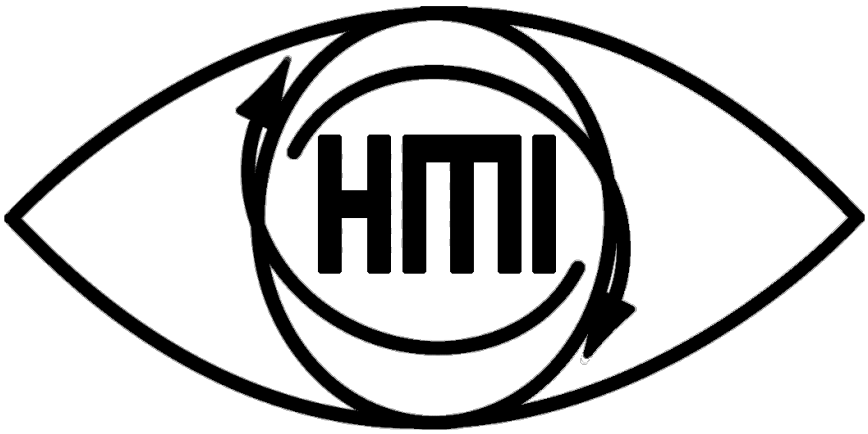

How is emotion change reflected in manual and automatic annotations of different modalities?

Master's thesis

Ye Xiang

HMI, University of Twente

March 2017

Graduation committee:

Dr.ir.D.Reidsma
Prof.dr.D.K.J. Heylen

ABSTRACT

The SEMAINE database consists of recordings of persons talking to different virtual characters. Each of the virtual characters speaks and behaves according to a repertoire of utterances. The repertoire designed the verbal and nonverbal behaviors of virtual character to influence the emotional state of persons via the conversational interaction. There are four characters - Obadiah, Poppy, Spike, and Prudence. The Obadiah character is gloomy and tries to make other person feeling depressed, whereas Poppy character is cheerful and is defined to make other person talking to it happily as well. The Spike character is angry and is intended to provoke other persons. The Prudence character is sensible. It tries to make other people sensible, too. The recordings have been annotated manually in several dimensions which indicating the emotional state of the person as it changes over the interaction.

The first goal of the thesis is to find out whether the emotions of the participants align with the character they interact with. Since there are 4 different characters, it is possible to explore how these characters differ in changing the emotions of participants.

In the manual annotations, the emotional states were indeed changed among characters. The changes in the dimensions of the Happiness, the Sadness, the Anger, and full rating dimensions (except the Intensity) were statistically different. The Poppy and Spike characters had unique impacts in the Happiness dimension. Compared to Poppy and Prudence characters, the Obadiah character had different impacts in the Sadness dimension. The Anger character had more advantages than Obadiah and Poppy characters in influencing the anger emotion of participants.

The second goal of the thesis is to investigate whether automatic emotion recognition tools assign the same emotions to the persons as the manual annotations. The selected tools included the FaceReader - a software program that recognizes the emotion of a person based on the detected facial expressions, and the LIWC (Linguistic Inquiry and Word Count) - a tool which is based on the words used by a person.

The emotions recognized by FaceReader match the annotated emotions more closely than the emotions recognized by LIWC. In the results of FaceReader, the dimensions of the happy, the sad, and the angry emotions were correctly correlated with the manual annotations in most characters. The FaceReader achieved almost 42% accuracy. In the results of LIWC, only the dimension of the Posemo (happy) well correlated with the Semaine annotation. The LIWC only had 33% accuracy.

The whole thesis is divided into four steps as follows. In the first step, we preprocess the samples to ensure the data is well organized. The second step is the Phase I which selected the facial expression and the pair of virtual characters (Obadiah and Poppy) with the strongest contrast of emotional impacts. The third step is the Phase II which increased the types of virtual characters (Spike and Prudence). The last step is the Phase III which added the modality of text into the analysis.

DEDICATION

I dedicate my master thesis to my family, my teachers, and my friends. A special feeling of gratitude to my parents who support me soul and heart. My girlfriend LinYa Teng whose words of encouragement ring in my ears.

I also dedicate this work to my teachers. I express my sincere gratitude to professor Dennis Reidsma, my first supervisor, who provided me the patient guidance as well as shaped my sprit of scientific rigor. I appreciate the encouragement and the professional assistant from professor Dirk Heylen.

At last, I want to thank professors Khiet Truong and Peter de Vries for the aids of experiment resources. Mrs.Sharon and Mrs.Hemo, who are my teachers as well as friends, give me many precious suggestions, encouragements, and help. Both of you drag me out of the trouble during the most desperate and darkest time.

LIST OF ACRONYM

LIWC	- Linguistic Inquiry and Word 2015
AVC	- Arousal-Valence-Control
NLP	- natural language processing
FACS	- facial action coding system
AU	- action units
PCA	- principal component analysis
MMI	- M&M initiative
USC	- University of South California
CERT	- computer expression recognition toolbox
GVC	- good vibrations company
MFCC	- mel frequency cepstral coefficient
LPC	- linear predictive coding
NAQ	- normalized amplitude quotient
QOQ	- quasi-open quotient
PSP	- parabolic spectral parameter
MDQ	- maxima dispersion quotient
SAL	- sensitive artificial listener
QA	- quantitive agreement
GCI	- Glottal Closure Instants
TE-MFCC	- Alternative MFCCs extracted from True-Envelope spectral representation
NaN	- not a number
WC	- word count
WPS	- words per sentence
CDI	- categorical-dynamic index

CONTENTS

1. Introduction

In this thesis, there are two research questions. First one is the investigation of the emotional state change of a person with a fixed emotion impact. Second one is the validation of the automatic emotion recognition softwares.

For the first research question, the form of daily conversation is the most basic method of human emotional interaction. The daily conversation contains many channels of communication such as linguistics, facial expression and speech. These channels convey the emotion stimuli/impact which could influence the psychological state of people. In order to keep the accuracy of emotion recognition, the manual annotation of the professional annotators are used in the thesis.

For the second research question, the first problem of automatic emotion recognition tools is the selection of modalities. Because the automatic tools trace the participants' emotion by using single/multiple modalities. The selected modalities of tools relate to the extraction and the comparison of emotions.

According to our current resources (section 2.5), the text, the speech, and the facial expression are selected as the input modalities of automatic tools. Since the tools of multiple modalities are more complex (section 1.1.3), the tools of single modality are chosen (section 3.2). In order to clearly describe the mental state of persons, dimensional approach is used to parameterize the person's emotion in each selected modality of automatic tools (section 2.1).

1.1. Problems

Compare to the manual annotations of emotions, automatic emotion recognition is still in the early phase. There are still some problems unsolved.

1.1.1. Deceptiveness

Compared to technical algorithms and classification methods, Wagner (2005) argued that the input data has a greater impact on the performance quality of current multimodal affect recognizers.

The subjective impression of persons is not the perfect indicator of their emotions. Because there are many factors (culture, social context, and etc) which could influence the expression of people. For example, participants, who have different cultural background, are possible to express or intercept the same emotional information differently. Additionally, participants can manipulate their emotional expression, especially when they want to please the annotators.

With the impacts of social context, cultural event and persuasive communication, the emotional expression of people can be complex and illusive, even cheated. After processing some disguised expressions of people, the final results of automatic tools could be biased, even misleading. It is necessary to well check the input data.

1.1.2. Measurement of emotion

Most automatic tools use the dimensional approach and the emotional dimensions (section 2.1.2) to indicate the emotion of persons. Due to the lack of a widely recognized annotation theory, many tools measure the emotion of participants in different dimensions. Because of the differences among dimensions, it is difficult to conduct the data comparison or transformation across automatic tools.

1.1.3. Modalities Selection

A person's emotion can be expressed by multiple modalities including face, voice, body movement and bio-signals. These channels could send primary, supplementary or even misleading information of emotion. Many modalities have been applied to extract the emotion. The text (Mairesse, Walker, Mehl, & Moore, 2007), the speech (Oudeyer, 2003; Schuller, Rigoll, & Lang, 2003), and the facial expression (P. Ekman & W. Friesen, 1978) have been widely used for emotion extraction and recognition. Besides that, bio-physiological signals (Cannon, 1927; Schachter, 1964) and bodily movement (S. Scherer et al., 2012) can also provide valuable information. These implementations have brought many benefits on selling and branding (iMotions, 2016), product promotion (Affectiva, 2016), improvement of the interaction experience (Picard, 2003; Sebe et al., 2007), and etc.

In terms of modality selection, the thesis mainly focuses on facial expression, text and speech. On the one hand, these modalities contributes a large proportion of semantic information of communication (Mehrabian, 2008). On the other hand, these modalities are widely applied in the industry. Consequently, the applications of these modalities created a massive scale of samples and annotations.

1.2. Related Work

1.2.1. Autonomous mechanism of machines

The autonomous mechanism of machines is the implementation of emotional capability (André, Klesen, Gebhard, Allen, & Rist, 2000). Researchers had made many contributions by utilizing machine learning algorithms (Blumberg et al., 2002; Pang, Lee, & Vaithyanathan, 2002) and other robot learning algorithms (Mavridis, 2015; C. Strapparava & Mihalcea, 2008; Trilla & Alías, 2009). By building the interaction configuration (digital mental state, optional action, selective utterances, and etc) of machine from the training samples, the machine can analyze current information and make the smart decision after the simulation of the human decision-making process.

Although it is far to reach the human level, the autonomous mechanism of machines creates the preliminary machine perception for interactive virtual personality. The interaction configuration and the perception of machines provide the fundamental guidance to transfer the raw emotional data into a computational psychological state. In the thesis, the emotion comparison also need the transferring and the calculation of emotional state.

1.2.2. The framework of emotion recognition

Emotion recognition is also a complex process which includes emotion detection, emotion extraction, and emotional analysis. The framework of the emotion recognition (Gunes, 2010) would fully use the context-specific interpretation of affective displays to improve the emotion analysis of the participants.

During the working process, the framework of emotion recognition integrally consider different issues, such as context interpretation, emotion representation, persons' information mapping between emotions and behaviors, and the annotation of participants' emotion. These work are all needed in the thesis.

1.3. Goals

This thesis is planed to investigate whether the participants' emotion align with their received emotional stimulus. It is also intended to find out whether the automatic emotion recognition softwares assign the same emotions to the participants as the annotators' judgement.

1.3.1. Research requirements

According to above goals, the key requirements of this research are displayed as follows:

I. To find a proper emotion representation which supports the computable state of participants' emotions, regardless the modality and the type of the participants' emotion.

II. To create multiple fixed emotional stimuli which aim to lastly influence the emotion of participants during the conversational interaction.

III. To find a trustworthy manual annotation which provides the authoritative analysis about the emotion of participants during the conversational interaction.

IV. To find a reliable source of participants' multimodal emotional expressions which ensures the usages of the automatic emotion recognition tools in each selected modality.

1.3.2. Hypotheses

Based on the above discussion, we hypothesize as followed:

I. In the manual judgements of annotators and the results of automatic tools, the emotion of participants would align with the fixed emotional feedbacks.

II. In the results of automatic tools, the assigned emotions of participants should be consistent with the assigned emotions of participants in the manual annotations.

III. In the results of automatic tools, the tools of different modalities should differ in assigning the emotions to participants.

2. Theory And Concept

2.1. Emotion Representation

The representation of emotion is the first issue to be solved for the manual annotation as well as automatic recognition. In order to analyze the emotion change of participants during the conversational interaction, it is better for the emotion to be represented and annotated continuously. Fortunately, there is a manual annotation (G. McKeown, Valstar, Cowie, Pantic, & Schröder, 2012) which had annotated the emotion of participants in a continuous way.

In terms of the automatic annotation, the most widely accepted theories of emotion representation can be classified into three approaches: categorical, dimensional, and appraisal-based approaches (Grandjean, Sander, & Scherer, 2008). Dimensional approach is used in the thesis. Because it provides a representation of the computable psychological state with large scale of implementations.

2.1.1. Categorical approach

The categorical approach assumes that there are some basic emotions (including happiness, sadness, surprise, fear, anger, and disgust) which are widely recognized (Ekman, Friesen, & Ellsworth, 2013). These emotions are hard-wired in the brain (Darwin, Ekman, & Prodger, 1998; Ekman, 1992; Tomkins, 1962). This approach use the basic categories/labels of emotions to classify the emotions of participants. It is one of the most popular method to evaluate the connection between the emotional nonverbal expression and the basic emotions. For example, Schwartz and his colleagues (2013) had applied the categorical approach in the research of personality impacts on the vocabulary usage of Facebook. They compared the words of participants with the basic categories of emotional linguistics via the software - LIWC (J. W. Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007).

The categorical approach cannot provide the access to blend or mix multiple labels/ categories. Therefore, many psychological researchers advocate the dimensional approach.

2.1.2. Dimensional approach

In the field of automatic emotion classification, the overall trend has slowly shifted from categorical classification to dimensional classification (Gunes, 2010). The dimensional classification is based on the dimensions which can indicate the emotional state of a person. The dimensional approach and its supporters believe that the affective states are dependent on each other and related to one another systematically (Grandjean et al., 2008; Mehrabian, 1996; Russell, 1980; K. R. Scherer, Schorr, & Johnstone, 2001). The dimensional approach can mix multiple emotional representations. The change of emotional state can also trigger a serial reaction among the emotion representations. The dimensional approach can easily parameterize the

emotion while allowing the observer to distinguish the emotional expression continuously.

The dimensions of dimensional approach vary with time. Firstly, the Circumplex model of affect, assumes that each basic emotion is part of the emotional continuum, represents emotions in a circle with two bipolar dimensions (active and pleasant; Nowlis & Nowlis, 1956; Russell, 1980; Schlosberg, 1952, 1954). Another example is the Five Factor model (Digman, 1990; McCrae & John, 1992) which is widely applied in psychological research. It uses five basic dimensions of personality traits to indicate the emotional state. Then, Dietz and Lang (1999) had invented the Arousal-Valence-Control (AVC) dimensional model. The dimensional classification can simplify the complexity of emotion representations. The most commonly deployed method is to simplify the basic emotion representations into a 3 classes valence-related description: positive, neutral and negative emotion (e.g., Yu, Aoki, & Woodruff, 2004). Similarly, it is also possible to transfer the emotion classification problem into a 4-quadrant measurement by two dimensions of positive-negative and active-passive (Fragopanagos & Taylor, 2005). According to Revelle and Scherer (2009), the emotion is a psychological state which contains feeling, action, appraisal and wants over a certain time period. In terms of the emotion decay, the intensity (Gary McKeown, Valstar, Cowie, & Pantic, 2010) could be used to describe the emotion existence.

2.1.3. Appraisal-based approach

Appraisal-based approach detects the emotion during the continuous, recursive and subjective evaluation of both people's internal state and the world external state (Grandjean et al., 2008). It views the emotion via both long-term and instant changes of all the relevant components including cognition, motivation, physiological reaction and etc. By linking the contextual information into an automatic emotion recognizer, it may improve the interpretation capability as well as enrich the choice of expressive behaviors (Mortillaro, Meuleman, & Scherer, 2012). The appraisals are used as an intermediate layer between expressive features (input) and emotion labeling (output).

Component process model (K. R. Scherer & Ekman, 1984; K. R. Scherer et al., 2001), which divides the emotion into five components as five distinctive functions, is the appraisal-driven synchronization of the changed component states as the response to the evaluation of a stimulus event (Sander, Grandjean, & Scherer, 2005). Component process model focuses on the variability of different emotional states and appraisal patterns. It enhanced the recognition of multiple emotions and integrated the contextual information with a better interpretation capability. There is a limited choice of automatic tools which are based on the appraisal-base approach.

2.2. Sentiment Analysis

Sentiment analysis (also known as opinions mining) represents the subjectivity which is extracted from the text. Sentiment analysis could use natural language processing (NLP) as well as text analysis. Sentiment analysis is supposed to study the

attitude of the writer. Sentiment analysis aims to provide a chance to map the overall contextual subjective information of the text into an emotion model. However, some words/sentences may cause confusion (metaphor, exaggeration) or paradox (like typing mistake, sarcasm). This specialty makes the emotional annotation with lexical semantics still difficult (Carlo Strapparava & Mihalcea, 2007).

2.2.1. Natural language processing

NLP enables computers to automatically understand or derive the meaning of the inputed human natural language. In the early phase, NLP systems were based on the set of hand-written rules or restricted vocabularies to translate the sentence based on the language grammar or transformational grammar (Chomsky, 1995). After the introduction of machine learning, NLP systems are released from the annotated training data to semi-annotated and non-annotated data. Machine learning makes the NLP system more powerful to extract the emotion from the language (social software, online comments and etc).

There are many models, such as Hidden Markov model (Baum, Petrie, Soules, & Weiss, 1970), Cache Language model (Kuhn & De Mori, 1990) and etc. These models focus on the probabilistic decision-making mechanism based on the weights or the features extracted from the input text. These models largely expanded the usage of the sentiment analysis from the extraction of subjective information in text to natural language understanding in the first-order logic. Therefore, NLP provides an easier format for the computer to parse the sentence regarding grammatical analysis, text simplification and so on.

2.2.2. Text analysis

Text analysis usually refers to the text mining, aims to derive novel and interesting information from the lexical analysis of text. Its tasks include text categorization, text clustering, entity relation modeling, etc. Text analysis parse the structure of sentence rather than the meaning of the sentence like NLP. Recently, text analysis shifts the focus from semantic structure to words classification. The main method of text analysis depends on the statical pattern learning and frequency distribution study. Its lexical analysis derives the linguistic features and remove irrelevant words.

Recently, Schwartz and his colleagues had conducted the research about the personality impacts on the vocabulary usage of Facebook (Schwartz et al., 2013) by using the dimensional approach of emotion representation, the Five Factor model of personality, and the standard software LIWC. It had been proven that different personalities as well as different emotions could influence the linguistic features of the vocabulary in the text (Hirsh & Peterson, 2009; Mairesse et al., 2007).

2.3. Facial Expression

Facial expression, as the direct reflection of mental state, is one of the most explicit expressions to extract the emotion. Hoffman (2006) also pointed out dynamic facial

expression is more beneficial to the emotional recognition than static facial image. Therefore, a short video is more persuasive than an image.

Currently, the automatic facial expression analysis has two approaches: facial affect detection and facial muscle action detection.

2.3.1. Facial affect detection

Since the basic emotions of facial expressions are universally interpreted (Ekman & Friesen, 2003; Keltner, Ekman, Gonzaga, & Beer, 2003), the annotation of six basic facial expressions is widely applied in the research (Cohn, 2006). The researchers have used the basic facial expressions to detect other cognitive states such as interests (El Kaliouby & Robinson, 2005), pain (Bartlett et al., 2006), and fatigue (Gu & Ji, 2005). Because the real world has fewer constraints and controls, the accuracy and reliability of facial emotion detection are seriously challenged.

2.3.2. Facial muscle action detection

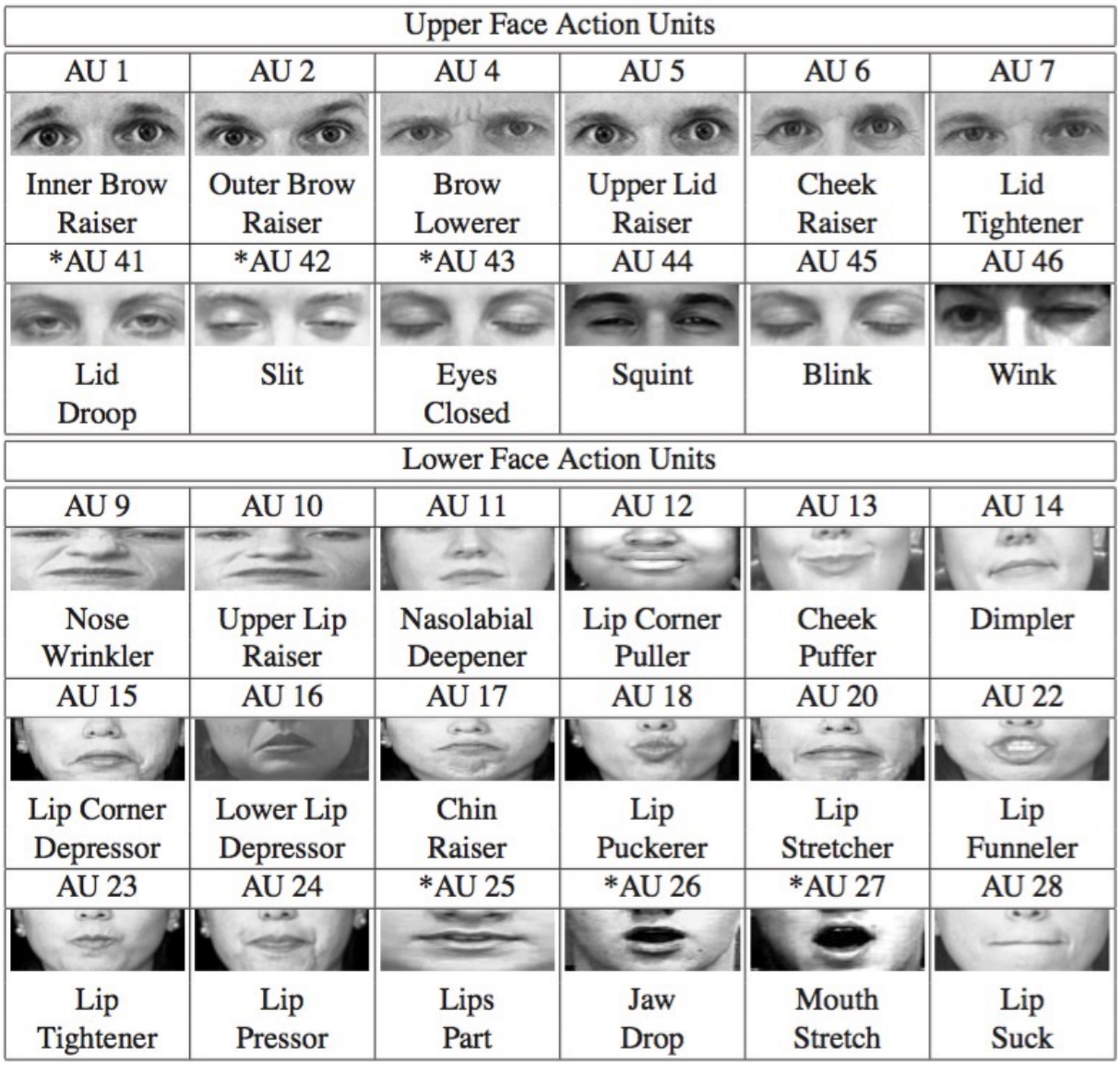

Figure 1. FACS action units (AU). AUs with "*" indicate that the criteria have changed for this AU. (Ekman, 1989b; Tian, Kanade, & Cohn, 2005)

Facial Action Coding System (FACS; P. Ekman & W. V. Friesen, 1978), which consists of 44 different action units (AUs) as illustrated in figure 1, is the framework of the facial muscle action detection. It has been widely applied. Because the FACS and its AUs are capable of coding nearly any anatomical facial movement. Most emotions and cognitive states of facial expressions can be independently interpreted. While comparing to the subjective facial expression labels, FACS is objective and comprehensive to describe facial expressions. FACS associates the facial expressions with the meaning as perceived from previous research and literature.

In order to objectively test the accuracy of facial expression analysis, there are several facial expression databases to use, such as MMI database[1] (Pantic, Valstar, Rademaker, & Maat, 2005), Yin Facial Expression Database (Yin, Wei, Sun, Wang, & Rosato, 2006), and Cohn-Kanade Facial Expression database (Kanade, Cohn, & Yingli Tian, 2000).

2.3.3. Posed vs spontaneous facial expression

Although Ekman (1989a) supports the universals in facial expressions of emotions, human can still recognize the posed and the spontaneous facial expressions accurately via the difference between truthful and deceptive visual expressions of face and body (Ambadar, Schooler, & Cohn, 2005; Darwin et al., 1998; Ekman, 2003; Schmidt & Cohn, 2001). The posed and the spontaneous facial expressions are mediated separately with significantly different methods (Miehlke, Fisch, & Eneroth, 1973). Those differences lead to the variances of facial muscle movement (Ekman, 2009). Valstar had explored the method to automatically distinguish the spontaneous facial behavior from posed one by detecting the movement of eye brows including speed, intensity, duration, and etc (Valstar, Pantic, Ambadar, & Cohn, 2006). With more related features and details, affective states of facial expression are more difficult to be hidden.

2.4. Speech

Researchers also have made a great work to mapping the audio expression with emotional representations like dimensional approach (R. Cowie et al., 2001) as well as appraisal-based approach (Grandjean et al., 2008).

2.4.1. Linguistic and paralinguistic

Speech, as a main communicative method for human, consists of two types of information including the explicit message (linguistic) and the implicit/vocal message (paralinguistic) (R. Cowie et al., 2001). Linguistic information aims to identify the targets which are related to linguistic functions in the voice pitch, intensity and intonation. Paralinguistic information means the information with no linguistic function. For example, vibrations are just related to spectral properties instead of word identity.

---

[1] The acronym MMI means M&M Initiative. The two ‘M’ represent the initials of the two main authors.

However, the boundaries between those two types information are controversial. It had been proven that these explicit messages and implicit messages share a logical link (Ladd 2008). The linguists assumed that some linguistics, which could be intuitively understood by users, are mistakenly classified as paralinguistic (Cowie, Douglas-Cowie et al. 2001). The paralinguistic cues are (at least) partly and contextually associated with linguistic (Roach, Stibbard et al. 1998).

2.4.2. Emotional related speech variables

There are four broad speech variables related to the emotional expression. They are tone type, pitch contours, continuous acoustic measures and voice quality. Table 1 illustrates how these variables associate with emotions. The tone-based level of description is the focus of linguistic tradition. The prosody is usually described by intonational phrases or tone groups. Each tone group contains a prominent/nuclear tone which is usually the tone on the last stressed syllable. The prominent/nuclear tone could be a rising or falling or combined or level one. Cowie et al. suggested that different types of tones have the association with different emotions (2001).

Pitch contours describe the geometric patterns of pitch variation. In the previous studies (Ladd, Silverman, Tolkmitt, Bergmann, & Scherer, 1985; Uldall, 1960), listeners were asked to indicate the emotion based on the contour types. In the study of Ladd et al. (1985), the results also showed that contour type was related to arousal states.

Table 1. Table of Speech and Emotion(a) (R. Cowie et al., 2001)

		Fear	Grief	Surprise/Astonishment
Acoustic	Pitch	Increase in mean F0, range F0, perturbation, variability F0 movement	Very low range, low median, raised mean F0, slow change	Below normal mean F0, range F0
	Intensity	Normal	Increased	Decreased
	Duration	Increased rate, reduced rate	Slow—due to high rate of pause to phonation time, longer vowels and consonants	Tempo normal, tempo restrained
	Spectral	Increase in high-frequency energy		
Contour		Disintegration of pattern and great number of changes in direction of pitch	Long sustained falling intonation throughout each phrase	Sudden glide up to a high level within the stressed syllables, then falls to mid-level or lower level in last syllable
Tone based		Falling tones		Fall rise nuclear tone with falling head (in questions), high fall preceded by rising head (in interjections), high rise tone
Voice quality		Tense	Whisper	Breathy
Other		Precise articulation of vowel/ consonant, voicing irregularity due to disturbed respiratory pattern	Voicing irregularities	

Table 1. Table of Speech and Emotion(b) (R. Cowie et al., 2001)

		Affection/Tenderness	Coolness/Hostility	Puzzlement
Acoustic	Pitch	Higher mean, lower mean, narrow range	Decrease in mean F0	High mean, wide range
	Intensity	Reduced		Low
	Duration	Slow rate		Slow
	Spectral			
Contour		Slightly descending melody, steady and slightly upward inflection		
Tone based			Low falling nuclear tone, high head followed by rise-fall nuclear tone	Rising tones
Voice quality		A little nasal articulation		
Other		Audible off-glide in long stressed syllables		

Table 1. Table of Speech and Emotion(c) (R. Cowie et al., 2001)

		Anger	Happiness	Sadness
Acoustic	Pitch	Increase in mean, median, range, variability	Increase in mean, range, variability	Below normal mean F0, range F0
	Intensity	Raised	Increased	Decreased
	Duration	High rate, reduced rate	Increased rate, slow tempo	Slightly slow, long pitch falls
	Spectral	High midpoint for av spectrum for nofric portions	Increase in high-frequency energy	Decrease in high-frequency energy
Contour		Angular frequency curve, Stressed syllables ascend frequently and rhythmically, irregular up and down inflection, level average pitch except for jumps of about a musical fourth or fifth on stressed syllables	Descending line, melody ascending frequently and at irregular intervals	Downward inflections
Tone based		Falling tones		
Voice quality		Tense, breathy, heavy chest tone, blaring	Tense, breathy, blaring	Lax, resonant
Other		Clipped speech, irregular rhythm basic opening and closing, articulatory gestures for vowel/consonant alternation more extreme	Irregular stress distribution, capriciously alternating level of stressed syllables	Slurring, rhythm with irregular pauses

Table 1. Table of Speech and Emotion(d) (R. Cowie et al., 2001)

		Excitement	Warmth
Acoustic	Pitch	Wide range	
	Intensity	High	
	Duration	Fast	
	Spectral		
Contour			
Tone based		Falling and rise-fall tones	Wide ascending and descending heads
Voice quality			
Other			

Table 1. Table of Speech and Emotion(e) (R. Cowie et al., 2001)

		Sarcasm/Irony	Boredom	Anxiety/Worry
Acoustic	Pitch		Decrease in mean F0	Increased in mean F0
	Intensity		Decreased	
	Duration	Restrained tempo	Increased rate, decreased rate	
	Spectral			
Contour		Stressed syllables glide to low level in wide arc		
Tone based		Low rise-fall tone preceded by rising glissando pretonic, level nuclear tone	Level tone	
Voice quality		Tense articulation leading to grumbling, creaky phonation		
Other				

Continuous acoustic variables also correlate with emotions (R. Cowie et al., 2001) Particularly, pitch, duration, intensity, and spectral makeup are relevant. Lieberman and Michaels (1962) used single acoustic parameters (like F0) to test. The listeners were asked to identify what emotion is being expressed. The results had shown that some emotional information is expressed paralinguistically.

Voice quality is usually described auditorily by terms of tense, harsh and breathy. These auditory qualities may map on spectral patterns (Banse & Scherer, 1996; Roddy Cowie & Douglas-Cowie).

2.5. Fusion of modalities

In the case of multiple modalities for emotional recognition, it is inevitable to deal with two issues : when to mix the modalities and how to mix the modalities. Currently, there are two approaches. The modalities are mixed either in the feature level (estimated with the maximum likelihood) or decision level (most joint statistical properties may lost) (Corradini, Mehta, Bernsen, Martin, & Abrilian, 2005). For both approaches, each single modality is assumed to be independent of others to track data. The decisional level approach uses separate classifier for each modality. Then, all the outputs are combined at a later stage to finalize the hypothesis of the detected affective behavior. The feature level approach assumes a strict time synchrony exists among different modalities. The decision level approach allows the asynchronous factors and the flexibility of modalities. Additionally, adaptive strategy can be used to weighing different modalities. Compared to decision level approach, feature level approach results in a higher and more complex dimensional data set. But feature level approach can perfectly contain the co-occurrence information of different modalities (multimodal cues exist at the same time).

Compared to automatic tools of single modality, the tools of multimodality provide a better result in terms of emotion classification and recognition (Jaimes & Sebe, 2007; Tan & Nareyek, 2009; Yang et al., 2008). But multimodal tools would lose the co-occurrence information of detected emotions in different modalities. The single modality tool just records the emotion change of persons in the selected modality. Therefore, single modality tools are better to classify the emotion change of people in different modalities/tools. In the thesis, the feature level approach and the single modal tools are more proper.

