# A tension-moderating mechanism for promoting speech-based human-robot interaction

Takayuki Kanda[*], Kayoko Iwase[*], Masahiro Shiomi[*†], Hiroshi Ishiguro[*†]

[*]*Department of Communication Robots*
*ATR Intelligent Robotics and Communication Laboratories*
*2-2-2 Hikaridai, Keihanna Science City, Kyoto, Japan*

[†]*Faculty of Engineering*
*Osaka University*
*Suita City, Osaka, Japan*

{kanda, kayoko-i, m-shiomi, ishiguro}@atr.jp

***Abstract*** *- We propose a method for promoting human-robot interaction based on emotion recognition with particular focus on tension emotion. There are two types of emotions expressed in a short time. One is autonomic emotion caused by a stimulus, such as joy and fear. The other is self-reported emotion, such as tension, that is relatively independent of a single stimulus. In our preliminary experiment, we observed that tension emotion (self-reported emotion) obstructs the expression of autonomic emotion, which has demerits on speech recognition and interaction. Our method is based on detection and moderation of tension emotion. If a robot detects tension emotion, it tries to ease it so that a person will interact with it more comfortably and express autonomic emotions. It also retrieves nuances from expressed emotions for supplementing insufficient speech recognition, which will also promote interaction.*

***Index Terms*** *– human-robot interaction; emotion recognition; tension emotion; speech-based interaction.*

## I. INTRODUCTION

Recent developments in humanoid robots enable us to use them for ideal human-interface. Since they can typically make sophisticated human-like expressions, we believe that humanoid robots will be suitable for communicating with humans. The human-like bodies of humanoid robots enable humans to intuitively understand their gestures and cause people to unconsciously behave as if they were communicating with humans. That is, if a humanoid robot effectively uses its body, people will communicate naturally with it. This could allow robots to perform communicative tasks in human society, such as route guidance.

There are several research efforts at speech-based interaction between humans and robots. With a microphone array, Asoh et al. have implemented a speech recognition function for a robot that is capable of working in real office environments [1]. Robot body properties are also utilized for natural, human-style, communication. Nakadai et al. developed a tracking function of human heads based on real-time sensing by vision and audition [2]. Matsusaka and his colleagues developed a robot that can gaze at the person who is talking to it [3]. As shown in these examples, previous research efforts have demonstrated the importance of real-world sensing and the effects of robot existence, which are quite different with speech-based interface in computers.

The importance of non-verbal information has been highlighted for human communication. As Mehrabian reported, 93% of impression of a message (for example, positive or negative) is conveyed non-verbally, while 7% is conveyed verbally [4]. We believe that the robots' human-like bodies make most people expect human-like communication. That is, a humanoid robot should acquire non-verbal information as well as verbal information from interacting people so that it can react to people as naturally as a human does.

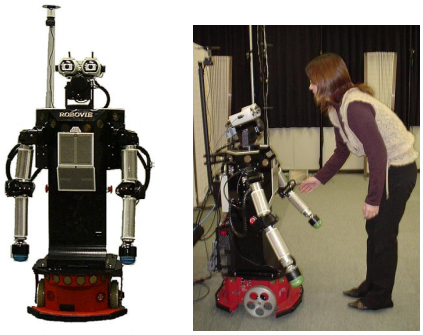

**Figure 1**: Robovie and a scene of interaction with it

Emotions have been utilized in human-robot interaction, such as for creating affective reaction [5], and estimating user context [6]. Moreover, a few papers have reported an integration of verbal and non-verbal information for robots to interact with people. Fujie et al. utilized para-linguistic information and motion, such as nodding, to recognize the attitude of the message that user expressed. Further, they planned to implement this system into a humanoid robot to detect subtle nuances of utterance from people [7]. On the other hand, Komatani et al. implemented an emotion recognition function in a humanoid robot and found that people often became tense and expressed tension emotion [8].

