Communication Atmosphere in Humans and Robots Interaction Based on the Concept of Fuzzy Atmosfield Generated by Emotional States of Humans and Robots

Submitted: 6th September 2012; accepted 6th December 2012

Zhen-Tao Liu, Min Wu, Dan-Yun Li, Lue-Feng Chen, Fang-Yan Dong, Yoichi Yamazaki, and Kaoru Hirota

Abstract:
Communication atmosphere based on emotional states of humans and robots is modeled by using Fuzzy Atmosfield (FA), where the human emotion is estimated from bimodal communication cues (i.e., speech and gesture) using weighted fusion and fuzzy logic, and the robot emotion is generated by emotional expression synthesis. It makes possible to quantitatively express overall affective expression of individuals, and helps to facilitate smooth communication in humans-robots interaction. Experiments in a household environment are performed by four humans and five eye robots, where emotion recognition of humans based on bimodal cues achieves 84% accuracy in average, improved by about 10% compared to that using only speech. Experimental results from the model of communication atmosphere based on the FA are evaluated by comparing with questionnaire surveys, from which the maximum error of 0.25 and the minimum correlation coefficient of 0.72 for three axes in the FA confirm the validity of the proposal. In ongoing work, an atmosphere representation system is being planned for casual communication between humans and robots, taking into account multiple emotional modalities such as speech, gesture, and music.

Keywords: *human-robot interaction, communication atmosphere, fuzzy logic, emotion recognition*

1. Introduction

Affective computing has attracted more interest in the field of human-robot interaction (HRI), where how to identify atmosphere as well as emotion, and how to create a comfortable environment to eliminate the barriers between humans and robots, have become the main objectives for the next generation of intelligent robots [1]. To realize smooth communication between humans and robots, most of previous works focus on emotion recognition of humans [2] [3] and/or robot emotion synthesis [4] [5] in the context of communication involving few individuals, e.g., one human to one robot communication. In [6], Taki et al. discuss the Interactive Emotion Communication that requires robot to not only recognize human emotion, but also generate and express emotion of robot in response to the human emotion. For many humans to many robots interaction, however, perceiving only the emotions is not enough to achieve smooth communication since various human emotions that exist at the same time make it difficult to obtain a consensus in generating robot emotion. To solve this problem, a concept of the FA is introduced to represent the communication atmosphere generated by all the emotional stats of individuals [7] [29], which could be used to represent the atmosphere and further facilitate smooth communication in humans-robots interaction as well.

Communication atmosphere in HRI is modeled based on emotional states of humans and robots by using the FA [7] [29]. In addition, human emotion recognition based on bimodal information, i.e., speech and gesture, is proposed, which is carried out in two steps. In the first, emotion of the interlocutor is classified into seven categories, i.e., "happiness," "surprise," "fear," "anger," "disgust," "sadness," and "neutral," which are used as the initial emotional states in Affinity-Pleasure-Arousal (APA) emotion space [8]; Second step involves the transition of emotional state, which is calculated using fuzzy logic, and acoustic features including volume, average fundamental frequency F0, and interval time are used as the inputs. The emotional states of robots can be pre-determined by robot emotion synthesis such as emotional speech and gesture synthesis [4], and facial expression synthesis [9]. To integrate the emotional states of humans and robots, the FA that is a tool for quantitative analysis of the communication atmosphere in a 3D space with "Friendly-Hostile," "Lively-Calm," and "Casual-Formal" axes [7], [29] is used, where fuzzy inference is introduced to map APA emotion space into the FA, and speech feature such as volume is used to calculate the weights of individual emotional states for constructing the FA.

The communication atmosphere that is the overall affective expression in a gathering is estimated by intelligently and quantitatively integrating the emotions of humans and robots, taking each ambiguous influence on the atmosphere into consideration. According to the real-time communication atmosphere, the robots could adjust their emotion expressions to the atmosphere by self-learning in order to realize smooth communication, for example, robots could adapt to the atmosphere in a short response time. Moreover, to obtain the emotional states of humans, the proposal takes advantage of emotional cues in both speech and gesture, better improved than only using speech for emotion recognition.

Experiments in a household environment with six scenarios are performed by four humans and five eye robots using Mascot Robot System (MRS) [10] which is a typical application of humans-robots interaction. In the experiments, "happiness," "surprise," "anger," "sadness," and "neutral" are used to express human

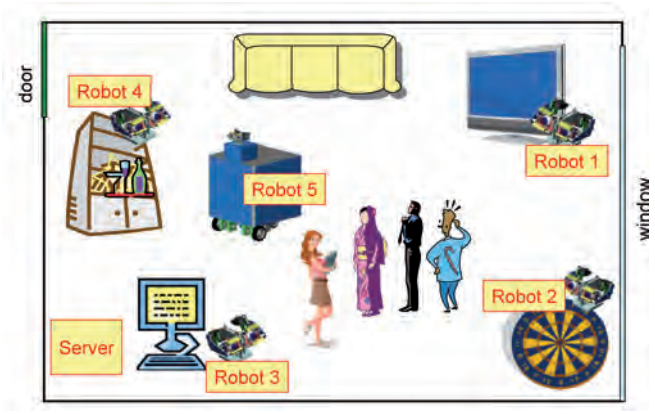


Fig. 1. *Mascot Robot System in a home party environment [29]*

emotion which is generated by speech and gesture. For emotion expression of eye robot, "happiness," "sadness," and "surprise" are used, and they are generated by emotional expression synthesis which controls eyelid motion and ocular motion [8]. The human emotion recognition is confirmed by using the results both from speech recognition [1], [11] and gesture recognition [12]. Experimental results from the model of communication atmosphere based on the FA compared with the results from subjective questionnaires verify the feasibility of the proposal.

Overview of the MRS developed for humans-robots interaction as well as quantitative representation of the communication atmosphere is mentioned in 2. Emotion recognition of humans based on bimodal cues and emotion synthesis of eye robot are presented in 3. In 4, communication atmosphere based on emotional states of humans and eye robots is modeled using the FA. Experiments on human emotion recognition and the FA based model of communication atmosphere generated by humans and eye robots are given in 5.

2. Mascot Robot System

Mascot Robot System, that is an information presentation system for casual communication between humans and robots, is an application of multi-human to multi-robot interaction developed as a part of the "Development Project for a Common Basis of Next-Generation Robots" supported by NEDO Society for the Promotion of Science [10].

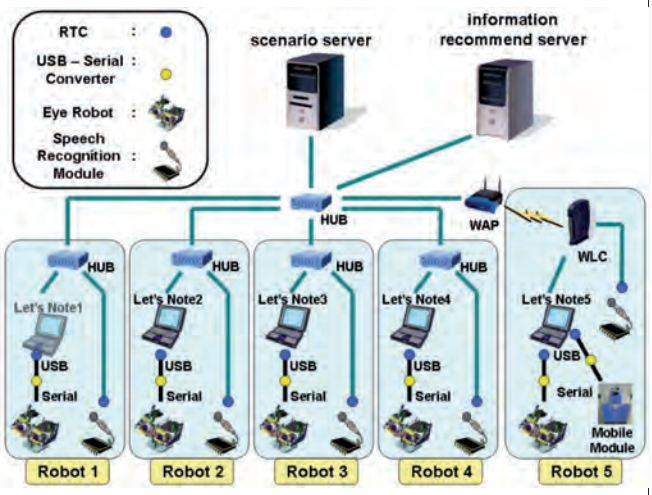


Fig. 2. *Block diagram of the Mascot Robot System*

The MRS in a household environment such as a living room, where there are four fixed type robots that are set up on furniture and appliances, e.g., the TV, dart game machine, information retrieval computer, mini-bar, and one self-propelled type mobile robot is shown in Fig. 1. Each robot is controlled by a laptop computer. The overall management server controls and supervises the whole system, allowing it to coordinate the actions of robots. The server connects the four fixed robots using a local area network via LAN cables, while it connects the mobile robot using a wireless LAN. The block diagram of the MRS is presented in Fig. 2.

