

A Computational Approach to Emotion Recognition in Intelligent Agent

Tee Connie¹, Goh Ong Sing², Goh Kah Ong Michael, Kam Lean Huat
Faculty of Information Science and Technology
Multimedia University
Jalan Ayer Keroh Lama, 75450, Melaka

1.0 Abstract

The ability to express emotions in an intelligent agent is a very important factor in creating an interesting and compelling interaction between human and the agent. This paper describes a computational approach to identify emotional setting in the conversation between human and the agent. Natural language is the main interaction channel in our research as we have developed a natural processing system called AINI (Artificial Intelligent Neural-network Identity). Human user can chat with the agent in almost any topics of interest. The agent will respond with different expressions appropriate to the conversation, just like real human-being displaying different expressions when they are talking. An agent character with humanoid face is used as the interface to display expression in our system. With this, the conversation appears more real and natural because the user can receive emotional feedback from the agent. The user will be more willing to share information with the agent because the agent seems to think, feel and act like a human companion.

2.0 Keywords

Emotion recognition, artificial intelligence, natural language processing, intelligent agent, believable agent

3.0 Introduction

Recently, emphasis on human-computer interaction is moving towards the emotion recognition field to identify human emotion. This is because research has discovered that emotion plays a very important role in human-to-human interaction. It can be satisfying when a conversation partner can accurately trace the feeling and mind-set of a person. People may achieve a sense of being “in synch” or “on the same wavelength” with a person just because that person can catch up with his or her internal state of mind. Therefore, the computer could appear to be more intelligent by having the skills to recognize human emotions and appropriately adapt to the emotional respond during the interaction. Most important of all, it can make people attach to it if it is able to create an image that it “understands” people and show a sense of “understanding” towards the person.

Many researches have been conducted to recognize emotions through speech and facialextraction method [1]. They collect a number of samples of human speech with different

¹tee.connie@mmu.edu.my

²osgoh@mmu.edu.my

emotions and extract the pitch patterns from the distinguishing emotional speech. Researchers in National University of Singapore try to analyze facial expressions through video images and speech through audio files in order to classify emotions [5]. They analyze the pitch contour and also video emotion output obtained by asking subjects to perform different emotional outbursts for each of the basic emotions. Apart from that, some researchers from Microsoft.com also try to classify emotions based on emotions derived from speech signal [14]. They find a set of coefficients which contain information of the features that could reflect the emotional state of a speaker.

However, we found that there are some limitations with both of these emotion recognition systems. One major problem of recognizing emotion through speech is that different people show variety of accent and pitch when they talk. Male and female also talk at different pitch and tones. Thus it is very difficult to find a standard speech signal or pattern to classify the different emotions. For emotion recognition through facial expression, there is problem because people react differently to different type of emotions. While speaking the same content, people may have significantly different facial expressions depending on their current internal state of mind. Some people do not even show changes in their facial expression as they do not want to reveal their internal emotional state.

That is why we want to use a different approach to identify emotions. We propose to identify and classify emotions from the *context* of conversation. We are able to do this as we have developed a natural processing system called Artificial Intelligent Neural-Network Identity (AINI), which allows human to chat with the agent in natural language [8, 9, 10]. AINI is able to interpret human speech and generate proper respond to steer the dialog as appropriate. AINI is based on ALICE (Artificial Linguistic Internet Computer Entity), the entertainment chatterbot created by Dr. Wallace in 1995. ALICE won the 2000 and 2001 Loebner Prize, a restricted Turing Test to evaluate the level of “humanity” of chatterbots [8]. The concept of communication between human and the agent through AINI is depicted by Figure 1.

