfaction in human-robot interaction, *in* 'Proceedings of the human factors and ergonomics society annual meeting', Vol. 59, SAGE Publications Sage CA: Los Angeles, CA, pp. 443–447.

Yang, E. & Dorneich, M. C. (2018), 'Evaluating human–automation etiquette strategies to mitigate user frustration and improve learning in affect-aware tutoring', *Applied Sciences* **8**(6), 895.

Zhang, B., Provost, E. M. & Essl, G. (2016), Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach, *in* '2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', IEEE, pp. 5805–5809.