

Emotional States Based 3-D Fuzzy Atmosfield for Casual Communication between Humans and Robots

Zhen-Tao Liu, Fang-Yan Dong, and
Kaoru Hirota

Department of Computational
Intelligence and Systems Science
Tokyo Institute of Technology
Yokohama, Japan
E-mail: {liuzhentao, tou, hirota}
@hrt.dis.titech.ac.jp

Min Wu, Dan-Yun Li
School of Information Science
and Engineering
Central South University
Changsha, P. R. China
E-mail: min@csu.edu.cn

Yoichi Yamazaki
Department of Electrical, Electronic,
and Information Engineering
Kanto Gakuin University
Yokohama, Japan
E-mail: yamazaki@kanto-gakuin.ac.jp

Abstract—Emotional states based three-dimensional (3-D) Fuzzy Atmosfield (FA) is proposed to express the feeling between humans and robots in communication, and is built with 3-D coordinates, i.e. “friendly-hostile”, “lively-calm”, and “casual-formal” axes. The FA aims to be an effective tool to proceed with the communication paying attention to the atmosphere generated by individual emotional state which is calculated from bimodal communication cues namely emotional-speech and emotional-gesture by using weighted fusion and fuzzy inference. A novel emotion recognition approach is presented which consists of two steps, i.e., classification of six basic emotions for initial emotional states in the Affinity-Pleasure-Arousal emotion space, and emotional transition from prior state by using fuzzy logic. Home party enjoying demonstration confirms FA’s availability empirically by questionnaires from unrelated persons of authors’ group, where smooth communication between four humans and five eye robots is realized in Mascot Robot System.

Keywords- human-robot communication; emotion recognition; emotional state; communication atmosphere; fuzzy inference

I. INTRODUCTION

Affecting computing, playing a significant role in human-robot interactions, has been paid more and more attention, where how to identify atmosphere amongst communicators as well as attitude, mood, and emotion of interlocutors, and how to provide a comfortable environment to eliminate the barriers between humans and robots, have become the main objectives for the next generation of intelligent robots.

Emotion recognition, a popular topic in recent years, has attracted increasing attention from researchers involved in a variety of discipline e.g. psychology, computer science, linguistics, neuroscience, and related disciplines [1]. Multiple channels of information during communication are used to analyze emotional states of humans including speech, facial expression, body gesture, posture, and so on [2-5]. Few studies, however, concentrate on communication atmosphere which is easily felt but hard to define, and even more difficult to evaluate, and it is never finished, static or at rest [6, 7]. In [8],

human expressive communication is analyzed by exploiting interrelation between speech and gestures. A 3-D communication atmosphere is built by using audio-visual modals, where only face-to-face communications between two participants are considered in recognition [7]. Most of previous researches concentrate on analysis of personal mood without understanding the atmosphere which is composed of all emotional states of humans in the communication.

Fuzzy Atmosfield is proposed to describe mood states among communicators being generated from not only the verbal and non-verbal information of speakers, but also background music or noise. Bimodal information based emotion recognition is carried out in two steps: 1. Emotion of the interlocutor is classified on the basis of six primary emotions, i.e. “happiness”, “surprise”, “fear”, “anger”, “disgust”, and “sadness”, which are predetermined as the initial emotional states in Affinity-Pleasure-Arousal emotion space; 2. Emotional state’s transition is executed by using fuzzy logic on speech features including volume, average frequency, and interval time. Fuzzy inference and expert rules are employed to build Fuzzy Atmosfield from individual emotion spaces, and a 3-D space defined as “friendly-hostile”, “lively-calm”, and “casual-formal” axes is presented to express Fuzzy Atmosfield. Experiments for Fuzzy Atmosfield are performed with four humans and five eye robots in a household environment by using Mascot Robot System. The results based on questionnaires verify availability of proposed methods.

The emotion recognition based on bimodal communication cues is mentioned in II. In III, emotional states based Fuzzy Atmosfield by using fuzzy inference and expert rules is presented. The experiments for Fuzzy Atmosfield in Mascot Robot System are given in IV.

