Development of a Speech-Driven Embodied Entrainment Character System with Pupil Response

Yoshihiro Sejima[1(✉)], Yoichiro Sato[1], Tomio Watanabe[1], and Mitsuru Jindai[2]

[1] Faculty of Computer Science and System Engineering, Okayama Prefectural University, 111 Kuboki, Soja-Shi, Okayama, Japan
sejima@ss.oka-pu.ac.jp
[2] Graduate School of Science and Engineering, University of Toyama, 3190 Gofuku, Toyama-Shi, Toyama, Japan

Abstract. We have developed a speech-driven embodied entrainment character called "InterActor" that had functions of both speaker and listener for supporting human interaction and communication. This character would generate communicative actions and movements such as nodding, body movements, and eyeball movements by using only speech input. In this paper, we analyze the pupil response during the face-to-face communication and non-face-to-face communication with the typical users of the character system. On the basis of the analysis results, we enhance the functionalities of the character and develop an advanced speech-driven embodied entrainment character system for expressing the pupil response.

Keywords: Human interaction · Nonverbal communication · Avatar-Mediated communication · Line-of-Sight · Pupil response

1 Introduction

In human face-to-face communication, not only verbal messages but also nonverbal behavior such as nodding, body movement, line-of-sight and facial expression are rhythmically related and mutually synchronized between talkers [1]. This synchrony of embodied rhythms in communication is called entrainment, and it enhances the sharing of embodiment and empathy unconsciously in human interaction [2].

In our previous work, focusing on the line-of-sight of embodied interaction, we analyzed the eyeball movement using line-of-sight measurement device and proposed an eyeball movement model [3]. In addition, we earlier developed a speech-driven embodied entrainment character called "InterActor" that has functions of both speaker and listener for supporting human interaction and communication. This character would generate not only communicative movements and actions such as nodding, body movements, and blinking that are coherently related to voice input, but also line-of-sight actions such as eye contact and glancing aside on the basis of the proposed model. The effectiveness of the proposed eyeball movement model and the developed

S. Yamamoto (Ed.): HIMI 2015, Part II, LNCS 9173, pp. 378–386, 2015.
DOI: 10.1007/978-3-319-20618-9_38

character was demonstrated by the sensory evaluation methodology adopted in a communication experiment [4].

On the other hand, it is confirmed that the pupil response of human is enlarged or reduced in order to adjust the amount of light in eyeball [5]. In addition, the pupil response has a function which relates human emotions such as human-interest and degree of stress [6, 7]. Moreover, during recent years, newer measurement methods that evaluate the level of human-interest based on the pupil response were proposed and a line-of-sight measurement system was developed without calibration [8, 9]. These previous researches were targeted at the interaction between human and artifact device such as display, picture, and poster. However, an approach that focuses on the pupil response during human interaction has not been designed thus far. Therefore, it is essential and imperative to develop an embodied communication system that enhances the empathy in human interaction on the basis of analyzing the pupil response unconsciously.

In this paper, focusing around the pupil response in human face-to-face communication, we perform an analysis of the pupil response for human interaction using an embodied communication system with a line-of-sight measurement device. On the basis of the analysis results, we enhance our existing character and develop an advanced speech-driven embodied entrainment character system for supporting human interaction and communication. The system uses only speech input to generate a character's pupil response as well as nodding and body movements.

2 Analysis of Pupil Response

In order to analyze the pupil response in human interaction and communication, a communication experiment was carried out using the embodied communication system with line-of-sight measurement device.

2.1 Experimental System

In this experiment, an embodied communication system was developed to measure the line-of-sight and analyze the typical pupil response of talkers during human interaction and communication. The experimental setup is shown in Fig. 1. This system consists of a Windows 7 workstation (CPU: Corei7 2.93 GHz, Memory: 8 GB, Graphics: NVIDIA Geforce GTS250), magnetic sensors (Polhemus FASTRAK), a headset (Logicool H330), and a line-of-sight measurement device. In this system, two talkers were seated face-to-face across tables. The distance between the talkers was 1200 mm based on a personal space. In addition, in order to compare the communication style, non-face-to-face communication scene was generated by inserting a partition between the talkers. The size of partition was 1820 x 910 mm. The talker's line-of-sight was measured by the developed line-of-sight measurement device [3]. Figure 2 shows the outline of the device. The dichroic mirror in the device has a function that transmits visible rays, and reflects infrared rays. The reflected image of a talker's eyeball was input to a PC through an A/D converter. The pupil movement was measured according to the following procedure. First, the binary image was generated using the reflected image on