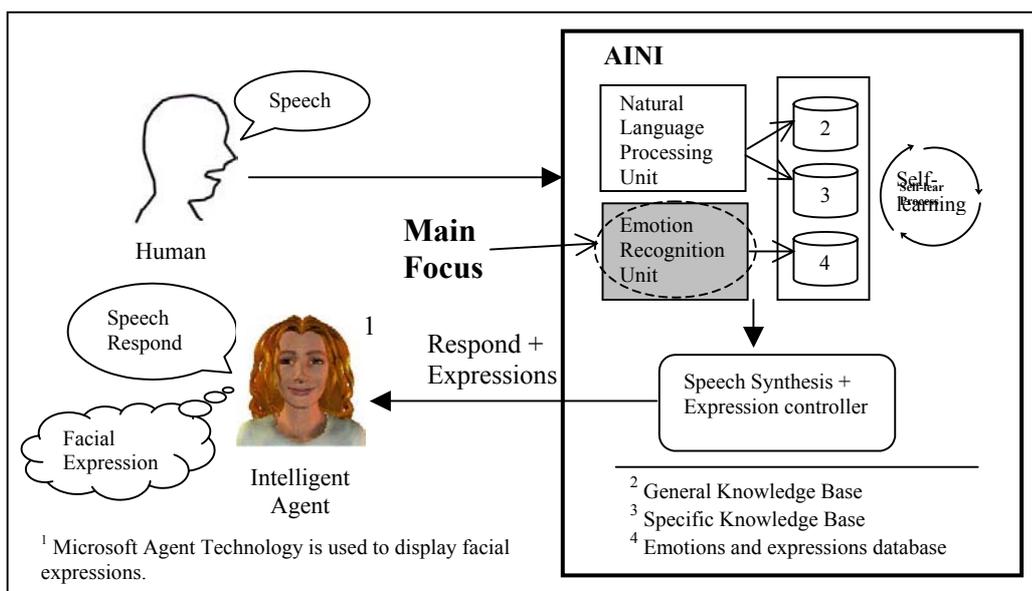


Figure 1: Communication between agent and human through AINI.

From the diagram above, human speech will be passed to the natural processing unit in AINI for analysis and processing. Proper response will be generated by the natural processing unit by extracting the knowledge stored in the database. The emotion recognition unit is responsible to identify the emotion found in the speech and instructs the agent to display appropriate facial expression. The interaction process can become more interesting and compelling as the agent is able to show different expressions when it talks. For example, the agent will display a happy face to greet the user when the conversation starts; it will display a sad face when it hears something miserable; and it will show an angry face when the user says some obscene words. Research is converging to the important implications of Affective Computing [2]. Therefore we are working hard to create a “believable agent” by giving it internally an “artificial heart” that can exhibit emotions and personality, and externally a model expression augmented with quasi-human characteristics. It is strongly believed that it has been the portrayal of emotions that has given agent characters the illusion of life. If the agent can feel about what happens in the world, that means it cares about the world, and this makes us care about it. An emotionless character is lifeless, and it can merely be treated as a machine [3].

4.0 Architecture

We propose a system to identify human’s emotion based on the context of their speech. When they talk, the keyword that conveys emotion will be captured so that the agent is able to display a facial expression to adapt to the emotion. We are using a humanized face from Microsoft Agent technology as our interface to show expression. In this way, the agent can appear to “think”, “feel” and “live”. Like real humans, the agent will change its expression when the topic of conversation changes. When the topic is a happy one, the agent will display a happy face; when it is a sad one the agent will display a sad face. This is very important because when the agent is able to change expression according to the conversation, people will pay more attention to it as it appears to be very “real” and “attentive” and make people want to share information with it, as with a friend or companion. The example below illustrates an example of conversation between human and the agent. The agent character beside every line of chat refers to the facial expressions shown when the agent talks.

