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

Clodagh Lynch

A dissertation submitted to the University of Dublin,
in partial fulfilment of the requirements for the degree of
Master of Science in Computer Science (Data Science)

Supervised by Professor Khurshid Ahmad and Dr. Carl Vogel
2021

Declaration

I declare that the work described in this dissertation is, except where otherwise stated, entirely my own work, and has not been submitted as an exercise for a degree at this or any other university

Signed: _____

August 31, 2021

Permission to Lend and/or Copy

I agree that the Trinity College Library may lend or copy this dissertation upon request.

Signed: _____

Clodagh Lynch

August 31, 2021

Acknowledgements

Firstly, I would like to thank my supervisors Professor Khurshid Ahmad and Dr. Carl Vogel for their continued guidance and support over the course of the past few months. Their feedback and advice were immensely valuable to the completion of this dissertation and are hugely appreciated.

I would also like to thank Shirui Wang for her assistance in providing me with access to systems and resources that were extremely beneficial to the completion of this work.

A special thank you to my Mom and Dad, Noreen and Bill Lynch, without whom this would not have been possible and whose love and encouragement guided me throughout this entire MSc program.

Finally, a thank you to the School of Computer Science and Statistics at Trinity College Dublin for their support and assistance throughout the academic year.

CLODAGH LYNCH

University of Dublin, Trinity College
August 2021

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

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

Supervisors: Professor Khurshid Ahmad and Dr. Carl Vogel

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

Table of Contents

List of Figures

List of Tables

Abbreviations

AER	Automatic Emotion Recognition
AU	Action Unit
FACS	Facial Action Coding System
F0	Fundamental Frequency
WER	Word Error Rate
ASR	Automatic Speech Recognition
STT	Speech-to-Text
MFCC	Mel-Frequency Cepstral Coefficients
LPCC	Linear Prediction Cepstral Coefficients
PLP	Perceptual Linear Prediction
LLD	Low-level Descriptors
SAVEE	Surrey Audio-Visual Expressed Emotion (Database)
TESS	Toronto Emotional Speech Set (Database)
EmoDB	Berlin emotional database
SER	Speech Emotion Recognition
DNN	Deep Neural Network
SVM	Support Vector Machine
UAR	Unweighted Average Recall
ML	Machine Learning
FPS	Frames per Second
CNN	Convolutional Neural Network
LIWC	Linguistic Inquiry and Word Count
FCPs	Facial Characteristic Points
GI	General Inquirer
NRC	National Research Council Canada
FAPs	Facial Animation Parameters
LDC	Linguistic Data Consortium
CK+	Extended Cohn-Kanade Database
API	Application Programming Interface
FABO	Bimodal Face and Body Gesture Database
ISEAR	International Survey of Emotion Antecedents and Reactions database

USC- IEMOCAP University of Southern California's Interactive Emotional Motion Capture

Chapter 1

Introduction

1.1. Automatic Emotion Recognition

Humans express emotion through a myriad of modalities, such as facial expressions, speech, language, gestures, and so on. Automatic Emotion Recognition (AER) is an emerging area of research in recent years. AER involves computer systems analyzing the physical and linguistic correlates of emotion from facial expression, speech, and text and using those to determine emotion. Some applications of AER include in elderly care, to alert caregivers to emotional distress of elderly patients [49], in robotics, to equip a robotics system with cognitive abilities [26], and in market research, to determine customers' emotional reactivity to new products [18].

The physical correlates of emotion in facial expression, being facial muscle movements, can be extracted from video and encoded in a computer system to determine emotion. For example, when a person is angry, they typically furrow their brow. Doing so causes the three facial muscles *Depressor Glabellae*, *Depressor Supercilii*, and *Currugator* [1], seen in Figure 1, to contract, resulting in frown lines between their eyebrows. This muscle movement is detectable by a computer and can be encoded in an AER system to determine a high chance that the person being analysed is expressing anger. One widely-used encoding system is FACS (Facial Action Coding System), developed by Paul Ekman [2], which encodes facial muscle movements as Action Units (AUs). Each AU encapsulates a specific set of facial muscle movements, with the AU corresponding to the muscle movement of furrowing your brow, described above, being AU4 [3]. AU4, or the 'brow lowerer', is typically characteristic of anger in facial expression and therefore, the detection of AU4 in a subject's facial expression will enhance the classification of anger by an AER system. AER systems use a linear combination of all AUs to determine emotion from facial expression, with those evident in the facial expression carrying more weight in the equation, hence contributing to the enhanced classification of the emotion they are characteristic of.

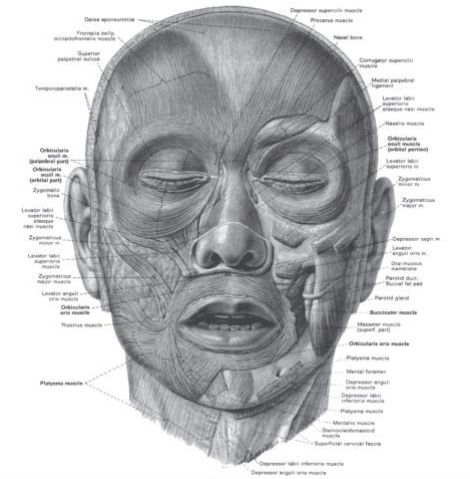

Figure 1 Muscles of the Face [1]

The physical correlates of emotion in speech, being acoustic cues, can be extracted from audio to determine emotion. Hundreds of acoustic cues are used to detect emotion from speech, including fundamental frequency (F0), intensity, mel-frequency cepstral coefficients (MFCC), and voice quality. F0 can be perceived as pitch [4] and is one of the most insightful acoustic cues in determining emotion [5]. Figure 2 and Figure 3 show the mean F0 across all 2000ms chunks of four respective joyful and sorrowful music instrumentals. It can be seen from these that joyful instrumentals produce significantly higher mean F0 values within a smaller range than the sorrowful instrumentals, which produce much lower mean F0 values across a much wider range. This illustrates how the mean F0 as well as variation in F0 of an audio excerpt can be characteristic of various emotional states.

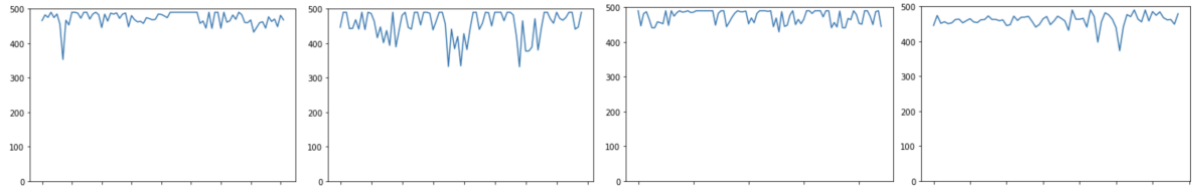

Figure 2 F0 (Hz) of four Joyful musical instrumentals

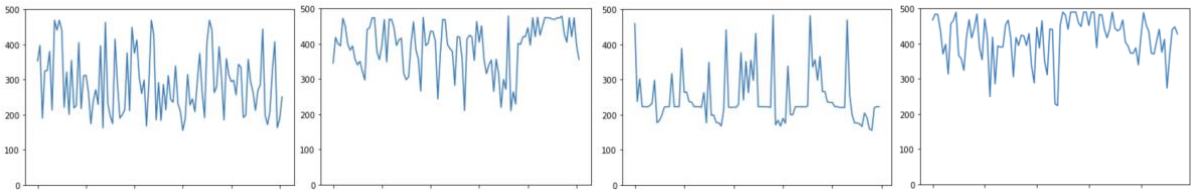

Figure 3 F0 (Hz) of four Sad musical instrumentals

The linguistic correlates of emotion in language can be extracted from text and used to determine emotion as well. However, emotion is near impossible to detect in entirety from text and therefore, text sentiment analysis used to extract positive or negative sentiment from text is more generally employed. Sentiment is used to enhance the detection of an emotion. For example, a sentence “we celebrated and laughed” would return positive sentiment on both ‘celebrated’ and ‘laughed’, whereas a sentence “we mourned and cried” would return negative sentiment on both ‘mourned’ and ‘cried’. The former would enhance the classification of emotions associated with positive sentiment, such as joy, in a multimodal AER system, while the latter would enhance the classification of emotions associated with negative sentiment, such as sadness or anger.

Multimodal emotion recognition has many advantages over unimodal emotion recognition. For example, should an AER system be analyzing the facial expression of a person who has undergone cosmetic procedure Botox, for example, the facial muscle movements characteristic of certain emotions may go undetected by a computer because the Botox injectable prevents the muscles from contracting. Therefore, although a person may be expressing an emotion, the AER system will fail to encode the distinguishing muscle movements as such and hence, the appropriate emotion will not be detected from their facial expression. However, should their voice be evidently expressing anger, it may be possible for the system to still output an accurate emotion estimate.

There have also been studies on the idea of ‘emotional leakage’, which is detectable through use of multiple modalities in emotion recognition. This occurs when a person is trying to conceal their true emotion. If someone is angry, for example, but trying to hide it, their facial expression and the words they are speaking may express non-anger emotion however, anger may be still be detected in their voice through increased pitch, a physiological response of the human body to feeling anger [55], which is more difficult to regulate and therefore, by using more than facial expression or spoken text alone, an AER system might be able to detect this

misalignment in emotion detected caused by this leakage of emotion and subsequently uncover the true emotional state of the subject.

In addition, the manner in which emotion is expressed varies from person to person, some people express themselves more vocally, others more visually, and others with more logic than emotion expression [6], which further substantiates the need for AER systems to employ multiple modalities for emotion recognition.

Therefore, although facial expression can be analysed in isolation, the use of additional modalities for emotion recognition within an AER system improves the overall accuracy of the system and the reliability of emotion estimates. Hence, research has more recently begun to focus on multimodal emotion recognition systems over unimodal emotion recognition systems. These studies mostly employ facial expressions, speech, and text as the modalities used within the multimodal system.

1.2. Research Contribution

One major contribution of this research is the specification of a modality pipeline for multimodal emotion recognition using facial expression emotion, speech emotion, and text sentiment. This in itself is a difficult task due to the complexity of both pre-processing and post-processing steps involved in implementing several emotion recognition software systems within a multimodal framework. The software systems employed are Emotient FACET, Affectiva AFFDEX, and the Microsoft Azure Face API for facial expression emotion recognition and OpenSMILE, OpenVokaturi and DeepAffects for speech emotion recognition. Lexicon-based sentiment analysis is employed for text sentiment analysis, using the lexicons The General Inquirer, used within the RockSteady text analysis tool, the NRC Emotion Lexicon, and the Opinion Lexicon. Pre-processing steps include editing the single MP4 video file input to the pipeline and extracting its frames to ensure compatibility with the requirements of each given facial expression processing system implemented. The MP4 video file is then converted to WAV audio file format and 2000ms/3000ms chunks extracted dependent on the given speech processing system being used. Finally, the WAV audio file is transcribed to text using ASR services Amazon Transcribe, Google Cloud Speech to Text, and Microsoft Azure Speech to Text so that text sentiment analysis can be subsequently conducted. Post-processing steps include ensuring the results of each modality are aligned with one another, so that decision-level data fusion can be conducted. Speech is chosen as the modality to base segmentation on, therefore each modality's results are segmented on a 2000ms basis. The results are then fused using a novel two-step decision-level data fusion process that determines an emotion to be strongly evident in each 2000ms chunk of a video or not.

Each of the software systems implemented in this pipeline are not fully understood systems and therefore, this research provides an insight into the calibration of each system with respect to each other, hence why three software systems or lexicons are implemented for each modality. This is another major contribution of this research. The dataset chosen for this work consists of 10 Irish politicians, 5 male and 5 female, with 5 videos per politician, a total of 50 in the dataset. This specific dataset used and the specification of a modality pipeline for multimodal emotion recognition produced as part of this work provide the experimental apparatus for examining the calibration of each system with respect to each other and analyzing the similarity or difference between emotion or sentiment detected in each system within each modality, as well as between emotion or sentiment detected in each gender by each system implemented.

They also provide the experimental apparatus for examining the physical correlates of emotion from an Irish context, the third major contribution of this work. A gap is identified in the literature surrounding the physical correlates of emotion in facial expression and speech of Irish nationals, which will be addressed as part of this work. There is a lack of results across studies on the relationship between F0 and emotion, F0 and gender, and F0 and age that pertain to Irish-English speakers. There also exists contradictory results for the primary AUs of various emotions. This study aims to address both of these within an Irish context. Joy and anger are the emotions focused on for the analysis of facial expression and speech emotion recognition systems in this study as well as for the analysis of the physical correlates of emotion in facial expression and speech, with both positive and negative sentiment being focused on for analysis of text sentiment analyses.

1.3. Dissertation Structure

The rest of this paper is structured as follows: Chapter 2 (Motivation and Literature Review) outlines previous studies conducted in the field of AER, including unimodal, bimodal, and trimodal emotion recognition system implementations and their data fusion methods as well as literature on the physical correlates of emotion in facial expression and speech. It also identifies gaps in the literature that this research aims to address. Chapter 3 (Design Solution and Validation) describes the specification of a modality pipeline for multimodal emotion recognition in detail, including software systems implemented, pre-processing steps taken, and the proposed novelty two-step data fusion process. A detailed description of the dataset used and an outline of how each experiment conducted in the next section is performed is provided, as well as the specific hypotheses and accept/reject criteria for each. Chapter 4 (Case Studies and Results) the specific subset of the dataset used for each part of the analysis and presents all results derived. The findings from each analysis are discussed, including findings for the physical correlates of emotion in facial expression and speech. The proposed data fusion process is illustrated using outputs from processing a single dataset video and the relationship between AUs and F0 investigated. Chapter 5 (Conclusion and Future Work) provides an overview of work conducted in this study and inferences made, including how these compare to those found in previous literature. Areas of future work are also identified that build on work carried out in this study.

Chapter 2

Motivation and Literature Review

There have been many works [7, 8, 9, 10] in using machines to identify human emotion via facial expression, speech, language, gestures, etc. According to [11], approximately 90% of the literature in the field of AER covers the three modalities of facial expression, speech, and text. Emotions can be defined as "short-term, complex, multidimensional behavioural responses" [12] however, this definition is subjective as there is no widely-accepted or objective definition of emotion [13]. Ekman [14] established the widely-accepted six basic emotional states as happiness, anger, sadness, fear, disgust, and surprise. In the case of non-verbal communication, emotion is detected by examining the physical correlates of that emotional state, being facial actions for facial expressions or vocal cues for speech [15]. In the case of verbal communication, spoken text is often analysed for sentiment, identifying individual words as either positive or negative [16]. Emotions are "near impossible to detect in totality from text" [17] and therefore, sentiment analysis is a more common practice when analysing spoken text [16]. Emotion detection through facial expressions, speech, and language can be unimodal [18, 19, 20], bi-modal [21, 22, 23, 24], or multi-modal [25, 6, 26].

2.1. Facial Expressions

Ekman has contributed greatly to the field of AER, particularly in the case of facial expression. Much of his work builds on that of Darwin's [27], another key contributor to the field of facial expressions and emotion. He developed the Facial Action Coding System (FACS), one of the "most comprehensive and widely used objective taxonomy for coding facial behaviour" [28] that "quantifies facial movement in terms of component actions" [2] by mapping facial muscle movements to various emotions using a set of Action Units (AUs) that correspond to the various facial actions. The original creator was Carl-Herman Hjortsjö in 1970 however, Ekman further developed it in 1978 along with Wallace Friesen and updated it again in 2002 [3]. FACS provides the basis of many emotion recognition software systems, such as Emotient FACET [29] and Affectiva AFFDEX [28]. A linear combination of these AUs are used to classify emotion in such systems [28].

Facial expressions of Ekman's six basic emotional states can be seen in Figure 4, with the AUs corresponding to upper face and lower face facial muscle movements seen in Figure 5. Previous studies [30, 5] have focused on the emotions joy and anger as a means of analysing the physical correlates of facial expressions, with [30] finding both joy and anger to be the emotions successfully detected above the level of chance by Affectiva AFFDEX. Ekman defines the "particular associations between movements of facial muscles and emotions" to be pan-cultural i.e. universal [31]. However, it can be seen from previous literature that different studies produced different results for which AUs, i.e. facial muscle movements, were most dominantly associated with various emotions. Looking at the physical correlates of joy and anger in facial expression, [32] found the most dominant or 'prototypical' AUs to be AU12 (lip corner puller) and AU25 (lips part) for joy, and AU4 (brow lowerer), AU7 (lid tightener), and AU24 (lip pressor) for anger. The database used to determine these results [32] includes subjects of many nationalities, such as Caucasians, Asians, African-Americans and Hispanics. The official website for iMotions [3], the SDK supporting both Emotient FACET and Affectiva AFFDEX software systems, states slightly different results for the prototypical AUs of joy and anger.

It includes AU6 (cheek raiser) but excludes AU25 for joy, and includes both AU5 (upper lip raiser) and AU23 (lip tightener) for anger, but excludes AU24 [3]. Another study [33] produced slightly different results again on the basis of the extended Cohn-Kanade (CK+) database which includes both Euro-American and Afro-American nationalities, among others. It found AU1 (inner brow raiser), AU6, AU12, and AU14 (dimpler) to be the most involved for joy, and AU2 (outer brow raiser), AU4, AU7, AU9 (nose wrinkler), AU10 (upper lip raiser), AU20 (lip stretcher), and AU26 (jaw drop) to be most involved for anger. These variances across different studies' results may be a function of gender, age, and ethnic biases in the databases used for facial recognition [34].

Figure 4 Expressions of six prototypical emotions; from left to right: happy, sad, fear, anger, surprise, disgust [32]

Figure 5 FACS action units (AUs) [35]

Facial Expressions can be spontaneous or posed, with the latter tending to be more exaggerated and hence, easier to identify, with the former being more difficult to identify [36]. Much of previous research has analysed posed, or 'acted', facial expressions collected in controlled environments, whilst more recent studies have researchers focusing on analysing spontaneous expressions, collected in naturalistic environments [37, 38].

Initial works in examining facial expression looked at manually classifying these facial actions using Ekman's FACS. One study [39] used a hybrid system of three methods: 'holistic spatial analysis', 'explicit measurement of features', and 'estimation of motion flow fields' to achieve an accuracy of 91% in classifying facial actions. Another study [2] experimented with methods such as 'optical flow', 'holistic spatial analysis' again, and 'methods based on the outputs of local filters' such as 'Gabor wavelet representations' to classify facial actions, the latter of which achieved 96% classification accuracy. However, both [2] and [39] address the need for automation in classifying facial actions in order to increase the accessibility of facial expression recognition systems. Work in [40] created an automatic 'feature-based' system, while work in [41] used

Independent Component Analysis (ICA) and study [42] used Machine Learning classifiers such as Support Vector Machines (SVMs) and AdaBoost to automatically classify facial actions from facial expressions, with the latter of the three focusing on spontaneous facial expressions specifically. Some of the first automatic facial expression recognition systems to be made freely available for use are CERT (Computer Expression Recognition Toolbox) [43] and OpenFace [44]. CERT can identify 16 AUs and 6 emotions, achieving 90.1% facial expression recognition accuracy on posed facial expressions, and nearly 80% on spontaneous facial expressions. More recent software systems used for automatic facial expression emotion recognition include Emotient FACET and Affectiva AFFDEX, two algorithms available via the iMotions software suite, developed by the technology companies Emotient and Affectiva respectively, with Emotient FACET having been built upon CERT [45]. It was seen that Emotient FACET outperforms Affectiva AFFDEX on posed facial expressions, achieving 97% and 73% classification accuracy on acted facial expressions, respectively, with Affectiva AFFDEX performing best on spontaneous expressions [46]. Emotient FACET was used in [47] to examine the relationship between the facial expressions of a subject and their physiological responses, such as galvanic skin response and skin conductance as well as in [7] to determine the emotional response of people watching presidential candidate debates via their facial expressions. Affectiva AFFDEX was used in [8] to examine the 'emotional reactivity' of people diagnosed with bipolar disorder in order to help with future diagnoses, as well as in [18] as part of product testing for different brand energy drinks to assess the consumers' emotional response to each. With the recent popularity and development of cloud computing, many cloud-based emotion recognition software systems have been developed that use computer vision to detect human emotion through facial expression, such as Microsoft Azure's Face API, Google Cloud's Vision API, and Amazon's Rekognition API [48]. Of these three, Microsoft Azure's Face API returned the highest accuracy and confidence scores in determining emotion on the Karolinska Directed Emotional Faces Database (KDEF) of images of both male and female subjects portraying Ekman's six basic emotional states [48]. The Microsoft Azure Face API has been used in studies such as [49] as a means of monitoring the facial expressions of patients in elderly care via video stream in order to alert caregivers when an elderly patient is displaying a critical emotion i.e. sadness, as well as in studies to determine the emotion of a group of people via their facial expressions [50].

2.2. Speech

Prosodic features of speech can be analysed to detect emotion from spoken utterances [51]. A study by Pereira and Watson [5] found the fundamental frequency (F0) of speech to be the most insightful prosodic feature in differentiating between emotions when compared with both speech duration and sound intensity. Findings in [52] supports this, by finding the mean F0 of a speaker to be the most important in accurately detecting emotion from an analysis on the Toronto Emotional Speech Set (TESS) database. F0 can be defined as "the frequency at which vocal chords vibrate" [53], and can be perceived as pitch [4]. Pereira and Watson [5] found expressions of both joy and anger in speech to have a high mean F0 and wide F0 range when compared with expressions of emotions such as sadness, which is supportive of findings in previous studies also reviewed by the authors. This finding is supported by with results found in [4], shown in Figure 6. From Figure 6 it can be seen that both joy and anger have a higher mean F0 and larger F0 variation than sadness.

Acoustic parameters	Happiness/elation	Anger/rage	Sadness
Voice source—F_0 and prosody			
F_0 mean	>	>	<*
F_0 variation	>	>	<

Figure 6 F0 Mean and Variation for emotions Happiness, Anger, and Sadness ('>' : high/large, '<': low/small) [4]

Studies [51] and [54] defend this finding further, with the latter reviewing a number of previous studies and aggregating each of their results to find both joy and anger to have a higher mean F0 than sadness, which has a much lower mean F0 across all studies' results for both 'emotion portrayals' and natural expressions i.e. acted and spontaneous speech. This F0 and emotion correlate arises as a result of how the human body reacts when feeling certain emotions. For example, when a person experiences joy or anger, their sympathetic nervous system is triggered and hence, their heart rate and blood pressure increase, among other things, which results in louder, faster speech at a higher pitch than when a person is feeling sadness, which results in their parasympathetic nervous system being triggered, causing decreased blood pressure and heart rate, among other things, and subsequently, slow, low-pitched speech [55].

Studies have found F0 to vary across age, gender, and ethnicity. Gender variations arise due to a difference in the size of the larynx in men versus women, causing women's voices to be almost 1.45 to 1.7 times higher than men's [56]. However, results of studies determining the mean F0 for women are contradictory, with some finding the mean F0 for women to be 190Hz and others, 220Hz. Studies defining the mean F0 for men are clearer, with the general agreement being that men's mean F0 averages around 120-130Hz [56]. A study on both Parisian-French and American-English speakers [57] found female speakers to have a slightly higher mean F0 overall than male speakers, with the mean F0 for Parisian-French speakers being 234Hz for females and 133Hz for males and the mean F0 for American-English speakers being 210Hz for females and 119Hz for males. A study on Japanese speakers [58] also found female speakers to have a higher mean F0 than male speakers. However, this study found male speakers to have a larger F0 range than female speakers, which is contradictory to the previous study on American-English and Parisian-French speakers, which found female speakers to have a larger F0 range than male speakers [57].

The mean F0 values for male and female speakers found in the former study [57] showed that, across both genders, American-English speakers overall have a lower mean F0 than Parisian-French speakers. Results from [58] also showed the mean F0 for both male and female Japanese speakers to be much lower than reported in for both American-English and Parisian-French speakers, suggesting a variation in F0 dependent on nationality. Another paper [59] analysed the mean F0 of male and female speakers from around the world between the ages of 20 and 40, including native-English speakers from Canada and the USA, identifying the mean F0 range for various age brackets within this age range. The mean F0 values identified for female speakers of various age brackets from Canada and the USA can be seen in Figure 7, and for male speakers in Figure 8. For both nationalities, it can be seen that the female mean F0 is higher than that of the males, which supports earlier findings. It can also be seen that the mean F0 of Canadian female speakers between the ages of 18-27, being 251Hz based on 104 subjects, is higher than that of American female speakers between the ages of 18-25, being 222.92Hz based on 25 subjects. Similarly, the mean F0 for Canadian male speakers between the ages of 18-28, being 128Hz based on 55 subjects, is higher than that of American male speakers aged 18-25, being 117.77Hz based on 24 subjects. This shows that overall, Canadian speakers have higher mean F0 than

American speakers across both genders, which further supports the notion of mean F0 variability dependent on nationality.

Country	age (#)	F0 ± SD (Hz)
Canada	18–27 (104)	251 ± 28
USA	21–34 (14)	209.68 ± 3.88
	19–35 (34)	232
	25.30 ± 1.70 (5)	227.99 ± 9.13
	18–25 (25)	222.92 ± 20.25
	20.35 ± 1.64	-

Figure 7 Mean F0 (Hz) of Female speakers aged 20-40 from Canada and the USA

Country	age (#)	F0 ± SD (Hz)
Canada	18–28 (55)	128 ± 21
USA	20–42 (10)	133 ± 18
	24.90 ± 1.9 (5)	127.56 ± 36.67
	18–25 (24)	117.77 ± 15.51

Figure 8 Mean F0 (Hz) of male speakers aged 20-40 from Canada and the USA

F0 can even vary dependent on the language being spoken, with one study [60] finding the F0 of Finnish speakers to increase when the speaker is speaking in a language "foreign" to them, due to the increased stress of speaking a language that is not native to them. However, this change in F0 was not as apparent for English speakers [60].

One study [61] finds a decrease in F0 with age in both male and female speakers, with a more significant decrease seen in female F0 due to hormonal changes that occur with the onset of menopause. This finding is supported by another study [62] that found the F0 of females to remain stationary up until menopause, when it decreases significantly. It also finds the male F0 to decrease significantly around puberty, and continue to decrease until approximately age 35, where it remains stationary until age approximately age 55, where it begins to increase again [62].

Few studies [63, 64, 65] have looked at F0 in native Irish and Irish-English speakers. The first study [63] focuses solely on the impact of broad, narrow, and contrastive focus on F0, and found F0 to increase on the "focal element" for four varieties of Donegal Irish. The second paper [64] also conducts experiments in this area, incorporating Donegal English as well, and the third paper [65] examines the low turning point of F0 in Belfast English, finding it more variable than previously reported. However, there exists no studies outlining the mean F0 nor the F0 range found for male or female Irish-English speakers.

Speech Emotion Recognition (SER) systems are used to detect emotion from speech signals [13]. A review of previous literature in the field of SER was conducted by [13] and many applications of SER outlined, including its use in call centres in order to analyse the behaviour of call attendants and improve customer service, or in a smart car where the emotional state of a driver can be detected, alerted, and thus, reduce the risk of a car accident occurring. SER systems uses a two-phased approach, phase one being feature extraction and phase two being emotion classification [55]. OpenSMILE and Praat are two commonly-used toolkits for feature extraction from speech prior to emotion classification [66]. The study in [67] uses OpenSMILE to extract prosodic features such as pitch and voice-quality as well as cepstral features such as Linear Prediction Cepstral Coefficients (LPCC), MFCC, and Perceptual Linear Prediction (PLP) from speech. There exist many emotional speech databases that are used to subsequently train SER systems to classify emotion based on such acoustic features extracted. One widely-used database is the Berlin Emotional Database (EmoDB), with others including the INTERFACE database, the Danish emotional database, the Surrey Audio-Visual Expressed Emotion (SAVEE) database, and the TESS database [55, 68]. However, out of the 17 emotional speech databases named in [55], EmoDB is the only database that is publicly available and free of any license fee, and out of the 24 databases discussed in [68], it is the only open-access database, hence its' popularity and widespread use in SER

systems. It's demonstrated in [66] that features extracted using OpenSMILE and Praat achieve a higher recognition rate in all categories of emotions – joy, sadness, and anger – using the EmoDB database over the Danish emotional database. It was shown in [69] that over a range of four different classification methods for speech emotion recognition, namely SVM, ranking SVM, and two late-fusion methods, all four achieved highest classification accuracy when trained using the EmoDB database as opposed to the LDC (Linguistic Data Consortium) database or the FAU Aibo Emotion Corpus of spontaneous speech. EmoDB is used in [70] to train a Support Vector Machine for the classification of joy, sadness, and neutral, achieving 95.1% accuracy. It is also used in [71] to train a Deep Neutral Network for emotions angry, neutral, and sadness, achieving 96.97% classification accuracy. The EmoDB database was used, along with the eNTERFACE database and others, to benchmark and train models for the open-source speech emotion recognition tooklit, OpenEAR [72], which builds on the OpenSMILE toolkit for feature extraction [73]. A review of existing SER systems, conducted in [74], identified OpenVokaturi, Beyond Learning, and OpenSMILE as three open-source SER software systems, with the latter solely being used for feature extraction, hence why classification toolkits such as OpenEAR must be used in conjunction with it. Another review of existing SER systems, conducted in [75], identified OpenVokaturi, Beyond Learning, Good Vibrations, and EmoVoice as four pre-built SER systems, of which only OpenVokaturi and EmoVoice are freely-available for use, with EmoVoice having a 'high difficulty of use'. OpenVokaturi was used in [9] to detect the emotion of a radio listener in a phone call to the radio station requesting a song, achieving 66% classification accuracy, as well as in [19] as part of the specification of a robotic system, used to give the robot the ability to detect emotion from speech. OpenEAR is used in [76] to extract the emotions from audio segments of a movie as part of an intelligent movie-trailer creation system.