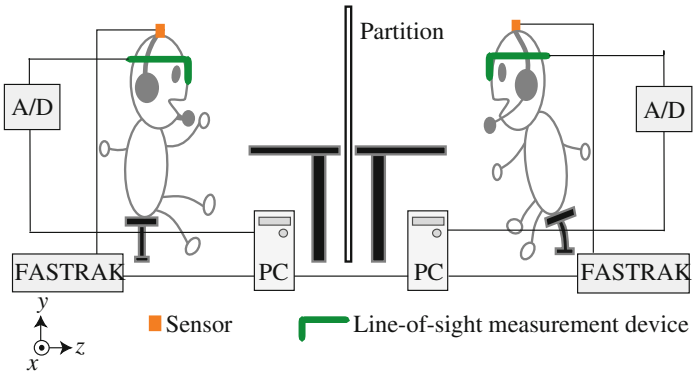

Fig. 1. Setup of the experimental system

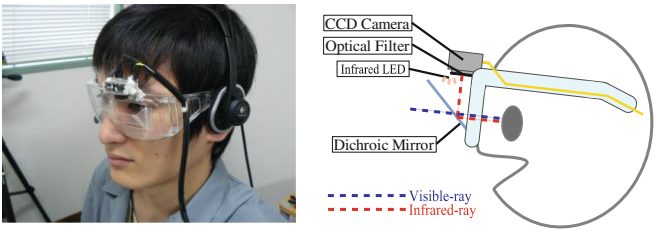

Fig. 2. Line-of-sight measurement device

the basis of the prepared threshold by brightness. Next, the position and size of the talker's pupil were calculated using the ellipse fitting function in Open Source Computer Vision Library (OpenCV). Here, the sample rate was 30 fps. The positions and angles of talker's head movement were measured by magnetic sensors placed on the top of talker's head at 30 Hz. The voice was sampled using 16 bits at 11 kHz with a headset. The measured data was recorded on an HDD in real-time.

2.2 Experimental Method

The experiment was performed under the conditions that the talkers interchangeably played the roles of a speaker and a listener called "Role play experiment," and was later engaged in a free conversation called "Free conversation experiment". In Role play experiment, children's stories were introduced as conversational topics. In Free conversation experiment, conversational topics were not specified. In these experiments, the following two modes were compared: in one mode (a), there was no partition between talkers and in the other (b), there was a partition between the talkers. The subjects were 10 pairs of talkers (10 males and 10 females). Each pair was presented with the two modes in a random order.

The experimental procedure adopted was as follows:

- First, the subjects selected two well-known children's stories and confirmed their summaries.
- Next, the devices for the pupil measurement were calibrated and the subjects used the system for around 2 min freely for familiarization.
- Then, they were asked to talk on conversational topics for 3 min in separate roles (speaker and listener) in each mode.
- Then, the roles of the speaker and listener were interchanged, and the subjects communicated in each mode for 3 min in the abovementioned manner.
- Lastly, the subjects engaged free conversation for 3 min in each mode.

2.3 Analysis of Pupil Response

Focusing on the amount of the pupil response change, talker's pupillary area was analyzed in the each experiment. Here, three subjects were removed from analysis, because the calibration data collected was not accurate. The pupillary area was calculated by the estimated ellipse size. Figure 3 shows the example of time changes of pupillary area. On the basis of this figure, we defined a "unit." The unit is the period between an eye-blink and the next eye-blink. We calculated the average of pupillary area in units. The evaluation of pupillary area was done by the ratio using the calibration data as the standard. Figure 4 shows the result of ratio of pupillary area. In this

