Computational Emotional Intelligence

From Facial Expressions of Emotion

A dissertation submitted in partial fulfilment of

the requirements for the degree of

BACHELOR OF SCIENCE in Computer Science

in

The Queen’s University of Belfast

by

Robert Palmer

Tuesday 9th May 2017

**Declaration of Academic Integrity**

Before signing the declaration below please check that the submission:

1. Has a full bibliography attached laid out according to the guidelines specified in the Student Project Handbook
2. Contains full acknowledgement of all secondary sources used (paper-based and electronic)
3. Does not exceed the specified page limit
4. Is clearly presented and proof-read
5. Is submitted on, or before, the specified or agreed due date. Late submissions will only be accepted in exceptional circumstances or where a deferment has been granted in advance.

**I declare that I have read both the University and the School of Electronics, Electrical Engineering and Computer Science guidelines on plagiarism - http://www.qub.ac.uk/schools/eeecs/Education/StudentStudyInformation/Plagiarism/ - and that the attached submission is my own original work. No part of it has been submitted for any other assignment and I have acknowledged in my notes and bibliography all written and electronic sources used.**

*Student’s signature* *Date of submission*

**Declaration of Academic Integrity**

Before signing the declaration below please check that the submission:

1. Has a full bibliography attached laid out according to the guidelines specified in the Student Project Handbook
2. Contains full acknowledgement of all secondary sources used (paper-based and electronic)
3. Does not exceed the specified page limit
4. Is clearly presented and proof-read
5. Is submitted on, or before, the specified or agreed due date. Late submissions will only be accepted in exceptional circumstances or where a deferment has been granted in advance.

**I declare that I have read both the University and the School of Electronics, Electrical Engineering and Computer Science guidelines on plagiarism - http://www.qub.ac.uk/schools/eeecs/Education/StudentStudyInformation/Plagiarism/ - and that the attached submission is my own original work. No part of it has been submitted for any other assignment and I have acknowledged in my notes and bibliography all written and electronic sources used.**

*Student’s signature* *Date of submission*

**Acknowledgements**

I wish to give thanks to my supervisor Neil Robertson, for all his advice and support throughout the project.

I would also like to thank Mark Burnett for his aid in creating a suitable dataset.

My thanks go out to all those who volunteered their time and efforts to create an interesting dataset to aid in proving the concept of such systems.

Many thanks to Ed Goldswain for his enthusiasm and his permission to use and publish adaptations of the work at nosweatshakespeare.com.

I would like to thank my parents, my family, my friends, my teachers, my lecturers and all those who aided my learning throughout my life, who have enabled me to pursue such research and find joy and excitement in my career.**Abstract**

This paper and the corresponding project act as a proof of concept of the utility of facial expressions of emotion as a modality. Their intention is to explore the entire process of creating useful applications based on this modality to uncover challenges and determine their viability.

This paper addresses issues including: the scientific basis of multimodal communication, affective computing and facial expressions of emotion; potential applications of derived emotional information such as improving chat bots or selecting key scenes from a speech and their implementation; issues with procuring a valid dataset and evaluating the reliability of a dataset’s labels; metrics for evaluating the success of implementations; the challenges facing such applications and what advances are necessary to bring such applications to fruition.

**Contents**

**Chapter 1 Literature Review7**

1.1 Multimodal Communication**7**

1.2 Affective Data & Facial Expressions**8**

1.3 Decoding Facial Expressions**8**

1.4 Criticisms of Facial Expressions of Emotion**9**

1.5 Applications**10**

**Chapter 2 Project Specification12**

2.1 Emotion Aware Chat Bot**12**

2.2 Sentiment & Emotion Aware Chat Bot**12**

2.3 Key Scene Selection**13**

2.4 Assumptions**13**

2.5 Dataset Requirements**14**

**Chapter 3 Chat Bot Implementation15**

3.1 Unimodal Chat Bot**15**

3.2 Emotion Aware Chat Bot**16**

3.3 Sentiment & Emotion Aware Chat Bot**17**

3.4 Creating a Dataset**18**

3.5 Evaluation of Results**18**

**Chapter 4 Key Scene Selection19**

4.1 Models**19**

4.1.1 Emotionally Charged Moments**19**

4.1.2 Stochastic Processes with the Markov Property **19**

4.1.3 Emotional State Transitions**20**

4.1.4 Emotional Intensity Transitions**21**

4.2 Establishing the Dataset Domain**24**

4.2.1 Genuine vs Acted Scenes**24**

4.2.2 Overcoming Acted Data Issues**25**

4.2.3 Use Case: Shakespeare**25**

4.2.4 Difficulties with Shakespeare**26**

4.3 Creating a Dataset**27**

4.3.1 Sourcing Speeches**28**

4.3.2 Volunteer Session**28**

4.4 Labelling Data**29**

4.4.1 Translating Feedback to Labels**30**

4.4.2 Producing Comparable Results**30**

4.5 Evaluating Dataset Label Reliability**31**

4.6 Evaluating Model Results**32**

4.6.1 Metrics**32**

4.6.2 Issues/Considerations**33**

4.6.3 Setting Model Parameters**34**

4.6.4 Results**35**

4.6.5 Explanations for Results**36**

**Chapter 5 Conclusion38**

**Bibliography40**

**Appendix A: Emotional State Transition Matrix42**

**Appendix B: Speech Extracts43**

**Appendix C: Feedback Form Template46**

**Appendix D: Dataset Label Reliability Results47**

**Appendix E: Key Scene Selector Model Results48**

**Appendix F: User Manual49**

**Chapter 1: Literature Review**

To explore the potential of facial expression of emotion as a modality to computationally infer emotional intelligence onto a system, it is important to first establish its scientific underpinnings. This chapter shall explore our understanding of human communication, the role which emotions play in such communication, the basis for using facial expressions to determine emotions, and potential applications from literature.

**1.1 Multimodal Communication**

Natural language understanding is a field which has seen dramatic improvements in recent years [2] and is now a part of our everyday lives [3]. However, with people’s use of sarcasm, innuendo, irony and disingenuous comments, the true meaning can be missed when analysing the chosen words alone [4]. Human communication is multi-modal [5] and complex, with verbal communication accompanied by changes in vocalisation such as intonation, pitch and rhythm, as well as non-verbal modalities such as gestures, movement and facial expressions [6] [4].

It is well established that multimodal systems can significantly outperform unimodal systems. The combination of aural and visual modalities provides richer information than a single modality alone. The combination of streams of data allows the brain or a system to compensate for incomplete information and to gain a deeper insight into the intent of a user by detecting the semantic and affective information they provide [4] [5]. A lack of visual cues can make communication difficult for humans, further emphasising the necessity of these modalities to understand the true intention of a person [7].

Of the possible audio or visual modalities which could be used in tandem with natural language understanding, [5] found that in 90% of the studies they analysed, the visual modality was superior, with the audio modality suffering from the presence of noise. Nonverbal streams of data may also be more reliable than verbal streams, as nonverbal communication is rarely “under the sustained, conscious control of the communicator” which limits the opportunity for deception [8].

**1.2 Affective Data & Facial Expressions**

It is believed that the affective (emotional) state of the brain plays an important role in collaboration with the cognitive aspects, to determine a human’s actions and behaviours. This is a key component of humanity which is often neglected in analysis and research, despite the knowledge that “the passions” can take control of the mind [9].

Nonverbal communication provides the richest and most accurate source of information regarding emotional states [8] with semantic analysis of the chosen words known to be insufficient for determining affective content [5].

Of the nonverbal modalities, it is known that facial expressions are primary cues for understanding emotions and sentiments, spontaneously expressing emotionality [5] [9]. It is believed facial expressions have evolved to be a near-optimal system of communication, providing important information as quickly as possible [10]. Facial expressions reveal a person’s intentions, provide social cues and aid us in establishing information about a person’s general character [11], with the face being as [9] phrases it “an exquisitely flexible communicative device”. It is clear that in tandem with natural language, facial expressions can be a powerful tool to enrich our understanding of a person’s intentions [5].

**1.3 Decoding Facial Expressions**

We have established above that facial expressions can be used as a rich source of information for gaining a much deeper understanding of a person’s character, and in the immediate moment, their intentions. One of the main challenges we face is in decoding the information expressed.

A lot of research has been carried out with regards to recognising emotion in facial expressions, and this ability is now widely available thanks to numerous APIs, namely: Emotient, EmoVu, nViso, Kairos, Project Oxford, Face Reader, Sightcorp, SkyBiometry and Affectiva [5]. All of these APIs concentrate on detecting between six to eight “basic”, “universal” emotions.

The basis for these basic, universal facial expressions of emotion comes from the work of Ekman et al. [12] considered pioneers in this research [5]. Ekman proposed that there are six basic emotions expressed through the face which are expressed across cultures: happiness, sadness, anger, fear, surprise and disgust.

Additional research has supported basic emotion theory to some degree. For example, high coherence has been found between amusement and smiling [13], and the data resulting from the work in [10] suggests these six emotions have arose from a simpler system of communication in early man. However, much work suggests that identification of these emotional displays does not represent the entirety of the message being conveyed through facial expressions. It would appear that analysis of the sequence of facial expressions transmitted over time is key to decoding the signals portrayed [10].

The support for basic emotion theory tends to derive from an evolutionary perception of facial expressions. The basic emotional expressions are treated as context-free, biological functions. It is assumed there are also context-bound sociocultural expressive functions [6] which this theory does not address.