2.3. Text

Human emotion is very complex and thus, it is very challenging to detect emotion from text [6], especially when a piece of text, or even a single word, contains more than one clear emotion [77]. Text sentiment is more general, and relates to whether linguistic content is of positive, negative, or neutral feeling [78]. Emotional states such as anger, fear, disgust, and sadness are associated with negative sentiment, while joy and surprise are associated with positive sentiment [79]. Text sentiment can be used to "increase confidence" in emotions predicted [78]. For example, information of negative sentiment can help in the prediction of anger [79]. However, there are some emotions that can be associated with both positive and negative sentiment, such as surprise [77]. This was established in a study that looked at classifying both emotion and sentiment in tweets extracted from Twitter and found that surprise can express both positive and negative sentiment i.e. positively-surprised or negatively-surprised [77]. Therefore, a more comprehensive understanding of human emotion can be achieved and output by AER systems by including both emotion and sentiment.

Studies have found gender differences in text production to exist, with women typically expressing more joy in text than men [80, 81]. One study [80] found women to express more positive emotion and joy in a sample of stroke survivor's tweets, while another [81] found women to use a much more words associated with joy and sadness in work-emails than men, who used much more words associated with trust and fear.

Text sentiment analysis is the process of extracting sentiment from text, and can be categorised into three main approaches, being machine learning-based, lexicon-based, as well as a hybrid approach [82]. Text sentiment analysis has many real-world applications in public health, politics, and customer relationship

management, such as tracking public sentiment during election periods or detecting cyber-bullying [78]. Machine learning-based text sentiment analysis makes use of supervised classifiers such as Naïve Bayes, Support Vector Machines, Decision Trees, and Rule-Based classifiers, as well as unsupervised classifiers [82], lexicon-based text sentiment analysis requires a dictionary of words, with each word annotated as positive or negative [83], while a hybrid approach uses a combination of both machine-learning and lexicons. One major advantage of the lexicon-based approach to text sentiment analysis is that it does not require a labelled dataset for training, as machine-learning approaches do [83] and therefore, they are more accessible and extensible. Some widely-used text sentiment lexicons include the General Inquirer, the NRC Emotion Lexicon, and SentiWordNet [78] which each contain a set of positive and negative words, every word in the lexicon being annotated as either positive or negative in sentiment. The General Inquirer is used in [84] to determine sentiment from employee company reviews found on Glassdoor and subsequently analyse the relationship between employee feeling and company earnings, as well as in [20] to determine the sentiment expressed in movie reviews. Rocksteady, a "text analytic system for extracting sentiment from text" [85] developed at Trinity College Dublin, uses the General Inquirer lexicon to compute the frequency of positive and negative words in a text document. It has been used in many studies [85, 86] as a means of performing sentiment analysis on text. The NRC Emotion Lexicon is used in [10] to extract sentiment from tweets posted before, during, and after a series of cricket matches in order to analyse the sentiment of fans throughout each match. It is also used in [87] as part of a movie recommender system to perform sentiment analysis on the subtitles of movies in order to derive the general emotion of the movie. SentiWordNet is used in [88] as part of an automatic classification system for film reviews, as well as in [89] for performing sentiment classification on news headlines. In an evaluation of six sentiment lexicons [90], including the General Inquirer and the NRC Emotion Lexicon, SentiWordNet performed worst, being excluded from the core evaluation due to its poor performance against the baseline. This evaluation [90] also evaluates the Hu & Liu Opinion Lexicon, which is found to achieve the highest sentiment classification accuracy on product reviews in the case where no training corpus is available. The Hu & Liu Opinion Lexicon has been found to perform comparably well to the General Inquirer lexicon, with only a marginal difference in classification performance [91].

2.4. Transcription

Text can be derived from audio using Automatic Speech Recognition (ASR). ASR is defined as "the independent, computer-driven transcription of spoken language into readable text in real time" [92]. Many studies have used ASR within a bimodal speech and text AER system that only takes audio as input in order to transcribe the audio for subsequent text analysis [93, 94]. The basis of modern ASR is DNNs (Deep Neural Networks) however, high-performing DNNs with substantially improved performance require huge amounts of labelled training data and therefore, cloud-based API services are a more efficient choice for implementing ASR [95]. Four cloud-based ASR services are evaluated in [95] based on their Word Error Rate (WER) scores, obtained by comparing the transcription derived from each ASR service with a manually-transcribed reference transcription [96]. The formula to calculate the WER score can be seen in Figure 9 below. The lower a WER score achieved by the ASR service, the more accurate its transcription is [95].

$$\text{WER} = \frac{\text{Insertions} + \text{Deletions} + \text{Substitutions}}{\text{Length of reference string}} \times 100\%$$

Figure 9 WER formula [96]

The four ASR services evaluated in this study [95] are Amazon Transcribe, Google Cloud Speech-to-Text (STT), IBM Watson STT, and Microsoft Azure Cognitive Services STT, which each return WER scores of 6.2%, 5.6%, 5.5%, and 5.1%, respectively, making Microsoft Azure's ASR service the highest performing of all four, reaching a WER comparable to the WER of human parity, being 5.1% as well [95]. Another paper [97] evaluated the IBM Watson STT, Google Cloud STT, and Microsoft Azure STT and also found Microsoft's ASR service to have the lowest WER score, being 3.24%. Other studies [98, 99], however, have found Google Cloud's STT service to be most accurate, with another finding Amazon Transcribe to be most accurate [100].

There have also been many studies that explore the impact of external factors on the evaluation of ASR systems that can lead to higher WER scores and less accurate transcriptions. These external factors include things such as speech disfluencies, background noise, and speaker variability in terms of the gender of the speaker, their dialect, and variances in their style of speaking due to their personality and the emotion being expressed [101, 102], as well as their rate of speech [102]. All of these factors pose a challenge to ASR systems [95] and hinder the accuracy of the transcription. Additional factors such as technical issues during the recordings of the audio, reduction in quality of the audio files as a result of data processing, or poor speech intelligibility due to meagre speaker articulation, for example, can also impact the accuracy of the transcriptions [98]. In addition, human-error that occurs in the manual reference transcription [103] can cause the WER of the ASR service-produced transcription to be higher. However, [104] proves the feasibility of using an ASR system to transcribe text for the subsequent purpose of emotion recognition in spite of imperfect transcription accuracy by showing that the performance of trained emotion recognition classifiers remains "roughly constant as long as word accuracy doesn't fall below a reasonable value" [104]. In addition, [105] states that a "perfect word chain" is not critical, and errors in the transcript are really only adverse should they change the "tone" of the piece. Therefore, it can be said that these ASR systems are a viable method of transcribing speech for subsequent text analysis. The Python package 'jiwer' is used in many studies to automatically calculate the WER scores of transcriptions [106, 107].

2.5. Multimodal Emotion Recognition

Facial expression, speech and text modalities used in emotion recognition each have their own advantages however, they also each face limitations, which can be compensated by integrating multiple modalities into a single multimodal emotion recognition system [108]. Instead of processing each of these modalities in isolation, recent studies have looked at combining two or more of them. Doing so results in more reliable emotion estimates and increased confidence in the system [21].

Data Fusion is the process of aggregating multiple data modalities to produce an overall result that is more accurate than that achieved by any of the data modalities individually [108]. It can be conducted at feature-level or decision-level and has many applications in multimodal emotion recognition systems [108, 109, 110]. Feature-level fusion, or 'early' fusion, involves all modalities being processed and classified synchronously to output a single emotion result [108], whereas decision-level fusion, or 'late' fusion, allows for

each modality to be processed independently and the emotion results output by each modality combined afterwards [110].

2.5.1. Facial Expression & Speech

Humans are 35% more accurate in identifying emotion via facial expression than speech [109]. However, facial expressions and speech have been proven complementary to each other in emotion expression [110] and therefore, AER systems should, at the very least, include facial expression analysis, along with at least one other modality, such as speech [109]. Work conducted in [21] supports this notion by showing that emotion recognition from facial expression is more accurate than emotion expression from speech, but that a decision-level fusion of both modalities is more accurate than each unimodally. One paper [21] reviews a number of existing approaches to bimodal emotion recognition using facial expressions and speech. It outlines one approach taken by Han et al. that implements feature-level fusion of facial expression and speech using binary SVM classifiers. Twelve geometric facial features are extracted along with twelve audio features which are subsequently fed to the classifier. The bimodal classifier results in a 5% improvement in classification accuracy over the unimodal facial expression classifier, and a 13% improvement over the unimodal speech classifier [21]. Another study reviewed in [21] conducted by Sebe et al. saw a bimodal classifier of facial and speech features achieve 90% classification accuracy, with each of the unimodal classifiers for facial expression and speech only achieving 56% and 45% classification accuracy, respectively [21]. This paper [21] conducts two experiments itself on the bimodal fusion of facial and speech modalities, the first performing feature-level fusion and the second performing decision-level fusion. Results showed that only the decision-level data fusion approach achieved higher classification accuracy over both the unimodal speech and unimodal facial expression classifiers [21].

One challenge to bimodal facial expression and speech emotion recognition is the impediment to a person's facial expression caused by the movement of their mouth as they are speaking. This can obstruct the appropriate detection of AUs from a person's facial expression by a machine and can lead to less accurate emotion predictions [111]. Pelachaud et al. [112] examine this relationship between facial expression and speech, and in particular the impact of rate of speech on facial muscle movement, with rapid speech causing the mouth to lose its characteristic shape, for example. They state that the appearance of facial action units, such as those used in Ekman's FACS, follow the voice pattern, and hence, their intensity is proportional to speech-rate. An equation, seen in Figure 10, is proposed by Pelachaud et al. that uses speech-rate to determine the true intensity of an AU.

$$intensity\text{-}AU = minimum\text{-}AU * speech\text{-}rate + maximum\text{-}AU * (1 - speech\text{-}rate).$$

Figure 10 Action Unit Intensity equation relative to speech rate [112]

2.5.2. Speech & Text

It has also been proven that combining text features with speech improves overall speech emotion recognition accuracy [23]. One study [23] implements feature-level fusion of speech and text by extracting acoustic features from speech and word embedding from text and inputting these features into a bimodal emotion recognition model. The bimodal model achieves much higher classification accuracy, being 75.49%, than each of the unimodal models, with the unimodal speech model achieving 71.34% and the unimodal text model achieving 66.09%. Another study [113] focuses on classifying negative and non-negative emotion using

decision-level fusion of speech and text information, finding a performance improvement of 45.7% classification accuracy over a unimodal speech model, and 32.9% improvement over a unimodal text model using a linear discriminant classifier. An ASR system is used in [93] to derive text from audio in order to classify positive and negative sentiment in a bimodal speech and text emotion recognition system that solely takes audio as input. Results found that fusing speech and text modalities improved the overall classification accuracy of the system. It was also found that text was more accurate in predicting positive or negative sentiment than speech [93]. A bimodal emotion recognition system is built in [24] that classifies both emotion and sentiment. It takes a single audio file as input and therefore, also uses an ASR to derive the text, with the ASR-derived transcript achieving a WER rate of 7.6%. A CNN (Convolutional Neural Network) is used to perform both the emotion classification on speech and sentiment classification on text, with each achieving a classification accuracy of 65.7% and f-measure of 82.5, respectively. [114] also analyses both emotion and sentiment in a bimodal speech emotion recognition system, extracting six emotions from speech and text, along with two sentiments from text also. The emotions joy and surprise are grouped with positive sentiment and the emotions angry, sad, disgusted, and scared are grouped with negative sentiment. Decision-level data fusion is performed using a weighted sum of probabilities, discussed in Section 2.5.4 (Data Fusion) below. The performance is measured in terms of Unweighted Average Recall (UAR), with the bimodal system achieving 72.01% UAR, a performance improvement of 38.29% over a unimodal speech model.

2.5.3. Facial Expression, Speech & Text

One study by Rozgić et al. [25] proposes a method of multimodal emotion recognition that uses facial, speech, and text features. The system architecture for this study [25] can be seen in Figure 11. Emotion recognition was conducted on a sentence-level basis, with a total of 5,531 sentences being analysed. The database used was the University of Southern California's Interactive Emotional Motion Capture (USC-IEMOCAP) corpus of audio and visual data. Facial features were extracted that “represent vectors between different points on the face”, based on Facial Animation Parameters (FAPs). Speech features such as F0, intensity, MFCCs, jitter, and shimmer were extracted from each audio clip using audio analysis tools such as Praat. The text of each sentence was obtained using ASR, with two sets of textual features subsequently extracted, namely sentiment and word stem features, using the LIWC (Linguistic Inquiry and Word Count) and General Inquirer lexicons. After feature extraction, an ensemble tree of binary SVM classifiers was created. Each SVM classifier is trained on either all or a subset of the available multimodal features, as well as on a subset of possible classified emotions, being anger, joy, sadness, or neutral. The proposed ensemble tree method outperforms two state-of-the-art baseline classifiers on each category of training features, with the biggest improvement in classification accuracy seen for the ensemble tree using all facial expression, speech and text features, achieving 69.4% accuracy as opposed to 66.4% and 66.5% achieved by each respective baseline classifier.

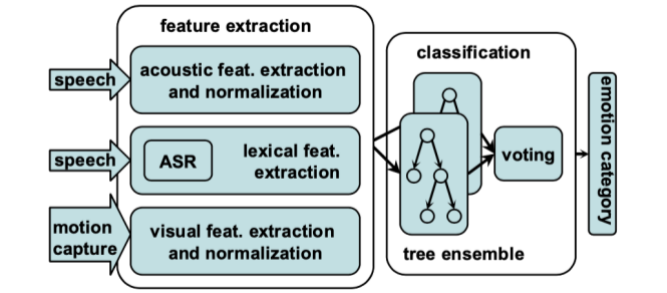

Figure 11 System architecture of multimodal emotion recognition using an ensemble tree of binary SVM classifiers [25]

[6] builds a multimodal sentiment and emotion recognition system to examine the hypothesis that fusing multiple modalities improves performance over unimodal systems. The target emotions to be extracted from each modality are based on Ekman's six basic emotional states, anger, disgust, fear, joy, sadness, and surprise. The International Survey of Emotion Antecedents and Reactions (ISEAR) database was used to extract text features such as emotion, using a new lexicon called EmoSenticSpace, and sentiment polarity, using the SenticNet lexicon. The eNTERFACE database was used for speech, with audio features such as MFCC and Spectral centroid extracted using audio analysis tool JAudio. The CK+ database was used for facial expressions, with features extracted by tracking facial characteristic points (FCPs) using the facial recognition software Luxand FSDK 1.74. The FCPs used to extract facial features from the data can be seen in Figure 12 below. Feature-level fusion was performed by concatenating each of the modalities' features together in a single feature vector used to train several classifiers, with the SVM classifier achieving highest classification accuracy. The multimodal classifier proposed in this paper [6] achieved 72.10% accuracy on the MMI database, and 61.21% on the FABO database, being 16.5% and 25.71% higher than accuracies achieved by current state-of-the-art classifiers on each database, respectively. The multimodal classifier also achieved higher accuracy in all six emotion categories than each of the unimodal facial, speech, and text classifiers. The authors have employed this multimodal framework in a real-time multimodal emotion recognition system, the architecture of which can be seen in Figure 13 below. This system additionally utilised ASR to derive text from audio.

Figure 12 Facial Characteristic Points detected in a facial image [6]

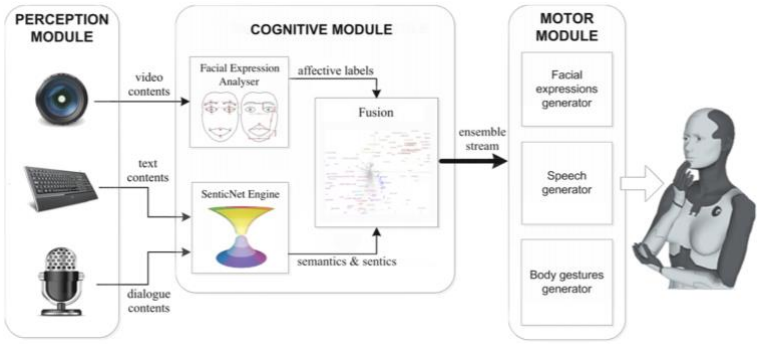

Figure 13 Real-time multimodal emotion recognition system architecture [6]

Another study [26] employs a multimodal approach to emotion recognition as part of a robotics system equipped with cognitive abilities. The architecture for this system is seen in Figure 14 below. The system accepts speech and audio-visual data as input, extracting facial expression features such as eyebrows, lips, and facial muscle movements from video, speech features such as intensity and quality of voice from audio, and text features such as syntactic structure and meaning from transcripts derived from audio. Decision-level data fusion is implemented by inputting the results of each of these individual modalities' classifiers into the input layer of a

Multilayer Perceptron (MLP) neural network. The MLP outputs a prediction for each of the target emotion classes, being anger, contempt, fear, joy, neutral, sadness surprise, and disgust. Classification accuracy achieved by this MLP model outperforms one state-of-the-art approach, achieving mean accuracy of 81.65% over 74.09% achieved by the state-of-the-art model. Furthermore, this study experiments with the inclusion of additional features, such as age, gender, and culture in the MLP neural network model, which resulted in further improved classification accuracy of 85.45%, an 11.36% improvement over the state-of-the-art. This study takes emotion recognition one step further, developing a semantic multimodal non-observable emotion recognition system by associating emotions recognised by the MLP model with their context using N-ary ontologies from event semantic descriptions [26].

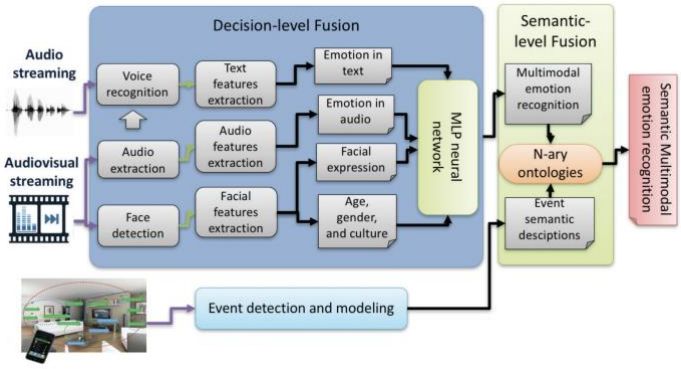

Figure 14 Multimodal Emotion Recognition architecture in an intelligible robotics system [26]

2.5.4. Data Fusion

Both feature-level and decision-level data fusion have been shown to outperform each other on different occasions. Decision-level fusion was found to have higher emotion recognition accuracy over feature-level fusion in [109, 21], while [115] showed feature-level fusion to have higher classification accuracy. Although many studies, such as [6], successfully employ feature-level fusion of facial expressions, speech, and text in multimodal emotion recognition systems, it is said that decision-level fusion is a more appropriate fusion strategy in systems where the various modalities' signals are asynchronous [108, 110]. The speed of light is $3x10^8$m/s, and the speed of sound is 343m/s, being the rates at which video and audio signals are received, and at which text can be transcribed in real-time via audio signals. The lack of a common timeframe across the various data modality signals implies they are "not sufficiently correlated to be fused at the feature-level" [109].

A decision-level fusion formula is put forward in [22] for the bimodal fusion of speech and facial expression results that calculates the weighted sum of probabilities output by each modality to determine an overall emotion probability for each respective emotion. The formula used can be seen in Figure 15, with w_0 being the weight given to facial expression emotion probabilities, s^{face}, and w_1 being the weight given to speech emotion probabilities, s^{speech}. The weights used in this study on the eNTERFACE'05 database classification results were 0.8 and 0.2 for facial expression and speech, respectively, with experimental results proving the fusion of both speech and text to be significantly more accurate than unimodal results [22].

$$S = w_0 S^{face} + w_1 S^{speech}$$

Figure 15 Decision-level weighted formula for facial and speech fusion [22]

Another formula is put forward in [113] that uses a logical "OR" function with a 'combined decision rule' for the bimodal fusion of speech and text results, seen in Figure 16 below. This rule determines the overall emotion result being chosen as such if either the speech model or text model output that emotion as it's result.

$$d = \begin{cases} +1, & \text{if } d_0 + d_1 \geq 0 \\ -1, & \text{else} \end{cases}$$

Figure 16 Decision-level Logical "OR" formula for speech and text fusion [113]

Decision-level data fusion is conducted in [111] on facial expression and speech emotion results by "finding the emotion that has highest sum of probabilities from all modalities" and having that be the overall emotion recognition result. The equation used can be seen in Figure 17 below, with *m* being the number of unimodal classifiers used, *p* being the probability, and *e* being the emotion being classified.

$$L = \quad argmax_i \sum_{j=1}^{m} p(e_i)_j \ , i=1,....5$$

Figure 17 Decision-level equation for bimodal fusion of facial and speech [111]

The formula in Figure 18 is used for decision-level fusion of facial expression, speech, and text results in a multimodal sentiment analysis system. q_1, q_2, and q_3 represent the respective weights attributed to each modality, with an equal value of 0.33 being used for each weight in this study [116]. s^a, s^v, and s^t represent the probability of sentiment for audio, visual and text, respectively, with *i* denoting the category of sentiment the probability score pertains to. In this study [116], three categories of sentiment are analysed, being positive, negative, and neutral, and therefore, the overall sentiment is determined by which sentiment category produces the largest value as per the formula below.

$$l' = \text{argmax}_i \left(q_1 s_i^a + q_2 s_i^v + q_3 s_i^t \right), \quad i = 1, 2, 3, \ldots, C$$

Figure 18 Decision-level multimodal data fusion formula [116]

A weighted sum of probabilities is used to conduct decision-level data fusion in an emotion recognition system that extracts speech emotion, text emotion, as well as text sentiment to classify emotion [114]. The results of all three are input into the equation seen in Figure 19, below. In this equation α and β are the weights given to speech emotion (P_a) and text emotion (P_t) probabilities, respectively, with the weight given to text sentiment probability (P_s) being inferred from the other two weights (1 - α – β). The highest accuracy of 72.01% UAR was achieved using weights α = 0.8 and β = 0.2, which actually cancel out P_s from the equation, meaning text sentiment is not encapsulated in the final data fusion result. The next-best results include text sentiment in the equation by using α = 0.7 and β = 0.1, inferring a 0.1 weight for P_s, achieving 71.44% UAR, only 0.57% lower than the highest result achieved.

$$P_b = \alpha * P_A + \beta * P_T + (1 - \alpha - \beta) * P_S, \quad \beta \leq 1 - \alpha$$

Figure 19 Decision-level weighted sum of probabilities from speech emotion, text emotion, and text sentiment [114]

In order for the results of emotion recognition across all data modalities to be fused at the decision-level, each modality must be segmented into equal "emotion units" [105] prior to fusion so that the results of each modality are aligned with each other. Emotion units are difficult to identify given the differences in rate of emotion recognition across the different data modalities. One study [117] examining recognition of fear, joy, anger, sadness, disgust, surprise, contempt and pride from facial expressions found that, on average, when emotions were correctly identified, they were recognised in under 650ms for all emotions except fear, which

was recognised in under 700ms on average. These results were derived from presenting participants with static images of facial expressions [117]. Another study [118] identified 400-1200ms as the range in which anger, disgust, joy, sadness, and neutrality can be recognised from speech. It was seen from one study [119] that shorter time intervals, such as 500ms, achieve lower classification accuracy when used in speech emotion recognition when compared with longer intervals, such as 1000ms, with each achieving 67.2% and 70.7% accuracy, respectively. It was also found that emotion segments from speech that contain seven syllables to range from 1260ms to 2126ms in duration dependent on the emotion being conveyed [120], with joy and anger being approximately 1665ms and 1799ms, respectively. With regards to text, it can be seen that the 'word' is the "smallest meaningful emotional unit" [105] of text. Looking at all three modalities, [105] identifies speech to be the best and most advantageous modality to base segmentation upon.

2.6. Gap in the Literature

One gap in the literature I have identified as an area for experimentation in this study is an analysis of the mean F0 for Irish-English speakers. This will include deriving the mean F0 for female Irish-English speakers and male Irish-English speakers and conducting a comparison between the two, looking at which gender has a higher mean F0 and which has a larger mean F0 range and comparing findings to what is seen across literature. There exist results on the mean F0 of both male and female American-English, Parisian-French and Japanese speakers however, there is a gap in the literature for such results for Irish-English male and female speakers. In addition, there appears to exist a variation in overall mean F0 of both genders across different nationalities and therefore, determining the mean F0 for Irish-English speakers will be a valuable contribution to this literature.

Building on this gap in the literature for results on Irish-English speakers, these exists a study that outlines the mean F0 of male and female speakers within various age brackets, including Canadian and American speakers [59]. However, this study did not include any Irish speakers, nor did it examine the mean F0 of any subjects older than 40 years of age, which is a notable age range that could be used to observe the variations in mean F0 that occur in male speakers around the age of 55 and in female speakers around the age of onset of menopause. Therefore, work conducted in this research will investigate the mean F0 of male and female speakers with respect to their age for the age range of 35 to 62 years, a range wide enough to ideally observe variations in mean F0 for both males and females as they age.

There exist contradictions across studies identifying the most dominant or prototypical AUs for joy and anger and therefore, this work will also address these contradictions by identifying the AUs 'most dominant' in facial expressions of joy and anger from an Irish context as well.

Many decision-level data fusion methods have been identified in previous studies [22, 111, 113, 114, 116] however, none of them conduct data fusion on the specific combination of facial expression emotion, speech emotion, and text sentiment, being the modalities used in this study. One method in [116] uses facial expression sentiment, speech sentiment and text sentiment, and another in [114] uses speech emotion, text emotion, and text sentiment. However, a gap exists in the literature for a data fusion method using facial expression emotion, speech emotion, and text sentiment. This work proposes a novelty two-step data fusion process that builds off existing methods outlined in this literature review in order to fuse the results of these three modalities.

Overall, this work proposes a specification of a modality pipeline for multimodal emotion recognition that implements many of the software systems, lexicons, and data fusion methods identified in the literature, including Emotient FACET, OpenSMILE, and RockSteady, for example.

2.7. Summary

This chapter has discussed previous studies and research conducted in the field of AER, including on the physical correlates of emotion in facial expressions and speech. Various decision-level data fusion techniques were reviewed, and software systems and lexicons used in facial expression and speech emotion recognition as well as text sentiment analysis outlined. The motivation for using multiple modalities in an AER system over individual modalities in isolation is presented, as well as gaps in the literature which this study aims to address. The next chapter (Chapter 3 Design Solution and Validation) will discuss in detail the specification of a modality pipeline for multimodal emotion recognition put forward as a result of this work, including the specific software systems and lexicons implemented, pre-processing and post-processing steps conducted, and the novel decision-level data fusion process proposed.

Chapter 3

Design Solution and Validation

A high-level system architecture for the specification of a modality pipeline for emotion recognition, explored as part of this research, is presented in Figure 20.

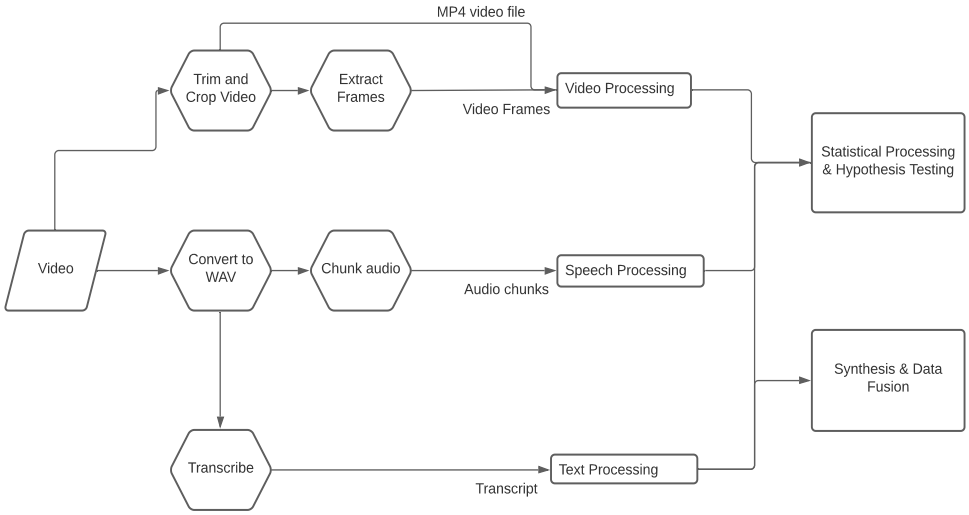

Figure 20 High-level system architecture

This high-level architecture conceptualises a multimodal emotion recognition system that uses facial expression, speech, and text extracted from a single MP4 video file. Pre-processing must be carried out on the MP4 video file in order to ensure data is in the appropriate format for processing each modality. The MP4 video file is first cropped and trimmed, and either used as input to the video processing software directly, or its frames extracted and those used as input, depending on the input requirements of the software system. The MP4 video file is converted to WAV audio file format to be input to the speech processing software systems, with the WAV audio file being chunked into either 2000ms or 3000ms segments prior to being input to the software system dependent on the input requirements of the system also. The text is transcribed from the audio WAV file using ASR transcription software, with the output used as input to subsequent text processing. Video processing extracts emotion per frame of an MP4 video file, the categories of emotions extracted dependent on the functionality of the individual video processing software systems. Speech processing extracts emotion per segmented audio chunk input to the system, with the categories of emotions also determined dependent on the software system used. Text processing determines the sentiment of each individual word in the transcription, being either positive or negative in sentiment. Hypothesis tests such as the Kruskal-Wallis Rank Sum test and the Wilcoxon Rank Sum test are performed in order to assess the distribution of emotion or sentiment across software systems and lexicons as well as across genders to determine if there is a significant difference between emotion or sentiment detected. Other statistical processing includes using Spearman's Rank-Order Correlation to determine the 'most dominant' AUs for various emotions and calculating z-scores to examine the relationship of mean F0 and F0 range with emotion, gender, and age. Data Fusion techniques are employed to determine if an emotion is strongly evident across each video chunk processed within the multimodal emotion recognition framework.

Statistical Processing and Hypothesis Testing are conducted solely on joy and anger results for facial expression and speech processing, and positive and negative sentiment results for text processing. Data Fusion is conducted solely on joy and positive sentiment for one video in the dataset as a means of illustrating the process. The choice of joy and anger as the emotions for analysis is reasoned in previous studies that focused on these emotions [30, 5] as well as the Affectiva AFFDEX only detecting joy and anger above the level of chance out of all seven emotions detected by this software system [30]. Positive sentiment is also chosen to illustrate the data fusion process given joy is solely associated with positive sentiment [79].