Robot Technology Middleware (RTM) is used to construct the network system which is called RTM-Network [11]. In the RTM-Network, each robot can be viewed as a network component and each function unit of the robot is called Robot Technology Component (RTC). There are eight modules in the RTM-Network, including the speech recognition module (SRM), gesture recognition module (GRM), eye-robot control module (ERCM), emotion processing module (EPM), display module, mobile robot control module, scenario server module, and information retrieval module [12]. The EPM receives the recognized results from the SRM and the GRM, and then transmits them to the server module for the further processing, such as the calculation of individual emotions and communication atmosphere. According to the human emotions and atmosphere, the server module transmits eye movement instructions to the ERCM for expressing corresponding emotions of eye robots through designated actions.

For the MRS, communication atmosphere as well as individual emotion plays an important role in facilitating smooth communication between humans and robots. Furthermore, the communication atmosphere is more important for multi-human to multi-robot interaction, since it not only reflects overall emotional expression but also impacts on the emotions or feelings of individuals.

3. Emotional States of Humans and Eye Robots

3.1 Affinity-Pleasure-Arousal Emotion Space

Emotion space is commonly used as a tool for quantitative analysis of emotion, and it can express emotional states in two dimensions, three dimensions, and so on. To express the emotional states of humans and eye robots in the MRS, a 3D emotion space, i.e., Affinity-Pleasure-Arousal emotion space [8] is employed, as shown in Fig. 3. Using APA emotion space, it is not only possible to express fixed emotional states but also takes in consideration rapid variations of the emotional state due to time involving communication [8].

The emotional state in APA emotion space is defined as

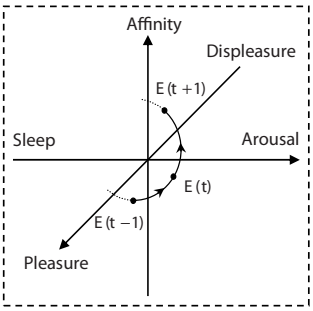


Fig. 3. *Affinity-Pleasure-Arousal emotion space [8]*

$$E = (e_{affinity},\ e_{pleasure},\ e_{arousal}), \quad \forall e \in [-1,\ 1], \qquad (1)$$

where E is the emotional state, and $e_{affinity}$, $e_{pleasure}$, and $e_{arousal}$ are the values on "Affinity," "Pleasure-Displeasure," and "Arousal-Sleep" axes, respectively.

"Affinity" axis represents degree of empathy, i.e., intimacy (positive affinity) and estrangement (negative affinity); "Pleasure-Displeasure" axis means happiness (positive values) and sadness (negative values); and "Arousal-Sleep" axis expresses excitement (positive values) and calm (negative values) [13].

Variation of emotional states is assumed to be a continuous process in the 3D space, and emotional state is time-dependent, which can be described by $E(t-1)$, $E(t)$, and $E(t+1)$ in Fig. 3. In addition, it is assumed that the intensity of emotional state will not disappear instantaneously, for instance, when a speaker stops talking, it will decrease until finally it reaches the neutral state as time goes by.

3.2 Emotion Recognition of Humans Based on Bimodal Cues

To recognize human emotion, multiple channels of information during human to human interaction are used in previous works, including speech, facial expression, body gesture, posture, and so on [14-16]. In the MRS, frequent head movements make it more difficult to capture the facial expression of humans. Hence, emotion recognition based on bimodal information, i.e., speech and gesture, is proposed. Fig. 4 shows the procedure of the emotion recognition module of the MRS, mainly consisting of three steps:

Step 1: Initial state of emotion is recognized based on semantic cues from speech and the gestures performed by communicators using weighted fusion.

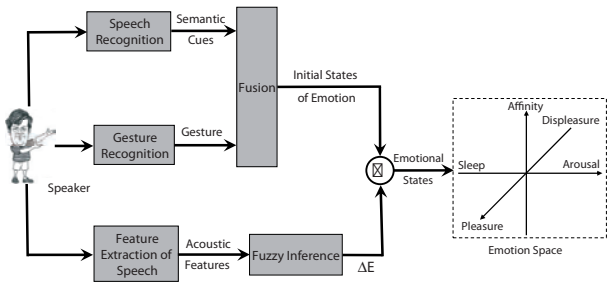


Fig. 4. *Emotion recognition based on bimodal information*

Step 2: Starting from the initial/previous state of emotion, transition of emotional states in APA emotion space is calculated based on acoustic features of speech by using a fuzzy inference system.

Step 3: At the sampling time t+1, if the scene changes in the MRS, then jump to Step 1, or else jump to Step 2 to calculate the transition of emotional state.

The number of sub-steps of Step 3 is based on duration of a scene, for example, there are five sub-steps in a scene that lasts ten seconds, in which the sampling period is two seconds. Thus, scene change is important for determining when to recalculate initial state of an emotion, i.e., jump to Step 1. And in one scene, emotional state can be calculated as

$$E(t) = \begin{cases} E_{Initial} + \Delta E & t = 1 \\ E(t-1) + \Delta E & t > 1 \end{cases}, \qquad (2)$$

where $E_{Initial}$ is initial state of emotion, ΔE is the transition of emotional state. And if $e_{i\text{Initial}} + \Delta e_i > 1$ or $e_i(t-1) + \Delta e_i > 1$, then $e_i(t) = 1$; if $e_{i\text{Initial}} + \Delta e_i < -1$ or $e_i(t-1) + \Delta e_i < -1$, then $e_i(t) = -1$, i="affinity," "pleasure," or "arousal."

Six primary emotions [17], i.e., "happiness," "surprise," "fear," "anger," "disgust," and "sadness," together with "neutral," are employed as the initial state of emotion in Step 1.

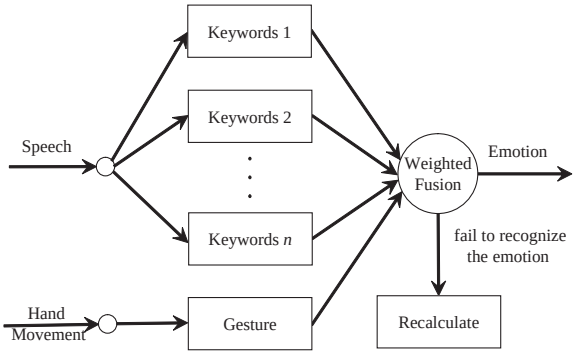


Fig. 5. *Emotion recognition by weighted fusion*

3.2.1 Semantic Cues from Emotional Speech

Recently, both acoustic and semantic analyses of speech are commonly used for emotion recognition [18] [19], where there is an assumption that talking words or phrases contain more or less emotional information. In order to reduce the complexity and computational cost of the SRM mentioned in 2, an open-source software – Julius [20] is adopted for speech recognition. The software contains major speech recognition techniques, and it can effectively perform a large vocabulary continuous and real-time speech recognition task [20].

Traditionally, emotion recognition using semantic cues of speech focuses on discovery and utilization of emotional keywords or phrases, that is, specific words that express and convey the speaker's emotion [21]. To take advantage of the function of single word recognition in SRM, emotional keywords or phrases from speech are selected for emotion recognition. The contribution from any emotional keyword/ phrase w

to one of the seven initial emotions $emotion_i$ is defined as $P(emotion_i|w)$, where $P\in[0, 1]$, $\Sigma P(emotion_i|w)=1$, i="happiness," "surprise," "fear," "anger," "disgust," "sadness," or "neutral." $P(emotion_i|w)$ could be calculated in a training phase by its frequency of occurrence under the observation of the emotion on the basis of large emotional speech corpus [22].

3.2.2 Emotional Gesture Recognition

Although semantic cues of speech could reflect emotional state to some extent, the variety and complexity of language makes it difficult for researchers to recognize emotional states from pure textual information [21]. Another available emotional cue, gesture, is involved in conveying inner emotions of humans [23].

