Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

School of Computer Science and Statistics

Performance Comparison of Emotion Recognition Systems, and Validation of Physical Correlates of Emotions: Across Different Ethnicities and Languages with Particular Focus on Chinese

Wanying Jiang

August 19, 2022

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

Supervised by Prof. Khurshid Ahmad

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

____

Wanying Jiang

August 19, 2022

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

____

Wanying Jiang

August 19, 2022

Abstract

Affectiva AFFDEX, Emotient FACET, openSmile and openVokaturi are four emotion recognition systems with different architectures and trained on different training datasets. The training datasets of these four systems contain only a small amount of Asian data, and there are no reports about the performance of the systems on Asian data. In this study, 45 speech videos and audios of Chinese politicians and 212 speech videos and audios of politicians from other countries were collected and processed using the four systems. In addition, two video datasets, RAVDESS and CASME II, and one audio dataset, EMO-DB, with emotion labels were used as baselines in this study. In the dataset of Chinese politicians, the speech texts were transcribed by using Amazon Transcribe. The emotions in the texts were identified through textual sentiment analysis with two Chinese sentiment lexicons, HowNet and National Taiwan University Sentiment Dictionary (NTUSD). The consistency and difference of the emotion recognition results were tested by using statistical methods including the Mann-Whitney U test, Spearman correlation coefficient, McNemar test, Cohen’s Kappa, etc. Ultimately, it was found that the consistency of the results from the systems on Chinese data was weaker than that on data from other countries. The consistency among the results of facial emotion recognition, speech emotion recognition, and textual sentiment analysis was very weak or even inconsistent. Finally, this study also validated the correlation between emotion and action unit (AU) and acoustic features using the Spearman correlation coefficient. The results of the experiment revealed correlations between Joy and AU12, correlations between Anger and AU6, AU4, AU23 and AU24, and differences in fundamental frequency (F0) in different languages and different genders, which validated previous findings. Ultimately, future work is proposed, including using approaches such as machine learning to improve the accuracy of textual sentiment analysis and implementing the fusion of different emotion analysis results based on fuzzy logic to improve the accuracy of emotion recognition.

Acknowledgements

Firstly, many thanks to my supervisor, Prof. Khurshid Ahmad, for his guidance on my experiments and dissertation over the past few months, which has enabled me to complete the research successfully.

Secondly, I would like to thank Prof. Carl Vogel, for all his help with the statistical treatment of the data during the research.

Thirdly, I would like to thank my colleagues, Deepayan Datta, Shirui Wang, and Subishi Chemmarathil, for working together to complete the experimental data collection.

Finally, I would like to thank my mum and dad, for their support and encouragement to get me through the toughest time.

Wanying Jiang

University of Dublin, Trinity College

August 2022

Contents

Abstract iii

Acknowledgements iv

Chapter 1 Introduction 1

1.1 Emotion Detection in Facial Expression, Speech and Text 1

1.2 Research Contributions 2

1.3 Dissertation Structure 2

Chapter 2 Motivation and Literature Review 4

2.1 Facial Expression 4

2.1.1 Facial Action Coding System 5

2.1.2 Facial Emotion Recognition Systems 8

2.1.3 Summary 9

2.2 Vocal Expression 9

2.2.1 Prosodic Features 10

2.2.2 Spectral Features 11

2.2.3 Voice Quality Features 11

2.2.4 Speech Emotion Recognition System 12

2.3 Verbal emotion extraction 12

2.3.1 ‘Sentiment’ in Text 12

2.3.2 Automatic Speech Recognition Technology 14

2.4 Summary 14

Chapter 3 Design Solution and Validation 16

3.1 Dataset Description 16

3.2 Data Pre-processing 18

3.2.1 Video Pre-processing 18

3.2.2 Audio Pre-processing 18

3.2.3 Speech-to-Text 18

3.3 Video Processing 19

3.3.1 Affectiva AFFDEX 19
3.3.2 Emotient FACET 21
3.3.3 Probability of Emotion Every 2000ms 22
3.4 Audio Processing 22
3.4.1 OpenSmile 22
3.4.2 OpenVokaturi 23
3.5 Text Processing 24
3.6 Hypothesis Testing and Statistical Methods 25
3.6.1 Mann-Whitney U Test 25
3.6.2 Kruskal-Wallis Test 25
3.6.3 Spearman Rank Correlation 26
3.6.4 McNemar Test 26
3.6.5 Cohen’s Kappa 27
3.7 Summary 28
Chapter 4 Case Study and Results 29
4.1 Affectiva AFFDEX and Emotient FACET 29
4.1.1 Performance on Datasets with Emotion Labels 29
4.1.2 Performance on Politician Datasets 33
4.1.2.1 Dataset of Chinese Politicians 35
4.1.2.2 Dataset of Other Politicians 36
4.1.3 Discussion 37
4.2 OpenSmile and OpenVokaturi 38
4.2.1 Performance on Datasets with Emotion Labels 38
4.2.2 Performance on Politician Datasets 40
4.2.2.1 Dataset of Chinese Politicians 42
4.2.2.2 Dataset of Other Politicians 43
4.2.3 Discussion 45
4.3 Facial Expression and Speech 45

4.3.1 Dataset of Chinese Politicians 45

4.3.2 Dataset of Other Politicians 47

4.3.3 Discussion 49

4.4 Text, Facial Expression and Speech 49

4.5 Physical Correlates of Emotions 52

4.5.1 Action Units 52

4.5.2 Acoustic Features 55

Chapter 5 Conclusion and Future Work 57

5.1 Key Findings 57

5.2 Future Work 59

Bibliography 61

List of Figures

Figure 1 Muscles of Facial Expression [11] 4

Figure 2 Word Error Rate Formula 14

Figure 3 System Pipeline 16

Figure 4 Character Error Rate Formula 19

Figure 5 The 34 Facial Landmarks in Affectiva AFFDEX [52] 20

Figure 6 The 19 Facial Expression Metrics Detected by Affectiva AFFDEX 20

Figure 7 The 7 Basic Emotions Detected by Affectiva AFFDEX 20

Figure 8 The 20 AUs Detected by Emotient FACET 21

Figure 9 The 7 Basic Emotions Detected by Emotient FACET 21

Figure 10 Formula for Calculating Intensity Score (Probability) from Log-likelihood Evidence Score in Emotient FACET 22

Figure 11 Formula for Geometric Mean 22

Figure 12 The 7 Emotions Detected by OpenSmile 23

Figure 13 Selected Acoustic Features Extracted by OpenSmile.............................23

Figure 14 The 5 Emotions Detected by OpenVokaturi..24

Figure 15 Textual sentiment analysis Results (Based on HowNet and NTUSD)......24

Figure 16 Formula for Calculating $\chi 2$ in the McNemar Test..................................27

Figure 17 Formula for Calculating Kappa 28

Figure 18 Confusion Matrix of Emotion Recognition Results of RAVDESS by Affectiva AFFDEX (Left) and Emotient FACET (Right) 30

Figure 19 Confusion Matrix of Emotion Recognition Results of CASME II by Affectiva AFFDEX (Left) and Emotient FACET (Right) 32

Figure 20 The Peak of **Anger** in Affectiva AFFDEX (Left) and Emotient FACET (Right) 34

Figure 21 The Peak of **Joy** in Affectiva AFFDEX (Left) and Emotient FACET (Right) 34

Figure 22 Confusion Matrix of Emotion Recognition Results of EMO-DB by OpenSmile (Left) and OpenVokaturi (Right) 39

Figure 23 The Peak of **Anger** in OpenSmile (Left) and OpenVokaturi (Right).........41

Figure 24 The Peak of **Happiness** in OpenSmile (Left) and OpenVokaturi (Right) .42

List of Tables

Table 1 Example Action Units and Possible Emotions (Images Are Obtained from Farnsworth [25]) 5

Table 2 Studies on the Relationship Between a Given Emotion and Action Units (AU) 6

Table 3 Relationship Between Specific Emotion and Action Units (AU) (Ranking Based on the Degree of Consensus in the 6 Previous Studies (From Highest to Lowest)) 7

Table 4 Details of Positive and Negative Words in HowNet 13

Table 5 Detail Description of Dataset 17

Table 6 Spearman's Rho Value and the Strength of Correlation 26

Table 7 Example Contingency Table for McNemar Test 27

Table 8 Kappa Value and the Strength of Consistency 27

Table 9 Accuracy and Recall of Affectiva AFFDEX and Emotient FACET for 'Joy' and 'Anger'. 30

Table 10 Mann-Whitney U Test Results (RAVDESS)....................31

Table 11 Spearman correlation coefficient (RAVDESS) 31

Table 12 Accuracy and Recall of Affectiva AFFDEX and Emotient FACET for 'Joy'. 32

Table 13 Mann-Whitney U Test Results (CASME II)....................32

Table 14 Spearman Correlation Coefficient (CASME II)....................33

Table 15 Test Results for Differences in the Distribution of 'Joy'/'Anger' Across Ethnicity, Gender and Age (Affectiva AFFDEX and Emotient FACET). 33

Table 16 Mann-Whitney U Test Results of Dataset of Chinese Politicians (Affectiva AFFDEX and Emotient FACET) 35

Table 17 Spearman Correlation Coefficient of Dataset of Chinese Politicians (Affectiva AFFDEX and Emotient FACET)....................36

Table 18 Mann-Whitney U Test Results of Dataset of Chinese Politicians (Affectiva AFFDEX and Emotient FACET) 36

Table 19 Spearman Correlation Coefficient of Dataset of Chinese Politicians (Affectiva AFFDEX and Emotient FACET)....................37

Table 20 Accuracy and Recall of OpenSmile and OpenVokaturi for 'Happiness' and 'Anger'. 39

Table 21 Mann-Whitney U Test Results (EMO-DB) 40

Table 22 Spearman Correlation Coefficient (EMO-DB) 40

Table 23 Test Results for Differences in the Distribution of 'happiness'/'anger' Across Ethnicity, Gender and Age (OpenSmile and OpenVokaturi). 41

Table 24 Mann-Whitney U Test Results of Dataset of Chinese Politicians (OpenSmile and OpenVokaturi) 42

Table 25 Spearman Correlation Coefficient of Dataset of Chinese Politicians (OpenSmile and OpenVokaturi) 43

Table 26 Mann-Whitney U Test Results of Dataset of Other Politicians (OpenSmile and OpenVokaturi) 44

Table 27 Spearman Correlation Coefficient of Dataset of Other Politicians (OpenSmile and OpenVokaturi) 44

Table 28 McNemar Test Results in Dataset of Chinese Politicians (**Joy**) 46

Table 29 McNemar Test Results in Dataset of Chinese Politicians (**Anger**) 46

Table 30 Kappa Results in Dataset of Chinese Politicians (**Joy**) 47

Table 31 Kappa Results in Dataset of Chinese Politicians (**Anger**) 47

Table 32 McNemar Test Results in Dataset of Other Politicians (**Joy**) 47

Table 33 McNemar Test Results in Dataset of Other Politicians (**Anger**) 48

Table 34 Kappa Results in Dataset of Other Politicians (**Joy**) 48

Table 35 Kappa Results in Dataset of Other Politicians (**Anger**) 48

Table 36 Top 10 most frequent words 49

Table 37 McNemar Test Results in Dataset of Chinese Politicians (**Joy/Positive**).50

Table 38 McNemar Test Results in Dataset of Chinese Politicians (**Anger/Negative**) 50

Table 39 Kappa Results in Dataset of Chinese Politicians (**Joy/Positive**) 51

Table 40 Kappa Results in Dataset of Chinese Politicians (**Anger/Negative**) 51

Table 41 Correlation between Action Units and **'Joy'** (Affectiva AFFDEX) 52

Table 42 Correlation between Action Units and **'Joy'** (Emotient FACET) 53

Table 43 Correlation between Action Units and **‘Anger’** (Affectiva AFFDEX)........54

Table 44 Correlation between Action Units and **‘Anger’** (Emotient FACET)..........54

Table 45 F0 Mean in Different Samples..55

Table 46 Correlation between Acoustic and **‘Joy’** 56

Table 47 Correlation between Acoustic and **‘Anger’**...56

Nomenclature

ADFES	Amsterdam Dynamic Facial Expressions Set
AU	Action Unit
CASME II	Chinese Academy of Sciences Micro-expression Database II
CER	Character Error Rate
CERT	Computer Expression Recognition Toolbox
CK	Cohn-Kanade dataset
CK+	Extended Cohn-Kanade dataset
CNSD	Chinese Network Sentiment Dictionary
EMO-DB	Berlin Speech Emotion Database
F0	Fundamental Frequency
FACS	Facial Action Coding System
FER	Facial Emotion Recognition
GI	General Inquirer
ISL	Intelligent Systems Laboratory Database
JAFFE	Japanese Female Facial Expression Dataset
MFCC	Mel Frequency Cepstral Coefficients
NTUSD	National Taiwan University Sentiment Dictionary
RaFD	Radboud Faces Database
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
SAVEE	Surrey Audio-Visual Expressed Emotion database
SER	Speech Emotion Recognition
Std	Standard Deviation
TEO	Teager Energy Operator
TESS	Toronto Emotional Speech Set
WER	Word Error Rate
WSEFEP	Warsaw Set of Emotional Facial Expression Pictures

Chapter 1 Introduction

1.1 Emotion Detection in Facial Expression, Speech and Text

Emotion detection is an important component in the field of artificial intelligence, which to a certain extent improves the intelligence of human-machine interaction. People can easily sense a person's emotions through facial expressions, gestures, voice and the content of their speech, etc. For machines, emotion recognition systems are used to detect emotions expressed by non-verbal communication, and textual sentiment analysis techniques can be used to analyse the emotions expressed in verbal communication. Common emotion recognition systems include facial emotion recognition (FER) systems and speech emotion recognition (SER) systems.

AFFDEX, developed by Affectiva Inc., and FACET, developed by Emotient Inc., are two commercial FER systems. Both systems are built on the Facial Action Coding System (FACS) developed by Ekman and Friesen [20]. Through face recognition and automatic facial expression recognition technology, both systems identify action units on the human face and make emotion predictions based on them. Previous studies have shown that both systems can recognize emotions well. Currently, Emotient was acquired by Apple and both systems are available through the iMotions software.

Examples of SER systems are openSmile and openVokaturi, both of which are currently available for free. When audio is passed into the openSmile and openVokaturi, the acoustic features of the audio are extracted, and based on these features, the system makes predictions about the emotion of the audio.

A simple approach for textual sentiment analysis is to use a sentiment lexicon to analyse the positivity and negativity of words in the text. In contrast to emotion recognition systems, text analysis is difficult to detect specific emotions and can only analyse whether the emotion is positive or negative. Currently, sentiment lexicons are being constructed in different languages, for example, English sentiment lexicon SentiWordNet [7], Chinese sentiment lexicons HowNet [15] and the National Taiwan University Sentiment Dictionary (NTUSD) [39].

In reality, there can be inconsistencies in emotions expressed by facial expression, voice and content of speech, which is known as ‘emotional leakage’. The term ‘emotional leakage’ was first introduced by Ekman and Friesen [19], who believed that when people try to lie to others, their true emotions may leak out from their facial expressions. In some studies [4][59], researchers have attempted to combine the results of multiple emotion detection to construct multimodal emotion recognition systems. This type of system can consider the problem of ‘emotion leakage’ and can also combine the advantages of individual unimodal emotion detection systems, thus improving the accuracy of the detection result.

1.2 Research Contributions

In this study, the performance of four emotion recognition systems, Affectiva AFFDEX, Emotient FACET, openSmile and openVokaturi, with different architectures and training datasets were tested to examine the difference and consistency of their emotion detection results for the same sample. Due to the lack of previous studies on Asian data, this study has collected speech videos of Chinese politicians from Mainland China. To ensure the integrity of the study, speech videos of politicians from other countries were also collected and analysed for comparison purposes. In addition, two video datasets with emotion labels, Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and Chinese Academy of Sciences Micro-expression Database II (CASME II), and one audio dataset with emotion labels, Berlin Speech Emotion Database (EMO-DB), are used as baselines. Experiments will be performed on data of different ethnicities, languages, genders and ages to observe whether the performance of the systems is independent of ethnicity, language, gender and age. Meanwhile, for the dataset of Chinese politicians, this study used Chinese sentiment lexicons including HowNet and NTUSD for sentiment analysis of speech texts and compared the results with those of emotion recognition systems to examine their consistency and differences. Finally, this study validated the physical correlates of emotion in the dataset used and examined whether the physical correlates of emotion are independent of ethnicity, language, gender, and age. The results of this study will provide suggestions for the construction of multimodal emotion detection systems.

This study contains the following research questions:

Research Question 1: Can emotion recognition systems with different architectures, training and testing methods have consistent emotion detection results for the same sample?

Research Question 2: Can the results of textual sentiment analysis and the emotion detection results of emotion recognition systems support each other for Chinese?

Research Question 3: Are the physical correlates of emotion independent of ethnicity, language, gender and age?

1.3 Dissertation Structure

This dissertation will include the following 4 sections. In Chapter 2, relevant previous studies on emotion detection in facial expression, speech and text will be reviewed and summarized, and contributions of this research will be discussed. In Chapter 3, the design of the experiment will be described in detail, including the collection of the videos, pre-processing of videos and audios, the acquisition of sentiment data, textual sentiment analysis, the statistical methods used, etc. In Chapter 4, the results of the experiments will be presented and discussed. Finally, Chapter 5 will summarize the findings of this study and discuss the limitations and future work.

Chapter 2 Motivation and Literature Review

Emotions can be expressed in a variety of ways, the more common of which is through facial expressions, voices and speech texts. In this chapter, a summary of previous research on the analysis of emotion in facial expression (2.1), vocal expression (2.2) and texts (2.3) will be presented, and the contribution of this research will be described (2.4).

2.1 Facial Expression

Facial expression is one of the relatively intuitive ways of expressing emotions. The human face contains many muscles, as shown in Figure 1, which play a significant role in controlling facial expressions. The realization of an expression can be the outcome of a single muscle or the combined effect of multiple muscles. For example, the occipitofrontalis muscle controls the raising of the eyebrows, and the levator labii superioris, levator labii superioris alaeque nasi and zygomaticus minor muscles cooperate to control the raising of the upper lip [11].

Next, I will introduce the Facial Action Coding System (FACS) and the studies on the relations between action units and emotions (2.1.1), the construction of the FER systems (2.1.2), and summarise the outcomes of the studies on the performance of the FER systems (2.1.3).

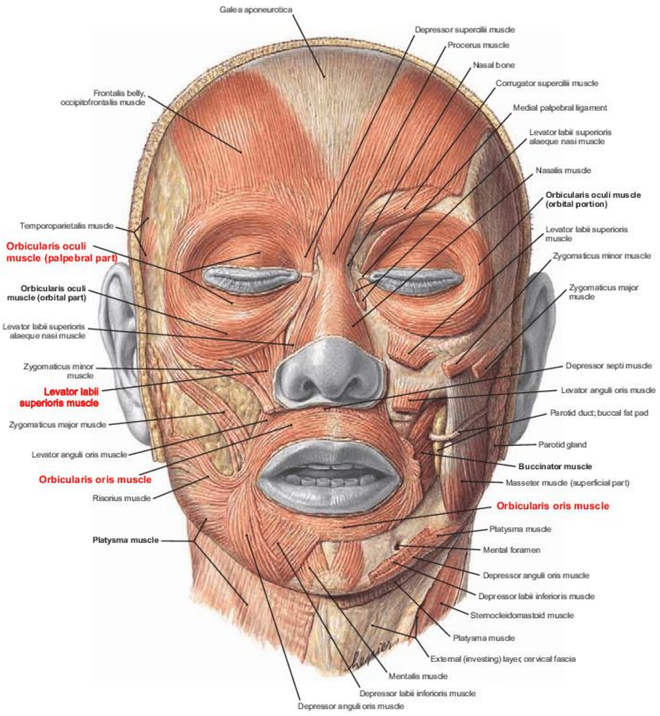
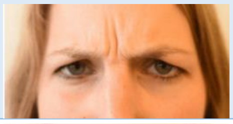
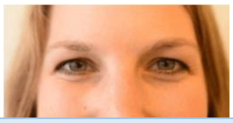
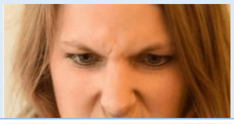
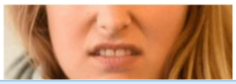
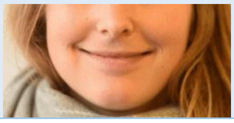
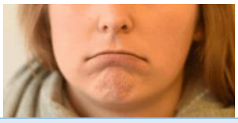
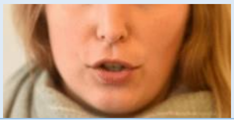
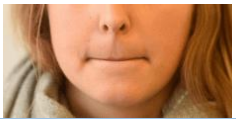

Galea aponeurotica
Depressor supercilii muscle
Procerus muscle
Nasal bone
Corrugator supercilii muscle
Medial palpebral ligament
Levator labii superioris alaeque nasi muscle
Nasalis muscle
Orbicularis oculi muscle (orbital portion)
Levator labii superioris muscle
Zygomaticus minor muscle
Zygomaticus major muscle
Frontalis belly, occipitofrontalis muscle
Temporoparietalis muscle
Orbicularis oculi muscle (palpebral part)
Orbicularis oculi muscle (orbital part)
Levator labii superioris alaeque nasi muscle
Zygomaticus minor muscle
Levator labii superioris muscle
Zygomaticus major muscle
Levator anguli oris muscle
Orbicularis oris muscle
Risorius muscle
Platysma muscle
Depressor septi muscle
Levator anguli oris muscle
Parotid duct; buccal fat pad
Parotid gland
Buccinator muscle
Masseter muscle (superficial part)
Orbicularis oris muscle
Platysma muscle
Mental foramen
Depressor anguli oris muscle
Depressor labii inferioris muscle
Sternocleidomastoid muscle
Platysma muscle
External (investing) layer, cervical fascia
Mentalis muscle
Depressor labii inferioris muscle
Depressor anguli oris muscle

Figure 1 Muscles of Facial Expression [11]

2.1.1 Facial Action Coding System

To facilitate the research and application of human facial expression, the Facial Action Coding System (FACS) was proposed. The first attempt to encode facial expressions was made in 1969 by the Swedish anatomist Hjortsjö, and he encoded 23 facial expressions [28]. Subsequently, Ekman and Friesen established FACS based on Hjortsjö's work in 1976 [20]. Eventually, Ekman et al. updated FACS further in 2002 [21]. The action units in FACS currently used can be divided into three categories, which are the main action units, the head movement action units and the eye movement action units. The facial muscles associated with each action unit may be single or multiple. These action units can encode each individual facial motion. Table 1 shows some of the action units from the main action unit and the possible emotions associated with them.