We focus on moderation of the tension emotion. Findings from psychology have classified emotions into categories relative to time [9]. One is "autonomic emotion" caused by a stimulus, such as joy and fear. Another is "self-reported emotion" that is independent of a single stimulus, such as tension. It is our hypothesis that if strong self-reported emotion is expressed, it is difficult to recognize autonomic emotion. For example, a person under tension does not smile broadly, which spoils the ability of robots to detect subtle nuances of utterance from people. In this paper, we will verify this hypothesis and propose a method to detect and moderate tension emotion to promote speech-based interaction. This method will also utilize other detected emotions to supplement the insufficient speech-recognition ability of the robot.

## II. USE OF NON-VERBAL INFORMATION

### A. Communication robot "Robovie"

**Figure 1** shows the humanoid robot "Robovie" [10]. This robot is capable of human-like expression and recognizes individuals by using various actuators and sensors. Its body possesses highly articulated arms, eyes, and a head, which were designed to produce sufficient gestures to communicate effectively with humans. The sensory equipment includes auditory, tactile, ultrasonic, and vision sensors, which allow the robot to behave autonomously and to interact with humans. All processing and control systems, such as the computer and motor control hardware, are located inside the robot's body.

### B. Difficulty of speech recognition in distant communication

The current ability of Robovie's speech recognition is not very good due to noise from the environment. Since an interacting person is expected to stand 50 cm to 100 cm away from the robot, such noise is critical for speech recognition, particularly in daily environments. Humans can, however, communicate with each other even under quite noisy conditions. We believe that this is because we are acquiring non-verbal information, such as facial expression, intonation, and bodily gestures, in addition to verbal information. For example, when we ask someone a question, such as "would you like a coffee," we easily understand the answer if he/she expresses joy.

### C. Psychological knowledge on emotions

Mehrabian has argued the importance of non-verbal information and established an equation on the impression of a message exchanged in human communication, which indicates that most of the impression is conveyed non-verbally [4]:

*Total Feeling = 7% Verbal Feeling +*
*38% Vocal Feeling + 55% Facial Feeling*

Inspired by this idea, we particularly focus on emotions that are meta-level non-verbal information based on other low-level non-verbal information, such as para-language.

In psychology, many research works have reported on emotions. Ekman argued the existence of basic emotions, which are the common emotions of humans, and proposed six basic emotions [11]: anger, disgust, fear, joy, sadness, and surprise. Russell assumed two basic dimensions of emotions, "pleasant – unpleasant" and "low arousal - high arousal", and proposed a circumplex model (**Fig**. 2), where the six basic emotions and other emotions are mapped on a circle [12].

Findings from psychology enabled the classification of emotions into several categories relative to time [9]. The shortest one is autonomic emotion mostly caused by a stimulus (for example, an utterance from a robot) and lasts for seconds, such as joy and fear. Another is self-reported emotion. We are conscious and can report this status. It is relatively independent of such a single stimulus and lasts for minutes, such as tension. (Others are related to moods and emotional disorders that change in hours, days, and so forth.)

### D. Hypothesis: disturbance caused by tension emotion

As Komatani et al. reported, people often became tense (nervous) and expressed tension emotion during human-robot interaction [8]. Similarly, Nomura et al. reported that people who have a negative attitude to a robot tend to avoid communication with it [13]. These findings indicate that it is important to moderate the tension emotion of interacting people.

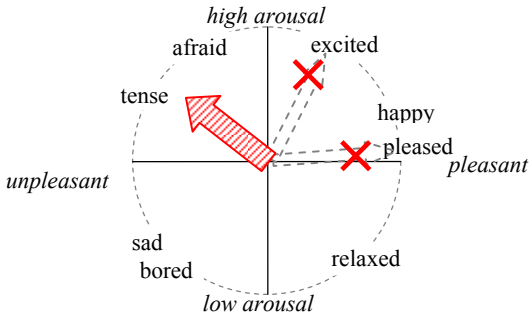


**Figure 2**: Hypothesis on the disturbance by tension emotion based on the Circumplex model proposed by Russell

Such moderation will also have merit in the recognition of emotion. As Russell's circular model suggests, if there is a strong emotion expressed, other emotions may be not observed (**Fig.** 2). That is, in the case of human-robot interaction, tension emotion, which is a self-reported emotion, could disturb the expression of other autonomic emotions, such as joy. For example, a person under tension does not smile broadly. This will hamper the ability of robots to detect subtle nuances of utterance. To summarize these discussions, it is our hypothesis that if strong self-reported emotion is expressed, it is difficult to recognize autonomic emotion.