2.6. Dimensionality of emotion recognition

In the emotion recognition, the feature level approach of modalities fusion usually use dimensional approach to represent emotions. It has to deals with a large number of dimensions. It is a data space with a high dimensionality which consists of the emotion representations in different modalities. For example, 2520 features have been extracted for each frame of the input facial video (Valstar & Pantic, 2007), 4843 features have been extracted from each utterance (Wöllmer et al., 2008), and 61 features extracted from speech segments (Kim, 2007). In this case, each feature has less training samples than the proper requirement for target classification.

Dimensionality simplification or feature selection method is appropriate to alleviate the problem. In this thesis, the emotional connection of features/dimensions is the fundamental principle to simplify the dimensionality. For example, the affective processes of LIWC dictionary , which has direct connection with emotions (positive emotion, sadness, etc), could be the simplified dimensions of the text.

2.7. Resources

2.7.1. Database/Corpus

Table 2. Databases summary for multiple modalities of affect recognition (G. McKeown et al., 2012) [2][3]

Database	Semaine database	The Vera am Mittal speech database	MHi-Mimicry database	USC Creative IT database
Reference	McKeown, Valstar et al., 2012	Grimm et al., 2008	Sun et al., 2011	Metallinou et al., 2010
Data Type	Spontaneous	Spontaneous	Posed	Posed
Modalities	Audiovisual, Facial expression, emotional speech text in transcripts	Audiovisual, facial and bodily expression, emotional speech	Audiovisual, facial and bodily expression	Audiovisual, speech, and bodily expression
Subjects	One user and one operator (human/ virtual agent) interact conversationally	various participants in the show	28 males and 12 females participate in one discussion and one role-playing game	19 actors
Categorical annotation	Basic categories of emotions	Not applicable	Social signaling cues	Not applicable
Dimensional Annotation	Valence, activation, power, anticipation/ expectation, intensity	Continuous annotation for valence, activation, and dominance	Not applicable	valence, activation, dominance, as well as interest, naturalness, creativity and actor verbs of theatrical performance rating
Appraisal annotation	Not applicable	Not applicable	Not applicable	Not applicable
Annotators	Feeltrace coders and observers	17 observers	MHi-Mimicry annotation tool	Feeltrace tool and audience
Content	Three types of recordings based on the SAL. In total, over hundreds videos have been added and annotated.	12 hours of audio-visual recordings of German TV talk show “Vera am Mittag” segmented into dialogue acts and utterances	The dataset consists of 54 recordings and 43 subjects imagery. 34 recordings are discussions and 20 ones are of the role-playing game.	9 sessions of audiovisual data which contain 40 two-sentence exercises and 19 paraphrases
Speech language	English	Germans	English	English
Public availability	Yes	Yes	Yes	Yes

[2] MHi represents the mimicry in human-human interaction

[3] USC means University of Southern California

Taking the consideration of our goals, a multiple modalities database is needed to simulate the real emotional interaction. The available options are illustrated in table 2.

2.7.2. Automatic Tools

According to the selection of modalities (section 1.3) and the fusion approach (section 2.5.1), a survey about existing tools was conducted. The tables from 3 to 5 display the tools of single modality in sentiment analysis, speech analysis and facial expression (G. McKeown et al., 2012).

Table 3. Sentiment analysis tools (Bradley & Lang, 1999; J. W. Pennebaker et al., 2007; Socher et al., 2013)

System	LIWC	Affective Norms for English Words	Deep learning sentiment analysis	Werfamous
Annotator	External annotation file (dictionaries, linguistic list)	Testers	Sentence parsing model based on sentiment treebank	Its own analysis datasets, categories and subjects
Features	Physiologically meaningful categories, linguistic frequency	The self-assessment manikin	Utterance of polarity/ objectivity	scored based on twitter and web searched results
Input source	Text files	individual word	Text input	Text input
Categorical annotation	Psychometrics of words and its usage	Not applicable	Sentiment treebank with emotionally labeled words	Not applicable
Dimensional annotation	Language dimensions based on linguistic categories (article, verb, noun and etc)	Dimensions of pleasure, arousal, and dominance	Not applicable	Positive/ Negative tone
Appraisal annotation	Not applicable	Not applicable	Not applicable	Not applicable
Platform	Mac/Windows/Linux	Windows	Windows/web-based	web-based
Public availability	Yes (web-based version)	Only public to research institution	Yes	Yes (online)

Table 4. Facial expression tools (iMotions, 2015; Littlewort et al., 2011; Vicarvision, 2015)[4]

Name	FaceReader	CERT	iMotions
Annotator	FaceReader coder, integrated classification models (including Asian people)	Multivariate logistic regression classier based on final AU parameters, extension modules,	iMotions' coders
Features	Dimensional and affective representation, facial action units, stimuli and event markers	Facial action units, facial features, expression intensity	Emotional stimuli, facial feature, facial action units
Input source	Video	Video	Video
Categorical annotation	6 basic facial expression category	6 basic facial expression category	7 basic facial expression and 2 advanced expressions
Dimensional annotation	Circumplex dimensions of valence, arousal, contempt as well as time-frame	Not applicable	Dimensions of valence and time-frame
Appraisal annotation	Not applicable	Not applicable	Not applicable
Platform	Windows	Windows	Windows
Public availability	Yes (only institution)	No (server shut down)	No (institution may accpeted)

[4] CERT equals the initials of Computer Expression Recognition Toolbox

Table 5. Speech analysis tools (Degottex, Kane, Drugman, Raitio, & Scherer, 2014; Eyben, Wöllmer, & Schuller, 2009; Vibrations, 2016; Vogt, André, & Bee, 2008)[5][6][7][8][9][10][11][12][13][14]

Name	EmoVoice	GVC emotion recognition	OpenEar	Covarep
Annotator	Classifiers of Naïve Bayes and SVM	Users, Good Vibrations' classifiers	SVM based on LibSVM library	SVM with radial basis function kernel.
Features	Pitch, energy, Mel-frequency cepstral coefficients (MFCCs), duration, voice quality and spectral information	Pitch, intensity, resonances, dullness, sharpness, softness, tempo, and phrasing	Signal energy, loudness (pseudo), mel-spectra, MFCCs, pitch, voice quality formats and LPC coefficient	Glottal source and spectral envelope
Input source	Audio	Voice	Audio	Audio
Categorical annotation	Predefined emotion categories by training speech database	Basic emotion category and user feedback	Pre-trained models	Not applicable
Dimensional annotation	Not applicable	Not applicable	Not applicable	NAQ, QOQ, PSP, H1–H2, peakSlope, MDQ and R_d
Appraisal annotation	Not applicable	Not applicable	Not applicable	Not applicable
Platform	Mac/Windows/Linux	Mac/Windows/ Linux/iPhone	Mac/Windows/Linux	Windows
Public availability	Yes	No	Yes	Yes

[5] GVC equals the initials of the company name Good Vibrations Company

[6] SVM means support vector machine learning algorithm

[7] MFCC means Mel Frequency Cepstral Coefficient . MFCC is the coefficient of mel-frequency cepstrum and represents the sound power spectrum in a short term.

[8] LPC Linear predictive coding as an audio signal processing and speech processing tool. Its coefficients represents the spectral envelope of speech signal.

[9] NAQ means normalized amplitude quotient

[10] QOQ means quasi-open quotient

[11] PSP means parabolic spectral parameter

[12] H1H2 means the difference of glottal harmonic amplitude

[13] MDQ means maxima dispersion quotient

[14] R_d means estimation of the Liljencrants-Fant glottal model. It represents the regression of the shape parameters.

3. Methodology

According to the research requirements (section 1.3.1), this research need a database that contains recordings of users interacting in a conversation. Then, the selected database should provide the manual annotation of the users' emotional expressions during the conversation. Such manual annotation should be computable. Additionally, the database should also should contain the fixed emotional stimuli which aims to influence the emotions of users. Each recorded session of the selected database should have multimodal recordings (section 1.1.3) of the emotional expressions: video recordings for facial expression analysis, audio recordings for speech analysis, and transcriptions of the speech for the text analysis. According to the table 2, Semaine database is the best option.

According to section 2.5, the selected automatic tools should be single modality. The dimensional approach is suggested to provide the continuous and computable representation of people's emotions for the automatic annotation in the section 2.1. Then, the automatic tools should also use the dimensional approach. FaceReader (version 6.0; Vicarvision, 2015), Covarep (1.4.1) and LIWC (version 2015) are selected.

3.1. Semaine Database

The Semaine database is based on the sensitive artificial listener (SAL) scenarios (Douglas-Cowie, Cowie, Cox, Amier, & Heylen, 2008) which could generate the emotionally colored conversation. In the scenario, a virtual character emotionally interacts with the participant via a conversation. There are four types of virtual characters (gloomy for Obadiah character, happy for Poppy character, angry for Spike character, and sensible for Prudence character). Each character can act as the emotional stimulus to influence the users' emotions. Moreover, the sample of Semaine database contains the video, the audio and the text files of the person's expression. Therefore, the automatic emotion analysis tools can annotate the emotion of participants during the interaction. In the database, there is a large amount of manual

Figure 2. Semi-automatic SAL screens for the user and the operator (G. McKeown et al., 2012).

annotations of the user emotions and other aspects, based on the judgement of the professional annotators.

The virtual characters of Semaine database respond to users according to a predefined script of emotional phrases. The script aims to enable the SAL character to emotionally influence the users.

3.1.1 Data type

Semaine Database recordings are built on SAL scenarios. These scenarios were designed to create conversations which have enough nonverbal features to sustain the verbal and emotional communication between two speakers. One of the speakers is the 'operator' who simulates a machine with very little competence of language. The 'operator' speaks according to a 'script' that composed of emotional phrases. The other speaker is called the 'user' and always acts as human. The 'operator' and the 'user' locate in separate rooms.

The database recordings can be classified into three types - Solid SAL, Semi-automatic SAL, and Automatic SAL.

Solid SAL: It is designed to record the behaviors of a conversation. But the 'operator' plays the SAL character and cannot answer any question asked by the 'user'. But the 'user' is encouraged to talk to the SAL character as naturally as possible. Meanwhile, the 'operator' and the 'user' can see and talk to each other by a teleprompter screen and speakers.

Semi-automatic SAL: The 'operator' chooses one of the predefined phrases via the graphical interface in figure 2. Then, the selected phrase, which was recorded in an audio file, is spoken by the computer. The audio file was recorded by an actor with an

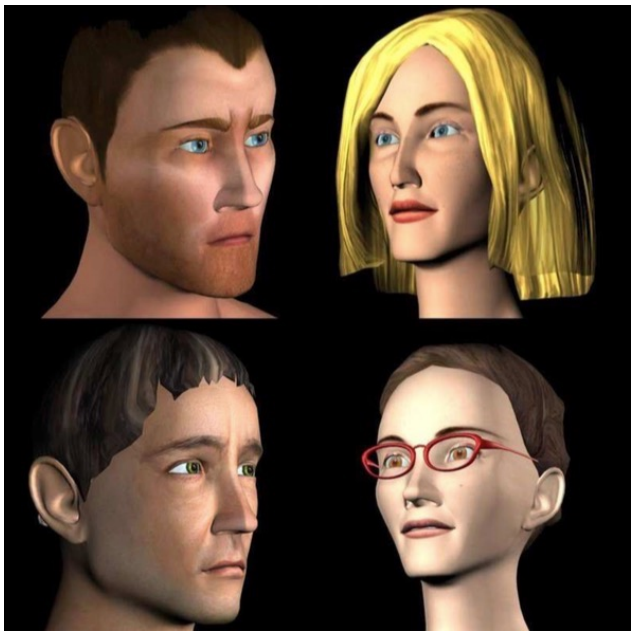

Figure. 3. The avatars of Automatic SAL characters (G. McKeown et al., 2012). Counter clockwise from Top-right: Poppy, Spike, Obadiah, and Prudence.

appropriate voice for the SAL character. Meanwhile, the 'user' can see an abstract face on the screen. The abstract face is illustrated in figure 2, too. Additionally, Semi-automatic SAL scenario has three cases. In the first one, the 'operator' can see and hear the 'user' before an appropriate utterance is chosen. In the second case, the 'operator' could only see the 'user'. In the third case, the 'operator' could see the 'user' with the audio which is filtered to remove verbal information.

Automatic SAL: The 'operator' is the SAL character which displays as a life-like avatar (figure 3). Each character has its own stereotypical appearance and voice. Before the 'operator' speaks, the Semaine project system (Schroder et al., 2012) automatically decides the utterances and non-verbal actions for SAL character.

Solid SAL is the most useful version for our experiment, because it provides the most natural and least constrained emotional interaction. Semi-automatic SAL uses a simplified screen to display an abstract face which may omit the details of the nonverbal behaviors (the facial movements of eyes contact, lips, eye lids, eyebrows, and etc) of the operator and make the user feel unnatural. What's worse, other two cases don't provide the speech or the linguistic feedbacks. In this way, the user has the problem of fully receiving the emotion impacts from the operator. Automatic SAL interacts with 'user' via an avatar. Its utterances and non-verbal actions are chosen automatically. It has the same problem of the first case in the Semi-automatic SAL. Similarly, the user also has the problem of perceiving all the emotion details from the avatar.

3.1.2. Character

Table 6 . Dimensions for Semaine database annotation.

Full rating dimensions	Optional dimensions			
	Basic emotions	Epistemic states	Interaction process analysis	Validity
Valence	fear	certain/not certain	shows solidarity	breakdown of engagement
Activation	angry	agreeing/not agreeing	shows antagonism	anomalous simulation
Power	happiness	interested/not interested	shows tension	marked sociable concealment
Expectation	sadness	at ease/not at ease	releases tension	marked sociable simulation
Intensity	disgust	thoughtful/not thoughtful	makes suggestion	
	contempt	concentrating/not concentrating	asks for suggestion	
	amusement		gives opinion	
			asks for opinion	
			gives Information	
			asks for Information	

Initially, SAL technique was used to generate the natural conversation which was emotionally colored by virtual character (Douglas-Cowie et al., 2008). Then, it was proved that the character with a coherent personality and agenda can prevent the conversation from breaking down. There are four characters created. They are Spike (angry), Poppy (happy), Prudence (sensible), and Obadiah (gloomy). Each character has one personality which was designed to be coherent and different from other three. The designed personality will drive the character to influence expressed emotions of users in a certain direction by giving well designed types of responses. For example, the Spike character with angry personality responds empathically to the user's angry expression, and critically to user's other emotional expressions.

3.1.3. Annotation

For the Solid SAL scenarios, human annotators continuously recorded the perceived users' emotion during the conversation. The Solid SAL manual annotation contains five full rating dimensions (Valence, Activation, Power, Expectation and Intensity) and other 27 optional dimensions. These dimensions are displayed in table 6. The optional dimensions include basic emotions, epistemic states, interaction process analysis and validity. Most items of the basic emotions are selected from Ekman's list (Ekman, 1992). Baron-Cohen and his colleagues (Baron-Cohen, 2003) defined the epistemic states. These states are used to label where a relatively clear epistemic state appears in the clip. The labels of interaction process analysis, as the subset of the system categories used in Interaction Process Analysis (Bales, 1950), are used to indicate when the issues become salient in the dialogue management. Validity indicates the cases when the user avoids a straightforward emotional communication.

Table 7. Distribution of Semaine optional traces for the 13 most used options. No others reach 5 per character or 10 across characters. (G. McKeown et al., 2012). Bold numbers are referred in the text.

	Optional Trace	Obadiah	Poppy	Prudence	Spike
Basic Emotions	Happiness	2	15	5	1
	Sadness	**13**	1	1	0
	Anger	1	0	2	**8**
	Amusement	8	14	13	12
	Contempt	0	0	1	5
Epistemic States	Certain	4	5	9	4
	Agreeing	15	11	15	15
	Interested	3	3	2	2
	At Ease	5	6	7	9
	Thoughtful	10	9	8	4
Interaction Process Analysis	Show Antagonism	0	1	1	6
	Gives Opinion	12	7	9	11
	Gives Information	10	20	19	9
Mention these concerns of annotations of users not operators					

The optional dimensions are only annotated when raters think the optional dimensions could apply to the situation. During the interaction, the usages of each optional dimensions vary with the character. For example, sadness (13) appears frequently only in the conversational interaction with Obadiah character in table 7 (table 7 only includes the data from the 6 raters who traced all the clips for the sake of balance). The Anger (8) dimension is frequently annotated only during the interaction with Spike character. This table also shows that some optional dimensions are rarely used in the annotations. Besides the dimensional scripts of emotional data, Solid SAL sessions were transcribed and time aligned. The utterances of the 'operator' and the 'user' were also recorded in the transcripts.

Table 8. Alpha coefficient for functionals associated with each trace dimension (* indicates alpha>0.6 – the lowest value commonly considered acceptable ** indicates alpha>0.7 – almost always considered acceptable † indicates non-acceptable values) (G. McKeown et al., 2012). Bold numbers are referred in the text.

	Intensity	Valence	Activation	Power	Expectation
Mean all	0.74**	0.92**	0.73**	0.68*	0.71**
sd bins	0.83**	0.75**	0.65*	0.61*	0.68*
min bin	0.23†	0.90**	0.43†	0.43†	0.43†
median bin	0.72**	0.91**	0.72**	0.67*	0.68*
max bin	0.74**	0.92**	0.73**	0.68*	0.71**
AveMagnRise	0.74**	0.49†	0.53†	0.39†	0.58†
SDMagnRise	0.74**	0.60*	0.63*	0.32†	0.59†
MaxMagnRise	0.75**	0.56†	0.64*	0.25†	0.63*
AveMagnFall	0.68*	0.45†	0.55†	0.55†	0.51†
SDMagnFall	0.66*	0.45†	0.63*	0.60*	0.49†
MinMagnFall	0.60*	0.46†	0.59†	0.60*	0.41†

Table 9. Reliability Analysis (G. McKeown et al., 2012).

	QA analysis	Correlational (α) analysis
Total no. of datasets (i.e. sets of 6 or 8 traces of particular clip on a particular dimension)	**305**	303
Fail stringent test (alpha > 0.85, p(QAg) < 0.01)	90	104
Fail moderate test (alpha > 0.75, p(QAg) < 0.05)	43	41
Fail minimal test (alpha > 0.7)	n/a	28

The annotations of the full rating dimensions are reliable. The reliability is measured in two levels. The first level of reliability is considered between clips. The reliability

between clips is based on the calculation of mean, standard deviation, etc. The reliability can measure the agreement of the same dimension (e.g. mean valence) between different ratings. Then, the standard Cronbach's alpha is used (Cronbach, 1951). The results are illustrated in table 8. Most of the results are reliable. Mean all, sb bins, median bin and max bin are rated reliably in all the traces. Intensity shows consistent patterns of rises and falls in different aspects. The second level of reality is to measure the intra-clip agreement. It means the agreement between different ratings/ raters of a single aspect (e.g. valence rises) during the same clip. It is possible to calculate the correlation and the alpha coefficients between different lists of these rating values. But researchers are wary of the correlation calculation (G. McKeown et al., 2012). Because the successive values of each trace are not independent. In other words, the measurement is ordinal rather than interval. Therefore, an alternative method has been used to avoid these problems. It is the quantitive agreement (QA; R. Cowie & McKeown, 2010). Table 9 summarizes the results of 305 sets of traces. In terms of correlational analysis, less than 10 percents fail to reach the standard criterion ($\alpha = 0.7$), and less than 30 percents fail to reach the stringent criterion ($\alpha > 0.85$). In terms of QA analysis, more than 70 percents meet the stringent criterion. Overall, the QA analysis is more stringent than correlational analysis.

Among three types of Semaine database recordings, Solid SAL has the largest body of annotations. Each Solid SAL sample contains media files (video, video without audio, and audio, transcript files) and annotation files (FeelTraced annotation files). The annotation files have 5 full rating dimensions and 27 optional dimensions. Sometimes, the dimension/optional dimension has more than one corresponding annotation file. Because each user clip maybe rated by more than one annotator. When the filename ends with a number, it means an extra attempt of annotation. The file, which has the largest number at the end of its filename, is the most correct.

3.2. Selected tools

Based on above discussion, FaceReader, Covarep and LIWC were chosen. All of them use single modality and dimensional approach to recognize emotions.

3.2.1. FaceReader

Table 10. FaceReader outputs.

Basic emotions	Full rating dimensions
Netural	Arousal
Happy	Valence
Sad	
Angry	
Surprised	
Scared	
Disgusted	

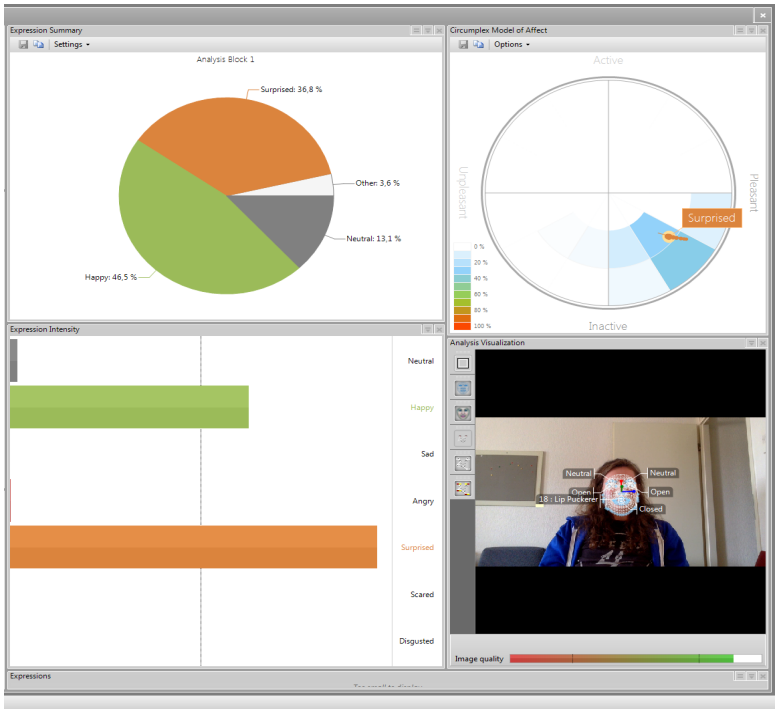

Figure 4. FaceReader: Project Analysis Module (Vicarvision, 2015).
Clockwise from Top-right: the Circumplex model of affect, facial movements analysis, the intensities of basic emotions, and accumulated percentage of basic emotions.

Figure 5. FaceReader analysis interface (Vicarvision, 2015).
Clockwise from Top-left: input video, face movement analysis and continuous signals of basic emotions and rating dimensions.

Facial expression is one of the most direct reflection of personal emotion. Hoffman and his team (2006) pointed out that dynamic facial expression is more beneficial to emotional recognition than static facial image. Although FaceReader could process both

videos and images, videos were selected as the sole form of input data for facial expression analysis in the thesis.

As a facial action coding system, FaceReader analyzes the facial movements/AUs to extract emotional information. The output dimensions are displayed in table 10. It includes basic emotional dimensions and full rating dimensions. In figure 4, the Circumplex model of emotion is used to label the percentage of basic emotions in a circular space with two bipolar dimensions (active and pleasant). In the figure 5, the analysis of facial expressions can be visualized as a group of continuous signals. In addition, the bar graph displays the intensity of each basic emotion. In the pie chart, the FaceReader calculates the total percentage of each basic emotion. The 'other' category of the pie chart contains any detected emotion with a total percentage less than 5%.

At last, the output file records the dimensional values of basic emotions and full rating dimensions. Each line of output file represents the detected results of one time frame. Some dimensions (Arousal, Valence, and etc) of FaceReader can directly compare with the corresponding dimensions of Semaine annotation. It reduces the dimensionality of emotional representation, breaks the boundary between the FaceReader and the Semaine database.

3.2.2. Covarep

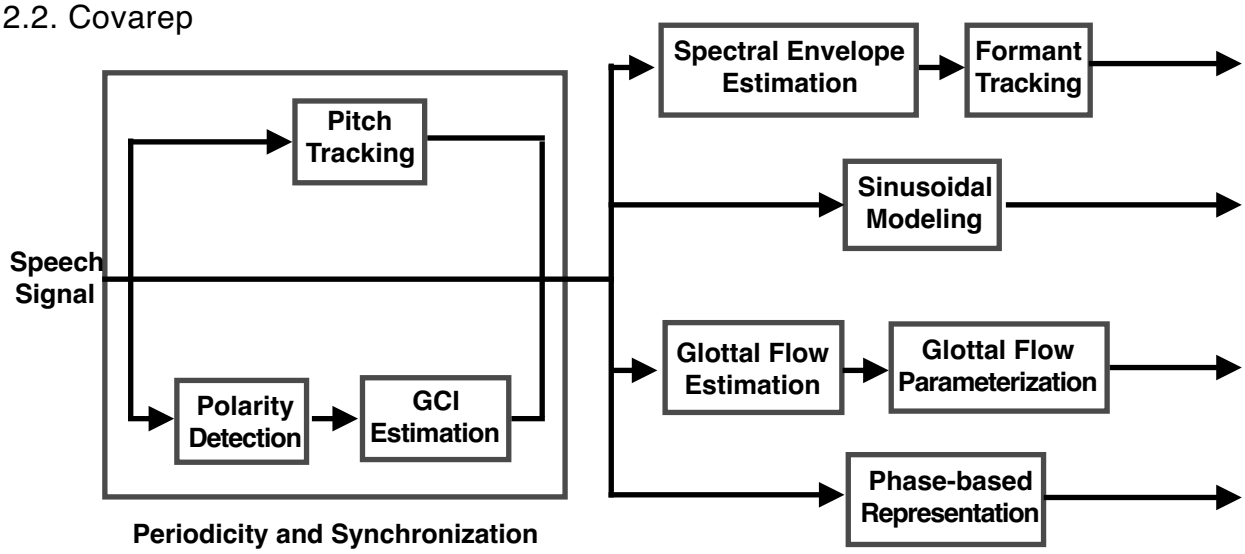

Figure 6. Workflow of the methods in Covarep (Degottex et al., 2014). GCI represents glottal closure instant.

Covarep is the collaborative repository for speech processing algorithms. It aims to support the research by providing an easy access to new speech processing algorithms. Currently, there are five methods in Covarep. The workflow is illustrated in figure 6.

Firstly, the parts of periodicity and synchronization extract the pitch-synchronous information (fundamental frequency, speech polarity and glottal closure instants) for further methods (sinusoidal modeling, spectral envelope estimation and formant tracking, glottal analysis, and phase processing).

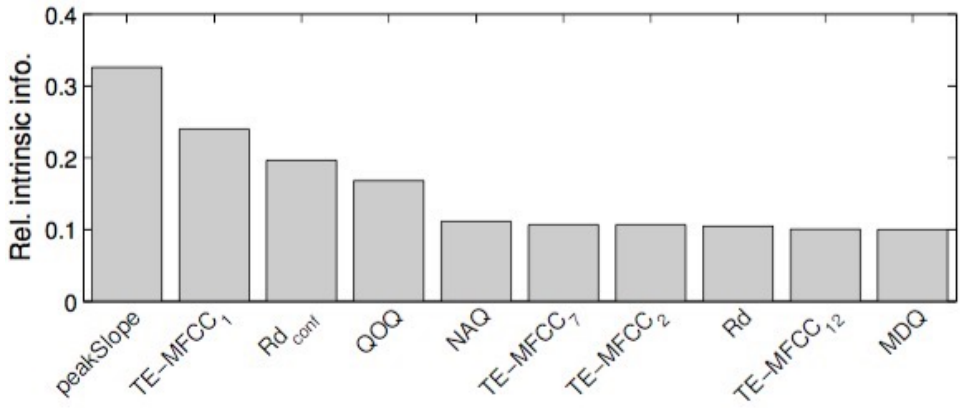

Figure 7. Top 10 features in terms of relative intrinsic information (Degottex et al., 2014).

A set of features, which is extracted by Covarep algorithms, includes MFCCs (O'Shaughnessy, 2008), TE-MFCCs (alternative MFCCs extracted from True-Envelope spectral representation), NAQ, QOQ, PSP, H1–H2, creaky voice, peakSlope, MDQ and Rd (Fant, 1995). Covarep conducted the feature selection based on the assessment of Drugman and Gurban (2007). The top 10 discriminative features are listed in figure 7. Unfortunately, it is difficult for these features to establish a direct psychological connection with Semaine annotation. Therefore, we had to regrettably give up speech analysis for the emotion analysis in the later study.

3.2.3. LIWC

LIWC, as one of the widely used text analysis tool, calculates the frequency of each word category within the targeted text file. LIWC application uses its internal dictionary/ library to classify the words. The internal dictionaries/libraries of words usually indicate the same linguistic domain/category/dimension. Psychologically, word category collects all the related words (e.g. Power category contains 'boss', 'underling', 'president', etc). All the word lists of dictionaries are iteratively edited and judged by LIWC research team. But the software still makes mistakes and errors in identifying and counting words (e.g. 'mad' is a positive word in 'he is mad for her'). The problem could be seldom in the probabilistic model of language use ('mad' related words rarely appears in the same text if 'mad' is positive). This probabilistic approach is useful to solve the errors caused by irony, sarcasm, or metaphor.

The LIWC output file consists of 80 word categories/dimensions. The output data has the file name, 4 general descriptor categories, 22 standard linguistic dimensions, 32 word categories tapping psychological constructs, 7 personal concern categories, 3 paralinguistic dimensions, and 12 punctuation categories. The full list is illustrated in table 11. In this table, word count (WC), words per sentence (WPS) and summary variables are calculated differently. For example, WC is the raw word number of the target text file. WPS represents the mean number of each sentence words in the target text file. Many of these dimensions have been used in the psychological research like the affective process. It would be easier to establish the psychological connection between the LIWC and Semaine database. During the emotional analysis between

LIWC and Semaine annotation, the simplification of dimensionality only includes WC, WPS, and the affective process.

Table 11. LIWC output variables/dimensions (a). Within the same category, "Alphas" means Cronbach alphas (Cronbach, 1951) as the internal reliability of each specific words. The binary alphas are computed on each dictionary word's occurrence/non-occurrence. The raw or uncorrected alphas are calculated by the usage percentage of each category word. (J. W. Pennebaker et al., 2007) [15]

LIWC Dimension					Abbrev	Alpha:Binary/raw
Linguistic Processes	Word Count				WC	
	Words per sentence				WPS	
	Dictionary words				dic	
	Words>6 letters				sixltr	
	Total function words				funct	.97/.40
		Total pronouns			pronoun	.91/.38
			Personal pronouns		ppron	.88/.20
				1st pers singular	i	.62/.44
				1st pers plural	we	.66/.47
				2nd person	you	.73/.34
				3rd pers singular	shehe	.75/.52
				3rd pers plural	they	.50/.36
			Impersonal pronouns		ipron	.78/.46
		Articles			article	.14/.14
	Common verbs				verb	.97/.42
		Auxiliary verbs			auxverb	.91/.23
		Past tense			past	.94/.75
		Present tense			present	.91/.74
		Future tense			future	.75/.02
		Adverbs			adverb	.84/.48
		Prepositions			prep	.88/.35
		Conjunctions			conj	.70/.21
		Negations			negate	.80/.28
		quantifiers			quant	.88/.12
		numbers			number	.87/.61
	Swear words				swear	.65/.48

[15] The function word category excludes common verbs.

Table 11. LIWC output variables/dimensions (b). Within the same category, "Alphas" means Cronbach alphas (Cronbach, 1951) as the internal reliability of each specific words. The binary alphas are computed on each dictionary word's occurrence/non-occurrence. The raw or uncorrected alphas are calculated by the usage percentage of each category word. (J. W. Pennebaker et al., 2007)

LIWC Dimension				Abbrev	Alpha:Binary/Raw
Psychological Processess	Social processes			social	.97/.59
		Family		family	.81/.65
		Friends		friend	.53/.12
		Humans		human	.86/.26
	Affective processes			affect	.97/.36
		positive emotion		posemo	.97/.40
		negative emotion		negemo	.97/.61
			Anxiety	anx	.89/.33
			Anger	anger	.92/.55
			Sadness	sad	.91/.45
	Cognitive processes			cogmech	.97/.37
		Insight		insight	.94/.51
		Causation		cause	.88/.26
		Discrepancy		discrep	.80/.28
		Tentative		tentat	.87/.13
		Certainty		certain	.85/.29
		Inhibition		inhib	.91/.20
		Inclusive		incl	.66/.32
		Exclusive		excel	.67/.47
	Perceptual processes			percept	.96/.43
		See		see	.90/.43
		Hear		hear	.89/.37
		Feel		feel	.88/.26
	Biological processes			bio	.95/.53
		Body		body	.93/.45
		Health		health	.85/.38
		Sexual		sexual	.69/.34
		Ingestion		ingest	.86/.68

[16]

[16] Social processes include all non-first-person-singular personal pronouns and human interaction verbs (talking, sharing).