There have been a number of approaches to try to address this issue, from treating emotion as multi-dimensional, such as with values for arousal (state of being awoken) and valence (positive or negative), or up to four dimensions. An alternative is the “Hourglass of Emotions” which has four affective dimensions with additional labels, and can represent cases where up to four emotions are expressed at the same time [5]. Emotient’s API provides additional information to aid in this way, as alongside analysis of seven basic emotions, it detects overall sentiment, action units and constructs two advanced emotions of confusion and frustration, assuming these can be built from the basic emotions [14].

**1.4 Criticisms of Facial Expressions of Emotion**

Other evidence completely disputes basic emotion theory and undermines the assumptions made in many research projects and industry systems. [13] claims that the “traditional” facial expressions for surprise and disgust appear only in a minority of cases. It additionally states that there is limited evidence concerning sadness, anger and fear and that what evidence does exist, points to low coherence between the emotion and the expressions for anger and fear.

Some suggest that rather than facial expressions of emotion being biologically based, that they are primarily social constructs, which cannot be interpreted without context and social analysis. One major criticism of the universality of Ekman’s studies comes from the relationship between emotions and language, and how Ekman’s labels cannot even be literally translated into some other languages, for example some do not have a word for “fear” [5].

[15] suggests that previous studies have been flawed, through evaluating success by giving volunteers emotions to match to images, which leads to volunteers acting on their expectations, thus giving unrealistic and biased results. Russell claims the “basic emotions” are simply “folk theory” passed down through the generations and influencing our expectations, but not how we truly interpret others when communicating. When observers see spontaneous rather than posed faces, the matching scores plummet.

Studies have shown that young children interpret faces in terms of valence rather than emotion, and as they age they differentiate into adult emotion concepts, which include valence, arousal and discrete emotions. However Russell found that context is far more powerful than facial expressions when determining valence or discrete emotions. [15]

Basic emotion theory suggests that a facial expression is an automatic signal of a specific emotion, or part of that emotion. In this case, context would simply give us a probabilistic likelihood of such an emotion occurring, since people react differently to the same situation. Russell found that this was not the case, as the interpretation of the emotion being displayed correlated with context to a much greater degree than with the facial expression. This leads Russell to disregard basic emotion theory, as he claims observers use facial information to determine not just emotion, but psychological state. Such interpretation relies heavily on context and many modalities such as words, gaze, body position and vocal cues. [15]

It is not in dispute that facial expressions are a source of key information when it comes to understanding human communication, or that facial expressions can and do express emotions, however, there is disagreement as to how they can be used to measure emotionality or derive a psychological state. [9]

**1.5 Applications**

The most popular application of affective analysis would appear to be analysing public sentiment for purposes such as marketing campaigns. For example, mining videos of product reviews on YouTube to evaluate a product’s popularity [16] [5].

There is huge potential to improve the field of Human Computer Interaction, by providing computers with emotional intelligence via Sociocognitive Language Processing. There are complications in human communication which cannot be overcome with traditional natural language understanding alone. A computer system could potentially handle sarcasm, irony, humour, and more by integrating “soft factors” of communication into a natural language processing system. This can enable us to much more accurately computationally analyse the true intentions of people during communication, and to create empathetic, socially competent machines [4].

An explicit example of when facial expression analysis could improve a natural language processing system, is when the chosen words and the facial expressions of the person portray a conflicting meaning. People employ a systematic and consistent approach to determine the meaning of conflicting cues, with facial expression being the most important indicator, followed by tone of voice, with the chosen words having the least value. Therefore, when a conflict in messages is detected, facial expression is relied upon as portraying the true message of the communicator [8]. Given this systematic approach by humans, it would be safe to assume it has merit and could be applied to a computer system.

**Chapter 2: Project Specification**

The intention of this project is to act as a proof of concept of the utility of analysing facial expression of emotion, and to demonstrate how current systems could be vastly improved right now with this readily available technology.

There are a number of example scenarios which will be explored.

**2.1 Emotion Aware Chat Bot**

With the dramatic rise of personal assistants [3], and the prediction of a booming market in the Internet of Things, “chat-bots” are systems which are becoming ever more prevalent and important. These systems rely heavily on context-aware communications [17] and could immediately benefit from the greater awareness extra modalities provide.

For the first example scenario, a typical unimodal chat bot will be created, which listens to the audio stream of a person speaking, parses the words, determines their intent and responds appropriately.

The system will then be improved by additionally analysing a video stream alongside the audio stream. This stream will be used to analyse the user’s facial expressions, and determine a sequential stream of facial expressions of emotion. The facial expressions of emotion will be combined with the determined intent to provide a fine-tuned, emotionally aware response.

**2.2 Sentiment & Emotion Aware Chat Bot**

The second example scenario will build upon the first example scenario, by taking the “chat bot” and integrating sentiment analysis of the chosen words, plus the sentiment of the expressed emotions. The response of the system will be adjusted based on whether these sentiments reinforce each other or contradict each other. This is an application of the discovery outlined in the literature review and in [8] that people rely on the meaning conveyed through the face when there is a conflicting message between the chosen words and the facial expression.

**2.3 Key Scene Selection**

To select key moments from a speech, a unimodal system with an audio stream will only be able to base its decisions on the semantics of the chosen words, picking out nouns, expressive adjectives and perhaps analysing the sentiment of sentences. In many cases this system would miss important moments which are emotionally charged, or have deeper nuances such as sarcasm, or have an unexpected change in portrayal.

For the third example scenario, a system to select the key moments of the communication will be created. The system will explore selecting key scenes by examining the sequential transition between different emotional states determined by the communicator’s facial expressions, as well as the strength and sentiment of the displayed emotions.

**2.4 Assumptions**

Example scenario one assumes that basic emotion theory, as outlined in the literature review, holds true. As established above, it is the principle theory in the analysis of the messages conveyed through facial expressions. The criticisms of this theory either suggest this model is too simplistic but has some merit as basic building blocks, or in the case of Russell, dismiss it due to “context” being a more valid signal of a person’s affective state. This paper argues that while not perfect, this assumption is valid for improving current models, as determining the context of a person is a much more challenging, arguably intractable problem. Russell argues that context is better than basic emotion theory, and that we need a more advanced theory regarding facial expressions, not that basic emotion theory cannot be used for insights. It is therefore assumed that these signals can offer some degree of insight and improvement.

Example scenario two makes the same assumptions as example scenario one, and builds on it to assume that the displayed basic emotions correlate with a positive or negative sentiment. In addition to this, it assumes phrases have valence which can be derived from semantic analysis.

Example scenario three assumes key moments of speech can be marked by moments of strong emotion, unexpected changes in emotion strength and sentiment and unexpected emotional state transitions.

**2.5 Dataset Requirements**

To evaluate the results of the above systems it is crucial to have an appropriate dataset with which to test them. As outlined in the literature review, an important next step for these example applications would be to integrate further modalities. Therefore initial testing was carried out against datasets intended for multi-modal analysis such as the Chalearn Multimodal Gesture Recognition dataset [18]. This is a publicly available dataset intended for research into multimodal systems, providing audio data, an RGB video stream, depth masks, user masks and a skeletal model, with manual annotation for the ground truth.

Testing the systems against the Chalearn Multimodal Gesture Recognition dataset uncovered a number of issues and limitations to consider moving forward. The Emotion API struggled to identify faces, and when it succeeded the overwhelming assessment was that neutrality was displayed throughout which does not make for a useful proof of concept. There are a number of potential causes for these issues. The camera is recording relatively far from the volunteers’ faces so as to record entire bodily movements and therefore does not provide a high quality data stream for facial analysis. Additionally, the volunteers are performing short displays with focus on the gestures; they therefore are not feeling or expressing the associated emotions through their facial expressions that would be present in real life scenarios such as happiness when expressing the gesture for “perfetto”. This emphasised the need for a dataset with a high quality stream of video data with focus on the face and which is based on emotive scenarios.

**Chapter 3: Chat Bot Implementation**

The proof of concept systems have been implemented using Microsoft’s Cognitive Services, otherwise known as Project Oxford. This decision was made due to the range of artificial intelligence APIs available in this collection. This allowed a unimodal natural language processing system with a range of features to be created with relative ease. It also allowed for this system to be extended to incorporate facial expression of emotion data, within the same environment to aid with development.

**3.1 Unimodal Chat Bot**

The basic unimodal natural language understanding chat bot which analyses only the audio stream to determine a response, was implemented as follows:

Videos of type mp4 are read from a folder, and using ffmpeg, the audio stream is extracted. The resulting wav file is passed to the Bing Speech API which parses the audio into textual form. The Bing Speech API is integrated with a LUIS (Language Understanding Intelligent Service) subscription, which was trained to determine the intent of the user for specific scenarios, such as asking a variety of questions. The system then takes the user intent and selects an appropriate response from a list of strings. This string is then passed to the Bing Speech API once again, but this time to convert it into audio so the system can verbally respond to the user.

Example usage:

User says “What are you doing?”

System responds “I am following your commands to the best of my ability.”