<p>A : Agent U : User</p>	 Agent Character
<p>A: Thanks for connecting. What is your name? <i>(Agent initiates the conversation. The default expression is a happy face)</i> U: My name is Mary.</p>	
<p>A: Hello Mary. How are you feeling today? <i>(Agent obtains the user’s name and continues asking question. The happy expression remains)</i> U: I feel very sad today.</p>	
<p>A: I am very sorry to hear that. What makes you so sad, Mary? <i>(The keyword sad is captured and the agent displays a sad expression)</i> U: I did not do well in my exam.</p>	

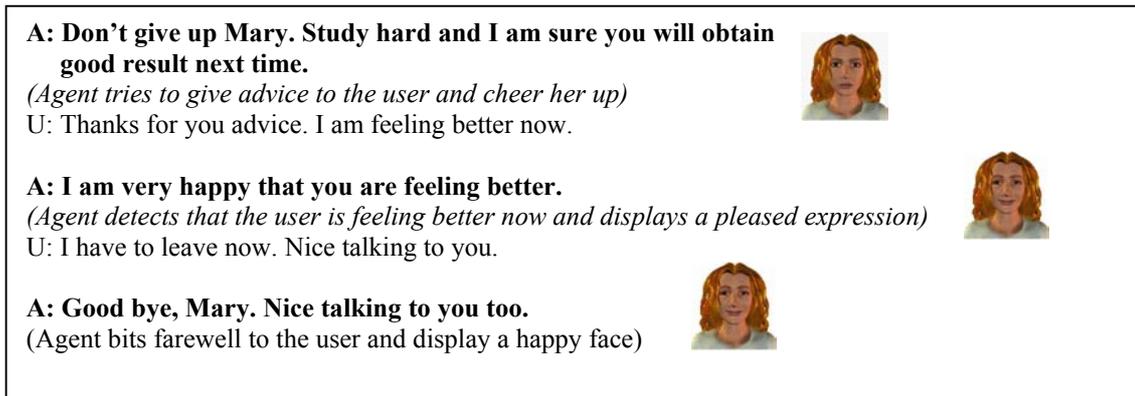


Figure 2: Example of chat between human user and agent character.

The diagram below shows the architecture of the emotion recognition system:

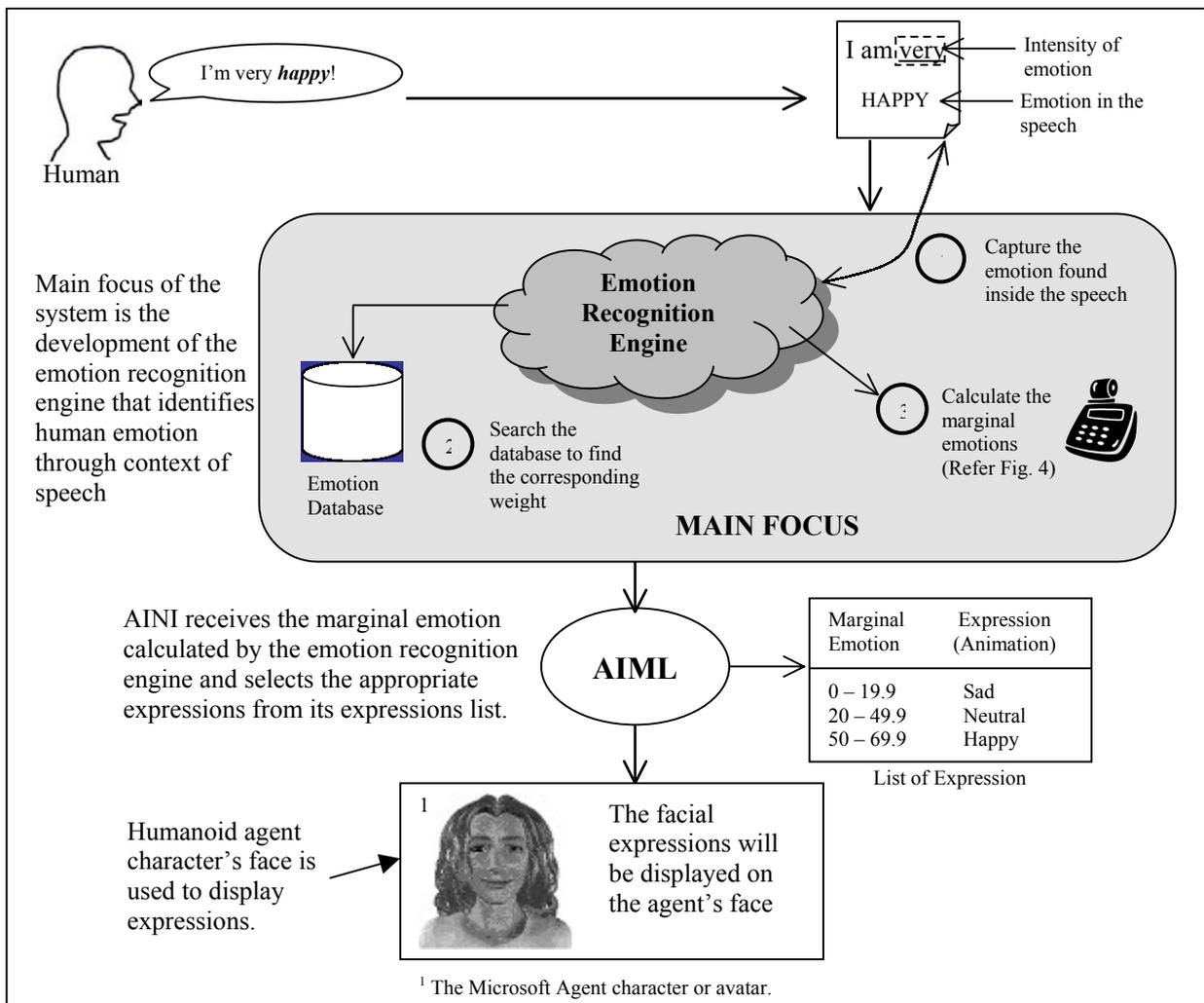


Figure 3: Conceptual diagram of emotion recognition system.

The emotion recognition engine will first identify the conversational context of the conversation to classify the emotional setting of the conversation. Then it will analyze the human speech and try to capture the emotions found in the speech. In the example provided in the diagram above, the keyword that shows emotion in the speech is “HAPPY”, with intensity “very”. After that, the engine will search the emotion database to find the corresponding weight for “HAPPY”. An emotional database is built to store all the words that convey emotions (such as happy, pleased, glad, and joyful) with varying weight. For example, the keyword happy will carry weight 50, whereas the keyword sad will carry weight 5. Extensive research has been carried out to collect all the words that convey emotions and careful consideration is placed to assign weight to these words. As the database will be very large, a good compression and searching technique (the Offline Pattern Matching) will be used to search the large emotions vocabulary so that the searching time will be less.

As speech can convey a few emotions at a time, the marginal or “net” emotions will be calculated by using the algorithm shown below. The words which show intensity, such as “very”, “quite”, “almost”, “extremely”, also carry some weight and they will also be counted. Another parameter which is included is the weight for conversational context. The weight of the conversational context varies according to its emotional setting. Please refer to example in Figure 6 for better understanding of the application of this algorithm. This algorithm can be expressed as

$$M_{\xi}(s_k) = \left[\sum_{i=1}^n [W_{\xi_i} * I_{\xi_i}] + f(N) \right] / n$$

Where,

$M_{\xi}(s_k)$ = Marginal emotion, ξ , due to the kth state of the world

W_{ξ_i} = The weight of the ith emotion, ξ .

I_{ξ_i} = Intensity related to the ith emotion.

$f(N)$ = Function that captures the conversational context, N.

Figure 4: Marginal emotion algorithm

The marginal emotion obtained from this algorithm will be passed to AINI to select the appropriate facial expression for that emotion. The <agplay> tag in AIML (Artificial Intelligence Markup Language) is responsible to animate the agent’s expression. AIML is represented in an XML specification, which is able to express the artificial intelligence concept elegantly. The <agplay> tag is responsible to add more attractive Microsoft Agent character expressions in the AIML answer.

