IEICE TRANS. INF. & SYST., VOL.E91–D, NO.6 JUNE 2008

PAPER Special Section on Human Communication III

Facial Expression Generation from Speaker's Emotional States in Daily Conversation

Hiroki MORI[†a)], *Member and* **Koh OHSHIMA**[†], *Nonmember*

SUMMARY A framework for generating facial expressions from emotional states in daily conversation is described. It provides a mapping between emotional states and facial expressions, where the former is represented by vectors with psychologically-defined abstract dimensions, and the latter is coded by the Facial Action Coding System. In order to obtain the mapping, parallel data with rated emotional states and facial expressions were collected for utterances of a female speaker, and a neural network was trained with the data. The effectiveness of proposed method is verified by a subjective evaluation test. As the result, the Mean Opinion Score with respect to the suitability of generated facial expression was 3.86 for the speaker, which was close to that of hand-made facial expressions.

key words: *dialogue, paralinguistic information, emotional state, facial action coding system, avatar*

1. Introduction

Humans are making use of various means to communicate with each other. Among all, spoken language plays a primary role in human communication. This fact can be understood if we remember how rapidly mobile phones have been popularized and influenced our whole life. In most cases, however, linguistic content alone is not the whole message. Non-verbal elements such as facial expression, gaze, and gesture also form indispensable part of communication. *Paralanguage* [1] is the term referring to such non-verbal elements of communication that are accompanied by verbal message and convey speaker's peripheral information like emotion, attitude and intention. Speech signals deliver rich paralinguistic information as well. Typical carrier of paralinguistic information in speech communication is prosody (pitch, intensity, rhythm, etc.) and voice quality.

Meanwhile, current virtual communication media on the internet, which includes text-based chat and *metaverse* (e.g. Second Life), cannot handle the paralinguistic aspect of participants' message. On the one hand, *emoticons* (characters representing facial expressions) are commonly used in chat in addition to standard punctuation marks. On the other hand, avatars in the metaverse have capabilities to display non-verbal, gestural actions. However, both of them are quite restrictive. Suppose a user wants to exhibit a certain emotion in the metaverse. Then he/she selects an emotive action from a menu, which may be assigned to a keyboard shortcut. Apparently, the inventory of emotion is finite and pre-defined (though customizable). If our emotions could be classified into a small number of categories, the selection and display of an emotion was straightforward. But that is a too simplified way in reality, as long as avatars are responsible for displaying paralanguage just as in our daily conversation. The fact that emotion labels such as fear, anger, happiness, sadness, surprise, and disgust (the "Big Six [2]") do not fit to most part of daily speech was also pointed out in a project for the massive collection of speech data in daily conversation [3].

In this paper, we propose a framework for generating facial expressions for daily conversation. The framework provides a mapping between emotional states and facial expressions, where the former is represented by vectors with psychologically-defined abstract dimensions, and the latter is coded by the Facial Action Coding System (or FACS [4]). Adopting abstract dimensions as representation of emotional states, our method enables avatars to express not only "full-blown" emotions [5] but also subtle ones that are difficult to describe with category labels of emotions.

As a realistic instance of the applications of the framework, we are developing a system where facial expressions of an avatar on screen reflect user's emotional states estimated from his/her utterances, as depicted in Fig. 1. The system is an integration of two modules, namely emotional state estimation module and facial expression generation module. The first module can be realized by establishing rules which map vocal cues into their perceived emotional states (e.g. [6]), which is beyond the scope of this paper. The facial expression generation method proposed in this paper is a realization of the second module.

This paper is organized as follows. Section 2 reviews our method of emotion description and its backgrounds. Section 3 illustrates the framework of facial expression generation. In the section, a brief outline of the Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies is given. The section also explains the building of facial expression data corresponding to each utterance in the database. Section 4 shows the method and examples of facial expression generation by machine learning. Section 5 describes the procedure and result of subjective evaluation test for validating the effectiveness of proposed method. Section 6 concludes the paper.

Manuscript received September 9, 2007.
Manuscript revised December 8, 2007.