Table 1 Example Action Units and Possible Emotions (Images Are Obtained from Farnsworth [25])

Action Unit	Description	Example	Possible Emotion
AU4	Brow Lowerer		Anger
AU6	Cheek Raiser		Joy
AU9	Nose Wrinkler		Disgust, Anger
AU10	Upper Lip Raiser		Disgust
AU12	Lip Corner Puller		Joy
AU17	Chin Raiser		Sadness
AU22	Lip Funneler		Not obvious
AU28	Lip Suck		Not obvious

The establishment of FACS has enabled various facial expressions to be represented concisely and clearly. Based on FACS, many studies have been conducted on the relationship between facial expressions and emotions in different cultures. Some of the studies focused on the relationship between emotions and posed expressions. Du et al. analysed the posed expression of 21 emotions in 230 subjects in their study, and these subjects included races such as Caucasian, Asian, African American, and Hispanic [17]. Cordaro et al. analysed the relationship between posed expressions and 22 emotions in 119 subjects from China, India, Japan, Korea, and the United States [12]. The experiment held by Keltner et al. involved subjects from 10 countries including China, Japan, Korea, and New Zealand, and the relationship between 18 emotions and posed expression was analysed [35]. Other studies have focused on the relationship between emotions and spontaneous expressions. The study of Lucey et al. was based on the Extended Cohn-Kanade dataset (CK+), which includes posed expression and spontaneous expression of 123 subjects with 81% Euro-Americans, 13% African-Americans, and 6% other groups [46]. This dataset was also used in the study of Velusamy et al., in addition to this, they used datasets including the Intelligent Systems Laboratory (ISL) database, FACS, the Japanese Female Facial Expression (JAFFE) Dataset, MindReading and video data they collected in the laboratory [67]. Matsumoto et al. also analysed the relationship between spontaneous expression and the 8 emotions through a summary of 26 previous studies [50]. Table 2 shows a summary of the relationship between the 6 basic emotions, including anger, joy, sadness, surprise, disgust and fear, and action units in the 6 previous studies mentioned above.

Table 2 Studies on the Relationship Between a Given Emotion and Action Units (AU)

Action Unit	Studies on					
	Anger	Joy	Sadness	Surprise	Disgust	Fear
AU1			[17][35][46][50][67]	[12][17][35][46][50][67]		[12][17][35][46][50][67]
AU2	[67]			[12][17][35][46][50][67]		[12][17][35][46][50]
AU4	[12][17][35][50][67]		[12][17][35][46][50][67]		[12][17][67]	[17][35][46][50][67]
AU5	[35][50]			[12][17][35][46][50][67]		[12][17][35][46][50][67]
AU6		[12][17][35][50][67]	[17][35][46]		[12][67]	
AU7	[12][17][50][67]	[12][35]			[12][35][67]	[12][35][67]
AU9					[12][17][35][46][50][67]	
AU10	[50]	[67]	[67]		[12][17][46][50]	
AU11			[17][46]			
AU12		[12][17][35][46][50][67]				
AU15			[17][35][46][50][67]			
AU16		[12]				
AU17	[17][35][67]		[17][35][50][67]		[17][67]	

AU20						[17][35][50][67]
AU22	[50]					
AU23	[17][35][46][50][67]	[67]				
AU24	[17][35][46][50]				[17]	
AU25		[12][17][35]		[12][17][35][50]	[12][35][50]	[12][17][35][50]
AU26		[12][35][67]		[12][17][35][50][67]	[12][35][50]	[12][17][50]
AU27		[12]		[12][67]	[12]	[12]
AU43			[12]			

Summarising Table 2 according to the degree of consensus on the relationship between particular emotions and AU in these six studies, Table 3 can be obtained. It can be observed that all 6 studies considered AU12 to be related to Joy and AU9 to be related to Disgust. Meanwhile, these two AU were found to be related only to the corresponding emotion suggesting that Joy and Disgust are two pure emotions. For Surprise and Fear, all six studies found that AU1 and AU5 were related to them. In addition to this, in Surprise, all six studies found AU2 to be associated with it. From the results, it seems that the facial expressions of these two emotions are relatively similar. For Sadness, AU4 was found to be related to this emotion in six studies, but five studies found that AU4 was also related to Anger and Fear, so the possibility of Sadness cannot be fully determined based on AU4. In addition, five studies found that AU15 was only associated with Sadness. In contrast, for Anger, the six studies had no consensus on the relationship between AU and it, and only five studies considered AU4 and AU23 to be associated with it.

Table 3 Relationship Between Specific Emotion and Action Units (AU) (Ranking Based on the Degree of Consensus in the 6 Previous Studies (From Highest to Lowest))

Degree of Consensus	Anger	Joy	Sadness	Surprise	Disgust	Fear
6		**AU12**	AU4	AU1, AU2, AU5	**AU9**	AU1, AU5
5	AU4, AU23	AU6	AU1, ***AU15***	AU26		AU2, AU4
4	AU7, AU24		AU17	AU25	AU10	AU20, AU25
3	AU17	AU25, AU26	AU6		AU4, AU7, AU25, AU26	AU7, AU26
2	AU5	AU7	AU11	AU27	AU6, AU17	
1	AU2, AU10, AU22	AU10, AU16, AU23, AU27	AU10, AU43		AU24, AU27	AU27

It is evident from the results of these studies that there is disagreement in the association between some of the action units and emotions, which may be due to the differences in gender, age and ethnicity in the datasets used in these studies.

Some studies have also shown differences in facial expression across ethnicity, gender and age. The study by Jack et al. [30] found that Western Caucasians showed greater variation in their facial expression of the six basic emotions, including anger, happiness, surprise, fear, disgust and sadness, and were able to form six different models, whereas East Asians showed greater overlap in their expression of these emotions, which suggests that there are differences in the expression of emotions across races. McDuff et al. [51] found differences in emotional expression across cultures, which include individualistic and collectivistic cultures, and genders through analysis of a large sample. One study showed that there are differences in facial expression between the elderly and the young, as facial muscles atrophy with age and facial expression is affected as a result [26].

2.1.2 Facial Emotion Recognition Systems

Initially, to use FACS to encode a face, humans had to be professionally trained and used manually to complete the encoding. However, this method is time-consuming and prone to bias. Recently, it is more common to use automatic facial expression recognition technology. Donato et al. compared the currently commonly used automatic facial recognition methods in their study, which include optical flow analysis, principal component analysis, local feature analysis, independent component analysis, Gabor wavelet-based methods, etc. [14]. The study showed that independent component analysis and Gabor wavelet-based methods performed the best, with an accuracy of 96%, which was consistent with the level of manual coding by experts.

Based on the establishment of FACS, the study of the association of emotion and action units and the development of automatic facial expression recognition techniques, FER systems have been developed. The system implements emotion recognition in 3 main steps. First, recognize the face in the image. Second, using automatic facial expression recognition techniques to annotate action units. Finally, analyse action units to predict emotions.

One example is FACET, developed by Emotient Inc., which was built on the Computer Expression Recognition Toolbox (CERT) [44]. CERT provides intensity detection of 19 action units and probability of 6 basic emotions, including anger, happiness, sadness, surprise, disgust and fear. The developers used a multiple logistic regression model on the Cohn-Kanade dataset (CK) [32] to train the emotion recognition module of CERT. The dataset included 97 subjects with an age range from 18 to 30, 65% are female, 82% are Euro-Americans, 15% are African-Americans, and 3% are Asian or Latino. CK+ was used to evaluate the emotion recognition module of CERT which had an additional 26 subjects over CK. The final average precision of the 7 emotions, which contains 6 basic emotions and neutral, for these 26 subjects was 87.21%.

Another example of the FER system is AFFDEX [52], developed by Affectiva Inc., which can recognize 7 emotions, which include anger, joy, sadness, surprise, disgust, fear and contempt. The developers collected about 1.8 million facial videos via the Internet using webcams from India, the United States, China and Indonesia. By labelling manually, 27,000 videos were finally obtained [64]. The system is currently used in video conferencing, remote education, health inspection, and gaming [52]. Both emotion recognition systems, FACET and AFFDEX, are now available through iMotions.

2.1.3 Summary

Since FACET and AFFDEX are trained and tested using different datasets and different models, some studies on their performance have been carried out in the past few years. Stöckli et al. [65] showed that the accuracy of emotion recognition of FACET was better than that of AFFDEX, where for posed facial expression, the accuracy of FACET was 97% compared to 73% for AFFDEX, and for spontaneous facial expression, the accuracy of FACET was 57% compared to 55% for AFFDEX. Ahmad et al. [2] demonstrated that the emotion recognition results of AFFDEX and FACET for semi-spontaneous emotional expressions were statistically significantly different for samples of a different races, gender and age, and the correlation between the emotion recognition results of the two systems for most emotions was weak for the same sample. Yang et al. [72] evaluated the performance of AFFDEX with 3 datasets, which include the Amsterdam Dynamic Facial Expressions Set (ADFES), the Radboud Faces Database (RaFD) and the Warsaw Set of Emotional Facial Expression Pictures (WSEFEP). The overall accuracy was 67%, with happiness being the best recognition (F1=0.954), followed by disgust (F1=0.82) and fear being the worst (F1=0.13). Krumhuber et al. [38] evaluated the performance of FACET with 14 datasets and overall FACET was more accurate in recognising posed expressions than spontaneous expressions. Joy was recognized best, followed by anger, sadness and disgust, and fear was recognized worst.

In this study, some of the above findings are validated using two video datasets with emotion labels.

2.2 Vocal Expression

Voice is another way to express emotions. By extracting and analysing the features in audio, the emotions of speech can be recognized. The features contained in audio can be divided into four categories, including prosodic, spectral, voice quality and Teager Energy Operator (TEO) based features [3].

I will discuss the prosodic features (2.2.1), spectral features (2.2.2) and voice quality features (2.2.3) used in SER systems, and this will lead to a discussion of two key SER systems, openSmile and openVokaturi (2.2.4).

2.2.1 Prosodic Features

The most commonly considered feature in the prosodic feature is the fundamental frequency (F0). Sound is emitted through the vibration of the vocal cords, which can be broken down into several sine waves of different frequencies, with F0 representing the lowest frequency. Studies have shown that emotions can have an impact on F0. Paeschke et al. [57] found that the mean F0 varied across emotions, with boredom, sadness, neutrality, disgust, anger, fear and happiness in order from lowest to highest. While for the standard deviation of F0, anger and happiness were the largest, followed by neutral and disgust, then fear and boredom, and the smallest was sadness. Probst and Braun [60] also show the differences in the F0 across emotions, and they similarly found that the mean F0 for sadness was lower than neutral, and the mean F0 for happiness and anger was higher than neutral. In addition, they examined F0 for different degrees of emotion and found that as the degree of emotion increased, the mean F0 increased.

Research have shown that language can also have an impact on F0. By studying females speaking English and German, Mennen et al. [53] found that although the mean F0 did not differ significantly between the two languages, there were significant differences in other aspects, for instance, the range, maximum and minimum values of F0. Mandarin Chinese as a tonal language is quite different from other languages. Keating and Kuo [34] found by comparison that the mean F0 was higher and the F0 range was greater in Mandarin than in English, while they noted that this difference is due to the high-falling tone (tone 4) of Mandarin. Ding et al. [13] also demonstrated differences in F0 between English and Mandarin, while they found differences in F0 patterns for native English speakers and those whose second foreign language was English.

The value of F0 is also influenced by gender, with Teixeira et al. [66] stating that the F0 of females ranges between 200 and 300 Hz, while that of males ranges between 80 and 200 Hz. This was supported by experiments in the study [13], where the mean F0 of English-speaking females was 205Hz and that of males was 124Hz, while the mean F0 of Chinese-speaking females was 229Hz and that of males was 141Hz.

Age is another factor that influences F0 values, and Eichhorn et al. [18] showed that F0 decreases with age in both males and females, with a more significant decline in females, one reason for which might be the change in females' hormone levels.

2.2.2 Spectral Features

Mel Frequency Cepstral Coefficients (MFCC) is one of the spectral features which is most widely used. Studies have shown that MFCC is associated with emotions in the voice. Likitha et al. [42] built a SER system based on MFCC. They identified three emotions, sadness, happiness and anger, by dividing the standard deviation of MFCC. For sadness, MFCC standard deviation ranged from 0.1700 to 0.2199, for happiness MFCC standard deviation ranged from 0.2200 to 0.2899 and for anger MFCC standard deviation ranged from 0.2900 to 0.3999. The final accuracy of the system was 80%. Kumbhar and Bhandari [40] trained a SER model using 39 MFCC factors and the LSTM algorithm, which recognized the emotions happy, angry, sad, fearful, surprised, disgusted, calm and neutral with an accuracy of 84.81%.

The work by Koolagudi et al. [36] demonstrated that MFCC can be used to distinguish languages. They constructed language classification models using different numbers of MFCC factors for distinguishing 15 Indian languages with an accuracy of up to 88.4%. Gunawan et al. [27] also demonstrated that MFCC can be used to differentiate between five languages including Arabic, Chinese, English, Korean and Malay, with an accuracy of 78%. The gender difference can also lead to differences in MFCC. Yücesoy and Nabiyev [73] identified the gender of the speaker using MFCC with an accuracy of 97.67%.

2.2.3 Voice Quality Features

Voice Quality Features include jitter, shimmer, etc. Jitter represents the variation of F0 between vibration cycles and shimmer represents the variation of F0 in amplitude [3]. Drioli et al. [16] found that for stressed vowels, shimmer values were higher for anger and jitter values were higher for surprise and joy, whereas for non-stressed vowels, both shimmer and jitter values were higher for disgust, and the difference in jitter and shimmer values for other emotions was not significant. Nunes et al. [55] found through their study of Portuguese speech that negative emotions such as anger, despair, sadness, and fear increased jitter values, while joy had the lowest jitter values and could be distinguished by jitter. Whereas shimmer can be used to distinguish anger and despair, anger has the highest shimmer value, followed by despair, and there is little difference in shimmer values for the other emotions. These two studies had inconsistent results on the value of jitter for joy.

A study by Wagner and Braun [68] on German, Italian and Polish found that shimmer values differed across languages, while differences in jitter values across languages were not found. Zhu et al. [75] found that when people spoke a second foreign language, jitter and shimmer were lower compared to when they spoke their native language, which was verified in Mandarin and English. Brockmann et al. [8] found that gender had a small but significant effect on jitter and shimmer, with men having lower jitter and shimmer values than women.

2.2.4 Speech Emotion Recognition System

By analysing and learning from some of the extracted voice features, a SER system can be created. OpenSmile [24] is a tool for speech feature extraction. The extraction of different acoustic features and the detection of emotions can be achieved by using different configuration files. The openSmile/openEar 'emobase' sets include several configuration files for feature extraction and emotion recognition. It can be used to detect the seven basic emotions, including angry, happy, scared, disgusted, sad, bored and neutral, which was trained on the Berlin Speech Emotion Database (EMO-DB) and the eNTERFACE database [23]. EMO-DB includes the emotional speech of 5 males and 5 females in German [10]. The eNTERFACE database collected the English speech of 42 subjects from 14 countries, of which 81% were male [49]. Ultimately, the system was tested with a recall rate of 89.5% at EMO-DB and 75.2% at eNTERFACE [23].

OpenVokaturi is another open-source SER system that detects 5 emotions, including angry, happy, scared, sad and neutral, and the model is trained on EMO-DB and the Surrey Audio-Visual Expressed Emotion (SAVEE) database [56]. The SAVEE dataset collected English emotional speech from 4 males [31]. Ley et al. [41] used the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) to test the performance of emotion recognition of openVokaturi, the results showed that the system had the highest accuracy for angry (42%), followed by happy (41.8%) and the lowest accuracy for sad (9.2%), and happy and angry were often confused. The authors also suggest that the error in this experiment may be due to the fact that the emotions of some of the samples in the RAVDESS were not accurately labelled. Bui and Chong [9] evaluated the performance of openVokaturi using the Toronto Emotional Speech Set (TESS) dataset, with an overall accuracy of 0.43, with the highest F1 score for neutral (0.67), 0.43 for sad, 0.42 for happy, 0.36 for angry and 0.26 for fear. Another study showed that the accuracy of openVokaturi in detecting emotions was 66.5% [48].

2.3 Verbal emotion extraction

This section will introduce methods for textual sentiment analysis, in particular, text analysis based on sentiment lexicons (2.3.1), and introduce automatic speech recognition technology for speech-to-text conversion (2.3.2).

2.3.1 'Sentiment' in Text

Texts can convey information as well as emotions. Nowadays, textual sentiment analysis is used in many fields. In the field of e-commerce, sentiment analysis of customer reviews allows companies to understand the condition of their products and to make timely improvements to the products [6][29]. Saranya et al. [63] suggest that sentiment analysis of tweets can identify health problems that occur in some countries or regions and contribute to the targeting of health services in the future. Ren et al. [61] included textual sentiment analysis on the traditional recommender system to enable better recommendations based on user preferences.

Textual sentiment analysis can be divided into fine-grained, which analyses words or phrases, and coarse-grained sentiment analysis, which analyses sentences or large amounts of text as a whole [43]. The current approaches to sentiment analysis are machine learning, sentiment lexicon-based analysis and ontology-based analysis [1].

The machine learning approach consists of the following steps. First, perform pre-processing, feature extraction and feature selection on the training data, then train text sentiment classifier using algorithms such as Naive Bayes, support vector machine and logistic regression [69], and finally use the classifier model for sentiment classification.

The lexicon-based analysis is an easier approach for sentiment analysis, and lexicons have been created for sentiment analysis in various languages. SentiWordNet is an English sentiment lexicon that ranks words in WordNet for positivity, negativity and neutrality through semi-supervised learning and random walk, with the latest update, SentiWordNet 3.0, improving the accuracy of both positivity and negativity ranking [7]. HowNet [15] is commonly used for Chinese textual sentiment analysis. It contains two languages, Chinese and English, with four subsets of sentiment words for each language, as shown in Table 4. The National Taiwan University Sentiment Dictionary (NTUSD) [39] is another commonly used Chinese sentiment dictionary, which was developed based on the General Inquirer (GI) and the Chinese Network Sentiment Dictionary (CNSD). With the opinion mining of news, blogs and other corpora, the dictionary eventually includes 2810 positive words and 8275 negative words.

Table 4 Details of Positive and Negative Words in HowNet

	Chinese	English
Positive Opinion Words	3730	3594
Negative Opinion Words	3116	3563
Positive Emotion Words	836	769
Negative Emotion Words	1254	1011

Compared to English sentiment lexicons, the construction of Chinese sentiment lexicons is more challenging. First, the ambiguity of the meaning of Chinese words expressed in different contexts leads to difficulties in calculating sentiment polarity. Second, the resources available for building sentiment dictionaries are limited [58]. These factors have resulted in fewer mature lexicons in Chinese.

2.3.2 Automatic Speech Recognition Technology

Automatic speech recognition technology enables text to be extracted from speech. Pattern recognition is one of the approaches for automatic speech recognition, where patterns are constructed for each word in the lexicon using acoustic signals during training, and during recognition, the most likely word is matched by comparing the pattern of the word to be recognised with the stored pattern. Artificial intelligence methods are also commonly used in automatic speech recognition which utilises the idea of pattern recognition and uses different models to achieve pattern matching [33].