Audio Input

Text Data

**Bing Speech**

User Intent

**LUIS**

Text Response

Audio Output

**Bing Speech**

**Fig. 1.** The unimodal chat bot implementation workflow

**3.2 Emotion Aware Chat Bot**

To implement the Emotion Aware Chat Bot, the unimodal chat bot system was taken, and updated so that the system asynchronously sends the video mp4 file to Microsoft’s Emotion API at the same time as sending the extracted audio to the Bing Speech API. The system then takes Emotion API response and determines the most prevalent or key emotion in the input. The user intent and the determined emotion are then used together to choose the most appropriate string as a response to output verbally.

Example one:

User says “What are you doing?” and user expression is neutral

System responds “I am following your commands to the best of my ability.”

Example two:

User says “What are you doing?” and user expression is sad

System responds “I’m sorry, I’ll stop it now.”

Audio Input

Text Data

**Bing Speech**

User Intent

**LUIS**

Text Response

Audio Output

**Bing Speech**

Video Input

Key Emotion

**Emotion API**

**Fig. 2.** Multimodal implementation of the emotion aware chat bot

**3.3 Sentiment & Emotion Aware Chat Bot**

To implement the Sentiment & Emotion Aware Chat Bot, the Emotion Aware Chat Bot implementation was taken and expanded upon. It was updated so that as well as sending the text data parsed from the audio to LUIS to determine intent, it is also sent to the Text Analytics API to analyse the sentiment of the chosen words. The intent, key emotion and sentiment of the chosen words is then all used together to choose the appropriate string with which to respond to the user. The sentiment is used to contrast against the sentiment of the expressed emotion when appropriate, with the facial expression of emotion assumed to be the true sentiment, but the conflict provides extra context and allows us to determine further information about the tone. For example if a negative comment is made about the system, while expressing a positive emotion, it could be determined that this statement was said in jest.

Example one:

User says “You’re awful.” and user expression is sad

System responds "I will do all I can to learn from this and not do it again."

Example two:

User says “You’re awful.” and user expression is happy

System responds "I know you love me really."

Audio Input

Text Data

**Bing Speech**

User

Intent

**LUIS**

Text Response

Audio Output

**Bing Speech**

Video

Input

Key Emotion

**Emotion API**

**Fig. 3.** Multimodal chat bot implementation of use case two

Word Sentiment

**Text Analytics**

**3.4 Creating a Dataset**

To create appropriate input videos, use cases were determined that demonstrate scenarios in which information derived from facial expression of emotion would enable a chat bot to provide a more appropriate response than would be determined with audio analysis alone. To support these use cases, an instance of Microsoft’s Language Understanding Intelligent Service was trained to recognise these scenarios and derive intent. The appropriate responses for each applicable intent, emotion and word sentiment combination, determined in the use cases, were inserted into the application implementation. The use cases were then performed and recorded, with a combination of different chosen phrases and expressed emotions.

The selected use cases display the impact of the additional modality on interaction with a chat bot via short to the point interactions, such as asking a question or making a statement. The response in this scenario is based on the most prevalent emotion during the phrase, which means there is room for error as the input is viewed in aggregate rather than frame by frame. This allowed for layman performances in front of a webcam, and is usable by anyone for demonstration purposes.

**3.5 Evaluation of the System**

In the dataset, only seven phrases are said, with one phrase per video. A unimodal system is clearly only able to provide a maximum of seven applicable responses for this dataset. When emotional awareness was introduced, the chat bot can then respond with a response not just for every intent, but for every intent-emotion pair. When text sentiment was introduced, the chat bot could respond with a response for every intent-emotion-sentiment triplet, allowing it to pick up on cases of heightened emotion, sarcasm and teasing. The created dataset demonstrates how the system fine-tunes its response in all these scenarios.

However, the videos in the dataset have exaggerated, acted expressions, and the emotion analysis API is far too slow for real time responses.

The implementation for this project available at [1] demonstrates that the required technology is readily available to create a crude, but possibly insightful, emotionally aware system.

**Chapter 4: Key Scene Selection**

Key Scene Selection is a much more complex application than the chat bots implemented above. Such a system must be able to process input with much longer speech, and extract interesting subsections, and therefore has a much smaller margin for error. The process of implementing this application is expected to identify challenges and opportunities for using facial expressions of emotion as a modality for many other nuanced forms of application.

**4.1 Models**

To select key scenes from a speech, there are a variety of models which can be explored. To utilise the facial expression of emotion modality, the video is split into multiple segments in which the emotional state is analysed, with probabilities given for the chance of each basic emotion occurring, including neutrality. The sequence of scenes containing facial expression of emotion data is obtained in the example applications via Microsoft’s Emotion API.

**4.1.1 Emotionally Charged Moments**

One of the simplest and most common sense models for choosing key moments from a speech based on facial expression of emotion, is to choose the scenes with the most intense displays of emotion.

To implement this model, the analysed video segments are ordered from the most neutral to the least neutral. The percentage of frames to select is provided as a threshold, and the segments the within this top range are selected as the key moments of the speech. This model assumes the most emotional moments are key to the message of the speech.

**4.1.2 Stochastic Processes with the Markov Property**

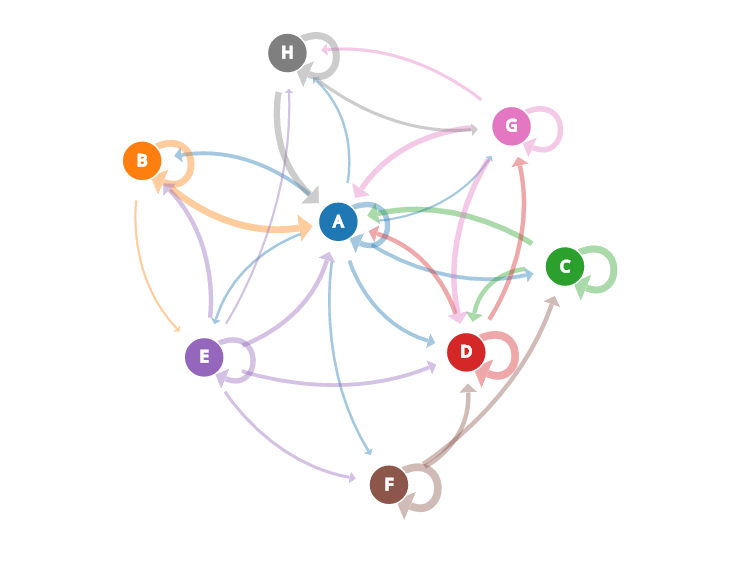
This paper proposes that the sequence of quantitative and qualitative results of emotional displays by humans over time, can be modelled as a stochastic process. The magnitude of change between states in the sequence is expected to fall within a certain range, which accounts for noise and natural transitions expected in communication. This paper also proposes that the expectation of the values of subsequent displays of emotion depends solely on the current state, and can therefore be said to have the Markov property. These assumptions are proposed due to common knowledge that observers find sudden, dramatic changes in expressed emotion shocking, while small shifts in emotion are expected. Key moments of a speech can be assumed to be those which appear unlikely to occur, as their appearance suggests a dramatic change in circumstance, thought or portrayal.

**4.1.3 Emotional State Transitions**

The sequence of a speaker’s facial expressions of emotion over time, can be viewed as transitions between discrete emotional states. Humans naturally categorise certain emotional states as being similar or dissimilar to others, such as via sentiment. It is clear that transitions to related emotional states are much more likely than to contrasting states, and a variety of contextual aspects such as style of communication determine the likelihood of these transitions. This can be modelled as a Markov Chain. There are a finite number of states, which are expressed in a sequence over time which is assumed to be a stochastic process for which the Markov property holds. A Markov Chain consists of discrete states and a transition matrix which holds the probability of transitions between states.

For implementation, the possible states of the Markov Chain consist of the basic emotions the Emotion API can detect; happy, sad, angry, surprised, fearful, disgusted, contemptuous or neutral. The transition matrix was created as an expert system. Based on common sense analysis of what emotional transitions are to be expected, probabilities were assigned to each possible transition. The chosen implementation is represented in figure four; ‘A’ represents the neutral state, ‘B’ the happy state, ‘C’ the sad state, ‘D’ the angry state, ‘E’ the surprised state, ‘F’ the fearful state, ‘G’ the contemptuous state and ‘H’ the disgusted state. Given a person is in one of the states, the width of each of the arrows from that state represents the likelihood of that transition being taken. For all states in the state space, the likelihood of staying in the same state is greater than or equal to the likelihood of any other transition. Details of the transition matrix values can be found in Appendix A.

This Markov Chain can then be used to determine key scenes from a video. The sequence of emotional states displayed in the video is derived from the analysed video segments, as the most likely emotion in each segment. The transition matrix is then used to analyse the probability of each state transition that occurred, providing a sequence of probabilities. The scenes involved in unlikely transitions are then selected using a threshold, and chosen as the key moments of the speech.



**Fig. 4.** Visualisation of the transition matrix. The circles represent the emotional states, and the width of the colour-coded arrows represent the weighted likelihoods of the transitions. Created with the aid of http://setosa.io/ev/markov-chains/

**4.1.4 Emotional Intensity Transitions**

Small changes in emotional intensity through time are to be expected, while dramatic changes of intensity and sentiment are shocking. This model assumes such scenes should be selected as key moments of a speech.