II. EMOTION RECOGNITION BASED ON BIMODAL COMMUNICATION CUES

A. Overview of the Model’s Architecture

Emotion recognition based on bimodal information, i.e. emotional speech and gesture, is proposed to express the emotional states of interlocutors in Affinity-Arousal-Pleasure space. Fig. 1 shows the procedure of the proposed emotion recognition method, where there are two steps.

- Initial states of emotion are determined by fusing semantic cues from speech and the gestures of communicators.
- Starting from the initial/previous state of emotion, transition of emotional states in the 3-D emotion space is executed by using fuzzy inference considering the influence from features of speech.

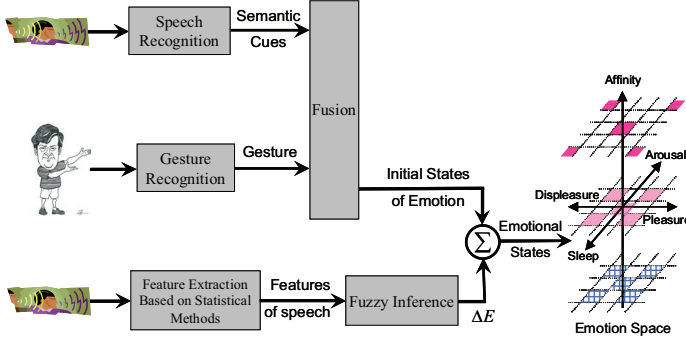


Figure 1. Emotion recognition based on bimodal information

B. Affinity-Pleasure-Arousal Emotion Space

A 3-D emotion space named Affinity-Pleasure-Arousal space is used to illustrate the mentality state of speakers during communication. By using the Affinity-Pleasure-Arousal plane, it is not only possible to express fixed mentality states but also to take into account rapid variations in the mentality state due to time involving communication [9], as shown in Fig. 2.

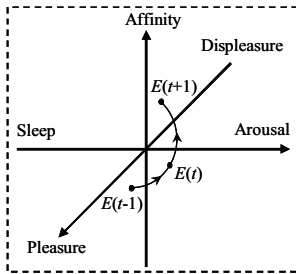


Figure 2. Emotion space expressed in 3-D axes

Affinity-Pleasure-Arousal emotion space is defined with fuzzy domains that can be indicated as

$$E = (e_{affinity}, e_{pleasure}, e_{arousal}), \quad (1)$$

$$\forall e_{affinity} \in [-1, 1], \forall e_{pleasure} \in [-1, 1], \forall e_{arousal} \in [-1, 1]$$

where $e_{affinity}$, $e_{pleasure}$, $e_{arousal}$ are the values in “affinity”, “pleasure-displeasure”, and “arousal-sleep” axes, respectively. The membership functions for these three axes are shown in Fig. 3.

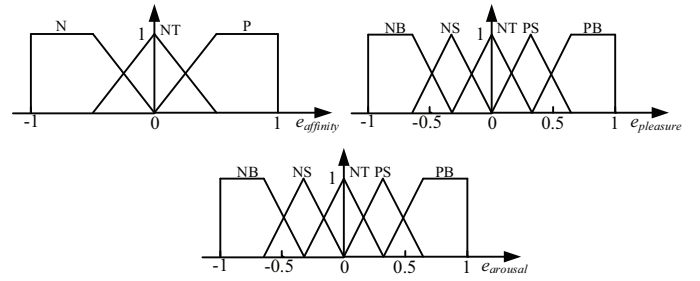


Figure 3. Membership functions for emotion space axes

C. Semantic Cues from Emotional Speech

Automatic speech recognition (ASR) system is developed for general robot systems by NEC Corporation in the Development Project for a Common Basis of Next-Generation Robots led by the NEDO organization. Its functions include single word speech recognition, noise canceller, sound source directional detection, speaker independent recognition and dictionary switching, i.e. it can select a specific dictionary fitting with current conversation topic. The speech recognition principle in the ASR is a word-matching search based on pre-defined dictionary.