Recent research has identified five main emotions in human: normal, happy, sad, afraid, and angry [19]. We are using only these five expressions in our system for research purpose. We have a distribution list that stores the five expressions and their corresponding marginal emotion, for example, marginal emotion from -20 – 29.9 corresponds to expression “neutral”, 30 – 79.9 for “angry” expression and others. We represent the five main emotions with the Microsoft agent character shown in figure 5.

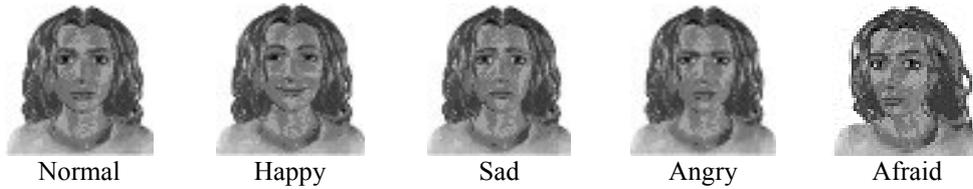


Figure 5: Various expressions shown by the avatar.

We provide an example to better illustrate our idea. In this example, the human user is talking to the agent about his examination result.

User says: "I am very *happy* today. But at the same time I feel *sad*".

First of all, context of the conversation is identified. The conversational context in this example is about examination. Different contexts carry different emotions. For example, when the context of conversation is about natural catastrophe, the emotional setting will be sad; whereas when the context of conversation is about leisure pursuit, the emotional setting will be happy. The emotional setting can change when the topic of conversation changes, therefore the emotion recognition engine must be aware of contextual change of topic all the time, so that the agent can change its emotion according to the contextual changes. In our example, the emotional setting for this topic (human user's examination) should be neutral, as it does not involve any emotional state. After that, the emotional engine can extract the keywords that convey emotions found in the speech. In this example the keywords are "happy" and "sad", and the intensity for "happy" is "very".

After obtaining the conversational context, keywords that convey emotion and the corresponding intensity, the emotion recognition engine searches the emotion database to find the matching weights for them.