The sequence of a speaker’s facial expressions of emotion over time, can be viewed as a continuous value representing emotional intensity and sentiment. The magnitude of the emotional intensity can be modelled as the likelihood that that emotion is being expressed in a specific scene. The sentiment of the overall emotion can be modelled as the sign of the assigned emotional intensity value, with a positive sign for positive sentiment and a negative sign for negative sentiment. Each “scene” is therefore assigned an emotional intensity value between -1 and 1. A value of -1 is assigned if the main emotion being expressed is negative, and it is certain neutrality is not being expressed; a value of 0 is assigned if the main emotion being expressed is neutrality and a value of 1 is assigned if the main emotion being expressed is positive and it is certain neutrality is not being expressed.

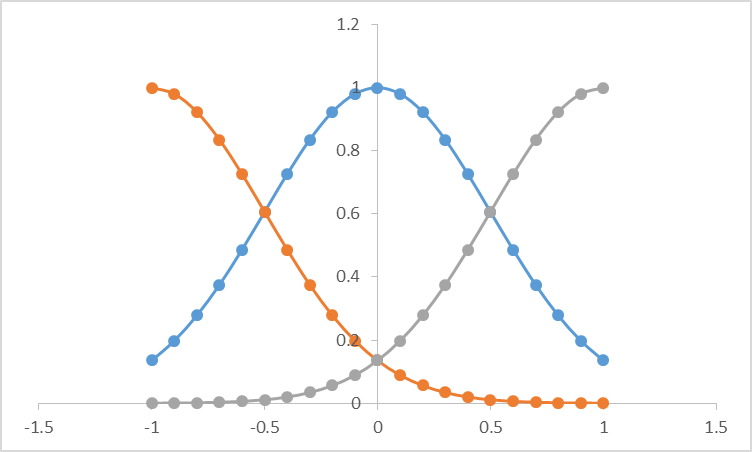
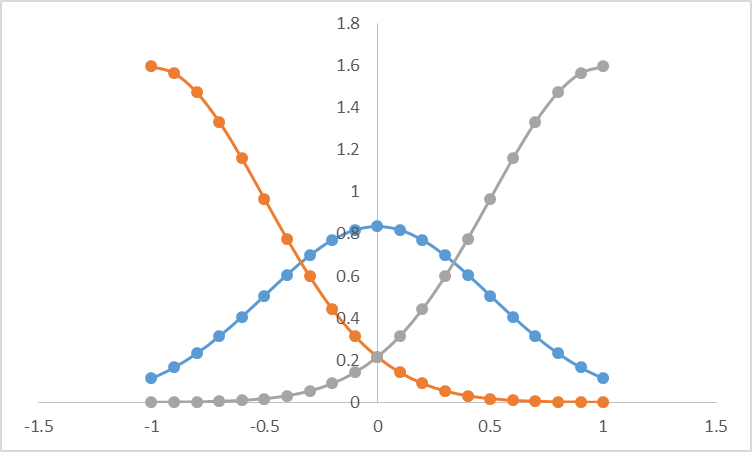
It is assumed that the intensity of the displayed facial expression of emotion over time is a stochastic process for which the Markov property holds. Given this assumption and the current value of emotional intensity, the value of emotional intensity in the following scene can be assumed to be close to the current value, with greater deviations in value being less likely. The deviation is likely to fall within a certain range to either side of the current value. The intensity from one scene to the next could increase or decrease, or in cases where there is low intensity, it is reasonably likely that the sentiment expressed could change. Large changes in emotional intensity, and changes between positive and negative sentiment are unexpected and would startle an observer.

This paper suggests that transitions in emotional intensity can be modelled as a normal distribution. The value of the standard deviation of this distribution is set to encompass the extent of change which could be reasonably expected to be observed in a person communicating. Since the possible values of emotional intensity are constrained to be between -1 and 1, the normal distribution must be truncated.

Given the emotional intensity in the current scene, a truncated normal distribution can be created about the current value of emotional intensity, by setting the current value as the mean or expectation of the distribution. The likelihood of a transition to some value in the following scene can be calculated by taking the probability density function of the truncated normal distribution at the point equal to the value of the next scene’s emotional intensity.

For implementation, the video segments are taken, and a sequence of emotional intensities calculated. The likelihood of transition from one state of emotional intensity to the next is calculated, giving a sequence of probabilities. Scenes involved in a transition with a probability below a threshold are selected as the moments of most shocking transitions. These scenes are selected as the key scenes of the speech.

**Fig. 5.** The graph to the left displays the probability distribution function for three truncated normal distributions. They are truncated at the values -1 and 1and they have a standard deviation of 0.5. The mean is set to -1, 0 and 1 to display the extremes. The graph to the right is the result when the pdfs are weighted by their maximum value.



This has the potential to be a powerful model due to it encompassing emotional strength in addition to sentiment over time.

To calculate the normal distribution in the example application, an open source implementation of the formula implemented in C++ was taken from [19] and translated into C# to integrate into the application. The probability distribution function of the value of emotional intensity in the following scene is assumed to represent the likelihood of the transition to this value as outlined above. However, the probability distribution function at a specific point can be greater than one, with the integral of the normal distribution function equalling one. A simple solution to scaling the pdf to a value which correlates with the predicted probability, is to take the pdf at that point and scale it to be a percentage of the maximum possible pdf for that normal distribution.

**4.2 Establishing the Dataset Domain**

The requirements for the dataset to be used for key scene selection are much stricter than those for the chat bot implementations. Accuracy of emotion display and recognition in each frame is very important, and an understanding of the “ground truth” of which scenes in a speech truly are key is essential.

**4.2.1 Genuine vs Acted Displays**

There are two approaches to ensuring accuracy of emotional display. The first approach is to find a dataset in which real life events have been recorded in an emotional context, and therefore the speaker is portraying genuine emotion. The second approach is to find or create a dataset in which actors are performing emotive scenes.

As outlined earlier, one of the limitations of the system is that it is necessary to have the recording camera focused on the face of the communicator. It is difficult to find videos of real life emotive events for which this is the case. One potential source is videos of public figures making speeches, but these tend to have cinematic features which would have to be taken into account when evaluating state transitions.

It is much easier to source videos of actors performing emotive scenes with focus on their face, as there are numerous examples of emotional monologues or soliloquys in plays and other sources. It is also feasible to gather volunteers with acting abilities to create a dataset. Creating a dataset provides flexibility to demonstrate interesting use cases which play to the strengths of such a system to demonstrate a strong proof of concept of the potential of such a system. Many of the issues with this approach are outlined in [5]. For example, it is suggested acted performances do not accurately replicate natural, spontaneous expressions, the data can suffer from inaccurate actions by subjects, and they can suffer from biased labelling. The latter two issues can corrupt the training data, and the former can prevent a model trained on the data from having the ability to apply to generic input.

For this project it was decided that acted data was preferable due to the ability to source it, the ease with which it can be processed, the ability to demonstrate interesting use cases and due to the belief that the shortcomings can be overcome to produce a useful system.

**4.2.2 Overcoming Acted Data Issues**

By using trained actors the potential for inaccurate actions by subjects should be reduced. This also has the potential to overcome the issue of performances not replicating natural expressions.

The issue of performances not replicating natural expressions can also be overcome by demonstrating the utility of such a system on a more constrained domain. If it is found that the key scene selector cannot be applied to general speeches containing natural displays of emotion due to it being trained on acted material, it will still be applicable to acted material. Example applications of such a model could be selecting dramatic scenes from a movie, or comparing and contrasting performances of speeches by different actors from different demographics.

**4.2.3 Use Case: Shakespeare**

Good use cases for such a system are scenes in which there are strong displays of emotion and scenes full of sarcasm, irony or general conflicting meaning between the sentiment of the chosen words and the expressed emotion.

Shakespeare speeches are renowned for their dramatic emotive scenes, and provide a number of examples of spiteful, sarcastic disdain. Shakespeare speeches are also well known which make them useful for proof of concept displays, and there is the potential for multiple interpretations when it comes to performing plays.

Videos made by the Guardian newspaper of Shakespeare speeches by professional actors were considered good sources of input. These videos are of high quality, with the camera focused close to the face, of renowned actors with emotive speeches.

Damian Lewis’ portrayal of Antony in Julius Caesar is of particular interest due to the conflicting meaning displayed between modalities [Available: [www.youtube.com/watch?v=q89MLuLSJgk](http://www.youtube.com/watch?v=q89MLuLSJgk)].

The context of the speech is that Julius Caesar has been murdered by Brutus, and Antony is making a vehement, spiteful, contemptuous speech to the crowds gathered around Caesar’s body, calling Brutus an honourable man with obvious sarcasm and distaste.

For a system to determine this context from the chosen words alone would require a very sophisticated natural language understanding model. This NLU model would need to be capable of interpreting the relationships between the characters and dismiss the misleading statements which appear positive such as “Brutus is an honourable man”.

The facial expression of emotion modality and its contrast versus the sentiment of the chosen words in this scenario can be used to provide a degree of context. The system can utilise this data to identify disingenuous statements, and in recognising the portrayal of disdainful sarcasm, the system can identify these moments as key scenes.

**4.2.4 Difficulties with Shakespeare**

Initial smoke testing of the videos sourced from the Guardian was conducted by carrying out feature extraction. These tests indicated potential for the displayed facial expressions of emotion to be useful, but found that the Bing Speech API was providing very inaccurate results for its speech recognition.

It is assumed this occurred because the speech recognition models have been trained to recognise modern English words and phrases, while the example videos are performed in old Shakespearian English.