[†]The authors are with the Faculty of Engineering, Utsunomiya University, Utsunomiya-shi, 321–8585 Japan.
a) E-mail: hiroki@speech-lab.org

DOI: 10.1093/ietisy/e91–d.6.1628

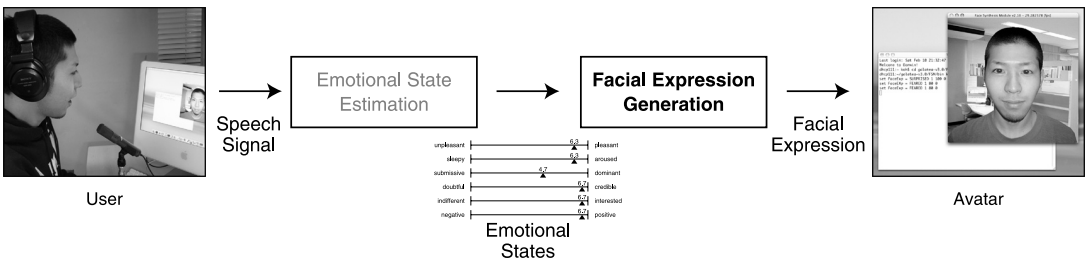

Fig. 1 Outline of an example application of proposed framework.

2. Emotion Description

The term "emotion" is used in multiple senses. Cowie [5] carefully distinguished "full-blown emotion" (or "emotion" in the narrow sense) from "underlying emotion," which refers to an aspect of mental states that may influence a person's thoughts and actions but the actions are more or less under control. In our ongoing studies, emotional states of speakers are considered to be continuous because our interest is not limited to full-blown emotions. Therefore, we adopt emotion dimensions instead of categorical descriptions.

Dimensional descriptions of emotions have a long history and are well established in psychology. Among all, *activation* and *evaluation* dimensions have been regarded as minimal representations of emotions [7], [8]. Some studies incorporates *power* as a third dimension. The important point is that emotion-related words can be approximately converted from/to positions in a two- or three-dimensional space. For example, happiness corresponds to a vector with positive activation (aroused) and positive evaluation (pleasant); anger with positive activation (aroused) but a negative evaluation (unpleasant); sadness with negative evaluation (unpleasant) and possibly negative activation (sleepy).

In this paper, it is assumed that emotional states that affect facial expression of virtual characters are described with the following six dimensions [9]:

1. pleasant-unpleasant
2. aroused-sleepy
3. dominant-submissive
4. credible-doubtful
5. interested-indifferent
6. positive-negative

The first three dimensions are compatible with evaluation, activation and power, respectively. Note that these dimensions were chosen as necessary for investigating paralinguistic aspects of dialogue speech and therefore not specifically intended for voluntary control of virtual characters.

There are several ways for acquiring dimensional descriptions of users' emotional states. For voice chat or interpreting telephony, various acoustic cues can be used to predict speaker's perceived emotional states [6]. Biosignals are also promising candidates to estimate speaker's state [10]. Even manual selection of emotion category from a menu does work using a presupposed mapping function (e.g. [11]).

3. Data Preparation

3.1 Parallel Data

Machine learning requires data. In order to obtain a mapping function from emotional states to facial expressions, we need a collection of facial expressions in various emotional states. Unfortunately, no such database is available as of now. However, as far as applied to facial expression generation from speech signal, speaker's actual facial expression matters little. Rather, consistency of facial and vocal expressions as paralinguistic cues does matter for the listeners. Therefore, we decided to get started with speech data.

Figure 2 illustrates the procedure of collecting parallel data for each utterance. On the one hand, perceived emotional state of the utterance is rated with the emotion dimensions described in the previous section. Section 3.2 gives detailed explanation of this. On the other hand, a facial expression is created by manipulating a character's neutral face. Operators are trying to make the facial expression as suitable to the utterance as possible. This process is explained in Sect. 3.3.

3.2 UU Database

Annotation of emotional states for utterances in daily conversation is provided by the Utsunomiya University (UU) Spoken Dialogue Database for Paralinguistic Information Studies [9] (henceforth UU Database). The UU database is especially intended for use in understanding the usage, structure and effect of paralinguistic information in expressive conversational speech. In this section, a brief overview of the UU database is given.

The UU Database is a collection of natural, spontaneous dialogues of college students consisting of seven pairs (12 females, 2 males). The participants and pairing were selected carefully to ensure that both people in each pair were of the same grade and able to get along well with each other.