Table 11. LIWC output variables/dimensions (c). Within the same category, "Alphas" means Cronbach alphas (Cronbach, 1951) as the internal reliability of each specific words. The binary alphas are computed on each dictionary word's occurrence/non-occurrence. The raw or uncorrected alphas are calculated by the usage percentage of each category word. (J. W. Pennebaker et al., 2007)

LIWC Dimension			Abbrev	Alpha:Binary/raw
Psychological Processess	Relativity		relative	.98/.51
		Motion	motion	.96/.41
		Space	space	.96/.44
		Time	time	.94/.58
Personal Concerns	Work		work	.91/.69
	Achievement		achieve	.93/.37
	Leisure		leisure	.88/.50
	Home		home	.81/.57
	Money		money	.90/.53
	Religion		relig	.91/.53
	Death		death	.86/.40
Spoken categories	Assent		assent	.59/.41
	Nonfluencies		nonflu	.28/.23.
	Fillers		filler	63/.18
Punctuation	Total punctuation		Allpunc	
		Periods	Period	
		Commas	Comma	
		Colons	Colon	
		Semicolons	SemiC	
		Question mark	QMrk	
		Exclamation mark	Exclam	
		Dashes	Dash	
		Quotation mark	Quote	
		Apostrophes	Apostro	
		Parentheses	Parenth	
		Other punctuation	OtherP	

4. Experiment Plan

The whole experiment plan contains the preprocess and three phases. The preprocess aims to prepare and format the data for the later processing. The Phase I is planned to conduct the research by exploring how Obadiah and Poppy characters impact the people's emotion as measured by the FaceReader, or as recorded in the manual annotation of the Semaine Database. More characters would be investigated in the Phase II if the results of Phase I had indicated that Obadiah and Poppy characters had different impacts on the expressed emotions of users. Then, the Phase II aims to investigate how the four SAL characters influence the people's emotion in the annotations of the FaceReader and the Semaine Database. After the exploration of the impacts from SAL characters on the facial emotions of users, the Phase III is planned to found out how these virtual characters influence the expressed emotion of participants in the text.

4.1. Preprocess of data

As mentioned above, LIWC and FaceReader were selected to process the multimodal recordings of Semaine database. The preprocess is used to ensure the data of Semaine recordings is well organized and formatted for automatic tools.

4.1.1. Normalizing Semaine transcripts

In the Semaine database, the multiple FeelTraced annotation files of the same recording could be generated by different raters or multiple rating attempts of the same annotator. But the annotated session is finished and the expressed emotions of each session are fixed. In order to reduce the bias of personal judgments, duplicated files were merged and time aligned according to the dimensions of the annotation. During the consolidation, the average value was calculated and kept by omitting the null and the non-numeric values of the merged files.

4.1.2. Discarding failed facial videos

Due to the failure of Facial expression detection, FaceReader generated null and non-numeric values in the automatic annotation. In this way, the consolidation and the calculation of FaceReader results can use the same method in section 4.1.1. If the video had too many time frames of detection failures (more than 30% of the whole video time), the video would be regarded as an unavailable file. Then, other related files (transcripts, FeelTraced annotation files, and audios) would be removed.

4.1.3. Filtering unnecessary text

During the session, only the spoken words of users were the targets. But the transcript files recorded the utterances of both SAL character and participant. Additionally, the transcript files also contained the concrete time of each sentence and the nonverbal descriptions of users and characters.

The speaking time and the nonverbal information were unrelated to the user

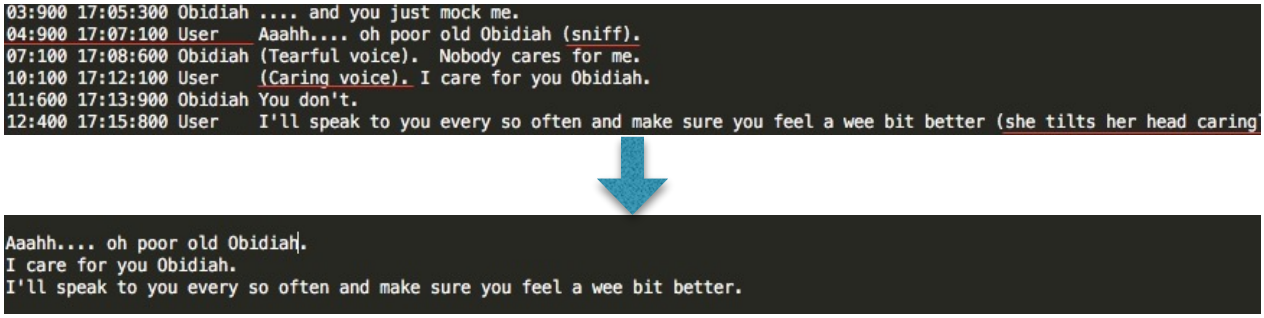

Figure 8. Preprocess of transcripts

emotion. Some words were wrongly spelled or written. Consequently, those types of information should be deleted. Figure 8 displays the process.

4.2. Phase I

By analyzing the annotations of FaceReader and Semaine database, Phase I aims to found out how Obadiah (Gloomy) and Poppy (Happy) characters influence the emotions of users. Because Obadiah character is gloomy, and Poppy character is happy. If the emotional interaction were effective to influence people in the Semaine annotation, the users, who talk to Obadiah character, should respond more gloomily than the users interacting with other characters. Similarly, the users should interact with Poppy character more happily than the users interacting with Obadiah character. In the results of FaceReader, the user, who speaks with Obadiah and Poppy characters respectively, should express contrary emotions. Other characters would be considered if the results of Phase I had succeed in supporting any hypothesis in the section 1.3.2.. Otherwise, it is unnecessary to continue the research.

Facial expression is the most explicit modality for emotional expression. Therefore, FaceReader was evaluated before LIWC.

4.2.1. Investigate the emotions/characters impacts as annotated in the Semaine annotation

Table 12. Experiments for section 4.2.1.
This table displays the codes for the datasets that we will compare. Each code has one letter and one number. The letter indicates the modality/database type. The number indicate the character Type. E.g. D1 means the modality/database D (Seamine database) and the character 1 (Obadiah)

Modalities		SAL character (Obadiah)	SAL character (Poppy)
Manually annotations	Semaine (Multiple modalities)	D1	D2

The comparison plan of Semaine annotation was designed as table 12. The comparison between D1 and D2 groups can analyze the character impacts on the happy and sad emotions as annotated in the Semaine annoation. If the users of D2 group had achieved higher value of happy emotion, it would be regarded as the evidence for the emotional impacts of Poppy character. Similarly, if the users of D1

group had expressed more sad emotion, it would support the assumption about the emotion impacts of Obadiah character.

Table 13. Functionals of basic emotions

Label	Meaning	Valence	Basic emotions …..	……
Min	Minimum value			
Mean	Mean value			
Max	Maximum value			
SD	Standard Deviation			
MinMagnRises	Minimum Magnitude of Rises			
MeanMagnRises	Mean Magnitude of Rises			
MaxMagnRises	Maximum Magnitude of Rises			
SDMagnRises	Standard Deviation of Magnitude of Rises			
MinMagnFalls	Minimum Magnitude of Falls			
MeanMagnFalls	Mean Magnitude of Falls			
MaxMagnFalls	Max Magnitude of Falls			
SDMagnFalls	Standard Deviation of Magnitude of Falls			
FreqChanges	Frequency of value Changes			
FreqRises	Frequency of Rises			
FreqFalls	Frequency of Falls			

During the analysis, the calculation of functionals in each dimension of Semaine annotation was the approach to measure the difference. These functionals included the minimum, maximum, standard deviation, and etc. They are displayed in the table 13. After the functionals analysis, T-test for two independent samples was used to check whether the expressed emotions of D1 and D2 are significantly different from each other.

4.2.2. Investigate the emotions/characters impacts as measured by the FaceReader

Table 14. Experiments plan for Section 4.2.2.
This table displays the codes for the datasets that we will compare. Each code has one letter and one number. The letter indicates the modality/database type. The number indicate the character Type. E.g. A1 means the modality/database A (FaceReader) and the character 1 (Obadiah)

Modalities		SAL character (Obadiah)	SAL character (Poppy)
Outputs of Automatic Tools	FaceReader (Facial Expression)	A1	A2

The investigation would continue if there were any difference in the manual annotation regarding how people respond to Obadiah and Poppy characters. It is

necessary to check how characters influence the emotional expression of people as measured by the automatic tools (table 14). The comparison between A1 and A2 groups was used. The procedures of functionals analysis and the T-test (section 4.2.1) were repeated to analyze the measurements of FaceReader. These analysis were used to explore how the characters influence the user emotion as measured by the FaceReader, and whether the expressed emotions of A1 and A2 groups were significantly different from each other.

4.2.3. Investigate the dimensional correlation between FaceReader and Semaine annotation

Table 15. Experiments plan for Section 4.2.3.
This table displays the codes for the datasets that we will compare. Each code has one letter and one number. The letter indicates the modality type. The number indicate the character Type. E.g. A2 means the modality A (FaceReader or Facial expression) and the character 2 (Poppy).

Characters	Automatic tools	Manually annotations
	FaceReader (Facial Expression)	Semaine (Multiple modalities)
SAL character (Obadiah)	A1	D1
SAL character (Poppy)	A2	D2

Table 15 displays the comparison plan between FaceReader and Semaine annotation (A1&D1, A2&D2). Because SAL characters relate to different personalities (Obadiah - gloomy and Poppy - happy). If these characters had the emotion impacts on the users, the expressed emotions of participants should have been aligned with virtual characters in the annotation of Semaine database. If facial expression and FaceReader were reliable to reflect and detect the user emotions, the emotion impacts of characters should be found in the measurement of FaceReader as well.

Table 16. Psychological connections of basic emotions between FaceReader and characters.

Character	Basic emotions of Semaine annotation	FaceReader and its basic emotions
Obadiah	Sadness	Sad
Spike	Anger	Angry
Poppy	Happiness	Happy
Prudence	Happiness & Anger	Happy & Angry

Table 7 lists the most used 13 optional traces for each character in the Semaine annotation. The table also displayed the frequency of each selected dimension. From the table 7, the impacts on the users' emotions of each character are partially reflected via the most frequently used emotional dimension for each character. If facial expression and FaceReader were reliable to reflect and detect the user emotions, the measurement of basic emotions from the FaceReader should be consistent with the emotion distribution in the table 7. According to the most frequently used emotional

dimension for each character in the table 7, the corresponding dimension of FaceReader annotation is concluded in the table 16. Table 17 lists all the common dimensions between Semaine annotation and the measurement of FaceReader.

Table 17. The common dimensions between FaceReader and Semaine database

Dimensions	FaceReader	Semaine database
Rating dimensions	Arousal	Activation
	Valence	Valence
		Power
		Expectation
		Intensity
Basic emotions	Netural	
	Happy	Happiness
	Sad	Sadness
	Angry	Anger
	Surprised	
	Scared	Fear
	Disgusted	Disgust
		Amusement
		Contempt

The emotionally corresponding dimensions in the table 16 is helpful to explore whether the characters have the emotion impacts on the facial expression of users. The common dimensions, which are listed in the table 17, are helpful to validate the emotion impacts of characters as measured by the FaceReader, and are beneficial to check whether FaceReader can recognize the user emotion as good as the Semaine annotation.

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2} \sqrt{\sum_i (y_i - \overline{y})^2}}$$

Figure 9. Pearson Correlation Calculator. i means the index of trace values. x and y represent the dimensions of FaceReader

Pearson correlation calculator (figure 9) was used to measure the agreement between the measurement of FaceReader and the manual annotation of Semaine database. The correlation calculation was based on the dimension match in the table 16. The final result of correlation analysis for each character were a Pearson correlation

matrix as table 18 (take Obadiah character for example). Each item of the matrix (table 18) represented the correlation between one dimension of Semaine annotation and one dimension of FaceReader measurement.

Table 18. Pearson correlation matrix of Obadiah.

		FaceReader						
Semaine annotation		Arousal	Valence	Happy	Sad	Angry	Scared	Disgusted
	Activation	1	0.72 (e.g.)	…	…	…	…	…
	Valence		1	…	…	…	…	…
	Happiness	…	…	1	…	…	…	…
	Sadness	…	…	…	1	…	…	…
	Anger					1	…	…
	Fear						1	…
	Disgust							1

4.3. Phase II

If the annotations of Semaine database and FaceReader indicated the character impacts on the users' emotion, another two characters (Spike and Prudence characters) should be added into the analysis. Similarly, Phase II should repeat the steps in section 4.2. The new experiment plan was designed as table 19. The functionals analysis and Pearson correlation calculator was still used. But the T-tests in the Phase I should all be replaced with the one-way analysis of variance (ANOVA). Because there are four different groups of participants.

Table 19. Experiments plan for section 4.3. The letter indicates the modality/database type. The number indicate the character Type.
E.g. A1 means the modality/database A (FaceReader) and the character 1 (Obadiah)

Characters	Outputs of Automatic tools	Manually annotations
	FaceReader (Facial Expression)	Semaine (Multiple modalities)
SAL character (Obadiah)	A1	D1
SAL character (Poppy)	A2	D2
SAL character (Spike)	A3	D3
SAL character (Prudence)	A4	D4

4.4. Phase III

Because the personal language and linguistic style are remarkably reliable across time and situations (J. Pennebaker, King, & Diener, 1999). LIWC, as a word account tool, could analyze the difference between participants' vocabulary usages during the interactions with SAL characters. Moreover, LIWC's subjective dictionaries were

independently rated by psychologist. In this way, LIWC provided a wide range of reliable dimensions to assess the user's personality or psychological state in the text. The linguistic preferences of participants, who speak with SAL characters, could be concluded. By comparing the emotional impacts of virtual characters on the linguistic style of users, the connection between the Semaine annotation and the linguistic preferences of participants could be established.

4.4.1. Investigate the emotions/characters impacts in LIWC

Table 20. Experiments plan for Section 4.4.1. The letter indicates the modality/database type. The number indicate the character Type.
E.g. B1 means the modality/database B (LIWC) and the character 1 (Obadiah)

Modalities		SAL character (Obadiah)	SAL character (Poppy)	SAL character (Spike)	SAL character (Prudence)
Outputs of Automatic Tools	LIWC (Text)	B1	B2	B3	B4

Table 20 displays the group codes in phase III. According to the LIWC results, Phase III repeated the steps in the Phase II to conduct the functionals analysis and the ANOVA. These analysis were also used to explore how the characters influence the user emotions in the text as measured by the LIWC, and whether the expressed emotions in the text among groups were significantly different from each other.

4.4.2. Investigate the dimensional correlation between LIWC and Semaine annotation

Table 21. The common dimensions between the annotations of the LIWC and the Semaine database

Dimensions	LIWC	Semaine database
Basic emotions	Affect	
	Posemo	Happiness
	Negemo	
	Anx	
	Anger	Anger
	Sad	Sadness
		Disgust
		Amusement
		Contempt

The LIWC output file consists of 80 word categories/dimensions (section 3.2.3). Only WC, WPS, and affective process were selected (table 11). Because the WC and WPS can reflect the social status of the participants. The affective process has the direct connection with the basic emotions. It is easy for LIWC to associate with Semaine annotation. Then, the common dimensions between the annotations of the LIWC and the Semaine database is listed in table 21. The correlation comparison of Phase III was based on the common dimensions in the table 21. According to the experiments plan

(table 22), the correlation coefficients for each character between the annotation of Semaine database and the measurement of LIWC were calculated respectively. Pearson correlation calculator was still used.

Table 22. Experiments plan for Section 4.4.2. The letter indicates the modality/database type. The number indicate the character Type.
E.g. B1 means the modality/database B (LIWC) and the character 1 (Obadiah)

Characters	Automatic tools	Manually annotations
	LIWC (Text)	Semaine (Multiple modalities)
SAL character (Obadiah)	B1	D1
SAL character (Poppy)	B2	D2
SAL character (Spike)	B3	D3
SAL character (Prudence)	B4	D4

4.5. Participants and samples

Table 23. Group information of participants. (a)

Session ID	Recording ID	Character	User,Operator	Gender
5	1	Obadiah	16,2	Female,Male
8	2	Obadiah	16,2	Female,Male
15	3	Obadiah	5,2	Female,Male
19	4	Obadiah	3,2	Male,Male
27	5	Obadiah	4,2	Male,Male
30	6	Obadiah	16,3	Female,Male
49	11	Obadiah	2,3	Male,Male
54	12	Obadiah	8,3	Female,Male
61	13	Obadiah	9,3	Female,Male
103	20	Obadiah	15,19	Female,Female
107	21	Obadiah	17,16	Male,Female
115	22	Obadiah	18,1	Female,Female
121	23	Obadiah	19,16	Male,Female
2	1	Poppy	16,2	Female,Male
11	2	Poppy	16,2	Female,Male
16	3	Poppy	5,2	Female,Male
21	4	Poppy	3,2	Male,Male
26	5	Poppy	4,2	Male,Male
29	6	Poppy	16,3	Female,Male
48	11	Poppy	2,3	Male,Male

Table 23. Group information of participants. (b)

Session ID	Recording ID	Character	User,Operator	Gender
55	12	Poppy	8,3	Female,Male
60	13	Poppy	9,3	Female,Male
100	20	Poppy	15,16	Female,Female
108	21	Poppy	17,16	Male,Female
112	22	Poppy	18,1	Female,Female
118	23	Poppy	19,16	Male,Female
122	23	Poppy	19,16	Male,Female
4	1	Prudence	16,2	Female,Male
10	2	Prudence	16,2	Female,Male
14	3	Prudence	5,2	Female,Male
22	4	Prudence	3,2	Male,Male
31	6	Prudence	16,3	Female,Male
46	11	Prudence	2,3	Male,Male
53	12	Prudence	8,3	Female,Male
58	13	Prudence	9,3	Female,Male
102	20	Prudence	15,18	Female,Female
106	21	Prudence	17,16	Male,Female
114	22	Prudence	18,1	Female,Female
119	23	Prudence	19,16	Male,Female
3	1	Spike	16,2	Female,Male
9	2	Spike	16,2	Female,Male
13	3	Spike	5,2	Female,Male
20	4	Spike	3,2	Male,Male
25	5	Spike	4,2	Male,Male
47	11	Spike	2,3	Male,Male
52	12	Spike	8,3	Female,Male
59	13	Spike	9,3	Female,Male
101	20	Spike	15,17	Female,Female
109	21	Spike	17,16	Male,Female
113	22	Spike	18,1	Female,Female
120	23	Spike	19,16	Male,Female

According to section 3.1, the recordings of Solid SAL were the targeted input data. There were no extra requirements about gender, age, nationality, and etc. But the Feeltraced annotations, frontal videos of users, and transcripts were the three fundamental principles to filter the targeted input data. During the filtering, a fixed set of users was another necessary experiment requirement. It was helpful to extract the character/modality impact on the emotion of the same user. Eventually, the user group included 11 participants (user IDs: 2, 3, 4, 5, 8, 9, 15, 16, 17, 18, and 19). The complete information of experiment samples are displayed in the table 23. This table explains the session ID, recording ID, character type, user ID, Operator ID, and user gender of these selected samples.

5. Data analysis of Phase I

5.1. Emotion analysis of the Semaine annotation in the Phase I

5.1.1. Dimension usages of Semaine annotation in the Phase I

The usage distributions of Semaine full rating and optional dimensions for Obadiah and Poppy characters are illustrated in table 24. According to the table 24, the main difference between D1 and D2 groups is the distribution of optional dimensions.

Table 24. The usage distribution of Semaine full rating and optional dimensions. The D letter indicates the Semaine database. The number indicates the character type, 1 means Obadiah character, 2 means Poppy character. Bold numbers are referred in the text.

Group code		**D1 group (13 samples)**	**D2 group (14 samples)**
Character		**Obadiah**	**Poppy**
Full rating dimensions	**Valence**	13	14
	Arousal	13	14
	Power	13	14
	Anticipation	13	14
	Intensity	13	13
Optional dimensions	**Fear**	1	0
	Anger	3	1
	Happiness	3	**11**
	Sadness	**11**	**3**
	Disgust	5	0
	Contempt	2	1
	Amusement	**6**	12

Compared to D2 group (users interacting with Poppy character, table 11), the users of D1 group (users interacting with Obadiah character, table 11) were more often annotated for negative emotions including sadness, disgust, anger, contempt, and fear. For the users in the D2 group who interacting with Poppy character, the annotators more often chose to annotate positive emotions such as happiness and amusement. But there were some exceptions. The participants of D1 group had many expressions which were annotated as amusement (6 of 13 samples), and the users in group D2 expressed sadness several times in the Semaine annotation (3 of 14 samples).

5.1.2. Value variations in Semaine annotation of Phase I

The functionals of each dimension in the Semaine annotation are listed in table 25. The functionals have the maximum, the minimum, the mean, the value range and the

standard deviation. The value range is the value gap between the maximum and the minimum.

Table 25. The functionals of D1 and D2 groups in Semaine annotation (a). Bold numbers are referred in the text.

	Max		Min		Mean	
Character	**Obadiah**	**Poppy**	**Obadiah**	**Poppy**	**Obadiah**	**Poppy**
Valence	0.6800	0.6825	-0.7950	-0.3988	-0.1813	**0.2728**
Arousal	0.6698	0.7490	-0.6493	-0.5375	-0.2922	**-0.0056**
Power	0.8282	0.8411	-0.4499	-0.8768	0.1956	**0.4702**
Anticipation	72.6230	87.4490	0.0000	0.0000	35.5054	29.0129
Intensity	0.5279	0.5409	-0.8819	-1.0000	-0.1978	**-0.1856**
Fear	0.5138		-1.0000		-0.6786	
Anger	0.1713	0.3911	-1.0000	-1.0000	**-0.7571**	-0.8060
Happiness	0.2480	0.4821	-1.0000	-0.9614	-0.6156	**-0.1695**
Sadness	0.5029	0.0026	-1.0000	-1.0000	**-0.2913**	-0.8770
Disgust	0.6518		-1.0000		-0.8316	
Contempt	0.7439	0.4576	-1.0000	-1.0000	**-0.2697**	-0.7353
Amusement	1.0000	0.6424	-1.0000	-1.0000	-0.6908	**-0.4866**

Table 25. The functionals of D1 and D2 groups in Semaine annotation (b). Bold numbers are referred in the text.

	Value range		Standard Deviation (SD)	
Character	**Obadiah**	**Poppy**	**Obadiah**	**Poppy**
Valence	1.4750	1.0813	0.1576	0.1138
Arousal	1.3191	1.2865	0.1147	0.1236
Power	1.2781	1.7179	0.1656	0.1639
Anticipation	72.6230	87.4490	7.9275	6.7655
Intensity	1.4098	1.5409	0.1265	0.1318
Fear	1.5138		0.4827	
Anger	1.1713	1.3911	**0.2820**	0.3510
Happiness	1.2480	1.4435	0.2399	**0.1991**
Sadness	1.5029	1.0026	**0.2240**	0.1441
Disgust	1.6518		0.2372	
Contempt	1.7439	1.4576	**0.4880**	0.4015
Amusement	2.0000	1.6424	0.2642	**0.2439**

In terms of full rating dimensions, the users of D2 group had a larger value range than the users of D1 group in the dimensions of Power, Anticipation, and Intensity. The users of D2 group had lower standard deviations than the users of D1 group in the dimensions of Valence, Power, and Anticipation. Moreover, the users of D2 group scored higher means in most full rating dimensions (except Anticipation). Compared to the users of D1 group, the users of D2 group scored higher in the dimensions of Semaine annotation in the Valence (MValence = 0.2728), the Arousal (MArosal = -0.0056), the Power (MPower = 0.4702), and the Intensity (MIntensity = -0.1856).

In terms of the optional dimensions, the situation was slightly different. Because the emotions of the Fear and the Disgust were not found in the annotation of the D2 group. Compare to the users of D1 group, the users of D2 group had some differences in the Semaine annotation. For example, the Anger and the Happiness had larger value ranges in the D2 group, and the Happiness (MHappiness = -0.1695) as well as the Amusement (MAmusement = -0.4866) had higher means in the D2 group. They had less amount of variation or dispersion (SDHappiness= 0.1991, SDAmusement= 0.2439). The Anger (MAnger = -0.7571), the Sadness (MSadness = -0.2913) and the Contempt (MContempt = -0.2697) had higher means in the D1 group. The Anger (SDAnger = 0.2820) and the Contempt (SDContempt = 0.488) in the D1 group were less randomly distributed. But the Sadness of D1 group were more randomly distributed (SDSadness= 0.224) in a larger value range.

According to the analysis of the Semaine annotation, the optional dimensions were consistent with the full rating traces. Because the Valence scored higher in the D2 group which had more expressions of positive emotions. The higher means of the Arousal and the Power in the D2 group related to fewer negative emotional expressions and fewer emotion types. Then, with lower means in the Arousal and the Power, the users of D1 group had more negative emotional expressions and more negative emotions in the Semaine annotation.

5.1.3. The significant analysis in Semaine annotation of Phase I

The sample size was small and each group was less than 30 samples. They cannot represent the distribution of the whole population. Meanwhile, two characters were independent of each other. Therefore, the independent two samples T-test was used to determine whether D1 and D2 groups had significant difference in the dimensions of Semaine annotation. The annotated means of each dimension were used in the test. There were three main steps for the test. First step was the data normalization. The second step was the normality test. Third step was the T-test calculation including Levene's Test for equality of variances. Because the normality and the variance equality were the two assumptions for T-test.

Step1: Among the database, there were some cases that one participant generated multiple sessions by interacting with the same character. The repeated sessions were different from each other (different selections of optional dimensions and variances of parametric values). The Semaine annotation of the repeated sessions should be

Table 26. Means of D1 group.

Sample ID	User Id	Valence	Arousal	Power	Anticipation	Intensity	Fear	Anger	Happiness	Sadness	Disgust	Contempt	Amusement
5	**16**	0.088	0.122	0.446	42.065	-0.055				-0.733	-0.594		-0.664
8	**16**	-0.143	-0.044	0.406	33.137	-0.055		-0.759	-0.689	-0.588		-0.102	
15	**5**	-0.233	-0.386	-0.153	40.329	-0.047			-0.789	-0.291			-0.746
19	**3**	-0.062	-0.401	0.113	33.190	-0.185			-0.369	0.126			-0.362
27	**4**	-0.210	-0.372	-0.005	36.648	-0.192				-0.233			-0.553
30	**16**	-0.394	-0.357	0.004	24.110	-0.031				-0.092	-0.863	-0.438	
49	**2**	-0.612	-0.501	0.072	30.935	0.088	-0.679	-0.798		-0.001	-0.887		
54	**8**	-0.327	-0.098	0.059	31.597	-0.024		-0.714		-0.283	-0.905		
61	**9**	-0.313	-0.371	0.097	38.485	-0.179				-0.150			
103	**15**	-0.051	-0.296	0.321	41.278	-0.281				-0.518			-0.890
107	**17**	-0.273	-0.354	0.265	40.240	-0.225				-0.442	-0.909		
115	**18**	0.179	-0.341	0.418	41.828	-0.588							-0.929
121	**19**	-0.005	-0.401	0.501	27.728	-0.796							

Table 27. Means of D2 group.

Sample ID	User Id	Valence	Arousal	Power	Anticipation	Intensity	Fear	Anger	Happiness	Sadness	Disgust	Contempt	Amusement
2	**16**	0.312	-0.006	0.643	32.064	-0.007			-0.393				-0.414
11	**16**	0.251	0.259	0.564	23.137	0.053			-0.342				-0.545
16	**5**	0.405	0.083	0.264	37.482	0.061			0.137				-0.134
21	**3**	0.348	0.042	0.498	25.932	-0.090			0.010				-0.248
26	**4**	0.317	0.035	0.431	26.111	-0.102			0.013				-0.284
29	**16**	0.151	-0.306	0.460	27.179	-0.147			-0.315	-0.721			-0.588
48	**2**	0.083	0.155	0.566	32.323	-0.011		-0.806	-0.464			-0.735	-0.508
55	**8**	0.346	0.116	0.437	24.624	-0.152			-0.184	-0.990			-0.586
60	**9**	0.300	-0.002	0.528	26.871	-0.166			-0.071				-0.306
100	**15**	0.275	-0.037	0.316	37.333	-0.200			-0.225				-0.659
108	**17**	0.274	-0.087	0.487	25.414	-0.176			-0.031	-0.919			
112	**18**	0.381	0.113	0.376	38.912								
118	**19**	0.204	-0.259	0.549	23.922	-0.880							-0.920
122	**19**	0.171	-0.186	0.462	24.878	-0.596							-0.648

Table 28. The means of the merged sessions in the Semaine annotation of the D1 group

Sample ID	User Id	Valence	Arousal	Power	Anticipation	Intensity	Fear	Anger	Happiness	Sadness	Disgust	Contempt	Amusement
49	**2**	-0.612	-0.501	0.072	30.935	0.088	-0.679	-0.798		-0.001	-0.887		
19	**3**	-0.062	-0.401	0.113	33.190	-0.185			-0.369	0.126			-0.362
27	**4**	-0.210	-0.372	-0.005	36.648	-0.192				-0.233			-0.553
15	**5**	-0.233	-0.386	-0.153	40.329	-0.047			-0.789	-0.291			-0.746
54	**8**	-0.327	-0.098	0.059	31.597	-0.024		-0.714		-0.283	-0.905		
61	**9**	-0.313	-0.371	0.097	38.485	-0.179				-0.150			
103	**15**	-0.051	-0.296	0.321	41.278	-0.281				-0.518			-0.890
5&8&30	**16**	-0.150	-0.093	0.285	33.104	-0.047		-0.759	-0.689	-0.471	-0.728	-0.270	-0.664
107	**17**	-0.273	-0.354	0.265	40.240	-0.225				-0.442	-0.909		
115	**18**	0.179	-0.341	0.418	41.828	-0.588							-0.929
121	**19**	-0.005	-0.401	0.501	27.728	-0.796							

Table 29. The means of the merged sessions in the Semaine annotation of the D2 group

Samples ID	User Id	Valence	Arousal	Power	Anticipation	Intensity	Fear	Anger	Happiness	Sadness	Disgust	Contempt	Amusement
48	**2**	0.083	0.155	0.566	32.323	-0.011		-0.806	-0.464			-0.735	-0.508
21	**3**	0.348	0.042	0.498	25.932	-0.090			0.010				-0.248
26	**4**	0.317	0.035	0.431	26.111	-0.102			0.013				-0.284
16	**5**	0.405	0.083	0.264	37.482	0.061			0.137				-0.134
55	**8**	0.346	0.116	0.437	24.624	-0.152			-0.184	-0.990			-0.586
60	**9**	0.300	-0.002	0.528	26.871	-0.166			-0.071				-0.306
100	**15**	0.275	-0.037	0.316	37.333	-0.200			-0.225				-0.659
2&11&29	**16**	0.238	-0.017	0.556	27.460	-0.034			-0.350	-0.721			-0.516
108	**17**	0.274	-0.087	0.487	25.414	-0.176			-0.031	-0.919			
112	**18**	0.381	0.113	0.376	38.912								
118&122	**19**	0.188	-0.222	0.506	24.400	-0.738							-0.784

merged. For example, user 16 had three samples (5, 8 and 30) in D1 group (table 26), and three samples (2, 11 and 29) in D2 group (table 27) as well. The merged result was the average value of the annotations for the repeated sessions. During the merging

process, the empty results were excluded from the calculation. The example was the Happiness of user 16 in D1 group. The final merged results are listed in table 28 and 29.