An arm gesture recognition system for the MRS where emotional gestures are recognized based on multi-modal sensing, i.e., the information fusion of the data from 3D acceleration sensor and images from camera is developed by authors' group in [12]. The gesture recognition system includes two modules, i.e., the accelerometer based recognition module and the camera image recognition module. The final result of gesture recognition is calculated by using Fuzzy Choquet Integral to fuse the results of above two independent modules [12].

In the same way as emotional keywords, the contribution from any gesture g to one of the seven initial emotions $emotion_i$ is defined as $P(emotion_i|g)$, where $P\in[0, 1]$, $\Sigma P(emotion_i|g)=1$, i="happiness," "surprise," "fear," "anger," "disgust," "sadness," or "neutral."

3.2.3 Emotion Recognition by Weighted Fusion

Weighted fusion is one kind of technique for decision-level fusion, which arrives at the final decision of classification 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 [1] such as emotional keywords and gestures. Emotion recognition using weighted fusion of bimodal communication cues, i.e., semantic cues of speech and gesture is proposed as shown in Fig. 5, in which the weights of all the cues are assumed to be same, thus they are set to 1.

Through calculating the sum of each emotion probability based on the contributions from keywords or phrases of speech and gestures, the emotion $emotion_i$ with the maximum probability, will be output as the final result of emotion recognition as

$$Emotion = \arg\max_{i=1,\ldots,7}\{\sum_{j=1}^{N} P(emotion_i|w_j) + \sum_{k=1}^{M} P(emotion_i \mid g_k)\} \quad (3)$$

where *Emotion* is final result of emotion recognition that is one of the seven initial emotions. N is the number of emotional keywords or phrases. M is the number of performed gestures. If it fails to obtain only one kind of emotion as the recognition result, then it will jump to the recalculation procedure.

To determine the coordinates of each primary emotion in APA emotion space, forty people (aged 20 to 30 years, from seven countries, six female), have participated in questionnaire survey with eighteen questions – How "Affinity"/"Pleasure"/ "Arousal" do you feel about each emotion [13]? Seven answer options are given for each question, for example, to describe the "Arousal-Sleep" attribute, seven items that can be expediently chosen by the respondents are 1-Extremely Sleep, 2-Very Sleep, 3-Sleep, 4-Medium, 5-Arousal, 6-Very Arousal, and 7-Extremely Arousal. Each item is assigned with value for statistical analysis, i.e., 1 (-1), 2 (-0.66), 3 (-0.33), 4 (0), 5 (0.33), 6 (0.66), and 7 (1).

Since the answers are responded by different people who have different understanding for specific emotion, a certain range of coordinate values for each emotion exists. Cao et al. [24] suppose that each emotion has a certain space, e.g., "happiness" exists uncertain situations as very happy or a little happy which would occupy a subspace in the emotion space. In this paper, the six primary emotions are assumed to locate in APA emotion space with settled coordinates, in order to not only quantitatively analyze the human emotion but also further estimate communication atmosphere between humans and robots in the MRS. Thus, average coordinates from the questionnaire survey are adopted as the initial states, including "happiness" (0.58, 0.63, 0.54), "surprise" (0.27, 0.49, 0.58), "fear" (-0.33, -0.3, 0.21), "anger" (-0.67, -0.72, 0.65), "disgust" (-0.34, -0.59, 0.27), and "sadness" (-0.47, -0.44, -0.33).

3.2.4 Transition of Emotional States in APA Emotion Space

Primary emotions can be used to generate the other emotional state, which is similar to basic colors that are the basis for the whole spectrum of possible colors [25]. In other words, emotions are not limited to isolated categories but can be described along three nearly independent continuous dimensions in emotion space [26]. For example, E=(0.58, 0.63, 0.54) that is interpreted as a medium happiness can be changed by adding the transition ΔE according to variations in acoustic features such as volume and frequency. If each Δe is a small value, e.g., $0<\Delta e\leq 0.1$, a high happiness will be obtained, or $-0.1\leq\Delta e<0$, a low happiness will be obtained.

To calculate ΔE, acoustic features of speech such as volume, fundamental frequency F_0, and interval time are used. For mapping from these three acoustic features to the transition in three dimensions in APA emotion space, the ratio of a speaker volume at time t+1 to volume at time t defined as ra_{volume}, ratio of F_0 of a speaker to the standard (i.e., 120 Hz for male, 200 Hz for female, 300 Hz for children) defined as ra_{F0}, and interval time between two utterances defined as $t_{interval}$ are extracted. The relationship between these three acoustic features and the variation of three axes from previous position in APA emotion space is established by using fuzzy inference, as shown in Fig. 6.Membership functions for above acoustic features and $\Delta e_{affinity}$, $\Delta e_{pleasure}$, and $\Delta e_{arousal}$ are shown in Fig. 7, where three

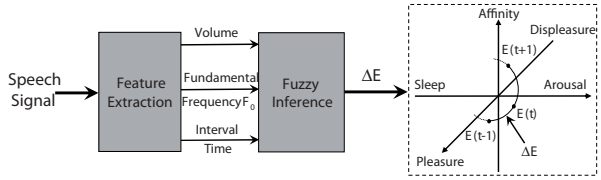


Fig. 6. *Transition of emotional state by using fuzzy inference*

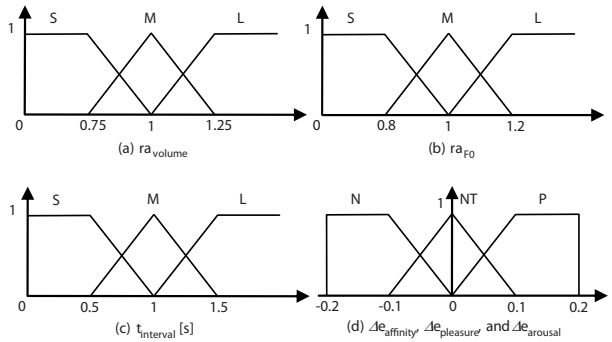


Fig. 7. *Membership functions for acoustic features of speech and DE*

linguistic values for acoustic features of speech, i.e., Small (S), Medium (M), and Large (L), and three linguistic values for Δe, i.e., Negative (N), Neutral (NT), and Positive (P) are formulated. Membership function of ra_{volume} is determined according to the level of volume, for example, 30 dB = whisper, 40 dB = quiet room, 50 dB = moderate rainfall, and 60 dB = typical conversation [27]. Using 40 dB as the baseline level, raising 10 dB is defined as Large and lowering 10 dB is defined as Small in Fig. 7 (a). Membership of ra_{F0} is determined based on the range of F_0, for example, F_0 of the male voice ranges from 100 to 150 Hz; F_0 ranges from 170 to 220 Hz for the female voice [28]. Compared with the standards mentioned above, increasing 20 Hz is defined as Large and decreasing 20 Hz is defined as Small in Fig. 7 (b). Membership of $t_{interval}$ is determined according to the interval in one-to-one communication, where continuous speech with less than 0.5 s interval is defined as Small, speaking with about 1 s interval is defined as Medium, and more than 1.5 s interval is defined as Large.

The fuzzy rules for calculating ΔE, namely, $\Delta e_{affinity}$, $\Delta e_{pleasure}$, and $\Delta e_{arousal}$ based on acoustic features are shown in Table 1, Table 2, and Table 3, respectively. The "IF-THEN" fuzzy production rules of fuzzy inference are designed by perceiving the relationship between the variations in acoustic features and three attributes of APA emotion space. For example, when ra_{volume} and ra_{F0} increase, feelings of affinity and pleasure are enhanced. To obtain the numerical values of $\Delta e_{affinity}$, $\Delta e_{pleasure}$, and $\Delta e_{arousal}$ in APA emotion space, the center of gravity method is used for the defuzzification.