Currently, many companies offer cloud-based automatic speech recognition services that allow for easier transcription of speech, these include Amazon Transcribe, Microsoft Azure Speech-to-Text, Google Cloud Speech-to-Text, etc. Word Error Rate (WER) is a metric used to measure the performance of speech-to-text services and is calculated using the formula in Figure 2. Mrozek et al. [54] compared the performance of automatic speech recognition services from Google, IBM, Microsoft and Amazon using the VoxForge dataset which includes speech in English, and found that Amazon Transcribe had the lowest WER (8.4%), followed by Microsoft Azure Speech-to-Text (9.9%). In another study [70], which also attempted to examine the performance of these four services on French, it was ultimately found that for good quality speech, Microsoft Azure had a lower WER (9.09%), followed by Amazon Transcribe (11.76%), while for poorer quality speech, Microsoft Azure still had the lowest WER (11.11%), followed by Amazon Transcribe and Google Cloud (20.00%). Experiments by Luzietti et al. [47] demonstrate that Amazon Transcribe has a lower WER than Google Cloud Speech-to-Text in Italian, French and Spanish, and Amazon Transcribe also performs better in different qualities of speech.

$$WER = \frac{Insertions + Substitutions + Deletions}{Total\ Words\ in\ the\ Reference} \times 100\%$$

Figure 2 Word Error Rate Formula

Previous studies have found that Amazon Transcribe and Microsoft Azure Speech-to-Text have better accuracy. However, there is a lack of research on the accuracy of these systems for Chinese transcription.

2.4 Summary

In this chapter, previous studies in facial, speech and text emotion recognition are reviewed, including the introduction of FACS, the association of action units and speech features with emotion, the introduction of FER and SER systems, methods for textual sentiment analysis, the introduction of emotion lexicons and the introduction of automatic speech recognition techniques for speech-to-text.

This chapter also reviews several experiments on the accuracy of emotion detection for FER systems, Affectiva AFFDEX and Emotient FACET, and SER systems, openSmile and openVokaturi. However, most of the experiments were performed on individual systems only, and very few of them tested the consistency of the detection results. Also, these emotion detection systems are built on different architectures, and different training datasets, and some studies have suggested that the physical correlates of emotion are related to ethnicity, language, gender and age. Studies have shown that only a small amount of Asian data is included in these training datasets, and the majority of subjects were young. Previous experiments have lacked research on the performance of these systems for emotion recognition in Asians, especially for Chinese speakers, in different gender groups, and in different age groups.

This study will examine the consistency and difference in the results of emotion detection across different emotion recognition systems for the same sample, and whether this is independent of ethnicity, language, gender and age. Meanwhile, the physical correlates of emotions will be validated within the dataset used in this experiment. The next chapter provides a detailed description of the experimental design.

Chapter 3 Design Solution and Validation

My research comprises facial emotion recognition, speech emotion recognition and textual sentiment analysis of the voice tracks of the video under analysis. The experimental process includes video collection, video pre-processing, audio pre-processing, acquisition of speech text, sentiment analysis of video, audio and text, data analysis and hypothesis testing. I will attempt to synthesise the three sources of emotion using 6 different software systems, plus the analysis of the outputs of these systems, which will be subject to statistical inference using Python language. Thus, a pipeline is created to implement my system (See Figure 3). A more detailed description of these processes is given in the rest of this chapter, which includes video collection and dataset description (3.1), video and audio pre-processing and speech-to-text (3.2), facial emotion recognition (3.3), speech emotion recognition (3.4), textual sentiment analysis (3.5), and introduction of statistical methods (3.6).

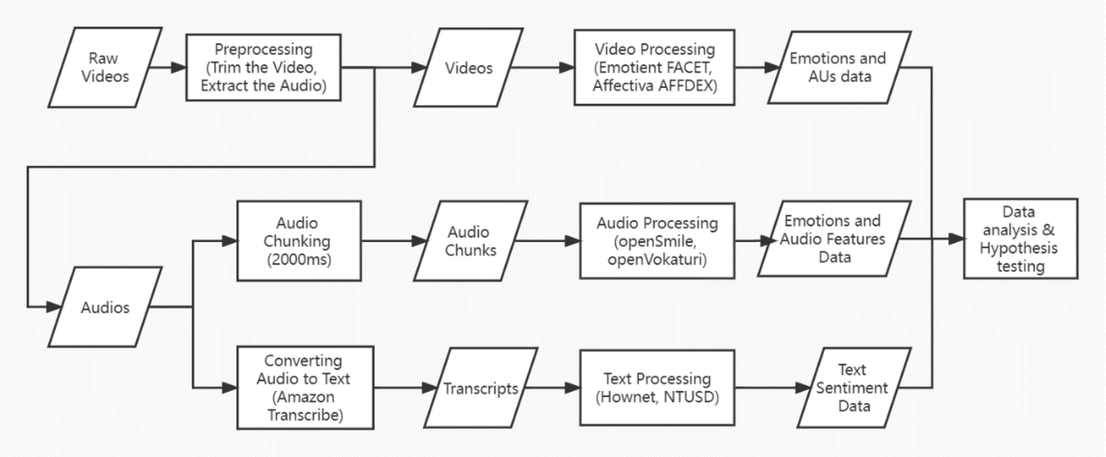

Raw Videos
Preprocessing (Trim the Video, Extract the Audio)
Videos
Video Processing (Emotient FACET, Affectiva AFFDEX)
Emotions and AUs data
Audio Chunking (2000ms)
Audio Chunks
Audio Processing (openSmile, openVokaturi)
Emotions and Audio Features Data
Data analysis & Hypothesis testing
Audios
Converting Audio to Text (Amazon Transcribe)
Transcripts
Text Processing (Hownet, NTUSD)
Text Sentiment Data

Figure 3 System Pipeline

3.1 Dataset Description

The video datasets used in this experiment include speech videos of politicians. They are semi-spontaneous expressions, where politicians usually control their expressions when speaking to the public, but there is also a degree of spontaneous expression. This type of video is recorded with the knowledge of the participants and is usually recorded in good lighting and radio conditions and i quiet environment, therefore the quality of the video and audio is good. For the videos collected, participants should face the camera directly, with their faces uncovered (e.g., masks, sunglasses, etc.).

One of the datasets used in this experiment is videos of speeches by Chinese politicians, and they all speak Chinese. The videos were collected from the official accounts of China Daily and China Net Live. All videos are in MP4 format. The final dataset included 45 videos with a total duration of 225 minutes and 8 seconds. There are 15 subjects, 11 male and 4 female, with an age range from 44 to 68 years old and an average age of 56 years old.

Another dataset contains speech videos of politicians from other countries. Some of the videos in this dataset were collected from video datasets already available in the lab, and some were collected by my colleagues. The final dataset contains 212 videos with a total duration of 774 minutes and 28 seconds. There are 56 subjects, 34 male and 22 female, from 10 countries including America, Germany, India, Japan, South Korea, New Zealand, Pakistan, the UK, Ireland and Bangladesh. The vast majority of them speak English, with only a few of them speaking Korean, Japanese, Bengali, etc. The average age is 56 years old, with an age range of 32 to 84 years old. Table 5 provides a detailed description of the video dataset of Chinese politicians and other national politicians used in the experiment.

Table 5 Detail Description of Dataset

Nationality	Male			Female		
	Number of Subjects	Number of Videos	Average Age	Number of Subjects	Number of Videos	Average Age
China	11	33	58.1	4	12	51.3
America	9	39	62.1	10	40	55.5
Germany	1	5	61.0	1	5	66.0
India	11	25	63.0	4	15	53.3
Japan	2	4	69.0	/	/	/
Korea	2	4	69.0	/	/	/
New Zealand	/	/	/	1	5	40.0
Pakistan	2	7	48.4	/	/	/
UK	1	5	56.0	/	/	/
Ireland	5	23	48.5	5	38	48.5
Bangladesh	1	4	50.5	1	3	67.0

In addition to this, two video datasets and one audio dataset that contained emotion labels were selected to examine the performance and consistency of two FER systems and two SER systems, and were used as the baseline for the experiment. The video datasets are RAVDESS and the Chinese Academy of Sciences Micro-expression Database II (CASME II). RAVDESS is a posed expression dataset containing 24 subjects, with equal numbers of males and females, with 20 Caucasians, 2 East Asians and 2 mixed races [45]. CASME II contains 247 samples of spontaneous emotional expressions from 26 participants, all of which are Asian, and the reliability of the emotion labels in the dataset is claimed to be 0.846 [71]. The audio dataset is EMO-

DB, which was chosen because both openSmile and openVokaturi were trained on this dataset.

3.2 Data Pre-processing

3.2.1 Video Pre-processing

As the videos in the video dataset were not recorded specifically for the emotion recognition experiments, the videos contain some irrelevant frames, which can affect the recognition of emotions. These irrelevant frames include frames that do not contain the face of the subject and frames that contain multiple faces. These frames were removed by using the Adobe Premiere software. The final processed videos ensured that there is and only the face of the subject and it appears in the centre of the screen. The processed videos remain in their original MP4 format.

3.2.2 Audio Pre-processing

The audio will be extracted from the processed video, this operation is done with Adobe Premiere software and the audio files are in WAV format. When processing speech audio of politicians from other countries, it was found that some of the audio contained both voices of the politician and the interpreter. These audios needed to be removed from the audio dataset as they would have an impact on the results of speech emotion recognition, but the videos could still be used for the analysis of facial emotions. Three audios were eventually excluded, which were audio of politicians from Germany, Japan and South Korea.

In order to analyse the emotion of each audio in more detail, the audio needs to be chunked. The duration of each chunk is 2000 milliseconds(ms), a time interval considered to be the shortest interval at which people can express their emotions through speech. Audio chunking is implemented via pydub, an audio processing library in Python.

3.2.3 Speech-to-Text

For the speech audio of Chinese politicians, the text content of the speech needs to be obtained for subsequent textual sentiment analysis. Based on the performance comparison of commonly used Speech-to-Text services in the previous chapter, Amazon Transcribe was ultimately chosen to be used as the text transcription tool for this experiment. Although the accuracy of this system in Chinese has not been verified in previous studies, research has shown that it performs better in numerous languages.

Amazon Transcribe supports 27 languages including Simplified Chinese, Australian English, British English, Indian English, French, etc., recommends FLAC or WAV format audio and offers 60 minutes per month for free for the first 12 months [5]. To use

Amazon Transcribe requires creating an S3 storage bucket and uploading audios to the bucket, then creating transcription jobs in the Amazon Transcribe console and running them. The final transcription can be exported and saved as a JSON file. The file contains the entire transcribed text, as well as individual words with information including timestamps, confidences, etc.

As Chinese is different from other languages where the smallest unit in a sentence is a character rather than a word, the Character Error Rate (CER) is used to measure the performance of Amazon Transcribe. The formula is shown in Figure 4. Reference texts are obtained by manual transcription.

$$CER = \frac{Insertions + Substitutions + Deletions}{Total\ Characters\ in\ the\ Reference} \times 100\%$$

Figure 4 Character Error Rate Formula

3.3 Video Processing

After the videos have been pre-processed, they are processed using the FER systems Affectiva AFFDEX and Emotient FACET, which are both available with the iMotions software. During the processing, the videos will be processed frame by frame to obtain the action units and emotion data for each frame. The interval between two frames may vary in different videos, with the shortest interval of 16ms, the longest interval of 50ms and the average interval of 35.38ms. Next, the video processing process and output of Affectiva AFFDEX (3.3.1) and Emotient FACET (3.3.2) are described, as well as the calculation method of the probability of emotion in every 2000ms (3.3.3).

3.3.1 Affectiva AFFDEX

After importing the video into the system, it is divided into several frames. For each frame, the system detects the face in the image and then obtains the distance between the eyes and labels the 34 facial landmarks (see Figure 5). Based on the location and orientation information of these facial markers, each action unit, and each emotion can be detected independently using different classifiers which ultimately return the probability of that action or emotion. The classification is implemented at a statistical level, where facial expressions are converted into numerical values according to facial landmarks and then compared with the values in the database [37].

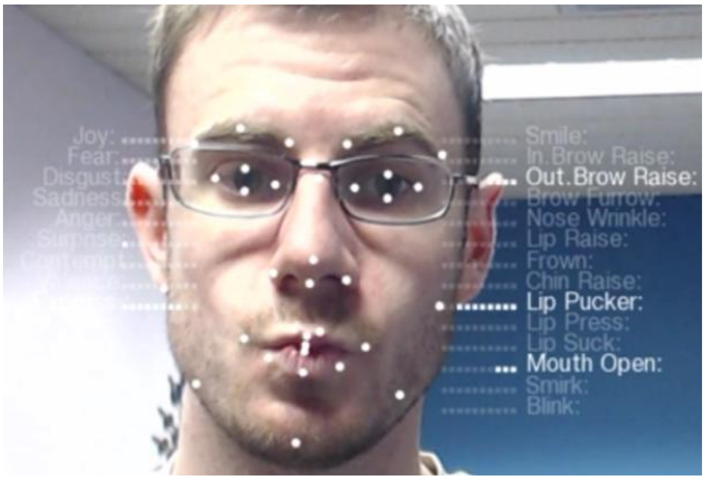
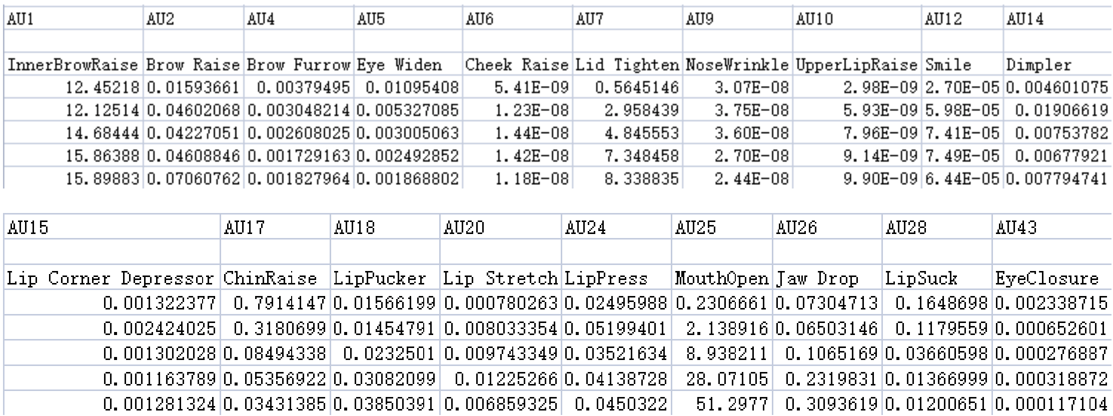


Figure 5 The 34 Facial Landmarks in Affectiva AFFDEX [52]

Figure 6 lists the 19 AUs recognised by Affectiva AFFDEX with the corresponding AU codes added in the first row as a reference, as well as the sample results for 5 frames. The result indicates the probability of the action and the value ranges from 0 to 100.

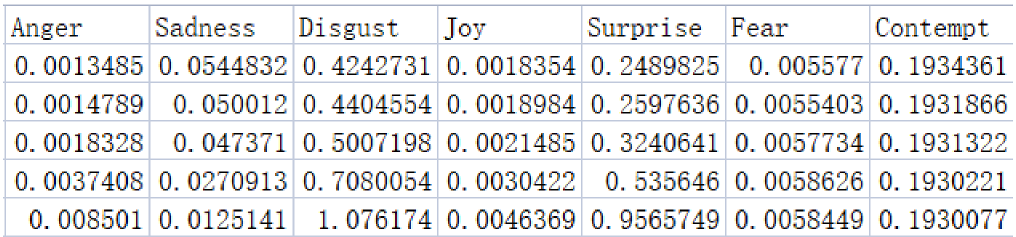


AU1	AU2	AU4	AU5	AU6	AU7	AU9	AU10	AU12	AU14
InnerBrowRaise	Brow Raise	Brow Furrow	Eye Widen	Cheek Raise	Lid Tighten	NoseWrinkle	UpperLipRaise	Smile	Dimpler
12.45218	0.01593661	0.00379495	0.01095408	5.41E-09	0.5645146	3.07E-08	2.98E-09	2.70E-05	0.004601075
12.12514	0.04602068	0.003048214	0.005327085	1.23E-08	2.958439	3.75E-08	5.93E-09	5.98E-05	0.01906619
14.68444	0.04227051	0.002608025	0.003005063	1.44E-08	4.845553	3.60E-08	7.96E-09	7.41E-05	0.00753782
15.86388	0.04608846	0.001729163	0.002492852	1.42E-08	7.348458	2.70E-08	9.14E-09	7.49E-05	0.00677921
15.89883	0.07060762	0.001827964	0.001868802	1.18E-08	8.338835	2.44E-08	9.90E-09	6.44E-05	0.007794741

AU15	AU17	AU18	AU20	AU24	AU25	AU26	AU28	AU43
Lip Corner Depressor	ChinRaise	LipPucker	Lip Stretch	LipPress	MouthOpen	Jaw Drop	LipSuck	EyeClosure
0.001322377	0.7914147	0.01566199	0.000780263	0.02495988	0.2306661	0.07304713	0.1648698	0.002338715
0.002424025	0.3180699	0.01454791	0.008033354	0.05199401	2.138916	0.06503146	0.1179559	0.000652601
0.001302028	0.08494338	0.0232501	0.009743349	0.03521634	8.938211	0.1065169	0.03660598	0.000276887
0.001163789	0.05356922	0.03082099	0.01225266	0.04138728	28.07105	0.2319831	0.01366999	0.000318872
0.001281324	0.03431385	0.03850391	0.006859325	0.0450322	51.2977	0.3093619	0.01200651	0.000117104

Figure 6 The 19 Facial Expression Metrics Detected by Affectiva AFFDEX

Affectiva AFFDEX recognises 7 basic emotions, including anger, sadness, disgust, joy, surprise, fear and contempt, and returns their probabilities in each frame, with values in the range of 0 to 100. Figure 7 shows these 7 emotions and sample results of 5 frames.



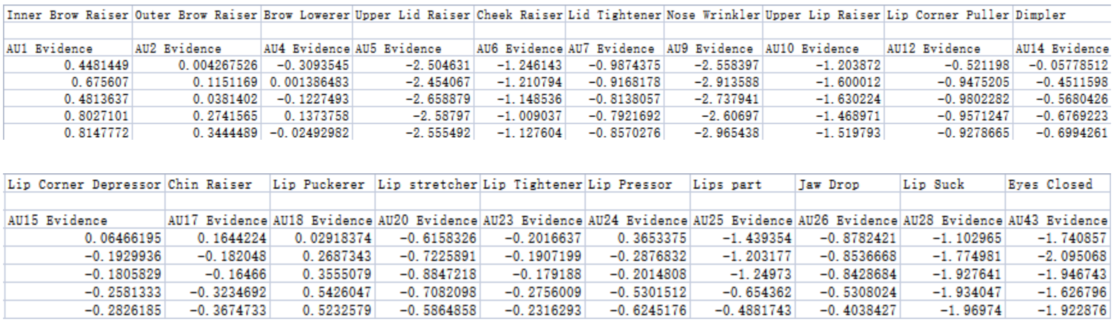
Anger	Sadness	Disgust	Joy	Surprise	Fear	Contempt
0.0013485	0.0544832	0.4242731	0.0018354	0.2489825	0.005577	0.1934361
0.0014789	0.050012	0.4404554	0.0018984	0.2597636	0.0055403	0.1931866
0.0018328	0.047371	0.5007198	0.0021485	0.3240641	0.0057734	0.1931322
0.0037408	0.0270913	0.7080054	0.0030422	0.535646	0.0058626	0.1930221
0.008501	0.0125141	1.076174	0.0046369	0.9565749	0.0058449	0.1930077

Figure 7 The 7 Basic Emotions Detected by Affectiva AFFDEX

3.3.2 Emotient FACET

Emotient FACET also processes the video based on frames. After recognising the face in the image, the system annotates 6 facial landmarks to obtain facial features. Then different AUs and expressions are classified using different classifiers. The classification process is the same as Affectiva AFFDEX, which is based on a statistical level, but there may be some differences in the results due to the different databases and the different number of facial landmarks used by the two systems [37].

There are 20 AUs recognised by Emotient FACET, with an additional AU23 (Lip Tightener) compared to Affectiva AFFDEX. Figure 8 lists these 20 AU codes, along with their descriptions as a reference on the first row, and gives sample results of 5 frames.