Step 2: These merged values were used for the Normality test (significance level = 0.05). The Shapiro-Wilk test suits the small size of samples. The test results are listed in table 30. Most dimensions passed the Test. Only the Arousal (in D1 group, $p = 0.019 < 0.05$), the Anticipation (in D2 group, $p = 0.01 < 0.05$) and the Intensity (in D2 group, $p = 0.002 < 0.05$) failed the normality test. The Fear, the Contempt, the Anger, and the Disgust had too few samples to complete the test. In the figure 10, the normal Q-Q plots of these dimensions still fitted in the linear model. They revealed no serious threats to the assumption of distribution normality. In this way, these tested dimensions would continue the following process.

Table 30. The results of normality test. Bold numbers are referred in the text.

Group code	D1			D2		
	Statistic	df	Sig.	Statistic	df	Sig.
Valence	0.970	11	0.890	0.936	11	0.471
Arousal	0.823	11	**0.019**	0.940	11	0.517
Power	0.971	11	0.897	0.925	11	0.359
Anticipation	0.919	11	0.310	0.802	11	**0.010**
Intensity	0.868	11	0.073	0.734	10	**0.002**
Fear						
Anger	0.999	3	0.928			
Happiness	0.917	3	0.441	0.957	9	0.767
Sadness	0.946	9	0.650	0.931	3	0.491
Disgust	0.721	4	0.020			
Contempt						
Amusement	0.970	11	0.890	0.958	9	0.779

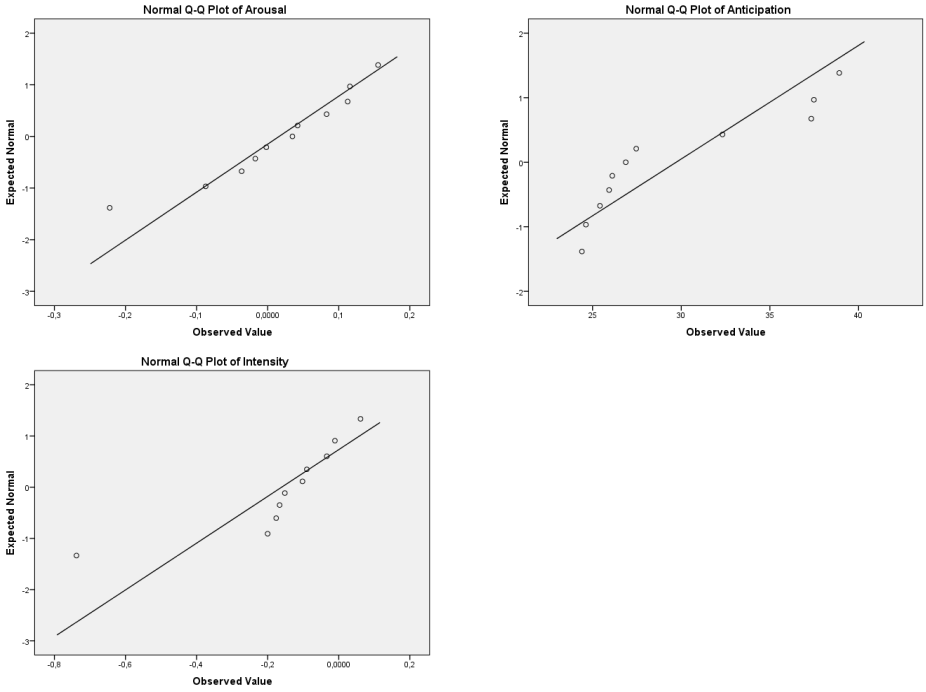

Figure 10. Normal Q-Q Plots of Arousal(D1), Anticipation(D2), and Intensity(D2).

Table 31. T-test results of Annotations between D1 and D2 groups. Bold numbers are referred in the text.

		Levene's Test for Equality of Variances		T-test for Equality of Means							Effect size
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference		
									Lower	Upper	
Valence	Equal variances assumed	3.986	0.060	-6.929	20.000	**0.000**	-0.474	0.068	-0.616	-0.331	6.929
	Equal variances not assumed			-6.929	13.821	0.000	-0.474	0.068	-0.621	-0.327	6.929
Arousal	Equal variances assumed	0.088	0.769	-6.915	20.000	**0.000**	-0.345	0.050	-0.449	-0.241	6.915
	Equal variances not assumed			-6.915	19.545	0.000	-0.345	0.050	-0.449	-0.241	6.915
Power	Equal variances assumed	6.518	**0.019**	-4.136	20.000	0.001	-0.272	0.066	-0.409	-0.135	4.136
	Equal variances not assumed			-4.136	14.743	**0.001**	-0.272	0.066	-0.412	-0.132	4.136
Anticipation	Equal variances assumed	0.631	0.436	2.766	20.000	**0.012**	6.227	2.252	1.530	10.924	2.766
	Equal variances not assumed			2.766	19.514	0.012	6.227	2.252	1.523	10.932	2.766
Intensity	Equal variances assumed	0.472	0.500	-0.613	19.000	0.547	-0.064	0.105	-0.284	0.155	0.613
	Equal variances not assumed			-0.618	18.920	0.544	-0.064	0.104	-0.283	0.154	0.618
Fear											
Anger	Equal variances assumed			1.007	2.000	0.420	0.049	0.049	-0.160	0.258	1.007
	Equal variances not assumed						0.049				
Happiness	Equal variances assumed	0.008	0.931	-3.865	11.000	**0.003**	-0.448	0.116	-0.703	-0.193	3.865
	Equal variances not assumed			-3.848	5.769	0.009	-0.448	0.116	-0.736	-0.160	3.848
Sadness	Equal variances assumed	0.662	0.435	4.617	10.000	**0.001**	0.626	0.135	0.324	0.927	4.617
	Equal variances not assumed			5.787	5.589	0.001	0.626	0.108	0.356	0.895	5.787
Contempt	Equal variances assumed			0.577	1.000	0.667	0.233	0.403	-4.891	5.356	0.577
	Equal variances not assumed						0.233				
Amusement	Equal variances assumed	0.099	0.758	-2.157	13.000	**0.050**	-0.244	0.113	-0.488	0.000	2.157
	Equal variances not assumed			-2.161	10.940	0.054	-0.244	0.113	-0.492	0.005	2.161

Step 3: The study of character impacts is based on the same user. Therefore, table 28 and table 29 have the same set of users. According to these two tables, the independent two samples T-test was calculated with significance level = 0.05.

According to the Levene's test in the table 31, only Power dimension violated the homogeneity of variance (sig = 0.019 < 0.05). For the Power dimension, the values of equal variances not assumed should be used. The Valence ($t(20) = 6.929$, $p = 0.000001 < 0.05$), the Arousal ($t(20) = 6.915$, $p = 0.000001 < 0.05$), the Power ($t(15) = 4.136$, $p = 0.001 < 0.05$), the Anticipation ($t(20) = 2.766$, $p = 0.012 < 0.05$), the Happiness ($t(11) = 3.865$, $p = 0.003 < 0.05$), and the Sadness ($t(10) = 4.617$, $p = 0.001 < 0.05$) indicated that they had significantly different means between D1 and D2 groups. The Amusement was also considered to have a significant difference ($t(13) = 2.157$, $p = 0.05 = 0.05$).

The effect sizes described the extent to the characters influence values of each dimension. The Valence ($d = 6.929$) and the Arousal ($d = 6.915$) had large effect sizes. Meanwhile, the Power ($d = 4.136$), the Happiness ($d = 3.865$) and the Sadness ($d = 4.617$) had medium ones. But the Intensity ($t(19) = 0.613$, $p = 0.547 > 0.05$) indicated that it did not have significant difference of means between D1 and D2 groups. Because Intensity described how far the emotion was away from the pure state or cool rationality. It relates to both positive and negative emotions. Although the users of D1 group shown less positive emotions (the Happiness), they had strong negative emotional expressions (Anger and Sadness) in the Semaine annotation. Due to the lack of samples, the Fear, the Anger, and the Contempt were unable to fully conduct the Levene's test or the T-test.

5.1.4. Summary

Overall, the character impacts were displayed in the Semaine annotation of participants' sessions. The users in D1 group had more annotations in the Sadness and negative emotions. Meanwhile, the people, who spoke with Poppy character, had more expressions which were annotated in the Happiness. In terms of full rating dimensions, Obadiah character enabled participants to feel less positively than the users of D2 group. In terms of the Semaine annotation, the users of D1 group had more emotion types such as Fear, Contempt, etc. According to above analysis, it was found that sad emotion was frequently annotated in D1 group, and happy emotion was well annotated in D2 group. Moreover, the T-test shown that the Sadness and the Happiness had significantly different distributions between these two groups. This emotion difference between characters was also supported by the analysis of full rating dimensions. The Valence and the Arousal positively influenced the positive emotions. The Power negatively associated with the diversity of emotions.

5.2. FaceReader of Phase I

The automatic recognition tool can help to record how users express emotions in each modality. Facial expression is the most explicit modality for detection, and the results of FaceReader was chosen for the comparison. FaceReader uses the basic emotional dimensions and full rating dimensions to annotate the expressed emotion of users. The basic emotional dimensions include the Neutral, the Happy, the Sad, the Angry, the Surprised, the Scared and the Disgusted. The full rating dimensions contain the Valence and the Arousal.

5.2.1. Value variations in FaceReader of Phase I

Table 32. The functionals of FaceReader (a). Bold numbers are referred in the text.

	Max		**Min**		**Mean**	
Group code	**A1**	**A2**	**A1**	**A2**	**A1**	**A2**
Character	**Obadiha**	**Poppy**	**Obadiha**	**Poppy**	**Obadiha**	**Poppy**
Neutral	0.994	0.993	0.000	0.000	**0.361**	0.275
Happy	0.999	0.999	0.000	0.000	0.153	**0.305**
Sad	0.988	0.993	0.000	0.000	**0.046**	0.043
Angry	0.994	0.992	0.000	0.000	**0.085**	0.043
Surprised	0.994	1.000	0.000	0.000	0.133	**0.148**
Scared	1.000	1.000	0.000	0.000	0.186	**0.210**
Disgusted	0.987	1.000	0.000	0.000	**0.045**	0.021
Valence	0.999	0.998	-0.998	-1.000	**-0.157**	**0.031**
Arousal	**0.899**	**0.955**	**0.020**	**0.036**	**0.310**	**0.366**

Table 32. The functionals of FaceReader (b). Bold numbers are referred in the text.

	Value Range		**Standard Deviation**	
Group code	**A1**	**A2**	**A1**	**A2**
Character	**Obadiha**	**Poppy**	**Obadiha**	**Poppy**
Neutral	0.994	0.992	**0.233**	0.200
Happy	0.999	0.999	0.222	**0.284**
Sad	0.988	0.993	**0.096**	0.088
Angry	0.994	0.992	**0.140**	0.102
Surprised	0.994	1.000	0.187	**0.207**
Scared	1.000	1.000	0.184	**0.170**
Disgusted	0.986	1.000	**0.074**	0.057
Valence	1.997	**1.998**	0.391	**0.415**
Arousal	0.879	**0.918**	0.111	**0.124**

According to table 14, A1 and A2 groups were analyzed. The functionals of each dimension are listed in table 32. The table 32 has the maximum value, minimum value, mean value, value range value and standard deviation.

In terms of full rating dimensions, the users of A2 group had larger value range than the ones of A1 group in the Valence and the Arousal of FaceReader. Particularly, the values of the value range in the Arousal had a significant positive shift from A1 group to A2 group. This trend went the same with the means in the Valence from A1 group (MValence= -0.1568 and MArousal= 0.3101) to A2 group (MValence= 0.0307 and MArousal= 0.3664). The users of A2 group also had higher standard deviations (SDValence= 0.4155 and SDArousal= 0.1242) in the Valence. These values displayed that most users of A2 group had more feelings which were annotated in the Valence and the Arousal. But there was a minority of users who did not have the same annotated feeling.

In terms of the basic emotions dimensions, all the means were positive values. Based on these functionals, the Scared scored higher with a larger standard deviation in the A2 group (MScared= 0.2103 and SDScared= 0.1698). The rest dimensions shared this consistency that higher means appeared with higher standard deviations. According to the values, these dimensions could be divided into two categories. The Neutral (SDNeutral = 0.2329), the Sad (SDSad = 0.0957), the Angry (SDAngry = 0.1401) and the Disgusted (SDDisgusted = 0.0741) were the first category which had higher means and standard deviations in A1 group. Meanwhile, the Happy and the Surprised were the second category which had higher means and standard deviations in A2 group.

According to the analysis, the full rating dimensions supported the emotional distribution. Because A2 group had higher Arousal and Valence. Higher values of the Valence reflected more positive feelings (for example Happy). The Arousal explained the less appearance of Neutral emotion in A2 group. The emotion impacts of characters also proved how SAL characters influence the emotion of participants. In the results of FaceReader, Poppy character enables users to express more facial feelings which were annotated as positive emotions (Happy), and Obadiah character enables the participants to display more facial expressions which were annotated as negative emotions (Sad, Angry, and Disgusted). But the Scared dimension of FaceReader was the only exception. Because the Poppy character had more emotion impacts on the scared emotion.

5.2.2. The significance difference in FaceReader of Phase I

Similarly, independent two samples T-test was used to testify whether A1 and A2 groups had any significant difference in the results of FaceReader. The test also had three steps - data normalization, normality test, and T-test calculation with Levene's Test for equality of variances.

Step 1: The samples of the repeated sessions should be merged together by re-calculating the means. The merged results are listed in table 33 and table 34.

Table 33. The means of the merged sessions in the FaceReader results of A1 group.

Sample ID	Subject /User Id	Neutral	Happy	Sad	Angry	Surprised	Scared	Disgusted	Valence	Arousal
49	**2**	0.410	0.057	0.001	0.129	0.016	0.002	0.195	-0.221	0.367
19	**3**	0.337	0.219	0.036	0.016	0.077	0.229	0.059	-0.046	0.332
27	**4**	0.314	0.448	0.002	0.338	0.066	0.001	0.001	0.108	0.346
15	**5**	0.318	0.082	0.056	0.015	0.137	0.328	0.078	-0.290	0.294
54	**8**	0.443	0.096	0.000	0.220	0.021	0.001	0.077	-0.172	0.338
61	**9**	0.399	0.070	0.177	0.001	0.062	0.223	0.005	-0.283	0.318
103	**15**	0.439	0.252	0.104	0.025	0.155	0.108	0.005	0.045	0.301
5&8&30	**16**	0.192	0.092	0.072	0.086	0.087	0.500	0.053	-0.512	0.297
107	**17**	0.499	0.102	0.003	0.056	0.321	0.008	0.002	0.036	0.284
115	**18**	0.395	0.252	0.005	0.030	0.320	0.009	0.002	0.209	0.280
121	**19**	0.561	0.138	0.004	0.013	0.289	0.008	0.001	0.113	0.279

Table 34. The means of the merged sessions in the FaceReader results of the A2 group.

Sample ID	Subject /User Id	Neutral	Happy	Sad	Angry	Surprised	Scared	Disgusted	Valence	Arousal
48	**2**	0.339	0.292	0.000	0.092	0.143	0.000	0.072	0.152	0.420
21	**3**	0.399	0.377	0.020	0.015	0.017	0.141	0.016	0.208	0.358
26	**4**	0.225	0.567	0.018	0.143	0.065	0.113	0.002	0.308	0.404
16	**5**	0.210	0.116	0.133	0.004	0.075	0.466	0.024	-0.392	0.318
55	**8**	0.399	0.204	0.000	0.136	0.130	0.000	0.061	0.033	0.375
60	**9**	0.350	0.345	0.019	0.014	0.192	0.060	0.013	0.252	0.360
100	**15**	0.263	0.451	0.009	0.027	0.223	0.033	0.018	0.374	0.404
2&11&29	**16**	0.091	0.095	0.124	0.032	0.056	0.681	0.009	-0.641	0.356
108	**17**	0.320	0.341	0.005	0.032	0.318	0.024	0.014	0.274	0.376
112	**18**	0.256	0.532	0.009	0.014	0.228	0.023	0.013	0.479	0.361
118&122	**19**	0.405	0.379	0.006	0.014	0.259	0.020	0.013	0.333	0.344

Step 2: These merged values were used for the normality test with significance level = 0.05. The results are displayed in table 35. The Valence (A2 group), the Happy (A1 group), the Sad, the Angry, the Surprised (A1 group), the Scared, and the Disgust failed the normality test. In the figure 11, the normal Q-Q plots of the Sad, the Angry, the Scared and the Disgust had some wrong value points which were unfit with linear model. They revealed serious threats to the assumptions of normal distribution. In this

way, these dimensions would stop the following process and participate the non-parameter test - Mann-Whitney U test. The Valence, the Happy and the Surprised still matched with linear model and continue the T-test.

Table 35. The results of normality test in FaceReader.
Bold numbers are referred in the text.

Group code	A1			A2		
	Statistic	df	Sig.	Statistic	df	Sig.
Valence	0.949	11	0.630	0.812	11	**0.014**
Arousal	0.918	11	0.304	0.956	11	0.724
Neutral	0.975	11	0.928	0.921	11	0.328
Happy	0.818	11	**0.016**	0.955	11	0.703
Sad	0.777	11	**0.005**	0.622	11	**0.000**
Angry	0.767	11	**0.004**	0.746	11	**0.002**
Surprised	0.845	11	**0.036**	0.958	11	0.745
Scared	0.787	11	**0.006**	0.672	11	**0.000**
Disgust	0.743	11	**0.002**	0.723	11	**0.001**

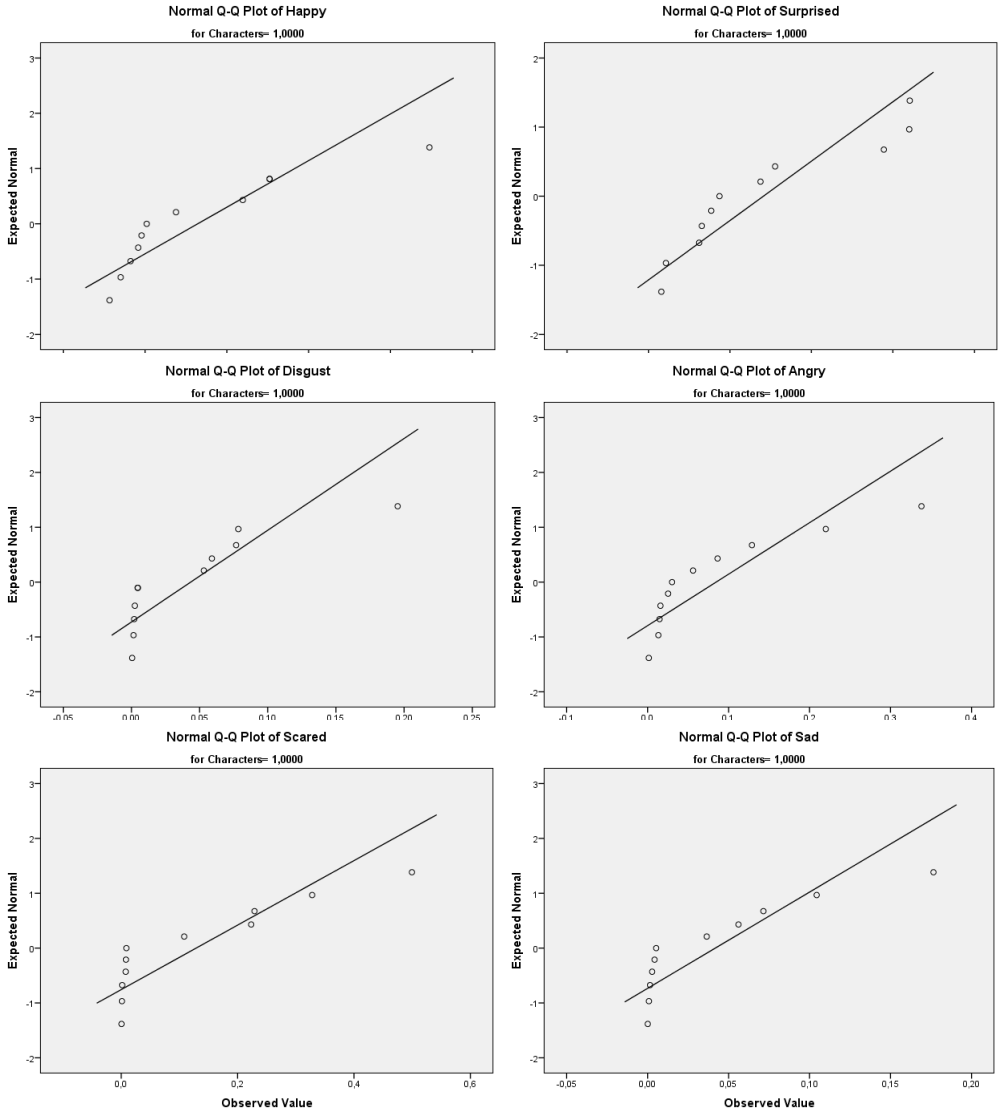

Figure 11. Normal Q-Q Plots of Failed dimensions (A1) in FaceReader (a).

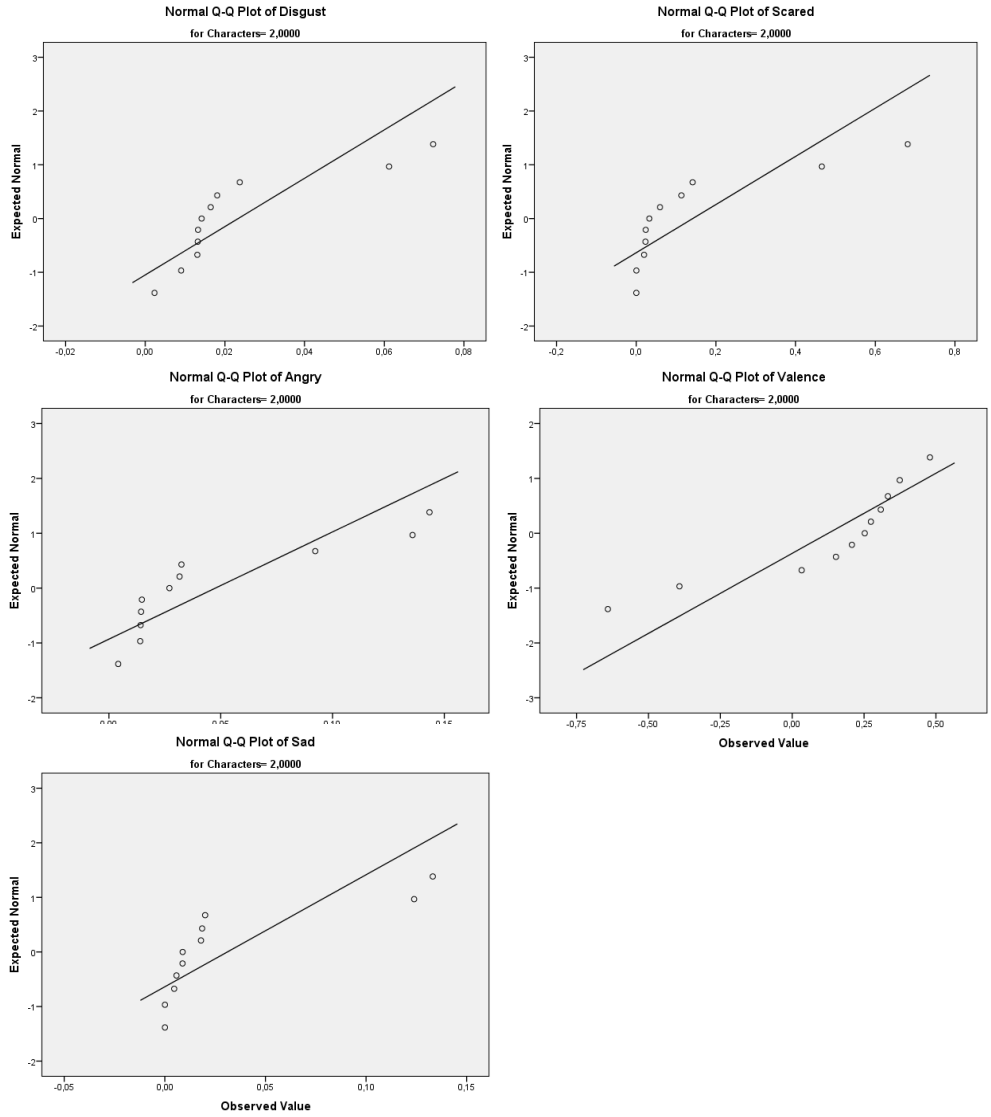

Figure 11. Normal Q-Q Plots of Failed dimensions (A2) in FaceReader (b).

Step 3: According to tables 33, 34 and 35, the independent two samples T-test was calculated with significance level = 0.05. The results are listed in table 36. According to the results, the Levene's test for equality of variances indicated that no tested dimension had significant violation of the assumption. Therefore, all the values of equal variances assumed should be used. After the T-test, the Arousal ($t(20) = 4.567$, $p = 0.0002 < 0.05$), the Neutral ($t(20) = 2.263$, $p = 0.035 < 0.05$), and the Happy ($t(18.8) = 2.947$, $p = 0.008 < 0.05$) indicated that they had significant different means between A1 and A2 groups in the results of FaceReader. In terms of the effect size, the Arousal ($d = 4.567$) had a medium one, and the Netural ($d = 2.263$) and the Happy ($d = 2.947$) had small ones. The results of non-parameter test are listed in the table 37. These dimensions were not significantly different across the characters according to the results of significance in table 37.

Compared to Semaine annotation, the results of FaceReader was less accurate. Because the Semaine annotation has more dimensions which indicate the significant

difference between groups. But the FaceReader is capable to analyze more types of emotions. For example, both D2 and A2 groups interacted with the Poppy character. Fear and Disgust emotions were not annotated in D2 group. But FaceReader could detect the Scared and Disgusted emotions for each sample of A2 group.

Table 36. T-test results of FaceReader between A1 and A2 groups. Bold numbers are referred in the text.

		Levene's Test for Equality of Variances		T-test for Equality of Means							Effect size
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference		
									Lower	Upper	
Valence	Equal variances assumed	0.792	0.384	-1.772	20.000	0.092	-0.218	0.123	-0.474	0.039	1.772
	Equal variances not assumed			-1.772	17.065	0.094	-0.218	0.123	-0.477	0.041	1.772
Arousal	Equal variances assumed	0.121	0.731	-4.567	20.000	**0.000**	-0.058	0.013	-0.084	-0.031	4.567
	Equal variances not assumed			-4.567	19.999	0.000	-0.058	0.013	-0.084	-0.031	4.567
Neutral	Equal variances assumed	0.051	0.823	2.263	20.000	**0.035**	0.095	0.042	0.007	0.183	2.263
	Equal variances not assumed			2.263	19.992	0.035	0.095	0.042	0.007	0.183	2.263
Happy	Equal variances assumed	0.425	0.522	-2.947	20.000	**0.008**	-0.172	0.058	-0.294	-0.050	2.947
	Equal variances not assumed			-2.947	18.821	0.008	-0.172	0.058	-0.294	-0.050	2.947
Surprised	Equal variances assumed	0.366	0.552	-0.308	20.000	0.761	-0.014	0.046	-0.109	0.081	0.308
	Equal variances not assumed			-0.308	19.330	0.761	-0.014	0.046	-0.109	0.081	0.308

Table 37. Mann-Whitney U test results of FaceReader between A1 and A2 groups

	Sad	Angry	Scared	Disgust
Mann-Whitney U	60.000	52.000	56.000	54.000
Wilcoxon W	126.000	118.000	122.000	120.000
Z	-0.033	-0.558	-0.295	-0.427
Asymp. Sig. (2-tailed)	0.974	0.577	0.768	0.669
Exact Sig. [2*(1-tailed Sig.)]	1.000_b	0.606_b	0.797_b	0.699_b

5.2.3. Correlations in FaceReader of Phase I

According to section 4.2.3, the common dimensions between FaceReader and Semaine database is illustrated in table 17. The correlation was based on the mean of each common dimension. Because not every user had the annotations in all optional dimensions of Semaine database. Therefore, the correlation calculation was based on

the actual data of Semaine annotation and FaceReader results. For example, only three users (8,16,17) of D2 group had values in the Sadness of Semaine annotation and FaceReader results. Then, only these users were considered in the later calculation of correlations in the Sadness between Semaine database and FaceReader results. According to the SAL character, the correlations between Semaine database and FaceReader were calculated respectively. By using Matlab function corr(), the final results are displayed in table 38 and table 40. In order to compare with the results of FaceReader, the correlations of Semaine annotation are also calculated as table 39 and table 41.

Table 38. The correlations of users interacting with Obadiah character between FaceReader and Semaine annotation. Bold numbers are referred in the text.

			FaceReader						
	ID	Dimension	Valence	Arousal	Happy	Sad	Angry	Scared	Disgusted
Semaine Database Annotations	2,3,4,5,8,9,15,16, 17,18,19	Valence	**0.526**	-0.689	**0.462**	-0.011	-0.349	0.063	-0.705
	2,3,4,5,8,9,15,16, 17,18,19	Arousal	-0.381	-0.164	-0.112	0.089	0.224	0.353	-0.167
	3,5,16	Happiness	**0.746**	0.987	**0.986**	-0.774	-0.275	-0.616	-0.505
	2,3,4,5,8,9,15,16, 17	Sadness	0.032	0.700	-0.003	-0.138	0.004	-0.141	0.487
	2,8,16	Anger	**0.170**	-0.373	0.894	-0.045	**0.694**	-0.038	-0.754
	2	Fear							
	2,8,16,17	Disgust	-0.907	-0.347	0.056	0.992	-0.326	0.993	-0.129

Table 39. Correlation within D1 group of Semaine annotation. Bold numbers are referred in the text.

Dimension	**ID**	**Valence**	**Arousal**	**Happiness**	**Sadness**	**Anger**	**Fear**	**Disgust**
Valence	2,3,4,5,8,9,15 ,16,17,18,19	**1.000**	0.149	**0.961**	-0.323	0.580		0.562
Arousal	2,3,4,5,8,9,15 ,16,17,18,19	0.149	1.000	-0.329	-0.571	0.841		0.497
Happiness	3,5,16	**0.961**	-0.329	**1.000**	0.863			
Sadness	2,3,4,5,8,9,15 ,16,17	-0.323	-0.571	0.863	1.000	-0.566		0.497
Anger	2,8,16	**0.580**	0.841		-0.566	**1.000**		-0.129
Fear	2							
Disgust	2,8,16,17	0.562	0.497		-0.436	-0.129		1.000

Compared to table 39, the results of D1 group were strongly as well as positively associated with the results of A1 group in Valence, Happiness (Happy), and Anger (Angry). Other negatively associated dimensions (in table 38) proved the variances

between FaceReader results and Semaine annotation. Among these three correctly correlated dimensions, only the Valence dimension of FaceReader results had reliably correlated with Semaine annotation in the Valence, the Happiness and the Anger. The Happy dimension reliably correlated with the Valence (r = 0.4615) and the Happiness (r = 0.9855) of the Semaine annotation. The Angry of FaceReader results reliably associated with the Anger (0.6944) of Semaine annotation. In the table 39, the correlation between the Anger and the Happiness still needed more evidence. In the Semaine annotation, the Happiness had a strong positive correlation with the Sadness. Combining with table 28, this correlation was true for such co-existence of multiple emotions like user 3, 5 and 16. In the FaceReader results, the Valence was the most reliable to detect the emotional expression. Although the FaceReader could reflect the features of the Happy and the Anger, it was incapable to distinguish the right relationship between the correctly detected emotions and other emotional dimensions.

Table 40. The correlations of users interacting with Poppy character between FaceReader and Semaine annotation. Bold numbers are referred in the text.