Fig. 8. *Structure of an eye robot [8]*

If the number of sub-steps is large in Step 3 in 3.2, the algorithm presented will not converge to a fixed point in APA emotion space, because the transition ΔE is used to reflect the variations in emotional states. The emotional states could be located in a small sub space, e.g., a sub space of happiness, but not a fixed point.

3.3 Emotional Expression Synthesis of Eye Robots

Emotional expression of eye robots using APA emotion space is designed to realize casual communication between humans and robots in a household environment [8]. In response to conversation content and emotions of humans, eye robots are instructed to speak and perform eye movements, conveying preset emotions to serve human requirements or improve the quality of communication.

For the above purposes, an eye robot is developed for the MRS, and its structure is shown in Fig. 8. The eye robot is equipped with a pair of eyelids and a pair of eyeballs, where the eyelid has two degree of freedom (D.O.F.) and the eyeball has three D.O.F. [8]. Both eyelid motion and ocular motion constitute the overall eye movement, which could express the preset primary emotions such as "happiness," "sadness," "surprise," and "anger."

Eyelid motion and ocular motion are decided according to the emotional state E in APA emotion space. To determine the relationship between parameters for eye movements and location in the emotion space, twenty-five different partitions of the pleasure-arousal plane are assigned based on psychological knowledge [8], as shown in Fig. 9.

Eyelid motion of an eye robot is defined by position θ_l and frequency t_l, while ocular motion is defined by ocular position in pitch direction θ_p, yaw direction θ_y, and frequency t_o. A look-up table for each parameter of eye movement based on the Pleasure-Arousal plane is formulated [8] are presented in Table 4.

Table 1. Fuzzy rules for $\Delta e_{affinity}$

$\Delta e_{affinity}$		ra_{F0}		
		S	M	L
ra_{volume}	S	N	N	NT
	M	NT	NT	P
	L	NT	P	P

Table 2. Fuzzy rules for $\Delta e_{pleasure}$

$\Delta e_{pleasure}$		ra_{F0}		
		S	M	L
ra_{volume}	S	N	NT	NT
	M	N	NT	P
	L	NT	P	P

Table 3. Fuzzy rules for $\Delta e_{arousal}$

$\Delta e_{arousal}$		$t_{interval}$		
		S	M	L
ra_{volume}	S	NT	N	N
	M	P	NT	N
	L	P	P	NT

4. Communication Atmosphere based on Emotional States of Humans and Eye Robots

Communication atmosphere is invisible but it is supposed to exist by occupying space and containing energy [29]. On one hand it reflects the emotional states of communicators, and on the other hand it affects the emotional states of humans. The FA [7] [29] is designed as a tool to express the communication atmosphere, which is associated with individual emotions, background music, noises, and other atmosphere-related factors. In this paper, only emotional states of humans and robots are considered for estimating the communication atmosphere.

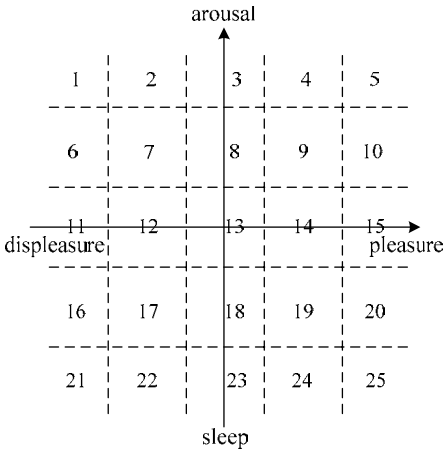


Fig. 9. *Twenty-five partitions of Pleasure-Arousal plane [8]*

4.1 Fuzzy Atmosfield Described in 3D Space

Atmosfield [29] – a contraction of *Atmos*phere and *Field* – is a concept that expresses the existence of communication atmosphere in the space surrounding us. Due to the ambiguity and uncertainty of the Atmosfield, it is further defined as *Fuzzy Atmosfield*.

The FA is described by 3D space with "Friendly-Hostile," "Lively-Calm," and "Casual-Formal" axes, as shown in Fig. 10.

The state of the FA in the 3D space is described as

$$FA = (a_{friendly},\ a_{lively},\ a_{casual}), \quad \forall a \in [-1,\ 1], \tag{4}$$

where *FA* is the state of the FA, and $a_{friendly}$, a_{lively}, and a_{casual} are the values on "Friendly-Hostile," "Lively-Calm," and "Casual-Formal" axes respectively.

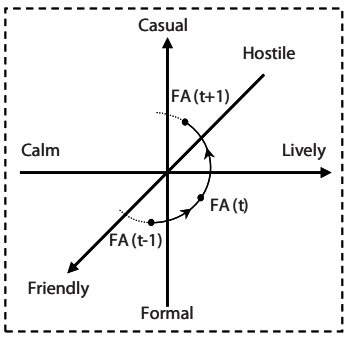


Fig. 10. *Fuzzy Atmosfield expressed in 3D space [7]*

Table 4. Look-up table for eye motion parameters based on Pleasure-Arousal plane [8]

θ_l[°]	value	150	120	90	60	30
	region	1~5,10,15	6~8,14	13	12,17~20	the rest
θ_p[°]	value	40	20	0	-20	-40
	region	5,10,15	3,4,8,9,14	13	2,6,12,17~20	the rest
θ_y[°]	value	40		20		0
	region	1,6,11,16,21		2,7,12,17,22		the rest
t_l[s]	value	2.5	2	1.2	1	0.5
	region	1~5	6~10	11~15	15~20	the rest
t_o[s]	value	0.1	0.3	0.4	0.5	1
	region	15~20	11~15	6~10	1~5	the rest

4.2 Model of Communication Atmosphere Based on the FA for Human-Robot Interaction

The framework of the FA based on emotional states of four humans and five eye robots in the MRS comprising of three steps [1] is presented in Fig. 11.

Step 1: Emotional states of humans are obtained by the proposed method for human emotion recognition, and emotional states of eye robots are pre-determined by emotional expression synthesis, respectively.

Step 2: The emotional states of humans and eye robots are integrated by an information fusion module, where a fuzzy inference system is used to express the relationship between APA emotion space and the FA, and expert rules are formulated for mapping from emotional states of four humans and five eye robots to the atmosphere state using the FA.

Step 3: The state of the communication atmosphere (i.e., state of the FA) is described in 3D space of the FA.

The model of communication atmosphere based on the FA [29] for humans-robots interaction in the MRS is defined as

$$FA(t) = \begin{cases} f(E_{H1}(t),\ \cdots,\ E_{H4}(t),\ E_{R1}(t),\ ...,\ E_{R5}(t)),\ t = 1 \\ (1-\lambda)FA(t-1)\cdot\gamma + \lambda f(E_{H1}(t),\ \cdots,\ E_{H4}(t), \\ E_{R1}(t),\ ...,\ E_{R5}(t)),\ t = 2,\ 3,\ ...,\ m \end{cases} \tag{5}$$

where f is the function of emotional states of humans $E_H(t)$ and eye robots $E_R(t)$; γ is a monotonically decreasing function which determines the decreasing speed of the intensity of $FA(t-1)$ when no new emotional element comes up, $0 \le \gamma \le 1$, for the experiments using the MRS, γ is set to $\exp(-0.1T)$ when all the emotional states E are located in the neighborhood of the origin, otherwise it is set to 1, and T is the sampling period for calculating the FA; λ is the correlation factor, $0 \le \lambda \le 1$.