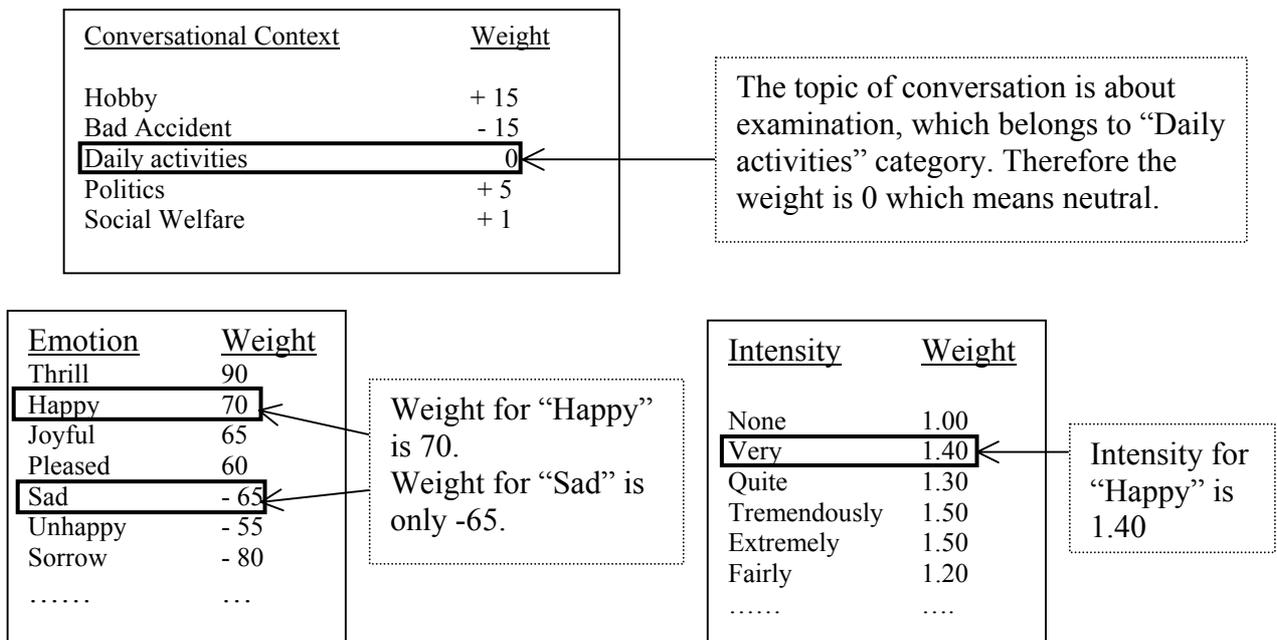


Figure 6: The emotional database.

The emotional database contains all the words that convey emotions such as happy, joyful, pleased, glad, sad, dejected, gloomy, depressing and others. Each of these emotions carries some weight depending on their degree of “emotioness”. For instance, the word *thrilled* shows more degree of “happiness” than *pleased*. Therefore the weight for *thrilled* is 90 and *pleased* is 60. The word *sorrow* shows more degree of “sadness” than *unhappy*, so the weight for *sorrow* is -80 and *unhappy* is -55. The context of conversation also carries some weights: happy topic will carry more weight than sad topic. Topic about a birthday party celebration will carry more weight than a topic about a fatal accident. We assign some weights to the conversational context because it also affects the human speech. Human emotion can’t go very far away from the emotional setting of the topic, for example we can’t feel too happy when we are talking about a fatal accident. Therefore we have decided that conversational context also carries some weight.

After obtaining the weight of the emotion, intensity and conversational context, the marginal or net emotion found in the speech can be calculated as follows:

$$M_{\xi}(s_k) = \left[\sum_{i=1}^n [W_{\xi_i} * I_{\xi_i}] + f(N) \right] / n$$

$$M_{\xi}(s_k) = [((70 * 1.4) + 0) + ((-65 * 1) + 0)] / 2$$

$$= 16.5 \text{ (Marginal emotion)}$$

After obtaining the marginal emotion, the corresponding facial expression is selected from a distribution list. The expression is passed to the <agplay> AIML tag so that the agent can display the corresponding facial expression when it speaks out the respond. In this example, the marginal expression is five which corresponds to a “neutral” expression. This sounds logical as human used to display neutral expression when he feel sad and happy at the same time.

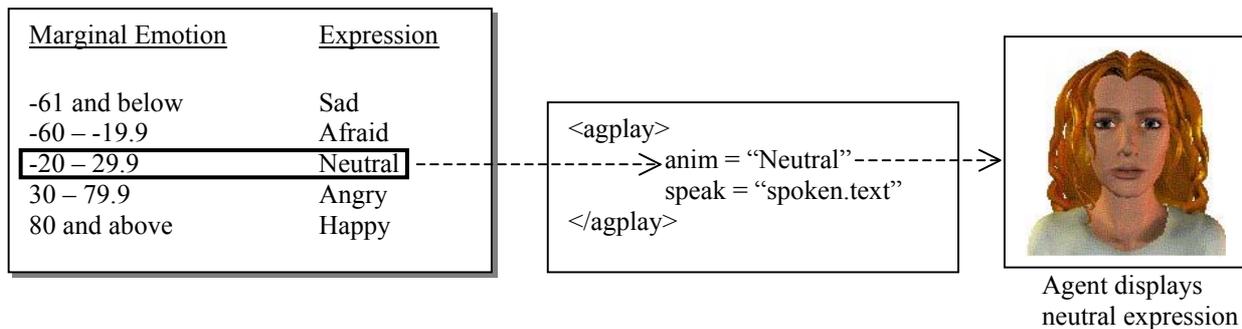


Figure 7: Emotion recognition system select the corresponding expression from the marginal emotion and pass to the <agplay> tag to display the expression on the agent character’s face.