### E. Hypothesis verification

We conducted a preliminary experiment for verification of the hypothesis. 45 university students participated. In the experiment, participants talked with Robovie from a distance of 50 cm. Robovie repeatedly asked some simple questions, such as "Let's play together, shall we?" and "Do you think Robovie is cute?" The participants were required to answer the questions (1) freely, (2) positively, and (3) negatively. We recorded the participants' faces observed from Robovie's camera for later analysis.

#### *Labeling of facial expression*

We selected 72 data items from obtained faces where the participants reported they expressed their emotions. Third persons other than the authors and participants rated the recorded faces with seven scales related to emotions: anger, disgust, fear, joy, sadness, surprise, and tension. These seven emotions were chosen because Ekman's six basic emotions are recognizable by the facial emotion recognition system we used for Robovie and the tension emotion effect is what we want to verify. The rating was conducted with *-3* (not match at all) to *3* (match very much) scales for each emotion ( $f_{anger}$ , $f_{disgust}$ , $f_{fear}$ , $f_{joy}$ , $f_{sadness}$ , $f_{surprise}$ , $f_{tension}$ ), and the most highly rated emotion was selected as the face label. (For example, if $f_{joy}$ was highest, the face was labeled as joy.)

#### *Results for hypothesis verification*

We classified all rated items into two categories of tension: *with tension* ( $f_{tension} > 0$) and *without tension* ( $f_{tension} \leq 0$),

and two categories of the attitude of the messages: *positive* and *negative* (**Table** 1). As a result, 23 items were classified into the *with tension* category among the 72 items. That is, there is tension emotion observed in 32% of the items. Most cases in the *with tension* category were labeled as tension emotion because it was highly rated over other emotions. This reinforces the importance of moderating tension emotions in human-robot interaction.

We believe that this result verifies our hypothesis. Often people become tense and express tension emotion, which is a kind of self-reported emotion. Tension represses the expression of other autonomic emotions. Thus, *with tension* cases, it is difficult to infer the nuance of a message from observed emotions.

On the contrary, in *without tension* cases, we often observed the joy emotion in *positive* answer cases, and anger and disgust emotions in *negative* answer cases. This suggests that we can observe the nuances of messages from emotions. For example, if the result of speech recognition on a message is ambiguous between "yes" and "no," positive emotions let us believe that the message is related to a positive answer.

## III. System configuration

We implemented the proposed method to detect and moderate tension emotion to promote speech-based interaction. The proposed method also utilizes other detected emotions to supplement insufficient speech-recognition ability.

### A. Overview

**Figure 3** shows the overview of the developed system for speech-based interaction based on emotion recognition. It consists of three recognition units: face tracking unit, speech recognition unit, and emotion recognition unit. The face tracking unit tracks the face of an interacting person so that it can observe facial emotions and direct its own directional microphone to that person. The emotion recognition unit detects tension emotion and other emotions, which are used for behavior selection and speech recognition units, respectively. If there is no tension emotion detected, the result of the speech recognition unit is used for behavior selection.

### B. Face tracking unit

Robovie can track the faces of interacting people by using an eye-camera and omnidirectional camera [14]. The direction of its head is controlled so that it can maintain visual contact of faces. Through this process, the face tracking unit acquires frontal face vision, which is used in the emotion recognition unit. At the same time, since it controls Robovie's head toward the interacting person, the direction of the attached directional microphone approaches this person, which has, as a result, merits in obtaining less noisy auditory input from him/her.