Fuzzy logic and weighted average method are used to construct the function f in

$$f:\ \sum_{i=1}^{4} w_{Hi} \cdot defuzzy(\underset{\sim}{E}_{Hi}(t) \circ R) + \sum_{j=1}^{5} w_{Rj} \cdot defuzzy(\underset{\sim}{E}_{Rj}(t) \circ R) \tag{6}$$

where R is the relationship between APA emotion space and the FA, which is obtained by using a fuzzy inference system with 75 fuzzy rules [29], as shown in Fig. 12; $\underset{\sim}{E}(t)$ is a fuzzy set of the emotional state of individual; the center of gravity method is used as

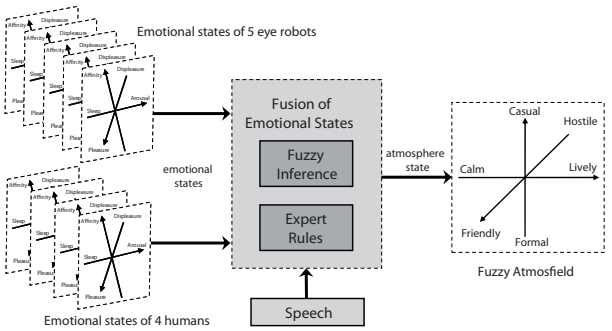


Fig. 11. *Framework of the FA based communication atmosphere in four humans to five eye robots interaction*

a defuzzification method in *defuzzy*; *w* is the weight of each emotional state contributing to the FA,

$$\forall w \in [0,\ 1],\ \sum_{i=1}^{4} w_{Hi} + \sum_{j=1}^{5} w_{Ri} = 1 .$$

In Fig. 12, the membership functions of APA emotion space and the FA are evenly distributed [29]. There are three linguistic values for the "Affinity" axis (i.e., **N**egative, **N**eu**t**ral, and **P**ositive), five linguistic values for "Pleasure-Displeasure" axis (i.e., **H**igh **D**ispleasure, **L**ow **D**ispleasure, **N**eu**t**ral, **L**ow **P**leasure, **H**igh **P**leasure) and "Arousal-Sleep" axis (i.e., **H**igh **S**leep, **L**ow **S**leep, **N**eu**t**ral, **L**ow **A**rousal, **H**igh **A**rousal), respectively. The seven linguistic values for each axis of the FA, are **E**xtremely **Fr**iendly, **V**ery **Fr**iendly, **Fr**iendly, **N**eu**t**ral, **H**ostile, **V**ery **H**ostile, and **E**xtremely **H**ostile for "Friendly-Hostile" axis; **E**xtremely **L**ively, **V**ery **L**ively, **L**ively, **N**eu**t**ral, **C**alm, **V**ery **C**alm, and **E**xtremely **C**alm for "Lively-Calm" axis; **E**xtremely **C**asual, **V**ery **C**asual, **C**asual, **N**eu**t**ral, **F**ormal, **V**ery **F**ormal, and **E**xtremely **F**ormal for "Casual-Formal" axis. The first letters of linguistic values as discussed here are shown in bold to highlight their use in Fig. 12.

In the MRS, when humans interact with eye robots, the main communication cues for emotional expression of humans are only speech and gesture, and for eye robots they are speech and eye movements. Speech is a mutual form of communication between humans and eye robots, in which the sound of speech influences on the atmosphere significantly [29], in other words, the sound determines the existence and state of an atmosphere to a great extent. Therefore, speech volume is employed for computing the weight w_i of each emotional state in Eq. (6) as

$$w_i = \frac{v_i}{\sum_{j=1}^{n} v_j},\ i = 1,2,\ ...,\ n,\ j = 1,2,\ ...,\ n, \qquad (7)$$

where *v* is the speech volume.

The correlation factor λ in Eq. (5) is calculated in

$$\lambda = \begin{cases} 0, & if \ \sum_{i=1}^{n} v_i(t) = 0 \\ 1, & if \ \sum_{i=1}^{n} v_i(t-1) = 0 \\ \frac{1}{2} \cdot ra_{volume}, & if \ ra_{volume} \le 1 \\ 1 - \frac{1}{2}\exp(-(\frac{ra_{volume} - 1}{\sigma})^2), & if \ ra_{volume} > 1 \end{cases}, \qquad (8)$$

where σ is a positive parameter determining the impact of ra_{volume} on λ, σ is set to 2 determined by the experiment, i.e., trial and error, and

$$ra_{volume} = \frac{\sum_{i=1}^{n} v_i(t)}{\sum_{i=1}^{n} v_i(t-1)} . \qquad (9)$$

5. Experiments on the Communication Atmosphere in Mascot Robot System

5.1 Experimental Environment

Using the MRS (Mascot Robot System) [10], the proposed model of the communication atmosphere based on the FA is evaluated by performing a home party experiment with four humans (i.e., Host, Guest 1, Guest 2, and Guest 3) and five eye robots, i.e., Robot 1 (TV), Robot 2 (dart game machine), Robot 3 (information terminal), Robot 4 (mini bar), and Robot 5 (mobile robot). An ideal household environment that is a relatively quiet place without background music or noises is created.

Six scenarios are designed in the experiment, Scenario 1: "Greeting guests at the door," Scenario 2: "Drinking at the mini bar," Scenario 3: "Playing darts," Scenario 4: "Watching TV," Scenario 5: "Querying information at the terminal computer," Scenario 6: "Farewell with guests at the door." Each scenario is designed with dialog between humans and robots. Fig. 13 shows the photo of Scenario 2, where a conversation between Host, Guest 1, Guest 2, Robot 4, and Robot 5 takes place when they are drinking at the mini bar. The dialog of Scenario 2 is shown in Fig. 14.

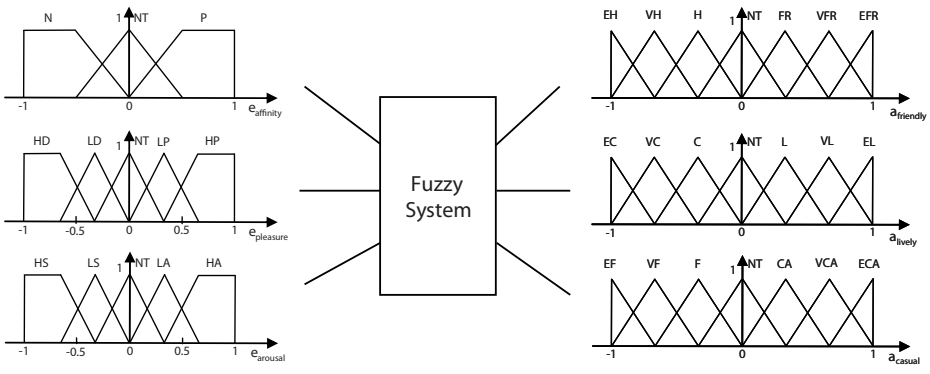


Fig. 12. *Fuzzy system for mapping Affinity-Pleasure-Arousal emotion space to Fuzzy Atmosfield*

Four participants (three male and one female, three Japanese and a Vietnamese, aged between 22 and 30) take part in home party scenarios. In the dialogue between humans and robots, there are seventy-four lines for human, in which sixty emotional keywords and phrases are used for expressing human emotions, and twenty-six lines for eye robots. Arm gestures are performed by four humans in the experiment, twenty-one times in total, where eight types of typical emotional gestures [12] are performed. The gestures include "toast" (G1), "throwing dart" (G2), "victory pose" (G3), "banzai pose" (G4), "squatting with hands over head" (G5), "face covering" (G6), "guiding" (G7), and "farewell" (G8). Three emotion categories are used for emotion expression conveyed by these gestures, i.e., "happiness" for G1, G3, G4, G7, and G8, "sadness" for G5 and G6, and "neutral" for G2.

All participants perform natural communication in speech and making a gesture. In addition, they are requested to make standard gestures; otherwise it will fail to obtain high accuracy of gesture recognition.

In the experiment, four of the six primary emotions mentioned in 3.2, i.e., "happiness," "surprise," "anger," "sadness," and "neutral" are used to represent human emotions. Since the purpose of the project of the MRS is to create a harmonious and friendly atmosphere between humans and robots, "anger" and "sadness" emotions are expressed only a few times, whereas "fear" and "disgust" emotions are not involved in any scenario.