Traditionally, research on the recognition of emotion from semantic cues of speech was focused on the discovery and utilization of emotional keywords, that is, specific words that express the speaker’s emotional state [10]. To take advantage of the function of single word recognition in ASR, emotional keywords from speech are selected for emotion recognition.

Six primary emotions [11], i.e., “happiness”, “surprise”, “fear”, “anger”, “disgust”, and “sadness”, are employed as the initial state of emotion recognition. The contribution from any emotional keyword w to one of the six basic emotion e_i is defined as $P(e_i|w)$ which is calculated in a training phase by its frequency of occurrence under the observation of the emotion on the basis of large emotional speech corpus.

D. Emotional Gesture Recognition

Although semantic cues of speech, to some extent, could reflect emotional state, the variety and complexity of language makes it difficult for researchers to recognize emotional states from pure textual information. Emotional-gesture, another available modal in communication, is necessary adopted for emotion recognition, which can originate from any bodily motion or state but commonly originate from the hand or face.

A multi-modal information based hand gesture recognition system is developed by authors’ group [12, 13], where emotional gestures are recognized based on multi-modal sensing, i.e., the combination of the 3-D acceleration sensor and web cameras. The gesture recognition system includes two modules, i.e. the acceleration sensor recognition module and the image recognition module. After fusing above two independent module’s results, emotional gestures are recognized finally. In the same way as emotional keywords, the contribution from any gesture g to one of the six basic emotion e_i is defined as $P(e_i|g)$ which is calculated in the data training of a large gesture database.

E. Emotion Recognition by Weighted Fusion

Weighted fusion is one kind of techniques for decision-level fusion, which arrives at the final classification decision by combining the weights of individual recognition. The basic idea behind this approach is with the hypothesis that the overall recognition rate will be improved by combining pieces of evidence provided by independent sources.

Information fusion of bimodal communication cues, i.e., semantic cues of speech and gesture is executed by using weighted fusion. Through calculating the sum of each emotion probabilistic from keywords of speech and gestures, the emotion e_i , receiving the maximum of probabilistic contribution, will be output as the final recognition result of emotion, as shown in Fig. 4. If it fails to receive only one kind of emotion, then jump to the recalculate procedure.

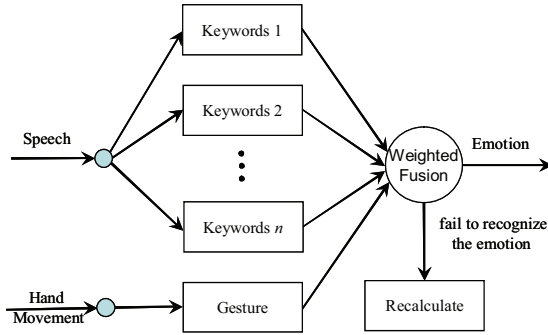


Figure 4. Emotion recognition by weighted fusion

Emotion recognition based on bimodal information outputs one of the six primary emotions as the initial coordinate in Affinity-Pleasure-Arousal emotion space, where is predetermined according to the relationship between six primary emotions and three axes of the emotion space.

F. Transition of Emotional States in Emotion Space

To recognize the real emotional states of communicators, features of speech such as volume, average frequency, and interval time are assigned to inputs for transition of emotional states. Fuzzy inference method is accepted to establish the relationship between voice features and the variation of three axes from previous position in Affinity-Pleasure-Arousal emotion space, as shown in Fig. 5.

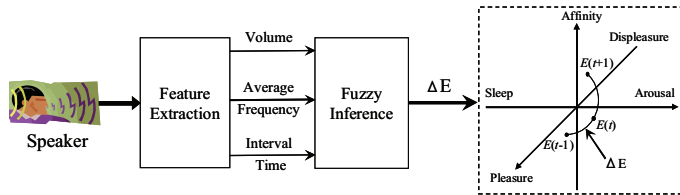


Figure 5. Emotion space of speaker using fuzzy inference

After completion of the feature extraction module, the ratio of two voices' volume, ratio of average frequency between speaker's and standard, and interval time between two utterances are used for fuzzy inference in transition of

emotional states. The membership functions for these features are shown in Fig. 6.