A triad of software systems/lexicons are implemented for each modality as a means of gaining an insight into the calibration of each with respect to each other, this being another major contribution of this research.

A low-level system architecture can be seen in Figure 45 below, the components of which will be discussed in more detail in this section.

3.1. Dataset Description

The dataset used in this research consists of videos of speeches and interviews of five Irish male politicians – Leo Varadkar, Michael Martin, Simon Coveney, Paschal Donohoe, and Stephen Donnelly, and five Irish female politicians – Arlene Foster, Michelle O'Neill, Mary Lou McDonald, Helen McEntee, and Mairead McGuiness. The age of all these politicians ranges from 35 to 62, with politicians from both the Republic of Ireland and Northern Ireland being included. All videos are publicly available on YouTube and were downloaded in MP4 video file format using an online YouTube downloader tool. The dataset consists of five videos per politician, the overall dataset made up of a total of 50 videos, with 25 pertaining to each gender. The duration of the videos ranges from 1 minute 58 seconds to 5 minutes 6 seconds, with a mean duration of 3 minutes 8 seconds. A detailed description of each politician in the dataset, including their age, gender, political position and nationality, is seen in Table 1.

Name	Age	Gender	Political Position	Nationality
Leo Varadkar	42	M	Tánaiste	ROI
Michael Martin	60	M	Taoiseach	ROI
Simon Coveney	49	M	Minister for Defence	ROI
Paschal Donohoe	46	M	Minister for Finance	ROI
Stephen Donnelly	45	M	Minister for Health	ROI
Arlene Foster	51	F	Former First Minister of Northern Ireland	Northern Ireland
Michelle O'Neill	44	F	Deputy First Minister of Northern Ireland	Northern Ireland
Helen McEntee	35	F	Member of the Dáil Eireann	ROI
Mary Lou McDonald	52	F	President of political party Sinn Féin	ROI
Mairead McGuinness	62	F	European Commissioner for Financial Stability, Financial Services and the Capital Markets Union	ROI

Table 1 Dataset description of politicians (M: male, F: Female, ROI: Republic of Ireland)

Given the "fundamental differences between posed and spontaneous stimuli in terms of their appearance and timing", facial expression emotion recognition software systems should be analysed on both spontaneous and acted/posed facial expressions in order to achieve more robust validation [46]. Software systems validated solely on posed expressions will be less reliable when analysing spontaneous facial

expressions [46]. Emotient FACET, for example, outperforms Affectiva AFFDEX on posed facial expressions whereas Affectiva AFFDEX outperforms Emotient FACET on spontaneous facial expressions because it is explicitly trained on these types of more naturalistic expressions [38]. Politicians were chosen as the dataset for this research, therefore, because they are semi-trained actors who have undergone media training and therefore, tend to regulate their emotions in public. Hence, they bridge the gap between acting and spontaneity in facial expression as well as in speech, given some video clips are from more casual settings such as interviews and others from more rehearsed setting such as speeches.

3.2. Pre-Processing

3.2.1. Video Pre-Processing

Each MP4 video file is trimmed and cropped prior to any emotion recognition processing. The videos are trimmed and cropped to ensure only one person is in the frame at all times and only one person is speaking at all times, that person being the politician who is the subject of the video. This cuts any audience members in video clips taken from political speeches as well as any interviewers in video clips taken from interviews with the politician from the video and ensures that neither their facial expression nor their speech will end up being processed for emotion in later stages of the pipeline, but solely the politicians' being analysed. This was done using the video editing software quickTime Player, available for Apple macOS.

Emotient FACET and Affective AFFDEX, two facial expression software systems used to analyse emotion, both accept MP4 video files as input and therefore, the edited videos can be input to these software systems directly. However, the Microsoft Azure Face API accepts only static images and therefore, the MP4 video files must be converted to a set of images prior to using this software. Because both Emotient FACET and Affectiva AFFDEX process videos on a frame-by-frame basis, all frames from each video are extracted and saved as JPG image files so they are compatible with the Microsoft Azure Face API as well as aligned with the rate of processing of both Emotient FACET and Affectiva AFFDEX. The frames are extracted using the OpenCV Python library for computer vision.

3.2.2. Audio Pre-Processing

An online MP4 to WAV conversion tool was used to convert the videos to audio file format and the resulting WAV files downloaded. These audio files are subsequently chunked prior to speech processing as audio processing software systems best recognise emotion on smaller segments of audio. This was deduced from findings in previous literature discussed in Section 2.5.4 (Data Fusion) above. This is also done to provide a foundation for decision-level data fusion, as discussed earlier in Section 2.5.4 (Data Fusion), with speech being identified as the best modality for data segmentation on.

Audio files are chunked using the Python pydub module, specifying the desired length of each chunk. 2000ms segments is deemed an appropriate length given findings on detecting emotion in shorter versus longer audio segments, and therefore 2000ms chunks are extracted for input to both OpenSMILE and OpenVokaturi software systems. However, DeepAffects requires audio chunks of minimum 3000ms in duration in order to detect emotion and therefore, 3000ms chunks are extracted for input into the DeepAffects software system. These audio chunks are input directly into each speech processing software in order to extract emotion per chunk of an audio file.

3.3. Transcription

Each converted WAV audio file is next input into various ASR software systems used to derive the text transcript for each audio, which will become the basis of text sentiment extraction later in the pipeline. The three ASR services utilized to transcribe the audio files are Amazon Transcribe [121], Microsoft Azure Speech-to-Text [122] and Google Cloud Speech-to-Text [123]. These were chosen as they were all found to be strong performing ASR services that achieved low WER rates in previous studies, as discussed in Section 2.4 (Transcription) above.

3.3.1. Amazon Transcribe

Amazon Transcribe is an ASR service provided by Amazon Web Services' (AWS) Cloud Computing Services that is available for use at no cost with a free-tier AWS account. According to its website, it uses "advanced machine learning technologies" [124] to recognize speech from an audio file and transcribe it to text. It allows for automatic punctuation and capitalisation within transcripts and the derivation of timestamps for each individual word, as well as support for 31 different languages, including Irish-English 'en-IE'. Amazon Transcribe accepts audio files that are of WAV, FLAC, MP4, MP3, OGG, WebM, or AMR file format. It also requires that audio files be 4 hours in length maximum and less than 2GB in size.

Transcription using Amazon Transcribe was conducted directly within the AWS portal, with an audio file being uploaded to an Amazon S3 cloud storage bucket and being loaded into a 'Transcription job' on the portal, with results being available to download from the portal in the form of a json file, containing the resulting transcript itself, as well as start and end times for each individual word. When creating the 'Transcription job', it is important to note that 'en-IE' (Irish English) was chosen as the language for the input audio file.

3.3.2. Microsoft Azure Speech to Text

Microsoft Azure STT [122] is part of Azure's Cognitive Services and is accessible via an API at no cost once the user holds a free or standard tier Microsoft Azure subscription. It uses "the same robust technology that powers speech recognition across Microsoft products" [122] and provides support for over 85 languages and variants, including 'en-IE'. It can also generate timestamps for each individual word as well as perform automatic formatting and punctuation. Microsoft Azure STT accepts audio files that are in 16kHz, 16-bit, mono PCM WAV file format and audio files can be transcribed directly from local storage or via a URL, there is no requirement to upload the file to cloud storage before transcribing.

'Continuous recognition', was utilized for transcription using Microsoft Azure STT as opposed to 'single-shot recognition'. 'Single-shot recognition' only recognises a single utterance at a time, putting a stop to the transcription once a silence is heard or once 15 seconds of audio is reached. With 'continuous recognition', multiple utterances can be transcribed and hence, an audio file can be transcribed start-to-finish. Code utilized to perform continuous recognition speech transcription is available on the Microsoft website [125]. An audio file is passed to the Speech SDK, along with the target audio language, being 'en-IE', as well as the Azure subscription details and a request for the timestamps of each individual word, with the resulting transcript and the requested timestamps being returned. These results were then written to both a text file containing only the transcript, and a csv file containing both the transcript and the individual word timestamps.

3.3.3. Google Cloud Speech to Text

Google Cloud Speech to Text uses Google's deep-learning neural network algorithms to transcribe speech to text [123]. It is accessible via an API on the free tier of Google Cloud once you have a Google account. It supports over 125 languages and variants, including 'en-IE'. Timestamps for each individual word can also be generated and automatic punctuation be performed upon request. It works best with FLAC file format or LINEAR16 encoded WAV file format audio files with a sampling rate of minimum 16,000Hz, with synchronous speech recognition being used for files with maximum length one minute, and asynchronous being used for files longer than one minute. If asynchronous speech recognition is being used, Google Cloud STT requires that the audio file be first uploaded to a Google Cloud Storage bucket and called for transcription from there which, similar to Amazon Transcribe's STT, adds further complexity to the operation. However, synchronous speech recognition allows for locally stored files of maximum one minute in length to be transcribed.

'Asynchronous speech recognition' is used for transcribing audio files longer than 1 minute, which those in this dataset are. The code for running asynchronous speech recognition is available on the Google Cloud website [126]. In order to run asynchronous speech recognition, the target audio file is first uploaded to a Google Cloud Storage bucket. Once the file has been uploaded, an audio transcription request is sent to the API containing the link to the audio file in Cloud Storage and the target language, being 'en-IE'. In addition, a request for the timestamps of each individual word to be returned as well as the transcript to be punctuated are sent in the API call. The resulting punctuated transcript and timestamps of each individual word are returned and subsequently stored in a txt and csv file, similarly to the output of the Microsoft Azure ASR as described above.

3.4. Facial Expression Processing

The three facial expression emotion recognition software systems implemented in this study are Emotient FACET, Affectiva AFFDEX, and the Microsoft Azure Face API. Facial expression processing is conducted on a frame-by-frame basis by Emotient FACET and Affectiva AFFDEX automatically, and by the Microsoft Azure Face API by sending it the static images of each frame in a video. The videos in this dataset have a median frame rate of 25 frames per second (fps) within a range of 24 to 50 fps, with only one video having a frame rate of 50 fps and 5 videos having a frame rate of 24 fps. This means that a video 3 minutes and 47 seconds in duration with a frame rate of 25 fps would have a total of 5675 frames. Each of these 5675 frames are processed individually by each of the three facial processing software systems and an emotion result returned for each, a total of 5675 emotion probabilities calculated and returned per emotion for this one video.

3.4.1. Emotient FACET

Emotient FACET is a "computer-based video classification algorithm" [45] developed by Emotient, available via the iMotions software suite. A license is needed to avail of iMotions and the individual Emotient FACET and Affectiva AFFDEX software systems. However, a license is granted to Trinity College Dublin therefore, enabling me to avail of these software systems in this study.

Emotient FACET is based upon the CERT (Computer Expression Recognition Toolbox) facial expression recognition software, which achieved classification accuracy of 90.1% on the extended Cohn-Kanade (CK+) database of posed expressions, and only 80% on a database of spontaneous facial expressions.

The CK+ database is built of over 2000 images from over 200 individuals [127], 81% Euro-American, 13% Afro-American, 6% other groups, and 61% female, 39% male [128].

Emotient FACET analyses facial expressions for the presence of seven emotions, namely joy, anger, surprise, fear, disgust, sadness, contempt, as well as neutral, returning an 'evidence score' for each [29]. These evidence scores are the "odds ratios in decimal logarithmic scale" of an emotion being evident in each frame of the video, with a positive-valued score meaning there is a greater than 50% chance of the emotion being present, a negative-valued score meaning there is a less than 50% chance, and a score of zero corresponding to a 50% chance of that emotion being present [29]. These scores can be translated to probabilities using the formula seen in Figure 21 [129]. A log-odds evidence score of 0.5, for example, would translate to a probability of 76% of an emotion being present in a frame [129].

$$P = 1/(1 + 10^{-\text{evidence score}})$$

Figure 21 Emotient FACET evidence score to probability formula [129]

Emotient FACET also produces evidence scores for 20 of the FACs set of identified AUs, being AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU18, AU20, AU23, AU24, AU25, AU26, AU28, and AU43, with their the corresponding facial movement and muscles described in Table 2. These evidence scores, again, correspond to the log-odds of the AU being present and can be translated to probabilities using the formula in Figure 21 also.

Action Unit	Facial Movement	Facial Muscle
1	Inner Brow Raiser	Frontalis, pars medialis
2	Outer Brow Raiser	Frontalis, pars lateralis
4	Brow Lowerer	Depressor Glabellae, Depressor Supercilli, Currugator
5	Upper Lid Raiser	Levator palpebrae superioris
6	Cheek Raiser	Orbicularis oculi, pars orbitalis
7	Lid Tightener	Orbicularis oculi, pars palpebralis
9	Nose Wrinkler	Levator labii superioris alaquae nasi
10	Upper Lip Raiser	Levator Labii Superioris, Caput infraorbitalis
12	Lip Corner Puller	Zygomatic Major
14	Dimpler	Buccinator
15	Lip Corner Depressor	Depressor anguli oris (Triangularis)
17	Chin Raiser	Mentalis
18	Lip Puckerer	Incisivii labii superioris and Incisivii labii inferioris
20	Lip stretcher	Risorius
23	Lip Tightener	Orbicularis oris
24	Lip Pressor	Orbicularis oris
25	Lips part	Depressor Labii, Relaxation of Mentalis (AU17), Orbicularis Oris
26	Jaw Drop	Masetter; Temporal and Internal Pterygoid relaxed
28	Lip Suck	Orbicularis oris
43	Eyes Closed	Relaxation of Levator Palpebrae Superioris

Table 2 Emotient FACET Action Units [3]

Both Emotient FACET and Affectiva AFFDEX use machine learning to classify emotions however, the exact statistical processes, facial landmarks, and databases used to train each of the ML algorithms, being FACET or AFFDEX, differs [45]. Emotient FACET identifies only 6 facial landmarks [45] and uses these to classify AUs that are subsequently encoded in a linear combination to determine an evidence score for each emotion. It can identify both micro and macro facial expressions, with micro expressions relating to those that

are 250 milliseconds (ms) or less, and macro expressions relating to those than can last from 500ms to 4000ms in duration [7].

To process the dataset using Emotient FACET, a study is set up in iMotions, setting 'RespondentCamera' as the sensor and specifying FACET as the algorithm for recognition. Next, 'Face Recording' is set as the stimuli and all respondents are added, being each video in the dataset, with the age and gender of the subject in the video being manually entered. The results are saved to a text file which can be imported into excel, with each row corresponding to a frame in the video and individual columns containing information such as the evidence scores for each of the emotions as well as each of the 20 AUs identified. The emotion evidence scores for the six emotions joy, anger, surprise, fear, contempt, disgust, and sadness for the first four frames of a video can be seen in Figure 22 below. These values correspond to the raw log-odds evidence scores, prior to being transformed to probabilities. Similar evidence scores for an excerpt of the AUs identified by Emotient FACET are seen in Figure 23 also.

Joy Evidence	Joy Intensity	Anger Evidence	Anger Intensity	Surprise Evidence	Surprise Intensity	Fear Evidence	Fear Intensity	Contempt Evidence	Contempt Intensity	Disgust Evidence	Disgust Intensity	Sadness Evidence	Sadness Intensity
-1.698435	0	0.2239795	0	-5.592932	0	-1.261584	0	-0.234102	0	-1.778603	0	-1.967477	0
-1.668684	0	0.04248488	0	-5.578319	0	-1.228765	0	-0.1729575	0	-1.767065	0	-1.893119	0
-1.683201	0	-0.1176701	0	-5.199615	0	-1.118216	0	-0.1402914	0	-1.511493	0	-1.806822	0
-1.66669	0	-0.4037824	0	-4.740931	0	-1.151072	0	-0.2659293	0	-1.180972	0	-1.809826	0

Figure 22 Emotient FACET emotion evidence scores

AU1 Evidence	AU2 Evidence	AU4 Evidence	AU5 Evidence	AU6 Evidence	AU7 Evidence	AU9 Evidence	AU10 Evidence	AU12 Evidence	AU14 Evidence	AU15 Evidence	AU17 Evidence	AU18 Evidence	AU20 Evidence	AU23 Evidence	AU24 Evidence	AU25 Evidence
-0.4130338	-0.5106336	-0.07063632	-0.7822819	0.0696516	-0.6039522	0.3277702	0.9180765	0.1670097	0.3620754	-0.7168502	0.06820106	-0.5213132	0.6321314	-0.9422564	0.02739765	-1.226918
-0.5491666	-0.5983337	-0.1222771	-0.6221043	-0.1191338	-0.724753	0.009526199	0.6548468	0.07521862	0.2806742	-0.5452267	-0.0124782	-0.7239716	0.646977	-1.038179	-0.08418383	-1.31238
-0.4379325	-0.4964582	0.01547149	-0.7760792	0.184127	-0.3370479	0.2605868	1.021711	0.1154746	0.06196779	-0.3581344	0.08896486	-0.8133584	0.7644244	-0.7705533	-0.1408294	-1.045606
-0.3475821	-0.1992016	-0.05824665	-0.7113031	-0.1247273	-0.1121369	-0.3208245	1.024762	-0.2148657	-0.868692	-0.9674948	-0.7461312	-1.555484	0.9347766	-1.081394	-1.106086	0.4901809

Figure 23 Emotient FACET AU evidence scores

3.4.2. Affectiva AFFDEX

Affectiva AFFDEX is an emotion classification algorithm developed by Affectiva Inc, also available via the iMotions software suite. Similar to Emotient FACET, it classifies seven emotions – anger, sadness, fear, disgust, joy, contempt, and surprise – from facial expressions based on FACS however, Affectiva AFFDEX is trained on a database of over 1.8 million spontaneous facial expressions of people from across the globe [38] and therefore, outperforms Emotient FACET on more naturalistic and spontaneous expressions than acted/posed ones.

Affectiva AFFDEX also detects 34 facial landmarks as opposed to Emotient FACET's 6 facial landmarks [45] to classify 19 AUs, with all AUs outlined in Table 2 except for AU23 (Lip Tightener) being identified. A value between 0 and 100 is returned for each AU identified as well as each subsequent emotion identified, with each frame in the video returning a value for each AU and each emotion as well as the xy pixel coordinates for each of the 34 facial landmarks. Affectiva AFFDEX extracts Histogram of Oriented Gradient (HOG) facial texture features from the 34 facial landmarks identified, using these to classify AUs using a Support Vector Machine (SVM) classifier – trained on the 1.8 million spontaneous facial expressions database – and subsequently determine emotions by coding a linear combination of the classified AUs [28]. This process can be seen in Figure 24 below and a screenshot of a real-time classification demo of the SDK seen in Figure 25, with the white dots corresponding to the 34 facial landmarks detected, the AU classifications annotated on the right hand side, and the emotion classifications on the left.

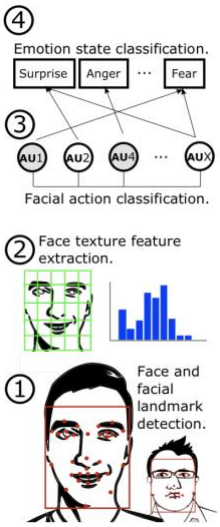

Figure 24 Affectiva AFFDEX emotion recognition process

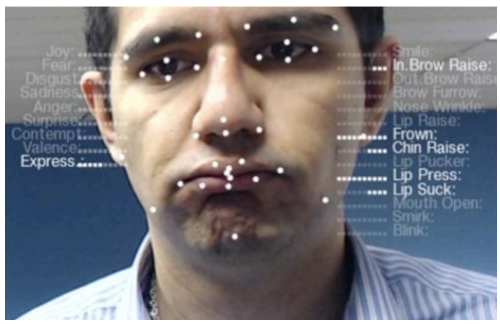

Figure 25 Affectiva AFFDEX SDK real-time demo

The same process is followed for Affectiva AFFDEX as Emotient FACET in the iMotions software study, with the AFFDEX algorithm being specified for the 'RespondentCamera' sensor instead. The results are also output in a text file that can be imported to excel, with each row corresponding to a single frame of the video input, and columns corresponding to AU and emotion probability values, as well as facial landmark coordinates. Probabilities output for the seven emotion classes for the first four frames of a video are seen in Figure 26, with probabilities for a subset of the AUs identified by Affectiva AFFDEX seen in Figure 27.

Anger	Sadness	Disgust	Joy	Surprise	Fear	Contempt
0.7107518	3.789378	0.4397394	2.15E-05	0.003292453	0.04370039	0.6844695
0.002656163	0.02758813	0.3747394	0.002047955	0.1563174	0.004457718	0.1759368
0.001536501	0.01676082	0.2837258	0.004941317	0.185203	0.003054723	0.1214097
0.001732148	0.01857617	0.3237732	0.003671323	0.1904089	0.003439237	0.1398528

Figure 26 excerpt of Affectiva AFFDEX emotion probabilities

Cheek Raise	Dimpler	Eye Widen	Lid Tighten	Lip Stretch	Jaw Drop
0.01498984	0.02404761	3.54E-06	4.476543	0.002206555	0.06752319
0.2903548	0.04217477	4.81E-06	2.668587	0.002206096	0.05203278
0.6616718	0.03252256	3.28E-06	1.998197	0.002007927	0.06352303
0.7110931	0.01580414	2.34E-06	2.219535	0.001374672	0.1428409

Figure 27 excerpt of Affectiva AFFDEX AU probabilities

3.4.3. Microsoft Azure Face API

The Microsoft Azure Face API is part of Microsoft Azure's Cognitive Services. It provides a set of AI algorithms that "detect, recognize, and analyse human faces in images" [130]. These algorithms have been trained by Microsoft however, the details of what database was used and whether subjects were expressing posed or spontaneous facial expressions is not publicly available. Similar to the Microsoft Azure's STT API, the Face API is available free of cost to Microsoft Azure free or standard tier subscribers. It conducts emotion recognition on a frame-by-frame basis in alignment with Emotient FACET and Affectiva AFFDEX processing however, this must be done by extracting the frames of each video prior to processing and inputting them as static images to the API. This is because the Face API only accepts image file formats JPEG, PNG, GIF for the first frame, and BMP [131]. Additionally, the API requires these image files to be of no more than 6 MB in size. Code to call this API using Python is provided by Microsoft on their website [132].

The API first detects a face in the image and assigns it a unique ID. It then identifies a set of 27 facial landmarks on the detected face, detailed and visualised in Figure 28 below. The xy pixel coordinates of each of these facial landmarks is returned by the API. The API can also be used to extract a number of facial attributes, including 'accessories', 'age', 'blur', 'emotion', 'exposure', 'facial hair', 'gender', 'glasses', 'hair', and 'head

pose'. The emotions extracted are anger, disgust, fear, happiness, sadness, and surprise, along with neutral, with the API returning a confidence score between 0 and 1 for each emotion per face detected. The confidence scores are normalised, and the sum of all scores is 1. [131]

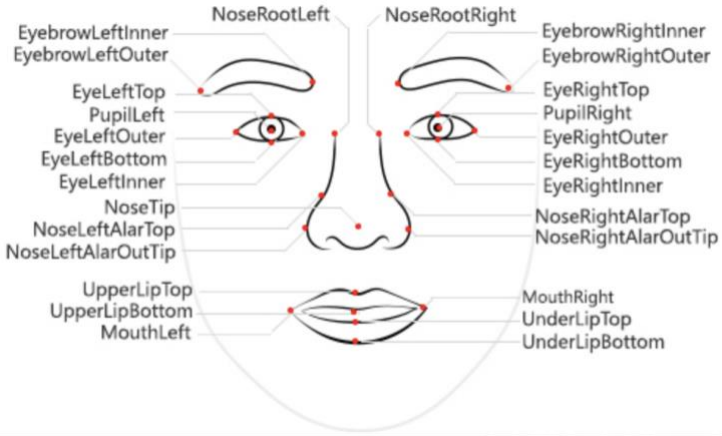

Figure 28 Microsoft Azure Face API facial landmarks [131]

To analyse emotion in a video file, each frame's JPG image file is sent with Azure subscription details and the requested facial landmarks and attributes in the API call. In this case, the API returns the pixel coordinates of all 27 facial landmarks along with the confidence scores for each emotion for each JPG frame passed to the API call. These results are written to a csv file, with one file per video, each row corresponding to each individual frame extracted, and the columns containing the facial landmark and emotion scores information. The emotion results for the first four frames of a video are seen in Figure 29.

happiness	surprise	fear	contempt	disgust	sadness	neutral	anger
0.013	0	0	0	0.001	0	0.001	0.985
0	0.01	0	0	0.001	0.001	0.003	0.984
0.006	0.011	0	0	0	0.001	0.002	0.98
0.001	0.189	0.001	0	0.002	0.012	0.002	0.793

Figure 29 Microsoft Azure Face API emotion output

3.5. Speech Processing

The three SER (Speech Emotion Recognition) software systems implemented in this study are OpenSMILE, OpenVokaturi, and DeepAffects. The WAV audio file converted from the MP4 video file provides the basis of speech emotion recognition. However, prior to processing this audio, the file is chunked into either 2000ms or 3000ms segments, depending on the speech processing software system being used, with 2000ms chunks being input to both OpenSMILE and OpenVokaturi, and 3000ms chunks being input to DeepAffects on the basis of reasons outlined in Section 3.2.2 (Audio Pre-Processing) above.

3.5.1. OpenSMILE

OpenSMILE is a "feature extraction and audio analysis tool" that allows you to "extract large audio features incrementally and fast and apply machine learning methods to classify and analyse your data in real-time" [73]. It is a free and open-source tool originally developed as part of the research project SEMAINE at the Technical University of Munich, whom still upkeep development of the tool, now owned by audEERING. The latest version is OpenSMILE 3.0, which can be downloaded from GitHub [133] or cloned and compiled from source [134]. When extracting features using OpenSMILE, a feature set is to be specified. There exist many feature sets for emotion recognition that can be used with OpenSMILE, each contained in a configuration file,

such as the INTERSPEECH 2009 Emotion Challenge feature set, the INTERSPEECH 2010 Paralinguistic Challenge feature set, the INTERSPEECH 2011 Speaker State Challenge feature set, the INTERSPEECH 2012 Speaker Trait Challenge feature set, the INTERSPEECH 2013 ComParE Challenge feature set, and the openSMILE/openEAR 'emobase' set [134].

OpenSMILE version 2.1 onwards supports the use of OpenEAR pre-trained models to carry out speech emotion classification [73]. OpenEAR emotion recognition models were trained on six databases, namely the Berlin Speech Emotion Database (EmoDB), the eNTERFACE database, the Audio Visual Interest Corpus, the Airplane Behaviour Corpus, the Belfast naturalistic corpus, and the Vera-am-Mitta corpus. The EmoDB and eNTERFACE databases are those used to train the models for emotion recognition of six basic emotional states – anger, boredom, fear, happiness, sadness, and neutral [72]. The EmoDB consists of 800 acted emotional utterances from 10 German speakers, 5 male and 5 female, expressing a range of emotions, namely happiness, anger, sadness, fear, disgust, boredom, and neutral [135]. The eNTERFACE database is made up of acted emotional speech from 42 individuals of 14 different nationalities [127], the distribution of which can be seen in Figure 30 below, 81% of which were men, and only 19% women, with emotions expressed being happiness, sadness, surprise, anger, disgust and fear.

Country	Number of subjects	Country	Number of subjects
Belgium	9	Cuba	1
Turkey	7	Slovakia	1
France	7	Brazil	1
Spain	6	U.S.A.	1
Greece	4	Croatia	1
Italy	1	Canada	1
Austria	1	Russia	1

Figure 30 Nationalities of subjects in eNTERFACE speech database

There are two OpenEAR configuration files used for emotion recognition, one for longer audio files that executes energy-based segmentation and returns an emotion classification result for each segment, and a second for shorter audio files that returns a single emotion classification result [72]. The latter, *config/emobase_live4_batch.conf,* is used for audio files that have been pre-segmented into shorter audio chunks before processing, as has been done in this study and therefore, it is the feature set configuration file that was used. This feature set contains 988 acoustic features extracted from audio, which will be used for emotion recognition. It contains low-level descriptors (LLD) such as intensity, loudness, 12 MFCC, pitch (*F0*), probability of voicing, *F0* envelope, 8 LSF (Line Spectral Frequencies), and zero-Crossing Rate [134].

Speech emotion recognition using OpenSMILE is executed using a command-line executable that takes the audio file, being the shorter audio segments, and configuration file, being *config/emobase_live4_batch.conf*, as input, and outputs both an arff file containing a single feature vector for that audio segment as well as a text file that contains the emotion classification result for that audio segment as well. The resulting arff file feature vector contains the values for that audio segment for the LLDs described above, including the mean F0 of each chunk, and the text file classification result contains a value between 0 and 1 corresponding to the probability of that emotion being present in the given audio chunk. Each segmented chunk of an audio file is input to this executable and the results of interests written to a csv file containing the results for emotion probabilities and mean F0 of each chunk in the entire audio file. A snippet of the emotion and mean F0 results for the first four 2000ms chunks in an audio file is seen in Figure 31.

anger	boredom	disgust	fear	happiness	neutral	sadness	f0
0.016346	0.469347	0.044905	0.016514	0.013655	0.097661	0.341572	308.1082
0.016545	0.485775	0.039219	0.010103	0.009371	0.359218	0.079769	278.71
0.00961	0.487136	0.006076	0.00301	0.012449	0.46211	0.01961	360.9765
0.002269	0.056588	0.00275	0.002238	0.001659	0.005971	0.928525	454.8713

Figure 31 OpenSMILE emotion probabilities and F0 output

3.5.2. OpenVokaturi

Vokaturi is an open-source speech emotion recognition software that is language-independent and has been validated with existing emotion databases [136]. The OpenVokaturi SDK can be used to easily integrate speech emotion recognition into applications [136]. OpenVokaturi is one of three libraries provided by Vokaturi [9], with all three being outlined in Table 3. The accuracy scores outlined are those derived from validating Vokaturi with existing emotional speech databases. It can be seen that OpenVokaturi returns the lowest accuracy of the three libraries, achieving 66.5%, with VokaturiPlus and VokaturiPro both achieving 76.1% however, it is the only library that is freely available at no cost and therefore, it is the most widely-used and is also the library that is implemented in this study.