**4.3 Creating a Dataset**

It was decided that a dataset should be created specifically to overcome the issues and limitations of the system as outlined above, and to demonstrate the strengths of the concept.

The goals of the dataset are as follows:

* To emphasise the strengths of a system integrating facial expression of emotion as a modality by consisting of performances:
  + That can have multiple interpretations
  + That contain emotive scenes
  + That contain scenes of contrasting messages between the chosen words and the displayed emotion
* To enable accurate analysis of the performer’s facial expressions by providing high quality recordings focused on the performer’s face
* To consist of high quality performances which can be said to represent natural spontaneous expressions, or at least be generalizable to generic performances
* To have accurate unbiased metadata associated with the data, representing the ground truth of the performances

The steps taken to create such a dataset are as follows:

1. Source appropriate segments of speech which:
   * Can have multiple interpretations, contain emotive scenes and contain scenes of contrasting messages between modalities
   * Are written in language understandable by the speech recognition system
2. Organise a session which attracts professional volunteers to offer their help to create videos of high quality performances, followed by high quality analysis
3. Have volunteers perform the segments of speech:
   * With a camera focused on their face
   * With different interpretations requiring different portrayals of emotion
4. Have volunteers analyse the performances of others and provide feedback as to their interpretation of the portrayal

**4.3.1 Sourcing Speeches**

As outlined above, Shakespeare speeches were identified as interesting use cases but caused issues due to the style of language. Modern translations of Shakespeare’s plays were sourced from nosweatshakespeare.com to ease the process of speech recognition. Short extracts were taken from the plays to make recordings easy and to provide impactful, to the point examples to prove the concept of the system.

Five extracts were chosen and adapted to be interesting use cases. Four of these come from Act 3 Scene 2 of Julius Caesar, as this is an extremely emotive scene with very pronounced contrasts between the messages portrayed in different modalities. The fifth is an adaptation of a speech in Act 2 Scene 7 in As You Like It, otherwise known as The Seven Ages of Man. These extracts with details of possible portrayals can be found in Appendix B.

**4.3.2 Volunteer Session**

Collaboration was established with Professor Mark Burnett of the School of Arts, English and Languages at Queen’s University Belfast who has published numerous papers on the analysis of Shakespearian plays on film. With the aid of Professor Neil Robertson and Professor Burnett, a session was organised to create a dataset with the help of further volunteers from the School of Arts, English and Languages. A high quality camera was positioned to focus on the volunteer’s faces as they performed the extracts with a variety of portrayals.

**4.4 Labelling Data**

Having gathered the data for training and testing the key scene selector, it is essential to label the data with the “ground truth” as to which scenes are the most important.

For this scenario, important scenes are defined as those which are key to understanding the message being depicted by the communicator. This is a difficult task, as determining what is important or key to understanding is very subjective. Additionally, the intention of the communicator may not always match the interpretations of an observer.

It is unrealistic to attempt to have a computer determine what scenes an actor intends to portray as most important, but it should be possible to mimic an observer’s behaviour to some degree. The data should therefore be labelled with the assessments of observers watching the video, not intention of the actor themselves. The actors also cannot be allowed to watch and assess themselves, as they will have inherent biases towards labelling their intentions.

Ideally each video should have a large set of observational analysis feedback from a range of volunteers, as the larger the set of labels the more reliable they should be. However given the limited time and resources of this project, this was not feasible.

To gather “labels” for the data, a form was created to gather the opinions of the observers. The volunteers watched each video through its length once to clearly observe the full message, then during a second screening filled in the forms with their opinions. The forms were developed to be easy to use and concise in an attempt to limit ambiguity, and best make use of the volunteer’s time. The forms ask the volunteers to identify the prevalent emotions displayed, and ask them to label the video as “genuine” or “sarcastic”. If there is much deviation in the responses to these questions for a video, it is a strong indication that there is extreme ambiguity in the fundamentals of the portrayal which could make the analysis invalid. The forms then provide a copy of the speech extract under consideration. The volunteers are asked to underline the words which correspond to the moments in the speeches that are of key importance. The generic template for the feedback forms is available to view in Appendix C.

**4.4.1 Translating Feedback to Labels**

A limitation to the form of metadata gathering carried out for this project, is that volunteers underlined sections of text from a speech, which does not necessarily translate directly to sections of a video portraying that speech. For this project, a scene is defined as a time period with start time and end-time/interval that corresponds to a time period in a video sequence.

In numerous videos, the volunteers had periods of speaking rapidly, which made it difficult to segment words. This is particularly problematic when underlined words coincide in time with words that were not underlined. Another issue is when there are large spaces in time between spoken words that were underlined, as it is not possible to know if the key scene should include these silent periods or not.

Underlined words to scene translation was carried out manually which further introduced inaccuracies. An attempt was made to find an API which could provide timestamps for words in an audio stream, but one was not discovered. This could be due to the predictive nature of speech recognition systems, as they do not detect words as audio segments. For this project, it was estimated that manual labelling is accurate to the nearest half second.

A systematic approach was taken for translation, based on the above limitations. The start time from manual labelling was rounded down to the previous half second, to take into account the moments leading up to the underlined words, and the end time rounded up to the next half second to take into account the following moments, while remaining in the margin of error for manual labelling.

The resulting metadata can be viewed in the dataset project located at [1].

**4.4.2 Producing Comparable Results**

A similar translation must also be carried out for the key moments identified by the statistical models. These models select sections of approximately between 0.3 and 0.5 seconds long from the video as key moments. This does not take into account that some of these frames might be very close in time to the others, and that sections between these frames should be taken into account as part of the scene.

To translate these frames into key scenes, those which are in immediate sequence with each other are merged together. The start time of these merged frames are then rounded down to the previous half second, and the end time rounded up to the next half second to accommodate for the lead up and finishing of these key moments, and to make the scenes comparable with the dataset labels. Scenes which overlap each other are then merged together.

**4.5 Evaluating Dataset Label Reliability**

From initial analysis of the dataset labels, it was clear that a number of videos were lacking consensus in their analysis. This paper argues that if there is not consensus in the determination of the intent and meaning of a speech, then there is not a strong basis for identification of key scenes.

It was therefore determined that a measurement of consensus should be carried out.

The two areas of the feedback which are quantifiable, are the weightings given for the display of each basic emotion, and the sections of the speech underlined as important scenes. Metrics analysing each of these should have a strong correlation with the reliability of the feedback. If there is not consensus in what emotions are being portrayed, or in which moments are key then it is very unlikely that there is consensus in the intent and meaning of the speech. In addition to this, consensus of identified scenes is essential as the success of a system’s predictions will be based on these results.

To measure the consensus between identified scenes, a metric named Key Scene Parity was defined as the percentage length of consensus scenes over suggested scenes. A consensus scene is defined as a scene identified as a key scene by all observers, and a suggested scene is defined as a scene identified as a key scene by any observer. If the observers selected identical scenes the Key Scene Parity would equal 100% and if the observers did not agree on any scenes then it would equal 0%.

To evaluate the consensus between emotion weightings, each basic emotion is treated as a dimension and each weighting as the magnitude of the vector in that direction. The feedback from each observer can therefore be viewed as a vector on an eight dimensional plane. A covariance matrix is then calculated as a measurement of the variance between the multi-dimensional vectors. This variance is then quantified as a single value using the Frobernius norm, also known as the Euclidean norm, which is the square root of the absolute sum of the squares of the elements of a matrix. For the purposes of this project, this value is defined as a metric named Emotion Variance.

These two metrics were calculated for every video in the dataset and videos with a low Key Scene Parity or high Emotion Variance were disregarded when it came to evaluation of the models. From the eighteen videos in the dataset, only seven were deemed reasonably reliable. The results of this analysis can be viewed in Appendix D.

**4.6 Evaluation Model Results**

Having determined a strong ground truth for a subset of the dataset, and implemented the ability to extract comparable key scenes using the statistical models as defined above, it is important to be able to quantify the success of each model.

Based on the gathered feedback from volunteers, it is assumed that an ideal model will identify all consensus scenes and will not identify any unmarked scenes, whereby an unmarked scene is a scene which is not a suggested scene, and consensus scenes and suggested scenes are defined as above.

**4.6.1 Metrics**

Two metrics are defined for the evaluation of the success of a model. The percentage of consensus scenes identified and the percentage of unmarked scenes identified. These metrics will be named Consensus Match and False Positive Match respectively.

The Consensus Match is calculated by determining the overlap between the scenes identified by the model and the consensus scenes. The duration of these overlap scenes is then calculated and divided by the duration of the consensus scenes. Therefore, if all consensus scenes are contained in the scenes identified by the model, the result will be 100%, or if none of the consensus scenes are identified, the result will be 0%.

The False Positive Match is calculated by determining the overlap between the scenes identified by the model and the unmarked scenes. The duration of these overlap scenes is then calculated and divided by the duration of the scenes identified by the model. Therefore, if all identified scenes are contained in the unmarked scenes, the result will be 100%. If, as desired, the model has only selected suggested scenes, then the result will be 0%.

A successful model will perform better at key scene selection than random selection of scenes. These metrics are not enough on their own to evaluate the success of a model, as by random chance, increasing the proportion of the video chosen will increase both the Consensus Match and the False Positive Match. To truly evaluate the success of a model, the values for these metrics must be compared against the values which would be achieved via random chance. The metrics for success are therefore the percentage increase in Consensus Match versus random chance, which will be referred to as Consensus Match Increase and the percentage reduction in False Positive Match versus random chance, which will be referred to as False Positive Reduction.