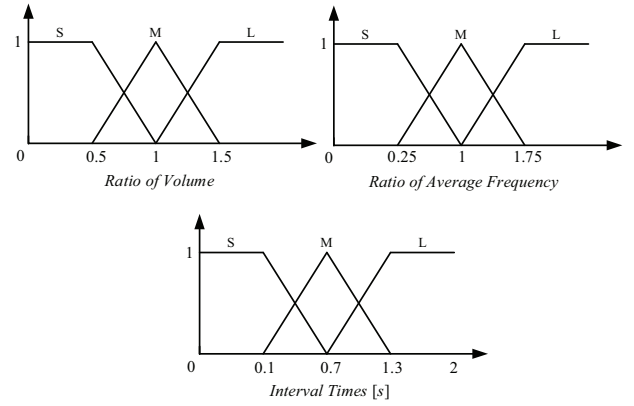


Figure 6. Membership functions for speech features

The fuzzy rules of ΔE , namely $\Delta e_{\text{affinity}}$, $\Delta e_{\text{pleasure}}$ and $\Delta e_{\text{arousal}}$ based on speech features are shown in Table I, Table II, and Table III, respectively.

TABLE I. FUZZY RULES OF $\Delta E_{\text{AFFINITY}}$

		Ratio of Average Frequency		
		S	M	L
Ratio of volume	S	N	P	P
	M	N	NT	P
	L	N	P	P

TABLE II. FUZZY RULES OF $\Delta E_{\text{PLEASURE}}$

		Ratio of Average Frequency		
		S	M	L
Ratio of volume	S	N	P	P
	M	N	NT	P
	L	N	P	P

TABLE III. FUZZY RULES OF $\Delta E_{\text{AROUSAL}}$

		Interval Time		
		S	M	L
Ratio of volume	S	NT	NT	N
	M	P	NT	N
	L	P	P	NT

To obtain the values of $\Delta e_{\text{affinity}}$, $\Delta e_{\text{pleasure}}$ and $\Delta e_{\text{arousal}}$ in Affinity-Pleasure-Arousal space, the center of gravity method is used as a defuzzification method, and then $E(t) = E(t-1) + \Delta E$.

The transfer of emotional states is one continuous process in the 3-D emotion space, as the smooth curve connecting $E(t-1)$, $E(t)$, and $E(t+1)$ in Fig 1. For instance, when the speaker stops talking, emotional state is in the process of diminishing but not at origin (neutral state).

III. FUZZY ATMOSFIELD BASED ON EMOTIONAL STATES

Fuzzy Atmosfield is invisible but it is supposed to exist by occupying space and containing energy. It reflects the mood states of communication between speakers, and affects the emotional states of humans as well. Thus, Fuzzy Atmosfield is not just a natural phenomenon, but also one kind of “psychological field”, associated with individual emotional states, background music, and noises, etc. In this paper, only emotional states are concerned.

A. Fuzzy Atmosfield described in a 3-D Space

Atmosfield (Atmosphere Field, i.e., atmos’ field) exists in the space surrounding us. It is created in the process of communication between speakers, producing “psychological feeling” that can give influence to the process and results of humans’ behavior. Due to its special characteristic of “psychological field”, it could not be calculated as accurately as some physic’s fields such as electrical field and magnetic field, and therefore, is defined as Fuzzy Atmosfield.

Fuzzy Atmosfield is illustrated in a 3-D space defined by “friendly-hostile”, “lively-calm”, and “casual-formal” axes, which is determined by questionnaires from 10 peoples on 20 different occasions [14], as shown in Fig. 7.