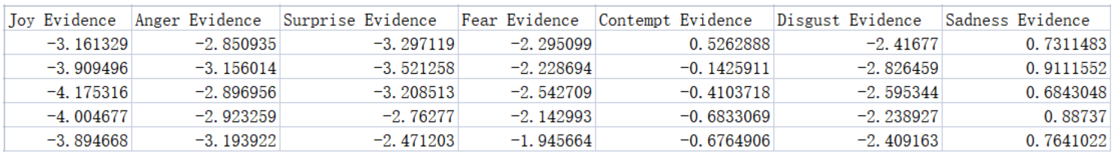


Inner Brow Raiser	Outer Brow Raiser	Brow Lowerer	Upper Lid Raiser	Cheek Raiser	Lid Tightener	Nose Wrinkler	Upper Lip Raiser	Lip Corner Puller	Dimpler
AU1 Evidence	AU2 Evidence	AU4 Evidence	AU5 Evidence	AU6 Evidence	AU7 Evidence	AU9 Evidence	AU10 Evidence	AU12 Evidence	AU14 Evidence
0.4481449	0.004267526	-0.3093545	-2.504631	-1.246143	-0.9874375	-2.558397	-1.203872	-0.521198	-0.05778512
0.675607	0.1151169	0.001386483	-2.454067	-1.210794	-0.9168178	-2.913588	-1.600012	-0.9475205	-0.4511598
0.4813637	0.0381402	-0.1227493	-2.658879	-1.148536	-0.8138057	-2.737941	-1.630224	-0.9802282	-0.5680426
0.8027101	0.2741565	0.1373758	-2.58797	-1.009037	-0.7921692	-2.60697	-1.468971	-0.9571247	-0.6769223
0.8147772	0.3444489	-0.02492982	-2.555492	-1.127604	-0.8570276	-2.965438	-1.519793	-0.9278665	-0.6994261

Lip Corner Depressor	Chin Raiser	Lip Puckerer	Lip stretcher	Lip Tightener	Lip Pressor	Lips part	Jaw Drop	Lip Suck	Eyes Closed
AU15 Evidence	AU17 Evidence	AU18 Evidence	AU20 Evidence	AU23 Evidence	AU24 Evidence	AU25 Evidence	AU26 Evidence	AU28 Evidence	AU43 Evidence
0.06466195	0.1644224	0.02918374	-0.6158326	-0.2016637	0.3653375	-1.439354	-0.8782421	-1.102965	-1.740857
-0.1929936	-0.182048	0.2687343	-0.7225891	-0.1907199	-0.2876832	-1.203177	-0.8536668	-1.774981	-2.095068
-0.1805829	-0.16466	0.3555079	-0.8847218	-0.179188	-0.2014808	-1.24973	-0.8428684	-1.927641	-1.946743
-0.2581333	-0.3234692	0.5426047	-0.7082098	-0.2756009	-0.5301512	-0.654362	-0.5308024	-1.934047	-1.626796
-0.2826185	-0.3674733	0.5232579	-0.5864858	-0.2316293	-0.6245176	-0.4881743	-0.4038427	-1.96974	-1.922876

Figure 8 The 20 AUs Detected by Emotient FACET

Similarly, Emotient FACET identifies 7 basic emotions, including joy, anger, surprise, fear, contempt, disgust and sadness, which are presented in Figure 9 along with sample results of 5 frames.



Joy Evidence	Anger Evidence	Surprise Evidence	Fear Evidence	Contempt Evidence	Disgust Evidence	Sadness Evidence
-3.161329	-2.850935	-3.297119	-2.295099	0.5262888	-2.41677	0.7311483
-3.909496	-3.156014	-3.521258	-2.228694	-0.1425911	-2.826459	0.9111552
-4.175316	-2.896956	-3.208513	-2.542709	-0.4103718	-2.595344	0.6843048
-4.004677	-2.923259	-2.76277	-2.142993	-0.6833069	-2.238927	0.88737
-3.894668	-3.193922	-2.471203	-1.945664	-0.6764906	-2.409163	0.7641022

Figure 9 The 7 Basic Emotions Detected by Emotient FACET

Unlike the Affectiva AFFDEX, the Emotient FACET outputs a log-likelihood evidence score for each AU or emotion. When the value is positive, it indicates that the probability of the action or emotion being present is greater than 50%, and when the value is negative, it indicates that the probability of the action or emotion being present is less than 50%. This evidence score can be converted to an intensity score (probability) in the range of 0 to 1 by using the formula in Figure 10.

$$Intensity = \frac{10^{Evidence}}{1 + 10^{Evidence}}$$

Figure 10 Formula for Calculating Intensity Score (Probability) from Log-likelihood Evidence Score in Emotient FACET

3.3.3 Probability of Emotion Every 2000ms

Since the emotion analysis of the speech is performed in units of 2000ms, the probability of emotion detected from the facial expression every 2000ms needs to be calculated to allow comparison of the emotion analysis results of the facial and speech.

The mean value was chosen to be used to calculate the probability of each emotion within 2000ms. For the calculation of the mean of probabilities, the geometric mean is more appropriate than the arithmetic mean, as the product of probabilities is more meaningful than the sum of probabilities. The formula for calculating the geometric mean is shown in Figure 11.

$$Geometric\ Mean = \sqrt[n]{\prod\nolimits_{i=1}^{n} x_n}$$

Figure 11 Formula for Geometric Mean

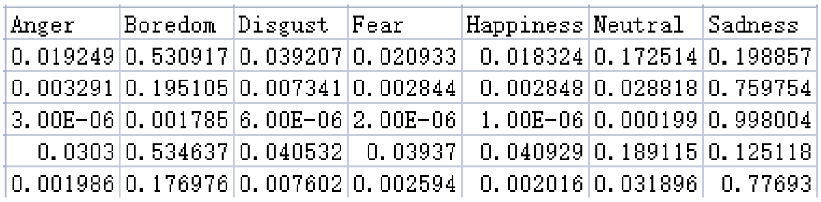
3.4 Audio Processing

For the 2000ms audio chunks, they are processed sequentially using openSmile and openVokaturi, which will eventually provide the audio features and emotion probabilities of each audio chunk. This section describes how to process audio using openSmile (3.4.1) and openVokaturi (3.4.2) and their outputs.

3.4.1 OpenSmile

OpenSmile is an open-source tool for audio feature extraction. The version used in this study is the latest version, openSmile 3.0, which can be downloaded from GitHub. A config folder is included in openSmile that contains configuration files for feature extraction and sentiment analysis.

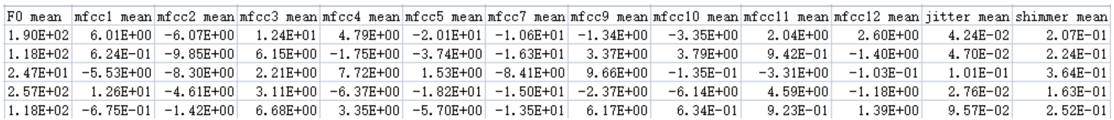
The configuration file in the openSmile/openEar 'emobase' set was used in this study. The set contains two configuration files for emotion recognition, which are 'emobase_live4_batch.conf' and 'emobase_live4_batch_single.conf'. The former applies to uncut audio, it will segment the audio based on energy and return a result for each segment. The latter applies to audio that has been chunked and it returns only a single result, hence it is more suitable for this study [22]. For each audio block, the system ultimately extracted 988 acoustic features and used them for sentiment recognition. These acoustic features include the 12 features of MFCC, F0, etc. The system recognised 7 emotions, including anger, boredom, disgust, fear, happiness, neutrality and sadness. Figure 12 shows the seven emotions detected by the system, along with sample results of 5 frames. The system gives the probability of each emotion and the sum of the probabilities of all emotions is 1.



Anger	Boredom	Disgust	Fear	Happiness	Neutral	Sadness
0.019249	0.530917	0.039207	0.020933	0.018324	0.172514	0.198857
0.003291	0.195105	0.007341	0.002844	0.002848	0.028818	0.759754
3.00E-06	0.001785	6.00E-06	2.00E-06	1.00E-06	0.000199	0.998004
0.0303	0.534637	0.040532	0.03937	0.040929	0.189115	0.125118
0.001986	0.176976	0.007602	0.002594	0.002016	0.031896	0.77693

Figure 12 The 7 Emotions Detected by OpenSmile

To obtain more acoustic features, the configuration file 'IS10_paraling.conf' is also used, which extracts 1582 acoustic features. In addition to MFCC and F0, it extracts jitterLocal and shimmerLocal, which will be used to analyse physical correlates of emotions. Based on the review of previous studies in Chapter 2, this study will validate the correlations between emotions and F0, MFCCs, jitterLocal and shimmerLocal. For each feature, the mean value will be chosen to represent the whole audio chunk. For the 12 MFCC feature values, their mean and standard deviation will be further calculated for the correlation analysis. Figure 13 shows the selected acoustic features, as well as an example of 5 frames.



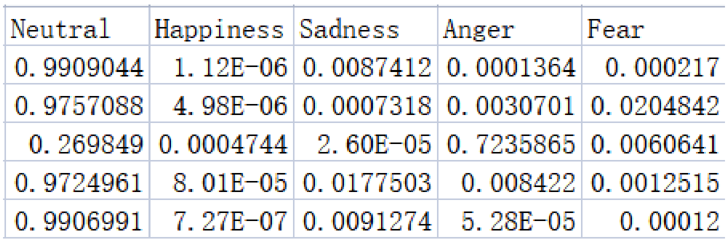
F0 mean	mfcc1 mean	mfcc2 mean	mfcc3 mean	mfcc4 mean	mfcc5 mean	mfcc7 mean	mfcc9 mean	mfcc10 mean	mfcc11 mean	mfcc12 mean	jitter mean	shimmer mean
1.90E+02	6.01E+00	-6.07E+00	1.24E+01	4.79E+00	-2.01E+01	-1.06E+01	-1.34E+00	-3.35E+00	2.04E+00	2.60E+00	4.24E-02	2.07E-01
1.18E+02	6.24E-01	-9.85E+00	6.15E+00	-1.75E+00	-3.74E+00	-1.63E+01	3.37E+00	3.79E+00	9.42E-01	-1.40E+00	4.70E-02	2.24E-01
2.47E+01	-5.53E+00	-8.30E+00	2.21E+00	7.72E+00	1.53E+00	-8.41E+00	9.66E+00	-1.35E-01	-3.31E+00	-1.03E-01	1.01E-01	3.64E-01
2.57E+02	1.26E+01	-4.61E+00	3.11E+00	-6.37E+00	-1.82E+01	-1.50E+01	-2.37E+00	-6.14E+00	4.59E+00	-1.18E+00	2.76E-02	1.63E-01
1.18E+02	-6.75E-01	-1.42E+00	6.68E+00	3.35E+00	-5.70E+00	-1.35E+01	6.17E+00	6.34E-01	9.23E-01	1.39E+00	9.57E-02	2.52E-01

Figure 13 Selected Acoustic Features Extracted by OpenSmile

3.4.2 OpenVokaturi

In this study, the latest version of openVokaturi, version 3.4, was used. The system can be freely downloaded on the official website and is free to use. There are other paid versions available on the website which have higher accuracy, but in this experiment, only the free version is considered.

OpenVokaturi recognises 5 basic emotions, including neutral, happiness, sadness, anger and fear. Figure 14 lists these 5 emotions and the sample results of 5 frames. Similar to openSmile, openVokaturi gives the probability of each emotion and the sum of the probabilities of all emotions is 1.



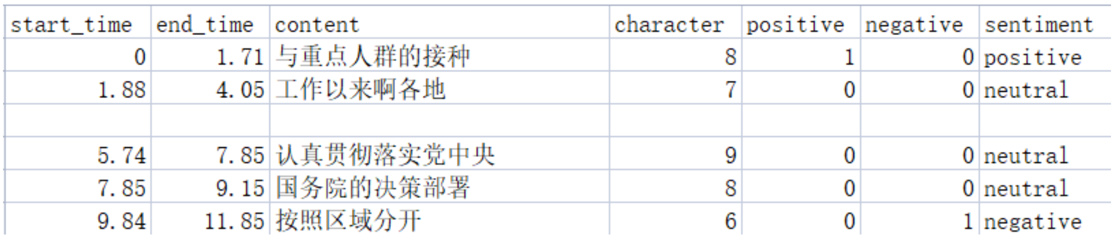
Neutral	Happiness	Sadness	Anger	Fear
0.9909044	1.12E-06	0.0087412	0.0001364	0.000217
0.9757088	4.98E-06	0.0007318	0.0030701	0.0204842
0.269849	0.0004744	2.60E-05	0.7235865	0.0060641
0.9724961	8.01E-05	0.0177503	0.008422	0.0012515
0.9906991	7.27E-07	0.0091274	5.28E-05	0.00012

Figure 14 The 5 Emotions Detected by OpenVokaturi

3.5 Text Processing

According to the JSON file returned by Amazon Transcribe, each word in the speech text and its timestamp can be obtained. After excluding stop words, two Chinese sentiment lexicons, HowNet and NTUSD, are used to label the sentiment polarity of these words. If one of the two lexicons marks the word as positive, the word has a positive sentiment polarity, and if one of the two lexicons marks the word as negative, the word has a negative sentiment polarity. If the two lexicons do not mark the sentiment polarity of the word as the same (i.e., one is positive and one is negative), the sentiment of the word is neutral.

As emotion analysis in audio is measured in 2000ms, to facilitate comparison of the results of the emotion analysis, the text data needs to be classified according to time. Texts within every 2000ms are combined according to their timestamp, and for texts which do not start and end in the same time period, compare the percentage of its time interval in the two time periods and divide it into the time period with the larger percentage. Then count the number of positive and negative words in each time period. If there are more positive words than negative words, the sentiment polarity in that 2000ms is marked as positive, if there are more negative words than positive words, the sentiment polarity in that 2000ms is marked as negative, and if there are no positive or negative words or both in equal numbers, the sentiment polarity in that 2000ms is marked as neutral. The final text analysis results are shown in Figure 15, which contains the text and the sentiment polarity analysed according to two sentiment lexicons for every 2000ms.



start_time	end_time	content	character	positive	negative	sentiment
0	1.71	与重点人群的接种	8	1	0	positive
1.88	4.05	工作以来啊各地	7	0	0	neutral
5.74	7.85	认真贯彻落实党中央	9	0	0	neutral
7.85	9.15	国务院的决策部署	8	0	0	neutral
9.84	11.85	按照区域分开	6	0	1	negative

Figure 15 Textual sentiment analysis Results (Based on HowNet and NTUSD)

3.6 Hypothesis Testing and Statistical Methods

In this study, the hypotheses are tested using non-parametric tests. In contrast to parametric tests, non-parametric tests can be used to test data where the distribution of the data is unknown, or where the distribution of the data does not satisfy a normal distribution. This is more applicable in this experiment. In the following subsections, the statistical methods used in the experiment will be described, which include the Mann-Whitney U test (3.6.1), Kruskal-Wallis test (3.6.2), Spearman rank correlation (3.6.3), McNemar test (3.6.4) and Cohen's Kappa (3.6.5). These methods are implemented via Python.

3.6.1 Mann-Whitney U Test

The Mann-Whitney U test is a non-parametric test used to test whether two groups of samples are from the same population, i.e., whether there is a significant difference in the distribution of two groups of samples. This test is applied to continuous variables and it determines whether there is a statistically significant difference in the distribution of two sets of data by using the size ranking of the data.

The H0 hypothesis of the test is that there is no difference in the distribution of two independent samples. When the p-value is greater than 0.05, H0 is accepted (i.e., there is no significant difference in the distribution of the two samples), and when the p-value is less than 0.05, H0 is rejected (i.e., there is a significant difference in the distribution of the two samples).

In this study, it is used to test whether there is a significant difference in the distribution of the detection results of 'anger'/'joy' for the same sample between two emotion recognition systems, etc. The test is implemented via the 'scipy.stats' library in Python.

3.6.2 Kruskal-Wallis Test

The Kruskal-Wallis Test is also a non-parametric test used to examine the distribution differences in samples. In contrast to the Mann-Whitney U test, it can be used to test for distribution differences in multiple groups of samples. It performs a test of distribution difference based on the ranking of the sample values.

The Kruskal-Wallis Test hypotheses that there is no significant difference in the distribution among multiple independent samples. When the P value is less than 0.05, the null hypothesis is rejected, i.e., the distributions of these samples are considered to be statistically significantly different from each other, otherwise, it is considered that the distributions of these samples are not significantly different from each other.

In this study, this test is used to detect whether the distribution of the detection results of ‘anger’/‘joy’ by the same emotion recognition system differs across age groups. The Kruskal-Wallis Test is implemented via the ‘scipy.stats’ library in Python.

3.6.3 Spearman Rank Correlation

The Spearman correlation coefficient can be used to test the correlation between two samples and is a non-parametric test. When the samples meet the conditions of being continuous, normally distributed and linearly correlated at the same time, the Pearson correlation coefficient can be used; otherwise, the Spearman correlation coefficient is used.

The correlation coefficient Rho ranges from -1 to 1, when it is positive, the two samples are positively correlated, when it is negative, the two samples are negatively correlated. The closer the absolute value of Rho is to 1, the stronger the correlation between the two samples, and the closer the absolute value of Rho is to 0, the weaker the correlation between the two samples. The strength of the correlation of the sample can be divided according to the Rho value as follows (See Table 6).

Table 6 Spearman's Rho Value and the Strength of Correlation

Absolute Value of Rho	Strength of the Correlation
0.00-0.20	Very Weak
0.21-0.40	Weak
0.41-0.60	Moderate
0.61-0.80	Strong
0.81-1.00	Very Strong

In this study, the Spearman correlation coefficient is used to analyse the correlation between ‘anger’/‘joy’ recognition results of different emotion recognition systems for the same sample, and verify the correlation between AUs and ‘anger’/‘joy’, etc.

The Spearman correlation coefficient can be calculated using the ‘corr’ function in Python, which calculates the Pearson correlation coefficient by default and needs to be set to the Spearman correlation coefficient. The ‘scipy.stats’ library also provides a method to calculate the Spearman correlation coefficient. In addition to the Rho value of the correlation coefficient, the method also returns a P value for determining the significance of the correlation. When the p-value is less than 0.05, the correlation between the two samples is significant.

3.6.4 McNemar Test

The McNemar test, also called paired chi-square test, which examines whether there is a statistically significant difference in the disagreement between two paired samples based on a 2×2 contingency table, and it is a non-parametric test.

In this study, the method is used to test whether the difference between the results of the textual sentiment analysis and the recognition results of the FER/SER system is statistically significant, etc. As the McNemar test is only applicable to binary categorical variables, the data needs to be processed before testing. Take the example of comparing whether the difference between the 'positive' emotion in the text analysis results and the 'joy' emotion in the emotion recognition system results is significant. For the FER/SER results, the possible emotions per 2000ms can be obtained by comparing the probabilities of each emotion. Next, the emotions other than 'positive' in the text analysis results are marked as 'not_positive' and the emotions other than 'joy' in the emotion recognition system results are marked as 'not_joy'. Finally, a contingency table can be obtained (See Table 7).

Table 7 Example Contingency Table for McNemar Test

		Facial/Speech Emotion	
		Joy	Not joy
Text Sentiment	Positive	A	B
	Not positive	C	D

The χ^2 can be calculated according to the formula in Figure 16. In Python, the 'statsmodels' library provides functions for the McNemar test that can be used directly.

$$\chi^2 = \frac{(B - C)^2}{B + C}$$

Figure 16 Formula for Calculating χ^2 in the McNemar Test

3.6.5 Cohen's Kappa

Kappa coefficients can be used to examine whether two samples are consistent or not, and this method is applied to category data. kappa coefficients range from -1 to 1, when the value is less than 0, the two samples are inconsistent, when the value is greater than 0, the two samples are consistent, the larger the value the better the consistency. The strength of consistency of the samples can be classified according to the Kappa values as follows (See Table 8).

Table 8 Kappa Value and the Strength of Consistency

Kappa	Strength of the Consistency
0.00-0.20	Very Weak
0.21-0.40	Weak
0.41-0.60	Moderate
0.61-0.80	Strong
0.81-1.00	Very Strong

In this study, Kappa will be used to test the consistency of the results of the textual sentiment analysis and the results of the sentiment recognition system for the same sample, etc.

Using the contingency table in Table 7 as an example, the formula for calculating the Kappa coefficient is shown in Figure 17. In Python, the ‘sklearn.metrics’ library provides functions to calculate Kappa coefficients.

$$P_o = \frac{A + D}{A + B + C + D}$$

$$P_e = \frac{(A + C) \times (A + B) + (B + D) \times (C + D)}{(A + B + C + D)^2}$$

$$Kappa = \frac{P_o - P_e}{1 - P_e}$$

Figure 17 Formula for Calculating Kappa

3.7 Summary

In this chapter, the pipeline for this study is presented. Then, a detailed description of the data collection and pre-processing, as well as the nationality, gender and age composition of the dataset is given. The approach to obtain emotion analysis data for facial, speech and text, as well as the content of the data are also described in detail. Finally, the statistical methods used for hypothesis testing in the experiment are described individually. In the next chapter, the results of the experiments are presented in detail and discussed.

Chapter 4 Case Study and Results

In this chapter, the consistency and differences in the emotion recognition results of Affectiva AFFDEX, Emotient FACET, openSmile and openVokaturi will be examined. Only two emotions were considered in this study, ‘anger’ and ‘joy’, which are two distinct emotions, and the other emotions are somewhere in between. Consistency and difference will be tested in samples of different races, genders and ages to see if these factors lead to differences in emotion recognition results. Also, two video datasets with emotion labels and one speech dataset with emotion labels will be used to compare the performance of two FER systems and two SER systems respectively, which provides a baseline for the whole experiment. For the Chinese dataset, the consistency and differences between the results of the textual sentiment analysis and the emotion recognition results of the four emotion recognition systems are also compared. Similarly, only ‘anger’ and ‘joy’ emotions are considered for this experiment, where ‘joy’ corresponds to positive emotions in textual sentiment analysis and ‘anger’ corresponds to negative emotions in the textual sentiment analysis. Finally, the physical correlates of emotions are verified with the Spearman correlation coefficient.

This chapter will show the results of the following case studies: comparison of emotion recognition results between Affectiva AFFDEX and Emotient FACET (4.1), comparison of emotion recognition results between openSmile and openVokaturi (4.2), comparison of emotion recognition results between FER systems and SER systems (4.3), comparison of results between textual sentiment analysis and emotion recognition systems (4.4), and verification of the physical correlates of emotions (4.5).

4.1 Affectiva AFFDEX and Emotient FACET

This section will show the performance of Affectiva AFFDEX and Emotient FACET on the datasets with emotion labels (4.1.1), and on the politician dataset (4.1.2) including the dataset of Chinese politicians (4.1.2.1) and the dataset of other politicians (4.1.2.2).