It is assumed that by randomly selecting x% of the duration of a video, x% of the consensus scenes will be chosen and x% of unmarked scenes. To calculate Consensus Match Increase and False Positive Reduction, the scenes selected by a model are taken, and the percentage duration of the video for which they comprise is calculated. The Consensus Match and False Positive Match are calculated and compared against the proportion of the video selected. The percentage change for each metric between random chance and the model is calculated as each metric divided by the proportion of the video chosen, minus one, times one hundred.

Consensus Match Increase equals this percentage change in Consensus Match.

False Positive Reduction equals zero minus this percentage change in False Positive Match.

**4.6.2 Issues/Considerations**

There are a number of issues and considerations regarding the models, the metrics and the API which need to be taken into account when evaluating the results.

1. The Emotion API does not analyse the end of the video. Based on the timestamps provided by the API, it was found that between 0.85 and 4.6 seconds of content at the end of each video was not evaluated. On average, approximately 29% of a video’s consensus scenes are contained in these unevaluated scenes, which the models cannot possibly select since no data is provided for them. To resolve this the segments of the dataset labels which correspond to times past the last timestamp must be discarded as they cannot be evaluated. This is unfortunate, as it greatly reduces the dataset content, and opportunities to identify key moments.
2. The Emotion API will not always evaluate the entirety of the middle of the video. For example, for one video in the dataset the period between 29 seconds into the video and 31 seconds into the video is not analysed. No immediate solution was found for this issue.
3. Due to the timescale, available resources, and evaluation of reliability, the resulting dataset is small. The resulting dataset also only contains performances from two actors, which clearly biases the data towards their style of performance. Results for this dataset therefore cannot necessarily be extrapolated to represent a larger range of scenarios.
4. The fact that the threshold chosen for a model may be too low to select any scenes from some videos should be taken into consideration.
5. The metrics outlined above only function as valid measures when the proportion of the video chosen is less than or equal to 50%. Above this value, the average results of the models becomes biased, as individual videos could produce a Consensus Match Increase of -100% but not +100% and a False Positive Reduction of +100% but not -100%. Models which select proportions greater than this therefore cannot be reliably evaluated. This is reasonable for this application as it can be argued that selecting more than half of a video’s contents does not qualify as selecting the key scenes.

The last two considerations both define times when a subsection of results for certain model-threshold pairs will not be valid. It is useful in these cases to know the success of a model-threshold pair for the videos to which it applies, but the fact that it yielded some invalid results cannot be disregarded. When averaging the results for a model-threshold pair, it is important to keep track of how many valid video results are being evaluated, as simply discarding those with no results would bias the average metrics.

**4.6.3 Setting Model Parameters**

Each of the models outlined in this paper have values which can be tuned to adjust which scenes they select. For the purposes of this project some will be varied to seek an optimal result, and others shall be kept constant to simplify analysis.

All models have a threshold which will be varied. For the Emotionally Charged Moments model, the percent of scenes it chooses is varied. For the Emotional State Transition model and the Emotional Intensity Transition model, the threshold under which the likelihood of a transition must fall to be considered key is varied.

The thresholds used to evaluate the Emotionally Charged Moments model and the Emotional Intensity Transition model will be between 0% to a value at which the proportion of the video chosen is greater than 50%. The thresholds used to evaluate the Emotional State Transition model will be the likelihoods in its transition matrix, as only at these thresholds will there be any change in the scenes chosen.

The transition matrix likelihoods of the Emotional State Transition model will equal the values demonstrated in Fig. 4 and be constant. Adjusting these likelihoods to be more accurate is beyond the remit of this project; it would require an analysis of a much larger and extensive dataset to have any justification for changes. The likelihoods chosen for this project stand as a proof of concept implementation.

For the Emotional Intensity Transition model, the lower and upper bounds of the truncated normal distribution is set to -1 and 1 respectively as outlined above, and the mean is set to the value of the emotional intensity at each scene to calculate each likelihood. The value of the standard deviation is left constant at a value of 0.5. Adjusting either the threshold of likelihood or the standard deviation would vary the proportion of the video chosen as key scenes. It was decided that for the purposes of this project the standard deviation would be left at a value which produces a graph that supports common sense reasoning, (that large jumps between emotion strengths and sentiments is very unlikely, and small changes are likely) and that the threshold was the most reasonable value to vary.

**4.6.4 Results**

The average results of each model type, with a range of example thresholds can be found in Appendix E, alongside the best result for each individual video in the dataset.

One of the videos in the dataset was labelled overwhelmingly neutral, and this is reaffirmed by the fact that neither the Emotional State Transition model nor the Emotional Intensity Transition model could yield valid results for this video. There is a very sharp transition between a threshold that is not large enough to choose any scenes, and a threshold which selects every single frame, due to all transitions being very similar.

The implementation of the Emotional State Transition model did not fare well for this dataset. It has been found that there was little commonality in the kind of emotion state transitions that occurred in the videos. Valid results are obtained for at most half of the dataset for ten of the twelve thresholds, with the remaining two obtaining results for all but the overwhelmingly neutral video mentioned above. The ‘best’ result of these two was with the threshold set to 0.1, with an average of 16.97% of the video selected, a False Positive Reduction of 37.98% and a Consensus Match Increase of -26.8%, which is far from ideal, being worse than random for the identification of consensus scenes.

The Emotional Intensity Transition model has shown some degree of success for videos which contain unlikely transitions, but the majority of videos in this dataset did not contain such changes thus rendering the model generally ineffective. When the threshold was set to 10% on average the model produced a Consensus Match Increase of 66.19% and a False Positive Reduction of -4.55%, selecting 5.55% of the video’s duration. The slight underperformance in False Positive Reduction is clearly more than made up for by the huge gains in the Consensus Match Increase, but at this threshold only 29% of the videos in the dataset yielded valid results. This therefore shows an indication that this model has potential in a subset of videos but not for generic application. The fact that small, “likely” transitions mainly occur in this dataset is demonstrated by how a small change in the threshold results in huge changes in the proportion of the video selected. When the threshold is set to 52%, 28% of the video is selected, and an increase to 53% on the threshold results in an average of 84% of the video being selected. This shows how the vast majority of transitions are in a small unremarkable range.

The Emotionally Charged Moments model fared much better, mostly because rather than trying to extract unlikely events, it will simply select a proportion of each video. This makes it applicable to every video, yielding valid results for a much larger range of thresholds. This model produced very successful results for every single video bar one, with up to a 366% Consensus Match Increase on individual videos. On average, the most successful of these results was when the threshold was set to select 11% of the most emotional frames. As the evaluated frames come in varying sizes, this did not correspond to selecting 11% of the video’s duration, but rather an average of 18%. This produced an average Consensus Match Increase of 50% and a False Positive Reduction of 19%. This huge gain indicates a strong correlation between emotionally charged moments and key scenes.

**4.6.5 Explanations for Results**

The failure of the Emotional State Transition and Emotional Intensity Transition models can be explained by their intent and the nature of the dataset. The basis for these models is the assumption that an unlikely change in expressed emotion correlates with a key scene. The dataset used for analysis is small, with very few such occurrences. In the cases when the Emotional Intensity Transition did apply, there are indications that this assumption is correct, but the dataset is far too small to say so emphatically. To test the true success of these models, a much larger dataset would need to be generated, focused on a domain in which such transitions would be more common.

An additional explanation for the failure of these models is due to their complexity, which is especially frank in the Emotional State Transition model. This model has 64 transition likelihoods set in its transition matrix, all based on naïve suggestion extrapolated from common sense reasoning. To generate successful models of this nature a much stronger basis must be determined for the underlying values of the constants on which the models depend.

The success of the Emotionally Charged Moments model can therefore be attributed to its simplicity, and the nature of its intent. This model varies simply one threshold, and rather than select scenes based on their determined likelihood, derived from emotional analysis, this model is simple in selecting scenes based on their emotional strength, reducing its margin of error.

**Chapter 5: Conclusion**

Through the course of this project, positive results were generated for using facial expressions of emotion as a modality. However, these results were obtained with a number of crucial limitations applied to the system and to the data which was processed.

The concept of how insight into the emotional state of a person and analysis of the sentiment of their words could provide much improved chat bots was demonstrated. However, it cannot be reliably said that the level of insight into the emotional state of a person, required for such a system can necessarily be determined from their facial expressions. The implemented system does not prove this due to the nature of the data on which it is used, which consists of exaggerated, acted expressions.

The Key Scene Selector application has shown that a useful, much more complex system can be created which relies on this modality. But, once again, this system was constrained to acted performances. For acted performances, this application proved that useful information can be extracted from facial expressions.

Key Scenes were selected with impressive accuracy when the state of the face was viewed as being quantifiably in two inverse states; neutrality and emotionality. Success was not found for models which introduced additional dimensions such as valence, or discrete emotion states. The result is that it has been proved that some important information is portrayed through facial expressions, but this study has not shown any indication that additional assumptions can be made. The failure of the multi-dimensional models can be explained by a variety of reasons, such as the nature of the scenes they try to extract, and the nature of the dataset on which they were tested, leaving the nature of the information portrayed inconclusive.