			FaceReader						
	ID	Dimension	Valence	Arousal	Happy	Sad	Angry	Scared	Disgusted
Semaine Database Annotations	2,3,4,5,8,9,15,16,17,18,19	Valence	-0.045	-0.507	0.067	0.262	-0.152	0.165	-0.389
	2,3,4,5,8,9,15,16,17,18,19	Arousal	-0.149	**0.255**	**-0.113**	0.072	0.393	0.026	0.560
	2,3,4,5,8,9,15,16,17	Happiness	0.117	**-0.579**	**0.190**	0.175	-0.242	0.033	-0.539
	8,16,17	Sadness	-0.871	**-0.964**	**-0.664**	**0.975**	-0.709	0.975	-0.764
	2	Anger							
		Fear							
		Disgust							

Table 41. Correlation within D2 group of Semaine annotation. Bold numbers are referred in the text.

Dimension	ID	Valence	Arousal	Happiness	Sadness	Anger	Fear	Disgust
Valence	2,3,4,5,8,9,15,16,17,18,19	1.000	0.236	0.860	-0.900			
Arousal	2,3,4,5,8,9,15,16,17,18,19	0.236	**1.000**	**-0.224**	-0.422			
Happiness	2,3,4,5,8,9,15,16,17	0.860	**-0.224**	**1.000**	-0.728			
Sadness	8,16,17	-0.900	**-0.422**	**-0.728**	**1.000**			
Anger	2							
Fear								
Disgust								

In the groups with Poppy character, FaceReader results and Semaine annotation were positively associated in dimensions of the Arousal, Happiness (Happy) and Sadness (Sad), and were only negatively associated in the dimension of Valence (table 40). These correctly correlated common dimensions also had right mutual relationships with each other. Compare to table 41, the Arousal of Semaine annotation correctly related to the Happy in the FaceReader. The Happiness of Semaine annotation negatively associated with FaceReader in the Arousal (r = -0.5794). The Sadness of Semaine annotation also associated negatively with FaceReader in the Arousal (r = -0.9643) and the Happy (r = -0.6635). In the A2 group, the Sadness of Semaine annotation were easy to be recognized by FaceReader, and the Arousal as well as the Happiness of FaceReader could reliably reflect the emotional information.

Based on the tables from 38 to 41, FaceReader was capable of distinguishing some emotions of users. But it also had many difference with Semaine annotation. Particularly, many users had multiple emotions with the character impacts. For example, the users of D1 group (user 3, 5, 16) and D2 group (user 8,16,17) expressed happy and sad feelings. Such fusion of emotions increased the difficulty of emotion recognition. This may explain why A1 and D1 groups were well associated in Happiness (Happy), as well as A2 and D2 groups were well correlated in Sadness (Sad). Because the Happiness of A1 group and the Sadness of A2 group were not aligned with the interacted characters, and the users of ambiguous emotions were filtered. According to these analysis, the happy emotion is the easiest to be detected by the emotional features across modalities and characters. In the results of FaceReader, the characters had no impacts on other emotions. For example, the Sadness was only effectively associated between the A2&D2 groups with Poppy character.

5.2.4. Summary

Compare to Semaine annotation, the FaceReader could detect the facial expressions and reflect the emotion impacts of characters. The A1 group had higher means in the Neutral, the Sad, the Angry and the Disgusted. Meanwhile, the A2 group had higher means in the Happy, the Scared, the Surprised and the full rating dimensions of the FaceReader. In the following T-test, the Arousal, the Neutral and the happy dimensions were proved to have statistically significant differences. Particularly, the happy emotion was widely effective for the emotional detection across characters (A1 and A2) and modalities (A1&D1 and A2&D2). Other emotions (e.g. the angry and the sad emotions) did not have the significant difference between groups. In terms of the emotion detection, they were limited by their interacted characters.

5.3. Conclusion of Phase I

According to the above analysis, the emotional impacts of characters are true. People would follow the emotional guidance. In the Semaine annotation and the FaceReader, the Poppy character enabled people to express more positive emotions which were annotated as happy emotion. Meanwhile, Obadiah character enabled

people to express more negative emotions which were usually annotated as sad, angry, and etc. But the FaceReader is not so reliable as the manual annotations in the Semaine database. In the perspective of statistics, the significant differences of the Valence and the Sadness were lost in the analysis of FaceReader.

The Happy (Happiness) and the Arousal are the most effective dimension across characters and modalities. Because they were best correlated between FaceReader and Semaine annotation. For other emotions, the annotation was still limited by the modalities. For example, the Sadness and the Valence were only distinguishable in the Semaine annotation. Because characters could influence the expressions of multiple emotions and the emotion fusion. Therefore, it is difficult for FaceReader to recognize the ambiguous emotion. Moreover, FaceReader only relies on the facial expression which may omit other information of speech, text, and etc.

6. Data analysis of Phase II

There are other two SAL characters which are different from Obadiah and Poppy characters. In the Phase II, other two characters - Spike and Prudence characters are added into the analysis. These two characters attempt to influence the emotion of users in a different way than Obadiah and Poppy characters, and may trigger different emotional reactions of the users.

6.1. Semaine annotation of Phase II

6.1.1. The dimensions distribution in Semaine annotation of Phase II

Table 42. The usage distribution of Semaine full rating and optional dimensions. The letter indicates the modality/database type. The number indicate the character Type. E.g. D1 means the modality database D (Semaine database) and the character 1 (Obadiah). Bold numbers are referred in the text.

Group code		**D1 (13 samples)**	**D2 (14 samples)**	**D3 (12 samples)**	**D4 (12 samples)**
Character		**Obadiah**	**Poppy**	**Spike**	**Prudence**
Full rating dimensions	**Valence**	13	14	12	12
	Arousal	13	14	12	12
	Power	13	14	12	12
	Anticipation	13	14	12	12
	Intensity	13	13	12	11
Optional dimensions	**Fear**	1	0	0	0
	Anger	3	1	**9**	2
	Happiness	3	**11**	1	7
	Sadness	**11**	3	1	3
	Disgust	5	0	1	0
	Contempt	2	1	**9**	1
	Amusement	6	**12**	10	10

The distribution of Semaine full rating and optional dimensions is illustrated in table 42. Full rating dimensions were well used in each group. But the appearances of optional dimensions varied with the interacted character. From table 42, only the Amusement is the most used dimension of the manual annotation among four groups. The users of D1 group frequently expressed the emotions of sadness and disgust. They also expressed other emotions like the Anger, the Contempt, the Fear, even the Happiness. Compared to other groups, the users of D1 group had expressed the largest number of emotion types, and had the highest frequency of negative emotions (the users of 11 samples expressed sadness). The users of D2 group had expressed the largest number of positive emotion types with the highest frequency (the users of 11 samples expressed happiness, the users of 12 samples expressed amusement). Additionally, the uses of D2 group also expressed a few negative emotions like sadness,

anger, and contempt. The users of D3 group, who interacting with Spike character, focused on the Anger, the Contempt, and the Amusement, and they rarely expressed other emotions. The users of D4 group, who interacting with Prudence character, also had expressed many positive emotional expressions like the users of D2 group. Meanwhile, the users of D1 and D3 groups had more similarities in the expressions of negative emotions.

According to above analysis, these characters had different impacts on the emotional expressions of the same set of users. This discovery basically supports the setting of SAL characters.

6.1.2. Variations in the Semaine annotation of Phase II

The functionals of each dimension of the Semaine annotation are listed in table 43. The functionals have the maximum, the minimum, the mean, the value range and the standard deviation. The value range is the value gap between the maximum and the minimum.

In terms of full rating dimensions, the users of D2 group had the highest means in the Valence (MValence = 0.2728) among four groups. Particularly, they also had the smallest standard deviation (SD = 0.1138) with the shortest value range. On the contrary, the users of D1 group had the lowest means in the Valence with the highest standard deviation and the longest value range. In addition, they had the lowest means with the least dispersion (SDArousal = 0.1147) in the Arousal, and the users of D2 group had the highest means ($M_{Arousal}$ = -0.0056). With similar standard deviations, the users of

Table 43. The functionals of Semaine annotation (a). Bold numbers are referred in the text.

	Max				**Min**			
Group code	**D1**	**D2**	**D3**	**D4**	**D1**	**D2**	**D3**	**D4**
Character	**Obadiah**	**Poppy**	**Spike**	**Prudence**	**Obadiah**	**Poppy**	**Spike**	**Prudence**
Valence	0.680	0.683	0.717	0.552	-0.795	-0.399	-0.726	-0.708
Arousal	0.670	0.749	0.658	0.644	-0.649	-0.537	-0.501	-0.587
Power	0.828	0.841	0.823	0.754	-0.450	-0.877	-0.711	-0.605
Anticipation	72.623	87.449	78.302	84.254	0.000	0.000	0.000	0.000
Intensity	0.528	0.541	0.867	0.399	-0.882	-1.000	-1.000	-0.872
Fear	0.514				-1.000			
Anger	0.171	0.391	0.727	0.386	-1.000	-1.000	-1.000	-1.000
Happiness	0.248	0.482	0.176	0.233	-1.000	-0.961	-0.667	-1.000
Sadness	0.503	0.003	0.335	0.023	-1.000	-1.000	-0.974	-1.000
Disgust	0.652		0.565		-1.000		-0.805	
Contempt	0.744	0.458	0.872	0.353	-1.000	-1.000	-1.000	-0.956
Amusement	1.000	0.642	1.000	0.897	-1.000	-1.000	-1.000	-1.000

Table 43. The functionals of Semaine annotation (b). Bold numbers are referred in the text.

	Value Range				Mean			
Group code	**D1**	**D2**	**D3**	**D4**	**D1**	**D2**	**D3**	**D4**
Character	**Obadiah**	**Poppy**	**Spike**	**Prudence**	**Obadiah**	**Poppy**	**Spike**	**Prudence**
Valence	1.475	1.081	1.443	1.260	**-0.181**	**0.273**	-0.095	0.067
Arousal	1.319	1.286	1.159	1.231	**-0.292**	**-0.006**	-0.010	-0.067
Power	1.278	1.718	1.535	1.359	**0.196**	**0.470**	0.310	0.379
Anticipation	72.623	87.449	78.302	84.254	35.505	**29.013**	**39.424**	33.633
Intensity	1.410	1.541	1.867	1.270	-0.198	-0.186	**-0.083**	**-0.238**
Fear	1.514				-0.679			
Anger	1.171	1.391	1.727	1.386	-0.757	**-0.806**	**-0.268**	-0.732
Happiness	1.248	1.444	0.844	1.233	**-0.616**	**-0.169**	-0.235	-0.526
Sadness	1.503	1.003	1.309	1.023	**-0.291**	**-0.877**	-0.656	-0.690
Disgust	1.652		1.370		**-0.832**		**-0.169**	
Contempt	1.744	1.458	1.872	1.309	**-0.180**	**-0.735**	-0.380	-0.339
Amusement	2.000	1.642	2.000	1.897	**-0.691**	**-0.487**	-0.528	-0.678

Table 43. The functionals of Semaine annotation (c). Bold numbers are referred in the text.

	Standard deviation			
Group code	**D1**	**D2**	**D3**	**D4**
Character	**Obadiah**	**Poppy**	**Spike**	**Prudence**
Valence	0.158	**0.114**	0.156	0.126
Arousal	**0.115**	0.124	0.161	0.133
Power	0.166	0.164	0.180	0.181
Anticipation	7.927	6.766	7.976	8.708
Intensity	0.126	0.132	**0.187**	**0.122**
Fear	0.483			
Anger	0.282	0.351	0.310	0.201
Happiness	0.240	**0.199**	0.316	0.237
Sadness	0.224	**0.144**	0.334	0.244
Disgust	0.237		0.322	
Contempt	0.488	0.402	0.348	0.450
Amusement	0.264	0.244	0.269	0.232

D2 group had the largest mean in the Power dimension, and the users of D1 group had the smallest one. In the Anticipation dimension, the users of D2 group had the smallest

mean. Meanwhile, the users of D3 group had the biggest mean in the Intensity with the smallest standard deviation ($M_{Intensity}$ = -0.083, $SD_{Intensity}$ = 0.1873). The users of D4 group had the lowest mean of the Intensity ($M_{Intensity}$ = -0.2381, $SD_{Intensity}$ = 0.1216) and the largest mean of the Anticipation.

Among the optional dimensions, the Fear was the special emotion with one appearance in D1 group (sample 49). With only one appearance, it is not enough to make any persuasive conclusion about the Fear emotion. Another possible reason is that characters do not influence the fear emotion of users. In the results of D1 group, all basic emotions were detected. The users of D1 group had the highest means in the Sadness and the Contempt, and the lowest means in the Happiness, the Disgust and the Amusement. The users of D2 group had the largest means in the Happiness and the Amusement, and had the smallest ones in the Anger, the Sadness and the Contempt. Additionally, the users of D2 group highly concentrated in the dimensions of the Happiness (SD = 0.1991) and the Sadness (SD = 0.1441). The users of the D3 group had the highest means in the Anger and the Disgust dimensions. The frequencies of the expressed emotions of the users in the D4 group were randomly distributed among the optional dimensions.

According to above analysis of Semaine annotation, we can draw the conclusion about the impacts that each character has on the emotions of users as annotated in the Semaine database. The users of D1 group expressed more the emotions of sadness and contempt, and expressed less the emotions of happiness, the disgust, and amusement. The users of D2 group had more expressions annotated in the Happiness and the Amusement, and had less annotations of sadness. The users of D3 group expressed more emotions in the Anger and the Disgust. Compared to other groups, the users of D4 group kept a balanced distribution of different types of emotional expressions. Except Prudence character, above discovery matched with the emotion-coloring definition of SAL characters (Obadiah character-gloomy, Poppy character-happy, Spike character-anger, Prudence character-sensible; G. McKeown et al., 2012).

Among the means of table 43, the Arousal dimension shown a positive relation with positive emotions (including the Happiness and the Amusement). The total number of expressed emotion types are also positively related with the mean value of the Power dimension. The Valence dimension is very interesting. Its mean value is positively associated with the mean values of the Happiness and the Amusement in D1 and D2 groups, but negatively associated with the ones in D3 and D4 groups.

6.1.3. The one-way analysis of variance in Semaine annotation of Phase II

ANOVA could determine whether three or more independent groups come from populations with different means. In other words, it could identify whether these groups represent the same population or not. Here, character type is the independent variable. The null hypothesis is that all the groups belong to the same population. Rejection of the null hypothesis does not necessarily indicate that the groups come from different groups. In order to precisely compare selected groups, multiple comparisons are

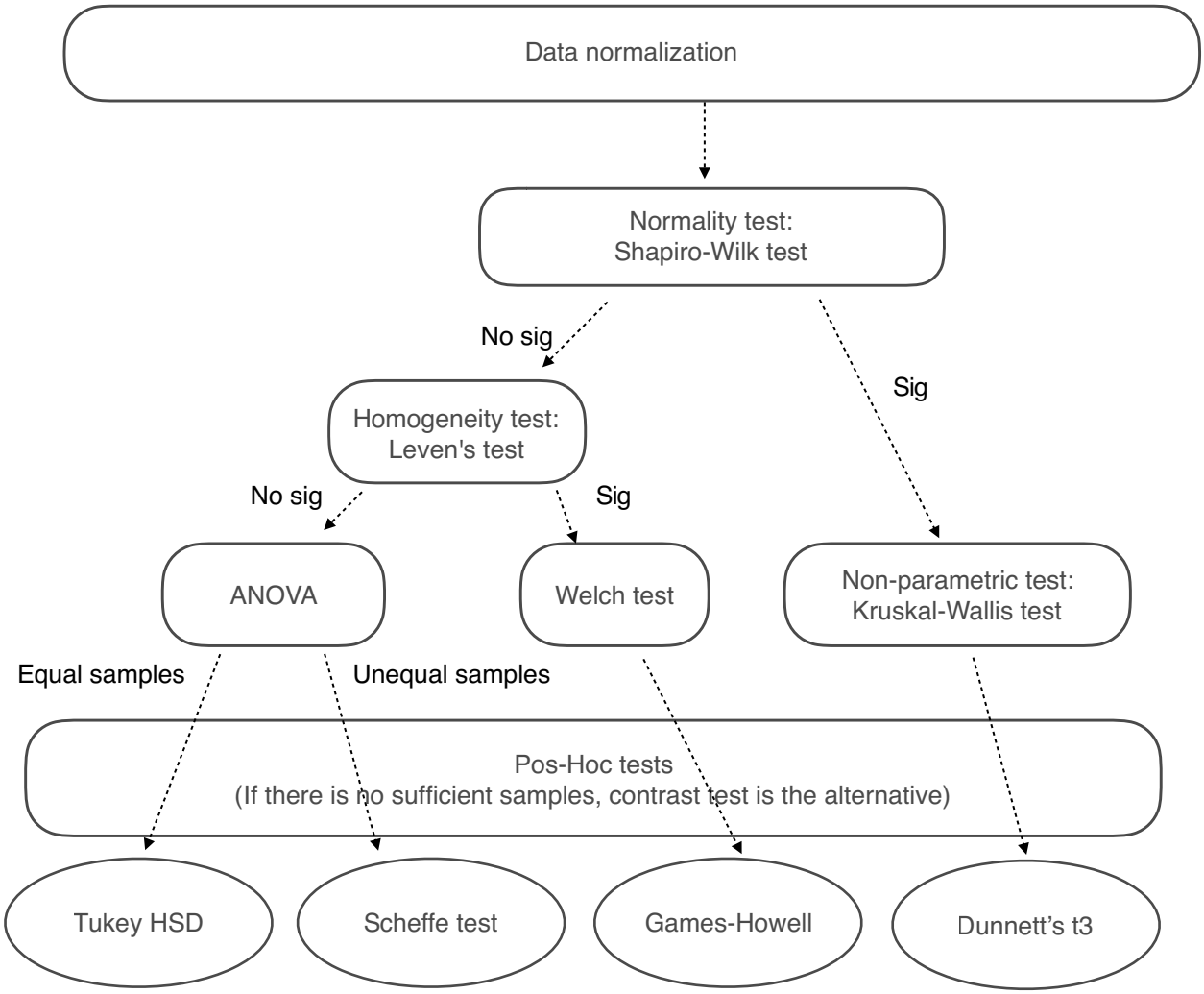

Figure 12. The chart of test selection (significance level = 0.05).

Table 44. The means of the merged sessions in the Semaine annotation of the D3 group.

Samples	User Id	Valence	Arousal	Power	Anticipation	Intensity	Fear	Anger	Happiness	Sadness	Disgust	Contempt	Amusement
47	**2**	-0.154	0.150	0.274	42.793	-0.000		-0.431				-0.765	-0.668
20	**3**	-0.037	-0.026	0.333	27.811	-0.123		-0.040				-0.050	-0.368
25	**4**	-0.108	-0.025	0.241	36.503	-0.010		0.109	-0.235			-0.009	-0.121
13	**5**	-0.165	-0.085	0.014	48.904	0.073		0.077				-0.236	-0.164
52	**8**	-0.437	0.218	0.215	34.359	0.191		0.014		-0.656		-0.297	
59	**9**	-0.233	-0.098	0.456	33.524	-0.175		-0.582				-0.573	-0.725
101	**15**	0.124	-0.214	0.261	46.900	-0.308							-0.618
3&9	**16**	-0.165	0.123	0.433	37.019	0.103		-0.571			-0.169	-0.101	-0.561
109	**17**	-0.102	-0.166	0.316	41.238	-0.082		-0.419				-0.418	-0.558
113	**18**	0.197	-0.036	0.413	45.248	-0.453							-0.784
120	**19**	0.112	-0.088	0.328	41.766	-0.313						-0.966	-0.710

Table 45. The means of the merged sessions in the Semaine annotation of the D4 group.

Samples	User Id	Valence	Arousal	Power	Anticipation	Intensity	Fear	Anger	Happiness	Sadness	Disgust	Contempt	Amusement
46.000	**2.000**	-0.327	0.067	0.074	43.287	-0.096		-0.464				-0.339	-0.746
22.000	**3.000**	0.041	-0.335	0.478	24.057	-0.425							-0.732
14.000	**5.000**	0.153	-0.071	0.308	34.351	-0.187			-0.263				-0.417
53.000	**8.000**	0.011	0.099	0.343	30.664	-0.201			-0.695	-0.653			-0.636
58.000	**9.000**	0.149	-0.196	0.474	28.397	-0.344			-0.452				-0.668
102.000	**15.000**	0.141	-0.205	0.465	29.317	-0.362			-0.410				-0.821
4&10 &31	**16.000**	0.046	0.086	0.424	32.839	-0.009			-0.658	-0.709			-0.602
106.000	**17.000**	0.107	-0.153	0.281	37.937	-0.322		-1.000	-0.548				
114.000	**18.000**	0.112	-0.303	0.400	35.000	-0.657							-0.954
119.000	**19.000**	0.285	0.032	0.458	42.065								

needed. But multiple comparisons will increase the Type I errors (i.e., rejecting the null hypothesis when the null hypothesis is true). In this way, the post hoc tests were selected to reduce the probability of the Type I errors. The whole working flow of analysis is illustrated in figure 12.

There are five main steps for the ANOVA. First step is the data normalization. The second step is the normality test (Sharpiro-Wilk test). Third step is the test of homogeneity of variance (Leven's test). Fourth step is the ANOVA/Welch/Non-parametric test. The test model is determined by the normality and the variances homogeneity of variances in each dimension of the Semaine annotation. Because the normality and the homogeneity of variances are the two assumptions for ANOVA test. These tests only indicate whether these four groups have the same population or not. In order to identify the emotion impacts of each SAL character, the last step is the multiple comparisons to analyze between paired groups. The test models include Tukey HSD, Scheffe test, Games-Howell, and the Dunnett's t3.

Step1: Before the test of normality, it was still necessary to normalize the data. Because there were some users who had generated multiple sessions. Each of them interacted with the same character for several times. These multiple sessions of the same user should be merged. The final results are listed in table 29, 30, 44 and 45.

Step 2: The Shapiro-Wilk test was conducted (significance level = 0.05). The results are listed in table 46. Most full rating dimensions passed the test. But the Valence ($p = 0.022 < 0.05$) and the Power ($p = 0.03 < 0.05$) failed the test in the D4 group. In terms of basic emotion dimensions, many dimensions had too few results to calculate the results.

In the figure 13, the normal Q-Q plots of these failed dimensions were still fitted in the linear model. They revealed no serious threats to the assumptions of distribution

normality. In this way, these dimensions would continue the following process. Combining with table 35 in the phase I (D1 and D2 groups), only full rating dimensions and the Amusement passed the normality test across the groups.

Table 46. Normality test of Semaine annotation in D3 and D4 groups. Bold numbers are referred in the text.

	Spike			Prudence		
	Statistic	df	Sig.	Statistic	df	Sig.
Valence	0.948	11.000	0.622	0.814	10.000	**0.022**
Arousal	0.933	11.000	0.440	0.908	10.000	0.267
Power	0.916	11.000	0.285	0.826	10.000	**0.030**
Anticipation	0.972	11.000	0.909	0.972	10.000	0.910
Intensity	0.969	11.000	0.876	0.968	9.000	0.879
Fear						
Anger	0.840	8.000	0.076			
Happiness				0.962	6.000	0.833
Sadness						
Disgust						
Contempt	0.931	9.000	0.487			
Amusement	0.872	10.000	0.104	0.982	8.000	0.973

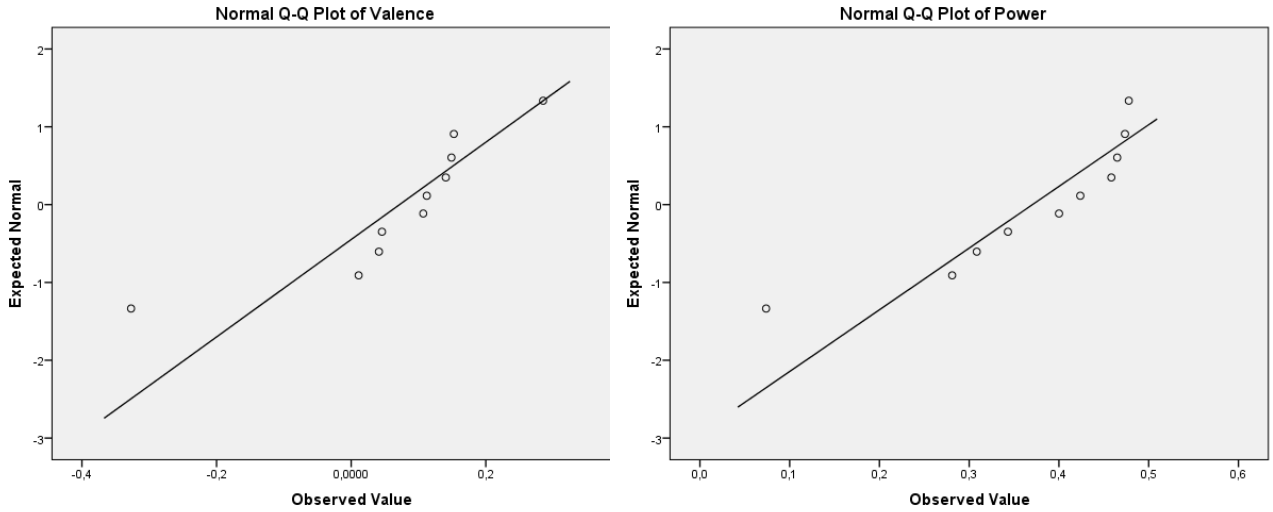

Figure 13. Normal Q-Q Plots of Valence and Power in D4 group.

Table 47. The results of test of Homogeneity of Variances.

	Levene Statistic	df1	df2	Sig.
Valence	1.351	3	39	0.272
Arousal	1.246	3	39	0.306
Power	2.620	3	39	0.064
Anticipation	0.279	3	39	0.840
Intensity	0.233	3	37	0.873
Amusement	0.709	3	29	0.554

Step 3: According to the means of the merged sessions in the Semaine annotations of four groups, table 47 illustrates the results of the test of homogeneity of variances. In this table, the full rating dimensions and the Amusement did not violate the assumption of homogeneity of variances.

Table 48. Kruskal Wallis test results. Bold numbers are referred in the text.

	Anger	Happiness	Sadness	Disgust	Contempt
Chi-Square	8.314	11.019	10.800	2.000	1.188
df	3	3	3	1	3
Asymp. Sig.	0.040	0.012	0.013	0.157	0.756
Exact Sig.	**0.008**	**0.002**	**0.000**	0.400	0.891
Point Probability	0.000	0.000	0.000	0.400	0.009

Step 4: Those dimensions, which failed in the test of normality, could be analyzed by non-parametric tests (Krusal Wallis test). The Krusal Wallis test was the alternative to the ANOVA test. Table 48 shown the results for those dimensions which failed in the step 2 (fear dimension does not have enough valid values). We assumed that the group size (10 or 11) was small, and chose exact significance for further analysis. From the table 48, the Disgust and the Contempt did not have significant difference (exact sig > 0.05). The Anger (0.008), the Happiness (0.002), and the Sadness (0.000) should further analyze the emotional impact of each character in the step 5.

Table 49. ANOVA results. Bold numbers are referred in the text.

		Sum of Squares	df	Mean Square	F	Sig.
Valence	Between Groups	1.410	3	0.470	17.086	**0.000**
	Within Groups	1.073	39	0.028		
	Total	2.483	42			
Arousal	Between Groups	0.785	3	0.262	14.729	**0.000**
	Within Groups	0.693	39	0.018		
	Total	1.479	42			
Power	Between Groups	0.438	3	0.146	7.377	**0.000**
	Within Groups	0.771	39	0.020		
	Total	1.209	42			
Anticipation	Between Groups	566.913	3	188.971	5.674	**0.003**
	Within Groups	1298.776	39	33.302		
	Total	1865.688	42			
Intensity	Between Groups	0.200	3	0.067	1.377	0.265
	Within Groups	1.791	37	0.048		
	Total	1.991	40			
Amusement	Between Groups	0.369	3	0.123	2.819	0.056
	Within Groups	1.265	29	0.044		
	Total	1.634	32			

ANOVA is prepared for the full rating and the Amusement dimensions. The ANOVA results are listed in table 49. From this table, the test results with significant level, which

equals 0.05 or less, indicate that four groups have different populations in the dimensions of the Valence, Arousal, Power, and Anticipation. These dimensions (the Valence, the Arousal, the Power, and the Anticipation) should continue the multiple comparisons in the step 5.

Table 50. Contrast coefficients[l] for contrast tests.

Contrast	Characters			
	Obadiah	Poppy	Spike	Prudence
A	1	-1	0	0
B	1	0	-1	0
C	1	0	0	-1
D	0	1	-1	0
E	0	1	0	-1
F	0	0	1	-1
l	1 and -1 mean the choice of character for comparison. 0 means unselected character.			

Table 51. The results of contrast tests. Bold numbers are referred in the text (a).

Contrast			Value of Contrast	Std. Error	t	df	Sig. (2-tailed)	Effect size
Anger	Assume equal variances	A	0.049	0.307	0.159	9	0.877	0.003
		B	-0.478	0.183	-2.608	9	0.028	0.430
		C	-0.025	0.243	-0.104	9	0.920	0.001
		D	-0.527	0.284	-1.856	9	0.096	0.277
		E	-0.074	0.326	-0.228	9	0.825	0.006
		F	0.453	0.213	2.126	9	0.062	0.334
	Does not assume equal variances	A	0.049	0.024	2.014	2.000	0.182	0.670
		B	-0.478	0.111	-4.325	6.573	**0.004**	0.740
		C	-0.025	0.269	-0.094	1.017	0.940	0.009
		D	-0.527	0.108	-4.887	6.000	**0.003**	0.799
		E	-0.074	0.268	-0.277	1.000	0.828	0.071
		F	0.453	0.289	1.570	1.345	0.312	0.647
Happiness	Assume equal variances	A	-0.468	0.128	-3.657	14	**0.003**	0.489
		B	-0.616	0.109	-5.636	14	**0.000**	0.694
		C	-0.111	0.134	-0.831	14	0.420	0.047
		D	-0.147	0.067	-2.201	14	**0.045**	0.257
		E	0.357	0.102	3.497	14	**0.004**	0.466
		F	0.504	0.077	6.531	14	**0.000**	0.753
	Does not assume equal variances	A	-0.468	0.145	-3.240	3.317	0.041	0.760
		B	-0.616	0.127	-4.865	2.000	0.040	0.922
		C	-0.111	0.143	-0.778	3.152	0.491	0.161
		D	-0.147	0.070	-2.107	7.000	0.073	0.388

Table 51. The results of contrast tests. Bold numbers are referred in the text (b).

Contrast			Value of Contrast	Std. Error	t	df	Sig. (2-tailed)	Effect size
Happiness	Does not assume equal variances	E	0.357	0.096	3.711	11.843	0.003	0.538
		F	0.504	0.066	7.614	5.000	0.001	0.921
Sadness	Assume equal variances	A	0.623	0.138	4.524	10	**0.001**	0.672
		B	0.402	0.216	1.865	10	0.092	0.258
		C	0.427	0.161	2.656	10	**0.024**	0.414
		D	-0.221	0.235	-0.939	10	0.370	0.081
		E	-0.196	0.186	-1.054	10	0.317	0.100
		F	0.025	0.249	0.100	10	0.922	0.001
	Does not assume equal variances	A	0.623	0.115	5.432	6.319	0.001	0.824
		B	0.402	0.082	4.927	7.000	0.002	0.776
		C	0.427	0.086	4.955	7.964	0.001	0.755
		D	-0.221	0.081	-2.740	2.000	0.111	0.790
		E	-0.196	0.085	-2.300	2.432	0.125	0.685
		F	0.025	0.028	0.900	1.000	0.534	0.447

Table 52. Results of Levene's test for the Anger, the Happiness, and the Sadness dimensions. Bold numbers are referred in the text.

	Levene Statistic	df1	df2	Sig.
Anger	17.983	2	10	**0.000**
Happiness	0.217	2	15	0.808
Sadness	1.444	2	11	0.278

Step 5: Not all the basic emotion dimensions had adequate samples/groups for multiple comparisons. Therefore, contrast tests were planned for the Anger, the Happiness, and the Sadness. The contrast coefficients are displayed in table 50. The results of contrast tests and related Leven's tests are illustrated in table 51 and 52 respectively. Only the Anger violated the assumption of equal variances. Then, the values in the bottom of the Anger dimensions in table 51 are valid.