The task of the dialogues, namely "four-frame cartoon sorting," was carefully designed to stimulate expressively-rich and vivid conversation. In this task, four cards each containing one frame of a four-frame cartoon were shuffled, and each participant had two cards out of the four. Then

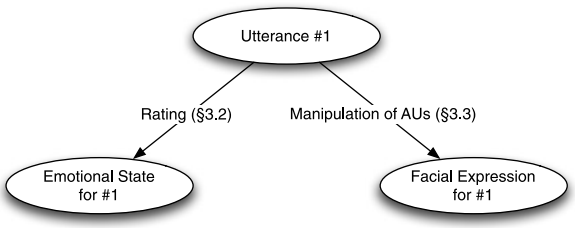

Fig. 2 Collecting parallel data for an utterance.

they were asked to estimate the original order by communicating by voice, without looking at the remaining cards. The task proved to motivate the participants quite well because most Japanese students like cartoons and would be eager to know the true story. Each pair participated in three to seven independent sessions, using different cartoon materials for different sessions.

In the UU database, each utterance is assigned a six-dimension vector that represents the perceived emotional state of the speaker in the identical way as described in Sect. 2. Multiple annotators rated the perceived emotional state of the speaker for each dimension with a value from 1 to 7, where 4 corresponds to neutral. They listened to the stimuli using an identical environment (PC, headphones, playback level). In our previous works [12], the consistency and inter-annotator agreement of the emotional state rating was extensively examined with 22 annotators for the "core" subset of the database, and the results were used to conduct a screening test for newly hired annotators. Consequently, three out of six annotators were selected according to our criteria (consistency, correlation with the average, distinction of the dimensions), who then rated the emotional states for the rest of the corpus. Complete (3 of 3) and partial (2 of 3) agreement were 22.09% (chance: 2.04%) and 83.92% (chance: 38.78%), respectively.

In this paper, utterances of a female speaker from the "core" is mainly used as training/test data, which is called "FUE set" hereinafter. FUE set contains 219 utterances in 3 sessions. In addition to this, "FMS set" of another female speaker and "MKK set" of a male speaker, each of which was composed of 100 utterances, were prepared for the evaluation test described later. The speakers were chosen because all of them are vivid and expressive speakers with wide variety of speaking style.

3.3 Creation of Facial Expression

In this paper, modules for face image fitting/synthesis that are offered as a part of Galatea Toolkit [13] are used as instruments for facial expression generation. The toolkit also provides a 3D wire frame model of a face, where a set of facial muscle movements are defined according to the Facial Action Coding System (FACS [4]). FACS is a method for decomposing facial expression into anatomically-based minimal actions, i.e. Action Units (AUs).

The face image fitting module offers a facial action control facility, which enables the manipulation of facial expression by independent control of 30 AUs.

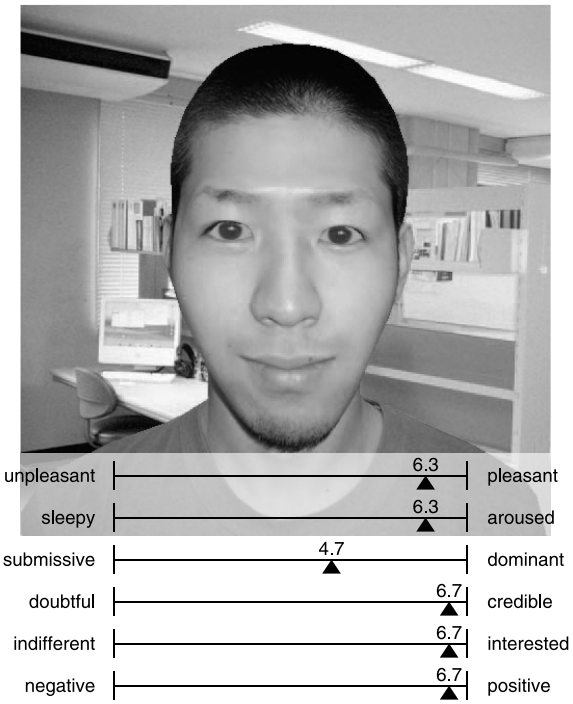

Fig. 3 An example of parallel data.

In order to match the facial expression with each utterance, one of the authors manipulated a neutral face image of himself. The matching procedure was as follows: (i) listen to an utterance, (ii) for each AU, adjust the control so as to fit best to the utterance, (iii) if not satisfied, go to (ii). Finally, a set of active AUs and their amplitude (0–100%) for 219 utterances in FUE set, 100 utterances in FMS set and 100 utterances in MKK set is obtained[†].