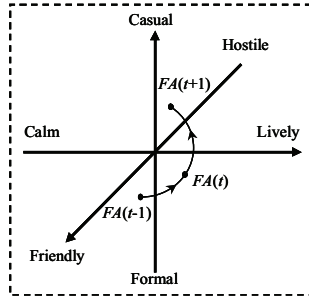


Figure 7. Fuzzy Atmosfield expressed in 3-D axes

To construct Fuzzy Atmosfield, the proposed 3-D space with fuzzy domains is described firstly as

$$FA = (a_{friendly}, a_{lively}, a_{casual}),$$

$$a_{friendly} \in [-1, 1], a_{lively} \in [-1, 1], a_{casual} \in [-1, 1] \quad (2)$$

where $a_{friendly}$, a_{lively} , a_{casual} are the values in “friendly-hostile”, “lively-calm”, and “casual-formal” axes, respectively. The membership functions for Fuzzy Atmosfield’s axes are shown in Fig. 8.

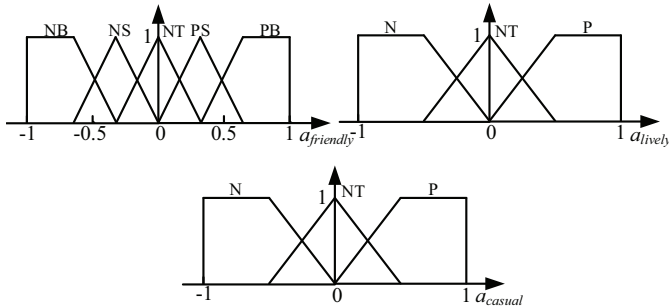


Figure 8. Membership functions for Fuzzy Atmosfield axes

B. Scheme of Fuzzy Atmosfield

The scheme of Fuzzy Atmosfield based on emotional states of communicators is presented as shown in Fig. 9. First of all, the emotional states of all participants in the conversation have been obtained in emotion recognition system (for humans) or emotion synthesis system (for robots). Secondly, information fusion module of emotional states is executed, where fuzzy inference is employed to express the relationship between Fuzzy Atmosfield and emotion space, and with the utilization of speaker’s volume, some expert rules are arranged to realize the mapping from n emotion spaces to Fuzzy Atmosfield. At last, the atmosphere state is given in the 3-D Fuzzy Atmosfield.

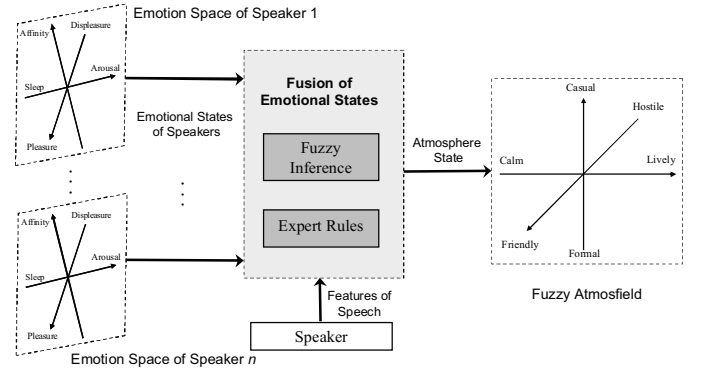


Figure 9. Scheme of Fuzzy Atmosfield

C. Emotional states based Fuzzy Atmosfield

As mentioned in II, Fuzzy Atmosfield is composed of individual emotion spaces. Meanwhile, the changing of atmosphere state is a continuous process as well as emotional state. In a word, at time t , the transfer of atmosphere state in Fuzzy Atmosfield is not only related to the atmosphere state at time $t-1$, but also the new emotional states of communicators stimulated from external environment. Therefore, the formula of Fuzzy Atmosfield can be defined as

$$FA(t) = \begin{cases} f(E_1(t), \dots, E_n(t)), & t=1 \\ (1-\lambda)FA(t-1) \cdot \gamma + \lambda f(E_1(t), \dots, E_n(t)), & t>1 \end{cases} \quad (3)$$

where FA is the atmosphere state in Fuzzy Atmosfield; f is the function of emotional states E at time t ; n is the number of speakers involved in the communication; γ is the decreasing function of $FA(t-1)$, and γ is an undetermined function, $0 \leq \gamma \leq 1$; λ is the correlation factor, $0 \leq \lambda \leq 1$; When $t = 0$, FA is at initial state where is set beforehand.