Name	*Accuracy*	*Pricing*
OpenVokaturi	66.5	Free
VokaturiPlus	76.1	Paid
VokaturiPro	76.1	Paid

Table 3 Vokaturi Libraries [9]

OpenVokaturi detects emotions happiness, sadness, anger, fear, and neutral from speech, and returns a probability of each emotion being present in an audio clip, with all probabilities summing to 1 [74]. It measures a set of 9 acoustic features from audio that relate to these five classes of emotion. Acoustic features include average pitch, pitch dynamics, average intensity, intensity dynamics, and spectral slope [136]. The means and standard deviations of each of these features are measured for each emotion and the results fed into a neural network with three linear connections that produces emotion probabilities using a softmax transformation. OpenVokaturi has been trained on two emotion-annotated speech databases – the EmoDB database, discussed in Section 3.5.1 (OpenSMILE) above, as well as the Surrey Audio-Visual Expressed Emotion (SAVEE) database [136]. The SAVEE emotional speech database consists of over 480 British-English acted emotional utterances by 4 male speakers, expressing anger, disgust, fear, happiness, sadness, surprise, and neutral [137].

Code to implement the OpenVokaturi SDK is provided on Vokaturi's developers website [136]. Each 2000ms chunk segmented from the original WAV audio file is passed to the API and emotion probabilities for each chunk returned, with all results being aggregated in a csv file that contains emotion probabilities per chunk for the entire audio file. An extract of the emotion probabilities for the first four 2000ms chunks of an audio file is seen in Figure 32.

Neutral	Happy	Sad	Angry	Fear
0.04	0.55	0.12	0.29	0
0.56	0.01	0.1	0.34	0
0.72	0	0.25	0.02	0
0	0.25	0	0.75	0

Figure 32 OpenVokaturi emotion probabilities output

3.5.3. DeepAffects

DeepAffects provides a set of APIs that provide support for audio, text & video recognition and analytics [138]. It is openly available for use with a free DeepAffects account and acquisition of an API key

from the developer's portal. Audio analysis offerings include separate APIs for speaker diarisation, speech-to-text, and emotion extraction, with the latter being implemented in this study. The emotions extracted by DeepAffects are anger, excitement, frustration, happiness, sadness and neutral. The training database used to classify emotions, or the exact classification process used are not disclosed by DeepAffects.

Calls to the audio emotion extraction API can be synchronous or asynchronous. Synchronous API calls are used when the audio file being analysed is less than 2 minutes in duration, such as the segmented audio chunks extracted earlier in this pipeline. Otherwise, for longer audio files asynchronous API calls are used, which require results to be sent to a webhook by specifying a webhook URL in the API call. There is a limit of 5 API calls per minute, and a total of 100 API calls a day [138].

The DeepAffects API python library can be installed from pip directly [139] and code to implement the audio emotion extraction API is available on the DeepAffects developer website, with code for both synchronous and asynchronous API calls provided [138]. Audio file formats supported by DeepAffects include WAV and MP3. The smallest audio chunk that can be effectively analysed for emotion is 3000ms in length, with anything shorter in duration than this returning emotion 'neutral' by default. Therefore, instead of segmenting the audio files into 2000ms chunks as was done for both OpenSMILE and OpenVokaturi speech processing, for DeepAffects, the WAV audio file is segmented into 3000ms chunks. Each of these chunks is sent in the API call and one single emotion returned for each chunk, being the emotion that is most dominant. The language code specified for analysis is 'en-IE', overwriting the default en-US language code. The results from each chunk are aggregated in a single csv file, with each row corresponding to an individual 3000ms chunk, and a column depicting the the most dominant emotion determined by DeepAffects for that chunk. Figure 33 shows the emotion output by DeepAffects for the first four 3000ms chunks, numbered starting at 0.

0	excited
1	excited
2	excited
3	neutral

Figure 33 DeepAffects emotion results

3.6. Text Processing

Following the transcription of a WAV audio file, text sentiment analysis is performed on the derived transcript. The derived transcripts in the dataset are short, and span approximately 500-700 words in length. As was outlined in Section 2.4 (Transcription) above, errors in the transcription do not have detrimental effects on subsequent text analysis and therefore, these derived transcripts can be used directly in text sentiment analysis without any further pre-processing or error-fixing. The three text sentiment analysis tools and lexicons implemented in this study are RockSteady, a text analytics tool that employs the General Inquirer lexicon, the NRC Emotion Lexicon, and the Opinion Lexicon, with each word in the transcript being analysed for positive or negative sentiment. The transcripts used in this text sentiment analysis are those derived using the Microsoft Azure STT service. This is because these are the transcripts that achieved the lowest mean WER rate overall of all three transcription services, as will be discussed in Section 4.1 (Transcription) below.

3.6.1. RockSteady

RockSteady, a sentiment analysis tool developed at Trinity College Dublin, uses the General Inquirer (GI) lexicon to carry out sentiment analysis. The GI lexicon contains 3,486 words, 1,915 positive and 2,291

negative [84]. RockSteady also provides an option of using another lexicon in addition to this general-purpose GI lexicon, one that is more domain-specific [140]. However, in this study solely the GI lexicon is used. The GI lexicon is built from four sources, namely the Harvard IV-4 dictionary, the Lasswell value dictionary, several categories recently constructed, and "marker" categories [141]. It is available for download at [141], with an excerpt of this lexicon seen in Figure 34. The 'Entry' column depicts each individual word, the 'Source' column denotes which of the four sources the word annotation came from, with the 'Positiv' and 'Negativ' columns being annotated if the source identifies that word as having positive or negative sentiment.

Entry	Source	Positiv	Negativ
A	H4Lvd		
ABANDON	H4Lvd		Negativ
ABANDONMENT	H4		Negativ
ABATE	H4Lvd		Negativ
ABATEMENT	Lvd		
ABDICATE	H4		Negativ
ABHOR	H4		Negativ
ABIDE	H4	Positiv	
ABILITY	H4Lvd	Positiv	
ABJECT	H4		Negativ
ABLE	H4Lvd	Positiv	
ABNORMAL	H4Lvd		Negativ

Figure 34 General Inquirer lexicon excerpt [141]

RockSteady can be downloaded as a desktop application and provides a GUI for sentiment analysis. The system accepts a folder corpus of multiple text files or a single LexisNexis text file as input, allowing the user to group text files from the folder corpus by minute, hour, day, week, month, year, source, custom, or leave them as individual text files. It also allows the user to specify the option to exclude text files that are shorter than a specified number of words in length. RockSteady outputs the total number of words in a text file, the number of positive words, negative words, active words, passive words, strong words, and weak words, along with a few more domain-specific outputs. It gives an option to express these results in terms of count of words, percentage of total words, or as a z-score.

RockSteady is implemented in this study by inputting a folder of all the derived transcripts in the database, leaving them as individual text files, obtaining results as a percentage of total words and saving those in a csv file. Sample output of RockSteady for the first five transcripts in the dataset is seen in Figure 35, with the values for Positiv, Negativ, Active, Passive, Strong, Weak, Econ@, POLIT, and MIlit corresponding to percentage of total number of words, which is entered under the column 'Terms'. The columns Econ@, POLIT, and Milit are domain-specific and are not used as part of sentiment analysis in this study. Each row in Figure 35 corresponds to one entire transcript.

Title ▲	...	...	Terms	URL	Positiv	Negativ	Active	Passive	Strong	Weak	Econ@	POLIT	Milit
1		1	600.00		3.83	2.50	8.17	2.50	10.17	2.00	3.33	4.67	0.00
2		1	492.00		4.47	3.05	6.10	4.88	9.55	2.03	0.20	2.64	0.20
3		1	490.00		4.29	1.02	4.90	5.31	6.12	2.24	0.20	1.22	0.00
4		1	492.00		4.27	4.88	7.32	6.91	9.76	4.88	0.41	1.22	0.00
5		1	486.00		7.00	3.50	8.02	4.12	7.41	3.09	1.23	1.03	0.00

Figure 35 Output of RockSteady sentiment analysis software for first five transcripts in dataset

3.6.2. NRC Emotion Lexicon

The NRC Emotion Lexicon, available at [142], contains approximately 14,182 words, each annotated with either a 0 or 1 for both positive or negative sentiment [10], 0 indicating the absence of positive or negative sentiment, and 1 indicating the presence of positive or negative sentiment. The lexicon was created and annotated through use of online crowdsourcing using Amazon's Mechanical Turk [143]. This lexicon also annotates each word with a 0 or 1 for a set of eight emotion classes – anger, anticipation, disgust, fear, joy,

sadness, surprise, trust. An excerpt of all ten annotations for a single word in the lexicon is seen in Figure 36 however, for the scope of this research, only positive and negative is extracted using the NRC Emotion Lexicon for sentiment analysis.

```
abandon  anger    0
abandon  anticipation     0
abandon  disgust  0
abandon  fear     1
abandon  joy      0
abandon  negative 1
abandon  positive 0
abandon  sadness  1
abandon  surprise 0
abandon  trust    0
```

Figure 36 NRC Emotion Lexicon for word 'abandon'; annotated 1 for fear, negative, and sadness [142]

The percentage of positive and negative words in a transcript based on this NRC Emotion Lexicon is derived using a Python script that loops through the NRC Emotion Lexicon, counting the number of positive and negative words that appear in the transcript and calculating these as a percentage of total words. Results are subsequently written to a csv file, with a percentage being output for both positive and negative sentiment for each transcript as a whole, an excerpt of which can be seen in Figure 37 below for the first five transcripts in the dataset.

Positive	Negative
0.0506	0.0168
0.0502	0.0162
0.0547	0.0317
0.0350	0.015
0.0485	0.0296

Figure 37 Output of NRC Emotion Lexicon sentiment analysis for first five transcripts in dataset

3.6.3. Opinion Lexicon

The Hu & Liu Opinion (or Sentiment) Lexicon, available at [144], contains 2,006 positive and 4,783 negative words, generated automatically using machine-learning methods on customer reviews [90]. Figure 38 shows an extract from the Opinion Lexicon, with the left column pertaining to positive words, and the right column to negative words.

```
a+            2-faced
abound        2-faces
abounds       abnormal
abundance     abolish
abundant      abominable
accessable    abominably
accessible    abominate
```

Figure 38 Hu & Liu Opinion Lexicon; (left) positive, (right) negative [144]

The process followed in Section 3.6.2 (NRC Emotion Lexicon) using a Python script to loop through the lexicon and calculate percentages of positive and negative words in a transcript is followed for the Opinion Lexicon sentiment analysis also, with results for the first five transcripts in the dataset shown In Figure 39 below.

Positive	Negative
0.0168	0.0318
0.0280	0.0147
0.0187	0.0129
0.0466	0.0250
0.0242	0.0269

Figure 39 Output of Opinion Lexicon sentiment analysis for first five transcripts in dataset

3.7. Statistical Processing & Hypothesis Testing

Hypothesis testing is mostly implemented using the R programming language for statistical processing. Hypothesis tests such as the Kruskal-Wallis Rank Sum test and the Wilcoxon Rank Sum test are implemented using R code. These are examples of non-parametric hypothesis tests and are used in place of parametric hypothesis tests if your data does not follow a normal distribution, which is an assumption of parametric tests. These tests rank data in order to assess the difference between two or more groups, determining a significant difference should the average rank of each group differ. However, should the two groups come from the same population with the same median, their average rank will be in and around the same value and thus, no significant difference exists between the groups. These tests are conducted on results derived from processing each modality. For tests conducted on emotion, emotions joy and anger are analysed, and for those on sentiment, positive and negative sentiment are analysed. Spearman's Rank-Order Correlation is another example of a non-parametric test, also implemented using R code, used to measure the association between two variables by ranking data also, used on Emotient FACET and Affectiva AFFDEX output to examine a correlation between emotions joy/anger and specific AUs. Statistical processing is carried out on both the ASR software systems' output as well as on the OpenSMILE mean F0 output, calculating WER scores for the ASR-derived transcriptions and z-scores for the mean F0 output along with summary statistics such as mean, minimum, maximum, and standard deviation.

3.7.1. Word Error Rate

One video per politician was transcribed by each ASR service in order to evaluate their accuracy on the current dataset. To evaluate each of their accuracies I derived a WER score on each transcript derived from all three ASR services using the jiwer Python package [145]. Figure 9 shows the formula used to calculate the WER rate of an ASR-derived transcript against a reference transcript. I manually created the reference transcripts for each audio transcription used in this evaluation myself. The ASR service that achieves the lowest WER rate is the best performing transcription service on this dataset. The WER rate of human parity is 5.1% and will be used as a baseline to compare the performance of the ASR services.

3.7.2. Kruskal-Wallis Rank Sum Test

The Kruskal-Wallis Rank Sum test is a non-parametric alternative to the one-way ANOVA parametric test used to ascertain if there exists a statistically significant difference between more than two groups. It is implemented using the *kruskal.test* R function and is used in this case to test if there exists a statistically significant difference between the detection of emotion or sentiment in the 3 various software implemented for each modality [146]. One example of the implementation of this test is to determine if there is a significant difference between joy detected in Emotient FACET, Affectiva AFFDEX, and Microsoft Azure Face API. The hypotheses for this test are:

- H_0: distribution of x is the same for all three systems/lexicons
- H_1: distribution of x is not the same for all three systems/lexicons

where x is joy/anger/positive sentiment/negative sentiment dependent on the software systems or lexicons being analysed. H_0 denotes the null hypothesis and H_1 the alternative hypothesis. The test outputs a test statistic known as a p-value, α, where if $\alpha \leq 0.05$ then there is a statistically significant difference between the three systems

[146]. If this is the case, the null hypothesis that the distribution of *x* is the same for all three systems is rejected in favour of the alternative hypothesis that the distribution of *x* is not the same for all three systems.

3.7.3. Wilcoxon Rank Sum Test

The Wilcoxon Rank Sum Test, also known as the Mann-Whitney U Test, is a non-parametric alternative to the unpaired two-samples t-test, used to ascertain if there is a statistically significant difference between two groups. The R function *wilcox.test* is used to implement this test [147] which is used in two ways in this analysis, being:

1. *to test if there is a statistically significant difference between emotion or sentiment detected in pairs of software systems or lexicons* e.g. if there a statistically significant difference between joy detected in Emotient FACET and Affectiva AFFDEX.
    - H_0: distribution of *x* is the same for both systems/lexicons
    - H_1: distribution of *x* is not the same for both systems/lexicons
2. *to test if there is a statistically significant difference between emotion or sentiment detected in males and females per software system or lexicon* e.g. if there a statistically significant difference between joy detected in male politicians and female politicians by Emotient FACET.
    - H_0: distribution of *x* is the same for both genders
    - H_1: distribution of *x* is not the same for both genders

where *x* is joy/anger/positive sentiment /negative sentiment dependent on the software system(s) being analysed. Similar to the Kruskal-Wallis Rank Sum test, a p-value test statistic, α, is output, where if $\alpha \leq 0.05$ there is a statistically significant difference and hence, the null hypothesis (H_0) is rejected in favour of the alternative hypothesis (H_1) that the distribution of *x* is not the same for both systems/genders.

3.7.4. Spearman's Rank-Order Correlation

Spearman's Rank-Order Correlation measures the strength and direction of a monotonic association between two ranked variables [148]. A monotonic relationship is 'less restrictive' than a linear relationship and refers to the increase or decrease in one variable as another increases [148]. The *cor.test* function with parameter '*spearman'* in R is used to derive the correlation coefficient, Rho, for Spearman's Rank-Order Correlation. The resulting Rho correlation coefficient value ranges from -1 to 1, with a negative value indicating that as one variable increases, the other decreases, and a positive value indicating that as one variable increases, the other also increases. The strength of association is depicted in Figure 40, with Rho values closer to -1 or 1 indicating a stronger positive or negative association between variables.

.00-.19 "very weak"
.20-.39 "weak"
.40-.59 "moderate"
.60-.79 "strong"
.80-1.0 "very strong"

Figure 40 Spearman's Rho Correlation Coefficient

Spearman's Rank-Order Correlation is used in this case to assess which AUs are 'most dominant' for emotions joy and anger. Strong, positive association is most desired, which indicates that as the probability of an AU increases, the probability of the associated emotion also increases. The criterion for acceptance as a 'most dominant' AU for an emotion in this study is a Rho value ≥ 0.4 returned for the association between the AU and the emotion on both Emotient FACET and Affectiva AFFDEX data, being the two software systems

that output probability values for both emotions and AUs. For example, previous literature found AU6 to be a primary AU associated with joy therefore, if the Rho values output by measuring the association between AU6 probabilities and joy probabilities from both Emotient FACET and Affectiva AFFDEX output are both ≥ 0.4, AU6 is moderately to very strongly associated with joy in both systems' output and hence, carries more weight in the linear combination of AUs used to detect the probability of joy in a video frame in both systems, making it a 'most dominant' AU. The criteria for acceptance or rejection of an AU as a 'most dominant' AU is seen in Figure 41 below, with *rho1* being the Rho coefficient of correlation attained using Emotient FACET data, and *rho2* being the Rho coefficient of correlation attained using Affectiva AFFDEX data.

$$Accept: \begin{cases} rho_1 \geq 0.4 \\ rho_2 \geq 0.4 \end{cases}$$

$$Reject: \begin{cases} rho_1 < 0.4 \\ rho_2 < 0.4 \end{cases}$$

$$Reject: \begin{cases} rho_1 \geq 0.4 \\ rho_2 < 0.4 \end{cases}$$

$$Reject: \begin{cases} rho_1 < 0.4 \\ rho_2 \geq 0.4 \end{cases}$$

Figure 41 Acceptance/Rejection criteria of a 'most dominant' AU for a given emotion

3.7.5. Z-score

A z-score is a statistical value that relates how far from the mean a specific value is. The formula for calculating a z-score is seen in Figure 42 below, with *x* being the specific value under analysis, μ being the overall mean of the group that the value comes from, and σ being the standard deviation of that group. The z-score tells you how many standard deviations above or below the mean the value in question is. A value of 0 would mean the value is the same as the mean, +1 or -1 mean the value is one standard deviation above or below the mean, respectively, and +2 or -2 mean two standard deviations above or below the mean, and so on.

$$z = \frac{x - \mu}{\sigma}$$

Figure 42 z-score formula

Z-scores are calculated in this case as a means of examining mean F0 and its variation with respect to emotion, gender, and age. For example, the overall mean F0 of the entire dataset will be calculated by getting the average of all mean F0 values extracted for each audio chunk across all 50 audios in the dataset, 4754 audio chunks in total, output by OpenSMILE. Then, the same will be calculated for the audio chunks pertaining solely to the male politicians and solely to the female politicians. These will each be input to the formula respectively, along with the overall mean and standard deviation to get a z-score for each gender. The z-scores can be interpreted to determine if gender variations occur i.e. if male or female mean F0 lie below or above the overall mean, etc. Similar experiments are carried out for the mean F0 of joy and anger as well as the mean F0 for each politician within each gender to examine the male and female F0 variations that occur with age.

3.8. Synthesis & Data Fusion

3.8.1. Data Fusion

A two-step decision-level data fusion process is implemented to illustrate a method of determining if an emotion is strongly evident for a 2000ms video segment on the basis of the results of all three individual modalities, being facial expression emotion, speech emotion, and text sentiment. The alignment of results from

each modality is conducted on this 2000ms basis because literature identified speech as the best modality to base data segmentation for decision-level data fusion upon [105]. Decision-level data fusion is implemented because each modality has been processed asynchronously, with separate emotion and sentiment results output from each which will be subsequently fused at the decision-level. To showcase how this data fusion process is conducted, solely the facial expression emotion, speech emotion, and text sentiment results obtained from Emotient FACET, OpenSMILE, and RockSteady for joy and positive sentiment on one video from the dataset, being a video of Irish male politician Leo Varadkar, are used. Joy is said to be solely associated with positive sentiment [79] and therefore, negative sentiment is not integrated into this data fusion equation for illustration purposes however, it is considered in Section 5.2 (Future Work). Each step in this data fusion process builds on work from previous literature, outlined in Section 2.5.4 (Data Fusion) earlier. The full data fusion process can be seen in Figure 43 below, including the equations implemented.

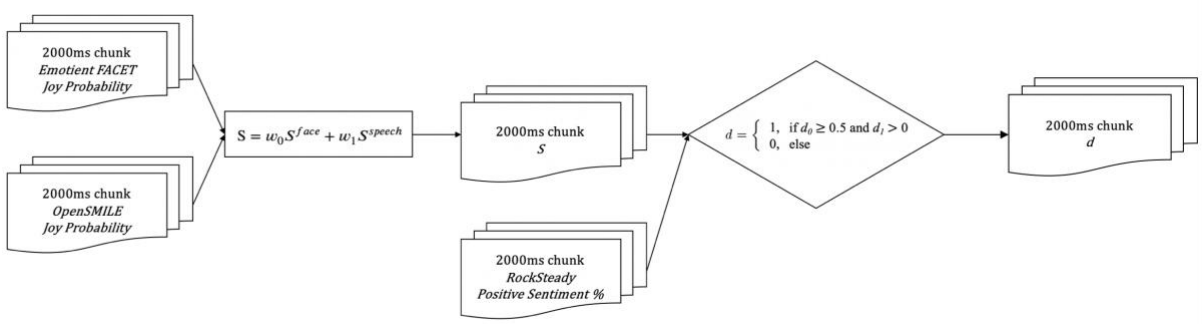

Figure 43 Two-step data fusion process

The first step fuses the facial expression and speech emotion probabilities together using the weighted sum of probabilities equation proposed by Wang et al. [22], seen in Figure 15 above. The weights used by Wang et al. are maintained here, being 0.8 for w_0 on facial expression probabilities, and 0.2 for w_1 on speech probabilities. The equation is applied to each 2000ms chunk of the video, returning a single fused probability value for each chunk. OpenSMILE outputs a single probability value for joy for each 2000ms chunk of the video however, Emotient FACET processes emotion on a frame-by-frame basis and therefore, multiple probability values for joy are output by Emotient FACET within a 2000ms chunk. For example, a video with a 30 frames per second (fps) frame-rate would have 60 frames and therefore, 60 probability values output for joy for each 2000ms chunk. Therefore, the mean of these probabilities is calculated in order to produce a single probability value for joy across all 60 frames in a single 2000ms that is appropriate aligned with speech and therefore, able to be input into the equation.

The result of this equation is subsequently fed into an equation derived from the logical "OR" formula proposed in [113] by Lee et al. The original equation is seen in Figure 16 above, and the adjusted equation below in Figure 44. In the adjusted equation, the combined decision rule is based on both the result of the first equation being greater than 0.5 i.e. joy is most dominant, and the percentage of positive sentiment being greater than 0. Positive sentiment enhances the determination of joy as a strongly evident emotion, hence why the criterion in the combined decision rule is > 0% positive sentiment. d_0 denotes the result from the first equation in this two-step process, and d_1 denotes the percentage of positive sentiment derived using RockSteady. Like the first equation, this logical "OR" formula is carried out on a chunk-by-chunk basis, with both values based on 2000ms segments. 2000ms-chunked transcript excerpts are each input to RockSteady in order to derive the percentage of positive sentiment per 2000ms chunk of the transcript, so that this modality is aligned with the

fused probability value derived above. An overall decision is made and returned for each 2000ms chunk in a video based on this logical "OR" and combined decision rule which returns a 1 or a 0, 1 if joy is determined to be strongly evident, and 0 if it's not.

$$d = \begin{cases} 1, & \text{if } d_0 \geq 0.5 \text{ and } d_1 > 0 \\ 0, & \text{else} \end{cases}$$

Figure 44 Adjusted Lee et al. equation

Where this logical "OR" formula returns '1' on a 2000ms chunk, the physical correlates of joy for that chunk are examined. The mean probability of the AU's determined in Section 4.5 (Facial Action Units) as 'most dominant' for joy will be computed for each chunk, as well as the mean F0 extracted from OpenSMILE for that chunk also. These values will be examined to determine if there is a link between the AUs 'most dominant' to joy and the mean F0 where joy is determined a strongly evident emotion.

3.8.2. AU & F0 synthesis

To further examine if there is a link between the physical correlates of facial expression, AUs, and one of the more insightful physical correlates of speech, mean F0, the average of all mean F0 values across each chunk of a video where the probability of the 'most dominant' AUs for joy and anger are > 50%, respectively, are computed. This will examine if where there is a high probability of the 'most dominant' AUs for jor or anger, do the mean F0 values corresponds to what is characteristically expected for that emotion or not, being a high mean F0 for both joy and anger. This is implemented on only one video from the dataset, being the one used to illustrate the data fusion process outlined above, with AU probabilities output by Emotient FACET and mean F0 values output by OpenSMILE being used. The 'most dominant' AUs used in this analysis will be determined by statistical analysis described in Section 3.7.4 (Spearman's Rank-Order Correlation) above. In addition, the mean F0 for a prototypical AU of sadness, identified from literature, will be used as a baseline for comparison to determine how 'high' the mean F0s for the joy and anger instances are. A prototypical AU of sadness is AU15 (lip corner depressor) [32] which will be used to compute a mean F0 value for where AU15 is > 50% probability across the video in order to derive this baseline comparison.

3.9. Summary

This chapter described each step of the modality pipeline for multimodal emotion recognition in detail, including what tools and software systems were implemented, and how each of these derive their results as well as what databases and techniques were used to train or build them. The dataset, which consists of videos of 10 Irish politicians, both male and female, is detailed, including the nationality of each politician, their gender, and their age. Five videos per politician are included in this dataset and are the basis of analysis for this study. Figure 45 shows a low-level architecture of the specification of a modality pipeline for multimodal emotion recognition output as part of this work. It summarises this chapter quite well, depicting what software systems and lexicons are used at each stage of the pipeline. The ASR services used to transcribe the WAV file to text are Microsoft Azure STT, Google Cloud STT, and Amazon Transcribe. Emotient FACET, Affectiva AFFDEX, and the Microsoft Azure Face API are used to extract emotion probabilities from video, with an MP4 video file being sent directly to the former two software systems and extracted video frames to the latter. The speech emotion recognition systems used are OpenSMILE, OpenVokaturi, and DeepAffects, with 2000ms audio chunks being input to the former two systems and 3000ms chunks to the latter. Rocksteady, which uses the

General Inquirer lexicon, the NRC Emotion Lexicon, and the Opinion Lexicon are used for text sentiment analysis. Emotion probabilities are output for facial expressions and speech, with sentiment percentages output for text. These results are used to determine statistically significant differences between the detection of joy, anger, positive sentiment or negative sentiment in each triad of systems/lexicons, each pair of systems/lexicons, and each gender per system/lexicon. Emotient FACET and Affectiva AFFDEX also output AU probabilities, and OpenSMILE mean F0 values. These are also used to determine the 'most dominant' AUs and the range of mean F0 for joy and anger, as well as variances in mean F0 across gender and age. Finally, one video from the dataset is input to a novelty two-step data fusion process and a link between AUs and mean F0 also examined.

All of the code used to implement each of these emotion recognition software systems and lexicon-based text sentiment analyses, as well as all pre-processing steps and post-processing statistical analyses and data fusion steps were conducted on a Macbook Pro laptop with a 2.3 GHz Dual-Core Intel Core i5 processor running macOS Big Sur, with the exception of the Emotient FACET and Affectiva AFFDEX systems, which were run on the Windows PC inside Trinity College Dublin that has a licensed version of the iMotions software suite installed. The duration of time it took to run code to implement each step in the pipeline ranged from a matter of seconds to extract audio chunks, for example, to over 30 minutes to process one set of video frames pertaining to a single dataset video file using the Microsoft Azure Face API. Over 4GB of disk storage was needed to store all the dataset files, results files, and code files created throughout the implementation of this research.

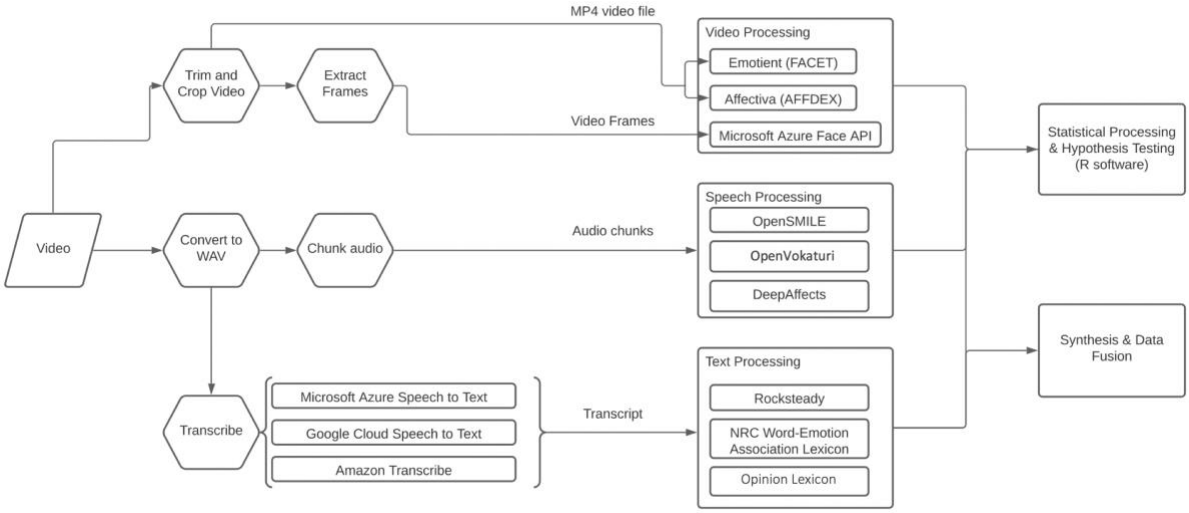

Figure 45 Low-level system architecture

Chapter 4

Case Studies and Results

4.1. Transcription

To evaluate the accuracy of the three ASR software systems on this specific dataset, one audio per politician was transcribed by all three ASR services, namely Microsoft Azure STT, Google Cloud STT, and Amazon Transcribe, with the WER scores achieved on each transcription presented in Table 4 below. The WER scores derived were computed by comparing each ASR-derived transcript to a manually-transcribed reference transcript using the Python jiwer package.