### C. Emotion recognition unit

There are two sources for the emotion recognition unit: facial emotions and vocal emotions. The facial emotions are recognized by using a system developed by Littlewort et al. [15]. This system is based on Ekman's FACS (Facial Action Coding Systems) and outputs likelihoods of six emotions (anger, disgust, fear, joy, sadness, and surprise) with an SVM (Support Vector Machine) so that we can recognize these six emotions and neutral emotion based on these likelihoods.

**TABLE 1**: Result of preliminary experiment

| Emotion | $f_{tension} > 0$ (with tension) | | $f_{tension} \leq 0$ (w/o tension) | |
|---|---|---|---|---|
| | *Positive* | *Negative* | *Positive* | *Negative* |
| No. of items | *14* | *9* | *34* | *15* |
| Anger | 0 % | 0 % | 6 % | 27 % |
| Disgust | 3 % | 0 % | 22 % | 45 % |
| Fear | 0 % | 7 % | 0 % | 0 % |
| Joy | 29 % | 7 % | 60 % | 27 % |
| Sadness | 6 % | 0 % | 12 % | 1% |
| Surprise | 0 % | 0 % | 0 % | 0 % |
| Tension | 62 % | 86 % | 0 % | 0 % |

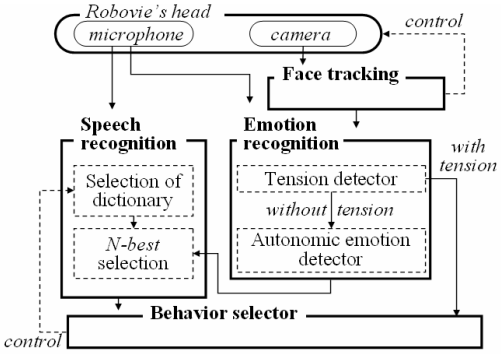


**Figure 3**: Robovie's speech-based interaction system

The vocal emotions are recognized using Komatani et al.'s method [8]. This method uses 29 features that are calculated based on fundamental frequency (F0), power, length of utterance, and duration between utterances. It detects joy and perplexity emotions with the SVM. Further, we added detection of tension emotions with a C5.0 decision tree with the same 29 characteristics. Trained with 400 data items obtained in an experiment with 15 participants in the same settings as reported in Section II, a performance for tension emotion detection at 67.1 % of correct answer for the test data was obtained (not including any training data).

### D. Speech recognition unit

***Situated recognition on speech recognition***

We adopted a speech recognition system, Julian [16], which allows us to switch its dictation dictionaries and grammars. Based on our constructive approach with situated recognition [10], a dictionary and grammar is chosen to conform to Robovie's current situation. For example, when Robovie asks about the name of a place, such as "where are you from," it uses a dictionary that includes names of places so that it can get a better recognition result in a noisy environment. Each dictionary contains 50 to 200 words.

It outputs *N-best* results ($1 \leq N \leq 5$) of recognition with a certain threshold on the likelihood score. These *N-best* results are compared with the output from the emotion recognition.

***Supplementing speech recognition from emotion recognition***

We supplemented the insufficient ability of the robot's speech recognition with the result from the emotion recogni-

tion. This supplement is conducted when tension emotion is not detected and a positive-negative answer is expected, such as an answer to a yes-no question.

The result from emotion detection is classified into three categories: positive emotion (denoted as *Pe*), negative emotion (*Ne*), and neutral (*Nt*). If a joy emotion is detected from either facial or vocal emotion recognition, the system classifies it as a positive emotion (*Pe*). If there is anger, disgust, fear, or sadness detected from facial expressions, or perplexity from vocal information, the system classifies it as a negative emotion (*Ne*). Otherwise, the system assumes a neutral emotion (*Nt*). If conflict occurs between the recognition of the facial and vocal emotions, the facial emotion is used.

We decided on this classification by referring to the analysis results reported in Section II. In addition, although there were no cases of fear and sadness related to negative utterances, we classified fear and sad emotions as negative emotions; because the number of analyzed data was too small to conclude that these emotions are not related to negative utterances and these emotions are usually related to negative situations.