Considering different contribution of individual emotional states to Fuzzy Atmosfield, weighted average method is accepted to construct function f in

$$f = \sum_{i=1}^n w_i \cdot defuzzy(E_i(t) \circ R), i = 1, 2, \dots, n \quad (4)$$

where R is the relationship between emotion space and Fuzzy Atmosfield, which is obtained by using fuzzy inference, and 75

fuzzy rules are arranged to build R ; $E_i(t)$ is fuzzy set of emotion space of i th communicator; the center of gravity method is used as a defuzzification method in *defuzzy*; w_i is the weight of i th emotion space contributing to the Fuzzy Atmosfield, $\forall w \in [0, 1]$, $\sum_{i=1}^n w_i = 1$. It can be illustrated as shown in Fig. 10.

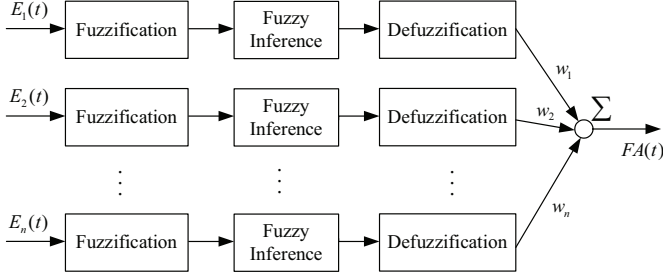


Figure 10. Flow chart from emotion spaces to Fuzzy Atmosfield

D. The weights of w_i

To analyze the contribution of each emotion space, the role of each participant for altering communication atmosphere should be considered, i.e., the actions such as speech, facial expression, and other body movements should be taken into account. In this paper, volume of speech is employed for computing the weights of each emotion space. At time t , w_i is calculated as

$$w_i = \frac{v_i}{\sum_{j=1}^n v_j}, \quad i = 1, 2, \dots, n, \quad j = 1, 2, \dots, n \quad (5)$$

where v is the volume of speaker.

E. The weight of λ

Considering the influencing factors for communication atmosphere, volume of speech is used to determine the correlation factor λ in equation (3). Some expert rules are formulated by using ratio of volume $v(t)$ and $v(t-1)$ as

$$ra_{volume} = \frac{\sum_{i=1}^n v_i(t)}{\sum_{i=1}^n v_i(t-1)} \quad (6)$$

where ra_{volume} is the ratio of volume $v(t)$ and $v(t-1)$, and

$$\lambda = \begin{cases} 0, & \text{if } \sum_{i=1}^n v_i = 0 \\ \frac{1}{2} \cdot ra_{volume}, & \text{if } ra_{volume} < 1 \\ \frac{1}{2}, & \text{if } ra_{volume} = 1 \\ 1 - \frac{1}{2} \exp\left(-\left(\frac{ra_{volume}-1}{\sigma}\right)^2\right), & \text{if } ra_{volume} > 1 \end{cases} \quad (7)$$

where σ is an undetermined parameter.

IV. EXPERIMENTS FOR FUZZY ATMOSFIELD

The proposed Fuzzy Atmosfield is evaluated by performing communication experiments with four humans and five eye robots in Mascot Robot System (MRS). The MRS is an information presentation system destined to be used mainly as an information terminal in a home environment for casual communication with humans [15]. The five eye robots include four fixed robots (placed on a TV, a darts game machine, an information terminal, and a mini-bar) and 1 mobile robot, as shown in Fig. 11.