	MS Azure	Google	Amazon
Arlene Foster	3.21%	20.64%	9.65%
Helen McEntree	3.53%	29.05%	13.12%
Mary Lou McDonald	6.65%	21.85%	19.23%
Mairead McGuinness	9.04%	25.65%	14.14%
Michelle O'Neill	13.04%	40.44%	40.71%
Leo Varadkar	8.81%	45.25%	21.01%
Michael Martin	5.95%	31.76%	14.98%
Simon Coveney	7.95%	24.54%	19.09%
Stephen Donnelly	7.2%	22.45%	16.31%
Paschal Donohoe	6.81%	26.74%	15.50%
MEAN	7.21%	28.83%	18.37%
STD	2.82%	8.18%	8.51%

Table 4 Transcription WER results

Figure 46 shows a snippet of the Arlene Foster audio chosen to be transcribed, with the transcript on top being the manual reference transcript, the bottom-left transcript being that transcribed by Microsoft Azure STT, the bottom-center by Google Cloud STT, and bottom-right by Amazon Transcribe. The red text denotes a deletion or substitution error made by the ASR service i.e. it transcribed a word incorrectly or included a word not in the reference transcript, and the green text denotes an addition error i.e. a word that was left out in the ASR-derived transcription.

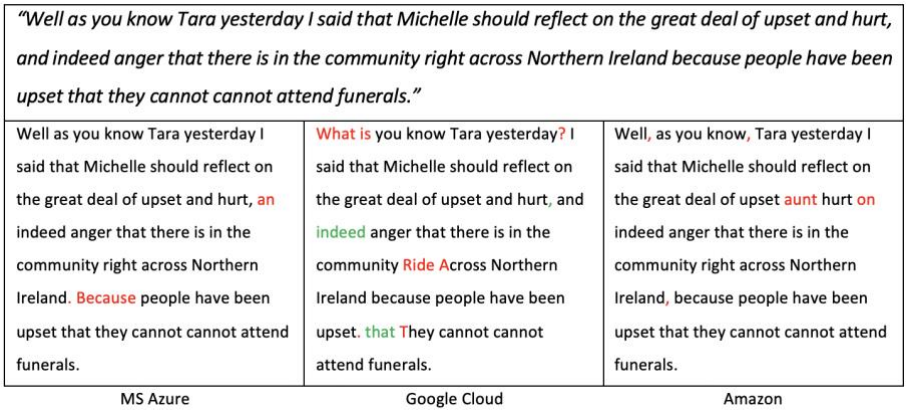

Figure 46 Transcription error (Red: deletion/substitution, Green: addition)

It can be seen from Table 4 that the Microsoft Azure STT transcription service performed best on this subset of the dataset, being 10 of the 50 total transcribed audios, one per politician. The Microsoft Azure STT service achieved a mean WER of 7.21%, while Google Cloud STT and Amazon Transcribe achieved a mean WER of 28.33% and 18.37%, respectively. Therefore, Microsoft Azure STT achieved the lowest mean WER score and that closest to the WER of human parity, being 5.1%, the baseline for comparison. In addition, Microsoft Azure STT has the smallest WER standard deviation compared to Google and Amazon's transcription services, meaning the WER values for each individual transcript are more clustered around the mean for Microsoft Azure STT than the other two ASR services, further supporting the better performance of this transcription service over Google's and Amazon's.

Each transcription service provides support for 'en-IE' however, the exact training databases and what nationality the speakers are that are used in them is not disclosed for any of the ASR services and therefore, it cannot be said for certain that they were trained on Irish speakers. This could be a contributing factor to the higher WER rates for both Google Cloud STT and Amazon Transcribe.

In addition, the Microsoft Azure STT service is more accessible and easier to use than both Amazon Transcribe and Google Cloud STT. Amazon Transcribe must be used within the AWS portal directly as opposed to Microsoft Azure STT, which can be accessed via an API call. Furthermore, Amazon Transcribe requires the target audio file to be stored in an Amazon S3 storage bucket for it to be transcribed directly via the AWS portal, which increases the complexity of using this ASR service. Although Google Cloud STT can be accessed via an API call, it also requires the target audio file to be uploaded to Google Cloud Storage prior to transcription, which makes it less straightforward to use than the Microsoft Azure STT. Therefore, Microsoft Azure is the best performing, most accessible, and easiest to use transcription service overall based on this dataset used and within the context of this study.

4.2. Facial Expression Processing

Emotion probabilities for both joy and anger are output on a frame-by-frame basis by Emotient FACET, Affectiva AFFDEX, and the Microsoft Azure Face API. Kruskal-Wallis Rank Sum tests and Wilcoxon Rank Sum tests are applied to the results of each as described in Section 3.7 (Statistical Processing & Hypothesis Testing) above. To carry out these tests, the percentage of frames that return > 50% probability of joy or anger out of the total number of frames in each video are determined, respectively, with one percentage calculated for joy and one percentage calculated for anger for each video in the dataset. This is conducted on the output of all three facial expression emotion recognition software systems, resulting in a total of 50 joy values and 50 anger values for each, there being 50 videos in the dataset in total. These are the values input to the hypothesis tests to examine if there is a difference between joy/anger detected of in each system. The generic hypotheses being tested and their evaluation criteria (p-value ≤ 0.5 reject the null hypothesis and p-value > 0.5 accept the null hypothesis) are outlined in Section 3.7 (Statistical Processing & Hypothesis Testing) however, the specific hypotheses being tested are outlined for each case study below.

4.2.1. Joy

Figure 47 shows the frame that returned the highest probability of joy for each software system for one video in the dataset, being one of Irish male politician, Leo Varadkar. The frame on the left is that which returned the highest probability of joy by the Microsoft Azure Face API, the frame in the middle by Affectiva

AFFDEX, and the frame on the right by Emotient FACET. The frame that returned the highest probability of joy by the Microsoft Azure FACE API was frame 2688, occurring around the 1 minute 48 second mark, with a 100% probability of joy. The frame that returned the highest probability of joy by Affectiva AFFDEX was frame 2716, occurring around the 1 minute 49 second mark with a 99.91899% probability of joy. These frames are approximately 1 second apart and therefore, it can be seen that the highest probability of joy detected by both the Microsoft Azure Face API and Affectiva AFFDEX was roughly at the same point in time of the video. This suggests that these software systems might detect joy from facial expressions similarly however, this is only 1 of 50 videos in the dataset, with only one frame looked at for each system and therefore, the Wilcoxon Rank Sum test conducted below will properly assess if there is a similarity between joy detected in the Microsoft Azure Face API and Affectiva AFFDEX. The frame that returns the highest probability of joy by Emotient FACET is frame 1884, occurring around the 1 minute 15 second mark with an evidence score of 2.713351, which can be converted to a 99.8% probability of joy using the formula in Figure 21. This frame does not occur around the same time frame as those for Microsoft Azure Face API or Affectiva AFFDEX.

(a) (b) (c)

Figure 47 Highest emotion probability for joy across (a) MS Azure (b) Affectiva AFFDEX (c) Emotient FACET

The results of the Kruskal-Wallis Rank Sum test are seen in Table 5. This test was used to examine if there is a statistically significant difference between joy detected in the three facial expression emotion recognition systems. The hypotheses are as follows:

- H_0: distribution of *joy* is the same for Emotient FACET, Affectiva AFFDEX, and Microsoft Azure Face API
- H_1: distribution of *joy* is not the same for Emotient FACET, Affectiva AFFDEX, and Microsoft Azure Face API

Joy	p-value (α)	H_0
Azure v Emotient v Affectiva	1.19e-11	REJECT

Table 5 Kruskal-Wallis Facial Processing (Joy)

It can be seen in Table 5 that the p-value returned by the Kruskal-Wallis Rank Sum test is 1.19e-11, which is ≤ 0.5 and therefore, the null hypothesis (H_0) is rejected in favour of the alternative hypothesis (H_1) that the distribution of joy is not the same for Emotient FACET, Affectiva AFFDEX, and Microsoft Azure Face API. This means there is a statistically significant difference between joy detected in the three software systems.

Table 6 and Table 7 show the results of Wilcoxon Rank Sum Tests conducted on this data, Table 6 shows the results of the Wilcoxon Rank Sum tests conducted on pairs of facial expression emotion recognition software systems to test if there is a statistically significant difference between joy detected in each pair. The hypotheses for this case study are:

- H_0: distribution of *joy* is the same for both systems
- H_1: distribution of *joy* is not the same for both systems

with ‘both systems’ referring to either Emotient FACET & Affectiva AFFDEX, Emotient FACET & Microsoft Azure Face API, or Affectiva AFFDEX & Microsoft Azure Face API.

Joy	p-value (α)	H_0
Emotient v Affectiva	1.416e-10	REJECT
Affectiva v Azure	1.859e-08	REJECT
Emotient v Azure	***0.05789***	***ACCEPT***

Table 6 Wilcoxon Facial Processing (Joy)

It is seen from results in Table 6 that the Wilcoxon Rank Sum test conducted on Emotient FACET and Microsoft Azure Face API joy detection percentages returns a p-value > 0.5, being 0.05789, therefore, the null hypothesis is accepted that the distribution of joy is the same for Emotient FACET & Microsoft Azure Face API. Both tests on Emotient FACET & Affectiva AFFDEX and Affectiva AFFDEX & Microsoft Azure Face API return p-values 1.416e-10 and 1.859e-08, respectively, which are both ≤ 0.5 and therefore, the null hypothesis is rejected in favour of the alternative hypothesis in both cases, that the distribution of joy is not the same for Emotient FACET & Affectiva AFFDEX nor Affectiva AFFDEX & Microsoft Azure Face API. This means there is a statistically significant difference between joy detected in both Emotient FACET & Affectiva and Affectiva AFFDEX & Microsoft Azure Face API but there is not a statistically significant difference between joy detected in Emotient FACET & Microsoft Azure Face API.

Table 7 shows the results of the Wilcoxon Rank Sum test carried out to test if there is a statistically significant difference between joy detected in males and females by each individual software system. The hypotheses are as follows:

- H_0: distribution of *joy* is the same for both genders
- H_1: distribution of *joy* is not the same for both genders

Male v Female - *Joy*	p-value (α)	H_0
Emotient	***0.2293***	***ACCEPT***
Affectiva	***0.2778***	***ACCEPT***
Azure	***0.2102***	***ACCEPT***

Table 7 Male v Female Wilcoxon Facial Processing (Joy)

Results in Table 7 show that there is a not a statistically significant difference between joy detected in males and females by each individual system as the p-values for all three tests are > 0.5, being 0.2293 for Emotient FACET, 0.2778 for Affectiva AFFDEX, and 0.2102 for Microsoft Azure Face API. Because each of these p-values are > 0.5, the null hypothesis is accepted in each case, that the distribution of joy is the same for both genders in Emotient FACET, Affectiva AFFDEX, and Microsoft Azure Face API, respectively.

4.2.2. Anger

Figure 48 shows the frame that returns the highest probability of anger per software system for the same dataset video of Leo Varadkar. The frame on the left is that which returned the highest probability of anger by the Microsoft Azure Face API, the frame in the middle by Affectiva AFFDEX, and the frame on the right by Emotient FACET. The frame that returned the highest probability of anger by the Microsoft Azure FACE API was frame 4651, which occurrs around the 3 minute 6 second mark and returns 99.9% probability of anger. For Affectiva AFFDEX, the frame that returned the highest probability of anger was frame 2517, occurring around the 1 minute 41 second mark with a 99.65369% probability of anger, and for Emotient FACET it was frame

1388, occurring around the 55 second mark with an evidence score of 2.402112 i.e. 99.6% probability of anger. None of these three frames occur within the same timeframe of each other in the video therefore suggesting that none of the software systems detect anger similarly solely on the basis of this video, however, hypothesis testing to come will examine this further.

(a) (b) (c)

Figure 48 Highest emotion probability for anger across (a) MS Azure (b) Emotient FACET (c) Affectiva AFFDEX

Table 8 shows the results of the Kruskal-Wallis Rank Sum test for the three facial expression emotion recognition software systems on anger. It is examining if there exists a statistically significant difference between anger detected in the three systems. The hypotheses are as follows:

- H_0: distribution of *anger* is the same for Emotient FACET, Affectiva AFFDEX, and Microsoft Azure Face API
- H_1: distribution of *anger* is not the same for Emotient FACET, Affectiva AFFDEX, and Microsoft Azure Face API

Anger	p-value (α)	H_0
Azure v Emotient v Affectiva	< 2.2e-16	REJECT

Table 8 Kruskal-Wallis Facial Processing (Anger)

The p-value returned by this test is ≤ 0.5, being < 2.2e-16, and therefore, the null hypothesis is rejected in favour of the alternative hypothesis that the distribution of anger is not the same for Emotient FACET, Affectiva AFFDEX, and Microsoft Azure Face API. Hence, there is a statistically significant difference between anger detected in the three software systems.

Table 9 shows the results of the Wilcoxon Rank Sum tests conducted on each pair of software systems to test if there's a statistically significant difference between anger detected in each pair, being Emotient FACET & Affectiva AFFDEX, Affectiva AFFDEX & Microsoft Azure Face API, and Emotient FACET & the Microsoft Azure Face API. The hypotheses being tested here are:

- H_0: distribution of *anger* is the same for both systems
- H_1: distribution of *anger* is not the same for both systems

Anger	p-value (α)	H_0
Emotient v Affectiva	***0.313***	***ACCEPT***
Affectiva v Azure	< 2.2e-16	REJECT
Emotient v Azure	2.722e-16	REJECT

Table 9 Wilcoxon Facial Processing (Anger)

Results presented in Table 9 show that the only pair of software systems that do not have a statistically significant difference between anger detected in them is Emotient FACET & Affectiva AFFDEX, as the Wilcoxon Rank Sum test returned a p-value of 0.5224, which is > 0.5 and therefore, the null hypothesis is accepted that the distribution of anger is the same for Emotient FACET and Affectiva AFFDEX. The test

conducted on Affectiva AFFDEX & the Microsoft Azure Face API returns a p-value of < 2.2.e-16 therefore, the null hypothesis is rejected in favour of the alternative hypothesis that the distribution of anger is not the same for both systems. Similarly, the third hypothesis test returns a p-value of 2.722e-16, which means that the null hypothesis that the distribution of anger is the same for Emotient FACET & Microsoft Azure Face API is rejected in favour of the alternative hypothesis. For both Affectiva AFFDEX & the Microsoft Azure Face API and Emotient FACET & the Microsoft Azure Face API there is a statistically significant difference between anger detected in each system.

The results of the Wilcoxon Rank Sum tests examining if there is a statistically significant difference between anger detected in males and females for each system are seen in Table 10. These results relate to the following hypotheses:

- H_0: distribution of *anger* is the same for both genders
- H_1: distribution of *anger* is not the same for both genders

Male v Female - *Anger*	p-value (α)	H_0
Emotient	0.003049	REJECT
Affectiva	***0.8041***	***ACCEPT***
Azure	0.006491	REJECT

Table 10 Male v Female Wilcoxon Facial Processing (Anger)

Table 10 shows that only Affectiva AFFDEX does not have a statistically significant difference between anger detected in males and females. The p-value returned for the Wilcoxon Rank Sum test using Affectiva AFFDEX joy detection percentages is 0.8041, which is > 0.5 and therefore, the null hypothesis is accepted that the distribution of anger is the same for both genders. However, for Emotient FACET and the Microsoft Azure Face API, both p-values are ≤ 0.5 and therefore, both reject the null hypothesis in favour of the alternative hypothesis that the distribution of anger is not the same for both genders, meaning there is a statistically significant difference between anger detected in males and females by both Emotient FACET and the Microsoft Azure Face API.

4.2.3. Discussion

Overall, it can be seen that the distribution of both joy and anger in all three facial expression emotion recognition systems is not the same as per the results of both Kruskal-Wallis Rank Sum tests conducted, i.e. there is a statistically significant difference between the detection of joy and anger in all three systems. However, looking more granularly at pairs of software systems, it can be deduced that for joy, there is no statistically significant difference between joy detected in Emotient FACET and the Microsoft Azure Face API and for anger, there is no statistically significant difference between anger detected in Emotient FACET and Affectiva AFFDEX. Looking at gender differences, there is no statistically significant difference between joy detected in males and females by any of the three systems. However, for anger, there is a statistically significant difference between anger detected in males and females by Emotient FACET and by the Microsoft Azure Face API, but not by Affectiva AFFDEX.

These differences may be a result of how each individual algorithm was trained. As discussed in Section 3.4 (Facial Expression Processing) earlier, each of the software systems use different databases, statistical procedures, and facial landmarks to train the classification algorithms and therefore, variances in how they process facial expressions and determine emotion are expected. In addition, for the gender results, the

number of males or females in each training database would have an effect on the detection of emotion in males versus females by each system. However, the exact gender breakdown for none of the training databases used to train these three software systems is available and therefore, the cause of these gender variations cannot be reasoned for certain.

It was also seen in previous literature that Emotient FACET performs better on spontaneous facial expressions and Affectiva AFFDEX on posed expressions. Given this dataset is of politicians i.e. semi-trained actors, it is interesting to note that there is no statistically significant difference between anger detected in Emotient FACET and Affectiva AFFDEX, but only for joy detected.

Another reason for variations between emotion detected in the various systems could be down to the different classes of emotions detected by each system. This could result in increased ambiguity around which class of emotion a facial expression most likely falls into should two emotions be similar in how they are expressed. However, in this case, all three systems detect the same range of emotions – joy, anger, fear, sadness, disgust, surprise, neutral – the only difference being Emotient FACET and Affectiva AFFDEX both detect contempt as well – therefore, this is not a reasonable cause of variations between emotion detected in each.

4.3. Speech Processing

Emotion is detected from audio file segments by speech emotion recognition software systems OpenSMILE, OpenVokaturi, and DeepAffects. 2000ms audio chunks are input to both OpenSMILE and OpenVokaturi, with each returning a probability value for various emotions including both joy and anger, while DeepAffects takes 3000ms audio chunks as input and outputs one singular categorical value as output, being the emotion that is most dominant for that segment. DeepAffects recognizes both joy and anger also. To apply the Kruskal-Wallis Rank Sum and Wilcoxon Rank Sum hypothesis tests, similar to the process followed for facial expression emotion detection analysis above, the percentage of 2000ms chunks that return > 50% probability of joy or anger out of the total number of 2000ms chunks in each audio is determined, respectively, with one percentage calculated for joy and one for anger for each video in the dataset for both OpenSMILE and OpenVokaturi. For the DeepAffects output, the percentage of 3000ms chunks that return joy or anger as the most dominant emotion out of the total number of 3000ms chunks is calculated, respectively. This results in each video in the dataset having a percentage value for joy and a percentage value for anger for each software system of the percentage frames where each respective emotion is most dominant across all three software systems, a total of 50 values per emotion per software system, there being 50 audios in the dataset. These are the values input to the hypothesis tests described below.

4.3.1. Joy

Table 11 shows the results for the Kruskal-Wallis Rank Sum test conducted on the OpenSMILE, OpenVokaturi, and DeepAffects joy detection percentages. This is examining if there is a statistically significant difference between joy detected in all three systems. The hypotheses for this test are:

- H_0: distribution of *joy* is the same for OpenSMILE, OpenVokaturi, and DeepAffects
- H_1: distribution of *joy* is not the same for OpenSMILE, OpenVokaturi, and DeepAffects

Joy	p-value (α)	H_0
OpenSMILE v OpenVokaturi v DeepAffects	9.12e-10	REJECT

Table 11 Kruskal-Wallis Speech Processing (Joy)

Results show a p-value of 9.12e-10 was returned by the Kruskal-Wallis Rank Sum test, which is ≤ 0.5 and therefore, the null hypothesis was rejected in favour of the alternative hypothesis that the distribution of joy is not the same for OpenSMILE, OpenVokaturi, and DeepAffects. This means there is a statistically significant difference between joy detected in OpenSMILE, OpenVokaturi, and DeepAffects.

Table 12 shows the results of a set of Wilcoxon Rank Sum tests on pairs of software systems for speech emotion recognition to examine if there is a statistically significant difference between joy detected in each pair. The hypothesis for each pair of systems tested are as follows:

- H_0: distribution of *joy* is the same for both systems
- H_1: distribution of *joy* is not the same for both systems

Joy	p-value (α)	H_0
OpenSMILE v OpenVokaturi	4.579e-10	REJECT
OpenSMILE v DeepAffects	2.704e-09	REJECT
OpenVokaturi v DeepAffects	***0.5666***	***ACCEPT***

Table 12 Wilcoxon Speech Processing (Joy)

From the results of these Wilcoxon Rank Sum tests it can be deduced that only OpenVokaturi and DeepAffects do not have a statistically significant difference between joy detected in each, as the p-value returned by the test is 0.5666, being > 0.5 and therefore, the null hypothesis is accepted that the distribution of joy is the same for both OpenVokaturi & DeepAffects. On the other hand, both Wilcoxon Rank Sum tests carried out on OpenSMILE & OpenVokaturi and OpenSMILE & DeepAffects, respectively, return p-values ≤ 0.5. Thus, the null hypothesis is rejected for both of these pairs in favour of the alternative hypotheses that the distribution of joy is not the same for OpenSMILE & OpenVokaturi nor for OpenSMILE & DeepAffects, respectively. In both of these cases, there is a statistically significant difference between joy detected in each system.

The results of the three Wilcoxon Rank Sum tests conducted on each individual speech emotion recognition software can be seen in Table 13, which examine if there is a statistically significant difference between joy detected in males and females by each software. Each is testing the following hypothesis:

- H_0: distribution of *joy* is the same for both genders
- H_1: distribution of *joy* is not the same for both genders

Male v Female - *Joy*	p-value (α)	H_0
OpenSMILE	NA	NA
OpenVokaturi	0.001548	REJECT
DeepAffects	0.009501	REJECT

Table 13 Male v Female Wilcoxon Speech Processing (Joy)

The results of the Wilcoxon Rank Sum test on OpenSMILE data were inconclusive as no p-value was returned by the *wilcox.test* R function on this data and therefore, it cannot be concluded if there is a statistically significant difference between joy detected in males and females by OpenSMILE. However, Table 13 shows the p-value for both the OpenVokaturi and DeepAffects hypothesis tests to be ≤ 0.5, with p-values of 0.001548 and 0.009501, respectively, and therefore, it can be concluded that there is a statistically significant difference between joy detected in males and females by both OpenVokaturi and DeepAffects as the null hypothesis was

rejected for each software system in favour of the alternative hypothesis that the distribution of joy is not the same for both genders.

4.3.2. Anger

The result of the Kruskal-Wallis Rank Sum test examining if there is a statistically significant difference between anger detected in OpenSMILE, OpenVokaturi and DeepAffects is seen in Table 14. The hypothesis for this test is:

- H_0: distribution of *anger* is the same for OpenSMILE, OpenVokaturi, and DeepAffects
- H_1: distribution of *anger* is not the same for OpenSMILE, OpenVokaturi, and DeepAffects

Anger	p-value (α)	H_0
OpenSMILE v OpenVokaturi v DeepAffects	< 2.2e-16	REJECT

Table 14 Kruskal-Wallis Speech Processing (Anger)

A p-value of < 2.2e-16 is returned by the Kruskal-Wallis Rank Sum test, causing the null hypothesis to be rejected in favour of the alternative hypothesis that the distribution of anger is not the same for OpenSMILE, OpenVokaturi, and DeepAffects. Therefore, it can be said that there is a statistically significant difference between anger detected in all three software systems.

Next, each pair of speech emotion recognition software systems is input to a Wilcoxon Rank Sum test to determine if there is a statistically significant difference between anger detected in each pair of systems, with the following hypothesis pertaining to each Wilcoxon Rank Sum test conducted:

- H_0: distribution of *anger* is the same for both systems
- H_1: distribution of *anger* is not the same for both systems

Anger	p-value (α)	H_0
OpenSMILE v OpenVokaturi	< 2.2e-16	REJECT
OpenSMILE v DeepAffects	***0.1257***	***ACCEPT***
OpenVokaturi v DeepAffects	5.282e-13	REJECT

Table 15 Wilcoxon Speech Processing (Anger)

The results of these Wilcoxon Rank Sum tests can be seen in Table 15. From these results, it can be concluded that only OpenSMILE & DeepAffects do not have a statistically significant difference between anger detected in each. The p-value returned for OpenSMILE & DeepAffects is 0.1257, which is > 0.5 and therefore, the null hypothesis that the distribution of anger is the same for OpenSMILE & DeepAffects is accepted. However, the p-value returned for OpenSMILE & OpenVokaturi is < 2.2e-16 and the p-value returned for OpenVokaturi & DeepAffects is 5.282e-13. Both of these p-values are ≤ 0.5 hence, the null hypothesis is rejected in both cases for the alternative hypothesis, and it can be said that there is a statistically significant difference between anger detected in OpenSMILE & OpenVokaturi and OpenVokaturi & DeepAffects, respectively.

The final hypothesis test conducted on speech emotion recognition software systems is to examine if there is a statistically significant difference between anger detected in males and females by each system. This is done by conducting three Wilcoxon Rank Sum tests, one for each software system being analysed, with the results derived shown in Table 16 below. The hypotheses for each of these tests is:

- H_0: distribution of *anger* is the same for both genders
- H_1: distribution of *anger* is not the same for both genders

Male v Female - *Anger*	p-value (α)	H_0
OpenSMILE	***0.1615***	***ACCEPT***
OpenVokaturi	0.02559	REJECT
DeepAffects	***0.09211***	***ACCEPT***

Table 16 Male v Female Wilcoxon Speech Processing (Anger)

The results of these Wilcoxon Rank Sum tests show that only OpenVokaturi has a statistically significant difference between anger detected in males and females. The p-value returned for OpenVokaturi is 0.02559, which is ≤ 0.5 and therefore, the null hypothesis is rejected in favour of the alternative that the distribution of anger is not the same for both genders. However, the p-values for both OpenSMILE and DeepAffects are both > 0.5, which means the null hypothesis is accepted in both cases, that the distribution of anger is the same for both genders. This means that there is not a statistically significant difference between anger detected in males and females by neither OpenSMILE nor DeepAffects.

4.3.3. Discussion

Similar to what was discussed regarding variances in joy and anger detected among the facial expression emotion recognition systems in Section 4.2.3 (Discussion), it can also be said that differences in training databases, acoustic features extracted, and statistical processes, etc. can account for variances in the detection of joy and anger in each of the speech emotion recognition software systems. For example, literature found the mean F0 of both joy and anger to be high and thus, a system may not be able to distinguish between these two emotions on the basis of mean F0 alone. Therefore, it would depend on what other acoustic features are extracted and used by the algorithm whether emotion is determined to be more dominantly joy or anger.

It can be seen that both the Kruskal-Wallis Rank Sum tests conducted for joy and anger reject the null hypothesis, meaning the three speech emotion recognition systems have statistically significant differences between joy and anger detected in each. However, conducting pairwise hypothesis tests showed both OpenVokaturi & DeepAffects to not have a statistically significant difference between joy detected in each and OpenSMILE & DeepAffects to not have a statistically significant difference between anger detected in each.

OpenSMILE is trained on two databases – the EmoDB, which consists solely of German speakers, and the eNTERFACE database, consisting of 14 nationalities, none of whom are Irish. OpenVokaturi is trained on the EmoDB also, as well as a second database, SAVEE, consisting of solely British speakers, which is not one of the nationalities in the second database used to train OpenSMILE, and could give reason as to why OpenSMILE and OpenVokaturi have statistically significant differences between both joy and anger detected in each. Findings across literature suggest mean F0 to, at the very least, vary across nationality therefore, systems trained on different nationalities would interpret F0 differently. If this is the case for mean F0, it could very much be the case for other acoustic features extracted by each system used to determine emotion as well, causing each system to encode acoustic cues of speech very differently when determining emotion, based on the characteristics of emotion in the acoustic features of different nationalities in the training databases. However, given both OpenSMILE and OpenVokaturi were trained on an overlapping database, EmoDB, it cannot be said for certain the differences in nationalities within the additional databases used to train each are the sole cause of differences between joy and anger detected in each. The demographic breakdown of the training database used to train the DeepAffects algorithm is not available and therefore, no inferences can be made as to why there are

no statistically significant differences between joy detected in DeepAffects & OpenVokaturi and between anger detected in DeepAffects & OpenSMILE.

The results of examining a statistically significant difference between joy detected in males and females by OpenSMILE were inconclusive however, there is no statistically significant difference between anger detected in males and females as per the results of OpenSMILE. There is a statistically significant difference between both joy detected and anger detected in males and females by OpenVokaturi, with OpenVokaturi being the only software system of the three to have a statistically significant difference between anger detected in males and females. These results for OpenVokaturi could potentially be attributed to the training database of emotional speech used to train OpenVokaturi, which consists solely of male speakers. This could cause bias in the algorithm as the training database contains no female speakers, resulting in a difference between emotion detected in male and female speakers, as seen in the results of this study. OpenSMILE, on the other hand, was trained on a database that contains both male and female speakers, giving reason as to why there was no statistically significant difference between anger detected in males and females by OpenSMILE. The gender breakdown of the database used to train DeepAffects is not available however, results show there is no statistically significant difference between anger detected in males and females by DeepAffects, but there is a statistically significant difference between joy detected in males and females.

As discussed for the facial expression software systems' results, the various classes of emotions detected by each system could also contribute to variations between systems by causing ambiguity around what acoustic cues fit into what category of emotion should two emotions be expressed very similarly in speech. All speech emotion recognition systems implemented in this study detect joy, anger, sadness, and neutral, with both OpenSMILE and OpenVokaturi detecting fear also, solely OpenSMILE detecting boredom, and solely DeepAffects detecting excitement and frustration.

4.4. Text Processing

Sentiment analysis can be conducted on either the entire transcript or the individual 2000ms chunk transcript excerpts, with the percentage of positive and negative words out of the total number of words in the text being returned by each sentiment lexicon, respectively. The lexicons employed in this study are the General Inquirer, used in the RockSteady sentiment analysis tool, the NRC Emotion Lexicon, and the Opinion Lexicon. In order to conduct hypothesis testing on the three methods of text sentiment analysis, each entire transcript in the dataset will undergo sentiment analysis. These transcripts have been derived using ASR service Microsoft Azure STT, given it was the ASR service to achieve the lowest mean WER rate on this dataset, as seen in Table 4. Two values are output per transcript in the dataset, one for the percentage of positive words and one for the percentage of negative words out of total words in each transcript. Hence, a total of 50 percentage values are computed by each sentiment analysis method per category of sentiment, being either positive or negative sentiment, with one value computed for each per transcript in the dataset. These are the values used in the hypothesis tests below. The word 'lexicon' is used in the description of these tests to represent the three sentiment analysis methods employed in this study as although the RockSteady sentiment analysis tool is used, it is solely based off the General Inquirer lexicon in this case, as no domain-specific secondary lexicons are used.