4.1.1 Performance on Datasets with Emotion Labels

RAVDESS contains 1440 videos recorded by 24 subjects. In this experiment, 720 videos from the first 12 subjects are used, with 96 videos for each of the 7 emotions ‘anger’, ‘calm’, ‘disgust’, ‘fear’, ‘happiness’, ‘sadness’ and ‘surprise’ and 48 videos for ‘neutral’. The videos are processed separately using Affectiva AFFDEX and Emotient FACET. There are 2 videos for which no data is obtained from Emotient FACET. The probability of each emotion is then calculated for each video using the geometric mean, and the emotion with the highest probability is selected as the predicted emotion for that video. A confusion matrix is eventually generated for each of the emotion recognition results of the two systems, with the horizontal axis representing the true emotions and the vertical axis representing the emotions predicted by the system (See Figure 18). It can be seen that Affectiva AFFDEX often confuses the facial expression of ‘joy’ with that of ‘surprise’, while Emotient FACET often identifies ‘calm’ as ‘joy’. For ‘anger’, both systems often misidentify the facial expression of ‘anger’ as ‘disgust’ and ‘surprise’. For the videos with emotion labels including ‘anger’, ‘disgust’, ‘fear’, ‘joy’, ‘sadness’ and ‘surprise’, the overall accuracy of Affectiva AFFDEX emotion recognition results is 37.33% and that of Emotient FACET is 59.03%.

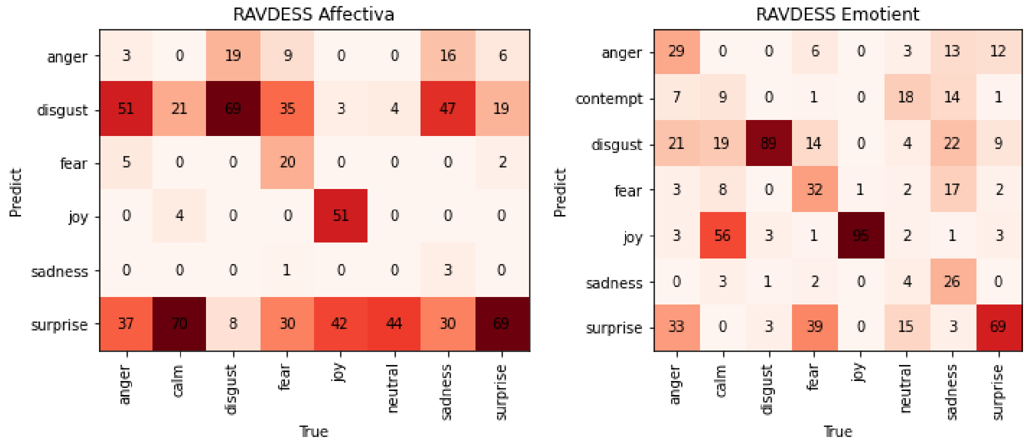


Figure 18 Confusion Matrix of Emotion Recognition Results of RAVDESS by Affectiva AFFDEX (Left) and Emotient FACET (Right)

Based on the confusion matrix the recognition accuracy and recall of the two systems for ‘joy’ and ‘anger’ can be calculated (See Table 9). It can be found that Affectiva AFFDEX has a higher recognition accuracy for ‘joy’, while Emotient FACET has a higher recognition recall for ‘joy’. For ‘anger’, both the recognition accuracy and recall are higher in Emotient FACET than in Affectiva AFFDEX.

Table 9 Accuracy and Recall of Affectiva AFFDEX and Emotient FACET for ‘Joy’ and ‘Anger’.

		Affectiva	**Emotient**
Joy	Accuracy	92.73%	57.93%
	Recall	53.13%	98.96%
Anger	Accuracy	5.66%	46.03%
	Recall	3.13%	30.21%

Next, the Mann-Whitney U test and Spearman correlation coefficient are used to analyse the differences and consistency between the detection results on ‘joy’ and ‘anger’ in two systems.

a) Mann-Whitney U test

The Mann-Whitney U test is used to test whether there is a significant difference between Affectiva AFFDEX and Emotient FACET in the detection of ‘joy’/’anger’. Following are the hypotheses for this test:

H0: There is no difference in the distribution of ‘joy’/‘anger’ detected by Affectiva AFFDEX and Emotient FACET.

H1: There exist differences in the distribution of ‘joy’/‘anger’ detected by Affectiva AFFDEX and Emotient FACET.

Table 10 Mann-Whitney U Test Results (RAVDESS)

	P Value	Result
Joy	3.0937e-84	Reject H0
Anger	3.2843e-39	Reject H0

The results of the Mann-Whitney U Test are shown in Table 10. The P-values obtained from the tests on ‘joy’ and ‘anger’ are 3.0937e-84 and 3.2843e-39, both of which are less than 0.05, therefore both reject the null hypothesis H0, which suggested that there is a statistically significant difference between Affectiva AFFDEX and Emotient FACET in the detection of ‘joy’ and ‘anger’.

b) Spearman Correlation Coefficient

Spearman correlation coefficient is used to check the consistency in detection of ‘joy’/‘anger’ by Affectiva AFFDEX and Emotient FACET. The results are shown in Table 11.

Table 11 Spearman correlation coefficient (RAVDESS)

	Rho		P Value
Joy	0.5719	Moderate	1.3848e-63
Anger	0.7990	Strong	2.4515e-160

It can be seen that the two systems have a moderate correlation for the detection of ‘joy’, a strong correlation for the detection of ‘anger’. The P values are all less than 0.05, proving that the correlation is statistically significant.

The other dataset, CASME II, contains 255 videos with the emotion label, 32 of which are ‘joy’, and the dataset does not contain videos with the ‘anger’ label. Using Affectiva AFFDEX and Emotient FACET to process these videos and calculate a predicted emotion for each video. There are 5 videos for which no data is obtained in both systems and 1 video for which no data is obtained in Affectiva AFFDEX. Finally, the two confusion matrices in Figure 19 are obtained. The confusion matrix shows that for the videos in CASME II, Affectivac AFFDEX tends to classify them as 'disgust' and Emotient FACET tends to classify them as ‘sadness’, ‘contempt’ and ‘disgust’, neither of which is ideal. For the videos with emotion labels including ‘disgust’, ‘fear’, ‘joy’, ‘sadness’ and ‘surprise’, the overall accuracy of Affectiva AFFDEX emotion recognition results is 44.35% and that of Emotient FACET is 15.32%.

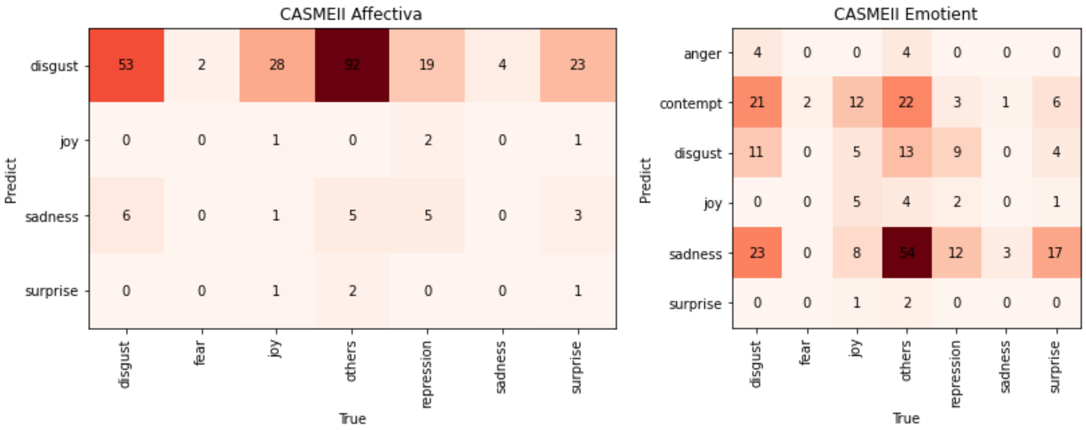


Figure 19 Confusion Matrix of Emotion Recognition Results of CASME II by Affectiva AFFDEX (Left) and Emotient FACET (Right)

Based on the confusion matrix the recognition accuracy and recall of the two systems for ‘joy’ can be calculated (See Table 12). It can be found that for the detection of ‘joy’, Emotient FACET has a high accuracy and recall rate.

Table 12 Accuracy and Recall of Affectiva AFFDEX and Emotient FACET for ‘Joy’.

		Affectiva	Emotient
Joy	Accuracy	25.00%	41.67%
	Recall	3.23%	16.13%

The differences and consistency between the detection results on ‘joy’/‘anger’ in the two systems are analysed using the Mann-Whitney U test and the Spearman correlation coefficient.

a) Mann-Whitney U test

Following are the hypotheses for this test:

H0: There is no difference in the distribution of ‘joy’/‘anger’ detected by Affectiva AFFDEX and Emotient FACET.

H1: There exist differences in the distribution of ‘joy’/‘anger’ detected by Affectiva AFFDEX and Emotient FACET.

Table 13 Mann-Whitney U Test Results (CASME II)

	P Value	Result
Joy	7.7040e-21	Reject H0
Anger	7.1001e-76	Reject H0

The results of the Mann-Whitney U Test are shown in Table 13. The P-values obtained from the tests on 'joy' and 'anger' are less than 0.05, therefore both reject the null hypothesis H0, which suggested that there is a statistically significant difference between Affectiva AFFDEX and Emotient FACET in the detection of 'joy' and 'anger'.

b) Spearman correlation coefficient

The consistency in the detection of 'joy'/'anger' by the 2 FER systems are shown in Table 14.

Table 14 Spearman Correlation Coefficient (CASME II)

	Rho		P Value
Joy	0.4981	Moderate	5.1205e-17
Anger	0.4297	Moderate	1.3076e-12

It can be seen that the two systems have a moderate correlation for the detection of 'joy' and 'anger', and the correlation on 'joy' detection is slightly stronger than the correlation on 'anger' detection. The P values are all less than 0.05, proving that the correlation is statistically significant.

4.1.2 Performance on Politician Datasets

The Mann-Whitney U Test and Kruskal-Wallis Test are used to test whether 'joy'/'anger' detected by the individual system is significantly different across ethnicities, genders and ages. The Mann-Whitney U Test is used for ethnicity and gender, and the Kruskal-Wallis Test is used for age. Following are the hypotheses:

H0: The distribution of 'joy'/'anger' in different ethnicities/genders/ages has no difference.

H1: The distribution of 'joy'/'anger' in different ethnicities/genders/ages has difference.

Table 15 Test Results for Differences in the Distribution of 'Joy'/'Anger' Across Ethnicity, Gender and Age (Affectiva AFFDEX and Emotient FACET).

	Affectiva			Emotient		
	Ethnicity (China, Others)	Gender (Female, Male)	Age (<45, 45-60, 60-75, >75)	Ethnicity (China, Others)	Gender (Female, Male)	Age (<45, 45-60, 60-75, >75)
Joy	P<<0.05	P<<0.05	P<<0.05	P<<0.05	P<<0.05	P<<0.05
Anger	P<<0.05	P<<0.05	P<<0.05	P<<0.05	P<<0.05	P<<0.05

From the results in Table 15, it can be found that the P value is less than 0.05 in all test results, i.e., in both Affectiva AFFDEX and Emotient FACET, the distribution of 'joy'/'anger' is significantly different in different ethnicities/genders/ages. The results suggest that ethnicity, gender and age influence the detection of ‘joy’ and ‘anger’ by Affectiva AFFDEX and Emotient FACET.

Figure 20 shows the peak of ‘anger’ detected for the same video in different systems. It can be noticed that there is a difference in the peak of ‘anger’ for the same video in different systems. In Affectiva AFFDEX, the peak of ‘anger’ occurs at 181.5 seconds with a probability of 69.90%. In Emotient FACET, the peak of ‘anger’ occurs at 222.24 seconds, with a probability of 61.68%. While one system considered the frame to be the peak of ‘anger’, another system gave the opposite result, considering the emotion of that frame to have a low probability of being ‘anger’.

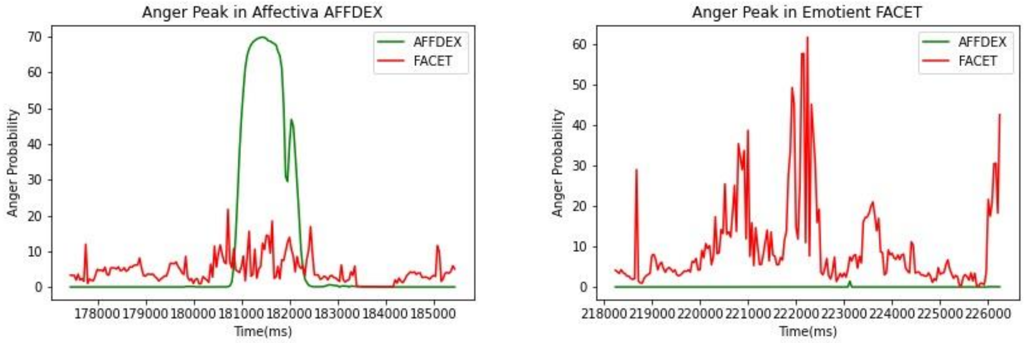


*Figure 20 The Peak of **Anger** in Affectiva AFFDEX (Left) and Emotient FACET (Right)*

A similar phenomenon is found in the peak of ‘joy’. As can be seen in Figure 21, for the same video, Affectiva AFFDEX considers the peak of joy occurs at 121.5 seconds, with a probability of 87.05%. Emotient FACET, on the other hand, believes that the peak of joy occurs at 169.08 seconds, with a probability of 64.58%.

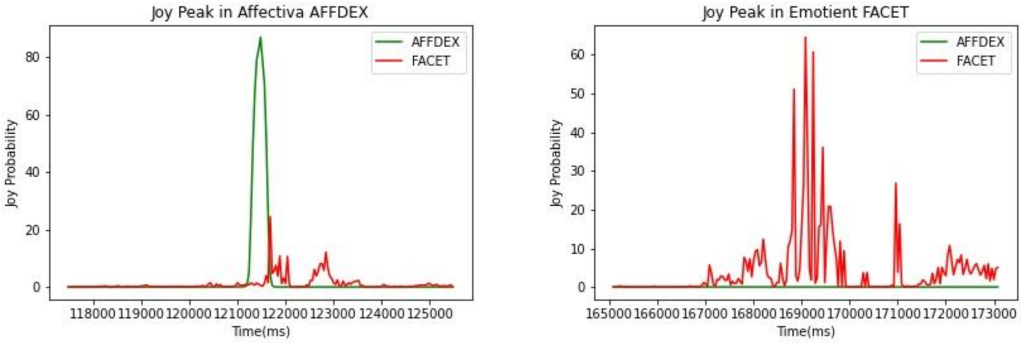


*Figure 21 The Peak of **Joy** in Affectiva AFFDEX (Left) and Emotient FACET (Right)*

It can be found that Affectiva AFFDEX and Emotient FACET differ in the detection of ‘joy’ and ‘anger’. Next, the differences and consistency between the two systems in the detection of ‘joy’ and ‘anger’ are examined with statistical methods.

4.1.2.1 Dataset of Chinese Politicians

After processing 45 speech videos of Chinese politicians with Affectiva AFFDEX and Emotient FACET, 364619 data (frames) are obtained from each system. The emotions in some of the frames are not correctly identified. After data cleaning, 340124 data (frames) from each system are retained.

Next, the differences and consistency of the detection results on 'joy'/'anger' in the two systems are analysed using the Mann-Whitney U test and Spearman correlation coefficient. The test is performed on samples from different genders and different ages.

a) Mann-Whitney U Test

Following are the hypotheses for this test:

H0: There is no difference in the distribution of 'joy'/'anger' detected by Affectiva AFFDEX and Emotient FACET.

H1: There exist differences in the distribution of 'joy'/'anger' detected by Affectiva AFFDEX and Emotient FACET.

Table 16 Mann-Whitney U Test Results of Dataset of Chinese Politicians (Affectiva AFFDEX and Emotient FACET)

	Joy	Anger
All Data	P<<0.05	P<<0.05
Female	P<<0.05	P<<0.05
Male	P<<0.05	P<<0.05
Age < 60	P<<0.05	P<<0.05
Age >= 60	P<<0.05	P<<0.05

As can be seen from Table 16, all p-values are less than 0.05 and therefore the null hypothesis H0 is rejected, i.e., for all samples, male samples, female samples, samples less than 60 years old, and samples more than 60 years old, the distribution of 'joy' and 'anger' detected by Affectiva AFFDEX and Emotient FACET all have statistically significantly difference.

b) Spearman Correlation Coefficient

As can be seen from Table 17, the consistency of 'joy'/'anger' detected by Affectiva AFFDEX and Emotient FACET is quite weak. The two systems have the best consistency in the detection of 'anger' for the male sample (Rho=0.3608). In most cases, the consistency between the two systems is better in 'anger', except in the female sample where the consistency in 'joy' (Rho=0.3029) is found to be better than the consistency in 'anger' (Rho=0.1089). All P values are less than 0.05, indicating that these correlations are statistically significant.

Table 17 Spearman Correlation Coefficient of Dataset of Chinese Politicians (Affectiva AFFDEX and Emotient FACET)

	Joy			Anger		
	Rho		P Value	Rho		P Value
All Data	0.1176	Very Weak	P<<0.05	0.3018	Weak	P<<0.05
Female	0.3029	Weak	P<<0.05	0.1089	Very Weak	P<<0.05
Male	0.0137	Very Weak	P<<0.05	0.3608	Weak	P<<0.05
Age < 60	0.1953	Very Weak	P<<0.05	0.2448	Weak	P<<0.05
Age >= 60	-0.0745	Very Weak	P<<0.05	0.2155	Weak	P<<0.05

4.1.2.2 Dataset of Other Politicians

The 212 videos are processed using Affectiva AFFDEX and Emotient FACET, resulting in 1372515 data (frames) from Affectiva AFFDEX and 1372512 data (frames) from Emotient FACET. Excluding frames where each system failed to correctly recognise the emotion, 1145859 data (frames) are ultimately retained from each system.

Then, using the Mann-Whitney U test and Spearman correlation coefficient to examine the differences and consistency of the detection results on ‘joy’/‘anger’ in the two systems.

a) Mann-Whitney U Test

The hypotheses for this test are as follows:

H0: There is no difference in the distribution of ‘joy’/‘anger’ detected by Affectiva AFFDEX and Emotient FACET.

H1: There exist differences in the distribution of ‘joy’/‘anger’ detected by Affectiva AFFDEX and Emotient FACET.

Table 18 Mann-Whitney U Test Results of Dataset of Chinese Politicians (Affectiva AFFDEX and Emotient FACET)

	Joy	Anger
All Data	P<<0.05	P<<0.05
Female	P<<0.05	P<<0.05
Male	P<<0.05	P<<0.05
Age < 45	P<<0.05	P<<0.05
Age: 45-59	P<<0.05	P<<0.05
Age: 60-74	P<<0.05	P<<0.05
Age >= 75	P<<0.05	P<<0.05

As can be seen from Table 18, all p-values are less than 0.05 and therefore the null hypothesis H0 is rejected, i.e., for all samples, male samples, female samples, samples under 45 years old, samples between 45 and 59 years old, samples between 60 and 74 years old, and samples over 75 years old, the distribution of ‘joy’ and ‘anger’ detected by Affectiva AFFDEX and Emotient FACET all have statistically significantly difference.

b) Spearman Correlation Coefficient

As can be seen from Table 19, the two systems have the best consistency in the detection of ‘anger’ for the sample over 75 years old (Rho=0.7119). The two systems have weak consistency for the detection of ‘joy’ and moderate to strong consistency for ‘anger’. The consistency for ‘anger’ is stronger than that of ‘joy’. These correlations are statistically significant as P values are all less than 0.05.

Table 19 Spearman Correlation Coefficient of Dataset of Chinese Politicians (Affectiva AFFDEX and Emotient FACET)

	Joy			Anger		
	Rho		P Value	Rho		P Value
All Data	0.3474	Weak	P<<0.05	0.5217	Moderate	P<<0.05
Female	0.2662	Weak	P<<0.05	0.4051	Moderate	P<<0.05
Male	0.3678	Weak	P<<0.05	0.4841	Moderate	P<<0.05
Age < 45	0.3423	Weak	P<<0.05	0.4692	Moderate	P<<0.05
Age: 45-59	0.3898	Weak	P<<0.05	0.5282	Moderate	P<<0.05
Age: 60-74	0.2851	Weak	P<<0.05	0.4654	Moderate	P<<0.05
Age >= 75	0.2755	Weak	P<<0.05	0.7119	Strong	P<<0.05

4.1.3 Discussion

The results of the experiments on dataset with emotion labels show that, for posed expressions in RAVDESS, Emotient FACET has higher overall accuracy than Affectiva AFFDEX in the recognition of six emotions, but for spontaneous expressions in CASME II, Affectiva AFFDEX has higher overall accuracy than Emotient FACET in the recognition of five emotions. This partially validates the experimental findings of Stöckli et al. [65] that for posed expressions and spontaneous expressions, Emotient FACET recognition accuracy was higher than Affectiva AFFDEX. The difference between the two experiments may be related to the different datasets used. For CASME II, Affectiva AFFDEX tended to identify the emotions of the videos as 'disgust' and the number of videos for each emotion in the dataset was inconsistent, which would affect the calculation of the overall accuracy. The experimental results also validated the finding of Krumhuber et al. [38] that Emotient FACET was less accurate in recognising spontaneous expression than posed expression.