An example of rated emotional states and facial expression for an utterance is shown in Fig. 3. The values on the scales denote the averaged values of rated emotional state by three annotators.

4. Machine Learning

Given a set of parallel data, a mapping function between emotional states and facial expressions can be acquired using some machine learning method. In this paper, 3-layer neural network (NN) is adopted. The NN was trained using the back propagation, with 6-dimensional emotional states as input and 30-dimensional AUs as output, for FUE set. The number of units of the intermediate layer was experimentally set to 6.

Root mean square error (RMSE) between the output

[†]Here a gender inconsistency arises, which should be resolved in future. Privacy-right-free Japanese female image is not available so far. We tried to collect dialogues of himself, but the paralinguistic variation of his utterances was not sufficient for the current study.

of the NN and given data converged to 0.085 after an iterative learning. A 10-fold cross validation test was conducted, which showed 0.118 in RMSE. From the RMSE values, it can be said that the mapping function obtained is sufficiently generalized for unseen data.

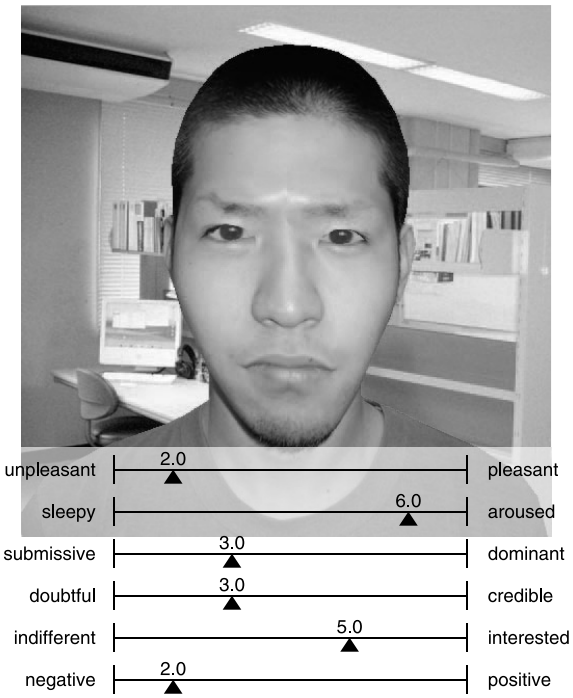

Fig. 4 An example of facial expression generated from an emotional state.

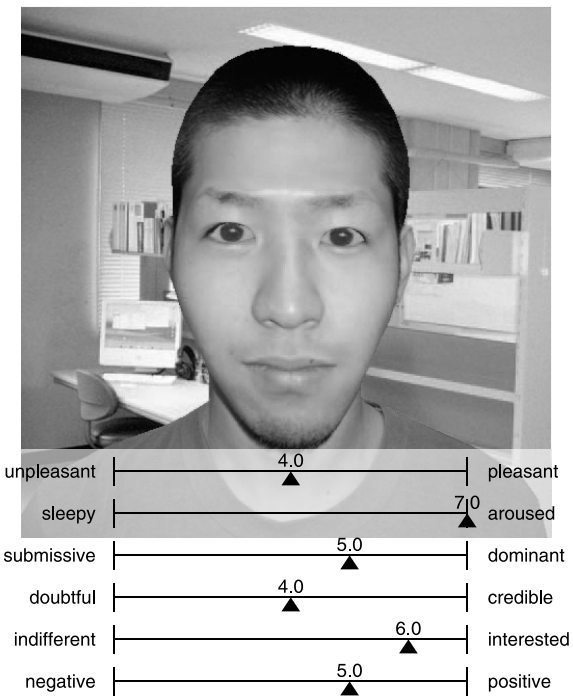

Fig. 5 Another example of generated facial expression.

Figure 4 exemplifies the facial expression generation from an arbitrary emotional state. Here, the specified state is unpleasant, aroused, somewhat submissive, somewhat doubtful, somewhat interested and negative, which may correspond to some emotion like "disgust." Another example is shown in Fig. 5. The state is neither pleasant nor unpleasant, very aroused, somewhat dominant, neither credible nor doubtful, interested and somewhat positive. Although it is difficult to find an appropriate word to explain, such state is typical and often found in the dialogue situations where a person is interested in another person's unexpected talk. From Figs. 4 and 5, it can be said that both facial expressions generated seem to fit well to the given emotional states.