In the table 51, D3 group had a population which is different from the ones of D1 and D2 groups in the Anger dimension (contrast B = 0.004, contrast D = 0.003), and the population of D3 group is also different from other three groups' populations in the Happiness dimension (contrast B = 0.0001, contrast D = 0.045, and contrast F = 0.00001). Similarly, the expressed emotions of D2 group was different from others in the Happiness dimension (contrast A = 0.003, contrast D = 0.045, contrast E = 0.004). Compared to D2 and D4 groups, the users of D1 group expressed the sad emotions differently (contrast A = 0.001, contrast C = 0.024). These statistics proved the connection between the characters and the emotional dimensions. It indicated that the characters could influence the expressed emotions of users.

Table 53. Multiple comparisons for full rating dimensions (a).
Bold numbers are referred in the text.

Dimensions	Character	Characters	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Valence	Obadiah	Poppy	-0.474*	0.071	**0.000**	-0.680	-0.267
		Spike	-0.099	0.071	0.586	-0.306	0.108
		Prudence	-0.259*	0.072	**0.011**	-0.471	-0.047
	Poppy	Obadiah	0.474*	0.071	**0.000**	0.267	0.680
		Spike	0.375*	0.071	**0.000**	0.168	0.582
		Prudence	0.215*	0.072	**0.045**	0.003	0.427
	Spike	Obadiah	0.099	0.071	0.586	-0.108	0.306
		Poppy	-0.375*	0.071	**0.000**	-0.582	-0.168
		Prudence	-0.160	0.072	0.200	-0.372	0.052
	Prudence	Obadiah	0.259*	0.072	**0.011**	0.047	0.471
		Poppy	-0.215*	0.072	**0.045**	-0.427	-0.003
		Spike	0.160	0.072	0.200	-0.052	0.372
Arousal	Obadiah	Poppy	-0.345*	0.057	**0.000**	-0.511	-0.179
		Spike	-0.306*	0.057	**0.000**	-0.472	-0.140
		Prudence	-0.230*	0.058	**0.004**	-0.401	-0.060
	Poppy	Obadiah	0.345*	0.057	**0.000**	0.179	0.511
		Spike	0.039	0.057	0.926	-0.127	0.205
		Prudence	0.114	0.058	0.295	-0.056	0.284
	Spike	Obadiah	0.306*	0.057	**0.000**	0.140	0.472
		Poppy	-0.039	0.057	0.926	-0.205	0.127
		Prudence	0.075	0.058	0.646	-0.095	0.246
	Prudence	Obadiah	0.230*	0.058	**0.004**	0.060	0.401
		Poppy	-0.114	0.058	0.295	-0.284	0.056
		Spike	-0.075	0.058	0.646	-0.246	0.095
Power	Obadiah	Poppy	-0.272*	0.060	**0.001**	-0.447	-0.097
		Spike	-0.119	0.060	0.283	-0.294	0.056
		Prudence	-0.191*	0.061	**0.033**	-0.371	-0.012
	Poppy	Obadiah	0.272*	0.060	**0.001**	0.097	0.447
		Spike	0.153	0.060	0.107	-0.022	0.328
		Prudence	0.081	0.061	0.633	-0.099	0.260
	Spike	Obadiah	0.119	0.060	0.283	-0.056	0.294
		Poppy	-0.153	0.060	0.107	-0.328	0.022
		Prudence	-0.072	0.061	0.713	-0.251	0.107
	Prudence	Obadiah	0.191*	0.061	**0.033**	0.012	0.371
		Poppy	-0.081	0.061	0.633	-0.260	0.099
		Spike	0.072	0.061	0.713	-0.107	0.251

*. The mean difference is significant at the 0.05 level.

Table 53. Multiple comparisons for full rating dimensions (b).
Bold numbers are referred in the text.

Dimensions	Character	Characters	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Anticipation	Obadiah	Poppy	6.227	2.461	0.111	-0.961	13.416
		Spike	-3.700	2.461	0.527	-10.889	3.489
		Prudence	2.151	2.521	0.866	-5.216	9.517
	Poppy	Obadiah	-6.227	2.461	0.111	-13.416	0.961
		Spike	-9.928*	2.461	**0.003**	-17.117	-2.739
		Prudence	-4.077	2.521	0.464	-11.443	3.290
	Spike	Obadiah	3.700	2.461	0.527	-3.489	10.889
		Poppy	9.928*	2.461	**0.003**	2.739	17.117
		Prudence	5.851	2.521	0.164	-1.516	13.217
	Prudence	Obadiah	-2.151	2.521	0.866	-9.517	5.216
		Poppy	4.077	2.521	0.464	-3.290	11.443
		Spike	-5.851	2.521	0.164	-13.217	1.516

*. The mean difference is significant at the 0.05 level.

In step 5, these four groups had different sizes. Therefore, the full rating dimensions used Scheffe test for multiple comparisons. The results of the multiple comparisons are displayed in the table 53. It is easy to find that D2 group has the significant difference with other groups in the Valence. It was the same with D1 group in the Arousal. Additionally, there were many cases that these four groups can be divided into two or three different teams in these analyzed dimensions.

Table 54. Homogenous subsets for full rating dimensions.
Bold numbers are referred in the text.

Modal	Characters	N	Valence			Arousal	
			Subset for alpha = 0.05			Subset for alpha = 0.05	
			1	2	3	1	2
Scheffe	Obadiah	11	-0.187			**-0.328**	
	Poppy	11			**0.287**		0.016
	Spike	11	-0.088	-0.088			-0.023
	Prudence	10		0.072			-0.098
	Sig.		0.596	0.191	1.000	1.000	0.285
Modal	Characters	N	Power			Anticipation	
			Subset for alpha = 0.05			Subset for alpha = 0.05	
			1	2		1	2
Scheffe	Obadiah	11	0.179			35.942	35.942
	Poppy	11		0.451		29.715	
	Spike	11	0.298	0.298			39.642
	Prudence	10		0.370		33.791	33.791
	Sig.		0.293	0.114		0.118	0.156

In the homogeneous subsets (table 54), the Valence divided the groups into three different teams. Except D2 group, the D1 and the D4 groups were different from each other. According to the results, other dimensions divided these four groups into two teams. Particularly, D1 group was totally different from other three groups in the Arousal.

Except D4 group, the emotion changes of participants in each group are well matched with the emotional definition of SAL characters. These SAL characters could influence multiple emotions of users. For example, Spike character influenced the user expressions of the Anger and the Happiness in the Semaine annotation. As annotated in the Semaine database, Obadiah character made users express more sad emotions, and Poppy character drove the users to display more happy emotions. In terms of full rating dimensions, the Valence and the Arousal also shown the character impacts in each group. The Valence, which relates to positive emotion, had a higher mean in D2 group than other groups. The D1 group had the highest mean in the Arousal. The Power kept a positive relationship with the total number of expressed emotion types in each group.

6.1.4. Summary in Semaine annotation of Phase II.

In this section, the distribution of emotions illustrated how each character influences the emotion of users in the Semaine annotation. The users of D1 group expressed more sad emotions, the users of D2 group expressed more happy emotions, and the users of D3 group got angry easily (D4 group needs further data). Then, the functional variation further explained the relationships among the dimensions as well as the character. For example, the Valence positively connected with Poppy character and the Happiness dimension. The Arousal positively related to the Obadiah character and the Sadness dimension. The Power positively related to the number of expressed emotion types.

The test of significant difference helps to support these discoveries. Additionally, the test proved that characters can simultaneously influence multiple emotions/dimensions. Fox example, the Spike character had the impacts on both the Happiness and the Anger. In the Semaine annotation, there were 5 dimensions (including the Arousal, the Power, the Anger, the Happiness, and the Sadness) which had statistically significant differences among/between groups. Due to the lack of adequate samples, other dimensions were not well analyzed. But these dimensions were still possible to behave differently in the annotation of other modality.

6.2. FaceReader of Phase II

Within the same group, each FaceReader sample has the same number of results in basic emotions dimensions as well as full rating dimensions. Therefore, the analysis of emotion distribution is unnecessary. It brings the hope to study those ignored emotions in the section 5.2.

6.2.1. Value variations in FaceReader of Phase II

Table 55. The functionals of FaceReader results (a). Bold numbers are referred in the text.

	Max				Min			
Group code	A1	A2	A3	A4	A1	A2	A3	A4
Character	Obadiah	Poppy	Spike	Prudence	Obadiah	Poppy	Spike	Prudence
Neutral	0.994	0.993	0.993	0.994	0.000	0.000	0.000	0.000
Happy	0.999	0.999	0.999	0.999	0.000	0.000	0.000	0.000
Sad	0.988	0.993	0.950	0.943	0.000	0.000	0.000	0.000
Angry	0.994	0.992	0.999	0.983	0.000	0.000	0.000	0.000
Surprised	0.994	1.000	0.986	0.999	0.000	0.000	0.000	0.000
Scared	1.000	1.000	1.000	0.999	0.000	0.000	0.000	0.000
Disgusted	0.987	1.000	0.696	0.933	0.000	0.000	0.000	0.000
Valence	0.999	0.998	0.998	0.999	-0.998	-1.000	-0.999	-0.999
Arousal	0.899	0.955	0.990	0.937	0.020	0.036	0.083	0.039

Table 55. The functionals of FaceReader results (b). Bold numbers are referred in the text.

	Value Range				Mean			
Group code	A1	A2	A3	A4	A1	A2	A3	A4
Character	Obadiah	Poppy	Spike	Prudence	Obadiah	Poppy	Spike	Prudence
Neutral	0.994	0.992	0.992	0.994	0.361	**0.275**	0.329	**0.423**
Happy	0.999	0.999	0.999	0.999	**0.153**	**0.305**	0.281	0.164
Sad	0.988	0.993	0.950	0.943	0.046	0.043	**0.049**	**0.036**
Angry	0.994	0.992	0.999	0.983	**0.085**	0.043	0.069	**0.035**
Surprised	0.994	1.000	0.986	0.999	0.133	**0.148**	**0.085**	0.143
Scared	1.000	1.000	1.000	0.999	0.186	**0.210**	**0.176**	0.198
Disgusted	0.986	1.000	0.696	0.933	**0.045**	0.021	0.019	**0.017**
Valence	1.997	1.998	1.997	1.998	**-0.157**	**0.031**	0.013	-0.082
Arousal	0.879	0.918	0.907	0.898	**0.310**	**0.366**	0.351	0.328

Table 55. The functionals of FaceReader results (c). Bold numbers are referred in the text.

	Standard Deviation			
Group code	A1	A2	A3	A4
Character	Obadiah	Poppy	Spike	Prudence
Neutral	0.233	**0.200**	0.207	**0.239**
Happy	0.222	**0.284**	0.244	0.218
Sad	0.096	0.088	0.092	**0.072**
Angry	**0.140**	0.102	0.117	**0.088**
Surprised	0.187	**0.207**	0.144	0.173
Scared	0.184	0.170	0.110	0.139
Disgusted	**0.074**	0.057	0.037	0.043
Valence	0.391	0.415	0.382	0.338
Arousal	0.111	0.124	0.111	0.112

After adding Spike and Prudence characters, the functionals of FaceReader results are listed in table 55. In the Arousal dimension, the value range and the mean kept a positive relationship and rose up together. Among these four groups, the A2 group had the highest means and the A1 group had the lowest ones in the full rating dimensions (the Arousal and the Valence). Meanwhile, the A2 (MValence = 0.0307, MArousal = 0.3664) and the A3 (MValence = 0.0132, MArousal = 0.3513) groups all had positive values in the Valence and the Arousal. The A1 (MValence = -0.1568, MArousal = 0.3101) group and the A4 (MValence = -0.0822, MArousal = 0.3276) group were similar to each other. Among these four groups, the means of the emotional dimensions were all positive. The users of A2 group had the lowest mean with the highest standard deviation in the Neutral (MNeutral = 0.2747, SD = 0.2). They also scored the highest mean in the Happy (MHappy = 0.305, SD = 0.2843), the Surprised (MSuprised = 0.1484, SD = 0.2066), and the Scared (MScared = 0.2103). Similarly, the users of A1 group had the smallest mean in the Happy (MHappy = 0.1531), and the largest means in the Angry (MAngry = 0.0849, SD = 0.1401), and the Disgusted (MDisgusted = 0.045, SD = 0.0741). The users of A3 group had the strongest feeling of sad (MSad = 0.0488). But they did not express too many feelings in the Surprised (MSurprised = 0. 853, SD = 0.1436) and the Scared (MScared = 0.1764, SD = 0.1104). The users of A4 group had more expressions of the neutral emotion (MNeutral = 0.4277, SD = 0.2388) rather than sad (MSad = 0.0363, SD = 0.0716), angry (MAngry = 0.0351, SD = 0.0879) or disgusted (MDisgusted = 0.0174) emotions.

According to above analysis, each character had its own impacts on the emotions of users. In the A2 group, the Poppy character displayed its impact on the happy emotion of users. The users of A1 group expressed sad, angry, and other negative emotions. The users of A3 group also frequently expressed sad and angry emotions. The Prudence character seemed to be less ‘sensible’ than it is expected with the largest

mean in the Neutral. The full rating dimensions were basically consistent with the basic emotions. The Valence kept a positive relation with positive emotions.

There were some emotion impacts of characters were different from the definition of SAL characters. For example, the users of A1 group had the largest mean in the Angry of the FaceReader results. Contrarily, the users of A3 group had the largest mean in the Sad. Compared to other groups, the users of A4 group was the least emotional and expressive with the largest mean in the Neutral.

6.2.2. Significant difference in FaceReader of Phase II

Table 56. The means of the merged sessions in the FaceReader results of the A3 group.

Sample ID	Subject /User Id	Valence	Arousal	Neutral	Happy	Sad	Angry	Surprised	Scared	Disgusted
47	2	-0.195	0.422	0.318	0.138	0.000	0.311	0.063	0.000	0.065
20	3	0.578	0.416	0.264	0.617	0.001	0.005	0.001	0.000	0.038
25	4	0.793	0.402	0.140	0.867	0.001	0.067	0.012	0.000	0.007
13	5	0.016	0.321	0.169	0.383	0.117	0.005	0.071	0.343	0.001
52	8	-0.203	0.418	0.317	0.136	0.000	0.317	0.063	0.000	0.065
59	9	0.052	0.299	0.562	0.165	0.046	0.038	0.079	0.045	0.004
101	15	0.221	0.349	0.475	0.305	0.038	0.016	0.119	0.038	0.009
3&9	16	-0.799	0.317	0.067	0.016	0.140	0.007	0.062	0.793	0.016
109	17	0.081	0.319	0.541	0.172	0.033	0.035	0.200	0.034	0.004
113	18	0.280	0.305	0.430	0.351	0.037	0.009	0.160	0.036	0.004
120	19	0.133	0.329	0.593	0.206	0.034	0.016	0.133	0.034	0.004

Table 57. The means of the merged sessions in the FaceReader results of the A4 group.

Sampl e ID	Subject /User Id	Valence	Arousal	Neutral	Happy	Sad	Angry	Surprised	Scared	Disgusted
53	2	-0.139	0.310	0.505	0.093	0.001	0.207	0.019	0.000	0.039
22	3	0.122	0.357	0.759	0.147	0.002	0.019	0.006	0.001	0.006
31	4	-0.334	0.348	0.367	0.051	0.072	0.028	0.116	0.338	0.015
14	5	-0.206	0.295	0.273	0.172	0.052	0.004	0.132	0.357	0.023
58	8	0.154	0.331	0.672	0.191	0.022	0.005	0.077	0.015	0.000
102	9	0.244	0.353	0.539	0.298	0.012	0.034	0.137	0.009	0.004
106	15	0.233	0.321	0.442	0.265	0.000	0.030	0.352	0.001	0.002
4&10& 46	16	-0.551	0.330	0.213	0.037	0.088	0.022	0.076	0.548	0.038
114	17	0.334	0.323	0.372	0.363	0.009	0.014	0.322	0.008	0.001
119	18	0.257	0.302	0.506	0.274	0.001	0.015	0.327	0.001	0.002

Similarly, one way ANOVA was used to testify whether these groups have any significant difference in the results of FaceReader. The final results still need to go through the process in figure 12. The test also has five steps - data normalization, normality test, test of homogeneity of variance (Levene's Test for equality of variances), ANOVA, and multiple comparison.

Table 58. Normality test for FaceReader results in A3 and A4 groups. Bold numbers are referred in the text.

		Shapiro-Wilk		
Dimensions	Characters	Statistic	df	Sig.
Valence	Obadiah	0.949	11	0.630
	Poppy	0.812	11	**0.014**
	Spike	0.954	11	0.699
	Prudence	0.871	10	0.102
Arousal	Obadiah	0.918	11	0.304
	Poppy	0.956	11	0.724
	Spike	0.824	11	**0.020**
	Prudence	0.939	10	0.541
Neutral	Obadiah	0.975	11	**0.928**
	Poppy	0.921	11	**0.328**
	Spike	0.942	11	**0.543**
	Prudence	0.977	10	**0.948**
Happy	Obadiah	0.818	11	**0.016**
	Poppy	0.955	11	0.703
	Spike	0.869	11	**0.075**
	Prudence	0.974	10	0.928
Sad	Obadiah	0.777	11	**0.005**
	Poppy	0.622	11	**0.000**
	Spike	0.791	11	**0.007**
	Prudence	0.740	10	**0.003**
Angry	Obadiah	0.767	11	**0.004**
	Poppy	0.746	11	**0.002**
	Spike	0.610	11	**0.000**
	Prudence	0.576	10	**0.000**
Surprised	Obadiah	0.845	11	**0.036**
	Poppy	0.958	11	0.745
	Spike	0.948	11	0.617
	Prudence	0.858	10	0.071
Scared	Obadiah	0.787	11	**0.006**
	Poppy	0.672	11	**0.000**
	Spike	0.553	11	**0.000**
	Prudence	0.654	10	**0.000**
Disgust	Obadiah	0.743	11	**0.002**
	Poppy	0.723	11	**0.001**
	Spike	0.718	11	**0.001**
	Prudence	0.805	10	**0.017**

Step 1: The repeated sessions, which had the same user interacting with the same character, should be merged together. Besides A1 and A2 groups, A3 and A4 groups also needed to merge repeated sessions. The merged results are listed in table 56 and table 57.

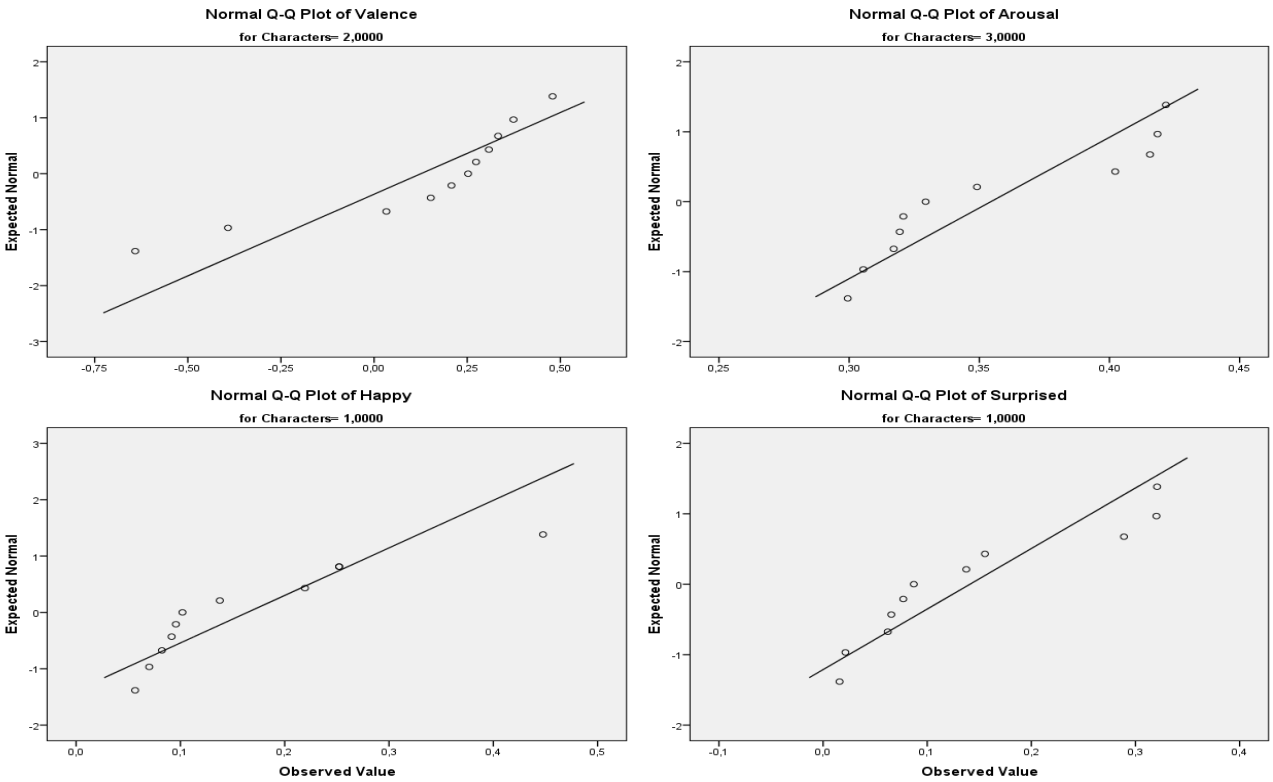

Figure 14. Normal Q-Q Plots of failed dimensions of FaceReader results in A3 and A4 groups.

Step 2: These merged values (tables 56, 57) were used for the normality test (Shapiro-Wilk test) with significance level (0.05). The results are displayed in table 58. Only the Neutral dimension passed the normality test (sig > 0.05). In the figure 14, the failed dimensions of the Valence, the Arousal, the Happy, and the Surprised presented their Q-Q plots for confirmation. But they revealed no serious threats to the assumptions of distribution normality. In this way, these dimensions would continue the following process. But the Sad, the Angry, the Scared, and the Disgust dimensions would stop the following process and participate the Krusal-Wallis test.

Table 59. Test of homogeneity of variance in FaceReader results.
Bold numbers are referred in the text.

	Levene Statistic	df1	df2	Sig.
Valence	0.450	3	39	0.719
Arousal	5.291	3	39	**0.004**
Neutral	2.512	3	39	0.073
Happy	1.945	3	39	0.138
Surprised	2.478	3	39	0.076

Step 3: According to tables 33, 34, 56 and 57, test of homogeneity of variance (Levene's Test for equality of variances) was calculated with significance level = 0.05.

The results are listed in table 59. According to the results, the Levene's test for equality of variances indicated that only Arousal dimension fail the test. Therefore, the Welch test was used for the Arousal. The rest dimensions would continue the ANOVA test.

Table 60. ANOVA of FaceReader results. Bold numbers are referred in the text.

		Sum of Squares	df	Mean Square	F	Sig.
Valence	Between Groups	0.299	3	0.100	0.925	0.438
	Within Groups	4.196	39	0.108		
	Total	4.495	42			
Neutral	Between Groups	0.168	3	0.056	2.773	**0.054**
	Within Groups	0.786	39	0.020		
	Total	0.954	42			
Happy	Between Groups	0.230	3	0.077	2.747	**0.056**
	Within Groups	1.087	39	0.028		
	Total	1.317	42			
Surprised	Between Groups	0.034	3	0.011	1.072	0.372
	Within Groups	0.416	39	0.011		
	Total	0.450	42			

Table 61. The Krusal-Wallis Test of emotional dimensions in FaceReader results.

	Sad	Angry	Scared	Disgust
Chi-Square	0.492	1.292	1.072	2.442
df	3	3	3	3
Asymp. Sig.	0.921	0.731	0.784	0.486

Table 62. Welch test of Arousal in FaceReader results.
Bold numbers are referred in the text.

		$Statistic_a$	df1	df2	Sig.
Arousal	Welch	7.603	3	21.413	**0.001**
a. Asymptotically F distributed.					

Step 4: The Valence, the Neutral, the Happy, and the Surprised were tested by ANOVA (table 60). The Neutral and the Happy dimensions failed the test (the values of sig are close to 0.05). It is possible that the Neutral and the Happy dimensions have significant differences in the subsets of four groups. Further multiple comparisons were needed to classify the difference between paired groups. The rest dimensions (except the Arousal) were tested by Krusal-Wallis test (table 61). The results indicated that four groups had no significant difference in the tested dimensions. The Arousal would go through the Welch test (table 62). The Arousal had a significant difference among its distributions in four groups (sig = 0.001 < 0.05).

Table 63. Multiple Comparisons in FaceReader. Bold numbers are referred in the text.

Dependent Variable				Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Arousal	Games-Howell	Obadiah	Poppy	-0.058*	0.013	**0.001**	-0.093	-0.022
			Spike	-0.042	0.017	0.115	-0.092	0.008
			Prudence	-0.016	0.012	0.552	-0.048	0.017
		Poppy	Obadiah	0.058*	0.013	**0.001**	0.022	0.093
			Spike	0.016	0.017	0.794	-0.034	0.066
			Prudence	0.042*	0.012	**0.008**	0.010	0.075
		Spike	Obadiah	0.042	0.017	0.115	-0.008	0.092
			Poppy	-0.016	0.017	0.794	-0.066	0.034
			Prudence	0.026	0.017	0.416	-0.022	0.074
		Prudence	Obadiah	0.016	0.012	0.552	-0.017	0.048
			Poppy	-0.042*	0.012	**0.008**	-0.075	-0.010
			Spike	-0.026	0.017	0.416	-0.074	0.022
Neutral	Scheffe	Obadiah	Poppy	0.095	0.061	0.487	-0.081	0.272
			Spike	0.039	0.061	0.936	-0.138	0.216
			Prudence	-0.079	0.062	0.660	-0.260	0.103
		Poppy	Obadiah	-0.095	0.061	0.487	-0.272	0.081
			Spike	-0.056	0.061	0.834	-0.233	0.121
			Prudence	-0.174	0.062	0.064	-0.355	0.007
		Spike	Obadiah	-0.039	0.061	0.936	-0.216	0.138
			Poppy	0.056	0.061	0.834	-0.121	0.233
			Prudence	-0.118	0.062	0.321	-0.299	0.063
		Prudence	Obadiah	0.079	0.062	0.660	-0.103	0.260
			Poppy	0.174	0.062	0.064	-0.007	0.355
			Spike	0.118	0.062	0.321	-0.063	0.299
Happy	Scheffe	Obadiah	Poppy	-0.172	0.071	0.138	-0.380	0.036
			Spike	-0.141	0.071	0.287	-0.349	0.067
			Prudence	-0.028	0.073	0.986	-0.241	0.185
		Poppy	Obadiah	0.172	0.071	0.138	-0.036	0.380
			Spike	0.031	0.071	0.978	-0.177	0.239
			Prudence	0.144	0.073	0.287	-0.069	0.357
		Spike	Obadiah	0.141	0.071	0.287	-0.067	0.349
			Poppy	-0.031	0.071	0.978	-0.239	0.177
			Prudence	0.113	0.073	0.502	-0.100	0.326
		Prudence	Obadiah	0.028	0.073	0.986	-0.185	0.241
			Poppy	-0.144	0.073	0.287	-0.357	0.069
			Spike	-0.113	0.073	0.502	-0.326	0.100

*. The mean difference is significant at the 0.05 level.

Table 64. Homogeneous subsets in FaceReader results of Phase II.

Characters		N	Arousal		Neutral		Happy	
			Subset for alpha = 0.05		Subset for alpha = 0.05		Subset for alpha = 0.05	
			1	2	1	2	1	2
Scheffe	1	11			0.296		0.164	
	2	11			0.352		0.192	
	3	11			0.392		0.305	
	4	10			0.470		0.336	
	Sig.				0.060		0.146	
Characters		N	Arousal					
			Subset for alpha = 0.05					
			1	2				
Games-Howell	1	11	0.313					
	2	11		0.370				
	3	11	0.354	0.354				
	4	10	0.328	0.328				
	Sig.		0.065	0.061				

Step 5: In order to verify the difference of emotion impacts among characters, multiple comparisons are needed. From step 4, there were three dimensions (the Arousal, the Neutral, and the Happy) had the significant differences of distributions among these four groups. The Arousal was the nonparametric dimension and should be tested in Games-Howell model. Due to the unequal group sizes, the Neutral and the Happy were tested in the Scheffe method. The results are listed in the table 63. From the table 63, no dimension shown a significant difference of distribution across groups. But there was a case that four groups can be divided into two teams. In the homogeneous subsets (table 64), only the Arousal significantly divided the groups into two teams. The characters did not have the emotion impacts on the users in the Neutral and the Happy dimensions.

According to above discussion, the emotion impacts of SAL characters are not statistically supported in the FaceReader. In a way, the impacts of characters on the users' emotions can be measured by comparing FaceReader results for participants interacting with different characters.

6.2.3. Correlations in FaceReader of Phase II

The common dimensions between FaceReader and Semaine database is illustrated in table 17. After merging the repeated sessions in the FaceReader results and the Semaine annotation, the correlation is based on the calculation of actual data in four groups. By using the means of each dimension in the FaceReader results and Semaine annotation, the correlations between FaceReader and Semaine database in each group were calculated. The A1 and A2 groups had been checked in section 5.2.3. In terms of A3 and A4 groups, the correlations between Semaine annotation and FaceReader

Table 65. The correlations of A3 group between FaceReader and Semaine annotation.

			FaceReader						
	ID	Dimensions	Valence	Arousal	Happy	Sad	Angry	Scared	Disgusted
Semaine Database Annotations	2,3,4,5,8,9,15,16,17,18,19	Valence	**0.398**	-0.343	0.237	-0.019	**-0.599**	-0.157	-0.522
	2,3,4,5,8,9,15,16,17,18,19	Arousal	-0.487	**0.540**	-0.251	-0.083	**0.737**	0.231	0.791
	4.000	Happiness							
	8.000	Sadness							
	2,3,5,8,9,16, 17	Anger	0.600	**0.515**	0.722	-0.284	**0.064**	-0.315	0.089
	2.000	Fear							
	16.000	Disgust							

Table 66. The correlations of D3 group in Semaine annotation.

Character	ID	Valence	Arousal	Happiness	Sadness	Anger	Fear	Disgust
Valence	2,3,4,5,8,9,15,16,17,18,19	**1.000**	0.236			**-0.965**		
Arousal	2,3,4,5,8,9,15,16,17,18,19	0.236	**1.000**			**0.999**		
Happiness	4							
Sadness	8							
Anger	2,3,5,8,9,16, 17	**-0.965**	**0.999**			**1.000**		
Fear	2							
Disgust	16							

Table 67. The correlations of A4 group between FaceReader and Semaine annotation.

			FaceReader						
	ID	Dimensions	Valence	Arousal	Happy	Sad	Angry	Scared	Disgusted
Semaine Database Annotations	2,3,4,5,8,9,15,16,17,18,19	Valence	**0.252**	0.215	0.347	0.176	-0.860	0.092	-0.635
	2,3,4,5,8,9,15,16,17,18,19	Arousal	-0.238	-0.726	-0.166	0.038	0.344	0.150	0.350
	5,8,9,15,16,17	Happiness	-0.079	0.871	-0.284	0.208	0.380	-0.049	-0.217
	8,16	Sadness							
	2,17	Anger							
	2	Fear							
	16	Disgust							

Table 68. The correlations of D4 group in Semaine annotation.

Character	ID	Valence	Arousal	Happiness	Sadness	Anger	Fear	Disgust
Valence	2,3,4,5,8,9,15,16,17, 18,19	**1.000**	-0.301	0.912				
Arousal	2,3,4,5,8,9,15,16,17, 18,19	-0.301	1.000	-0.649				
Happiness	5,8,9,15,16,17	0.912	-0.649	1.000				
Sadness	8,16							
Anger	2,17							
Fear	2							
Disgust	16							

results were calculated in table 65 and table 67. The correlations between Semaine annotation in the D3 and D4 groups were also calculated in table 66 and table 68.