The home party experiment is recorded using a SONY HDR-UX1 video camera located in the position where it can just capture all the humans and robots, and saved in AVI format at 720 480 pixels resolution at 30 fps rate. The sound track is recorded in 44100 Hz, 16 bit, and mono. Human gestures are recorded by wearable accelerometers manufactured by Microstone Inc. [12]. The accelerometer captures the acceleration data of gesture in x, y, and z axis, and transmits data to the server via a Bluetooth connection [12].

5.2 Experiments on Human Emotion Recognition

In the home party experiment mentioned in 5.1, natural communication takes place between 4 humans and 5 eye robots, in which human emotion is recognized based on speech and gesture, and humans are not asked to express over-exaggerated emotions, e.g., making specific facial expressions. Moreover, emotion recognition is carried out after one person finishes speaking, or after performing gestures – in general, gestures are performed by some one while he is speaking or after speaking.

Experimental results of emotion recognition are illustrated in Table 5 and Table 6, where the rows represent ground truth emotions and the columns show the emotion recognition results. The ground truth emotions are determined by carrying out questionnaire survey which is responded by twenty post-graduate students (aged between 22 and 30 years, from five different countries). The question is "which emotion does this performer express?", and the answer options include "happiness," "surprise," "anger," "sadness," and "neutral." The majority of choices are assigned as the ground truth emotions.

Fig. 13. *A photo of Scenario 2*

Emotion recognition based on speech obtains recognition rate of 74.36% in average, as shown in Table 5. Table 6 shows that emotion recognition based on speech and gesture obtains 84.16% accuracy in average by improving approximate 10% compared with that only using speech. Because the gestures performed in the experiment are not many and they expresses only three kinds of emotions i.e., "happiness," "sadness," and "neutral," experimental results of emotion recognition of "happiness" and "sadness" are improved significantly. In addition, in both Table 5 and Table 6, the "neutral" obtains the highest accuracy, since in the case of neutral emotion expression, no emotional keywords and phrases as well as emotional gestures are performed in the scenarios.

Table 5. *Confusion matrix of emotion recognition based on speech*

	Happiness	Surprise	Anger	Sadness	Neutral	Recognition rate
Happiness	33	0	0	0	11	75%
Surprise	0	9	0	0	2	81.8%
Anger	0	0	3	0	1	75%
Sadness	0	0	1	2	2	40%
Neutral	0	0	0	0	10	100%
Average recognition rate						74.36%

Table 6. Confusion matrix of emotion recognition based on bimodal cues

	Happiness	Surprise	Anger	Sadness	Neutral	Recognition rate
Happiness	37	1	0	0	6	84%
Surprise	0	9	0	0	2	81.8%
Anger	0	0	3	0	1	75%
Sadness	0	0	0	4	1	80%
Neutral	0	0	0	0	10	100%
Average recognition rate						84.16%

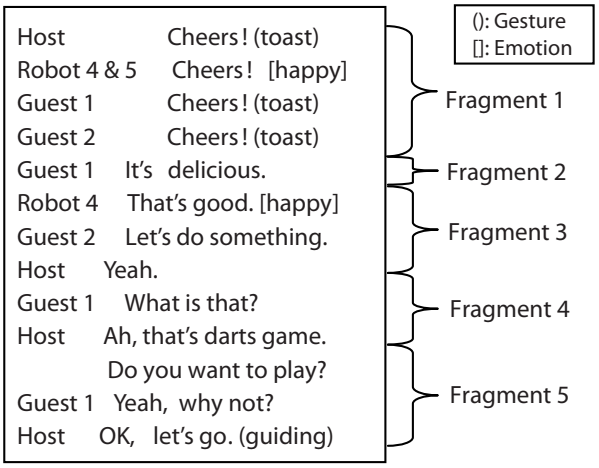


Fig. 14. *The dialog of Scenario 2*

In Table 5, eleven separate occasions of "happiness" emotion are recognized as "neutral," where actually no emotional keywords or phrases are involved in these dialogs, whereas participants express their own emotions by using the tone of voice which is related to the acoustic features of speech. With the addition of results from gesture recognition, recognition rate of "happiness" is improved by 9%, achieving 84%.

"Anger" and "sadness" are expressed only four times and five times respectively, which to some extent result in relatively low recognition rate. In addition, "anger" and "sadness" are easily confused with each other. For example, in Scenario 3, the Host failed to hit the dart target and said "Oh, *no*." Only using emotional keywords (i.e., *no*), the possibility of "anger" is higher than "sadness." In the case of using both keywords (i.e., *no*) and gesture (i.e., "face covering" that conveys "sadness" emotion), the "sadness" with higher possibility is output as the final recognition result.

5.3 Emotion Synthesis of Eye Robots

Emotion of eye robots are pre-determined and performed by eye movements as mentioned in 3.3 and imitation of human voices. The voice of each eye robot is designed as follow, Robot 1: "little boy;" Robot 2: "little girl;" Robot 3: "young woman;" Robot 4: "young man;" and Robot 5: "middle-aged man."

In the experiment, "happiness," "sadness," and "surprise" are used for emotion expression of eye robot. The "happiness" takes place more often than other two emotions, and the number of emotion expression of each eye robot is illustrated in Table 7. The mobile robot that accompanies the host when walking around the house, performs more emotional expressions than other fixed robots that only appear in a specific scenario.

Emotion of eye robot is conveyed by using both emotional keywords and eye movements. For example, in Scenario 2, the Host, Guest 1, and Guest 2 are toasting at the mini-bar, saying "*Cheers*." Robot 4 and Robot 5 respond "*Cheers*" which is an emotional keyword that represents "happy" emotion, and perform corresponding eye movement to express happy. As mentioned in 3.2.3, the "sadness" state in Affinity-Pleasure-Arousal space is estimated as (-0.47, -0.44, -0.33), which is located in the Part 21 in the Fig. 9. According to the look-up table in Table 4, the parameters for eye movements are θ_l=30°, θ_p=-40°, θ_y=40°, t_l=0.5s, and t_o=1s, as shown in Fig. 15 (b).

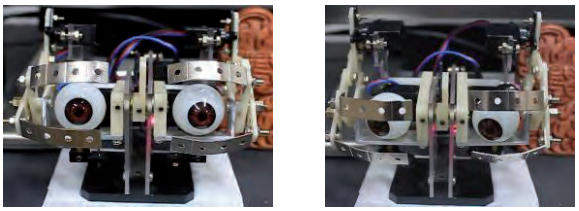
To make sure that eye robots are clearly exposed on video and help audiences perceive the emotions of eye robots, close-up views of eye robots are added to the video as shown in Fig. 13.

5.4 Experiments on the Communication Atmosphere using the FA in Human-Robot Interaction

Part of the experimental results has been given in [29], in which the home party experiment video is divided into thirty-nine fragments, and each fragment lasts six seconds (i.e., T = 6s in γ, in Eq.(5)) according to the exchange time of interlocutors in this experiment, namely, conversations between two or three individuals are made within six seconds so as to create the atmospheres. The state of the communication atmosphere in each fragment is estimated by integrating the emotional states of humans and eye robots, and the emotional state of human is calculated every two seconds (i.e., the sampling period for emotion recognition in Eq.(2)).