5. Subjective Evaluation Test

In order to verify the suitability of facial expression that the proposed method generates, subjective evaluation tests were performed.

The evaluation scheme was the opinion test. The tests were performed for FUE, FMS and MKK sets, each of which was composed of 219, 100 and 100 utterances. Because the NN described in Sect. 4 was trained with FUE set, the test for FUE set was in a speaker-matched condition. Contrastively, the test conditions for FMS and MKK sets were speaker-unmatched, which were meant to verify the generality of NN trained with a single speaker's parallel data. For each utterance, auditory stimulus and visual stimulus were presented simultaneously. Subjects were asked to judge the suitability of the presented facial expression to the presented utterance with the 5-grade scale (1:unsuitable, 3:neutral, 5:suitable).

Detailed descriptions of the procedure was as follows. The neutral face image was displayed on a monitor. When a subject was ready, he was to press a key. Then, the image changed to the facial expression for the first utterance. At the same time, the utterance was played back through a headphone. No lip-sync was performed in displaying the image. Once the playback finished, the face image was reset to the neutral. The subject was allowed to playback the sound and see the facial expression again if needed. Finally, he rated the suitability of the facial expression with 1 to 5.

The subjects included one graduate school student and 4 undergraduates with preliminary knowledge of speech science but no specific knowledge of avatar, human-like agent, or facial image processing.

The presented order of the stimuli was the same as the original sets. In other words, the subjects were given its discourse context. The facial expression for each utterance was randomly chosen as one of the following:

a) (manual) hand-made facial expression,
b) (proposed) automatically-generated facial expression,

where a) corresponds to the data described in Sect. 3.3, and b) is generated with the proposed method using the averaged emotional state rating given by the UU database. In generat-

ing facial expressions for FUE set, the NN was trained with nine-tenth of the whole set which did not include the utterance, just like the way that generalization for unseen data was examined with 10-fold cross validation as described in Sect. 4. The subjects were presented from utterance #1 to #219, and again from #1 to #219, in FUE set. In total, the number of stimuli was 219×2 = 438, which exactly includes one facial expression for both a) and b) for each utterance. The tests for FMS set and MKK set were performed in the same way except that the NN was trained with the whole FUE set.

The results of the subjective evaluation tests are shown in Table 1. The histograms of Mean Opinion Score (MOS) distribution for 219 utterances in FUE set are also illustrated in Figs. 6 and 7. Evaluation result for FUE set, where training and test data are of same speaker, shows that the averaged MOS over the set was 3.97 and 3.86 for a) and b), respectively. Although the difference is statistically significant (paired t-test, $p < 0.05$), Figs. 6 and 7 imply that the difference is not quite apparent. To judge the difference, Effect Size [14] for facial expression generation methods was computed, which is also shown in Table 1. Effect Size d value is obtained by the following formula:

$$d = \frac{|\mu_{\text{manual}} - \mu_{\text{proposed}}|}{\sqrt{(\sigma^2_{\text{manual}} + \sigma^2_{\text{proposed}})/2}}, \quad (1)$$

where μ_{manual} and μ_{proposed} denote the mean of MOSs, and σ^2_{manual} and $\sigma^2_{\text{proposed}}$ denote the variance of MOSs. According to Cohen's suggestion that d values of 0.2, 0.5 and 0.8 represent small, medium and large [14], the d value of 0.195 for matched condition can be considered as a small difference. Therefore, we can conclude that the quality difference between manual and automatically-generated facial expressions is small with respect to the consistency of facial and vocal expressions as paralinguistic cues.

We further looked into the relationship between MOS and emotional states for the utterances to find out factors that potentially affecting the quality of generated facial expression. Figure 8 illustrates the relationship between MOS and rated pleasant-unpleasant value given by the UU database for FUE set. It can be said that overall quality is high for emotionally "prominent" utterances (unpleasant ones with below 3 and pleasant ones above 5). On the other hand, all of the utterances whose suitability of generated facial expressions were evaluated as around or below 3 in MOS are emotionally non-prominent (around 4; neither pleasant nor unpleasant). This fact implies that it is relatively difficult to generate facial expressions for utterances with subtle emo-