4.4.1. Positive

Table 17 contains the results of a Kruskal-Wallis Rank Sum test conducted on the percentages of positive sentiment returned by RockSteady (which uses the General Inquirer lexicon), the NRC Emotion Lexicon, and the Opinion Lexicon. It examines if there is a statistically significant difference between positive sentiment detected using all three lexicons. The hypothesis being tested is:

- H_0: distribution of *positive sentiment* is the same for all three lexicons
- H_1: distribution of *positive sentiment* is not the same for all three lexicons

Positive	p-value (α)	H_0
RockSteady v NRC v Opinion	1.375e-06	REJECT

Table 17 Kruskal-Wallis Text Processing (Positive)

The p-value returned by this test is 1.375e-0, which is ≤ 0.05, meaning the null hypothesis is rejected in favour of the alternative hypothesis that the distribution of positive sentiment is not the same for all three lexicons. This means there is a statistically significant difference between positive sentiment detected using all three lexicons.

Next, a set of Wilcoxon Rank Sum tests were conducted to test if there is a statistically significant difference between positive sentiment detected by each pair of lexicons, being RockSteady & the NRC Emotion Lexicon, RockSteady & the Opinion Lexicon, and the NRC Emotion Lexicon & the Opinion Lexicon. The hypothesis for each test conducted is:

- H_0: distribution of *positive sentiment* is the same for both lexicons
- H_1: distribution of *positive sentiment* is not the same for both lexicons

Positive	p-value (α)	H_0
RockSteady v NRC	***0.3242***	***ACCEPT***
RockSteady v Opinion	1.9e-06	REJECT
NRC v Opinion	4.479e-05	REJECT

Table 18 Wilcoxon Speech Processing (Positive)

Results show that only for the RockSteady & NRC Emotion Lexicon pairing the null hypothesis is accepted. This test returned a p-value of 0.3242, which is > 0.5 and therefore, the null hypothesis that the distribution of positive sentiment is the same for both RockSteady (the General Inquirer lexicon) & the NRC Emotion Lexicon is accepted and hence, there is no statistically significant difference between positive sentiment detected using these lexicons. For both the RockSteady & Opinion Lexicon pair and the NRC Emotion Lexicon & Opinion Lexicon pair, the null hypothesis is rejected in favour of the alternative hypothesis that the distribution of positive sentiment is not the same for each pair. The p-values returned are 1.9e-06 and 4.479e-05, respectively, and it can be deduced that there is a statistically significant difference between positive sentiment detected using RockSteady & the Opinion Lexicon as well as between positive sentiment detected using the NRC Emotion Lexicon & the Opinion Lexicon.

In order to examine if there is a statistically significant difference between positive sentiment detected in male transcripts and female transcripts for each lexicon used, another set of Wilcoxon Rank Sum tests are carried out. The hypothesis for each test is:

- H_0: distribution of *positive sentiment* is the same for both genders
- H_1: distribution of *positive sentiment* is not the same for both genders

Male v Female - *Positive*	p-value (α)	H_0
RockSteady	1.308e-07	REJECT
NRC	9.507e-08	REJECT
Opinion	0.0001804	REJECT

Table 19 Male v Female Wilcoxon Speech Processing (Positive)

Results of these tests are detailed in Table 19, with all three systems rejecting the null hypothesis on the basis of p-values of 1.308e-07, 9.507e-08, and 0.0001804 for RockSteady, the NRC Emotion Lexicon, and the Opinion Lexicon, respectively. Therefore, for each lexicon, the alternative hypothesis that the distribution of positive sentiment is not the same for both genders' transcripts is accepted, and it can be said that there is a statistically significant difference between positive sentiment detected in male transcripts and female transcripts using RockSteady, the NRC Emotion Lexicon, as well as the Opinion Lexicon.

4.4.2. Negative

The same tests are conducted on negative sentiment detected using each lexicon, the first being a Kruskal-Wallis Rank Sum test to examine if there is a statistically significant difference between negative sentiment detected using the three lexicons the General Inquirer (via RockSteady), the NRC Emotion Lexicon, and the Opinion Lexicon, with the following hypothesis:

- H_0: distribution of *negative sentiment* is the same for all three lexicons
- H_1: distribution of *negative sentiment* is not the same for all three lexicons

Negative	p-value (α)	H_0
RockSteady v NRC v Opinion	***0.3202***	***ACCEPT***

Table 20 Kruskal-Wallis Text Processing (Negative)

Results of this Kruskal-Wallis Rank Sum test are seen in Table 20. The p-value returned is 0.3202, which is > 0.5 and therefore, the null hypothesis that the distribution of negative sentiment is the same for all three lexicons is accepted and it can be said that there is no statistically significant difference between negative sentiment detected using the three lexicons.

This result is corroborated by the results of the Wilcoxon Rank Sum tests conducted on each pair of lexicons, used to determine if there is a statistically significant difference between negative sentiment detected using each pair. The results of these tests are seen in Table 21, with each testing the following hypothesis:

- H_0: distribution of *negative sentiment* is the same for both lexicons
- H_1: distribution of *negative sentiment* is not the same for both lexicons

Negative	p-value (α)	H_0
RockSteady v NRC	***1***	***ACCEPT***
RockSteady v Opinion	***0.2009***	***ACCEPT***
NRC v Opinion	***0.1856***	***ACCEPT***

Table 21 Wilcoxon Speech Processing (Negative)

The p-value returned by each individual test is > 0.5 and therefore, the null hypothesis that the distribution of negative sentiment is the same for both lexicons is accepted for each pair of lexicons. There is no statistically significant difference between negative sentiment detected using RockSteady & the NRC Emotion Lexicon, nor RockSteady & the Opinion Lexicon, nor the NRC Emotion lexicon & the Opinion Lexicon.

Looking at the difference between negative sentiment detected in the transcripts of male and female politicians, it can be seen that there is a statically significant difference between negative sentiment detected in male transcripts and female transcripts for each lexicon used. These results are shown in Table 22, with each testing the following hypothesis for each individual lexicon:

- H_0: distribution of *negative sentiment* is the same for both genders
- H_1: distribution of *negative sentiment* is not the same for both genders

Male v Female - *Negative*	p-value (α)	H_0
RockSteady	2.71e-08	REJECT
NRC	3.577e-08	REJECT
Opinion	8.432e-06	REJECT

Table 22 Male v Female Wilcoxon Speech Processing (Negative)

The p-value returned for each Wilcoxon Rank Sum test is ≤ 0.05, meaning the null hypothesis is rejected in favour of the alternative hypothesis that the distribution of negative sentiment is not the same for both male and female transcripts for each lexicon used, being the General Inquirer (via RockSteady), the NRC Emotion Lexicon and the Opinion Lexicon.

4.4.3. Discussion

Overall, it can be seen for positive sentiment that there was a statistically significant difference between positive sentiment detected using all three lexicons, however, there was no statistically significant difference between positive sentiment using RockSteady and the NRC Emotion Lexicon, being the only pair of lexicons not to return a statistically significant difference between positive sentiment detected. For negative sentiment, on the other hand, it was seen that there were no statistically significant differences between negative sentiment detected using any of the three lexicons, with results for both the Kruskal-Wallis Rank Sum test on all three lexicons and the Wilcoxon Rank Sum tests on each pair of lexicons accepting the null hypothesis that the distribution of negative sentiment is the same for all three or each pair of lexicons.

These differences and similarities can be put down to the manner in which each individual lexicon was built as well as the size of each. For example, RockSteady uses the General Inquirer lexicon, which contains 3,486 words, whereas the NRC Emotion lexicon contains 14,182 words and the Opinion Lexicon 6,789 words. The sheer difference in number of words in each lexicon means that a word annotated as ‘positive’ or ‘negative’ in one lexicon may not even be included in another, giving rise to differences in text sentiment analysis results. In addition, the way in which the lexicons are created may have an impact. The NRC Emotion Lexicon was built through crowdsourcing, for example, which allows people to determine words as positive or negative and may give rise to differences in the annotation of words across lexicons as a result of human mis-judgement or bias in the creation of this NRC Emotion Lexicon.

In addition, it was seen that for both positive and negative sentiment, there were statistically significant differences between positive sentiment detected as well as negative sentiment detected in male and female transcripts by each individual lexicon. This supports findings from literature that state gender differences exist between the expression of emotion and sentiment in text [80, 81].

4.5. Facial Action Units

Emotion in facial expression is detected through a linear combination of AUs [28], with those typical of a particular emotion carrying more weight and thus, being more dominant in the linear combination of AUs and the subsequent detection of that emotion. To determine which AUs from Paul Ekman' FACS are 'most dominant' for this given dataset of male and female Irish subjects, Spearman's Rank-Order Correlation was used. This test produces a Rho value, being the correlation coefficient, used to determine which AUs are more strongly positively correlated with joy and anger. Only positive Rho values are considered because an increase in an emotion probability as the probability of an AU increases is desired in this case.

Emotions joy and anger and the AUs seen in Table 2 are used in this analysis. Emotient FACET and Affectiva AFFDEX both output probabilities for both emotions and AUs and therefore, the data output by these two facial expression emotion recognition systems is used. Emotient FACET outputs an evidence score for both joy and anger as well as for each AU in Table 2, which are subsequently converted to probabilities using the formula in Figure 21. Affectiva AFFDEX outputs a probability for joy and anger as well as for each AU in Table 2 except for AU23. All results for each video in the dataset are used in this analysis, comprising a total of 233,702 video frames across all 50 dataset videos. For each of the 233,702 frames there is a joy probability, anger probability, and set of AU probabilities output by each software system. These values will be input into the Spearman's Rank-Order Correlation tests, with one test being conducted for each emotion-AU pair using probabilities output by each software system e.g. AU1 & joy for Emotient, AU1 & joy for Affectiva, ..., AU43 & joy for Emotient, AU43 & joy for Affectiva, with the same conducted for anger also. As stated in Section 3.7.4 (Spearman's Rank-Order Correlation) above, an AU will be determined as a 'most dominant' AU for joy or anger if it returns a Rho value ≥ 0.4 on both Emotient FACET and Affectiva AFFDEX data.

Table 23 shows the results of Spearman's Rank-Order Correlation for joy where the AU-joy correlation returned a Rho value ≥ 0.4 for either Emotient FACET or Affectiva AFFDEX, or both. The AU that returned the strongest positive correlation with joy was AU12, with a Rho value of 0.59 on Affectiva AFFDEX data and 0.89 on Emotient FACET data. AU14 also returned strong positive correlation with joy on Emotient FACET data, with a Rho value of 0.62 however, it failed to return a Rho value ≥ 0.4 on Affectiva AFFDEX data and therefore, it does not meet the criteria to be deemed a 'most dominant' AU for joy. Only AU6 and AU12 return a Rho value of ≥ 0.4 for both Emotient FACET and Affectiva AFFDEX and therefore, these are the 'most dominant' AUs for joy as per the criteria in Section 3.7.4 (Spearman's Rank-Order Correlation).

Joy	System	Rho	
6	Affectiva	0.48	Moderate
	Emotient	0.40	Moderate
12	Affectiva	0.59	Moderate
	Emotient	0.89	Very Strong
14	Emotient	0.62	Strong
20	Emotient	0.45	Moderate
28	Emotient	0.52	Moderate

Table 23 Spearman's Correlation Rho for AUs and Joy

Table 24 shows the Rho values returned by Spearman's Rank-Order Correlation for AU-anger pairs using either Emotient FACET and Affectiva AFFDEX data where the Rho value was ≥ 0.4. AU7 and AU24

return Rho values 0.43 and 0.44 with anger, respectively, on Emotient FACET data however, the Rho value for these AUs and anger on Affectiva AFFDEX data are < 0.4 and therefore, they do not meet the criteria to be 'most dominant' AUs for anger. Similarly, AU10 returns a Rho value of 0.40 with anger on Affectiva AFFDEX data but a Rho value < 0.4 on Emotient FACET therefore, this AU also does not meet the criterion for a 'most dominant' AU with anger. AU4, however, returned strong positive correlation with anger on Affectiva AFFDEX data, with a Rho value of 0.66 as well as on Emotient FACET data, with a Rho value of 0.52, and therefore will be determined a 'most dominant' AU for anger. AU9 also returned Rho values ≥ 0.4 with anger for both Emotient FACET and Affectiva AFFDEX data and therefore, will also be deemed a 'most dominant' AU as per the criteria in Section 3.7.4 (Spearman's Rank-Order Correlation).

Anger	System	Rho	
4	Affectiva	0.66	Strong
	Emotient	0.52	Moderate
7	Emotient	0.43	Moderate
9	Affectiva	0.52	Moderate
	Emotient	0.72	Strong
10	Affectiva	0.40	Moderate
24	Emotient	0.44	Moderate

Table 24 Spearman's Correlation Rho for AUs and Anger

The AUs that meet the criteria to be 'most dominant' for either joy or anger are detailed in Table 25, including their corresponding muscle movement descriptions and facial muscles involved [149]. The AUs that are 'most dominant' for joy are AU6 and AU12. AU6 corresponds to the muscle movements that cause a subject's cheeks to raise and eyes narrow, known as the 'cheek raiser', while AU12, known as the 'lip corner puller', corresponds to a subject's lip corners being pulled up and laterally by the movement of the facial muscle *zygomaticus major* [149]. The AUs deemed 'most dominant' for anger are AU4 and AU9, with AU4, known as the 'brow lowerer' relating to the movement of a subject's eyebrows being drawn medially and down, caused by the muscle movement of the *corrugator supercilii* and *depressor supercilii* facial muscles and AU9, known as the 'nose wrinkler', corresponding to the movement of the *levator labii superioris alaeque nasi*. Each of these AUs are used in a linear combination of all AUs to determine emotion probabilities in an automatic facial expression emotion recognition system, such as Affectiva AFFDEX [28], with those deemed 'most dominant' for an emotion carrying more weight when that emotion is present.

Emotion	Dominant AU	AU Description	Muscle Movement	Facial Muscle
Joy	6	Cheek raiser	Cheeks raised; eyes narrowed	Orbicularis oculi, pars orbitalis
	12	Lip Corner Puller	Lip corners pulled up and laterally	Zygomaticus major
Anger	4	Brow Lowerer	Eyebrows drawn medially and down	Corrugator supercilii, depressor supercilii
	9	Nose Wrinkler	Upper lip raised and inverted; superior part of the nasolabial furrow deepened; nostril dilated by the medial slip of the muscle	Levator labii superioris alaeque nasi

Table 25 Most Dominant Action Units for Joy and Anger

4.6. Fundamental Frequency (*F0*)

To examine the range of values and mean value in which the mean F0 lies depending on the emotion being expressed in speech, the gender of the speaker, as well as their age, various statistical analyses based on the z-score value were conducted. The z-score is detailed in Section 3.7.5 (Z-score) above, with the formula to calculate a z-score seen in Figure 42. A z-score essentially tells you how far from the mean of a population a singular value is and will be used in this context to assess the behaviour of mean F0 when joy and anger are expressed in speech, as well as the overall mean F0 for male and female Irish-English speakers, and the variation of mean F0 with respect to age within each gender.

The mean F0 of each 2000ms audio chunk in the dataset is extracted and output by OpenSMILE. There are a total of 4,755 2000ms audio chunks segmented from all 50 audio files in the dataset with a mean F0 value being output by OpenSMILE for each of these chunks, resulting in 4,754 mean F0 values in total for the entire dataset. The overall mean of the population can then subsequently be derived by computing the mean of these 4,754 mean F0 values. The overall mean of all mean F0s from each 2000ms audio chunk in the dataset is 305Hz, and the standard deviation 98Hz. These values will be used in the computation of z-scores in Section 4.6.1 (Emotion) and Section 4.6.2 (Gender) below to examine the variation of mean F0 across instances of joy and anger in speech, as well as the mean F0 for both male and female Irish speakers. Summary statistics of the mean F0 for the overall population are seen in Table 26.

Min	Max	Mean	Median	Standard Deviation
90Hz	490Hz	305Hz	289Hz	98Hz

Table 26 Population summary statistics for F0 in Hz

4.6.1. Emotion

To determine the range of mean F0 values that pertain to the expression of joy and anger in speech, respectively, for this dataset, the audio files that return instances of joy and anger detected above the level of chance, being 1/7th as OpenSMILE detects seven emotions from speech, are first identified. Only 5 audio files of a total 50 audio files in the dataset returned instances of joy detected above the level of chance, being only 10% of the dataset. These audios are Arlene3, Mairead1, Mairead2, Mairead5, and Stephen5, with ‘Arlene3’, for example, being the third Arlene Foster audio in the dataset, and so on. Similarly, only six audios, being 12% of the total dataset, returned instances of anger detected above the level of chance, being the videos Arlene2, Arlene3, Mairead1, Mairead2, Mairead5, and Stephen5. Each instance relates to a 2000ms chunk within an audio file.

Next, summary statistics for the mean F0 values of all instances within each audio file identified as returning probabilities of joy or anger above the level of chance were computed and are presented in Table 27. The overall mean of each individual audio file’s overall mean of all mean F0 values for instances of joy/anger > 1/7th were also computed to obtain an overall mean F0 value for joy and anger, respectively, and a z-score calculated for each using the overall mean F0 and standard deviation of the entire population, being 305Hz and 98Hz, respectively, seen in Table 26. These summary statistics give an insight into the range of mean F0 where joy or anger are detected above the level of chance by OpenSMILE.

	JOY (Hz)				Anger (Hz)			
	Min	Max	Mean	Median	Min	Max	Mean	Median
Arlene2	-	-	-	-	202	296	254	259
Arlene3	276	441	357	338	215	446	310	297
Mairead1	277	490	443	469	253	490	379	354
Mairead2	227	460	342	341	289	433	340	331
Mairead5	215	473	360	354	351	371	361	361
Stephen5	207	479	373	409	190	466	319	300
Mean			**375**				**327**	
z-score			**0.7142**				**0.2244**	

Table 27 F0 (Hz) statistics for instances of Joy and Anger

From Table 27, it can be seen that the overall mean of each individual audio file's overall mean of all mean F0 values where joy is detected above the level of chance is 375Hz. The z-score derived for this value is 0.7142, meaning it is 0.7142^{th} of a standard deviation above the overall population mean. The same is calculated for anger, finding the overall mean F0 to be 327Hz, with a z-score of 0.2244, meaning the overall mean F0 for anger is 0.2244^{th} of a standard deviation above the overall population mean. Therefore, both the overall mean F0 for joy and overall mean F0 for anger lie higher than the overall population mean, which is indicative of findings in previous literature than the mean F0 for both joy and anger is 'high'.

The lowest mean F0 returned by any instance of a 2000ms chunk that returns a probability of joy above the level of chance in the dataset is 207Hz and of anger, 190Hz. Both of these are significantly higher than the overall population minimum F0, being 90Hz, as seen in Table 26. The maximum mean F0s found for any instance of joy or anger are both 490Hz, the same as the overall population. Therefore, the range of mean F0 for joy and anger can be seen to be much narrower and lie much higher than the overall population mean F0 range, further supporting literature finding the mean F0 of joy and anger to be 'high'. The mean F0 range for joy detected above the level of chance is 207Hz to 490Hz, and for anger 190Hz to 490Hz, whereas the overall mean F0 range for the entire population is 90Hz to 490Hz.

The median of mean F0s for the entire population is 289Hz. Looking at the median of mean F0s where joy is detected > $1/7^{th}$ in each of the five audio files, all median values lie much higher than the overall population median, ranging from 338Hz to 469Hz, compared to 289Hz for the overall population. For anger, five of the six audios have median values for the mean F0 higher than the overall population median, being 297Hz, 354Hz, 331Hz, 361Hz, and 300Hz. These median values additionally support the notion of the mean F0 for joy and anger being high as they are all higher than the median mean F0 across the entire dataset. There is only one audio with a median value of 259Hz for the mean F0 for anger that lies only slightly below the median mean F0 for the entire dataset.

In conclusion, analysis conducted on the mean F0 for instances of joy and anger detected above the level of chance within this dataset appear to reinforce findings from literature that this physical correlate of emotion in speech exerts high values when joy and anger are being expressed. Looking at the mean, minimum, and maximum values of mean F0 for both joy and anger, it appears both emotions have higher overall mean F0s compared to the overall population mean F0 and higher lower-bounds for the mean F0 range than the overall population mean F0 range, resulting in higher and narrower mean F0 ranges overall.

4.6.2. Gender

Next, the mean F0 for both male and female Irish-English speakers in this dataset was examined. Figure 49 plots the average mean F0 of each audio in the dataset, being the overall mean of the individual mean F0 values extracted by OpenSMILE for each 2000ms chunk per audio file. The blue dots represent the average mean F0 of each of the 25 audios of male speakers in the dataset, and the purple the 25 audios of female speakers in the dataset. There is no clear pattern derived from looking at Figure 49 and no clear delineation between the average mean F0 or mean F0 range of male Irish-English speakers versus female Irish-English speakers.

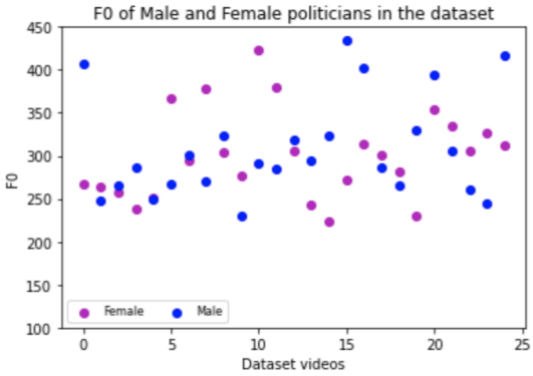

Figure 49 Mean F0 (Hz) of each video in the dataset (Blue - Male; Purple - Female)

		F0 (Hz)	z-score
Male	Min	90	-2.1938
	Max	490	1.18877
	Mean	308	0.0306
Female	Min	161	-1.4693
	Max	490	1.8877
	Mean	303	-0.0204

Table 28 F0 (Hz) statistics for Male and Female audios in the dataset

The 25 audios of male speakers in the dataset are comprised of 2,373 individual 2000ms chunks and the 25 audios of female speakers of 2,382 individual 2000ms chunks. Each chunk returns a single mean F0 value by OpenSMILE. A z-score is calculated for the minimum mean F0 value, maximum mean F0 value, and average of all mean F0 values across all 2000ms chunks, per gender, as seen in Table 28. The z-scores are calculated using the overall population values, being 305Hz mean F0 and 98Hz standard deviation, seen in Table 26. From these values, it can be seen that the overall mean of all male politician audios chunks' mean F0 values to be 308Hz and for females, 303Hz. Calculating each of these derives the z-scores 0.0306 for male mean F0 and -0.0204 for female mean F0, pointing to a higher overall mean F0 in Irish-English male speakers than Irish-English female speakers. However, these z-scores are miniscule values, meaning each overall mean F0 for males and females are within 5% of a standard deviation above or below the mean. Essentially, the overall mean F0s of male and female speakers, being 308Hz and 303Hz, lie very close to the overall population mean F0, being 305Hz.

The minimum mean F0 recorded for a male politician 2000ms audio chunk is 90Hz, and for a female politician 2000ms audio chunk 161Hz. The corresponding z-scores are -2.1938 and -1.4693, which mean the male minimum mean F0 falls over two standard deviations below the overall population mean, while the female minimum mean F0 falls only one and half standard deviations below. This indicates that the female mean F0 generally falls higher than the male mean F0, which supports literature that the mean F0 of females is higher than that of males. However, the maximum mean F0 found across both male audios and female audios is 490Hz, thus the same as the overall population maximum mean F0. From this, the range of mean F0 for males and females can be inferred, being 90Hz to 490Hz for males and 161Hz to 490Hz for females. This shows the mean F0 range of male speakers to be much larger than the mean F0 range of female speakers, which is supportive of studies on Japanese speakers however, it contradicts studies on American-English or Parisian-French speakers.

The average mean F0 for male speakers is 308Hz, which is higher than the female average mean F0, which is 303Hz. This contradicts all findings from literature, including scientific statements that the female mean F0 is higher than the male mean F0 because of the difference in the size of the larynx across genders [56]. This finding, therefore, is not reflective of what is known to be true in previous literature and therefore, might be a function of the small sample size of only 25 audios per gender in this dataset. However, the minimum mean F0 gives an indication that the female mean F0 range lies higher than the range for male mean F0 because the minimum mean F0 for females is much higher than the male minimum mean F0, being 161Hz and 90Hz, respectively.

4.6.3. Age

The variation in mean F0 that comes with age is analysed on a gender-basis, with literature finding the female mean F0 to decrease significantly with the onset of menopause, and the male mean F0 to remain stationary from the ages of 35 to 55, when it begins to increase again slightly. Therefore, z-scores calculated use the individual overall mean F0 and standard deviation values calculated for each gender, and not the overall population mean F0 and standard deviation, as used in the analyses in both Section 4.6.1 (Emotion) and Section 4.6.2 (Gender) above.

The overall mean of the 2,373 2000ms audio chunks' mean F0 values pertaining to male speakers in the dataset is 308Hz, and standard deviation 105Hz. The minimum mean F0, maximum mean F0, and overall mean F0 for all audio chunks pertaining to each individual male politician in the dataset are calculated and their z-scores derived, with all statistics presented in Table 29. Table 29 also shows the age of each politician, which provides the basis for this analysis. From Table 29 it can be seen that the youngest male politician, Leo Varadkar – aged 42, has the lowest minimum mean F0, lying over two standard deviations below the overall male mean F0, with a z-score of -2.0761. All four other male politicians have similar values for minimum mean F0, with each having a z-score in and around 1 and a half standard deviations below the overall male mean F0. However, the oldest politician, Michael Martin – aged 60, has a marginally higher minimum mean F0, which is narrowly indicative of previous literature that finds male mean F0 to increase after age 55, with Michael Martin being the only politician above this age. However, this difference in minimum mean F0 is extremely minimal and therefore, further analysis on a much larger dataset of male Irish-English speakers would be beneficial to corroborate this finding. The maximum mean F0s of each male politician are within 5Hz of each other, all being within 1.68^{th} to 1.74^{th} of a standard deviation above the overall male mean F0, which is a negligible variation. The results of z-scores on the average mean F0 for each politician do not show any coherent trend attributable to the age of the politicians. The politicians aged 42, 45, and 49 have negative z-scores as their average mean F0 values fall below the overall male mean F0, while the mean F0 of politicians aged 46 and 60 both fall above the mean. However, all these z-scores indicate the mean F0 values to be within half a standard deviation above or below the mean and therefore, the variances in mean F0 here are also negligible.

The finding from previous literature that the male F0 is stationary between ages 35 and 55 however, is indicative of findings in this study. There is only a marginal difference in the minimum and maximum mean F0s for each male politician in the dataset, with the exception of Leo Vardakar's minimum mean F0 being much lower. The average mean F0 values for each politician are all within half a standard deviation above or below the mean also and therefore, reflect quite a narrow range of mean F0s for male speakers that are quite clustered around the overall male mean F0 value. There is one politician aged 60 however, this age is only slightly above

the upper-bound of the supposed age range of stationary mean F0 proposed in previous literature and therefore, increased mean F0 after age 55 may not yet be recognised in the F0 of Michael Martin.

	Age	Min		Max		Mean	
		F0 (Hz)	z-score	F0 (Hz)	z-score	F0 (Hz)	z-score
Leo Varadkar	42	90	-2.0761	485	1.6857	298	-0.0952
Stephen Donnelly	45	155	-1.4571	490	1.7333	302	-0.0571
Paschal Donohoe	46	155	-1.4571	488	1.7142	347	0.3714
Simon Coveney	49	150	-1.5047	486	1.6952	279	-0.2761
Michael Martin	60	157	-1.4380	487	1.7047	320	0.1142

Table 29 Fundamental Frequency (F0) statistics of Male Irish Politicians

The same analysis is conducted for the female politicians in the dataset, with the z-scores being calculated using the overall mean F0 and standard deviation of all 2,382 individual 2000ms chunks' mean F0 values. The same analysis is conducted for the female politicians in the dataset, with the z-scores being calculated using the overall mean of the 2,373 2000ms audio chunks' mean F0 values pertaining to male speakers in the dataset, being 303Hz and standard deviation 90Hz. These z-scores are seen in Table 30 and are calculated on the minimum mean F0, maximum mean F0 and overall mean of all 2000ms mean F0s for each female politician in the dataset. The age of each female politician is included in Table 30 also and is the basis of this analysis.

These results show the youngest and oldest female politicians, Helen McEntee and Mairead McGuinness, aged 35 and 62, to have the highest minimum mean F0 of the five female politicians in the dataset, with both having z-scores larger than -1.5, meaning they are within one and a half standard deviations below the mean. The other three politicians, Michelle O'Neill, Arlene Foster, and Mary Lou McDonalds, aged 44, 51, and 52, have z-scores smaller than -1.5, meaning they are all over one and a half standard deviations below the mean. The former two politicians have minimum mean F0s in the 170Hz's, with the latter three having minimum mean F0s's in the 160Hz's. The difference in maximum mean F0 across the five female politicians is marginal, with each of their z-scores being either 2.0333, 2.0444, or 2.0777, all approximately two standard deviations above the mean. The overall mean F0 of each 2000ms chunks' mean F0 for each female politician, similar to what was said for the male results, does not show any clear trend or causation attributable to age of the politicians. The youngest politician, aged 35, has an overall mean F0 of 318Hz with a positive z-score of 0.1666. The eldest two politicians, aged 52 and 62, have positive z-scores also, being 0.2666 and 0.2555 while the politicians aged 44 and 51 have negative z-scores of -0.2777 and -0.5222. There appears to be a dip in overall mean F0 for Michelle O'Neill and Arlene Foster however, given there is only 1 year in difference between Arlene Foster and Mary Lou McDonald, whom each have negative and positive z-score, respectively, there is no evident age-related reason for why the dip in overall mean F0 occurs.