As a result, if a positive (negative) emotion is detected, the method refers the *N-best* results from speech recognition and chooses words with positive (negative) meanings. Thus, it chooses the word that fits best with the nuance estimated from the non-verbal information. The meaning of the words, whether positive or negative, is defined in advance.

### *E. Behavior selector: use of recognized emotions*

As reported in [15], Robovie always exhibits interactive behavior, such as shake-hands, greeting, and asking simple questions. The behavior selector receives recognition results from the speech recognition unit and emotion recognition unit in order to switch the interactive behaviors. In particular, when tension emotion is detected, Robovie exhibits tension-moderating behaviors, such as self-introduction and talking about the weather, which humans often do when meeting a person for the first time. Otherwise, it chooses its interactive behaviors based on the result from speech recognition, which is supplemented with the result from emotion recognition.

## IV. EXPERIMENT

We conducted an experiment to verify the effect of the proposed method and developed system.

### *A. Settings*

**[Participants]** The participants in our experiment were 27 university students (12 men and 15 women). Their average age was 19.7 years old.

**[Methods] Figure 1** (right) shows a scene of the experiment, which was conducted in a room in our laboratory, in which the participants and Robovie talked. Each participant stood about 50 cm in front of Robovie. At first, Robovie showed the normal behavior "hello" (N1, shown in **Table 2**). If there was a tension emotion detected in the participant's utterance in response to the "hello," it initiated a tension-moderating behavior (T1); otherwise it started the next normal behavior (N2). After it spoke a sentence as listed in Table 2, Robovie expected a response from the participant, and it spoke shortly in reply to the response, such as "thank you," "I'm glad," and "that's disappointing," according to the speech-recognition result. After that, if it detected a tension emotion from the response, it performed the next tension-moderating behavior; otherwise it performed the next normal behavior. For example, after it performed S2 after N2, it executed S3 or N3, which was decided by referring to the results for tension detection. The experiment ended after the execution of the last normal behavior (N6). If the last tension-moderating behavior (T5) was performed, even if it detected a tension emotion, it performed the next normal behavior.

**Table 2**: Robot's utterances for normal behaviors and tension-moderating behaviors used in the experiment

| Normal behaviors | | Tension-moderating behaviors | |
|---|---|---|---|
| N1 | Hello | T1 | I'm Robovie. What is your name? |
| N2 | Let's talk together. Shall we? | T2 | I'm from ATR. Where are you from? |
| N3 | Let's play together. Shall we? | T3 | What do you think today's weather? |
| N4 | Do you think Robovie is your friend? | T4 | Let's play a game of paper scissor rock. Shall we? (It plays the game) |
| N5 | Do you think Robovie is cute? | T5 | Do you know the song of "a flower smile"? (It sings the song.) |
| N6 | Bye-bye | | |

**[Measurement]** We video-taped the experiment to record the faces and utterances of the participants. Robovie's recognition results were also recorded for later analysis. After the experiment, we asked the participants to answer the following questionnaire:

Q1. "Did Robovie understand your utterances?"

Q2. "Do you feel it is easier to speak to Robovie after this session, than it was before the session?"

Q3. "Are there problems in communication with Robovie?"

### *B. Results*

We asked a third party to label the emotions of the participants during each of their utterances in the experiment. There were three classes of emotions (positive emotion: *Pe*, negative emotion: *Ne*, neutral: *Nt*) for facial emotions and vocal emotions, and two classes for tension (*with tension*, *without tension*). As a result, there were 220 utterances of the 27 participants analyzed. This was used as a ground-truth of emotion recognition to evaluate performance of the developed system.

***Results for performance of emotion recognition***

We compared the emotion recognition output from the system with the labeled emotions. **Table 3** shows the results of the comparison, where "success rate" represents the rate that the system output correctly matched the labeled emotions for all classes among 220 utterances. As a result, the success rate for the tension detection (denoted as *Tension* in **Table 3**), emotion detection from face (*Facial*), and that from vocal (*Vocal*) was 55.0%, 36.8%, and 31.7%, respectively.