Meanwhile, it can be found that for both RAVDESS and CASME II datasets, there is a significant difference in the distribution between the results of Affectiva AFFDEX and Emotient FACET for the detection of ‘joy’/‘anger’. The two systems have better consistency in the detection of ‘joy’/‘anger’ emotion in RAVDESS, which may be related to the fact that the videos in CASME II are spontaneous micro-expressions while videos in RAVDESS are posed expressions. In RAVDESS, the two systems have higher consistency for the detection of ‘anger’, while in CASME II, the two systems have higher consistency for the detection of ‘joy’.

From the experiments on the dataset of politicians, it can be found that the distribution of ‘joy’/‘anger’ detected by the two systems is found to be significantly different in samples of different ethnicities, genders and ages. In most cases, the two systems are more consistent in the detection of ‘anger’ than of ‘joy’. Stronger consistency in the detection of ‘joy’ than ‘anger’ is found only in the CASME II dataset and in the sample of Chinese female politicians.

By comparing the consistency of emotion recognition results between the two systems on RAVDESS and other datasets, it can be seen that the two systems are less consistent in recognizing spontaneous expressions than posed expressions. Based on the review of previous studies, this may be caused by the higher accuracy of both systems for recognising posed expressions and the lower accuracy of recognising spontaneous expressions. The two systems have different biases in the recognition of spontaneous expressions, which leads to a decrease in the consistency of the recognition results between the two systems.

Another finding is that the two systems are less consistent in recognising the facial expressions of Chinese politicians than those of politicians from other countries. This may be due to the fact that the training datasets of both systems contain only a small amount of Asian data, and therefore they are not trained enough for Asian facial expressions.

4.2 OpenSmile and OpenVokaturi

This section will show the performance of openSmile and openVokaturi on the datasets with emotion labels (4.2.1), and on the politician dataset (4.2.2) including the dataset of Chinese politicians (4.2.2.1) and the dataset of other politicians (4.2.2.2).

4.2.1 Performance on Datasets with Emotion Labels

The EMO-DB contains 535 audios with the emotion label, of which 71 are labelled as ‘happiness’ and 127 are labelled as ‘anger’. Processing these audios using openSmile and openVokaturi, both systems successfully recognised the emotion of all audios. The emotion with the highest probability is selected as the predicted emotion for that audio, and eventually, two confusion matrices are obtained in Figure 22. As can be seen from the figure, for most of the audio, both systems correctly recognised their emotions, with an overall accuracy of 82.43% for openSmile across the seven emotions and 84.80% for openVokaturi across the five emotions. In openSmile, some audio with the label 'fear' is incorrectly classified as ‘anger’/‘happiness’, and some audio with the label ‘happiness’ is incorrectly classified as ‘anger’.

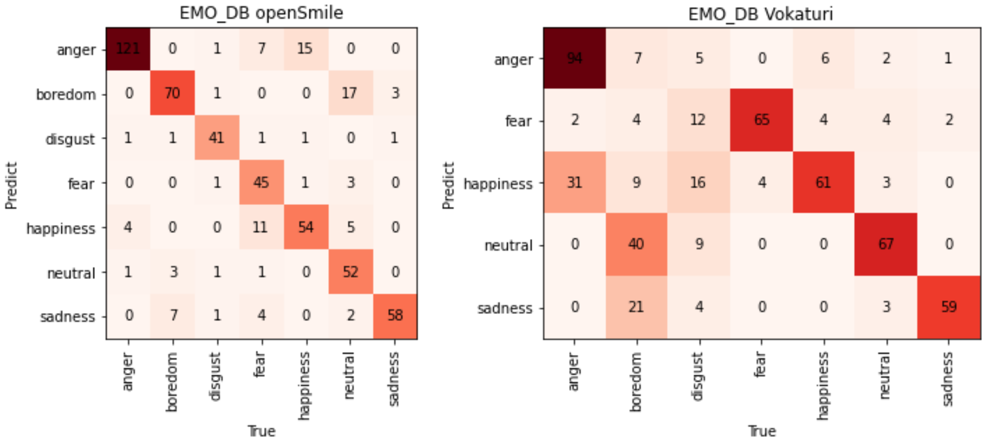


Figure 22 Confusion Matrix of Emotion Recognition Results of EMO-DB by OpenSmile (Left) and OpenVokaturi (Right)

Based on the confusion matrix the recognition accuracy and recall of the two systems for ‘happiness’ and ‘anger’ can be calculated (See Table 20). It can be found that openSmile is better than openVokaturi in the accuracy and recall of the recognition of ‘anger’. OpenVokaturi performs better than openSmile in the recall of ‘happiness’, but is weaker than openSmile in the accuracy of ‘happiness’.

Table 20 Accuracy and Recall of OpenSmile and OpenVokaturi for ‘Happiness’ and ‘Anger’.

		OpenSmile	OpenVokaturi
Happiness	Accuracy	72.97%	49.19%
	Recall	76.06%	85.91%
Anger	Accuracy	84.03%	81.74%
	Recall	95.28%	74.02%

Next, using Mann-Whitney U test and Spearman correlation coefficient to examine the differences and consistency between the detection results on ‘happiness’ and ‘anger’ in openSmile and openVokaturi.

a) Mann-Whitney U test

The Mann-Whitney U test is used to test whether there is a significant difference between openSmile and openVokaturi in the detection of ‘happiness’/‘anger’. The hypotheses are as follows:

H0: There is no difference in the distribution of ‘happiness’/‘anger’ detected by openSmile and openVokaturi.

H1: There exist differences in the distribution of ‘happiness’/‘anger’ detected by openSmile and openVokaturi.

Table 21 Mann-Whitney U Test Results (EMO-DB)

	P Value	Result
Happiness	0.1741	Accept H0
Anger	0.0242	Reject H0

The results of the Mann-Whitney U Test are shown in Table 21. The P values obtained from the tests on ‘happiness’ is 0.1741, therefore accepting the null hypothesis H0, i.e., no statistically significant differenexistsist in the distribution of ‘happiness’ detected by openSmile and openVokaturi. The P values obtained from the tests on ‘anger’ is 0.0242, therefore it suggests that significant difference exists in the distribution of ‘anger’ detected by openSmile and openVokaturi.

b) Spearman correlation coefficient

Spearman correlation coefficient is used to examine the consistency in detection of ‘happiness’/‘anger’ by openSmile and openVokaturi. The results are shown in Table 22.

Table 22 Spearman Correlation Coefficient (EMO-DB)

	Rho		P Value
Happiness	0.6065	Strong	4.8161e-55
Anger	0.6946	Strong	2.8682e-78

It can be seen that the two systems have a strong correlation for the detection of ‘happiness’/‘anger’. The consistency of the two systems in the detection of ‘anger’ is slightly better than the consistency of the detection of ‘happiness’. The P values are all less than 0.05, proving that the correlation is statistically significant.

4.2.2 Performance on Politician Datasets

Using Mann-Whitney U Test and Kruskal-Wallis Test to test whether ‘happiness’/‘anger’ detected by openSmile or openVokaturi is significantly different across ethnicities, genders and ages. The Mann-Whitney U Test is used for ethnicity and gender, and the Kruskal-Wallis Test is used for age. The hypotheses are as follows:

H0: The distribution of ‘happiness’/‘anger’ in different ethnicities/genders/ages has no difference.

H1: The distribution of ‘happiness’/‘anger’ in different ethnicities/genders/ages has difference.

Table 23 Test Results for Differences in the Distribution of ‘happiness’/‘anger’ Across Ethnicity, Gender and Age (OpenSmile and OpenVokaturi).

	OpenSmile			OpenVokaturi		
	Ethnicity (China, Others)	Gender (Female, Male)	Age (<45, 45-60, 60-75, >75)	Ethnicity (China, Others)	Gender (Female, Male)	Age (<45, 45-60, 60-75, >75)
Happiness	P<<0.05	P<<0.05	P<<0.05	P<<0.05	P<<0.05	P<<0.05
Anger	P<<0.05	P<<0.05	P<<0.05	P<<0.05	P<<0.05	P<<0.05

As can be seen in Table 23, the P value is less than 0.05 in all test results, i.e., in both openSmile and openVokaturi, the distribution of ‘happiness’/‘anger’ is significantly different in different ethnicities/genders/ages. Therefore, ethnicity, gender and age can have an impact on the detection of ‘happiness’ and ‘anger’ by openSmile and openVokaturi.

Figure 23 shows the peak of ‘anger’ detected for the same audio in openSmile and openVokaturi. In openSmile, the peak of ‘anger’ occurs at the 2nd chunk (2-4 seconds) with a probability of 2.61%. In openVokaturi, the peak of ‘anger’ occurs at the 113th chunk (226-228 seconds), with a probability of 99.99%. Differences are found in the two systems, with the probability of anger being considered small in openSmile, while in openVokaturi, anger is detected in some audio chunks with a high probability.

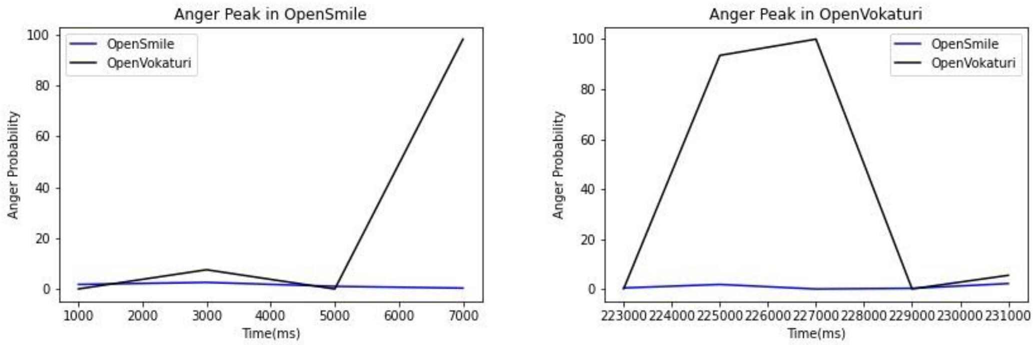


*Figure 23 The Peak of **Anger** in OpenSmile (Left) and OpenVokaturi (Right)*

For ‘happiness’, as can be observed from Figure 24, openSmile considers the peak of joy occurs at the 115th chunk (230-232 seconds), with a probability of 3.49%. OpenVokaturi believes that the peak of joy occursthe at 48th chunk (96-98 seconds), with a probability of 75.59%. In openSmile, no happiness is detected in the audio, while in openVokaturi, happiness is detected in some chunks with a high probability.

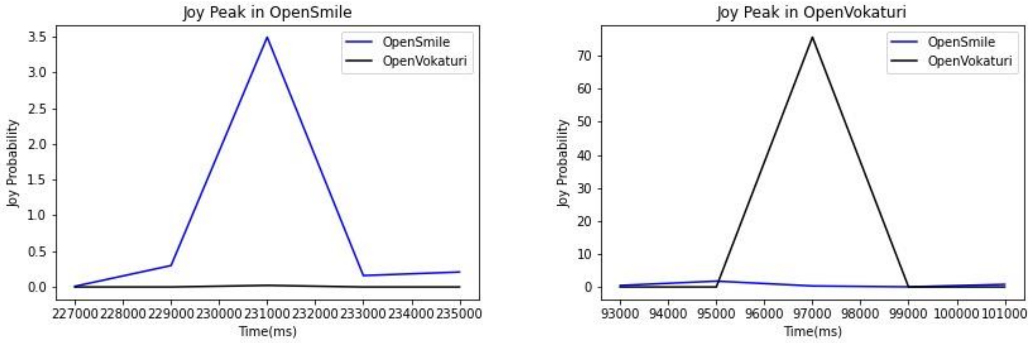


Figure 24 The Peak of ***Happiness*** *in OpenSmile (Left) and OpenVokaturi (Right)*

It can be found that openSmile and oopenVokaturi differ in the detection of ‘happiness’ and ‘anger’. Next, the differences and consistency between the two systems in the detection of ‘happiness’ and ‘anger’ are examined with statistical methods.

4.2.2.1 Dataset of Chinese Politicians

There are 6785 audio chunks in the audio dataset of Chinese politicians, which are generated from 45 audios. These audio chunks are processed using openSmile and openVokaturi, ultimately obtaining 6595 audio chunks with emotion recognition results in both systems.

Then, the Mann-Whitney U test and Spearman correlation coefficient are used to check the difference and consistency in ‘happiness’/‘anger’ detection by the two systems.

a) Mann-Whitney U Test

The hypotheses are as follows:

H0: There is no difference in the distribution of ‘happiness’/‘anger’ detected by openSmile and openVokaturi.

H1: There exist differences in the distribution of ‘happiness’/‘anger’ detected by openSmile and openVokaturi.

Table 24 Mann-Whitney U Test Results of Dataset of Chinese Politicians (OpenSmile and OpenVokaturi)

	Happiness	Anger
All Data	P<<0.05	P<<0.05
Female	P<<0.05	P<<0.05
Male	P<<0.05	P<<0.05
Age < 60	P<<0.05	P<<0.05
Age >= 60	P<<0.05	P<<0.05

As can be observed from Table 24, all P values are less than 0.05, thus the null hypothesis H0 is rejected, i.e., for all samples, male samples, female samples, samples under 60 years old, and samples over 60 years old, the distribution of ‘joy’ and ‘anger’ detected by openSmile and openVokaturi all have statistically significantly difference.

b) Spearman Correlation Coefficient

As can be seen from Table 25, For ‘happiness’, the P value detected in the male sample is greater than 0.05, and for ‘anger’, the P value detected in the male sample and the sample older than 60 years old is greater than 0.05, indicating that the correlation is not statistically significant in these samples. The results in the other samples showed that the two systems are weakly consistent in the detection of ‘happiness’/‘joy’, with slightly stronger consistency in the detection of ‘happiness’ than ‘anger’.

In addition, a negative correlation is found between the two systems for the detection of ‘happiness’ in the male sample, and for the detection of ‘happiness’ and ‘anger’ in the sample older than 60 years old.

Table 25 Spearman Correlation Coefficient of Dataset of Chinese Politicians (OpenSmile and OpenVokaturi)

	Happiness			**Anger**		
	Rho		P Value	Rho		P Value
All Data	0.1791	Very Weak	P<<0.05	0.1028	Very Weak	P<<0.05
Female	0.2698	Weak	P<<0.05	0.1861	Very Weak	P<<0.05
Male	-0.0231	Very Weak	0.1021	0.0107	Very Weak	0.4501
Age < 60	0.3030	Weak	P<<0.05	0.1616	Very Weak	P<<0.05
Age >= 60	-0.1242	Very Weak	P<<0.05	-0.0409	Very Weak	0.0795

4.2.2.2 Dataset of Other Politicians

There are 209 speech audios of politicians from other countries. After segmentation, these audios generate 22,865 audio chunks of 2000ms in duration. These audio chunks are processed using openSmile and openVokaturi, ultimately obtaining 22207 audio chunks with emotion recognition results in both systems.

Then, the Mann-Whitney U test and Spearman correlation coefficient are used to check the difference and consistency in ‘happiness’/‘anger’ detection by the two systems.

a) Mann-Whitney U Test

The hypotheses are as follows:

H0: There is no difference in the distribution of ‘happiness’/‘anger’ detected by openSmile and openVokaturi.

H1: There exist differences in the distribution of ‘happiness’/‘anger’ detected by openSmile and openVokaturi.

Table 26 Mann-Whitney U Test Results of Dataset of Other Politicians (OpenSmile and OpenVokaturi)

	Happiness	Anger
All Data	P<<0.05	P<<0.05
Female	P<<0.05	P<<0.05
Male	P<<0.05	P<<0.05
Age < 45	P<<0.05	P<<0.05
Age: 45-59	P<<0.05	P<<0.05
Age: 60-74	P<<0.05	P<<0.05
Age >= 75	P<<0.05	P<<0.05

As can be observed from Table 26, all p-values are less than 0.05, which indicates the null hypothesis H0 is rejected, i.e., for all samples, male samples, female samples, samples under 45 years old, samples between 45 and 59 years old, samples between 60 and 74 years old, and samples over 75 years old, the distribution of ‘joy’ and ‘anger’ detected by openSmile and openVokaturi all have statistically significantly difference.

b) Spearman Correlation Coefficient

As can be seen from Table 27, the P values are smaller than 0.05 in all samples tested, indicating that the correlation is statistically significant. Consistency on ‘happiness’ is stronger than that on ‘anger’ in the whole sample, the female sample, the sample below 45 years old, the sample between 45 and 59 years old and the sample between 60 and 74 years old. In contrast, in the male sample and the sample above 75 years old, the consistency on ‘anger’ is stronger than that on ‘happiness’.

Table 27 Spearman Correlation Coefficient of Dataset of Other Politicians (OpenSmile and OpenVokaturi)

	Happiness			Anger		
	Rho		P Value	Rho		P Value
All Data	0.2978	Weak	P<<0.05	0.2273	Weak	P<<0.05
Female	0.1709	Very Weak	P<<0.05	0.1216	Very Weak	P<<0.05
Male	0.2169	Weak	P<<0.05	0.2411	Weak	P<<0.05
Age < 45	0.3826	Weak	P<<0.05	0.2238	Weak	P<<0.05
Age: 45-59	0.2528	Weak	P<<0.05	0.2091	Weak	P<<0.05
Age: 60-74	0.2702	Weak	P<<0.05	0.2197	Weak	P<<0.05
Age >= 75	0.1432	Very Weak	P<<0.05	0.2485	Weak	P<<0.05

4.2.3 Discussion

From the above experimental results, it can be found that for the trained dataset EMO-DB, there is a strong correlation between the two systems on the detection of ‘happiness’ and ‘anger’, the correlation on ‘anger’ is stronger than that on ‘happiness’, and the distribution of ‘happiness’ is not significantly different in the two systems. For the dataset that the systems had not learned, the two systems showed weaker consistency in the detection of ‘happiness’ and ‘anger’. In tests with samples of different ethnicities, languages, genders, and ages, it is found that the distribution of ‘happiness’/‘anger’ detected by the two systems for the same sample is significantly different. In most of the samples tested, the consistency of the two systems for ‘happiness’ detection is found to be better than that for ‘anger’.

By comparing the outcome of the experiments on the dataset of Chinese politicians, which contains only Mandarin audio, and the dataset of other politicians, where the vast majority of the audio is in English, it can be found that the two systems have worse consistency in the emotion detection of Mandarin audio, with the results of the two systems being negatively correlated in some samples, and the correlation is not significant in some samples. This may be related to the fact that neither of the two systems has been trained on a dataset containing Mandarin audio.

4.3 Facial Expression and Speech

Video and audio have different time units when recognising emotions, in the video, the emotion is recognised on average every 35ms for one frame and in the audio, the emotion is recognised every 2000ms. Therefore, the emotion data from the Affectiva AFFDEX and Emotient FACET need to be processed. The probability of emotion per 2000ms in videos is calculated using the geometric mean. Then, in all 4 systems, the emotion with the highest probability is selected as the predicted emotion for that 2000ms. Finally, the predicted emotion is converted to a binary variable based on whether it is ‘joy’/‘anger’ or not. The McNemar test and Cohen’s Kappa are used to detect differences and consistency in the detection of ‘joy’/‘anger’ between FER systems and SER systems. Results of tests on the dataset of Chinese politicians (4.3.1) and the dataset of other politicians (4.3.2) will be presented.

4.3.1 Dataset of Chinese Politicians

Emotion recognition data from 6522 audio/video blocks are used in this experiment, and the emotions of these audio/video blocks are recognised in all four systems.

a) McNemar Test

The McNemar test is used to detect the difference between the detection results of the FER system and the SER system. The hypothesis is as follows:

H0: There is no difference in the detection of 'joy'/'anger' between the FER system and the SER system.

H1: There is a difference in the detection of 'joy'/'anger' between the FER system and the SER system.

*Table 28 McNemar Test Results in Dataset of Chinese Politicians (**Joy**)*

	Affectiva and openSmile	Affectiva and openVokaturi	Emotient and openSmile	Emotient and openVokaturi
All Data	P<<0.05	P<<0.05	P<<0.05	0.4647
Female	P<<0.05	P<<0.05	P<<0.05	P<<0.05
Male	P<<0.05	P<<0.05	P<<0.05	P<<0.05
Age < 60	P<<0.05	P<<0.05	P<<0.05	0.1753
Age >= 60	P<<0.05	P<<0.05	P<<0.05	P<<0.05

From the results in Table 28, it can be found that in the test of detection results for 'joy' by Emotient FACET and openVokaturi, P values of 0.4647 and 0.1753 are obtained for the whole sample and the sample below 60 years old, which indicates that in these two samples, Emotient FACET and openVokaturi have no significant difference in the detection results of 'joy'. In all other tests, significant differences are found in the detection of 'joy' between the FER system the and SER system.