Table 1 Averaged MOSs over FUE (speaker-matched), FMS (speaker-unmatched) and MKK (speaker-unmatched) sets. The effect size (ES) for facial expression generation (manual/proposed) is also shown for each speaker.

		averaged MOS		ES
		a) manual	b) proposed	
matched	FUE	3.97	3.86	0.195
unmatched	FMS	4.01	3.76	0.467
	MKK	4.17	3.80	0.767

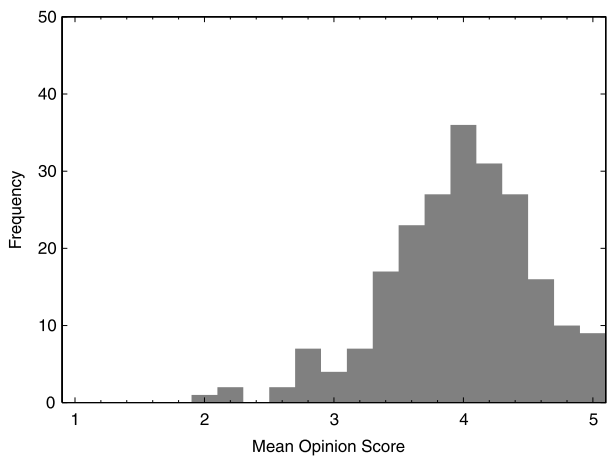

Fig. 6 MOS distribution for hand-made facial expressions for FUE set.

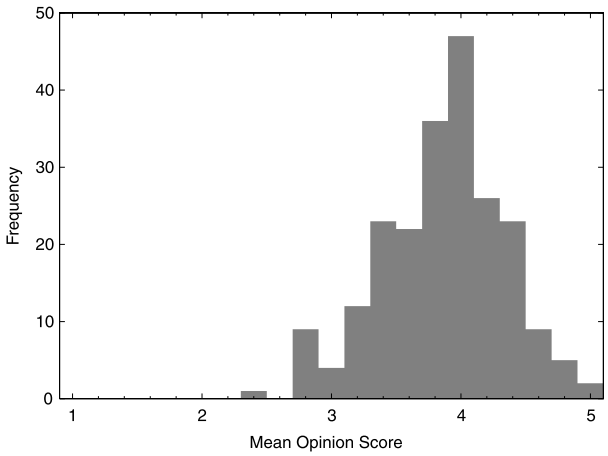

Fig. 7 MOS distribution for the facial expressions generated by the proposed method for FUE set.

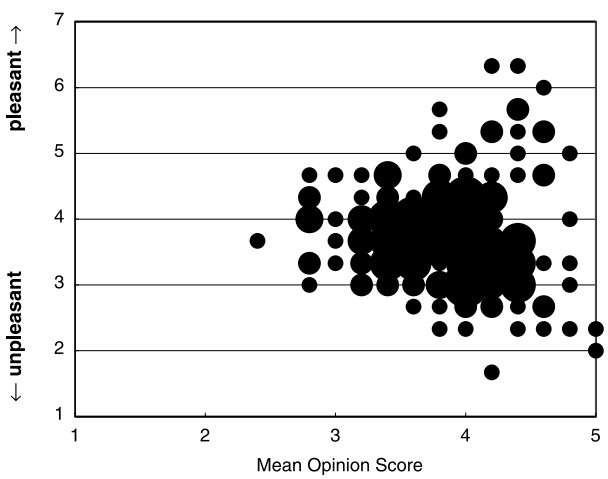

Fig. 8 MOS Distribution for FUE set with regard to perceived pleasantness of utterances. The size of each bubble is proportional to the number of instances.

tion. To achieve higher quality for such utterances, detailed analysis of face images in subtle emotional states will be necessary.

Also for unmatched conditions, the averaged MOSs (3.76 and 3.80) were not so worse than that of matched condition, as shown in Table 1. For both cases, however, the difference between a) and b) is statistically significant ($p < 0.01$), and their Effect Sizes can be interpreted as medium to large. These results suggest that personality factors should be taken into account for optimizing the mapping function between emotional states and facial expressions.

6. Conclusion

In this paper, we proposed a framework for generating facial expressions in daily conversation. It enables virtual characters to express facial expressions from arbitrary emotional states that are described in the form of dimensions. The effectiveness of proposed method was verified by subjective evaluation tests using a series of conversational speech of two females and a male. As the result, the Mean Opinion Score for the matched speaker with respect to the suitability of generated facial expressions was 3.86, which was close to that of hand-made facial expressions.