Further, we analyzed the detailed failed rate. In Table 3, "opposite", "false neutral", and "false from neutral" represents the rate of each case among all 220 utterances. As a result, the labeled results and the output from the system often mismatched around the boundary on neutral emotion, which low-

**Table 3:** Result for emotion recognition

| | | *Tension* | *Facial* | *Vocal* |
|---|---|---|---|---|
| No. of classes | | **2** | **3** | **3** |
| No. of analyzed data items | | 170 | 212 | 170 |
| No. of error data items (omitted from analysis) | | 50 | 8 | 50 |
| **Success rate** | | 55.0% | 36.8% | 31.7% |
| Failed rate | **Opposite:** *Pe* (*Ne)*, classified as *Ne* (*Pe)* | — | 11.8% | 6.5% |
| | **False neutral:** *Pe*, *Ne,* classified as *Nt* | — | 19.8% | 6.5% |
| | **False from neutral:** *Nt,* classified as *Pe*, *Ne* | — | 31.6% | 55.3% |

**Table 4**: Effect of tension-moderating behaviors

| | Normal behavior | Tension-moderating behavior |
|---|---|---|
| Success rate of tension-moderation | 12% ( 7 / 54 ) | 42% ( 21 / 50 ) |
| Success rate of improving emotion-expression | 31% ( 15 / 48 ) | 54% ( 19 / 35 ) |

ered the system's success rate. On the other hand, the number of "opposite" items was relatively small (11.8% for facial and 6.5% for vocal).

Meanwhile, eight data items for facial emotion caused errors and were omitted from this analysis. Since participants sometimes looked away or their faces were sometimes occluded by their arms, Robovie could not observe their faces, resulting in error outputs for facial emotion recognition. In addition, there were 50 data items for vocal emotion omitted due to errors in the low-level analysis program for retrieving F0 and pitch, which is probably due to background-noise in inputs.

***Moderation of tension***

Next, we compared the effects of tension-moderating behaviors for moderating tension emotion with that of normal behaviors (the behaviors used for the experiment are shown in Table 2). **Table 4** shows the results of the comparison. "Success rate of tension-moderation" represents the rate of the disappearance of the tension emotion after the execution of a behavior in situations where tension emotion was observed before the execution. For example, there were 50 cases of tension emotion observed before the execution of tension-moderating behavior, while there were 21 cases of no-tension emotion observed after these behaviors. (These evaluations were, of course, based on the labeled emotions).

We also compared the effects of tension-moderating behaviors for improving positive-negative emotions with that of normal behaviors. **Table 4** also shows the results of the comparison for the improvement of emotions, where "success rate of improving emotion-expression" represents the rate of the appearance of positive or negative emotions after the execution of a behavior in situations where there were no positive or negative emotions observed before the execution.

***Improvement of speech recognition***

**Table 5** shows the results of supplementing speech recognition from emotion recognition. There were 136 utterances in reply to Robovie's "yes" or "no" questions for the participants (N2, N3, N4, N5, T4, and T5 in Table 2). We analyzed the performance of speech recognition for these utterances, because the supplementation mechanism currently works for such utterances when the interacting person answers either positively or negatively.

**Table 5**: Results for supplementing speech recognition from emotion recognition

| | | No. of utterances | Supplementation | |
|---|---|---|---|---|
| | | | with | w/o |
| Success rate | All utterances for answering positive (yes) – negative (no) questions | 136 | 70% | 60% |
| | Only the utterances where participants expressed positive or negative emotions | 68 | 50% | 43% |

**Table 6:** Results for subjective evaluation

| | Yes | No | |
|---|---|---|---|
| Q1. Understanding | 20 | 7 | p<.05 |
| Q2. Easiness | 24 | 3 | p<.01 |
| Q3. Difficulty | 11 | 16 | n.s. |

In total, Robovie detected a correct answer for 70% of the utterances with the supplementation mechanism, but for 60% of the utterances without it. A correct answer was when it detected the correct keyword in the utterance, such as "yes," "ok," "cute," or "let's play," among eight to nine sets of keywords. Furthermore, we focused on the utterances where the participants expressed positive or negative emotions. As shown in the table, the performance for these utterances was 50% with the supplementation, while that without the supplementation was 43%. There were three cases where the supplementation mechanism for speech recognition failed.