If the speech does not have keyword that shows emotion, for example “I want to buy a comic book from the store”, the agent will maintain the previous expression although the topic conversational has been changed. For instance when the user says “Let us talk about football”, the agent will track this topic change and display a happy emotion because football is a stirring topic. There is no problem with context change within the conversation because AINI already has the function to keep tract of the topic change.

Careful consideration must be placed in the process of assigning weight for the emotions and also the context of conversation. We have prepared a few hypotheses to prove the precision of our recognition system. Below are some samples of the hypothesis for some of the expressions:

Hypothesis I:

Conversational context – Hobby (Playing Football)

Speech – “I am so excited that Brazil won the Cup eventually. But I hate the judge very much”.

Hypothesis: The agent should display “happy” expression.

Justification: We hypothesized that the agent should display a happy expression because the speech is dominated by the excitement of the winning of Brazil in the World Cup. The conversational context is related to a happy context and that is why the agent should display a happy face although the user mention about “hatred” in the speech.

Calculation: $[(85 * 1.4) + 15] + [(20 * 1.5) + 15] / 2 = 89.5$

Result: The marginal emotion is 89.5. Therefore the expression for the agent is happy.

Hypothesis II:

Conversational context – Social chatting

Speech – “You dummy machine. You do not understand a single thing”.

Hypothesis: The agent should display “angry” expression.

Justification: We hypothesized that the agent should display an angry expression because the user is speaking obscene words to the agent. In this situation, the conversational context does not have any emotional effects on the agent.

Calculation: $(48 * 1) + 1 / 1 = 49$

Result: The marginal emotion is 49. Therefore the expression for the agent is angry.

Hypothesis III:

Conversational context – Hobby (Movie)

Speech – “The movie is very scary. I had a nightmare after watching it”.

Hypothesis: The agent should display a “frightened” expression.

Justification: We hypothesized that the agent should display a frightened expression although the topic of discussion is about movie, which should be an enjoying one. However, because the user says that he is terribly scared by the movie, that is why the agent should display a “frightened” expression. In this case, the conversational context does not have any effect on the agent’s expression.

Calculation: $(-40 * 1.4) + 15 / 1 = -41$

Result: The marginal emotion is -41. Therefore the expression for the agent is afraid.

Hypothesis II:

Conversational context – Accident

Speech – “I feel very sad that Rave was admitted to the hospital”.

Hypothesis: The agent should display “sad” expression.

Justification: We hypothesized that the agent should display a sad expression because the conversational context is about an accident which relates to sad emotion. Moreover the user mentioned the word “sad” which add more degree of sadness to the emotional setting.

Calculation: $(-65 * 1.4) - 15 = -106$

Result: The marginal emotion is -106. Therefore the expression for the agent is sad.

Figure 8: Examples of emotion hypothesis according to the conversation context

Hypothesis provided above can justify our approach to identify appropriate facial expression for the agent. However many studies have to be carried out to make the emotion recognition system more accurate.

5.0 Conclusion

Emotion recognition research is not easy because understanding emotion is one of the most difficult area has been explored. No human can perfectly recognize others innermost emotions, and sometimes people cannot even recognize their own emotions. Therefore we can only develop an emotion recognizer based on what can be observed and reasoned about, that is the context of conversation in this case. However, there are many considerations needed to be taken into account. Ekman and colleagues acknowledge that even simply labeled emotions like “joy” and “anger” can have different interpretations across individuals within the same cultures. Therefore the weight we assigned to the emotions can vary to different subject. Sometimes, what is said is not what it means [7]. For example an utterance “Good” spoken in a harsh tone of voice is likely to infer that the speaker does not really mean it, in fact, he or she wishes to convey something else. “Please kindly open the door” can convey different meanings: politeness in UK but enforcement in USA. That is why we stress that our emotion recognition system is still in its preliminary stage and much effort needs to be put to enhance the system. In future, we will try to improve the system by combining another two methods: emotion recognition through vocal and facial measure. We hope to increase the accuracy of identifying human emotion by analyzing the human’s facial expression, his/her speech frequency and intensity, and also the content of his speech. With this “tri-model” emotion recognition system, we have higher confidence that the result obtained can closely describe the human’s emotion.