Were a more extensive, reliable and applicable dataset created, the multi-dimensional models could be revisited. Using this data to train their parameters, the feasibility of relying on higher level constructs borne from the low-level emotion analysis could be further explored.

This project has identified a number of issues, limitations and also opportunities with regards to employing facial expressions of emotion as a modality. The nature of its underlying scientific basis needs further exploration, but it is clear that useful information is obtainable. The utility of such information has also been displayed, with a demonstration of applications which could greatly benefit from such data streams. The models explored in the Key Scene Selector could additionally aid in the analysis of the nature of the information portrayed through facial expressions.

In conclusion, were we able to decode the emotional state of a person, or the information expressed through their facial expressions, we could enrich many systems. We could be on the verge of gaining such insights through the application of psychology and computer science.

# **Bibliography**

|  |  |
| --- | --- |
| [1] | R. Palmer, "CSC3002 40106318 - Computational Emotional Intelligence," 2017. [Online]. Available: https://gitlab.eeecs.qub.ac.uk/40106318/nlu-using-emotions. [Accessed 2017]. |
| [2] | E. Cambria and B. White, "Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]," *IEEE Computational Intelligence Magazine,* vol. 9, no. 2, pp. 48-57, April 2014. |
| [3] | R. Steinbrook, "Smarter Smartphones," *JAMA Internal Medicine,* vol. 176, no. 5, p. 625, 2016. |
| [4] | B. W. Schuller and M. F. McTear, "Sociocognitive Language Processing – Emphasising the Soft Factors," in *Proceedings of the Seventh International Workshop on Spoken Dialogue Systems*, Saariselkä, Finland, 2016. |
| [5] | S. Poria, E. Cambria, R. Bajpai and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion,* vol. 37, pp. 98-125, September 2017. |
| [6] | M. Anderson, "Nonverbal Communication," *Encyclopedia of Language & Linguistics,* no. 2, pp. 690-692, 2006. |
| [7] | S.-P. e. a. Wang, "KinChat: veiling your face without suppressing facial expression in text communication," in *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, Toronto, 2014. |
| [8] | D. G. Leathers and M. Eaves, Successful Nonverbal Communication: Principles and Applications, Routledge, 2015. |
| [9] | J. Panksepp, Affective Neuroscience: The Foundations of Human and Animal Emotions, Oxford University Press, 2004. |
| [10] | J. E. Rachael, O. G. Garrod and P. G. Schyns, "Dynamic Facial Expressions of Emotion Transmit an Evolving Hierarchy of Signals over Time," *Current Biology,* vol. 24, no. 2, pp. 187-192, January 2014. |
| [11] | A. J. Fridlund, Human Facial Expression: An Evolutionary View, Academic Press, 2014. |
| [12] | P. e. a. Ekman, "Universal Facial Expressions of Emotion," *California Mental Health Research Digest,* vol. 8, no. 4, pp. 151-158, 1970. |
| [13] | R. e. a. Reisenzein, "Coherence between Emotion and Facial Expression: Evidence from Laboratory Experiments," *Emotion Review,* vol. 5, no. 1, pp. 16-23, 2013. |
| [14] | iMotions, "Emotient," iMotions, 2017. [Online]. Available: https://imotions.com/emotient/. [Accessed April 2017]. |
| [15] | J. A. Russell, "Moving on from the basic emotion theory of facial expressions," *Emotion Researcher,* August 2015. |
| [16] | E. Cambria, "Affective Computing and Sentiment Analysis," *IEEE Intelligent Systems,* vol. 31, no. 2, pp. 102-107, 2016. |
| [17] | C. e. a. Perera, "A Survey on Internet of Things From Industrial Market Perspective," *IEEE Access,* vol. 2, pp. 1660 - 1679, January 2015. |
| [18] | S. e. a. Escalera, "Multi-modal Gesture Recognition Challenge 2013: Dataset and Results," 2013. |
| [19] | J. Burkardt, "Florida State University - Dept. of Scientific Computing - John Burkardt - The Truncated Normal Distribution," Florida State University, 11 September 2013. [Online]. Available: https://people.sc.fsu.edu/~jburkardt/cpp\_src/truncated\_normal/truncated\_normal.html. [Accessed April 2017]. |

**Appendix A: Emotional State Transition Matrix**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Neutral | Happy | Sad | Angry | Surprised | Fearful | Contemptuous | Disgusted |
| Neutral | 0.3 | 0.15 | 0.15 | 0.15 | 0.07 | 0.07 | 0.05 | 0.06 |
| Happy | 0.475 | 0.475 | 0 | 0 | 0.05 | 0 | 0 | 0 |
| Sad | 0.3 | 0 | 0.5 | 0.2 | 0 | 0 | 0 | 0 |
| Angry | 0.2 | 0 | 0 | 0.6 | 0 | 0 | 0.2 | 0 |
| Surprised | 0.2 | 0.2 | 0 | 0.15 | 0.3 | 0.1 | 0 | 0.05 |
| Fearful | 0 | 0 | 0.2 | 0.2 | 0 | 0.6 | 0 | 0 |
| Contemptuous | 0.3 | 0 | 0 | 0.3 | 0 | 0 | 0.3 | 0.1 |
| Disgusted | 0.4 | 0 | 0 | 0 | 0 | 0 | 0 | 0.5 |

Below is a table detailing the ascribed likelihoods of transitions between states for the Emotional State Transition model.

Each row corresponds to the current state in a sequence of emotion states, with the column corresponding to the next state. The values represent the percentage chance of transition to this emotion state. Each row therefore sums to a value of 1.

**Appendix B: Speech Extracts**

Speech Extract One

“Good friends, sweet friends, don’t let me stir you up to such a sudden flood of civil disorder. Those who have done this deed are honourable. What personal grievances they may have had that made them do it I don’t know. They are wise and honourable and will no doubt answer you with their reasons.”

This sentiment of these words are very positive; “Good friends, sweet friends”, “They are wise and honourable”.

The emotional portrayal can take on multiple interpretations. This extract can be portrayed as:

* Sarcastic, snide and venomous
* Contemptuous and angry
* Genuine concern for calming the crowd

Speech Extract Two

“The bad things that men do are remembered after their deaths: the good are often buried with their bones. Let it be so with Caesar. Here, with permission of Brutus and the rest – for Brutus is an honourable man, so are they all, all honourable men – I have come to speak in Caesar’s funeral order. He was my friend. Faithful and true to me. But Brutus says he was ambitious. And Brutus is an honourable man.”

This passage is more morbid than the first extract, but continues to praise Brutus and the other conspirators.

This emotional portrayal for this extract can be interpreted as:

* Sarcastic, snide and venomous
* Contemptuous and angry
* Sincere, accepting, sombre and morbid
* Distressed and broken

Speech Extract Three

“Only yesterday, Caesar’s word was the most powerful in the world. Now he’s lying there. And now we’re all superior to him. Oh, people of Rome, if I wanted to stir your hearts to rioting and rage I would be doing Brutus wrong, and Cassius wrong, who you all know are honourable men. I will not wrong them. I choose rather to wrong the dead, to wrong myself and you, than to wrong such honourable men.”

This passage is very similar to the first, but invites emphasis on different segments of the speech.

This emotional portrayal for this extract can be interpreted as:

* Sarcastic, snide and venomous
* Contemptuous and angry
* Genuine concern for calming the crowd

Speech Extract Four

“You all loved him once, not without cause. What cause do you now have to refrain from mourning for him? Oh reason, you have entered the bodies of animals and men have lost you. Bear with me. My heart is there with Caesar’s body and I must pause till it comes back to me.”

The text in this extract does not use sarcasm or any other technique that might mislead a unimodal system; the speaker could be seen to be anguished from semantic analysis of the words alone. The reason this passage is interesting as a use case is instead because textual analysis cannot detect moments of high emotion that are likely to be expressed at this time. It also would not detect dramatic changes of emotion that are likely in such a portrayal, such as a shift from intense rage to meek sorrow.

This emotional portrayal for this extract can be interpreted as:

* Rage and distress followed by sorrow and despair

Speech Extract Five

“One man plays many parts. At first a baby, crying and puking in the nurse’s arms. Then the whining school-boy, with his satchel and shining morning face, creeping like a snail, unwillingly to school. Then the lover, sighing like a furnace, with a tearful song. In the last scene, that ends his strange, eventful history; is childishness and simply oblivion. Without teeth, nor eyes, nor taste, nor anything.”

This extract is an interesting use case, as semantic analysis of the words could easily lead to a system interpreting this as simple a series of neutral statements, rather than a pointed story. Jacques, the character who gives this speech in the play, is often interpreted as sarcastic, snarky and spiteful; many would therefore interpret this speech as having a more subtle meaning.

This emotional portrayal for this extract can be interpreted as:

* Sarcastic, snarky and spiteful e.g. mocking others or trying to fill them with dismay
* Bitter and full of disgust at the negative aspects of life
* Amused, nostalgic and fond of the different aspects of life

**Appendix C: Feedback Form Template**

**Performance Analysis Form**

Use this form to provide your analysis of one of the performances you will view.

After your first viewing, choose the overall emotion and intention the actor portrays.

After/during your second viewing, determine which scenes you believe are the most important and underline the words said during these scenes.