One interesting notion of why Michelle O'Neill and Arlene Foster have negative z-scores, with all other female politicians having positive z-scores may be the fact that both of these female politicians are from Northern Ireland, and the rest are from the Republic of Ireland. Hence, their accent may be a cause of their lower mean F0 values, as nationality is seen to be a cause for variance in F0 within previous literature. However, that is a large claim that cannot be substantiated by the small sample sized dataset used in this analysis therefore, it is only a thought for future analysis and not a conclusion substantiated in this work.

Previous literature has stated the mean F0 of female speakers to decrease with the onset of menopause, so roughly between ages 44 to 55 however, this is not reflected in the results of this analysis. This may again be a function of the small sample size used in this dataset or maybe reflective of an underlying reason that is not clear in this analysis.

	Age	Min		Max		Mean	
		F0 (Hz)	z-score	F0 (Hz)	z-score	F0 (Hz)	z-score
Helen McEntee	35	170	-1.4777	486	2.0333	318	0.1666
Michelle O'Neill	44	162	-1.5666	490	2.0777	278	-0.2777
Arlene Foster	51	164	-1.5444	487	2.0444	256	-0.5222
Mary Lou McDonald	52	161	-1.5777	490	2.0777	327	0.2666
Mairead McGuinness	62	176	-1.4111	490	2.0777	326	0.2555

Table 30 Fundamental Frequency (F0) statistics of Female Irish Politicians

4.7. Data Fusion

Decision-level data fusion is implemented in this study using the joy probabilities output by Emotient FACET and OpenSMILE as well as the positive sentiment analysis results output by RockSteady to determine if joy is strongly evident per 2000ms chunk of the video. The software systems Emotient FACET, OpenSMILE, and RockSteady, and emotion joy were selected as a means of illustrating this data fusion process. Positive sentiment was subsequently chosen for simplicity of illustration because it is solely associated with joy, unlike surprise which can be associated with both positive or negative sentiment. The data fusion process is illustrated using the results output by these three software systems for each modality on a single dataset video, being one of Irish male politician Leo Varadkar. This novelty two-step decision-level data fusion process is explained in detail in Section 3.8.1 (Data Fusion) above however, it essentially involves using a weighted sum of probabilities to fuse facial expression and speech probabilities for joy, and a logical "OR" formula using a combined decision rule to determine joy as strongly evident should the fused probability be > 0.5 and the percentage of positive sentiment be > 0% per 2000ms chunk of the dataset video.

In order to implement this two-step decision-level data fusion process, the modalities must first be aligned, with speech having been chosen as the most appropriate modality to base data segmentation for data fusion upon. As data fusion will therefore be conducted on this 2000ms segmented basis, the speech probabilities are already sufficiently prepared to be input to the first equation. To prepare the facial expression joy probabilities for input to the equation, the overall mean probability for joy across all the frames that occur in each 2000ms time segment is computed, as described in Section 3.8.1 (Data Fusion). To prepare the positive text sentiment results, 2000ms-lengthed transcript excerpts are input to RockSteady and the percentage of positive words out of the total number of words in each excerpt output. Now, all three modalities are aligned and prepared for data fusion. An excerpt of all modalities' results for the first five 2000ms chunks of the video can be seen in Figure 50.

2000ms	Emotient_Joy_MEAN	OpenSMILE_Joy	Positiv
0	0.079975062	0.005101	0
1	0.080801771	0.01554	20
2	0.033581019	0.016938	0
3	0.012505855	0.001286	15.38461538
4	0.02005936	0.003877	0

Figure 50 Excerpt of first five 2000ms chunks modalities outputs

The first step in this proposed data fusion process is to find the weighted sum of the facial expression and speech joy probabilities, output by Emotient FACET and OpenSMILE, respectively. This is done on a chunk-by-chunk basis and uses the formula in Figure 15. The results of applying this formula are seen in Figure 51 for the first five 2000ms chunks in the video again, with the '*S*' column being the result of adding the 'Emotient_Joy_MEAN' value and 'OpenSMILE_Joy' value together after each has been multiplied by a weight, being 0.8 and 0.2, respectively.

2000ms	Emotient_Joy_MEAN	OpenSMILE_Joy	S
0	0.079975062	0.005101	0.06500025
1	0.080801771	0.01554	0.067749417
2	0.033581019	0.016938	0.030252415
3	0.012505855	0.001286	0.010261884
4	0.02005936	0.003877	0.016822888

Figure 51 Extract of first five 2000ms chunks for results of data fusion step #1

The next step is to feed these results in column 'S' into the adjusted logical "OR" combined decision rule formula seen in Figure 44. If the value for 'S' is > 0.5 and the percentage of positive sentiment is > 0, then joy will be deemed a strongly evident emotion in the 2000ms chunk in question. The result of this combined rule will be a '1' or a '0' value, for strongly evident or not. The results of applying this formula for the first five 2000ms chunks in the video is seen in Figure 52, with the '1' or '0' decision depicted in the column '*d*'.

2000ms	Emotient_Joy_MEAN	OpenSMILE_Joy	S	Positiv	d
0	0.079975062	0.005101	0.06500025	0	0
1	0.080801771	0.01554	0.067749417	20	0
2	0.033581019	0.016938	0.030252415	0	0
3	0.012505855	0.001286	0.010261884	15.38461538	0
4	0.02005936	0.003877	0.016822888	0	0

Figure 52 Extract of first five 2000ms chunks for results of data fusion step #2

Figure 53 shows the results of this two-step data fusion process sorted in order of values largest to smallest by columns:

1. '*d*'
2. '*S*'
3. '*Positiv*'

hence, showing the results in order of which 2000ms chunk returns '1' for the combined decision rule in the second step of the data fusion process first, then, for the chunks that return '0', those that return highest results from the first step of the decision fusion process are shown next. Next, they're ordered by those that have the highest percentage of positive sentiment because both facial expressions and speech probabilities are embedded in the 'S' value, whereas sentiment is not.

2000ms	Emotient_Joy_MEAN	OpenSMILE_Joy	S	Positiv	d
72	0.659252267	0.018156	0.531033013	20	1
55	0.92449903	0.015805	0.742760224	0	0
73	0.899481865	0.015271	0.722639692	NA	0
84	0.71506052	0.00602	0.573252416	0	0
71	0.672719829	0.015913	0.541358463	0	0

Figure 53 Sorted results of data fusion process

By ordering the results it can be seen that only one 2000ms chunk in the video returns '1' on the combined decision rule. This is chunk number 72 in the video, which returns a mean probability of 0.6592 for joy on facial expression by Emotient FACET, and 0.018156 on speech by OpenSMILE. This probability of joy in speech is very low, with joy essentially not being detected, however, implementing the first step of the data

fusion process results in a weighted sum of both these probabilities being 0.5310. Given this weighted sum probability is > 0.5 and the percentage of positive sentiment in the 2000ms chunk is > 0 (20%), '1' is returned on the combined decision rule. Where 'NA' is seen under the 'Positiv' column, there were no complete words spoken in that 2000ms chunk, although incomplete audio may still be heard and therefore, is why an OpenSMILE joy probability is still output for this chunk.

Looking at this 2000ms chunk in more detail, the physical correlates of emotion in facial expression and speech can be examined. This 2000ms chunk spans frames 3600 to 3649 of the video, with three of the frames within that range than return a > 50% probability of joy by Emotient FACET displayed in Figure 54. The mean fundamental frequency across the entire video is plotted in Figure 55, with the red point being the mean F0 pertaining to this 2000ms chunk.

Figure 54 High probability of Joy frames (top to bottom: frame 3602, 3629, 3649)

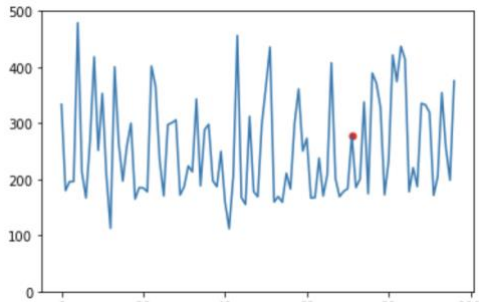

Figure 55 F0 (Hz) of Leo Varadkar video

Across these 50 frames, the mean probability for the AUs deemed 'most dominant' for joy in Section 4.5 (Facial Action Units) above, being AU6 and AU12, are 0.8519 and 0.6734, respectively, shown in Table 31. Both of these mean probabilities correspond to a > 50% probability of that AU being detected across the frames, which is what is expected when looking at what is characteristic for the physical correlates of joy in facial expression.

The mean F0 for this 2000ms chunk was 278Hz, seen in in Table 31 also, which lies above the overall mean of all mean F0s for each chunk in the corresponding audio for this video, being 255Hz. The standard deviation for this audio is 89Hz, resulting in a z-score of 0.2584 for this 2000ms chunk's mean F0. The mean F0 for this chunk, therefore, lies 0.2584th of a standard deviation above the mean. A high mean F0 implies the potential existence of joy in this chunk however, a high mean F0, as seen earlier, also relates to anger and therefore, it cannot be deduced that this mean F0 value directly corresponds to joy in the 2000ms chunk, especially because the probability of joy output by OpenSMILE for this chunk, seen in Figure 53, is so low.

The extremely low detection of joy in OpenSMILE could be attributed to the fact that OpenSMILE is trained on German speakers (EmoDB), among 14 other nationalities (eNTERFACE Database), none of which are Irish. In addition, F0 is only one of 988 acoustic features extracted by OpenSMILE and therefore, it cannot be solely relied on to detect emotion, but only examined to gain an insight into the physical correlates of speech

given the strong presence of an emotion in speech, with F0 being identified in previous literature as one of the most insightful acoustic cues of emotion. To examine the relationship between the physical correlates of emotion in facial expression and speech examined in this research, further analysis is conducted in the next section.

Mean AU6	Mean AU12	F0 (Hz)
0.851953487	0.673426885	278.2786

Table 31 Chunk 72 dominant Joy AU values and mean F0 (Hz)

4.8. AU & F0 Synthesis

As described in Section 3.8.2 (AU & F0 synthesis) above, the mean of all individual 2000ms chunks' mean F0s output by OpenSMILE for this Leo Varadkar dataset video where the probability of AUs output by Emotient FACET that are 'most dominant' to joy and anger, being AU6 and AU12 for joy, and AU4 and AU9 for anger, are computed and can be seen in Table 32. The overall mean F0 derived for where the probability of AU6 is > 50% is 256Hz. Where the probability of AU12 is > 50%, the overall mean F0 derived is 248Hz, and where the probability of AU4 and AU9 are > 50%, the overall mean F0s derived are 249Hz and 247Hz, respectively. The overall mean F0 across this entire audio is 255Hz, and therefore, only the mean F0 where AU6 is > 50% lies above the mean, suggesting the rest of these mean F0 values are not characteristic of joy and anger in speech, as they do not lie above the overall mean F0 for this audio. However, in order to better interpret these values, the overall mean of individual 2000ms chunks' mean F0 values where the probability of AU15 is > 50% is calculated. This is done because AU15 is an AU prototypical to the expression of sadness in facial expression, derived from literature, and therefore, is used as a baseline against which to compare results of mean F0 where joy and anger AUs are found > 50%. The overall mean F0 where AU15 is detected > 50% is 186Hz, which is much lower than the mean F0s derived for both the 'most dominant' joy AUs and anger AUs, which better reflects what is found in previous literature, that joy and anger have much higher mean F0 than sadness, which typically has low mean F0. The overall mean F0 values for AU12, AU4, and AU9 falling below the overall mean F0 of the audio, therefore, could be a function of the individuality of Leo Varadkar's F0 range, or reflective of the average emotion expressed in the audio. For example, if joy is the average emotion expressed in the audio, the overall mean F0 will be quite high to begin with and therefore, the mean F0 for instances where the AUs 'most dominant' to joy return > 50% probability will not be much higher than the overall mean F0. However, if the average emotion expressed was sadness, the overall mean F0 for joy's 'most dominant' AUs would be evidently much higher than the overall mean F0 for the entire audio. However, these insights only scratch the surface of a whole analysis and body of work to be carried out in the field of AER on this direct relationship between the physical correlates of emotion in facial expressions and speech, and therefore, much more work is needed to validate any inferences made.

		Mean F0 (Hz)
Joy	AU6	256 Hz
	AU12	248 Hz
Anger	AU4	249 Hz
	AU7	255 Hz
	AU9	247 Hz
Baseline (Sadness)	AU15	186 Hz

Table 32 Mean F0 where most dominant Joy and Anger AUs are >.5 for single dataset video

4.9. Summary

This chapter presented results derived from each experiment conducted on data output as a result of implementing each of the emotion recognition software systems and lexicon-based sentiment analysis methods outlined in Chapter 3 (Design Solution and Validation) above. These experiments included hypothesis tests such as the Kruskal-Wallis Rank Sum test and the Wilcoxon Rank Sum test examining if there is a significant difference between emotion detected in each software system and lexicon within each modality as well as between genders. Potential reasons for variances and similarities found were subsequently discussed. The 'most dominant' AUs for joy and anger were determined using Spearman's Rank Order Correlation and found to be AU6 and AU12 for joy, and AU4 and AU9 for anger. The mean F0 was examined with respect to emotions joy and anger, gender, and age. Findings were supportive of literature that joy and anger exert high mean F0 values however, results were less conclusive to previous literature for F0 variations across gender and age. Finally, the proposed novelty two-step decision-level data fusion process was illustrated on a single dataset video, and the physical correlates of emotion in facial expression and speech for this video examined also. The next chapter, being Chapter 5 (Conclusion and Future Work), will provide an overview of work carried out in this study and future work to be conducted that builds on this research.

Chapter 5

Conclusion and Future Work

5.1. Conclusion

This work presents a specification of a modality pipeline for multimodal emotion recognition using facial expression, speech, and text. Various software systems were implemented that provided the experimental apparatus to conduct subsequent analysis on the physical correlates of emotion in facial expression and speech from an Irish context. A novel two-step decision-level data fusion process is proposed that fuses emotion detected in facial expression with emotion detected in speech and supporting sentiment derived from text to determine if an emotion is strongly evident in a 2000ms video segment or not.

The three modalities used were facial expression, speech, and text, with emotions being extracted from the former two, and sentiment from the latter. Sentiment is used to enhance the determination of emotion as strongly evident, with emotions such as joy being solely associated with positive sentiment. The dataset used comprised of 50 videos of Irish politicians, 5 videos per politician, with 5 male politicians and 5 female politicians aged 35 to 62 from both Northern Ireland and the Republic of Ireland being included. Politicians were chosen as they bridge the gap between posed and spontaneous facial expressions, being semi-trained actors who have undergone media training. It is advised that systems be validated on both acted and spontaneous facial expressions to achieve more robust validation, with some facial expression emotion recognition software systems such as Affectiva AFFDEX being trained solely on a database of spontaneous facial expressions and therefore, although it outperforms Emotient FACET on more naturalistic facial expressions, it is less reliable when analysing posed/acted facial expressions. These politicians also bridge the gap between acted and spontaneous emotional speech, although it was found that the mean F0 follows the same behavioural characteristics across both acted and spontaneous speech [54]. Therefore, using only speech emotion recognition software systems trained on acted emotional speech databases in this study poses less of an issue than it would for facial expression recognition.

As the first stage in the modality pipeline, appropriate pre-processing steps were taken to transform the dataset video to the appropriate format for emotion recognition or sentiment analysis. Videos were edited, and frames extracted prior to facial expression emotion recognition processing using software systems Emotient FACET, Affectiva AFFDEX, and Microsoft Azure Face API. The videos were converted to audio file format and chunks extracted prior to speech emotion recognition processing using software systems OpenSMILE, OpenVokaturi, and DeepAffects. ASR services were used to transcribe audio files to text prior to conducting lexicon-based text sentiment analysis using text analytics tool RockSteady, which employs the General Inquirer lexicon, the NRC Emotion Lexicon, and the Opinion Lexicon. The ASR services used were Microsoft Azure STT, Google Cloud STT, and Amazon Transcribe however, the transcripts used for text sentiment analysis were those derived using Microsoft Azure STT, being the ASR service that achieved the lowest mean WER rate of 7.21%, and that closest to the WER of human parity, being 5.1%. A triad of software systems or lexicons were implemented at each stage of the pipeline in order to examine the calibration of each of these not fully-understood systems with respect to each other, this being another valuable contribution of this work.

Emotion probabilities are output by both the facial expression and speech emotion recognition software systems, with emotions joy and anger being focused on in this study, those being the two emotions detected above the level of chance by Affectiva AFFDEX. The percentage of positive and negative sentiment in text was derived through using lexicon-based sentiment analysis. The output from each of these software systems and lexicons implemented was subsequently analysed to investigate significant differences between joy or anger detected, or positive or negative sentiment detected in each. Results showed both differences and similarities between emotion detected in the facial expression emotion recognition systems and the speech emotion recognition systems, highlighting variances in the calibration of these software systems that process the same modalities. Such variances are a consequence of different databases, statistical processes and features being extracted by and used to train each software system to determine emotion. For facial expression emotion recognition systems, for example, both Emotient FACET and Affectiva AFFDEX determine emotion as a linear combination of AUs, which are first classified by the systems, however, they each extract a different number of facial landmarks subsequently used to classify these AUs, giving rise to variances among the classification of AUs based on these different sets of facial landmarks. The Microsoft Azure Face API additionally extracts a further set of different facial landmarks from facial expression. For speech emotion recognition systems, for example, OpenSMILE and OpenVokaturi are both trained on the EmoDB database however, they are each trained on an additional database that contains speakers of different nationalities, which may provide an explanation as to why there is a statistically significant difference between both joy detected and anger detected in each system. F0, for example, being one of the many acoustic features used to detect emotion in speech, has been shown to vary across nationalities and therefore, having different nationalities in the databases used to train different speech emotion recognition algorithms might result in acoustic cues being encoded by these systems differently depending on the nationalities in the dataset and thus, outputting different emotion results for the same speech. It can be seen, therefore, that the calibration of each of these systems has an impact on the manner in which they detect emotion. In the case of sentiment analysis lexicons, their variations are down to the means by which each lexicon was sourced, built and validated. For example, the General Inquirer lexicon used in RockSteady contains 3,486 words, while the NRC Emotion Lexicon contains 14,182 and the Opinion Lexicon contains 6,789. The vast difference in the size of each lexicon would lead to variances in sentiment detection using each as one may contain words annotated as positive or negative that aren't even included in another. These are all examples of how variances arise between emotion and sentiment detected across each of these systems and lexicons and potential causes of such, which are discussed in more detail in Chapter 4 (Case Studies and Results).

Within each modality, it was examined whether individual software systems and lexicons implemented detected emotion or sentiment differently across males and females. No statistically significant difference between joy detected in males and females was found for any of the facial expression emotion recognition systems however, a significant difference between anger detected in males and females was found for both Emotient FACET and the Microsoft Azure Face API. The gender breakdown of training databases used in these systems are not publicly disclosed however, and therefore, inferences cannot be made as to why such gender variances arose. For the speech emotion recognition systems, a significant difference between both joy and anger detected in males and females was found for OpenVokaturi, yet no significant difference between anger in males and females was found for OpenSMILE. Given the additional databases used to train these systems

contained solely male speakers for OpenVokaturi but both male and female speakers for OpenSMILE, this might provide an explanation as to why OpenVokaturi results show variances across genders, and not OpenSMILE however, this cannot be determined for certain as both are trained on the EmoDB which contains both male and female speakers. DeepAffects additionally shows variances in its results however, the gender breakdown of the training database used in this system is not available and hence, inferences about why this occurs are unable to be derived. Findings from previous studies in the literature showed a difference in emotions and sentiment typically expressed in text depending on the gender of the author [80, 81]. Results from this study showed a significant difference between both positive sentiment and negative sentiment detected in male transcripts and female transcripts for each sentiment lexicon implemented, which supports those findings from previous literature. What is interesting to note, however, is how some of the videos in the dataset are of speeches by politicians which could be scripted by an external author and therefore, future analysis could include determining the gender of the author, should it not be the politician in question, when looking at gender differences in text production in terms of positive and negative sentiment.

There exist contradictory findings in the literature regarding which AUs from Paul Ekman's FACS were prototypical to joy and anger, respectively. Results derived in this work found AU6 and AU12 to be the 'most dominant' for joy and AU4 and AU9 to be the 'most dominant' for anger. This supports what is published on the official iMotions website [3], which is the SDK supporting both Emotient FACET and Affectiva AFFDEX however, it contradicts the other studies' findings. It is important to note that these are the AUs found to be 'most dominant' for this particular dataset, being male and female Irish politicians. Paul Ekman stated the association between specific AUs and corresponding emotions to be universal [31], however, the contradiction across previous studies' results and now this study's results also, suggests the derivation of prototypical AUs for emotions may vary across ethnicities which would contradict Ekman's findings. However, these variances may also be a consequence of other factors, such as the experimental setting of the study, even, and therefore, a large study dedicated to investigating the prototypical AUs of various emotions in facial expressions of people of multiple ethnicities would have to be conducted to examine this notion further.

There exists a gap in the literature, also, around the mean F0 of Irish-English speakers. One study [57] found the mean F0 of Parisian-French speakers to be 234Hz for females and 133Hz for males and the mean F0 of American-English speakers to be 210Hz for females and 119Hz for males. Another more generally stated the mean F0 for females to range from 190Hz to 220Hz and for males to range from 120Hz to 130Hz [56]. The results from this study found the overall mean of all mean F0 values output by OpenSMILE for the 2000ms audio chunks pertaining to Irish-English male speakers in the dataset to be 308Hz and for audio chunks pertaining to Irish-English female speakers in the dataset to be 303Hz. These mean F0 values are significantly higher than what has been seen across literature which could be a consequence of the small sample size used in this study, being 5 male speakers and 5 female speakers, with only 25 audio files per gender. However, it could also be an inclination that the mean F0 of Irish speakers is higher than that of other nationalities, given it has also been seen in previous literature for mean F0 to vary across nationalities. Although a much larger study with more audio samples would need to be carried out to investigate this further.

These results saw the overall mean F0 for female Irish-English speakers to be lower than that of male Irish-English speakers, which contradicts findings for both Parisian-French speakers, American-English speakers, as well as Japanese speakers [58, 57]. It also contradicts scientific findings that the female mean F0 is

higher than the male mean F0 due to the variation in the size of the larynx across genders [56]. Therefore, it is assumed that the results of this study that saw the overall mean F0 of male Irish-English speakers to be higher than the overall mean F0 of female Irish-English speakers to be a consequence of the small sample size of the dataset, and not a nationality-based variance. On the other hand, this study also found the male mean F0 range of Irish-English speakers to be much wider than that of female Irish-English speakers, which is contradictory to findings for Parisian-French speakers and American-English speakers, but supportive of findings on Japanese speakers therefore, this could potentially be a nationality-based variance and is something that should be analysed on a much larger dataset of Irish-English speakers to determine for certain.

The overall mean F0 of Irish-English speakers where joy and anger were detected above the level of chance by OpenSMILE were also examined, finding the overall mean F0 where both of these emotions were detected above this level to be high, which is supportive of findings from previous literature.

The variation in mean F0 with respect to age in both male and female Irish-English speakers was also examined, with previous studies stating that the mean F0 for females decreases after the onset of menopause, and the mean F0 of males remains stationary between the ages of 35 and 55, after which it begins to increase again. Findings in this study on Irish-English speakers did not support those in previous literature for the relationship between the female mean F0 and age, with no clear trend seen in the results for a decrease in mean F0 around the age of menopause onset. Results did, however, appear to support the finding that the male mean F0 remains stationary between the ages of 35 to 55 on Irish-English male speakers also, although no trend was seen for an increase in mean F0 after the age of 55. Again, the small sample size of this dataset is a limitation of this study that may cause inadequate findings to be derived in terms of the variation of mean F0 with respect to age for both male and female speakers.

Upon the output of results from each software system, post-processing steps were undertaken to align the three modalities for data fusion. Decision-level data fusion was implemented using facial expression emotion probabilities, speech emotion probabilities, and text sentiment percentages on a 2000ms chunk basis. There exist many decision-level data fusion techniques in previous literature however, none covered the fusion of the specific combination of facial expression emotion, speech emotion, and text sentiment. Therefore, a two-step data fusion process was proposed that first fuses facial expression and speech emotion probabilities using an existing weighted sum of probabilities equation found in previous literature [22], the result of which is subsequently input into a logical “OR” formula using a combined decision rule, adjusted from literature [113], to determine an emotion as strongly evident or not for each 2000ms chunk of a video dependent on the fused facial expression and speech emotion probability and the percentage of associated sentiment detected. This process was illustrated using joy probabilities output by Emotient FACET and OpenSMILE, and the percentage of positive sentiment output by RockSteady on one dataset video, finding only one 2000ms chunk of this video to determine joy as strongly evident using this decision-level data fusion process.

The mean probabilities of the AUs determined ‘most dominant’ for joy, being AU6 and AU12, output by Emotient FACET, and the mean F0 value, output by OpenSMILE, for the 2000ms chunk of the dataset that determined joy as strongly evident as per the data fusion process were next examined. This was done to investigate if the physical correlates of joy examined in this study were characteristic of what was expected of them for this 2000ms chunk. It was seen that for both AU6 and AU12, the mean probabilities were > 50%, and the mean F0 higher than the overall mean F0 for the entire audio of this dataset video. Therefore, it can be

deduced that both the physical correlates of emotion for facial expression and speech were characteristic of what is expected of them when joy is present.

The direct relationship between these physical correlates of emotion was then examined for both joy and anger across the entire video, where it was seen that where the AUs 'most dominant' to joy and anger, respectively, were detected above 50% probability, the mean F0 was high. A baseline comparison was derived by computing the overall mean F0 for where AU15 was detected above 50% probability, with AU15 being a prototypical AU for expressions of sadness, derived from literature. This returned a low overall mean F0 value, which was subsequently used to consolidate the 'high'-ness of the overall mean F0 values found where joy and anger AUs were detected above 50% probability. This indicates that a direct relationship between these physical correlates of emotion exists however, it was deduced that a much larger body of work would be required in order to derive more substantial conclusions.

One important thing to note when looking at the physical correlates of emotion throughout this research is that the mean F0 is focused on solely because it has been found to be 'most insightful' [5] in detecting emotion from speech and therefore, due to the scope of this work, it was used to illustrate the physical correlates of joy and anger in speech. However, there are 988 other features extracted by OpenSMILE, for example, that are used to detect emotion and therefore, it should not be relied on entirely. Similarly, emotion is detected in facial expressions by applying a linear combination of AUs and therefore, the AUs identified as 'most dominant' in this study are not to be understood as the only AUs used to detect joy and anger, but moreover those that carry the most weight in this linear combination of all AUs when determining joy or anger to be present in a facial expression.

5.2. Future Work

The modality pipeline specified in this work is replicable as only pre-built software systems, APIs and sentiment lexicons are implemented. This means no extensive training needs to be performed for the ML algorithms implemented, cutting out considerable time and complexity associated with building the emotion recognition systems from scratch. In addition, the proposed data-fusion process, being the only novel element of this pipeline, is explained in great detail, hence making it replicable in future studies as well. In this study, there was a focus on the physical correlates of emotion in facial expression and particularly in speech from an Irish context. This work is replicable for the application of data from subjects of other nationalities to examine variations in these physical correlates of emotions on the basis of nationality as well. Other areas of future work are outlined below.

The results of the various transcription services are seen in Table 4. These WER scores could be improved through use of automatic multi-language recognition by specifying both 'en-IE', being Irish-English, and 'ga-IE', being Irish, as languages spoken in the audio. Microsoft Azure STT supports the recognition of ga-IE however, it does not yet provide multi-language recognition capabilities. Google Cloud STT does have multi-language recognition capabilities however, it does not yet support the recognition of ga-IE. An area of future work would be to avail of this multi-language recognition using both en-IE and ga-IE once it becomes available. Therefore, terms and place names in the Irish language used throughout speech primarily in the English language of Irish politicians and other Irish nationals may be recognised correctly, resulting in an improvement in overall transcription accuracy.

The two-step decision-level data fusion process detailed in Section 4.7 (Data Fusion) above was illustrated using joy probabilities from facial expression and speech, and positive sentiment percentages from text. The combined decision rule was based on a combination of the weighted sum of facial expression and speech probabilities for joy being > 0.5 and the percentage of positive sentiment for the same 2000ms chunk being > 0. If both conditions are met, joy was determined as a strongly evident emotion for that 2000ms segment. What wasn't considered in this decision, however, was if the percentage of negative sentiment was > 0, or greater than the percentage of positive sentiment. Incorporating this could have a very beneficial impact on the accurate detection of emotion, as discussed in Section 2.3 (Text) above. A previous study discussed in this section determined that some emotions, such as surprise, can express both positive or negative sentiment e.g. positively-surprised or negatively-surprised [77]. Therefore, incorporating negative sentiment into this combined decision rule is an area of future work that builds upon the current proposed decision-level data fusion process in order to derive a more comprehensive emotion estimate for each 2000ms chunk.

Another area of future work would be to use machine-learning based sentiment analysis instead of lexicon-based sentiment analysis as was implemented in this study. By using machine-learning based sentiment analysis, probability scores would be output for positive and negative sentiment in a text excerpt, resulting in all three modalities' results being obtained as probabilities. With each of the three modalities' results being in the form of probabilities per 2000ms chunk of a video, different decision-level data fusion formulas can be experimented with, such as those in [116] and [114], which each use a weighted sum of probabilities to derive an overall emotion probability incorporating all three modalities. The decision-level data fusion equation proposed in [116], seen in Figure 17, uses facial expression sentiment, speech sentiment, and text sentiment, while the equation proposed in [114], seen in Figure 18, uses speech emotion, text emotion, and text sentiment. Therefore, should machine-learning based text sentiment analysis be implemented, these equations can be experimented with and adjusted to conduct decision-level data fusion on facial expression emotion, speech emotion, and text sentiment, using probabilities derived in the modality pipeline.

Another interesting area of future work for this research is to take the physical correlates of emotion output by the various emotion recognition software systems, such as AUs for facial expression by Emotient FACET and the mean F0 for speech output by OpenSMILE, and use these directly to implement feature-level fusion in a bimodal emotion recognition system. Building on this, the equation put forward by Pelachaud et al. [112], seen in Figure 10, could be experimented with as part of the feature engineering process for this feature-level fusion to overcome the limitation of bimodal facial expression and speech emotion recognition, being the impediment to a person's facial expression as a result of them speaking. Audio analysis tool Praat can be used to extract the speech-rate from a piece of audio, which can be subsequently input into this equation to determine the true intensity of each AU proportional to the speech rate, which is later fed to the feature-level fusion.