***Subjective evaluation***

The participants answered the questionnaire after the experiment. As a result, 20 of the 27 participants answered that Robovie understood their utterances (Q1: understanding) and 24 participants answered that communication with Robovie became easier as they communicated with it (Q2: easiness), as shown in **Table 6**. A Chi-square test proved that the number of participants who answered "yes" for Q1 and Q2 were statistically more than that of the participants who answered "no", which seems to suggest that the majority of participants enjoyed communication with Robovie.

On the other hand, there were 11 participants who responded that there were problems in communicating with Robovie, providing comments such as "it was difficult to communicate, once I recognized it as a machine," "it was difficult to anticipate what it would say," and "I don't think it understands what I say." This seems to suggest that its communication abilities are still far below humans.

### *C. Discussions*

As we intended, tension-moderating behavior has an effect for moderating the tension emotion of the interacting person. Moreover, it improves their expression of positive or negative emotions, which fits with the model proposed in Section II-D (Figure 3). The supplementation mechanism also worked well to improve the performance of the speech recognition. The questionnaire results showed that most of the participants enjoyed communication with Robovie.

On the other hand, the success rates for emotion recognition were relatively poor and far lower than their original performances. Since the failed rate for "opposite" cases was not so

large, we believe that one major difficulty was the distinction of subtle expressions, while noisy input also probably decreased the performance. Often an emotion-recognition system is trained and evaluated with very expressive examples under noise-free conditions, such as a face with a big smile and a voice in an anechoic room. However, our practical use for the robot highlighted the weakness of the emotion-recognition system for detecting subtle expressions in a noisy environment.

We believe that the lower failed-rate of emotion recognition in "opposite" cases also explains why the supplementation mechanism worked with a poor success-rate of emotion recognition. When speech-recognition works well, the *N-best* result from speech recognition only includes a few candidates with a higher likelihood score, thus the poor result from emotion recognition is scarcely affected. On the contrary, when the output from speech-recognition is ambiguous, there are both positive and negative words included in the *N-best* candidates. Since emotion recognition does not often fall in the "opposite" case, it improves the performance of speech-recognition.

One interesting finding was that the speech-recognition performance was low for utterances where the participants expressed their emotions (43%, without the supplementation mechanism, in Table 5). Because the speech-recognition system is usually trained for utterances with neutral emotions, perhaps it does not work well in such a situation without specific training.

The tension-moderating mechanism worked well for this particular experiment, because almost all the participants were tense at the beginning. Thus, although the tension detection unfortunately behaved in a nearly random fashion, Robovie sometimes exhibited tension-moderating behaviors and, as a result, moderated the tension. We believe that the effects of tension-moderating behaviors suggest the usefulness of our framework; however, for practical and effective use, we should improve the performance of tension detection so that it does not exhibit unneeded tension-moderating behaviors.

## V. CONCLUSION

This paper reported our approach to promoting human-robot interaction based on emotion recognition, with particular focus on tension emotion. The preliminary experiment revealed that tension emotion obstructs the expression of autonomic emotions, which degrades emotion recognition and interaction. We focused on tension emotion moderation and proposed a method to detect and moderate tension emotion to promote speech-based interaction, which we implemented in Robovie. This system utilizes other detected emotions to supplement insufficient speech-recognition ability. The effectiveness of proposed method was verified through the experiment. In summary, it revealed the following four points:

- *The tension problem:* People became under tension faced with the robot, which obstructed express of other emotions.
- *Effect of tension-moderation:* It moderated the tension emotion so that people expressed other emotions more, such as joy.
- *Effect of supplementation mechanism:* The expressed emotions improved the performance of speech recognition.
- *Insufficient recognition performance:* It also unfortunately showed insufficient performance of emotion recognitions for the robot, although recognition worked well on the computer.