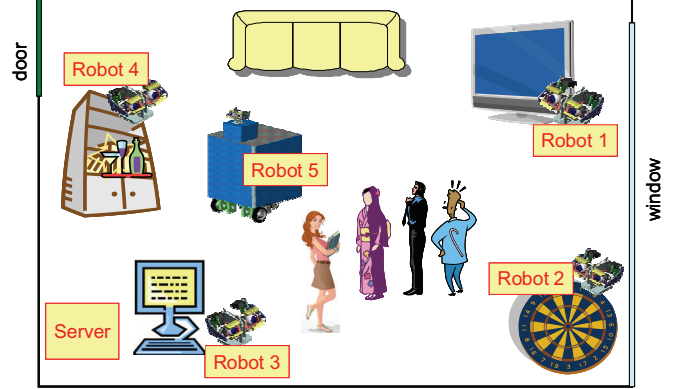


Figure 11. Household environment for humans and robots

An ideal household environment is created that is a relatively quiet place without background music or noises. Six scenarios are performed in the experiment, i.e., Scenario 1: “Greeting guests at the door”, Scenario 2: “Drinking at the bar”, Scenario 3: “Playing dart game”, Scenario 4: “Watching TV”, Scenario 5: “Checking timetable of train at the terminal computer”, Scenario 6: “Farewell with guests at the door”. Each scenario is designed with dialog between humans and robots, where eight types of typical human emotional gestures [11], including “Toast”, “Throw dart”, “Victory pose”, “Banzai pose”, “Put both hands on head and squat”, “Put hand on one’s face”, “Guiding”, and “Goodbye”, are performed as well. Fig. 12 shows the photo and dialog of Scenario 5, where a conversation is taken place between Host, Guest 1, Guest 2, and Robot 3 (the information terminal robot).



a) The photo of Scenario 5

Robot 3: [Which station?]
Host: Shibuya.
Robot 3: [Trains heading out of the city?]
Host: Yamanote line. To Shinagawa.
[Display PC shows the timetable.]
Guest 1: That's just right.
Guest 2: Thank you.
Robot: You're welcome.
Host: So let's go to the door.

b) The dialog of Scenario 5

Figure 12. The scenario 5 of “Checking timetable of train”

Emotional state of each person is obtained by proposed emotion recognition approach with high recognition rates, while robots’ emotion spaces are established in expression synthesis system [9] which is developed by the authors’ group.

To integrate individual emotional state of humans and robots, in the experiments, some parameters of Fuzzy Atmosfield's model are decided empirically, i.e., $\gamma = e^{0.1}$, $\sigma = 2$. The Fuzzy Atmosfield of Scenario 5 in home party is built by using proposed method, as shown in Fig. 13, where the coordinate of the sphere's center is $FA(t)$, and the diameter d which represents the variation of atmosphere state is calculated as

$$d = \min\left\{\frac{1}{\sqrt{3}}\sqrt{[a_{friendly}(t) - a_{friendly}(t-1)]^2 + [a_{lively}(t) - a_{lively}(t-1)]^2 + [a_{casual}(t) - a_{casual}(t-1)]^2}, 0.1\right\}. \quad (8)$$

The results of Fuzzy Atmosfield based on questionnaires by 10 peoples verify availability of proposed methods. Smooth communication between 4 humans and 5 eye robots in home party proceeds with harmonious atmosphere, where robots can recognize the atmosphere state in real time. And then, through classification, identification, and response, robots could adjust talking contents and tones to improve communication atmosphere.

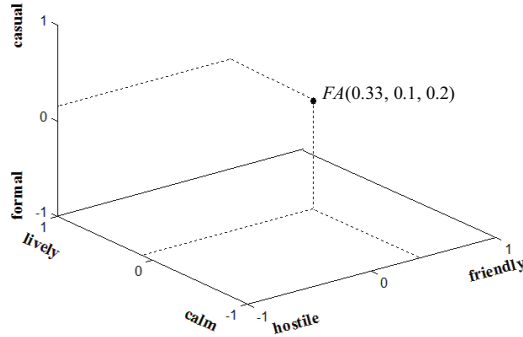


Figure 13. Fuzzy Atmosfield of Scenario 5 in home party

V. Conclusions

To express affective states during human-robot communication, the concept of Fuzzy Atmosfield is proposed, where it is composed of individual emotional states of humans and robots.

On the basis of bimodal communication cues including speech message and gesture message, first emotion classification is carried out in accordance with six basic emotions which are predetermined as the initial states in Affinity-Pleasure-Arousal emotion space; second transition of emotional state is executed by using fuzzy logic on some speech features. With the fusion idea of emotional states, Fuzzy Atmosfield is created by using fuzzy inference and expert rules. Being convenient to observe for users, a 3-D space with "friendly-hostile", "lively-calm", and "casual-formal" axes is introduced to express Fuzzy Atmosfield. Results of experiment demonstrate the availability of Fuzzy Atmosfield in a home party with smooth communication proceeded between 4 humans and 5 eye robots.