*Table 29 McNemar Test Results in Dataset of Chinese Politicians (**Anger**)*

	Affectiva and openSmile	Affectiva and openVokaturi	Emotient and openSmile	Emotient and openVokaturi
All Data	P<<0.05	P<<0.05	P<<0.05	P<<0.05
Female	P<<0.05	P<<0.05	0.0578	P<<0.05
Male	P<<0.05	P<<0.05	P<<0.05	P<<0.05
Age < 60	P<<0.05	P<<0.05	P<<0.05	P<<0.05
Age >= 60	P<<0.05	P<<0.05	0.3173	P<<0.05

As can be seen from the results in Table 29, in the test of the detection results for 'anger' by Emotient FACET and openSmile, P values of 0.0578 and 0.3173 are obtained in the female sample and the sample older than 60 years old, demonstrating that Emotient FACET and openSmile do not differ significantly in the detection of 'anger' in these two samples. In all other tests, significant differences are found in the detection of 'anger' between the FER system the and SER system.

b) Cohen's Kappa

Kappa is used to check the consistency of 'joy'/'anger' detected by the FER system and the SER system.

*Table 30 Kappa Results in Dataset of Chinese Politicians (**Joy**)*

	Affectiva and openSmile	Affectiva and openVokaturi	Emotient and openSmile	Emotient and openVokaturi
All Data	-0.0003	0.0119	-0.0003	-0.0529
Female	-0.0012	0.0102	-0.0013	-0.0228
Male	0.0	-0.0023	0.0	-0.1053
Age < 60	-0.0004	0.0194	-0.0004	-0.0288
Age >= 60	0.0	-0.0075	0.0	-0.1246

*Table 31 Kappa Results in Dataset of Chinese Politicians (**Anger**)*

	Affectiva and openSmile	Affectiva and openVokaturi	Emotient and openSmile	Emotient and openVokaturi
All Data	-0.0024	0.0156	-0.0021	0.0001
Female	-0.0011	-0.0013	-0.0021	-0.0026
Male	0.0	0.0278	0.0	0.0018
Age < 60	-0.0034	-0.0027	-0.0029	-0.0015
Age >= 60	0.0	0.0642	0.0	0.0032

As can be seen from Table 30 and Table 31, the kappa values are all very low, indicating that these FER systems and SER systems are very weak in consistency or even inconsistent in the detection of ‘joy’/‘anger’.

4.3.2 Dataset of Other Politicians

Emotion recognition data from 20862 audio/video blocks are used in this experiment, and the emotions of these audio/video blocks are recognised in all four systems.

c) McNemar Test

The hypothesis is as follows:

H0: There is no difference in the detection of ‘joy’/‘anger’ between the FER system and the SER system.

H1: There is a difference in the detection of ‘joy’/‘anger’ between the FER system and the SER system.

*Table 32 McNemar Test Results in Dataset of Other Politicians (**Joy**)*

	Affectiva and openSmile	Affectiva and openVokaturi	Emotient and openSmile	Emotient and openVokaturi
All Data	P<<0.05	P<<0.05	P<<0.05	P<<0.05
Female	P<<0.05	P<<0.05	P<<0.05	P<<0.05
Male	P<<0.05	P<<0.05	P<<0.05	P<<0.05
Age < 45	P<<0.05	P<<0.05	P<<0.05	P<<0.05
Age: 45-59	P<<0.05	P<<0.05	P<<0.05	P<<0.05
Age: 60-74	P<<0.05	P<<0.05	P<<0.05	P<<0.05

Age >= 75	P<<0.05	P<<0.05	P<<0.05	P<<0.05

*Table 33 McNemar Test Results in Dataset of Other Politicians (**Anger**)*

	Affectiva and openSmile	Affectiva and openVokaturi	Emotient and openSmile	Emotient and openVokaturi
All Data	P<<0.05	P<<0.05	P<<0.05	P<<0.05
Female	P<<0.05	P<<0.05	P<<0.05	P<<0.05
Male	P<<0.05	P<<0.05	P<<0.05	P<<0.05
Age < 45	P<<0.05	P<<0.05	P<<0.05	P<<0.05
Age: 45-59	P<<0.05	P<<0.05	P<<0.05	P<<0.05
Age: 60-74	P<<0.05	P<<0.05	P<<0.05	P<<0.05
Age >= 75	P<<0.05	P<<0.05	P<<0.05	P<<0.05

From the results in Table 32 and Table 33, it can be found that all the tests obtained P values less than 0.05, indicating that the null hypothesis H0 should be rejected, i.e., there are statistically significant differences in the detection of ‘joy’/‘anger’ between these facial emotion detection systems and speech emotion detection systems.

d) Cohen’s Kappa

The consistency of ‘joy’/‘anger’ detected by the FER system and the SER system are as follows.

*Table 34 Kappa Results in Dataset of Other Politicians (**Joy**)*

	Affectiva and openSmile	Affectiva and openVokaturi	Emotient and openSmile	Emotient and openVokaturi
All Data	0.0245	0.0067	0.0113	0.0358
Female	0.0259	-0.0087	0.0085	-0.0225
Male	-0.0003	-0.0134	-0.0003	-0.0496
Age < 45	0.0372	-0.0142	0.0029	-0.059
Age: 45-59	0.0264	0.0151	0.0222	0.004
Age: 60-74	0.0173	-0.0084	0.0139	0.0611
Age >= 75	-0.0024	0.0818	-0.0025	0.1778

*Table 35 Kappa Results in Dataset of Other Politicians (**Anger**)*

	Affectiva and openSmile	Affectiva and openVokaturi	Emotient and openSmile	Emotient and openVokaturi
All Data	-0.018	0.0334	-0.0143	0.0214
Female	-0.0254	-0.0279	-0.0079	-0.0044
Male	-0.0013	0.0669	0.0005	0.0332
Age < 45	-0.0485	-0.0487	-0.0459	-0.0023
Age: 45-59	-0.0058	0.0314	-0.0042	0.0102
Age: 60-74	-0.0105	0.0571	0.0047	0.0177
Age >= 75	0.0	0.1886	0.0	0.0832

As can be seen from Table 34 and Table 35, the kappa values are very low in all cases, indicating that these FER systems and SER systems are very weak in consistency or even inconsistent in the detection of ‘joy’/‘anger’. Only for the sample older than 75 years old, where a relatively better consistency is found between Emotient FACET and openVokaturi for the detection of ‘joy’ (Kappa=0.1778) and Affectiva AFFDEX and openVokaturi for the detection of ' anger' (Kappa=0.1886), but these consistencies are still very weak.

4.3.3 Discussion

The above experiments show that there is no significant difference found between Emotient FACET and openSmile for the detection of ‘anger’ and between Emotient FACET and openVokaturi for the detection of ‘joy’ in a few samples. Meanwhile, the FER systems and the SER systems have a very weak consistency in the recognition of ‘joy’/‘anger’. This phenomenon exists in samples of different races, languages, genders and ages.

This may be related to the fact that different emotion recognition systems have different biases when performing emotion recognition, and these biases increase the differences in emotion recognition results.

In addition to this, it can also be caused by ‘emotion leakage’, where politicians may try to control their facial expressions and voices to some extent when speaking in public, leading to a situation where the emotions detected by the systems do not match the actual emotions.

4.4 Text, Facial Expression and Speech

Using Amazon Transcribe to transcribe the content of the speech, 45 transcribed texts were obtained, along with each word and its corresponding timestamp, and other information, ultimately 27462 words are obtained. Among these words, the most frequent occurrences are words like ‘的’(of), ‘我们’(we) and some intonation words. After removing stop words, Table 36 shows the 10 most frequent words.

Table 36 Top 10 most frequent words

Order	Word	Frequency	Order	Word	Frequency
1	工作 (job)	135	6	防 (prevent)	99
2	发展 (develop)	117	7	方面 (aspects)	98
3	疫情 (epidemic)	114	8	健康 (health)	90
4	控 (control)	109	9	人员 (personnel)	85
5	中国 (China)	100	10	问题 (problem)	84

Based on the timestamps, the duration of saying each word and the gap between each word are also calculated. The average duration of each word is 0.38 seconds and the average gap between words is 0.11 seconds. In addition, reference texts are obtained by manual transcription, and the character error rate (CER) of the Amazon Transcribe transcription results is calculated to be 5.30%.

Using HowNet and NTUSD, the sentiment polarity of the words is annotated, resulting in 3452 words being annotated, of which 2339 are positive and 1113 are negative.

Next, the McNemar test and Cohen's kappa are used to examine the differences and consistency between the results of textual sentiment analysis and emotion recognition systems on the detection of 'joy'('positive')/'anger'('negative').

a) McNemar Test

The McNemar test is used to detect the difference between the results of the textual sentiment analysis and emotion recognition systems. The hypothesis is as follows:

H0: There is no difference in the detection of 'joy'('positive')/'anger'('negative') between the textual sentiment analysis and the emotion recognition system.

H1: There is a difference in the detection of 'joy'('positive')/'anger'('negative') between the textual sentiment analysis and the emotion recognition system.

*Table 37 McNemar Test Results in Dataset of Chinese Politicians (**Joy/Positive**)*

	Text and Affectiva	Text and Emotient	Text and openSmile	Text and openVokaturi
All Data	P<<0.05	P<<0.05	P<<0.05	P<<0.05
Female	P<<0.05	P<<0.05	P<<0.05	0.0812
Male	P<<0.05	P<<0.05	P<<0.05	P<<0.05
Age < 60	P<<0.05	P<<0.05	P<<0.05	P<<0.05
Age >= 60	P<<0.05	P<<0.05	P<<0.05	P<<0.05

The results in Table 37 show that most of the tests obtained a P value lower than 0.05, indicating that there is a significant difference between the textual sentiment analysis and the emotion recognition system in the detection of 'joy' ('positive'). Only in the female sample, no significant difference is found between the results of openVokaturi and text analysis on 'joy' ('positive').

*Table 38 McNemar Test Results in Dataset of Chinese Politicians (**Anger/Negative**)*

	Text and Affectiva	Text and Emotient	Text and openSmile	Text and openVokaturi
All Data	P<<0.05	P<<0.05	P<<0.05	P<<0.05
Female	P<<0.05	P<<0.05	P<<0.05	P<<0.05
Male	0.1652	P<<0.05	P<<0.05	P<<0.05
Age < 60	P<<0.05	P<<0.05	P<<0.05	P<<0.05

Age >= 60	P<<0.05	P<<0.05	P<<0.05	P<<0.05

It can be observed from the results in Table 38 that most of the tests obtained a P value lower than 0.05, indicating that there is a significant difference between the textual sentiment analysis and the emotion recognition system in the detection of ‘joy’ (‘positive’). Only in the male sample, no significant difference is found between the results of Affectiva AFFDEX and text analysis on ‘joy’ (‘positive’).

b) Cohen’s Kappa

Kappa is used to check the consistency of ‘joy’(‘positive’)/‘anger’(‘negative’) detected by the textual sentiment analysis and the emotion detection system.

*Table 39 Kappa Results in Dataset of Chinese Politicians (**Joy/Positive**)*

	Text and Affectiva	Text and Emotient	Text and openSmile	Text and openVokaturi
All Data	0.004	0.0653	-0.0003	0.0054
Female	0.0106	0.0578	-0.0013	0.0333
Male	0.0024	0.0692	0.0	-0.0015
Age < 60	0.006	0.0761	-0.0004	0.0077
Age >= 60	0.0024	0.0593	0.0	0.0001

*Table 40 Kappa Results in Dataset of Chinese Politicians (**Anger/Negative**)*

	Text and Affectiva	Text and Emotient	Text and openSmile	Text and openVokaturi
All Data	-0.0208	-0.0034	-0.0024	-0.0274
Female	0.0074	-0.0026	-0.01	-0.0483
Male	-0.0169	-0.0023	0.0	-0.0229
Age < 60	-0.0204	-0.0056	-0.0034	-0.0183
Age >= 60	-0.0362	-0.0011	0.0	-0.0612

From the Kappa in Table 39 and Table 40, it can be found that the textual sentiment analysis and emotion recognition systems are very poor in consistency in the detection of ‘joy’ (‘positive’)/‘anger’ (‘negative’). In ‘anger’ (‘negative’), most of the sentiment analysis results are not consistent, while in ‘joy’ (‘positive’), the results are slightly better, but the consistency of the sentiment analysis results is still very weak.

From the above experimental results, it can be found that in most cases, the textual sentiment analysis and emotion recognition systems differ significantly in their results and have a very weak consistency. There are several possible reasons for this result.

First, due to the limitations of the Chinese sentiment lexicon, HowNet and NTUSD used in this experiment were created at an earlier time, and some new sentiment words have not been added to the lexicon. Meanwhile, NTUSD was created by a university in Taiwan, while the politicians collected in the experiment were all from mainland China and there may be differences in language expressions and vocabulary usage. These may lead to inaccurate results of the textual sentiment analysis.

Secondly, the emotion recognition system also has biases in emotion recognition, as demonstrated by previous experimental results, the system may confuse some of the emotions, resulting in emotions being classified incorrectly. Also, previous experiments have found that these emotion recognition systems have poor consistency in the recognition of emotions of Chinese politicians compared to other politicians.

Third, differences in emotion recognition results may be due to 'emotional leakage'. When politicians speak in public, they may control their facial expressions and speech to a certain extent, which leads to the fact that when words with positive emotions are spoken, joy may not be detected in the face or voice.

4.5 Physical Correlates of Emotions

In this chapter, the physical correlates of emotions will be validated in the politician dataset, and the Spearman correlation coefficient is used to examine the correlation of ‘joy’/‘anger’ with action units (4.5.1), and acoustic features (4.5.2).

4.5.1 Action Units

The Spearman correlation coefficient is used to detect the correlation between the action units and ‘joy’/‘anger’ in the emotion recognition results of Affectiva AFFDEX and Emotient FACET. A threshold value of 0.4 is chosen, which indicates that the correlation strength is moderate or above. If there is no correlation greater than 0.4, the action unit with the highest correlation is listed. The final results are shown below, with the corresponding Spearman correlation coefficient labelled after each action unit.

*Table 41 Correlation between Action Units and **‘Joy’** (Affectiva AFFDEX)*

	Chinese		Others	
	AU	Rho	AU	Rho
All Data	**AU25**	**0.35**	**AU12**	**0.47**
Female	**AU25** AU6 AU9 AU12 AU4 AU10	**0.64** 0.45 0.45 0.45 0.41 0.4	**AU12** AU6	**0.48** 0.42

Male	**AU25**	**0.32**	**AU12**	**0.44**
Age < 60	AU12	0.38	**AU12**	**0.5**
Age >= 60	**AU25**	**0.51**	**AU12**	**0.43**

Table 41 shows results from Affectiva AFFDEX. It can be found that for politicians from other countries, 'joy' is found to be correlated with AU12 in different samples. In contrast, for Chinese politicians, 'joy' is only found to be better correlated with AU12 in the female sample and the sample younger than 60 years old, while AU25 is found to be better correlated with 'joy' in most of the samples.

Table 42 Correlation between Action Units and ***'Joy'*** *(Emotient FACET)*

	Chinese		Others	
	AU	Rho	AU	Rho
All Data	**AU12**	**0.86**	**AU12**	**0.88**
	AU14	**0.72**	**AU14**	**0.57**
	AU28	0.66	AU28	0.4
	AU24	0.46		
Female	**AU12**	**0.86**	**AU12**	**0.91**
	AU14	**0.55**	AU6	0.63
	AU20	0.55	**AU14**	**0.61**
	AU10	0.49	AU28	0.46
	AU28	0.48		
Male	**AU12**	**0.83**	**AU12**	**0.86**
	AU14	**0.73**	**AU14**	**0.51**
	AU28	0.69		
	AU24	0.5		
Age < 60	**AU12**	**0.87**	**AU12**	**0.87**
	AU14	**0.67**	**AU14**	**0.55**
	AU28	0.6	AU6	0.43
			AU28	0.4
Age >= 60	**AU12**	**0.88**	**AU12**	**0.89**
	AU14	**0.85**	**AU14**	**0.6**
	AU28	0.83		
	AU9	0.65		
	AU24	0.63		
	AU23	0.52		
	AU4	0.49		
	AU43	0.42		
	AU6	0.41		

In Emotient FACET (See Table 42), on the other hand, AU12 and AU14 are found to be strongly correlated with 'joy' across samples of different ethnicities, genders and ages, and AU28 is also found to be correlated with 'joy' in most samples.

Table 43 Correlation between Action Units and ***'Anger'*** *(Affectiva AFFDEX)*

	Chinese		Others	
	AU	Rho	AU	Rho
All Data	**AU4**	**0.56**	**AU4**	**0.71**
	AU9	0.41	AU7	0.48
Female	AU25	0.53	**AU4**	**0.55**
Male	**AU4**	**0.62**	**AU4**	**0.8**
	AU9	0.46	AU7	0.57
			AU9	0.4
Age < 60	**AU4**	**0.59**	**AU4**	**0.75**
			AU7	0.4
Age >= 60	**AU4**	**0.51**	**AU4**	**0.68**
			AU7	0.59
			AU9	0.44

As can be seen in Table 43, AU4 is found to be well correlated with 'anger' in most of the samples in Affectiva AFFDEX, except for the Chinese female sample. AU7 is also found to be better correlated with 'anger' for politicians from other countries, but the correlation was not found for Chinese politicians.

Table 44 Correlation between Action Units and ***'Anger'*** *(Emotient FACET)*

	Chinese		Others	
	AU	Rho	AU	Rho
All Data	**AU9**	**0.67**	AU4	0.61
	AU18	0.46	**AU9**	**0.57**
	AU23	0.42	AU23	0.41
	AU24	0.4	AU18	0.4
Female	AU4	0.61	**AU9**	**0.52**
	AU5	0.48	AU18	0.46
	AU9	**0.48**		
	AU18	0.41		
	AU17	0.4		
Male	**AU9**	**0.72**	AU4	0.67
	AU23	0.54	AU23	0.54
	AU24	0.49	**AU9**	**0.5**
	AU18	0.47	AU24	0.41
	AU43	0.45	AU7	0.4
Age < 60	**AU9**	**0.58**	AU4	0.63
			AU9	**0.57**
			AU7	0.41
			AU23	0.4
Age >= 60	**AU9**	**0.81**	**AU9**	**0.59**
	AU24	0.72	AU4	0.55
	AU23	0.67	AU18	0.54

	AU14	0.63	AU24	0.44
	AU18	0.6	AU23	0.42
	AU28	0.59		
	AU12	0.52		
	AU17	0.44		
	AU43	0.4		

From the results of Emotient FACET (See Table 44), it can be found that AU9 has a better correlation with 'anger' across samples of different ethnicities, genders and ages. For politicians from other countries, AU4 was found to correlate with 'anger' in the majority of the sample, except for the female sample.

As can be seen from the above results, there are differences in the physical correlates of emotion obtained by Affectiva AFFDEX and Emotient FACET, and also differences in the correlations found in samples of different ethnicities, genders, and ages. Overall, AU12 is found to have a strong correlation with 'joy' in the majority of samples and both systems, while AU14 is also found to be well correlated with 'joy' in different samples in Emotient FACET. AU4 is found to be well correlated with 'anger' in the vast majority of samples in Affectiva AFFDEX, and AU9 is found to be well correlated with 'anger' in different samples in Emotient FACET.

4.5.2 Acoustic Features

F0 has been shown to vary across language and gender in previous studies, and this is verified on the politician dataset in this study.

Table 45 F0 Mean in Different Samples

	Chinese	Others
All Data	134.48	115.88
Female	157.85	135.58
Male	127.39	102.60

As can be seen from Table 45, F0 is higher in Mandarin than in other languages (mainly English), and it is found to be higher in females than in males in different languages, which is consistent with the findings of previous studies.

The correlation of F0, the mean of MFCCs, the standard deviation of MFCCs, Jitter and Shimmer with emotions is examined in samples of different ethnicities, languages, genders, and ages using the Spearman correlation coefficient. In the following tables, the acoustic features with absolute values of Spearman's correlation coefficient greater than 0.3 and their corresponding correlation coefficients are listed.

*Table 46 Correlation between Acoustic and **'Joy'***

	Chinese		Others	
	Feature	Rho	Feature	Rho
All Data	F0 MFCCs Mean MFCCs Std	0.6192 -0.5164 0.3442	F0 MFCCs Mean	0.3367 -0.3357
Female	F0 MFCCs Mean MFCCs Std MFCCs Std	0.6013 -0.7001 0.3366 -0.3376	F0 MFCCs Mean	0.4011 -0.3524
Male	F0 MFCCs Mean MFCCs Std	0.5668 -0.3045 0.3049		
Age < 60	F0 MFCCs Mean MFCCs Std	0.665 -0.5704 0.3234	F0	0.3362
Age >= 60	F0 MFCCs Std	0.5067 0.4728	F0 MFCCs Mean	0.3431 -0.3873

As can be seen from Table 46, F0 is found to be correlated with 'joy' in samples with different languages, genders and ages, and the correlation is stronger in the Chinese sample. MFCCs mean is also found to be correlated with 'joy' in most of the samples, with the strongest correlation in the Chinese females (Rho = -0.7001). The standard deviation of MFCCs was also found to be weakly correlated with 'joy' in the Chinese sample. In these samples, Jitter and Shimmer are found to have a very weak correlation with 'joy'.