**Overall Emotion: Tick the box that matches the overall emotion**

(If you feel the need to tick multiple boxes, please put weightings next to them e.g. 75% Happy, 25% Sad)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Neutral** | **Happy** | **Sad** | **Angry** | **Surprised** | **Fearful** | **Contemptuous** | **Disgusted** |
|  |  |  |  |  |  |  |  |

**Overall Intention: Select if you interpreted this portrayal as genuine or sarcastic**

Genuine / Sarcastic

**Important Scenes: Select if you interpreted this portrayal as genuine or sarcastic**

**Insert Text Here**

**Additional Comments:**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Appendix D: Dataset Label Reliability Results**

|  |  |  |
| --- | --- | --- |
| **Video ID** | **Key Scene Parity** | **Emotion Variance** |
| Patrick1 | 0.528094634558513 | 878.969172130374 |
| Mark1 | 0.305810397553517 | 467.874181805323 |
| Mark4 | 0.279255319148936 | 2728.95461432341 |
| Patrick6 | 0.228571428571429 | 2469.43116481869 |
| Mark2 | 0.185414091470952 | 869.077176601071 |
| Mark3 | 0.181598062953995 | 800 |
| Patrick4 | 0.163256955810147 | 3172.90230353501 |
| Chris1 | 0.0740740740740741 | 2641.71924237901 |
| Patrick2 | 0.0598802395209581 | 2450 |
| Mark6 | 0.0588235294117647 | 1110.65702526828 |
| Mark5 | 0.0267165375367352 | 1197.98007078582 |
| Patrick3 | 0 | 4710.6944669667 |
| Patrick5 | 0 | 6078.1713258731 |
| Chris2 | 0 | 2090.7576045804 |
| Lauren1 | 0 | 3195.7029525772 |
| Lauren2 | 0 | 3725.065249242 |
| Rachel1 | 0 | 2560.3352042310 |

**Appendix E: Key Scene Selector Model Results**

Below is a sample of the average results for model-threshold pairs.

The format of the output is model name, followed by the number of valid results for this model-threshold pair, followed by the threshold used. The consensus match increase, false positive reduction and percentage of total video selected are then displayed.

The peak model is an alternative name for the Emotionally Charged Moments model, state model refers to the Emotional State Transition model and strength model refers to the Emotional Intensity Transition model.

The sample below shows the range in which the results of the peak model dramatically increases, reaches its best performance and then falls again. All insightful thresholds are displayed for the state model. The sample chosen for the strength model shows the range in which the model first begins selecting scenes in some videos, up until 0.2 when it can get results for the majority of the dataset. The results stay approximately the same and very negative until the threshold changes from 0.52 to 0.53, though at this point the model can only select valid results from a single video.

Peak Model 7 0.08 Consensus Match Increase = 2.26% False Positive Reduction = 16.54% Perc of Total Video = 13.09%

Peak Model 7 0.09 Consensus Match Increase = 38.05% False Positive Reduction = 21.70% Perc of Total Video = 14.50%

Peak Model 7 0.10 Consensus Match Increase = 36.35% False Positive Reduction = 16.70% Perc of Total Video = 16.64%

Peak Model 7 0.11 Consensus Match Increase = 49.58% False Positive Reduction = 18.55% Perc of Total Video = 17.51%

Peak Model 7 0.12 Consensus Match Increase = 34.76% False Positive Reduction = 24.34% Perc of Total Video = 20.26%

Peak Model 7 0.13 Consensus Match Increase = 25.66% False Positive Reduction = 15.58% Perc of Total Video = 21.29%

Peak Model 7 0.14 Consensus Match Increase = 25.73% False Positive Reduction = 16.70% Perc of Total Video = 21.83%

Peak Model 7 0.15 Consensus Match Increase = 24.83% False Positive Reduction = 14.56% Perc of Total Video = 22.96%

Peak Model 7 0.16 Consensus Match Increase = 28.17% False Positive Reduction = 12.78% Perc of Total Video = 23.91%

Peak Model 7 0.17 Consensus Match Increase = 19.36% False Positive Reduction = 11.66% Perc of Total Video = 25.48%

Peak Model 7 0.18 Consensus Match Increase = 18.46% False Positive Reduction = 15.69% Perc of Total Video = 26.65%

Peak Model 7 0.19 Consensus Match Increase = 16.10% False Positive Reduction = 14.20% Perc of Total Video = 26.97%

Peak Model 7 0.20 Consensus Match Increase = 12.54% False Positive Reduction = 15.13% Perc of Total Video = 29.40%

Peak Model 7 0.21 Consensus Match Increase = 6.93% False Positive Reduction = 15.34% Perc of Total Video = 30.83%

State Model 2 0.00 Consensus Match Increase = - 100.00% False Positive Reduction = 100.00% Perc of Total Video = 2.82%

State Model 2 0.05 Consensus Match Increase = - 100.00% False Positive Reduction = 100.00% Perc of Total Video = 2.82%

State Model 3 0.06 Consensus Match Increase = - 100.00% False Positive Reduction = 100.00% Perc of Total Video = 3.09%

State Model 3 0.07 Consensus Match Increase = - 100.00% False Positive Reduction = 100.00% Perc of Total Video = 3.55%

State Model 3 0.10 Consensus Match Increase = - 100.00% False Positive Reduction = 100.00% Perc of Total Video = 3.55%

State Model 6 0.15 Consensus Match Increase = - 26.80% False Positive Reduction = 37.98% Perc of Total Video = 16.97%

State Model 6 0.20 Consensus Match Increase = - 36.13% False Positive Reduction = 35.76% Perc of Total Video = 22.83%

State Model 1 0.30 Consensus Match Increase = - 100.00% False Positive Reduction = 100.00% Perc of Total Video = 7.14%

State Model 1 0.40 Consensus Match Increase = - 100.00% False Positive Reduction = 100.00% Perc of Total Video = 7.14%

State Model 0 0.48 Consensus Match Increase = NaN% False Positive Reduction = NaN% Perc of Total Video = NaN%

State Model 0 0.50 Consensus Match Increase = NaN% False Positive Reduction = NaN% Perc of Total Video = NaN%

State Model 0 0.60 Consensus Match Increase = NaN% False Positive Reduction = NaN% Perc of Total Video = NaN%

Strength Model 0 0.08 Consensus Match Increase = NaN% False Positive Reduction = NaN% Perc of Total Video = NaN%

Strength Model 1 0.09 Consensus Match Increase = - 100.00% False Positive Reduction = - 318.18% Perc of Total Video = 4.35%

Strength Model 2 0.10 Consensus Match Increase = 66.19% False Positive Reduction = - 4.55% Perc of Total Video = 5.55%

Strength Model 3 0.11 Consensus Match Increase = 10.80% False Positive Reduction = 30.30% Perc of Total Video = 4.67%

Strength Model 3 0.12 Consensus Match Increase = 10.80% False Positive Reduction = 30.30% Perc of Total Video = 4.67%

Strength Model 5 0.13 Consensus Match Increase = - 33.52% False Positive Reduction = - 66.06% Perc of Total Video = 3.30%

Strength Model 6 0.20 Consensus Match Increase = - 22.36% False Positive Reduction = - 54.75% Perc of Total Video = 5.83%

Strength Model 6 0.52 Consensus Match Increase = - 20.01% False Positive Reduction = 36.47% Perc of Total Video = 27.84%

Strength Model 1 0.53 Consensus Match Increase = 3.70% False Positive Reduction = 48.15% Perc of Total Video = 16.07%

**Appendix F: User Manual**

**Chat Bot**

The ChatBot executable processes all MP4 files in a folder one by one. When each video is processed, it prints the result to console and plays an audio response.

To use this program, ffmpeg must be downloaded and installed from: https://ffmpeg.org/download.html

Subscription keys must be obtained from https://www.microsoft.com/cognitive-services/ for a number of APIs. These are the Emotion API, Bing Speech API and Text Analytics API. The subscription keys must be entered into the ChatBot.exe.config file as the values of Emotion-API-Sub-Key, Speech-API-Sub-Key and Text-Analytics-API-Sub-Key respectively.

File paths to point to the input folder and to point to a location to cache information must be provided to the application. This is done by specifying the values of DatasetFolder and CacheFilePath in the ChatBot.exe.config file.

This program is designed to run on a 64 bit architecture.

**Key Scene Dataset Evaluator**

The DatasetEvaluator executable accesses the metadata of all the videos specified by the KeySceneDatset project. It evaluates the reliability of the feedback for each video by calculating metrics and outputting the results to files.

A file path pointing to a folder to contain the output files must be specified as the value of EvaluationOutput in the DatasetEvaluator.exe.config file.

**Key Scene Selector**

The KeySceneSelectorModelEvaluator executable accesses the video data of those deemed reliable from the KeySceneDatset project. It applies a number of models to each video, evaluates the results using metrics and outputs results for each video and on average for a range of model-threshold pairs.

To use this program, ffmpeg must be downloaded and installed from: https://ffmpeg.org/download.html

Subscription keys must be obtained from https://www.microsoft.com/cognitive-services/ for the Emotion API. The subscription key must be entered into the DatasetEvaluator.exe.config file as the values of Emotion-API-Sub-Key.

The file paths of the folders containing the dataset, for storing cached information and for outputting the analysis must be provided to the program. This is done by specifying the values of DatasetFolder, AllScenesOutput and ModelAnalysisOutput respectively in the DatasetEvaluator.exe.config file.

This program is designed to run on a 64 bit architecture.

To understand the output of this program, refer to chapter 4.6.1 of this document.