Thus, in our future research, we intend to develop a robust emotion recognition system for robots.

## ACKNOWLEDGMENTS

We wish to thank Prof. Masuzou Yanagida at Doushisha University and Prof. Tatsuya Kawahara at Kyoto University for their valuable advices. This research was supported by the National Institute of Information and Communications Technology of Japan.

## REFERENCES

[1] H. Asoh, S. Hayamizu, I. Hara, Y. Motomura, S. Akaho, and T.Matsui, Socially Embedded Learning of the Office-Conversant Mobile Robot Jijo-2, *Int. Joint Conf. on Artificial Intelligence (IJCAI'97),* 1997.

[2] K. Nakadai, K. Hidai, H. Mizoguchi, H. G. Okuno, and H. Kitano, Real-Time Auditory and Visual Multiple-Object Tracking for Robots, *Int. Joint Conf. on Artificial Intelligence (IJCAI'01)*, pp. 1425-1432, 2001.

[3] Y. Matsusaka, et al., Multi-person Conversation Robot using Multi-modal Interface, *Proc. World Multiconference on Systems, Cybernetics and Informatics,* pp. 450-455, 1999.

[4] A. Mehrabian, *silent messages*, Thomson Learning College, 1980.

[5] C. Breazeal and L. Aryananda, Recognizing Affective Intent in Robot Directed Speech, *Autonomous Robots*, 12:1, pp. 83-104, 2002.

[6] S-M Baek, D. Tachibana, F. Arai, and T. Fukuda, Situation Based Task Selection Mechanism for Interactive Robot System, *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS 2004)*, pp.3738-3743, 2004.

[7] S. Fujie, Y. Ejiri, Y. Matsusaka, H. Kikuchi, T. Kobayashi, Recognition of Para-linguistic Information and its Application to Spoken Dialogue System, *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '03),* pp. 231 – 236, 2003.

[8] K. Komatani, R. Ito, T. Kawahara, H. G. Okuno, Recognition of Emotional States in Spoken Dialogue with a Robot. *Int. Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems (IEA/AIE'04),* pp. 413-423, 2004.

[9] K. Oatley and J. M. Jenkins, *Understanding emotions*, Blackwell, 1996.

[10] T. Kanda, H. Ishiguro, M. Imai, T. Ono, Development and Evaluation of Interactive Humanoid Robots, *Proceedings of the IEEE*, Vol.92, No.11, pp. 1839-1850, 2004.

[11] P. Ekman, An argument for basic emotions, *Cognition and Emotions*, pp. 169-200, 1992.

[12] J. A. Russell, A circumplex model of affect, *Journal of Personality and Social Psychology*, pp1161-1178, 1980.

[13] T. Nomura, T. Kanda, and T. Suzuki, Experimental Investigation into Influence of Negative Attitudes toward Robots on Human-Robot Interaction, *SID (Social Intelligence Design),* 2004.

[14] M. Shiomi, T. Kanda, N. Miralles, T. Miyashita, I. Fasel, J. Movellan, and H. Ishiguro, Face-to-face Interactive Humanoid Robot, *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS2004)*, pp. 1340-1346, 2004.

[15] G. Littlewort, M. S. Bartlett, I. Fasel, J. Chenu, T. Kanda, H. Ishiguro and J. R. Movellan, Towards Social Robots: Automatical Evaluation of Human-robot Interaction by Face Detection and Expression Classification, *Int. Conf. on Advances in Neural Information Processing Systems*, 2003.

[16] A. Lee, T. Kawahara, K. Takeda, M. Mimura, A. Yamada, A. Ito, K. Itou, and K. Shikano, Continuous Speech Recognition Consortium --- an open repository for CSR tools and models ---, *IEEE Int' Conf. on Language Resources and Evaluation*, 2002.