In the A3 group, the Valence, the Arousal, and the Angry were effectively computed. Compare to table 66, only the Angry positively correlated with the Valence, the Arousal, and the Anger dimensions in the Semaine annotation (table 65). But the accuracy of the Angry-Anger (r = 0.0635) needed further improvement. Although the Valence and the Arousal were also effectively computed, it was still difficult for them to reflect their right correlations with other dimensions of Semaine annotation. For example, the Valence of FaceReader wrongly correlated with the Arousal of Semaine annotation. In the A4 group, only the Valence was positively associated between FaceReader and Semaine annotation. Due to the lack of samples, the correlations of other dimensions were unable to calculate.

Combing with tables from 38 to 41, FaceReader is able to extract the emotions, but it is greatly influenced by the characters. The FaceReader is reliable to annotate user emotion by using the Happiness and the Anger dimensions in the A1 group. In the A2 group, FaceReader was better to detect the values of the Happiness and the Sadness. In the A3 group, FaceReader had the advantage to reliably recognize the anger emotion. The A4 group lacked adequate samples to support such analysis.

6.2.4. Summary in FaceReader of Phase II

The FaceReader can detect the facial expressions and reflect the emotion impacts of each character. For example, the users of the A1 group had more expressions in the Angry and the Disgusted. The users of the A2 group had more expressions in the Happy, the Surprised, the Scared, the Valence, and the Arousal. Meanwhile, the users of the A3 group frequently displayed the sad emotion, and the users of the A4 group often displayed the neutral emotion. The interesting thing was that Obadiah and Spike characters were expected to influence the sad and the angry emotions of users respectively. Additionally, the sensibility of Prudence character did not match with its SAL definition of ‘sensible’. Because the users of the A4 group had the largest value in

the Neutral dimension. Only Poppy character matched its definition and kept influencing the happy emotion of users.

Although the characters had clear variations in the functionals analysis in the FaceReader dimensions, they did not have the significant difference among the tested groups. The results of FaceReader dimensions in each group are not so distinguishable as expected. Only the Arousal displayed the significant difference among A1, A2 and A4 groups.

In the Semaine annotation, the Valence is the most powerful dimension. It effectively correlated with the Valence of FaceReader dimensions in A1, A3, and A4 groups. Besides that, the Arousal, the Happiness, the Sadness, and the Anger of Semaine annotation were also reliable in different groups. In the A1 group, Semaine annotation were well correlated with FaceReader results in the Valence, the Happy (the Happiness), and the Angry (the Anger). In the A2 group, the Semaine annotation and FaceReader were well correlated in the Arousal, the Happy (the Happiness), the Sad (the Sadness). In the A3 group, the Valence, the Arousal, and the Anger (the Angry) were correctly correlated between the Semaine annotation and FaceReader results. It is surprising to find that the Sadness dimension was well correlated with Semaine annotation in A2 group rather A1 group.

6.3. Conclusion of Phase II

According to above analysis, there are some evidences that SAL characters have the impacts on the emotion of users.

In the Semaine annotation, the emotion impacts of characters are well reflected by the differences of emotion distributions, largest means, and statistical significances among different groups. The Happiness, the Sadness, the Anger, and other full rating dimensions (except the Intensity) are proved to be significantly different among groups. For example, the characters (Poppy and Spike) had the totally different impacts on the Happiness in the D2 and D3 groups. D2 group had the most happy expressions, and D3 group had only one happy expression. The D1 group was different from D2 and D4 groups in the Sadness. The anger emotion was more expressed in the D3 group than D1 and D2 groups. In the full rating dimensions, the character impacts were more clear. D2 group was statistically different from the rest groups in the Valence. Similarly, D1 group had significant difference with other groups in the Arousal. In the Power and the Anticipation, the groups were divided into two teams, such as D1 group against D2 and D4 groups in the Power, or D2 group against D3 group in the Anticipation.

In the FaceReader, the characters also displayed their impacts on the emotion of users in the functionals analysis. The A1 group had the largest mean in the Angry and the Disgust. A2 group was the most emotional group. It had the largest mean in the Valence, the Arousal, the Happy, the Surprised, and the Scared. A3 group just focused on sad emotion. A4 group was the least emotional and had the largest mean in the dimension of the Neutral. Unfortunately, most of these descriptive dimensions were

invalid in the statistical perspective. Only the Arousal dimension had the significant difference among groups. The Neutral and the Happy dimensions were close to the significance level (0.05).

In the correlation analysis between the Semaine annotation and the FaceReader results, the correlation of each common dimension between FaceReader and Semaine database varied with the emotion types and the characters. In terms of the full rating dimensions of FaceReader, the Valence was the most effective. It correctly correlated with the Valence dimension of Semaine annotation in the A1, A3, and A4 groups. In terms of the basic emotions in the FaceReader results, the happy, the sad, and the anger emotions were well detected. The Happiness and the Happy dimensions were correctly correlated in the A1 and A2 groups. Meanwhile, the Anger dimension of Semaine database well correlated with the Angry dimension of FaceReader results in the A1 and A3 groups. The correlation of sad emotion between Semaine annotation and FaceReader results was only valid in the A2 group.

Overall, the character could influence the people's expressed emotion. In the Semaine annotation, the emotional impacts of characters (except Prudence character) follows the Semaine definition. In the FaceReader results, these emotion impacts of characters only exists in the functionals analysis. Compared to the Semaine annotation, the impacts of these characters in the FaceReader results is damaged. In the statistical perspective, only the Arousal had the significant difference in the FaceReader. In the correlation analysis, some common dimensions between Semaine annotation and FaceReader results were reliably correlated in each group. But it is limited by the characters. For example, the Anger, which is related to the Spike character, was only correct in the correlation between A3 and D3 groups.

7. Data analysis of Phase III

After analyzing the Semaine annotation and FaceReader results, there are still some character impacts which were not reflected by the results of FaceReader. It is possible that these character impacts could be observed in the emotional expressions of other modalities. Therefore, the LIWC is added into the analysis.

The linguistic style of users is the long-term preferences which are displayed by the continuous oral/written language. In this study, we conducted the analysis of the external impacts on linguistic emotional expression based on the results of LIWC. Due to the rhetorics of text (metaphor, exaggeration, spelling mistakes, or sarcasm), it is difficult for LIWC to detect the contextual or underlying meaning. But the misleading impacts of such difficulties can be alleviated by the large set of data. Statistically, the word count strategy of LIWC is acceptable.

According to the section 3.2.3, there are almost 80 word categories/dimensions in the LIWC results. In the section 4.4.2, only linguistic process and affective process were selected. They contained 8 dimensions (Linguistic process - Word Count (WC) and Words per Sentence (WPS); affective process - affect, positive emotion, negative emotion, anxiety, anger, and sadness).

7.1. Distribution and Variations in LIWC of Phase III

Table 69. The distribution of emotional words. Bold numbers are referred in the text.

Group code		**B1 (13 samples)**	**B2 (14 samples)**	**B3 (12 samples)**	**B4 (12 samples)**
Character		**Obadiah**	**Poppy**	**Spike**	**Prudence**
Affective process	**Affect**	13	14	12	0
	Postive	**13**	**14**	**12**	2
	Negative	**13**	**14**	**11**	7
	Anxeity	**12**	7	6	3
	Anger	7	2	**10**	0
	Sad	**13**	8	6	1

The distribution of emotional words is illustrated in the table 69. Except B4 group, the users of each group fully expressed both positive and negative emotions. Particularly, participants in the B1 group focused on the Anxiety and the Sad. The people of B2 group payed attention to the positive emotion words. In the B3 group, users concentrated on the Anger. B4 group was not so emotional as expected. People of B4 used the fewest emotional words in each linguistic dimension of the affective process.

The functionals of LIWC results are listed in table 70. Similarly, the table includes the maximum, the minimum, the value range, the mean and the standard deviation.

Table 70. Functionals of LIWC (a). Bold numbers are referred in the text.

	Max				Min			
Group code	**B1**	**B2**	**B3**	**B4**	**B1**	**B2**	**B3**	**B4**
Character	**Obadiah**	**Poppy**	**Spike**	**Prudence**	**Obadiah**	**Poppy**	**Spike**	**Prudence**
WC	1044.000	944.0000	1003.000	1172.000	83.000	213.000	61.000	179.000
WPS	44.710	33.290	31.840	40.560	7.390	5.700	4.360	12.480
affect	10.840	12.940	13.020	9.250	3.880	5.693	4.830	2.620
posemo	4.950	12.770	9.840	7.210	1.090	4.767	1.610	2.310
negemo	6.220	1.8100	7.990	3.960	2.330	0.105	0.0000	0.240
anx	2.050	0.520	1.490	1.120	0.000	0.0000	0.0000	0.0000
anger	0.660	0.5200	4.140	0.947	0.000	0.0000	0.0000	0.0000
sad	3.610	0.520	3.230	0.880	0.553	0.0000	0.0000	0.0000
	Value Range				**Mean**			
Group code	**B1**	**B2**	**B3**	**B4**	**B1**	**B2**	**B3**	**B4**
Character	**Obadiah**	**Poppy**	**Spike**	**Prudence**	**Obadiah**	**Poppy**	**Spike**	**Prudence**
WC	961.000	731.000	942.000	993.000	475.111	532.576	**422.455**	**551.367**
WPS	37.320	27.5900	27.480	28.080	19.627	19.713	**13.465**	**25.254**
affect	6.960	7.247	8.190	6.630	7.168	7.675	**8.640**	**5.704**
posemo	3.860	8.003	8.230	4.900	**3.061**	**6.808**	5.021	4.448
negemo	3.890	1.705	7.990	3.720	**4.063**	**0.829**	3.581	1.173
anx	2.050	0.520	1.490	1.120	**0.784**	**0.172**	0.319	0.349
anger	0.660	0.520	4.140	0.947	0.251	**0.068**	**1.862**	0.279
sad	3.057	0.520	3.230	0.880	**1.788**	0.231	0.477	**0.170**
	Standard deviation							
Group code	**B1**	**B2**	**B3**	**B4**				
Character	**Obadiah**	**Poppy**	**Spike**	**Prudence**				
WC	314.047	235.285	327.844	300.683				
WPS	12.679	8.910	13.465	11.643				
affect	2.109	2.144	8.640	2.261				
posemo	1.222	2.357	5.021	1.734				
negemo	1.300	0.509	3.581	1.086				
anx	0.581	0.207	0.319	0.411				
anger	0.280	0.165	1.862	0.418				
sad	1.084	0.205	0.477	0.281				

In terms of linguistic process, the users of B4 group had the highest means in the words count (MWc = 551.4) and words per sentence (MWps = 25.3). The participants of B4 group is the most talkative than other groups. Meanwhile, the people in the B3 group had the lowest means in these two dimensions (MWc = 422.5 and MWps = 13.5).

Among these emotional words, the users of B2 group earned the highest mean in the positive emotion (M = 6.808), and the lowest ones in the words of the negative emotion (M = 0.829), the anxiety (M = 0.172) and the anger (M = 0.068). The people in the B1 group had the highest means in the linguistic categories of negative (M = 4.068) and sad (M = 1.788) emotions. The participants of B3 group was the most emotional. It had the largest means in the LIWC dimensions of the affect (M = 8.64) and the anger (M = 1.862). In addition, they had the second largest means in the dimensions of the positive words, the negative words, and the sad words. Although the users of B4 group were the most talkative, they expressed much less emotional words than other users.

7.2. Word clouds of LIWC in Phase III

The emotional words were too abstract to analyze the character impacts on the emotions of users. Therefore, it was beneficial to check the frequency of each words or word phases. Figure 15 to figure 18 shown the word clouds for each group. In the cloud, the size of word or phase varied with its frequency (the bigger, the higher frequency). Each color represented one category of the affective process (table 71).

According to these images, all these groups like to use the positive emotion words. But it is hard to deny that some positive emotional words overlap with common words of linguistic function. For example, the ‘nice’ of B1 group, the ‘great’ of B2 group, and the ‘good’ of B3 and B4 groups were usually used. These words like to occur in common phases like ‘I am great’ or ‘It is a good point’ rather than emotional expression.

The users of B1 group used many negative and sad emotion words. The people in the B2 group focused on the expressions of positive emotion. It used many detailed emotional words like ‘interesting’, ‘sunny’, ‘laugh’, and etc. These words basically connect with happy emotion. The B3 and B4 groups had many positive emotion words, too. But they had more negative words like the anger words like ‘rage’, ‘aggressive’, ‘mad’, and etc.

Table 71. Colors of categories in psychological process.

Category	Hex Color	Color
Affect	df3838	
Posemo	1e8bc3	
Negemo	be90d4	
Anx	ffa904	
Anger	8000	
Sad	95493c	

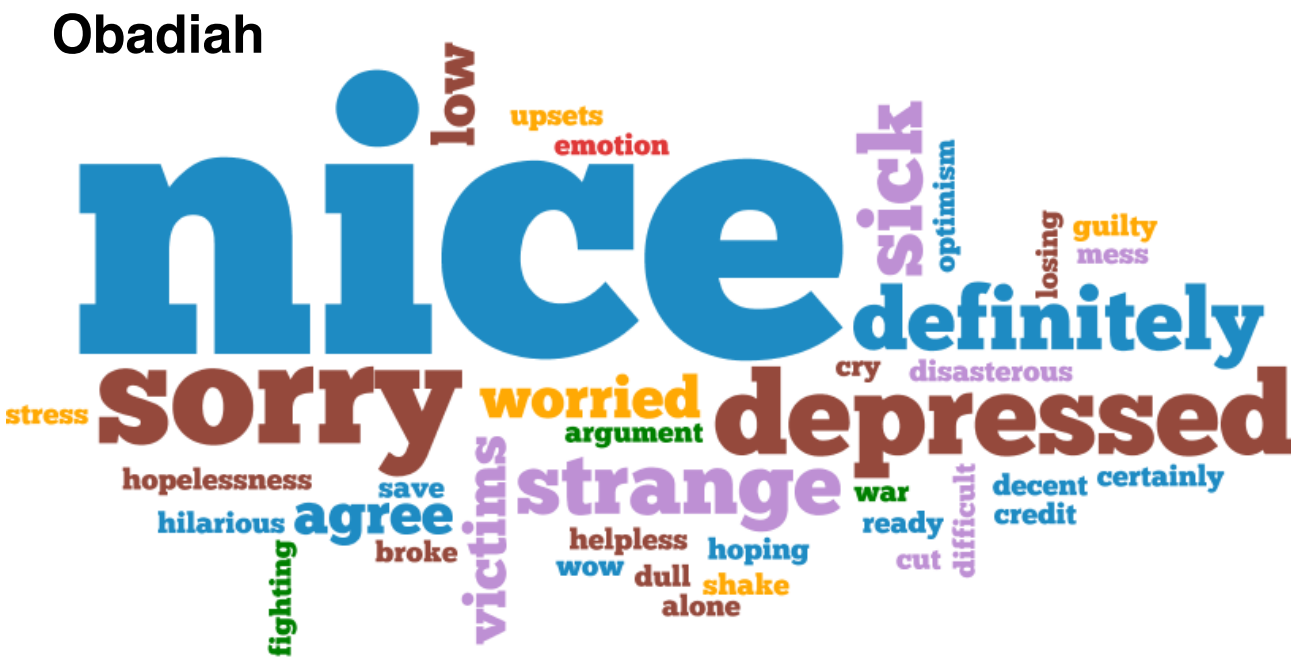

Figure 15. Obadiah Words Cloud Image.

Figure 16. Poppy Words Cloud Image.

Spike

Figure 17. Spike Words Cloud Image.

Prudence

Figure 18. Prudence Words Cloud Image.

According to above analysis, the users of each group liked to use the positive words. Particularly, the top frequent words often combined both emotional and linguistic functions. After the remove of most frequent words, the linguistic usage of each group (except B4) was consistent with the definition of characters in the Semaine database. The users of B1 group focused on the sad words, and used more negative words like the Anxiety, the Anger, and etc. The users of B2 group used many the positive emotion words with high frequency. The participants of B3 and the B4 groups used many the positive and the anger words. The users of B3 group also frequently used one sad word - 'sorry'.

7.3. Significant difference in LIWC of Phase III

According to the usage means of linguistic categories, the ANOVA can be calculated to specify the differences among different groups in the LIWC results. Similarly, the results still need to go through the work flow in figure 12. This process has five steps - data normalization, normality test, test of homogeneity of variance (Levene's Test for equality of variances), ANOVA, and multiple comparison.

Step 1: According to the user ID, the repeated sessions of LIWC should re-calculate the means of linguistic categories in each group (table 72).

Table 72. The merged means of LIWC in each groups (a).

Characters	Sample ID	User ID	WC	WPS	affect	posemo	negemo	anx	anger	sad
Obadiah	**49**	**2**	760.000	44.710	6.320	3.030	3.160	0.530	0.260	0.920
Obadiah	**19**	**3**	83.000	8.300	10.840	4.820	6.020	1.200	0.000	1.200
Obadiah	**27**	**4**	303.000	8.660	5.940	2.970	2.970	0.990	0.000	1.980
Obadiah	**15**	**5**	166.000	15.090	6.630	2.410	4.220	0.600	0.600	3.610
Obadiah	**54**	**8**	916.000	33.930	4.260	1.090	3.170	0.980	0.660	0.870
Obadiah	**61**	**9**	675.000	17.310	8.300	2.070	6.220	1.480	0.590	3.260
Obadiah	**103**	**15**	303.000	7.390	10.560	4.950	5.610	0.330	0.000	3.300
Obadiah	**5&8&30**	**16**	462.333	18.527	6.727	3.863	2.863	0.380	0.427	0.553
Obadiah	**107**	**17**	1044.000	37.290	7.090	2.390	4.600	0.290	0.480	1.250
Obadiah	**115**	**18**	132.000	7.760	7.580	3.790	3.790	0.000	0.000	2.270
Obadiah	**121**	**19**	342.000	13.150	7.890	3.800	3.800	2.050	0.000	1.460
Poppy	**48**	**2**	775.000	25.000	7.610	5.810	1.810	0.520	0.520	0.520
Poppy	**21**	**3**	699.000	33.290	5.720	4.860	0.860	0.140	0.000	0.290
Poppy	**26**	**4**	342.000	5.700	6.730	5.850	0.880	0.000	0.000	0.000
Poppy	**16**	**5**	213.000	14.200	9.860	9.390	0.470	0.000	0.000	0.470
Poppy	**55**	**8**	944.000	28.610	6.250	5.300	0.950	0.110	0.000	0.420
Poppy	**60**	**9**	595.000	10.620	12.940	12.770	0.170	0.000	0.000	0.170

Table 72. The merged means of LIWC in each groups (b).

Characters	Sample ID	User ID	WC	WPS	affect	posemo	negemo	anx	anger	sad
Poppy	**100**	**15**	219.000	11.530	8.680	7.310	1.370	0.460	0.000	0.000
Poppy	**2&11&29**	**16**	624.333	20.860	5.693	4.767	0.880	0.043	0.000	0.363
Poppy	**108**	**17**	644.000	30.670	6.990	6.370	0.470	0.160	0.000	0.310
Poppy	**112**	**18**	433.000	18.040	7.390	6.000	1.150	0.460	0.230	0.000
Poppy	**118&122**	**19**	370.000	18.320	6.565	6.465	0.105	0.000	0.000	0.000
Spike	**47**	**2**	600.000	16.220	4.830	3.170	1.670	0.000	1.000	0.170
Spike	**20**	**3**	201.000	11.170	11.440	8.960	2.490	1.490	0.000	0.000
Spike	**25**	**4**	61.000	4.360	9.840	9.840	0.000	0.000	0.000	0.000
Spike	**13**	**5**	62.000	6.890	8.060	1.610	6.450	0.000	1.610	3.230
Spike	**52**	**8**	987.000	31.840	6.890	2.840	3.950	0.000	3.140	0.200
Spike	**59**	**9**	1003.000	18.920	7.380	3.390	3.990	0.200	2.690	0.500
Spike	**101**	**15**	338.000	6.150	13.020	5.030	7.990	0.890	4.140	0.590
Spike	**3&9**	**16**	311.000	17.070	6.975	4.200	2.775	0.235	1.810	0.000
Spike	**109**	**17**	534.000	17.800	9.930	5.240	4.680	0.370	2.810	0.560
Spike	**113**	**18**	240.000	8.000	9.580	7.080	2.500	0.000	1.670	0.000
Spike	**120**	**19**	310.000	9.690	7.100	3.870	2.900	0.320	1.610	0.000
Prudence	**46**	**2**	2227.000	32.430	9.250	5.290	3.960	0.880	0.880	0.880
Prudence	**22**	**3**	648.000	38.120	2.620	2.310	0.310	0.150	0.150	0.000
Prudence	**14**	**5**	179.000	16.270	7.820	6.700	1.120	1.120	0.000	0.000
Prudence	**53**	**8**	730.000	40.560	3.700	2.740	0.960	0.270	0.000	0.270
Prudence	**58**	**9**	763.000	19.080	8.260	7.210	1.050	0.390	0.000	0.000
Prudence	**102**	**15**	369.000	12.720	4.340	3.520	0.810	0.000	0.810	0.000
Prudence	**4&10&31**	**16**	651.667	25.213	6.820	4.940	1.630	0.000	0.947	0.293
Prudence	**106**	**17**	1172.000	40.410	5.030	3.580	1.370	0.680	0.000	0.260
Prudence	**114**	**18**	362.000	12.480	5.800	5.520	0.280	0.000	0.000	0.000
Prudence	**119**	**19**	412.000	15.260	3.400	2.670	0.240	0.000	0.000	0.000

Step 2: These merged values were used for the normality test (Shapiro-Wilk test) with significance level (0.05). The results are displayed in table 73. From the table, there are many categories, which failed the test (sig < 0.05), includes the Affect, the Posemo, the Negemo, the Anx, the Anger, and the Sad. In the figure 19, the failed categories present their Q-Q plots. These failed ones still fitted with the linear model in the Q-Q plots. It means that there was no serious threat to the assumption of the distribution normality. These categories would continue the analysis.

Table 73. The Normality test of LIWC Resutls. Bold numbers are referred in the text.

Dimension	Charaters	Shapiro-Wilk		
		Statistic	df	Sig.
WC	Obadiah	0.937	12	0.466
	Poppy	0.955	11	0.704
	Spike	0.873	11	0.084
	Prudence	0.928	10	0.431
WPS	Obadiah	0.868	12	0.061
	Poppy	0.967	11	0.850
	Spike	0.894	11	0.157
	Prudence	0.858	10	0.072
Affect	Obadiah	0.947	12	0.589
	Poppy	0.830	11	**0.023**
	Spike	0.960	11	0.769
	Prudence	0.953	10	0.707
Posemo	Obadiah	0.962	12	0.815
	Poppy	0.776	11	**0.005**
	Spike	0.917	11	0.294
	Prudence	0.925	10	0.405
Negemo	Obadiah	0.916	12	0.251
	Poppy	0.958	11	0.743
	Spike	0.955	11	0.708
	Prudence	0.763	10	**0.005**
Anx	Obadiah	0.934	12	0.423
	Poppy	0.781	11	**0.005**
	Spike	0.729	11	**0.001**
	Prudence	0.840	10	**0.045**
Anger	Obadiah	0.782	12	**0.006**
	Poppy	0.496	11	**0.000**
	Spike	0.954	11	0.694
	Prudence	0.668	10	**0.000**
Sad	Obadiah	0.876	12	0.079
	Poppy	0.866	11	0.070
	Spike	0.552	11	**0.000**
	Prudence	0.672	10	**0.000**

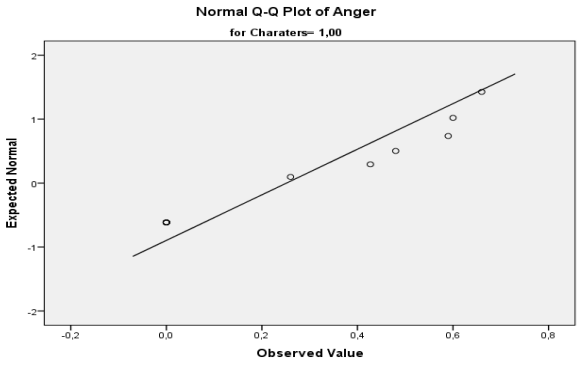

Figure 19. Q-Q plots of Failed categories in LIWC (a).

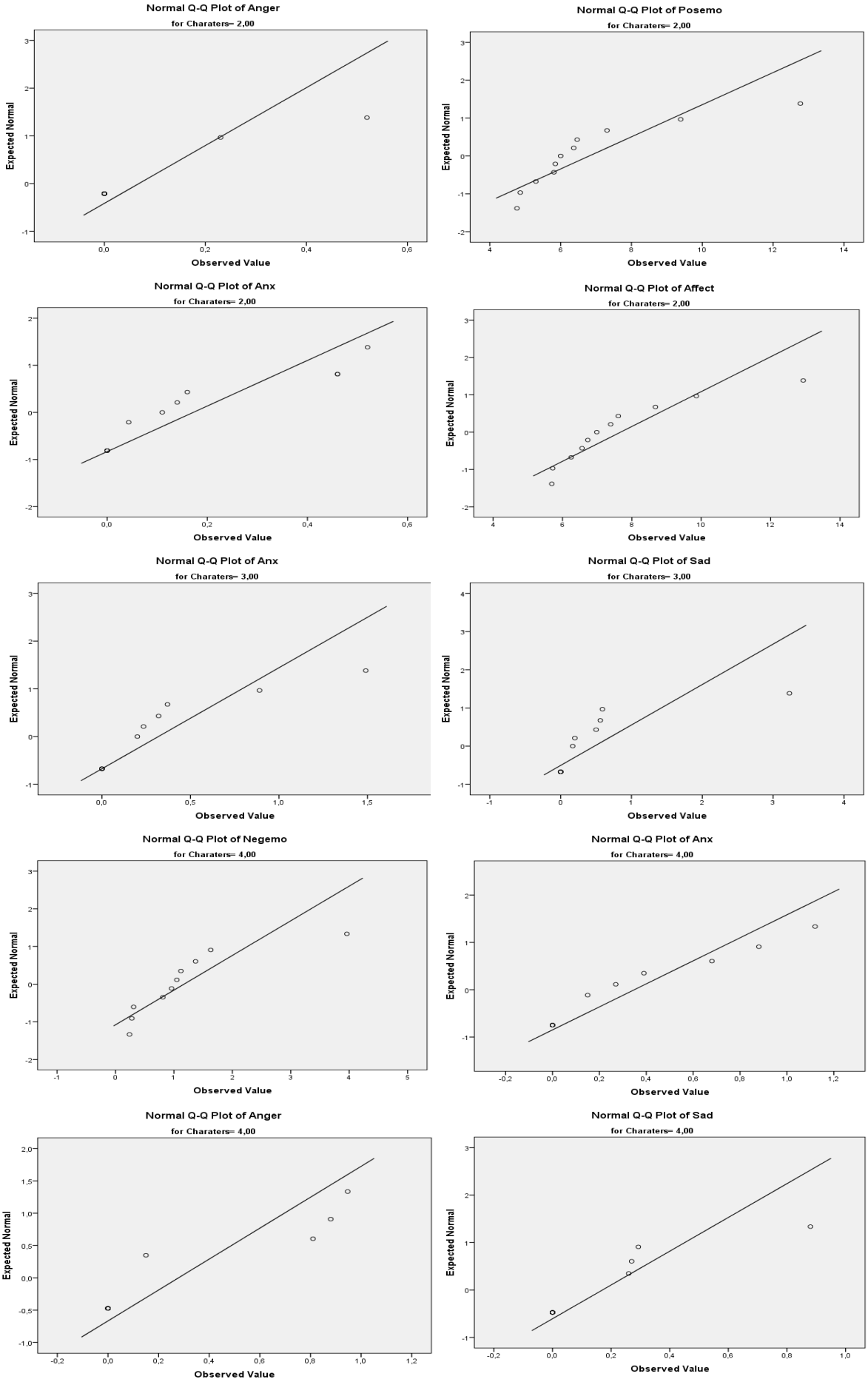

Figure 19. Q-Q plots of Failed categories in LIWC (b).

Table 74. Test of Homogeneity of variances in LIWC. Bold numbers are referred in the text.

	Levene Statistic	df1	df2	Sig.
WC	0.375	3	40	0.771
WPS	1.495	3	40	0.231
Affect	0.297	3	40	0.827
Posemo	1.413	3	40	0.253
Negemo	4.345	3	40	**0.010**
Anx	2.468	3	40	0.076
Anger	9.845	3	40	**0.000**
Sad	6.021	3	40	**0.002**

Step 3: According to step 1 and step 2, test of homogeneity of variance (Levene's Test for equality of variances) was calculated with significance level = 0.05. The results are listed in table 74. According to the results, the Leven's test indicated that the Negemo, the Anger, and the Sad categories failed the homogeneity test. These failed categories would be validated by the Welch test. The other categories would continue the ANOVA test.

Table 75. Results of Welch test in LIWC. Bold numbers are referred in the text.

	Statistica	df1	df2	Sig.
Negemo	23.475	3	19.668	**0.000**
Anger	7.823	3	19.740	**0.001**
Sad	8.025	3	20.253	**0.001**

Table 76. Results of ANVOA test in LIWC. Bold numbers are referred in the text.

		Sum of Squares	df	Mean Square	F	Sig.
WC	Between Groups	109860.513	3	36620.171	0.415	0.743
	Within Groups	3526982.261	40	88174.557		
	Total	3636842.773	43			
WPS	Between Groups	731.942	3	243.981	2.209	0.102
	Within Groups	4418.039	40	110.451		
	Total	5149.980	43			
Affect	Between Groups	46.955	3	15.652	3.185	**0.034**
	Within Groups	196.560	40	4.914		
	Total	243.515	43			
Posemo	Between Groups	82.430	3	27.477	6.599	**0.001**
	Within Groups	166.559	40	4.164		
	Total	248.989	43			
Anx	Between Groups	2.430	3	0.810	4.108	**0.012**
	Within Groups	7.887	40	0.197		
	Total	10.317	43			

Step 4: The Negemo, the Anxiety, and the Sad categories were validated by Welch test. The results are listed in table 75. The rest categories were checked by ANOVA analysis (table 76). From the these tables, all the categories of psychological process had the significant difference among four groups. In order to specify the different between paired groups, the multiple comparison were conducted.

Table 77. Multiple comparisons in LIWC (a). Bold numbers are referred in the text.

Dependent Variables	Comparison model	Characters		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Affect	Scheffe	Obadiah	Poppy	-0.507	0.925	0.959	-3.208	2.193
			Spike	-1.472	0.925	0.478	-4.173	1.228
			Prudence	1.464	0.949	0.505	-1.306	4.234
		Poppy	Obadiah	0.507	0.925	0.959	-2.193	3.208
			Spike	-0.965	0.945	0.791	-3.724	1.793
			Prudence	1.971	0.969	0.263	-0.855	4.798
		Spike	Obadiah	1.472	0.925	0.478	-1.228	4.173
			Poppy	0.965	0.945	0.791	-1.793	3.724
			Prudence	2.936	0.969	**0.039**	0.110	5.763
		Prudence	Obadiah	-1.464	0.949	0.505	-4.234	1.306
			Poppy	-1.971	0.969	0.263	-4.798	0.855
			Spike	-2.365	0.969	**0.039**	-5.763	-0.110
Posemo	Scheffe	Obadiah	Poppy	-3.747	0.852	**0.001**	-6.233	-1.262
			Spike	-1.960	0.852	0.169	-4.446	0.526
			Prudence	-1.387	0.874	0.480	-3.937	1.163
		Poppy	Obadiah	3.747	0.852	**0.001**	1.262	6.233
			Spike	1.787	0.870	0.255	-0.752	4.327
			Prudence	2.360	0.892	0.088	-0.242	4.962
		Spike	Obadiah	1.960	0.852	0.169	-0.526	4.446
			Poppy	-1.787	0.870	0.255	-4.327	0.752
			Prudence	0.573	0.892	0.937	-2.029	3.175
		Prudence	Obadiah	1.387	0.874	0.480	-1.163	3.937
			Poppy	-2.360	0.892	0.088	-4.962	0.242
			Spike	-0.573	0.892	0.937	-3.175	2.029
Negemo	Games-Howell	Obadiah	Poppy	3.234	0.405	**0.000**	2.061	4.407
			Spike	0.481	0.766	0.921	-1.713	2.676
			Prudence	2.890	0.509	**0.000**	1.466	4.314
		Poppy	Obadiah	-3.234	0.405	**0.000**	-4.407	-2.061
			Spike	-2.753	0.686	**0.009**	-4.815	-0.691
			Prudence	-0.344	0.376	0.797	-1.455	0.766
		Spike	Obadiah	-0.481	0.766	0.921	-2.676	1.713
			Poppy	2.753	0.686	**0.009**	0.691	4.815
			Prudence	2.408	0.751	**0.027**	0.240	4.577
		Prudence	Obadiah	-2.890	0.509	**0.000**	-4.314	-1.466
			Poppy	0.344	0.376	0.797	-0.766	1.455
			Spike	-2.408	0.751	**0.027**	-4.577	-0.240

Table 77. Multiple comparisons in LIWC (b). Bold numbers are referred in the text.