Finally, to determine the accuracy of the novel decision-level data fusion process proposed in this study, this pipeline could be applied to a dataset of videos pre-labelled with emotions evident in order to establish the accuracy of determining emotions as strongly evident by this fusion method. Alternatively, the videos could be assessed manually by a number of human judges however, should this method be undertaken, the extent to which individual judges agree on the whether each emotion is strongly evident or not in each 2000ms chunk of video should also be examined.

Bibliography

[1] S. Khanal, A. Reis, J. Barroso and V. Filipe, "Using emotion recognition in intelligent interface design for elderly care," in *Trends and Advances in Information Systems and Technologies*, Springer, 2018, pp. 240-247.

[2] N. Ayari, H. Abdelkawy, A. Chibani and Y. Amirat, "Towards semantic multimodal emotion recognition for enhancing assistive services in ubiquitous robotics," *2017 AAAI Fall Symposium Series,* 2017.

[3] A. Mehta, C. Sharma, M. Kanala, M. Thakur, R. Harrison and D. D. Torrico, "Self-reported emotions and facial expressions on consumer acceptability: A study using energy drinks," *Foods,* vol. 10, no. 2, p. 330, 2021.

[4] J. F. Cohn, Z. Ambadar and P. Ekman, "Observer-based measurement of facial expression with the Facial Action Coding System," *The handbook of emotion elicitation and assessment,* vol. 1, no. 3, pp. 203-221, 2007.

[5] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman and T. J. Sejnowski, "Classifying facial actions," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 21, no. 10, pp. 974-989, 1999.

[6] B. Farnsworth, "Facial Action Coding System (FACS) – A Visual Guidebook," 18 August 2019. [Online]. Available: https://imotions.com/blog/facial-action-coding-system/. [Accessed 4 August 2021].

[7] . K. R. Scherer, J. Sundberg, B. Fantini, S. Trznadel and F. Eyben, "The expression of emotion in the singing voice: Acoustic patterns in vocal performance," *The Journal of the Acoustical Society of America,* vol. 142, no. 4, pp. 1805-1815, 2017.

[8] C. Pereira and C. Watson, "Some acoustic characteristics of emotion," *Fifth International Conference on Spoken Language Processing,* 1998.

[9] M. El Ayadi, M. S. Kamel and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern recognition,* vol. 44, no. 3, pp. 572-587, 2011.

[10] S. Poria, E. Cambria, A. Hussain and G.-B. Huang, "Towards an intelligent framework for multimodal affective data analysis," *Neural Networks,* vol. 63, pp. 104-116, 2015.

[11] K. Fridkin and S. A. Gershon, "Nothing More than Feelings? How Emotions Affect Attitude Change during the 2016 General Election Debates," *Political Communication,* vol. 38, no. 4, pp. 370-387, 2021.

[12] I. Broch-Due, H. L. Kjærstad, L. V. Kessing and K. Miskowiak, "Subtle behavioural responses during negative emotion reactivity and down-regulation in bipolar disorder: A facial expression and eye-tracking study," *Psychiatry research,* vol. 266, pp. 152-159, 2018.

[13] E. B. Setiawan, "Song Recommendation Application Using Speech Emotion Recognition," *IJID (International Journal on Informatics for Development)* , vol. 10, no. 1, pp. 15-22, 2021.

[14] A. Agarwal and D. Toshniwal, "Application of lexicon based approach in sentiment analysis for short tweets," *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE),* pp. 189-193, 2018.

[15] 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, 2017.

[16] E. A. Clark, J. Kessinger, S. E. Duncan, M. A. Bell, J. Lahne, D. L. Gallagher and S. F. O'Keefe, “The Facial Action Coding System for Characterization of Human Affective Response to Consumer Product-Based Stimuli: A Systematic Review.,” *Frontiers in psychology,* vol. 11, p. 920, 2020.

[17] S. G. Koolagudi and K. S. Rao, “Emotion recognition from speech: a review,” *International journal of speech technology,* vol. 15, no. 2, pp. 99-117, 2012.

[18] S. Gu, F. Wang, J. A. Bourgeois, J. H. Huang and N. P. Patel, “A model for basic emotions using observations of behavior in Drosophila.,” *Frontiers in psychology,* vol. 10, p. 781, 2019.

[19] L. Barrett, R. Adolphs, S. Marsella, A. Martinez and S. Pollak, “Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements.,” *Psychological science in the public interest,* vol. 20, no. 1, pp. 1-68, 2019.

[20] M. Taboada, “Sentiment analysis: An overview from linguistics.,” *Annual Review of Linguistics,* vol. 2, pp. 325-347, 2016.

[21] M. Munezero, C. S. Montero, E. Sutinen and J. Pajunen, “Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text,” *IEEE transactions on affective computing,* vol. 5, no. 2, pp. 101-111, 2014.

[22] J. H. Lui, H. Samani and K.-Y. Tien, “An affective mood booster robot based on emotional processing unit,” *2017 International Automatic Control Conference (CACS),* pp. 1-6, 2017.

[23] A. Kennedy and D. Inkpen, “Sentiment classification of movie reviews using contextual valence shifters,” *Computational intelligence,* vol. 22, no. 2, pp. 110-125, 2006.

[24] D. Datcu and L. J. Rothkrantz, “Emotion recognition using bimodal data fusion,” *Proceedings of the 12th International Conference on Computer Systems and Technologies,* pp. 122-128, 2011.

[25] X. Wang, X. Chen and C. Cao, “Human emotion recognition by optimally fusing facial expression and speech feature,” *Signal Processing: Image Communication,* vol. 84, p. 115831, 2020.

[26] B. T. Atmaja, K. Shirai and M. Akagi, “Speech emotion recognition using speech feature and word embedding,” *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC),* pp. 519-523, 2019.

[27] D. Bertero, F. B. Siddique, C.-S. Wu, Y. Wan, R. H. Y. Chan and P. Fung, “Real-time speech emotion and sentiment recognition for interactive dialogue systems,” *roceedings of the 2016 conference on empirical methods in natural language processing,* pp. 1042-1047, 2016.

[28] V. Rozgić, S. Ananthakrishnan, S. Saleem, R. Kumar and R. Prasad, “Ensemble of svm trees for multimodal emotion recognition,” *Proceedings of the 2012 Asia Pacific signal and information processing association annual summit and conference,* pp. 1-4, 2012.

[29] P. Ekman, “Darwin's contributions to our understanding of emotional expressions.,” *Philosophical Transactions of the Royal Society B: Biological Sciences,* vol. 364, no. 1535, pp. 3449-3451, 2009.

[30] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot and R. el Kaliouby, "AFFDEX SDK: a cross-platform real-time multi-face expression recognition toolkit.," *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems,* pp. 3723-3726, 2016.

[31] M. Del Líbano, M. G. Calvo, A. Fernández-Martín and G. Recio, "Discrimination between smiling faces: Human observers vs. automated face analysis," *Acta psychologica ,* vol. 187, pp. 19-29, 2018.

[32] L. Kulke, D. Feyerabend and A. Schacht, "A comparison of the Affectiva iMotions Facial Expression Analysis Software with EMG for identifying facial expressions of emotion.," *Frontiers in psychology,* vol. 11, p. 329, 2020.

[33] P. Ekman and D. Keltner, "Universal facial expressions of emotion," *Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture,* vol. 27, p. 46, 1997.

[34] S. Du, Y. Tao and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences ,* vol. 111, no. 15, pp. E1454-E1462, 2014.

[35] M. Ghayoumi and A. K. Bansal, "Unifying geometric features and facial action units for improved performance of facial expression analysis," *arXiv preprint arXiv:1606.00822,* 2016.

[36] T. Xu, J. White, S. Kalkan and H. Gunes, "Investigating bias and fairness in facial expression recognition," *European Conference on Computer Vision,* pp. 506-523, 2020.

[37] R. Zhi, M. Liu and D. Zhang, "A comprehensive survey on automatic facial action unit analysis," *The Visual Computer,* vol. 36, no. 5, pp. 1067-1093, 2020.

[38] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern recognition,* vol. 36, no. 1, pp. 259-275, 2003.

[39] E. Sariyanidi, H. Gunes, M. Gökmen and A. Cavallaro, "Local Zernike Moment Representation for Facial Affect Recognition," *BMVC,* vol. 2, p. 3, 2013.

[40] T. Senechal, D. McDuff and R. Kaliouby, "Facial action unit detection using active learning and an efficient non-linear kernel approximation," *Proceedings of the IEEE International Conference on Computer Vision Workshops,* pp. 10-18, 2015.

[41] M. S. Bartlett, J. C. Hager, P. Ekman and T. J. Sejnowski, "Measuring facial expressions by computer image analysis," *Psychophysiology,* vol. 36, no. 2, pp. 253-263, 1999.

[42] Y.-I. Tian, T. Kanade and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on pattern analysis and machine intelligence,* vol. 23, no. 2, pp. 97-115, 2001.

[43] M. S. Bartlett, J. R. Movellan and T. J. Sejnowski, "Face recognition by independent component analysis," *IEEE Transactions on neural networks,* vol. 13, no. 6, pp. 1450-1464, 2002.

[44] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel and J. R. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *J. Multim,* vol. 1, no. 6, pp. 22-35, 2006.

[45] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan and M. Bartlett, "The computer expression recognition toolbox (CERT)," *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG),* pp. 298-305, 2011.

[46] T. Baltrušaitis, P. Robinson and L.-P. Morency, “Openface: an open source facial behavior analysis toolkit,” *2016 IEEE Winter Conference on Applications of Computer Vision (WACV),* pp. 1-10, 2016.

[47] S. Stöckli, M. Schulte-Mecklenbeck, S. Borer and A. C. Samson, “Facial expression analysis with AFFDEX and FACET: A validation study,” *Behavior research methods,* vol. 50, no. 4, pp. 1446-1460, 2018.

[48] D. Dupré, E. Krumhuber, D. Küster and G. McKeown, “Emotion recognition in humans and machine using posed and spontaneous facial expression,” *PsyArXiv [Preprint],* 2019.

[49] J. Lei, J. Sala and S. K. Jasra, “Identifying correlation between facial expression and heart rate and skin conductance with iMotions biometric platform,” *Journal of Emerging Forensic Sciences Research ,* vol. 2, no. 2, pp. 53-83, 2017.

[50] O. M. Al-Omair and S. Huang, “A comparative study on detection accuracy of cloud-based emotion recognition services,” *Proceedings of the 2018 International Conference on Signal Processing and Machine Learning,* pp. 142-148, 2018.

[51] S. Cosentino, E. I. Randria, J.-Y. Lin, T. Pellegrini, S. Sessa and A. Takanishi, “Group emotion recognition strategies for entertainment robots,” *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),* pp. 813-818, 2018.

[52] A. Paeschke, M. Kienast and W. F. Sendlmeier, “F0-contours in emotional speech,” *Proc. 14th Int. Congress of Phonetic Sciences,* vol. 2, pp. 929-932, 1999.

[53] M. K. Pichora-Fuller, K. Dupuis and P. V. Lieshout, “Importance of F0 for predicting vocal emotion categorization.,” *The Journal of the Acoustical Society of America,* vol. 140, no. 4, pp. 3401-3401, 2016.

[54] L. Z. Stan and A. Jain, “Fundamental Frequency, Pitch, F0,” in *Encyclopedia of Biometrics*, Springer, 2009, p. 26.

[55] P. N. Juslin and K. R. Scherer, Vocal expression of affect, Oxford University Press, 2005.

[56] P. Torre III and J. A. Barlow, “Age-related changes in acoustic characteristics of adult speech,” *Journal of communication disorders,* vol. 42, no. 5, pp. 324-333, 2009.

[57] E. Pépiot, “Male and female speech: a study of mean f0, f0 range, phonation type and speech rate in Parisian French and American English speakers.,” *Speech Prosody,* vol. 7, pp. 305-309, 2014.

[58] C. Guillemot and S.-i. Sano, “Gender-and register-biased patterns in f0 use: How does prosody contribute to (in) formality in Japanese?,” *10th International Conference on Speech Prosody, 725,* vol. 729, pp. 2020-148, 2020.

[59] G. Saggio and G. Costantini, “Worldwide healthy adult voice baseline parameters: a comprehensive review,” *Journal of Voice,* 2020.

[60] K. Järvinen, A.-M. Laukkanen and O. Aaltonen, “Speaking a foreign language and its effect on F0,” *Logopedics Phoniatrics Vocology ,* vol. 38, no. 2, pp. 47-51, 2013.

[61] J. T. Eichhorn, R. D. Kent, D. Austin and H. K. Vorperian, “Effects of aging on vocal fundamental frequency and vowel formants in men and women,” *Journal of Voice,* vol. 32, no. 5, pp. 644-e1, 2018.

[62] H. Traunmüller and A. Eriksson, “The frequency range of the voice fundamental in the speech of male and female adults,” *Unpublished manuscript,* 1995.

[63] A. Dorn and A. Ní Chasaide, “Effects of Focus on f0 across Four Varieties of Donegal Irish,” *focus,* vol. 12, p. 13, 2018.

[64] M. O'Reilly, A. Dorn and A. Ní Chasaide, “Focus in Donegal Irish (Gaelic) and Donegal English bilinguals,” *Speech Prosody 2010-Fifth International Conference,* 2010.

[65] J. Sullivan, “Variability in F0 valleys: the case of Belfast English,” *CamLing 2007: Proceedings of the Fifth University of Cambridge Postgraduate Conference in Language Research,* pp. 245-252, 2007.

[66] S. Ramakrishnan and I. M. El Emary, “Speech emotion recognition approaches in human computer interaction,” *Telecommunication Systems,* vol. 52, no. 3, pp. 1467-1478, 2013.

[67] Q. Jin, C. Li, S. Chen and H. Wu, “Speech emotion recognition with acoustic and lexical features,” *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP),* pp. 4749-4753, 2015.

[68] M. B. Akçay and K. Oğuz, “Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers,” *Speech Communication,* vol. 116, pp. 56-76, 2020.

[69] H. Cao, R. Verma and A. Nenkova, “Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech,” *Computer speech & language,* vol. 29, no. 1, pp. 186-202, 2015.

[70] Y. Pan, P. Shen and L. Shen, “Speech emotion recognition using support vector machine,” *International Journal of Smart Home,* vol. 6, no. 2, pp. 101-108, 2012.

[71] P. Harár, R. Burget and M. K. Dutta, “Speech emotion recognition with deep learning,” *2017 4th International Conference on Signal Processing and Integrated Networks (SPIN),* pp. 137-140, 2017.

[72] F. Eyben, M. Wöllmer and B. Schuller, “OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit,” *2009 3rd international conference on affective computing and intelligent interaction and workshops,* pp. 1-6, 2009.

[73] F. Eyben and B. Schuller, “openSMILE:) The Munich open-source large-scale multimedia feature extractor,” *ACM SIGMultimedia Records,* vol. 6, no. 4, pp. 4-13, 2015.

[74] L. Burzagli and S. Naldini, “Affective Computing and Loneliness: How This Approach Could Improve a Support System,” *International Conference on Human-Computer Interaction,* pp. 493-503, 2020.

[75] J. M. Garcia-Garcia, V. M. Penichet and M. D. Lozano, “Emotion detection: a technology review,” *Proceedings of the XVIII international conference on human computer interaction,* pp. 1-8, 2017.

[76] J. R. Smith, D. Joshi, B. Huet, W. Hsu and J. Cota, “Harnessing ai for augmenting creativity: Application to movie trailer creation,” *Proceedings of the 25th ACM international conference on Multimedia,* pp. 1799-1808, 2017.

[77] K. Sailunaz and R. Alhajj, “Emotion and sentiment analysis from Twitter text,” *Journal of Computational Science,* vol. 36, 2019.

[78] S. M. Mohammad, “Sentiment analysis: Detecting valence, emotions, and other affectual states from text,” in *Emotion measurement*, Woodhead Publishing, 2016, pp. 201-237.

[79] M. S. Akhtar, D. S. Chauhan, D. Ghosal, S. Poria, A. Ekbal and P. Bhattacharyya, “Multi-task learning for multi-modal emotion recognition and sentiment analysis,” *arXiv preprint arXiv:1905.05812,* 2019.

[80] S. Laxe, J. Saurí, M. B. Guitart and A. Garcia-Rudolph, “Stroke survivors on Twitter: sentiment and topic analysis from a gender perspective,” *Journal of medical Internet research,* vol. 21, no. 8, p. e14077, 2019.

[81] S. M. Mohammad, “Tracking sentiment in mail: How genders differ on emotional axes,” *arXiv preprint,* vol. arXiv:1309.6347, 2013.

[82] W. Medhat, A. Hassan and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams engineering journal,* vol. 5, no. 4, pp. 1093-1113, 2014.

[83] A. Ortigosa, J. M. Martín and R. M. Carro, “Sentiment analysis in Facebook and its application to e-learning,” *Computers in human behavior,* vol. 31, pp. 527-541, 2014.

[84] A. Moniz and F. de Jong, “Sentiment analysis and the impact of employee satisfaction on firm earnings,” in *Advances in Information Retrieval* , Springer, pp. 519-527.

[85] J. A. Cook and K. Ahmad, “Behaviour and Markets: The Interaction Between Sentiment Analysis and Ethical Values?,” in *International Conference on Intelligent Data Engineering and Automated Learning,* Springer, 2015, pp. 551-558.

[86] Z. Zhao, S. Kelly and K. Ahmad, “Finding sentiment in noise: Non-linear relationships between sentiment and financial markets,” in *ernational Conference on Intelligent Data Engineering and Automated Learning,* Springer, 2017, pp. 580-591.

[87] P. Gaag, D. Granvogl, R. Jackermeier, F. Ludwig, J. Rosenlöhner and A. Uitz, “FROY: exploring sentiment-based movie recommendations,” *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia,* pp. 345-349, 2015.

[88] B. Ohana and B. Tierney, “Sentiment classification of reviews using SentiWordNet,” *Proceedings of IT&T,* 2009.

[89] A. Agarwal, V. Sharma, G. Sikka and R. Dhir, “Opinion mining of news headlines using SentiWordNet,” *2016 Symposium on Colossal Data Analysis and Networking (CDAN),* pp. 1-5, 2016.

[90] C. S. Khoo and S. B. Johnkhan, “Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons,” *Journal of Information Science,* vol. 44, no. 4, pp. 491-511, 2018.

[91] A. Sharma and S. Dey, “Performance investigation of feature selection methods and sentiment lexicons for sentiment analysis,” *IJCA Special Issue on Advanced Computing and Communication Technologies for HPC Applications,* vol. 3, pp. 15-20, 2015.

[92] Docsoft, “What is Automatic Speech Recognition?,” June 2009. [Online]. Available: http://docsoft.com/Resources/Studies/Whitepapers/whitepaper-ASR.pdf. [Accessed 15 June 2021].

[93] S. Sahu, V. Mitra, N. Seneviratne and C. Y. Espy-Wilson, “Multi-Modal Learning for Speech Emotion Recognition: An Analysis and Comparison of ASR Outputs with Ground Truth Transcription,” *Interspeech,* pp. 3302-3306, 2019.

[94] V. Rozgić, S. Ananthakrishnan, S. Saleem, R. Kumar, A. Vembu and R. Prasad, “Emotion recognition using acoustic and lexical features.,” *Thirteenth Annual Conference of the International Speech Communication Association.,* 2012.

[95] B. Xu, C. Tao, Z. Feng, Y. Raqui and S. Ranwez, “A Benchmarking on Cloud based Speech-To-Text Services for French Speech and Background Noise Effect.,” *arXiv preprint arXiv:2105.03409,* 2021.

[96] G. Kallirroi, A. Leuski, V. Yanov and D. Traum, “Evaluation of Off-the-shelf Speech Recognizers Across Diverse Dialogue Domains.,” *Proceedings of The 12th Language Resources and Evaluation Conference ,* pp. 6469-6476, May 2020.

[97] A. Besim, N. Prljača, M. Kimmel and M. Schultalbers, “Speech recognition system for a service robot-a performance evaluation,” *2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV) ,* pp. 1171-1176, December 2020.

[98] C. Liu, J. Y. Kim, R. A. Calvo, K. McCabe, S. C. Taylor, B. W. Schuller and K. Wu, “A comparison of online automatic speech recognition systems and the nonverbal responses to unintelligible speech.,” *arXiv preprint arXiv:1904.12403,* 2019.

[99] V. Këpuska and G. Bohouta, “Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx).,” *Int. J. Eng. Res. Appl,* vol. 7(03), pp. 20-24, 2017.

[100] A. Woollacott and V. I. Design, “Benchmarking Speech Technologies,” vol. 2021.

[101] S. J. Arora and R. P. Singh, “Automatic speech recognition: a review.,” *International Journal of Computer Applications,* vol. 60(9), 2012.

[102] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris and R. Rose, “Automatic speech recognition and speech variability: A review.,” *Speech communication,* Vols. 49(10-11), pp. 763-786, 2007.

[103] M. Namazifar, J. Malik, L. Li, G. Tur and D. Tür, “Correcting Automated and Manual Speech Transcription Errors using Warped Language Models.,” *arXiv preprint arXiv:2103.14580.,* 2021.

[104] F. Metze, A. Batliner, F. Eyben, T. Polzehl, B. Schuller and S. Steidl, “Emotion recognition using imperfect speech recognition.,” 2010.

[105] A. Batliner, D. Seppi, S. Steidl and B. Schuller, “Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach.,” *Advances in Human-Computer Interaction, 2010.,* 2010.

[106] V. Pérez-Rosas and R. Mihalcea, “Evaluating Automatic Speech Recognition Quality and Its Impact on Counselor Utterance Coding.,” *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access,* pp. 159-168, June 2021.

[107] F. Catania, P. Crovari, M. Spitale and F. Garzotto, “Automatic Speech Recognition: Do Emotions Matter?,” *2019 IEEE International Conference on Conversational Data & Knowledge Engineering (CDKE),* pp. 9-16, December 2019.

[108] S. Nemati, “Canonical correlation analysis for data fusion in multimodal emotion recognition.,” *2018 9th International Symposium on Telecommunications (IST),* pp. 676-681, December 2018.

[109] M. Pantic, N. Sebe, J. Cohn and T. Huang, “Affective multimodal human-computer interaction.,” *Proceedings of the 13th annual ACM international conference on Multimedia,* pp. 669-676, November 2005.

[110] C. Wu, J. Lin and W. Wei, “Survey on audiovisual emotion recognition: databases, features, and data fusion strategies.,” *APSIPA transactions on signal and information processing,* vol. 3, 2014.

[111] M. Shah, D. G. Cooper, H. Cao, R. C. Gur, A. Nenkova and R. Verma, “Action unit models of facial expression of emotion in the presence of speech,” *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction,* pp. 49-54, 2013.

[112] C. Pelachaud, N. I. Badler and M. Steedman, “Generating facial expressions for speech,” *Cognitive science,* vol. 20, no. 1, pp. 1-46, 1996.

[113] C. M. Lee, S. S. Narayanan and R. Pieraccini, “Combining acoustic and language information for emotion recognition,” *Seventh international conference on spoken language processing,* 2002.

[114] O. Verkholyak, A. Dvoynikova and A. Karpov, “A Bimodal Approach for Speech Emotion Recognition using Audio and Text,” *J. Internet Serv. Inf. Secur,* vol. 11, no. 1, pp. 80-96, 2021.

[115] S. Planet and I. Iriondo, “Comparison between decision-level and feature-level fusion of acoustic and linguistic features for spontaneous emotion recognition,” *7th Iberian Conference on Information Systems and Technologies (CISTI 2012),* pp. 1-6, 2012.

[116] S. Poria, E. Cambria, N. Howard, G.-B. Huang and A. Hussain, “Fusing audio, visual and textual clues for sentiment analysis from multimodal content,” *Neurocomputing,* vol. 174, pp. 50-59, 2016.

[117] J. L. Tracey and R. W. Robins, “The Automaticity of Emotion Recognition,” *Emotion,* vol. 8, no. 1, pp. 81-95, 2008.

[118] S. Rigoulot, E. Wassiliwizky and M. Pell, “Feeling backwards? How temporal order in speech affects the time course of vocal emotion recognition.,” *Frontiers in psychology,* vol. 4, p. 367, 2013.

[119] B. Schuller and G. Rigoll, “Timing levels in segment-based speech emotion recognition,” *Proc. INTERSPEECH 2006, Proc. Int. Conf. on Spoken Language Processing ICSLP, Pittsburgh, USA,* 2006.

[120] M. D. Pell and S. A. Kotz, “On the time course of vocal emotion recognition,” *PLoS One,* vol. 6, no. 11, 2011.

[121] “Amazon Transcribe,” [Online]. Available: https://aws.amazon.com/transcribe/. [Accessed 16 June 2021].

[122] “Speech to Text,” [Online]. Available: https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/. [Accessed 16 June 2021].

[123] “Speech-to-Text,” [Online]. Available: https://cloud.google.com/speech-to-text. [Accessed 16 June 2021].

[124] “What is Amazon Transcribe?,” [Online]. Available: https://docs.aws.amazon.com/transcribe/latest/dg/what-is-transcribe.html. [Accessed 16 June 2021].

[125] trevorbye, J. Wells, J. Dempsey and D. Berry, “Get started with speech-to-text,” 15 September 2020. [Online]. Available: https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/get-started-speech-to-text?tabs=windowsinstall&pivots=programming-language-python. [Accessed 16 June 2021].

[126] “Transcribing long audio files,” [Online]. Available: https://cloud.google.com/speech-to-text/docs/async-recognize. [Accessed 16 June 2021].

[127] O. Martin, I. Kotsia, B. Macq and I. Pitas, “The eNTERFACE'05 audio-visual emotion database,” *22nd International Conference on Data Engineering Workshops (ICDEW'06),* 2006.

[128] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, “The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression,” *2010 ieee computer society conference on computer vision and pattern recognition-workshops,* pp. 94-101, 2010.

[129] M. Beringer, F. Spohn, A. Hildebrandt, J. Wacker and G. Recio, “Reliability and validity of machine vision for the assessment of facial expressions,” *Cognitive Systems Research,* vol. 56, pp. 119-132, 2019.

[130] Microsoft, “What is the Azure Face service?,” 19 April 2021. [Online]. Available: https://docs.microsoft.com/en-us/azure/cognitive-services/face/overview. [Accessed 16 August 2021].

[131] P. Farley, J. Wells, T. Christiani and P. Thurman, “Face detection and attributes,” Microsoft, 26 April 2019. [Online]. Available: https://docs.microsoft.com/en-us/azure/cognitive-services/face/concepts/face-detection. [Accessed 16 August 2021].

[132] P. Farley, J. Wells, W. Pickett and K. Brockschmidt, “Quickstart: Use the Face client library,” Microsoft, 25 May 2021. [Online]. Available: https://docs.microsoft.com/en-us/azure/cognitive-services/face/quickstarts/client-libraries?tabs=visual-studio&pivots=programming-language-python. [Accessed 16 August 2021].

[133] C. Hausner, “openSMILE 3.0,” 21 October 2020. [Online]. Available: https://github.com/audeering/opensmile/releases. [Accessed 16 August 2021].

[134] audEERING, “Get started,” [Online]. Available: https://audeering.github.io/opensmile/get-started.html#. [Accessed 16 August 2021].

[135] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier and B. Weiss, “A database of German emotional speech,” *Ninth european conference on speech communication and technology,* 2005.

[136] Vokaturi, “Overview,” [Online]. Available: https://developers.vokaturi.com/getting-started/overview. [Accessed 16 August 2021].

[137] J. Philip and S. ul haq, *Surrey Audio-Visual Expressed Emotion (SAVEE) database,* 2011.

[138] DeepAffects, “Introduction,” 2020. [Online]. Available: https://developers.deepaffects.com/docs. [Accessed 17 August 2021].

[139] S. Hiray, V. Nimkarde, V. Duppada and S. Tajbakhsh, “deepaffects-python,” [Online]. Available: https://github.com/SEERNET/deepaffects-python. [Accessed 17 August 2021].

[140] K. Ahmad, N. Daly and V. Liston, “What is new? News media, General Elections, Sentiment, and named entities,” *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011),* pp. 80-88, 2011.

[141] “Descriptions of Inquirer Categories and Use of Inquirer Dictionaries,” [Online]. Available: http://www.wjh.harvard.edu/~inquirer/homecat.htm. [Accessed 17 August 2021].

[142] S. Mohammad, “NRC Word-Emotion Association Lexicon,” [Online]. Available: https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm. [Accessed 17 August 2021].

[143] S. M. Mohammad and P. D. Turney, “NRC Emotion Lexicon,” National Research Council, 2013.

[144] B. Liu and M. Hu, “Opinion Mining, Sentiment Analysis, and Opinion Spam Detection,” 15 May 2004. [Online]. Available: https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon. [Accessed 17 August 2021].

[145] N. B. Van Andel, b. N. Stoker, P. Manuel and A. Fortin, “JiWER: Similarity measures for automatic speech recognition evaluation,” [Online]. Available: https://github.com/jitsi/jiwer/. [Accessed 17 June 2021].

[146] STHDA, “Kruskal Wallis test in R,” [Online]. Available: http://www.sthda.com/english/wiki/kruskal-wallis-test-in-r. [Accessed 22 August 2021].

[147] STHDA, “Unpaired Two-Samples Wilcoxon Test in R,” [Online]. Available: http://www.sthda.com/english/wiki/unpaired-two-samples-wilcoxon-test-in-r. [Accessed 22 August 2021].

[148] “Spearman's Rank-Order Correlation,” [Online]. Available: https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php. [Accessed 22 August 2021].

[149] T. Kanade, J. F. Cohn and Y. Tian, “Comprehensive database for facial expression analysis,” *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580),* pp. 46-53, 2000.

[150] Y. I. Tian, T. Kanade and J. F. Cohn, “Recognizing action units for facial expression analysis.,” *IEEE Transactions on pattern analysis and machine intelligence,* vol. 23, no. 2, pp. 97-115, 2001.