6.0 Reference

- [1] H. Sato, Y.Mitsukura, M. Fukumi, N.Akamatsu. Emotional Speech Classification with Prosodic Prameters by using Neural Network. *In proceedings of The Seventh Intelligent Information Systems Conference*, pages: 395 -398. Australian and New Zealand, 2001.
- [2] Helmut Prendinger, Mitsuru Ishizuka. Social Role Awareness in Animated Agents. *In proceedings of the fifth international conference on Autonomous agents*, May 28-June 1, 2001. Montreal, Quebec, Canada, 2001.
- [3] J. Bates. The role of emotion in believable agents. *Communications of the ACM*, pages: 37(7):122-125, 1994.
- [4] Jeffrey F. Cohn, Gary S. Katz. Bimodal Expressions of Emotion by Face and Voice. *In proceedings of the sixth ACM international conference on Multimedia: Face/gesture recognition and their applications*. Bristol, United Kingdom, 1998.
- [5] Liyanage C. De Silva, Pei Chi Ng. Bimodal Emotion Recognition. *In proceedings of Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pages: 332-335, 2000.

- [6] Michael Riley, William Byrane, Michael Finke, Sanjeev Khudanpur, Andrej Ljolje, John McDonouh, Harriet Nock, Murat Saraclar, Charles Wooters, George Zavaliagos. Stochastic pronunciation modeling from hand-labelled phonetic corpora. *Speech Communication* 29 (1999) 209-224, 1999.
- [7] Shinobu Kitayama and Keiko Ishii. World and voice: Spontaneous attention to emotional utterances in two languages. *Cognition & Emotion*, Vol 16(1), Jan 2002, pages: 29-59, 2002.
- [8] Loebner Prize, "Home Page of the Loebner Prize — The First Turing Test". URL: <http://www.loebner.net/Prizef/loebner-prize.html>
- [9] Goh Ong Sing, "Artificial Intelligent Neural-network Identity (AINI)–The Next Generation of the Virtual Advisor", MSC–Asia Pacific ICT Award (MSC-APICTA), 4-5 March 2002, Sri Pentas, TV3, Bandar Utama, Kuala Lumpur.
- [10] Goh Ong Sing, Teoh Kung Keat, "Intelligent Agent for E-Management" at IEEE International Conference on Artificial Intelligent in Engineering and Technology (ICAIET 2002)" Universiti Sabah Malaysia, Sabah on 17-18 June 2002.
- [11] Goh Ong Sing and Teoh Kung Keat, "Intelligent virtual doctor system" at 2nd IEE Seminar on Appropriate Medical Technology for Developing Countries on 6th February 2002, London, UK. More information at URL: <http://www.iee.org/Oncomms/pn/healthtech/06Feb.cfm>
- [12] Rosalina W. Picard, Elias V, and Jennifer Healey. Towards Machine Emotional Intelligence: Analysis of Affective Physiological State. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 23, Issue 10 (October 2001), 2001.
- [13] Wataru Tsukahara, Nigel Ward, Responding to Subtle, Fleeting Changes in the User's Internal State. *In proceedings of the SIGCHI conference on Human factors in computing systems 2001*. Seattle, Washington, United States, 2001.
- [14] Yan Li, Feng Lu, Ying-Qing Xu, Eric Chang, Heung-Yeung Shum. Speech-Driven Cartoon Animation with Emotions. *In proceedings of the ninth ACM international conference on Multimedia*, 2001. Ottawa, Canada, 2001.
- [15] Yasmine Arafa, Abe Mamdani. Virtual Personal Service Assistants: Towards Real-time Characters with Artificial Hearts. *In proceedings of the 5th international conference on Intelligent user interfaces 2000*. New Orleans, Louisiana, United States, 2000.