Dependent Variables	Comparison model	Characters		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Anx	Scheffe	Obadiah	Poppy	0.612	0.185	**0.021**	0.071	1.153
			Spike	0.466	0.185	0.115	-0.075	1.006
			Prudence	0.435	0.190	0.173	-0.120	0.990
		Poppy	Obadiah	-0.612	0.185	**0.021**	-1.153	-0.071
			Spike	-0.147	0.189	0.896	-0.699	0.406
			Prudence	-0.177	0.194	0.842	-0.743	0.389
		Spike	Obadiah	-0.466	0.185	0.115	-1.006	0.075
			Poppy	0.147	0.189	0.896	-0.406	0.699
			Prudence	-0.030	0.194	0.999	-0.597	0.536
		Prudence	Obadiah	-0.435	0.190	0.173	-0.990	0.120
			Poppy	0.177	0.194	0.842	-0.389	0.743
			Spike	0.030	0.194	0.999	-0.536	0.597
Anger	Games-Howell	Obadiah	Poppy	0.183	0.095	0.252	-0.085	0.452
			Spike	-1.610	0.393	**0.008**	-2.795	-0.426
			Prudence	-0.027	0.155	0.998	-0.473	0.419
		Poppy	Obadiah	-0.183	0.095	0.252	-0.452	0.085
			Spike	-1.794	0.388	**0.004**	-2.972	-0.615
			Prudence	-0.211	0.141	0.474	-0.633	0.211
		Spike	Obadiah	1.610	0.393	**0.008**	0.426	2.795
			Poppy	1.794	0.388	**0.004**	0.615	2.972
			Prudence	1.583	0.406	**0.010**	0.381	2.785
		Prudence	Obadiah	0.027	0.155	0.998	-0.419	0.473
			Poppy	0.211	0.141	0.474	-0.211	0.633
			Spike	-1.583	0.406	**0.010**	-2.785	-0.381
Sad	Games-Howell	Obadiah	Poppy	1.557	0.319	**0.002**	0.608	2.505
			Spike	1.310	0.423	**0.026**	0.131	2.490
			Prudence	1.617	0.325	**0.001**	0.660	2.575
		Poppy	Obadiah	-1.557	0.319	**0.002**	-2.505	-0.608
			Spike	-0.246	0.291	0.832	-1.123	0.631
			Prudence	0.061	0.108	0.942	-0.248	0.370
		Spike	Obadiah	-1.310	0.423	**0.026**	-2.490	-0.131
			Poppy	0.246	0.291	0.832	-0.631	1.123
			Prudence	0.307	0.298	0.736	-0.579	1.193
		Prudence	Obadiah	-1.617	0.325	**0.001**	-2.575	-0.660
			Poppy	-0.061	0.108	0.942	-0.370	0.248
			Spike	-0.307	0.298	0.736	-1.193	0.579

Step 5: According to step 4, the Negemo, the Anger, and the Sad categories used the Games-Howell model. The other ones use the Scheffe model. The categories of the linguistic process were not included in the table 77. Because no significant difference was found in the ANOVA analysis. In the affective process, the users of B3 group was totally different from other users in the Anger category. The people in B1 group also had the same significant difference in the usage of the Sad category. In terms of the negative emotion categories, the difference only existed in certain pairs of groups. For example, the Anxiety category only had the difference between the B1 and B2 groups.

Table 78. Homogenous subsets of categories in LIWC. Bold numbers are referred in the text.

Modal	Charaters	N	Affect		Posemo		Anx	
			Subset for alpha = 0.05		Subset for alpha = 0.05		Subset for alpha = 0.05	
			1	2	1	2	1	2
Scheffe	Obadiah	12	7.168	7.168	**3.061**			**0.784**
	Poppy	11	7.675	7.675		**6.808**	**0.172**	
	Spike	11		**8.640**	5.021	5.021	0.319	0.319
	Prudence	10	**5.704**		4.448	4.448	0.349	0.349
	Sig.		0.244	0.498	0.186	0.078	0.832	0.128
Modal	Charaters	N	Negemo		Anger		Posemo	
			Subset for alpha = 0.05		Subset for alpha = 0.05		Subset for alpha = 0.05	
			1	2	1	2	1	2
Games-Howell	Obadiah	12		**4.063**	**0.251**			**1.788**
	Poppy	11	**0.829**		**0.068**		**0.231**	
	Spike	11		**3.581**		**1.862**	**0.477**	
	Prudence	10	**1.173**		**0.279**		**0.170**	
	Sig.		0.956	0.889	0.916	1.000	0.825	1.000

In the homogenous subsets (table 78), the categories of affective process were divided into two parts. Except the Anger, the Sad, and the Negemo categories, not all paired groups had significant differences in each category. For example, the users of B1 and B2 groups were significantly different in the Posemo and the Anx. But the users of B3 and B4 groups did not have the significant difference in these tested categories. The affective words (M = 5.704) were least used in the B4 group. The results did not match with the SAL definition of Prudence character.

According to above discussion, the character impacts on the user emotion in the text were statistically supported in the LIWC. The Obadiah character had a great impact on the Sad category. The emotion impacts of Poppy character on the Posemo category were partially reflected, and were only displayed in the comparison between group B1 and group B2. Meanwhile, the Spike character influenced the Anger category. The sensibility of Prudence character was not well supported by the results.

7.4. Correlations in LIWC of Phase III

Table 79. the dimensions match between LIWC and Semaine annotation

Dimensions	LIWC	Semaine annotation
Basic emotions/Psychological prorcess	Affect	
	Negemo	
	Anx	
	Posemo	Happiness
	Ang	Anger
	Sad	Sadness
		Fear
		Disgust
		Amusement
		Contempt

Although there are only three common dimensions between LIWC and Semaine annotation, it is still helpful to illustrate the correlation of annotations. As table 79 displayed, the Posemo, the Sad and the Anger categories matched with the dimensions of Semaine annotation. The correlations were based on the actual values of tables 29, 30, 44, 45, and 72. After correlating with Semaine annotation, there are several empty elements in the tables.

Table 80. The correlations between B1 and D1 groups. Bold numbers are referred in the text.

			LIWC		
	User ID	Dimension	Posemo	Ang	Sad
Semaine annotation	3,5,16	Happiness	**0.9167**	-0.9986	-0.5246
	2,8,16	Anger	-0.7088	**0.9984**	-0.0881
	2,3,4,5,8,9,15,16,17	Sadness	0.0407	-0.2192	-0.1734

Table 81. The correlations between B2 and D2 groups. Bold numbers are referred in the text.

			LIWC		
	User ID	Dimension	Posemo	Ang	Sad
Semaine Database Annotations	2,3,4,5,8,9,15,16, 17	Happiness	**0.3487**	-0.6514	**-0.2407**
	2	Anger			
	8,16,17	Sadness	-0.5640		-0.2743

Table 82. The correlations between B3 and D3 groups. Bold numbers are referred in the text.

			LIWC		
	User ID	Dimension	Posemo	Ang	Sad
Semaine Database Annotations	4	Happiness			
	2,3,5,8,9,16, 17	Anger	0.3211	-0.4031	0.3186
	8	Sadness			

Table 83. The correlations between B4 and D4 groups. Bold numbers are referred in the text.

			LIWC		
	User ID	Dimension	Posemo	Ang	Sad
Semaine Database Annotations	5,8,9,15,16,17	Happiness	**0.5977**	-0.1852	-0.8824
	2,17	Anger			
	8,16	Sadness			

The correlation results are displayed in the tables of 39, 41, 66, 68, 80, 81, 82, and 83. According to these tables, only the Happiness (Posemo category) is reliably correlated across the characters (Due to insufficient samples, B3 group is not included). The correlation of the Happiness dimension (Posemo category) between B1 and D1 groups was more than 0.9. Besides that, the Sadness dimension (Sad category) was only correctly correlated between the B2 and D2 groups. The Anger dimension (Ang category) was only positively correlated between the B1 and D1 groups.

According to above analysis, the results of the LIWC and the Semaine annotation are partially associated across the characters. The happy emotion is the most widely recognized emotional expression. The Anger dimension is only positively correlated with the Ang category between the B1 and D1 groups instead of B3 and D3 groups.

7.5. Conclusion in LIWC of Phase III

In this chapter, the text modality had been tested. It revealed the functional variance and the distribution of linguistic usage among different groups. The linguistic usages of users in each group are highly consistent with the SAL definition of characters. The people in B1 group kept focusing on the sad and other negative words. The users of B2 group used more happy/positive emotion related words. The participants of B3 group used many words of anger emotion. The users of B4 group were the least emotional with the fewest usage of emotional words.

In the view of the statistics, the emotion impacts of characters were partially reflected. Because only the Sad and the Ang categories have the significant differences across groups. For example, B1 group was totally different from other groups in the category of the Sad. Similarly, the B3 group differs from others in the Ang category. In the rest categories of affective process, the emotion change of users in the text varied with the type of character. For example, B1 and B3 groups are different from B2 and B4

groups in the usage of the negative emotion words. In the categories of positive emotion and anxiety words, the difference only exists between the B1 and B2 groups. After all these analyses, it is still difficult to prove and measure the sensibility of B4 group and Prudence character.

In terms of correlations, only the Happiness dimension of Semaine annotation was reliably correlated with the Posemo category of LIWC across the groups (except B3&D3). The Happiness of Semaine annotation also well correlated with the Sad category in the B2 group. In the rest emotional dimensions of the Semaine annotation, the Anger dimension reliably associated with the Ang category in the B1 group. The interesting thing is that the users of B1 group failed to strengthen the correlation between the Sadness and the sad words. The users of B3 group had the same problem between the Anger and the anger words, too. According to current data, the correlation coefficients of common dimensions between Semaine and LIWC annotations were not aligned with the definition of SAL characters.

According to above analysis, the LIWC is able to reflect the character impacts. But the statistical significances of LIWC categories and the correlations of common dimensions between different annotations were partially damaged. For example, the significant difference of the Posemo between B1 and B3 groups was not displayed in the LIWC. But the statistical differences of the Anger and the Sad categories still matched with the definition of SAL characters. Only the Posemo category of LIWC widely correlated with the Happiness of Semaine annotation across groups. The correlations of other dimensions between Semaine and LIWC annotations are limited by the interacted characters and the expression modalities.

8. Conclusion

8.1. Emotion impacts of characters

During the analysis of Semaine annotation, it is easy to see the emotional impacts of each character. The majority of participants displayed the emotional expressions that are to some extent aligned with the definition of the characters. It is often to see the same set of users who expressed different emotions while interacting with different characters. Sometimes, these expressed emotions of users conflicted with each other. But the users' expressions of the whole group still match with the SAL definition of characters. In the Semaine annotation, most of these character impacts were supported by the statistical analysis. These statistical differences further proved the emotion impacts of characters matched with the SAL definition of characters. But it is difficult to measure the sensibility of Prudence character.

In the results of FaceReader, the emotion impacts of the characters were partially reflected in the functionals analysis (not fully reflected in the statistical analysis). The users of A2 group had the largest mean in the Happy. The interesting thing is that the emotion impacts of characters in the A1 and A3 groups were different with the definition of SAL definition. Because the users of A1 group had a higher mean than A3 group in the Angry (table 55b). Similarly, the users of A3 group had a higher mean than A1 group in the Sad (table 55b). The the users of A4 group, who interacted with Prudence character, were contrary with "sensible" by scoring the highest mean of neutral emotion. After the significance testing, only the Arousal dimension of FaceReader results kept the difference among four groups (A1 to A4 groups). The rest full rating dimensions and other emotional dimensions were not significantly different any more across groups.

In the analysis of LIWC, the emotion impacts of characters were reflected. LIWC shown the emotion impacts of characters in the linguistic distribution, the functional analysis, and the statistical analysis. The usages of linguistic categories for users were highly consistent with the definition of SAL characters. For example, people told to Obadiah character with more negative words like the categories of the Negemo, the Anx, and the Sad. Participants, who told to Poppy character, used more positive words like the Posemo category. The users, who interacted with Spike and Prudence characters, used more the words of the Posemo and the Ang categories. In terms of he statistical analysis, the Obadiah and the Spike characters shown their unique impacts on the linguistic style of users respectively. The people, who told to Obadiah character, used the most sad words, and behaved differently from others. The users, who interacted with Spike character, used the most angry words, and were significantly different from the users in other groups. In the rest categories, the character impacts on the users' linguistic usages were partially reflected. For example, the significant differences of the Posemo and the Anx categories only exist between B1 and B2 groups, the Negemo category divides the groups into two teams in table 78 (B1 and B3 groups - B2 and B4 groups).

Overall, the character impacts on the user emotions were reflected in the results of Semaine annotation, FaceReader and LIWC. In terms of statistical analysis, the reflections of such impacts varied with the types of character, emotion and modality. Compared to manual annotation in the Semaine database, the statistical significance of automatic annotations is less clear, For example, the FaceReader is only statistically significant in the Arousal. The LIWC is statistically significant in multiple emotional categories like the Posemo, the Sad, the Ang, and etc.

8.2. The correlation between Semaine annotation and the results of automatic tools

The dimensions of FaceReader results reliably correlated with Semaine annotation in the emotional dimensions of happiness (happy), sadness (sad), and anger (angry). But the correlation between the annotations of Semaine database and FaceReader is limited by the type of character and emotion. For example, the anger emotion was only well correlated between the A3 and D3 groups (table 65). The happy emotion is well associated in both A1&D1 and A2&D2 groups (tables 38 and 40). In addition, the SAL definition of characters is not fully supported in the correlation analysis between FaceReader and Semaine database. Because the Sadness annotations correctly correlated with FaceReader only between A2 and D2 groups (table 40) rather than A1 and D1 groups. Meanwhile, the angry emotions are well correlated between A1 and D1 groups instead of A3 and D3 groups.

In terms of the correlation analysis between the annotations of LIWC and Seamine database, the Posemo (happy emotion) of LIWC is the most effectively correlated across the characters and modalities (the pair of B3 and D3 groups is ignored due to the insufficient samples). The Ang category only well correlated with the anger dimension of Semaine annotation between B1 and D1 groups. The sad category failed to correlated with the dimensions of Semaine annotation.

Compared to FaceReader, LIWC is more reliable to the character impacts on certain user emotion (only Posemo/Happiness) as annotated in the Semaine annotation. But the FaceReader has more dimensions which were also correlated with the Semaine annotation within the subsets of groups. In the correlation analysis with Semaine annotation, LIWC and FaceReader have different advantages.

8.3. Limitations and future work

During the experiments, we have to admit that these groups had the same set of users. It is possible for them to adapt to the research context to please the researchers. As mentioned above, some emotions are rare to be found in the Semaine annotation. Such absence or lack of certain emotions may have some connections with the users' adaptation. It could cause the unbalanced distribution of emotional expressions.

In the FaceReader, the fusion of emotions is another issue which could undermine the accuracy of automatic annotation. FaceReader heavily relies on the facial feature. The emotion fusion, which mainly influences the quality of facial expression, could lead

the biases, even errors in the emotion recognition. Compared to facial expression, the accuracy of text analysis is also challenged by the rhetorics like sarcasm, metaphor, and etc.

The recognition quality of automatic tools also needs further improvement. For example, there are many empty elements in the samples of Semaine database. Because researchers are uncertain about the judgement of these elements. FaceReader assigns values to all processed samples which contain those uncertain elements. But there is another case. Some emotions are expressed by other modalities rather than facial expression. In this way, FaceReader still make mistakes with correct emotional recognition.

In terms of LIWC, there are many emotional words which are usually used as the common linguistic words, such as 'nice', 'good', 'great', and etc. Due to the lack of contextual analysis, these words are often regarded as emotional words. It undermines the frequency distribution of emotional words, and heavily influenced the image of word clouds (section 7.2). Although the quantitative and the probabilistic analysis is helpful, the final results are not as good as expected.

Apart from the above problems, the sensibility of Prudence group is very difficult to be measured under current experimental setting. Because there is no dimension or metric which has the direct relation with the emotional sensibility.

In terms of future work, there are three points:

1. The increment of experiment scale. It is one of the important methods to improve the statistical probability of detection accuracy. It is helpful to reduce the mistake caused by the emotion fusion or the rhetoric. If the samples are randomly selected, such increment is also beneficial to compensate the unbalanced distribution of emotion expressions in the samples of Semaine database.

2. The generic measurement across modalities. During the analysis of different modalities, the metrics of measurements vary with tools. It generates the variances of polarity or values for the comparisons between the annotations of different tools. In this experiment, the LIWC only has the Posemo category to contain all the positive emotional words. The Posemo category has a larger emotion domain than the Happy dimension of the FaceReader. In this way, the generic measurement should provide a standard emotional description for emotion comparison between different modalities.

3. More types of emotional stimuli of characters can be added. In the Semaine database, only four characters were created and aligned with four different personalities. More types of emotional stimuli are helpful to investigate how different emotions/personalities influence the users' emotions. What's more, the experiment will be more real and sophisticated if the 'operator' is able to respond to the emotional feedbacks of 'user'.

9. References

Affectiva. (2016). Affdex. Retrieved from http://www.affectiva.com/solutions/affdex/

Ambadar, Z., Schooler, J. W., & Cohn, J. F. (2005). Deciphering the enigmatic face the importance of facial dynamics in interpreting subtle facial expressions. *Psychological science, 16*(5), 403-410.

André, E., Klesen, M., Gebhard, P., Allen, S., & Rist, T. (2000). Integrating Models of Personality and Emotions into Lifelike Characters. In A. Paiva (Ed.), *Affective Interactions* (Vol. 1814, pp. 150-165): Springer Berlin Heidelberg.

Bales, R. F. (1950). Interaction process analysis; a method for the study of small groups.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*(3), 614.

Baron-Cohen, S. (2003). Mind reading : the interactive guide to emotions: Jessica Kingsley Publishers.

Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., & Movellan, J. (2006). *Fully automatic facial action recognition in spontaneous behavior*.

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics, 41*(1), 164-171.

Blumberg, B., Downie, M., Ivanov, Y., Berlin, M., Johnson, M. P., & Tomlinson, B. (2002). *Integrated learning for interactive synthetic characters.* Paper presented at the 29th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '02, San Antonio, TX.

Bradley, M. M., & Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings. Retrieved from

Cannon, W. B. (1927). The James-Lange theory of emotions: A critical examination and an alternative theory. *The American journal of psychology, 39*(1/4), 106-124.

Chomsky, N. (1995). *The minimalist program* (Vol. 1765): Cambridge Univ Press.

Cohn, J. F. (2006). Foundations of human computing: facial expression and emotion.

Corradini, A., Mehta, M., Bernsen, N. O., Martin, J., & Abrilian, S. (2005). Multimodal input fusion in human-computer interaction. *NATO Science Series Sub Series III Computer and Systems Sciences, 198*, 223.

Cowie, R., & Douglas-Cowie, E. (1996). Automatic statistical analysis of the signal and prosodic signs of emotion in speech.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine, 18*(1), 32-80. doi:10.1109/79.911197

Cowie, R., & McKeown, G. (2010). Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme. *SEMAINE Report D6b*.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika, 16*(3), 297-334.

Darwin, C., Ekman, P., & Prodger, P. (1998). The expression of the emotions in man and animals: Oxford University Press, USA.

Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, S. (2014). COVAREP—A collaborative voice analysis repository for speech technologies.

Dietz, R., & Lang, A. (1999, 1999). Affective agents: Effects of agent affect on arousal, attention, liking and learning.

Digman, J. M. (1990). Personality structure: emergence of the five-factor model. *Annual Review of Psychology, 41*(1), 417-440.

Douglas-Cowie, E., Cowie, R., Cox, C., Amier, N., & Heylen, D. K. J. (2008). The sensitive artificial listner: an induction technique for generating emotionally coloured conversation.

Drugman, T., Gurban, M., & Thiran, J.-P. (2007). Relevant feature selection for audio-visual speech recognition.

Ekman, P. (1989a). The argument and evidence about universals in facial expressions. *Handbook of social psychophysiology*, 143-164.

Ekman, P. (1989b). The argument and evidence about universals in facial expressions. *Handbook of social psychophysiology*, 143-164.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*(3-4), 169-200. doi:10.1080/02699939208411068

Ekman, P. (2003). Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences, 1000*(1), 205-221.

Ekman, P. (2009). Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage (Revised Edition): WW Norton & Company.

Ekman, P., & Friesen, W. (1978). Facial Action Coding System Investigator's Guide: Consulting Psychologists Press.

Ekman, P., & Friesen, W. V. (1978). Facial action coding system: a technique for the measurement of facial movement.

Ekman, P., & Friesen, W. V. (2003). Unmasking the face: A guide to recognizing emotions from facial clues: Ishk.

Ekman, P., Friesen, W. V., & Ellsworth, P. (2013). Emotion in the human face: Guidelines for research and an integration of findings: Elsevier.

El Kaliouby, R., & Robinson, P. (2005). Real-time inference of complex mental states from facial expressions and head gestures *Real-time vision for human-computer interaction* (pp. 181-200): Springer.

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

Fant, G. (1995). The LF-model revisited. Transformations and frequency domain analysis. *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm, 2*(3), 40.

Fragopanagos, N., & Taylor, J. G. (2005). Emotion recognition in human–computer interaction. *Neural Networks, 18*(4), 389-405.

Grandjean, D., Sander, D., & Scherer, K. R. (2008). Conscious emotional experience emerges as a function of multilevel, appraisal- driven response synchronization. *Consciousness and Cognition, 17*(2), 484-495. doi:10.1016/j.concog.2008.03.019

Grimm, M., Kroschel, K., & Narayanan, S. (2008, 2008). The Vera am Mittag German audio-visual emotional speech database.

Gu, H., & Ji, Q. (2005). Information extraction from image sequences of real-world facial expressions. *Machine Vision and Applications, 16*(2), 105-115.

Gunes, H. (2010). Automatic, dimensional and continuous emotion recognition.

Hirsh, J. B., & Peterson, J. B. (2009). Personality and language use in self-narratives. *Journal of Research in Personality, 43*(3), 524-527. doi:http://dx.doi.org/10.1016/j.jrp.2009.01.006

Hoffmann, H., Traue, H., Bachmayr, F., & Kessler, H. (2006). Perception of Dynamic Facial Expressions of Emotion. In E. André, L. Dybkjær, W. Minker, H. Neumann, & M. Weber (Eds.), *Perception and Interactive Technologies* (Vol. 4021, pp. 175-178): Springer Berlin Heidelberg.

iMotions. (2015). Emotient Module : Facial Expression Emotion Analysis. Retrieved from https://imotions.com/software/add-on-modules/attention-tool-facet-module-facial-action-coding-system-facs/

iMotions. (2016). Applications for Neuromarketing Research. Retrieved from https://imotions.com/solutions/

Jaimes, A., & Sebe, N. (2007). Multimodal human–computer interaction: A survey. *Computer vision and image understanding, 108*(1), 116-134.

Kanade, T., Cohn, J. F., & Yingli Tian, J. F. (2000). Comprehensive database for facial expression analysis (pp. 46-53).

Keltner, D., Ekman, P., Gonzaga, G. C., & Beer, J. (2003). Facial expression of emotion.

Kim, J. (2007). Bimodal emotion recognition using speech and physiological changes: Citeseer.

Kuhn, R., & De Mori, R. (1990). A cache- based natural language model for speech recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 12*(6), 570-583. doi:10.1109/34.56193

Ladd, D. R. (2008). *Intonational phonology*: Cambridge University Press.

Ladd, D. R., Silverman, K. E., Tolkmitt, F., Bergmann, G., & Scherer, K. R. (1985). Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. *The Journal of the Acoustical Society of America, 78*(2), 435-444.

Lieberman, P., & Michaels, S. B. (1962). Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. *The Journal of the Acoustical Society of America, 34*(7), 922-927.

Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., & Bartlett, M. (2011). *The computer expression recognition toolbox (CERT)*.

Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research, 30*, 457-500.

Mavridis, N. (2015). A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems, 63*, 22-35.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of personality, 60*(2), 175-215.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schröder, M. (2012). The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing, 3*(1), 5-17. doi:10.1109/T-AFFC.2011.20

McKeown, G., Valstar, M. F., Cowie, R., & Pantic, M. (2010). The SEMAINE corpus of emotionally coloured character interactions.

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology, 14*(4), 261-292.

Mehrabian, A. (2008). Communication without words. *Communication Theory*, 193-200.

Metallinou, A., Lee, C.-C., Busso, C., Carnicke, S., & Narayanan, S. (2010). The USC CreativeIT database: A multimodal database of theatrical improvisation. *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, 55.

Miehlke, A., Fisch, U., & Eneroth, C.-M. (1973). *Surgery of the facial nerve*: Saunders.

Mortillaro, M., Meuleman, B., & Scherer, K. R. (2012). Advocating a Componential Appraisal Model to Guide Emotion Recognition. *International Journal of Synthetic Emotions (IJSE), 3*(1), 18-32. doi:10.4018/jse.2012010102

Nowlis, V., & Nowlis, H. H. (1956). The description and analysis of mood. *Annals of the New York Academy of Sciences, 65*(1), 345-355.

O'Shaughnessy, D. (2008). Invited paper: Automatic speech recognition: History, methods and challenges. *Pattern Recognition, 41*(10), 2965-2979.

Oudeyer, P.-Y. (2003). The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies, 59*(1), 157-183.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques.

Pantic, M., Valstar, M., Rademaker, R., & Maat, L. (2005). Web- based database for facial expression analysis (pp. 5 pp.). USA.

Pennebaker, J., King, L., & Diener, E. (1999). Linguistic Styles: Language Use as an Individual Difference. *Journal of Personality and Social Psychology, 77(6)*, 1296-1312.

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The development and psychometric properties of LIWC2007.

Picard, R. W. (2003). Affective computing: challenges. *International Journal of Human-Computer Studies, 59*(1), 55-64.

Revelle, W., & Scherer, K. R. (2009). Personality and emotion. 304-306.

Roach, P., Stibbard, R., Osborne, J., Arnfield, S., & Setter, J. (1998). Transcription of prosodic and paralinguistic features of emotional speech. *Journal of the International Phonetic Association, 28*(1-2), 83-94.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161-1178. doi:10.1037/h0077714

Sander, D., Grandjean, D., & Scherer, K. R. (2005). A systems approach to appraisal mechanisms in emotion. *Neural Networks, 18*(4), 317-352.

Schachter, S. (1964). The interaction of cognitive and physiological determinants of emotional state. *Advances in experimental social psychology, 1*, 49-80.

Scherer, K. R., & Ekman, P. (1984). On the nature and function of emotion: A component process approach. *Approaches to emotion, 2293*, 317.

Scherer, K. R., Schorr, A., & Johnstone, T. (2001). Appraisal processes in emotion: Theory, methods, research: Oxford University Press.

Scherer, S., Glodek, M., Layher, G., Schels, M., Schmidt, M., Brosch, T., . . . Palm, G. (2012). A generic framework for the inference of user states in human computer interaction. *Journal on Multimodal User Interfaces, 6*(3-4), 117-141.

Schlosberg, H. (1952). The description of facial expressions in terms of two dimensions. *Journal of Experimental Psychology, 44*(4), 229-237. doi:10.1037/h0055778

Schlosberg, H. (1954). Three dimensions of emotion. *Psychological Review, 61*(2), 81-88. doi:10.1037/h0054570

Schmidt, K. L., & Cohn, J. F. (2001). Human facial expressions as adaptations: Evolutionary questions in facial expression research. *American journal of physical anthropology, 116*(S33), 3-24.

Schroder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., . . . Pantic, M. (2012). Building autonomous sensitive artificial listeners. *IEEE Transactions on Affective Computing, 3*(2), 165-183.

Schuller, B., Rigoll, G., & Lang, M. (2003). Hidden Markov model-based speech emotion recognition.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., . . . Ungar, L. H. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE, 8*(9). doi:10.1371/journal.pone.0073791

Sebe, N., Lew, M. S., Sun, Y., Cohen, I., Gevers, T., & Huang, T. S. (2007). Authentic facial expression analysis. *Image and Vision Computing, 25*(12), 1856-1863.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). *Recursive deep models for semantic compositionality over a sentiment treebank.* Paper presented at the Proceedings of the conference on empirical methods in natural language processing (EMNLP).

Strapparava, C., & Mihalcea, R. (2007). *SemEval-2007 task 14: affective text*. Paper presented at the Proceedings of the 4th International Workshop on Semantic Evaluations, Prague, Czech Republic.

Strapparava, C., & Mihalcea, R. (2008). *Learning to identify emotions in text.* Paper presented at the 23rd Annual ACM Symposium on Applied Computing, SAC'08, Fortaleza, Ceara.

Sun, X., Lichtenauer, J., Valstar, M., Nijholt, A., & Pantic, M. (2011). A multimodal database for mimicry analysis *Affective Computing and Intelligent Interaction* (pp. 367-376): Springer.

Tan, S. C. G., & Nareyek, A. (2009). Integrating facial, gesture, and posture emotion expression for a 3D virtual agent.

Tian, Y.-L., Kanade, T., & Cohn, J. F. (2005). Facial expression analysis *Handbook of face recognition* (pp. 247-275): Springer.

Tomkins, S. S. (1962). Affect, imagery, consciousness: Vol. I. The positive affects.

Trilla, A., & Alías, F. (2009). *Sentiment classification in English from sentence-level annotations of emotions regarding models of affect.* Paper presented at the 10th Annual Conference of the International Speech Communication Association, INTERSPEECH 2009, Brighton.

Uldall, E. (1960). Attitudinal meanings conveyed by intonation contours. *Language and Speech, 3*(4), 223-234.

Valstar, M. F., & Pantic, M. (2007). Combined support vector machines and hidden markov models for modeling facial action temporal dynamics *Human–Computer Interaction* (pp. 118-127): Springer.

Valstar, M. F., Pantic, M., Ambadar, Z., & Cohn, J. F. (2006, 2006). Spontaneous vs. posed facial behavior: automatic analysis of brow actions.

Vibrations. (2016). GVC emotion recognition. Retrieved from http://www.good-vibrations.nl/innovations/emotionrecognition

Vicarvision. (2015). Facial action coding system : FaceReader 6.0. Retrieved from http://www.vicarvision.nl/facereader/productdescription/

Vogt, T., André, E., & Bee, N. (2008). EmoVoice—A framework for online recognition of emotions from voice *Perception in multimodal dialogue systems* (pp. 188-199): Springer.

Wagner, J., Kim, J., & André, E. (2005). From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. Paper presented at the Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on.

Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., & Cowie, R. (2008, 2008). Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies.

Yang, Y.-H., Lin, Y.-C., Cheng, H.-T., Liao, I. B., Ho, Y.-C., & Chen, H. H. (2008). *Toward multi-modal music emotion classification*.

Yin, L., Wei, X., Sun, Y., Wang, J., & Rosato, M. J. (2006). *A 3D facial expression database for facial behavior research*.

Yu, C., Aoki, P. M., & Woodruff, A. (2004). Detecting user engagement in everyday conversations. *arXiv preprint cs/0410027*.