*Table 47 Correlation between Acoustic and **'Anger'***

	Chinese		Others	
	Feature	Rho	Feature	Rho
All Data	F0 MFCCs Mean MFCCs Std	0.5531 -0.5066 0.3201	MFCCs Mean	-0.3226
Female	F0 MFCCs Mean Jitter	0.4292 -0.729 -0.3263	F0 MFCCs Mean	0.3213 -0.3556
Male	F0	0.5118		
Age < 60	F0 MFCCs Mean	0.5902 -0.5698		
Age >= 60	F0 MFCCs Std	0.4487 0.4487	MFCCs Mean	-0.3682

From Table 47, it can be found that F0 has a moderate correlation with 'anger' in Chinese samples, while it has a weaker correlation with 'anger' in samples from other countries. MFCCs mean is found to have a strong correlation with ‘anger’ in Chinese female samples (RHO=-0.729). The standard deviation of MFCCs is found to have a moderate correlation with ‘anger’ in a sample of Chinese people who are older than 60 years old (RHO=0.4487). Jitter is found to have a weak correlation with ‘anger’ in the Chinese female sample (RHO=-0.3263).

From the above results, it can be found that the physical correlates of emotions differed in samples with different ethnicities, languages, genders and ages. The correlations of Jitter and Shimmer with ‘joy’ and ‘anger’ are found to be very weak in the majority of the samples. In most of the samples, the correlations of acoustic features with ‘joy’ are slightly stronger than those with ‘anger’. The correlation between these acoustic features and emotions is slightly stronger in the Chinese sample compared to the samples from other countries.

Chapter 5 Conclusion and Future Work

5.1 Key Findings

This study reviews previous research on the correlation between facial expression and acoustic features and emotions, and reviews relevant research on the 4 emotion recognition systems Affectiva AFFDEX, Emotient FACET, openSmile and openVokaturi. In these studies, differences in facial expression and acoustic features were found in different ethnicities, languages, genders and ages. While the FER systems, Affectiva AFFDEX and Emotient FACET, were trained using datasets containing only a very small amount of Asian data, the SER systems, openSmile and openVokaturi, were trained only on audio in English and German. In addition, previous studies on the performance of these emotion recognition systems have lacked research on Asians and Mandarin. In this study, the performance of 4 emotion recognition systems was compared using speech videos and audios of Chinese politicians, and videos and audios of politicians from other countries were used as controls to find out if there was any difference in the performance of these systems in facial expression and SER of people in different ethnicities, languages, genders and ages. In addition to this, an attempt was made to analyse the emotions in Chinese speech texts and to compare the results with those of facial and SER.

In this study, 45 speech videos of Chinese politicians, and 212 speech videos of politicians from other countries were collected and the audio was extracted from them. These videos and audios were eventually processed using two FER systems and two SER systems, and the results of these systems were analysed for differences and consistency in the detection of the two emotions ‘joy’ and ‘anger’. In addition, the two video datasets with emotion labels, RAVDESS and CASME II, and one audio dataset with emotion labels, EMO-DB, were used for analysis and as a baseline. The text of the Chinese speech was then transcribed using Amazon Transcribe, and sentiment analysis was performed on the text with HowNet and NTUSD, finally comparing the results with the results of the emotion recognition systems.

Ultimately, Affectiva AFFDEX and Emotient FACET were found to be less consistent in recognising emotions in Chinese politicians than in politicians from other countries. Meanwhile, the two systems were also less consistent in the recognition of spontaneous expressions than in the recognition of posed expressions. It was also found in openSmile and openVokaturi that they were less consistent in identifying the emotions of Chinese politicians than in identifying the emotions of other politicians. Moreover, the consistency in the recognition of emotions in the speech audio of these politicians was worse than that in the EMO-DB. These phenomena may be due to the following reasons. Firstly, spontaneous emotional expressions are less likely to be recognised compared to posed expressions. Secondly, different emotion recognition systems have different biases in recognising emotions. Third, the lack of training of the systems on Asian facial expressions or Mandarin may lead to a decrease in their accuracy in recognising emotions on these data, thus increasing the inconsistency between systems.

A comparison of the results of the FER system and the SER system revealed that the two systems had very weak consistency, or even inconsistency, in the recognition results across samples of different ethnicities, languages, genders and ages. In the comparison of Chinese text sentiment and the results of the emotion recognition system, poor consistency or even inconsistency was also found. Possible reasons for this phenomenon are, firstly, different emotion recognition systems have different accuracy in emotion recognition, which may result in an inconsistency between systems. Secondly, there may be poor consistency due to 'emotion leakage'. Politicians may control their facial expressions or voices when speaking in public in order to portray a certain image or convey a certain idea, which may cause the emotion detected by the system from one aspect not to be the real emotion, thus causing inconsistency between systems.

Finally, this study verified the physical correlates of emotions. For action units, AU12 was found to be correlated with ‘joy’ across samples of different races, genders, and ages, and the correlation was strong, which is consistent with previous findings. AU14 was found to be strongly correlated with ‘joy’ in results of Emotient FACET, which has not been mentioned in previous studies. In addition to this, AU6 was found to be correlated with ‘joy’ in previous studies and that correlation was also found in some samples in this study. For ‘anger’, most previous studies have suggested that AU4, AU7, AU23 and AU24 are correlated with it. In this study, AU4, AU23 and AU24 were found to be correlated in the sample of Chinese politicians and politicians from other countries, while AU7 was only found to be better correlated with ‘anger’ in politicians from other countries. For acoustic features, it was found that the mean F0 in Mandarin was higher than the mean F0 in other languages, and the mean F0 in females was found to be higher than the mean F0 in males in different languages, which validates the previous study. In addition, F0 was found to correlate with ‘joy’ and ‘anger’ in all Chinese politician samples, while for politicians from other countries, it was found to have better correlations with ‘joy’/‘anger’ in some gender and age groups only. The mean and standard deviation of MFCCs were also found to correlate with ‘joy’/‘anger’ in some samples.

5.2 Future Work

The text analysis method used in this study is a simple textual sentiment analysis based on sentiment lexicons. Due to the limitation of the quantity and quality of Chinese sentiment lexicons, the sentiment lexicons used for this study are HowNet and NTUSD, which are two Chinese sentiment lexicons developed at an earlier time, so some new words have not been added to the lexicons. Furthermore, NTUSD is a lexicon developed by Taiwan University, and some words and expressions are different from those in mainland China, the results of textual sentiment analysis may not be ideal. One direction for future work is to build a more complete sentiment lexicon or to use machine learning methods to achieve a more accurate sentiment analysis of texts.

In this study, differences are found among the emotion recognition results for facial expression, speech and text, which may be due to the accuracy of the emotion recognition of the systems and ‘emotion leakage’. In order to improve the emotion detection accuracy, the emotion recognition results from several aspects can be combined to build a multimodal emotion recognition system. Therefore, another future work is to find a way to combine facial emotion recognition results, speech emotion recognition results and textual sentiment analysis results to build a more comprehensive and accurate emotion recognition system.

An implementation of this is through fuzzy logic [74]. Taking ‘joy’ as an example, it is difficult to set a clear boundary to distinguish the two concepts of ‘joy’ and 'not joy', while in fuzzy logic the two concepts can be fuzzed by the Degree of Membership. The process of calculation involves, firstly, defining the Degree of Membership function. According to this function, the probability of ‘joy’ input by the two systems can be transformed into the Degree of Membership of ‘joy’, ‘emotion in the middle’ and ‘not joy’. Next, a 3 x 3 table is created in which the Degree of Membership obtained by the two systems is converted into 9 outputs, which are called Fire Strength, by using either the maximum or minimum rule. Then a calculation is defined that allows these 9 Fire Strength values to be converted into one final output value, one common calculation is the weighted average method. Eventually, a threshold can be set to determine whether the emotion is ‘joy’ based on the output value. When using this method, emotions can be subdivided depending on the situation, for example, ‘joy’ and ‘not joy’ can be further divided into ‘joy’, ‘little bit joy’, ‘emotion in the middle’, ‘little bit not joy’, ‘not joy’. Additional inputs are also possible, for example using three emotion recognition results as input.

In order to test the effectiveness of this data fusion approach, the emotions of the samples need to be manually labelled first. In addition, tests can also be carried out on datasets that already have emotion labels.

Bibliography

[1] Abirami, A. M., & Gayathri, V. (2017, January). A survey on sentiment analysis methods and approach. In *2016 Eighth International Conference on Advanced Computing (ICoAC)* (pp. 72-76). IEEE.

[2] Ahmad, K., Wang, S., Vogel, C., Jain, P., O'Neill, O., & Sufi, B. H. (2021, November). Comparing the Performance of Facial Emotion Recognition Systems on Real-Life Videos: Gender, Ethnicity and Age. In *Proceedings of the Future Technologies Conference* (pp. 193-210). Springer, Cham.

[3] Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, *116*, 56-76.

[4] Alonso-Martin, F., Malfaz, M., Sequeira, J., Gorostiza, J. F., & Salichs, M. A. (2013). A multimodal emotion detection system during human–robot interaction. Sensors, 13(11), 15549-15581.

[5] Amazon Transcribe Document. [Online]. Available: What is Amazon Transcribe? - Amazon Transcribe. (Accessed on 10 August 2022).

[6] Anju, A., Barath, M., Maheshwaran, R., Vignesh, K., & SR, K. K. (2022, April). Sentimental Analysis for E-Commerce Website. In *2022 10th International Conference on Emerging Trends in Engineering and Technology-Signal and Information Processing (ICETET-SIP-22)* (pp. 1-4). IEEE.

[7] Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

[8] Brockmann, M., Drinnan, M. J., Storck, C., & Carding, P. N. (2011). Reliable jitter and shimmer measurements in voice clinics: the relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task. *Journal of voice*, *25*(1), 44-53.

[9] Bui, H. D., & Chong, N. Y. (2018, September). An integrated approach to human-robot-smart environment interaction interface for ambient assisted living. In *2018 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)* (pp. 32-37). IEEE.

[10] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005, September). A database of German emotional speech. In *Interspeech* (Vol. 5, pp. 1517-1520).

[11] Clemente, C. D. (2011). Anatomy: a regional atlas of the human body. *Lippincott Williams & Wilkins*.

[12] Cordaro, D. T., Sun, R., Keltner, D., Kamble, S., Huddar, N., & McNeil, G. (2018). Universals and cultural variations in 22 emotional expressions across five cultures. *Emotion*, *18*(1), 75.

[13] Ding, H., Hoffmann, R., & Hirst, D. (2016). Prosodic transfer: A comparison study of f0 patterns in l2 english by chinese speakers. In *Speech Prosody* (Vol. 2016, pp. 756-760).

[14] Donato, G., Bartlett, M. S., Hager, J. C., Ekman, P., & Sejnowski, T. J. (1999). Classifying facial actions. *IEEE Transactions on pattern analysis and machine intelligence*, *21*(10), 974-989.

[15] Dong, Z., & Dong, Q. (2003, October). HowNet-a hybrid language and knowledge resource. In *International conference on natural language processing and knowledge engineering, 2003. Proceedings. 2003* (pp. 820-824). IEEE.

[16] Drioli, C., Tisato, G., Cosi, P., & Tesser, F. (2003). Emotions and voice quality: experiments with sinusoidal modeling. In *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*.

[17] Du, S., Tao, Y., & Martinez, A. M. (2014). Compound facial expressions of emotion. *Proceedings of the national academy of sciences*, *111*(15), E1454-E1462.

[18] Eichhorn, J. T., Kent, R. D., Austin, D., & Vorperian, H. K. (2018). Effects of aging on vocal fundamental frequency and vowel formants in men and women. *Journal of Voice*, *32*(5), 644-e1.

[19] Ekman, P., & Friesen, W. V. (1969). Nonverbal leakage and clues to deception. Psychiatry, 32(1), 88-106.

[20] Ekman, P., & Friesen, W. V. (1978). Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

[21] Ekman, P., Friesen, W. V., & Hager, J. C. (2002). Facial action coding system [E-book]. *Salt Lake City, UT: Research Nexus*.

[22] Eyben, F., & Schuller, B. (2015). openSMILE:) The Munich open-source large-scale multimedia feature extractor. ACM SIGMultimedia Records, 6(4), 4-13.

[23] Eyben, F., Wöllmer, M., & Schuller, B. (2009, September). OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit.

In *2009 3rd international conference on affective computing and intelligent interaction and workshops* (pp. 1-6). IEEE.

[24] Eyben, F., Wöllmer, M., & Schuller, B. (2010, October). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia* (pp. 1459-1462).

[25] Farnsworth, B. Facial Action Coding System (FACS) – A Visual Guidebook [Online]. 18 August 2019. Available: https://imotions.com/blog/facial-action-coding-system. (Accessed on 17 July 2022).

[26] Grondhuis, S. N., Jimmy, A., Teague, C., & Brunet, N. M. (2021). Having difficulties reading the facial expression of older individuals? blame it on the facial muscles, not the wrinkles. *Frontiers in Psychology*, *12*, 620768.

[27] Gunawan, T. S., Husain, R., & Kartiwi, M. (2017, November). Development of language identification system using MFCC and vector quantization. In *2017 IEEE 4th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA)* (pp. 1-4). IEEE.

[28] Hjortsjö, C. H. (1969). Man's face and mimic language. *Studentlitteratur*.

[29] Jabbar, J., Urooj, I., JunSheng, W., & Azeem, N. (2019, May). Real-time sentiment analysis on E-commerce application. In *2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC)* (pp. 391-396). IEEE.

[30] Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, *109*(19), 7241-7244.

[31] Jackson, P., & Haq, S. (2014). Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*.

[32] Kanade, T., Cohn, J. F., & Tian, Y. (2000, March). Comprehensive database for facial expression analysis. In *Proceedings fourth IEEE international conference on automatic face and gesture recognition (cat. No. PR00580)* (pp. 46-53). IEEE.

[33] Karpagavalli, S., & Chandra, E. (2016). A review on automatic speech recognition architecture and approaches. International Journal of Signal Processing, Image Processing and Pattern Recognition, 9(4), 393-404.

[34] Keating, P., & Kuo, G. (2012). Comparison of speaking fundamental frequency in English and Mandarin. *The Journal of the Acoustical Society of America*, *132*(2), 1050-1060.

[35] Keltner, D., Sauter, D., Tracy, J., & Cowen, A. (2019). Emotional expression: Advances in basic emotion theory. *Journal of nonverbal behavior*, *43*(2), 133-160.

[36] Koolagudi, S. G., Rastogi, D., & Rao, K. S. (2012). Identification of language using mel-frequency cepstral coefficients (MFCC). *Procedia Engineering*, *38*, 3391-3398.

[37] Krosschell, K. Facial Expression Analysis: The Complete Pocket Guide [Online]. 10 March 2020. Available: https://imotions.com/blog/facial-expression-analysis. (Accessed on 10 August 2022)

[38] Krumhuber, E. G., Küster, D., Namba, S., & Skora, L. (2021). Human and machine validation of 14 databases of dynamic facial expressions. *Behavior research methods*, *53*(2), 686-701.

[39] Ku, L. W., & Chen, H. H. (2007). Mining opinions from the Web: Beyond relevance retrieval. *Journal of the American Society for Information Science and Technology*, *58*(12), 1838-1850.

[40] Kumbhar, H. S., & Bhandari, S. U. (2019, September). Speech emotion recognition using MFCC features and LSTM network. In *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)* (pp. 1-3). IEEE.

[41] Ley, M., Egger, M., & Hanke, S. (2019, November). Evaluating Methods for Emotion Recognition based on Facial and Vocal Features. In *AmI (Workshops/Posters)* (pp. 84-93).

[42] Likitha, M. S., Gupta, S. R. R., Hasitha, K., & Raju, A. U. (2017, March). Speech based human emotion recognition using MFCC. In *2017 international conference on wireless communications, signal processing and networking (WiSPNET)* (pp. 2257-2260). IEEE.

[43] Lin, P., & Luo, X. (2020). A survey of the applications of sentiment analysis. *International Journal of Computer and Information Engineering*, *14*(10), 334-346.

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

[45] Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English.

PloS one, 13(5), e0196391.

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

[47] Luzietti, R. B., Pretto, N., Kaplan, F., Dufaux, A., & Canazza, S. (2021). FONTI 4.0: Evaluating Speech-to-Text Automatic Transcription of Digitized Historical Oral Sources. In CLiC-it.

[48] Lytridis, C., Vrochidou, E., & Kaburlasos, V. (2018). Emotional speech recognition toward modulating the behavior of a social robot. In *The Proceedings of JSME annual Conference on Robotics and Mechatronics (Robomec) 2018* (pp. 1A1-B14). The Japan Society of Mechanical Engineers.

[49] Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006, April). The eNTERFACE'05 audio-visual emotion database. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)* (pp. 8-8). IEEE.

[50] Matsumoto, D., Keltner, D., Shiota, M. N., O'Sullivan, M., & Frank, M. (2008). Facial expressions of emotion.

[51] McDuff, D., Girard, J. M., & Kaliouby, R. E. (2017). Large-scale observational evidence of cross-cultural differences in facial behavior. *Journal of Nonverbal Behavior*, *41*(1), 1-19.

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

[53] Mennen, I., Schaeffler, F., & Docherty, G. (2012). Cross-language differences in fundamental frequency range: A comparison of English and German. *The Journal of the Acoustical Society of America*, *131*(3), 2249-2260.

[54] Mrozek, D., Kwaśnicki, S., Sunderam, V., Małysiak-Mrozek, B., Tokarz, K., & Kozielski, S. (2021, June). Comparison of Speech Recognition and Natural Language Understanding Frameworks for Detection of Dangers with Smart Wearables. In International Conference on Computational Science (pp. 471-484). Springer, Cham.

[55] Nunes, A., Coimbra, R. L., & Teixeira, A. (2010, April). Voice quality of european portuguese emotional speech. In *International Conference on Computational Processing of the Portuguese Language* (pp. 142-151).

Springer, Berlin, Heidelberg.

[56] OpenVokaturi. [Online]. Available: https://vokaturi.com/algorithms. (Accessed on 2 August 2022).

[57] Paeschke, A., Kienast, M., & Sendlmeier, W. F. (1999, August). F0-contours in emotional speech. In *Proc. 14th Int. Congress of Phonetic Sciences* (Vol. 2, pp. 929-932).

[58] Peng, H., Cambria, E., & Hussain, A. (2017). A review of sentiment analysis research in Chinese language. *Cognitive Computation*, *9*(4), 423-435.

[59] Priyasad, D., Fernando, T., Denman, S., Sridharan, S., & Fookes, C. (2020, May). Attention driven fusion for multi-modal emotion recognition. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3227-3231). IEEE.

[60] Probst, L., & Braun, A. (2019). The effects of emotional state on fundamental frequency. In *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia* (pp. 67-71).

[61] Ren, Q., Zheng, Y., Guo, G., & Hu, Y. (2019, February). Resource recommendation algorithm based on text semantics and sentiment analysis. In *2019 Third IEEE International Conference on Robotic Computing (IRC)* (pp. 363-368). IEEE.

[62] Sapra, A., Panwar, N., & Panwar, S. (2013). Emotion recognition from speech. *International journal of emerging technology and advanced engineering*, *3*(2), 341-345.

[63] Saranya, G., Geetha, G., Meenakshi, K., & Karpagaselvi, S. (2020, December). Sentiment analysis of healthcare Tweets using SVM Classifier. In *2020 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS)* (pp. 1-3). IEEE.

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

[65] Stöckli, S., Schulte-Mecklenbeck, M., Borer, S., & Samson, A. C. (2018). Facial expression analysis with AFFDEX and FACET: A validation study. *Behavior research methods*, *50*(4), 1446-1460.

[66] Teixeira, J. P., Oliveira, C., & Lopes, C. (2013). Vocal acoustic analysis–jitter, shimmer and hnr parameters. *Procedia Technology*, *9*, 1112-1122.

[67] Velusamy, S., Kannan, H., Anand, B., Sharma, A., & Navathe, B. (2011, May). A method to infer emotions from facial action units. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2028-2031). IEEE.

[68] Wagner, A., & Braun, A. (2003). Is voice quality language-dependent? Acoustic analyses based on speakers of three different languages. *Language*, *6*(4), 2.

[69] Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 1-50.

[70] Xu, B., Tao, C., Feng, Z., Raqui, Y., & Ranwez, S. (2021). A benchmarking on cloud based speech-to-text services for french speech and background noise effect. arXiv preprint arXiv:2105.03409.

[71] Yan, W. J., Li, X., Wang, S. J., Zhao, G., Liu, Y. J., Chen, Y. H., & Fu, X. (2014). CASME II: An improved spontaneous micro-expression database and the baseline evaluation. PloS one, 9(1), e86041.

[72] Yang, K., Wang, C., Sarsenbayeva, Z., Tag, B., Dingler, T., Wadley, G., & Goncalves, J. (2021). Benchmarking commercial emotion detection systems using realistic distortions of facial image datasets. *The visual computer*, *37*(6), 1447-1466.

[73] Yücesoy, E., & Nabiyev, V. V. (2013, November). Gender identification of a speaker using MFCC and GMM. In *2013 8th International Conference on Electrical and Electronics Engineering (ELECO)* (pp. 626-629). IEEE.

[74] Zadeh, L. A. (1988). Fuzzy logic. Computer, 21(4), 83-93.

[75] Zhu, S., Chong, S., Chen, Y., Wang, T., & Ng, M. L. (2022). Effect of Language on Voice Quality-An Acoustic Study of Bilingual Speakers of Mandarin Chinese and English. *Folia Phoniatrica et Logopaedica*.