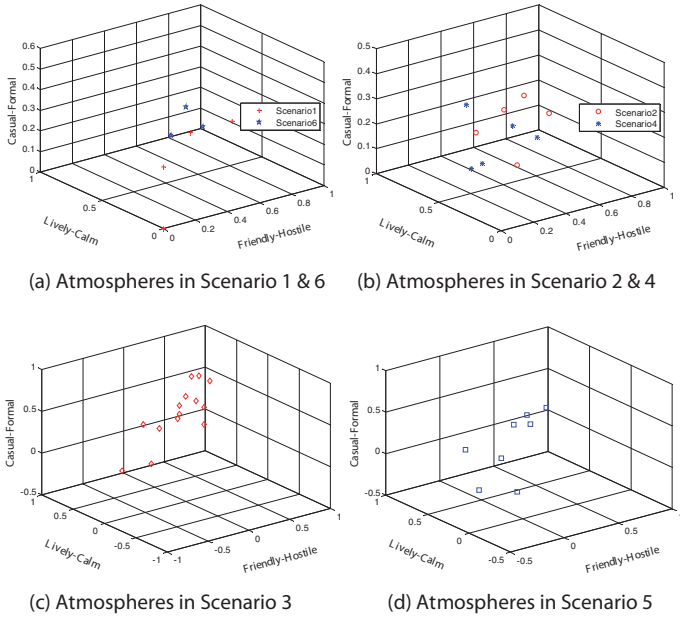
In this experiment, various atmospheres are involved as shown in Fig. 16, for example, in Scenario 1 and Scenario 6, there are friendly, lively, and casual atmospheres when the host welcomes and farewells the guests; in Scenario 3 (i.e., competition of dart game), there are hostile, lively, and casual atmospheres; and there exist friendly, lively, and formal atmospheres when they are querying information such as timetable at the terminal computer in Scenario 5.

To give a more detailed explanation of the experimental results, the Scenario 2 is taken as an example, where the FA is associated with the emotions of Host (E_H), Guest 1 (E_{G1}), Guest 2 (E_{G2}), and Robot 4 (E_{R4}) and Robot 5 (E_{R5}) when they are toasting at the mini-bar. This Scenario is cut into five fragments, as shown in Fig. 14. In the Fragment 1, with keyword *cheers* and the gesture "toast" that are correlated to "happy", initial emotional states including E_H(0.58, 0.63, 0.54), E_{G1}(0.58, 0.63, 0.54), E_{G2}(0.58, 0.63, 0.54), E_{R4}(0.5, 0.6, 0.2), and E_{R5}(0.5, 0.6, 0.2), construct the FA as *FA*(0.6, 0.6, 0.3), representing a very friendly, very lively and casual atmosphere; in the Fragment 3, as the emotional states decreasing, E_H(0.33, 0.38, 0.42), E_{G1}(0.46, 0.51, 0.29), E_{G2}(0.2, 0.13, 0.54), E_{R4}(0.3, 0.35, 0.4), and E_{R5}(0, 0, 0), output *FA*(0.2, 0.13, 0.2) to show



(a) Neutral expression (b) Sadness

Fig. 15. *Eye movement for "sadness" emotion*



(a) Atmospheres in Scenario 1 & 6 (b) Atmospheres in Scenario 2 & 4

(c) Atmospheres in Scenario 3 (d) Atmospheres in Scenario 5

Fig. 16. *Communication atmospheres represented by the FA in a home party*

Table 7. Emotion of each eye robot

	Happiness	Sadness	Suprise
TV Robot	1	1	1
Dart Robot	2	0	0
Information Terminal Robot	3	1	3
Mini-bar Robot	2	0	1
Mobile Robot	6	2	2

a friendly, lively, and casual atmosphere; and in the Fragment 5, E_H(0.46, 0.63, 0.66), E_{G1}(0.58, 0.63, 0.42), and three emotional states that are close to neutral state, i.e., E_{G2}(0.1, 0, 0.29), E_{R4}(0, 0, 0), and E_{R5}(0, 0, 0), creates the FA with *FA*(0.4, 0.5, 0.3) which expresses a very friendly, very lively, and casual atmosphere.To evaluate the experimental results of communication atmosphere based on the FA psychologically, a questionnaire survey is conducted [29]. After watching the experimental video, forty postgraduate students (aged between 20 and 35 years from ten countries, including four females) are asked to evaluate the atmosphere generated by four humans and five eye robots in each fragment. Three questions are formulated [29], i.e., "Is the atmosphere friendly or hostile?", "Is the atmosphere lively of calm?", "Is the communication atmosphere casual or formal?" Seven answer options are provided for each question. For example, to describe "friendly/hostile", optional answers are 1 – extremely hostile, 2 – very hostile, 3 – hostile, 4 – neutral, 5 – friendly, 6– very friendly, and 7– extremely friendly. For further quantitative assessment of the results, each sequence number is assigned with numerical number between -1 and 1, i.e., 1(-1), 2(-0.66), 3(-0.33), 4(0), 5(0.33), 6(0.66), 7(1). The average values of the results from questionnaire survey are used as the reference values for the FA.

To assess the estimation results from the proposed model of the communication atmosphere compared to the reference, two measures, i.e., mean linear error and empirical correlation coefficient [30] are employed. The experimental results show that the errors of 0.18, 0.14, and 0.25, and correlation coefficients of 0.92, 0.86, and 0.72 for "Friendly-Hostile," "Lively-Calm," and "Casual-Formal" axes, respectively, which demonstrates the consistency between the proposed model and human estimators from questionnaires. From the experimental results, it is noted that the "friendly" and "lively" are easier to perceive than "casual" and "formal." The level of "casual" and "formal" is somewhat difficult to understand by the people from different countries in particular, quantitative differences exist in estimating the atmosphere states consequently.

The experiments presented so far in this section demonstrate that using the proposed model of communication atmosphere, smooth communication between humans and robots in a household environment could be realized. This is attributed to the ability of robots to adapt to or even change the atmosphere state in a short response time. A future improvement would be to incorporate an atmosphere representation system for classification, identification, and response, where the robots could execute corresponding tasks such as adjusting spoken content or tone of voice to change/improve the communication atmosphere. Furthermore, music that can inspire emotion of humans could be indispensable for constructing the communication atmosphere, and it could be considered as an independent emotional object as well as humans and robots [13].

In addition, the proposal could be extended to other human-robot interaction using the FA that is a tool for representing the atmosphere in communication among many participants [29], e.g., multi-human to multi-robot interaction, where the situation doesn't need to be predefined. Moreover, human emotion recognition based on multiple cues such as facial expression, speech, body languages could be considered. Robot emotion could be synthesized using speech, sound, gesture, and so on. What's more, emotional states of humans or robots could be represented by other emotion spaces, and the fuzzy rules for mapping emotion space to the FA in Fig. 12 should be adjusted.

6. Conclusion

A home party environment is created for the experiments proceeded with four humans and five eye robots using the MRS (Mascot Robot System) [10]. Based on the concept of the FA (Fuzzy Atmosfield) [7] [29], the communication atmosphere is modeled quantitatively by integrating each emotional state of humans and eye robots, and the results are evaluated by comparing with those from questionnaire surveys. Errors of 0.18, 0.14, and 0.25, and correlation coefficients of 0.92, 0.86, and 0.72 for "Friendly-Hostile," "Lively-Calm," and "Casual-Formal" axes in the FA, respectively are obtained. According to the experimental results, the feasibility of quantitative representation of the communication atmosphere in humans-robots interaction is validated by the degree

of correlation (^{3}0.72) between the results of proposed model and those of subjective estimations. It is also confirmed in an attempt to facilitate smooth communication with several humans and several robots by a home party enjoying demonstration using the MRS.

To evaluate the emotion recognition of humans, emotional keywords/phrases are extracted by using Julius, and emotional gestures are recognized by fusing web camera images and 3D accelerometers data. Recognition rate of human emotion achieves 84% in average, improved by 10% compared to using only semantic cues from speech, and in particular the emotion recognition of "happiness" and "sadness" are improved by 9% and 40%, respectively. The results show that the proposed emotion recognition achieves complementary of emotional information in both speech and gesture, especially when the speech only relates to "neutral" emotion, the gesture contributes to the improvement of emotion recognition accuracy by embodying specific emotions, e.g., "happiness" and "sadness" in different scenarios.

In terms of future works, an atmosphere representation system is planning to be developed for decision-making of robots' behaviors to express corresponding emotions by adjusting the content of speech, improving the way of communication, playing music, dancing [31], and so on, in order to change the communication atmosphere towards designated target. In addition, the proposal aims to be used for an automatic music recommendation system [32] to accommodate the atmosphere to various occasions in humans-robots interaction, where the music could be regarded as an independent emotional object as well as humans.