The proposed Fuzzy Atmosfield aims to realize casual communication in multi-human and multi-robot environments, not only integrating the mood expression of individuals, but

also providing good guidance for robots to improve communication atmosphere by adjusting the content of speech, improving the way of communication, or playing harmonious music, and so on.

In-depth study on assignment of appropriate parameter in the model of Fuzzy Atmosfield will continue in the further researches. In addition, more verbal or non-verbal information except speech and hand gesture would be considered to construct Fuzzy Atmosfield, for example, facial expression, head movements, and other body languages. Furthermore, the background music or noise would be essential factors for Fuzzy Atmosfield.

REFERENCES

- [1] Z. Zeng, et al, "A Survey of Affect Recognition Methods--Audio, Visual, and Spontaneous Expressions," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31, No. 1, pp. 39-58, 2009.
- [2] T. L. Pao, Y. T. Chen, J. J. Lu, J. H. Yeh, and W. Y. Liao, "Detecting Emotions in Mandarin Speech," International Journal of Computational Linguistics and Chinese Language Processing, Vol. 10, No. 3, pp. 347-362, 2005.
- [3] Y. Wang and L. Guan, "Recognizing Human Emotional State from Audiovisual Signals," IEEE Transactions on Multimedia, Vol. 10, No. 5, pp. 936-946, 2008.
- [4] M. L. Song, et al, "A Robust Multimodal Approach for Emotion Recognition," Neurocomputing, Vol. 71, No. 10, pp. 1913-1920, 2008.
- [5] M. J. Han, J. H. Hsu, and K. T. Song, "A New Information Fusion Method for Bimodal Robotic Emotion Recognition," Journal of Computers, Vol. 3, No. 7, pp. 39-47, 2008.
- [6] B. Anderson, "Affective Atmospheres," Emotion, Space and Society, Vol. 2, No. 2, pp. 77-81, 2009.
- [7] T. M. Rutkowski and D. P. Mandic, "Modelling the Communication Atmosphere: A Human Centered Multimedia Approach to Evaluate Communicative Situations," AI for Human Computing, Berlin, Springer Press, pp. 155-169, 2007.
- [8] C. Busso, "Multimodal Analysis of Expressive Human communication: Speech and Gesture Interplay," PhD thesis, University of Southern California, 2008.
- [9] Y. Yamazaki, Y. Hatakeyama, F. Y. Dong, and K. Hirota, "Fuzzy Inference based Mentality Expression for Eye Robot in Affinity Pleasure-Arousal Space," Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol. 12, No. 3, pp. 304-313, 2008.
- [10] C. H. Wu, Z. J. Chuang, and Y. C. Lin, "Emotion Recognition from Text Using Semantic Labels and Separable Mixture Models," ACM Transactions on Asian Language Information Processing, Vol. 5, No. 2, pp. 165-182, 2006.
- [11] P. Ekman, "Are There Basic Emotions?" Psychological Review, Vol. 99, No. 3, pp. 550-553, 1992.
- [12] Y. Yamazaki, et al, "Gesture Recognition Using Combination of Acceleration Sensor and Images for Casual Communication between Robots and Humans," IEEE World Congress on Evolutionary Computation, 2010.
- [13] Y. Tang, et al, "Multimodal Gesture Recognition for Mascot Robot System Based on Choquet Integral Using Camera and 3D Accelerometers Fusion," Journal of Advanced Computational Intelligence and Intelligent Informatics, unpublished.
- [14] K. Fukuda, "Conversation Atmosphere Recognition using Fuzzy Inference on Speech Information," Master Thesis, Tokyo Institute of Technology, 2009.
- [15] K. Hirota and F. Y. Dong, "Development of Mascot Robot System in NEDO project," In Proc. 4th IEEE Int. Conf. Intelligent Systems, pp. 38-44, 2008.