The learning of the mapping function and evaluation was based on the UU database, which is a collection of task-oriented dialogues. Although the speaking style of the utterances used is much closer to daily conversation than acted emotional speech, still there is difference. For example, the database contains few utterances in relaxed, calm states. Additional data will be needed to cover the emotional states of broader range for some applications, but we think that required size will not be larger than that of FUE set.

In the current work, a facial expression is assumed to be a still image. As recent studies suggested the importance of dynamic properties of facial expressions [15], [16], incorporation of facial movement is a promising direction for improvement.

Currently, we are working on integrating emotional state recognition from speech signal [6] with the facial expression generation described in this paper.

References

[1] G.L. Trager, "Paralanguage: A first approximation," Studies in Linguistics, vol.13, nos.1–2, pp.1–12, 1958.
[2] P. Ekman, "Basic emotions," in Handbook of Cognition and Emotion, ed. T. Dalgleish and M. Power, pp.169–200, John Wiley, New York, 1999.
[3] N. Campbell, "The JST/CREST ESP project — A mid-term progress report," 1st JST/CREST Intl. Workshop Expressive Speech Processing, pp.61–70, 2003.
[4] P. Ekman and W. Friesen, Facial Action Coding System: A technique for the measurement of facial movement, Consulting Psychologists Press, Palo Alto, 1978.
[5] R. Cowie and R.R. Cornelius, "Describing the emotional states that are expressed in speech," Speech Commun., vol.40, no.1–2, pp.5–32, 2003.
[6] H. Mori and H. Kasuya, "Voice source and vocal tract variations as cues to emotional states perceived from expressive conversational speech," Proc. Interspeech 2007, pp.102–105, 2007.
[7] J.A. Russell, "How shall an emotion be called?," in Circumplex Models of Personality and Emotions, ed. R. Plutchik and H.R. Conte, pp.205–220, APA, Washington, 1997.
[8] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellentz, and J.G. Taylor, "Emotion recognition in human-computer interaction," IEEE Signal Process. Mag., vol.18, no.1, pp.32–80, 2001.
[9] H. Mori, H. Kasuya, M. Nakamura, and M. Amanuma, "Some considerations for designing spoken dialogue database from the viewpoint of paralinguistic information," Acoust. Sci. & Tech., vol.24, no.6, pp.376–378, 2003.
[10] R.W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," IEEE Trans. Pattern Anal. Mach. Intell., vol.23, no.10, pp.1175–1191, 2001.
[11] J.A. Russell and G. Lemay, "Emotion concepts," in Handbook of Emotions, 2nd ed., ed. M. Lewis and J.M. Haviland-Jones, pp.491–503, Guilford, New York, 2000.
[12] H. Mori, H. Aizawa, and H. Kasuya, "Consistency and agreement of paralinguistic information annotation for conversational speech," J. Acoust. Soc. Jpn., vol.61, no.12, pp.690–697, 2005.
[13] H. Prendinger and M. Ishizuka, eds., Life-Like Characters — Tools, Affective Functions, and Applications, Springer, Berlin, 2004.
[14] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, 2nd ed., Lawrence Erlbaum, Hillsdale, NJ, 1988.
[15] K.L. Schmidt and J.L. Cohn, "Dynamics of facial expression: Normative characteristics and individual differences," Proc. Intl. Conf. Multimedia & Expo, pp.728–731, 2001.
[16] J.F. Cohn, L.I. Reed, T. Moriyama, J. Xiao, K. Schmidt, and Z. Ambadar, "Multimodal coordination of facial action, head rotation, and eye motion during spontaneous smiles," Proc. Intl. Conf. Automatic Face & Gesture Recognition, pp.129–135, 2004.

Hiroki Mori received B.E., M.E. and Ph.D. degrees from Tohoku University, in 1993, 1995 and 1998, respectively. He was with the Graduate School of Engineering, Tohoku University in 1998. He is now an Associate Professor of Utsunomiya University. His research interests include speech recognition, speech synthesis, spoken dialogue systems, and natural language processing. He is a member of the Acoustical Society of Japan and IPSJ.

Koh Ohshima received B.E. degree from Utsunomiya University in 2007. He is currently working toward the Master's degree. His research interests include speech technology application to communication devices.