Acknowledgment

This work is supported by the Japan Society for the Promotion of Science (JSPS) under grant KAKENHI 21300080 and the National Natural Science Foundation of China under grant 61210011. The authors wish to thank Abdullah M. Iliyasu and Martin L. Tangel for their valuable advices in producing this article.

AUTHORS

Zhen-Tao Liu*, Lue-Feng Chen, Fang-Yan Dong, and Kaoru Hirota – Dept. C.I. & S.S., Tokyo Institute of Technology, G3-49, 4259 Nagatsuta, Midori-ku, Yokohama, Kanagawa 226-8502, Japan,
{liuzhentao, chen, tou, hirota}@hrt.dis.titech.ac.jp

Min Wu and Dan-Yun Li – School of Information Science and Engineering, Central South University, Yuelu Mountain, Changsha, Hunan 410083, China,
min@csu.edu.cn, lidanyun0606@yahoo.com.cn

Yoichi Yamazaki – Dept. E. E. & I. E., Kanto Gakuin University, 1-50-1 Mutsuura-higashi, Kanazawa-ku, Yokohama, Kanagawa 236-8501, Japan,
yamazaki@kanto-gakuin.ac.jp

*Corresponding author

REFERENCES

[1] Z.-T. Liu, M. Wu et al., "Emotional states based 3-D Fuzzy Atmosfield for casual communication between humans and robots". In: *IEEE Int. Conf. on Fuzzy Systems*, Taipei, Taiwan, 2011, pp. 777–782.

[2] P. Rani, C. Liu et al., "An empirical study of machine learning techniques for affect recognition in human-robot interaction", *Pattern Analysis & Applications*, vol. 9, no. 1, 2006, pp. 58–69.

[3] D. Kuli´c and E. A. Croft, "Affective state estimation for human-robot interaction," *IEEE Trans. on Robotics*, vol. 23, no. 5, 2007, pp. 991–1000.

[4] P. Robbel, M. E. Hoque et al., "An integrated approach to emotional speech and gesture synthesis in humanoid robots". In: *Proc. of the Int. Workshop on Affective-Aware Virtual Agents and Social Robots*, Boston, USA, 2009.

[5] X. Li, B. MacDonald et al., "Expressive facial speech synthesis on a robotic platform". In: *Int. Conf. on Intelligent Robots and Systems*, St. Louis, USA, 2009.

[6] R. Taki, Y. Maeda et al., "Personal preference analysis for emotional behavior response of autonomous robot in interactive emotion communication", *Jounal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 14, no. 7, 2010, pp. 852–859.

[7] Z.-T. Liu, F.-Y. Dong et al., "Proposal of Fuzzy Atmosfield for mood expression of human-robot communication". In: *Int. Symp. on Intelligent Systems*, Tokyo, Japan, 2010.

[8] Y. Yamazaki, Y. Hatakeyama et al. "Fuzzy inference based mentality expression for eye robot in Affinity Pleasure-Arousal space", *Jounal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 12, no. 3, 2008, pp. 304–313.

[9] L. D. Riek, "Toward natural human-robot interaction exploring facial expression synthesis on an android robot". In: *Proc. of the Doctoral Consortium at the IEEE Conf. on Affective Computing and Intelligent Interaction*, Amsterdam, Netherlands, 2009.

[10] K. Hirota and F.-Y. Dong, "Development of Mascot Robot System in NEDO project". In: *Proc. 4th IEEE Int. Conf. Intelligent Systems*, 2008, pp. 38–44.

[11] H. A. Vu, Y. Yamazaki et al., "Emotion recognition based on human gesture and speech information using RT middleware". In: *IEEE Int. Conf. on Fuzzy Systems*, Taipei, Taiwan, 2011, pp. 787–791.

[12] Y.-K. Tang, H. A. Vu 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*, vol. 5, no. 5, 2011, pp. 563–572.

[13] Z.-T. Liu, Z. Mu et al., "Emotion recognition of violin music based on strings music theory for Mascot Robot System". In: *The 9th Int. Conf. on Informatics in Control, Automation and Robotics*, Rome, Italy, 2012, pp. 5–14.

[14] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals", *IEEE Transactions on Multimedia*, vol. 10, no. 5, 2008, pp. 936–946.
[15] M.-L. Song, M.-Y. You et al., "A robust multimodal approach for emotion recognition", *Neurocomputing*, vol. 71, no. 10, 2008, pp. 1913–1920.
[16] M.-J. Han, J.-H. Hsu et al., "A new information fusion method for bimodal robotic emotion recognition", *Journal of Computers*, vol. 3, no. 7, 2008, pp. 39–47.
[17] P. Ekman, "Are there basic emotions?", *Psychological Review*, vol. 99, no. 3, 1992, pp. 550–553.
[18] Z.-J. Chuang and C.-H. Wu, "Emotion recognition using acoustic features and textual content". In: *IEEE Int. Conf. on Multimedia and Expo*, Taipei, Taiwan, 2004.
[19] B. Schuller, G. Rigoll et al., "Speech emotion recognition combining acoustic features and linguistic information in a hybrid Support Vector Machine - belief network architecture". In: *IEEE Int. Conf. on Acoustics*, *Speech*, *and Signal Processing*, Quebec, Canada, 2004.
[20] A. Lee, T. Kawahara et al., "Recent development of open-source speech recognition engine Julius". In: *Proc. of Asia-Pacific Signal and Information Processing Association*, Sapporo, Japan, 2009.
[21] C. Wu, Z. Chuang et al., "Emotion recognition from text using semantic labels and separable mixture models", *ACM Transactions on Asian Language Information Processing*, vol. 5, no. 2, 2006, pp. 165–182.
[22] B. Schuller, S. Reiter et al., " Speaker independent speech emotion recognition by ensemble classification". In: *IEEE Int. Conf. on Multimedia and Expo*, Amsterdam, Netherlands, 2005.
[23] D. Glowinski, A. Camurri et al., "Technique for automatic emotion recognition by body gesture analysis". In: *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008.
[24] J. Cao, H. Wang et al., "PAD model based facial expression analysis", *Advances in Visual Computing*, *Lecture Notes in Computer Science*, vol. 5359, 2008, pp. 450–459.
[25] J. Martínez-Miranda and A. Aldea, "Emotions in human and artificial intelligence", *Computers in Human Behavior*, vol. 21, no.2, 2005, pp. 323–341.
[26] S. Zhang, Z. Wu et al., "Facial expression synthesis using PAD emotional parameters for a Chinese expressive avatar", *Affective Computing and Intelligent Interaction*, *Lecture Notes in Computer Science*, vol. 4738, 2007, pp. 24–35.
[27] http://www.asha.org/public/hearing/noise/.
[28] http://www.annelawrence.com/voicesurgery.htm.
[29] Z.-T. Liu, M. Wu, D.-Y. Li, L.-F. Chen, F.-Y. Dong, Y. Yamazaki, and K. Hirota, "Concept of Fuzzy Atmosfield for Representing Communication Atmosphere and Its Application to Humans-Robots Interaction", *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 17, no. 1, 2013, pp. 3–17.
[30] M. Grimm and K. Kroschel, "Emotion estimation in speech using a 3D emotion space concept", *Robust Speech Recognition and Understanding*, *I-Tech Education and Publishing*, 2007, pp. 281–300.
[31] F. Michaud, A. Duquette et al., "Characteristics of mobile robotic toys for children with pervasive developmental disorders". In: *IEEE Int. Conf. on Systems*, *Man and Cybernetics*, Washington, USA, 2003.
[32] C.-Y. Chang, C.-Y. Lo et al., "A music recommendation system with consideration of personal emotion". In: *Int. Computer Symp.*, Tainan, Taiwan, 2010.