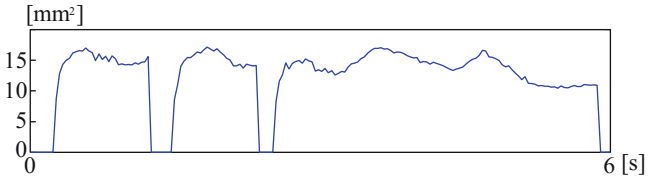

Fig. 3. Example of changes of pupillary area with time

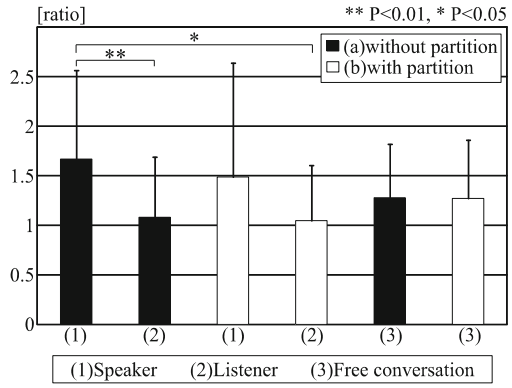

Fig. 4. Result of average of pupillary area

figure, it is showed that the speaker's pupillary area was enlarged about 1.5 times in both two modes. It is considered that the stress of task as "lecture" was prevalent with the speaker. In addition, from the results of the t-test, in the Role play experiment, the significance level between the speaker and the listener is 1% in mode (a). However, mode (b) which does not visualize the body of the talkers (owing to the partition between the speakers) has no significance level between the speaker and the listener. This result shows that recognizing the partner enhances the effectiveness of the conversation. In Free conversation experiment, there was no significance level, even though the role play of speaker and listener was interchanged between the subjects.

Thus, these results demonstrated that the pupil response has relations with the speaker's speech in human interaction and communication.

3 Development of a Speech-Driven Embodied Entrainment Character System with Pupil Response

3.1 Concept

The core concept of this research is shown in Fig. 5. In this research, an advanced speech-driven embodied entrainment character called "InterActor" is developed to express the pupil response unconsciously on the basis of speech input. InterActor is an interactive avatar that represents the talker's nonverbal behaviors. The talkers can realize embodied communication with or through the InterActors during which the pupil of character enlarged. By expressing the pupil responses, the impression of character such as vividness and interest level is improved. Therefore, it is expected that the human embodied interaction is supported for enhancing sharing of embodied rhythms such as nodding, body movements, and pupil response unconsciously.

3.2 Interaction Model

In order to support human interaction and communication, we have already developed a speech-driven embodied entrainment character called InterActor, which has the functions of both speaker and listener. The listener's interaction model includes a nodding reaction model which estimates the nodding timing from a speech ON–OFF pattern and a body reaction model linked to the nodding reaction model [10]. The

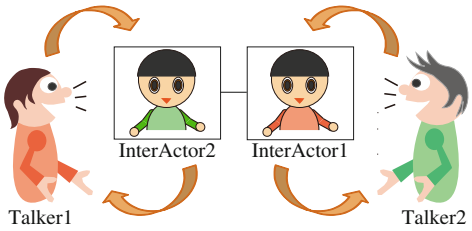

Fig. 5. Concept

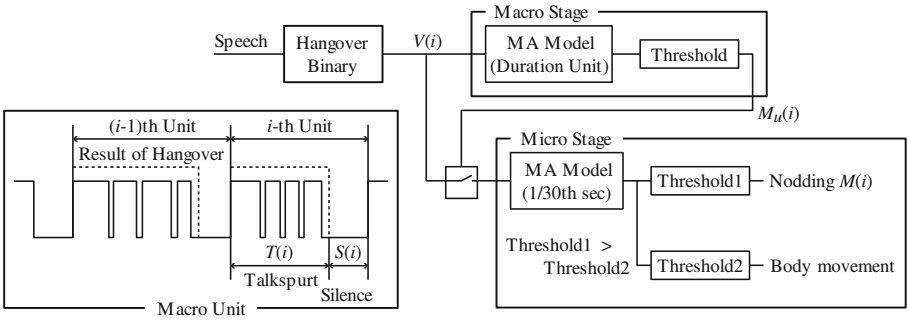

Fig. 6. Interaction model

timing of nodding is predicted using a hierarchy model consisting of two stages - macro and micro (Fig. 6). The macro stage identifies a nodding response, if any, in a duration unit which consists of a talkspurt episode $T(i)$ and the following silence episode $S(i)$ with a hangover value of 4/30 s. The estimator M_u (i) is a moving-average (MA) model, expressed as the weighted sum of unit speech activity $R(i)$ in Eqs. (1) and (2). When $M_u(i)$ exceeds a threshold value, nodding $M(i)$ is also an MA model, estimated as the weighted sum of the binary speech signal $V(i)$ in Eq. (3).

$$M_u(i) = \sum_{j=1}^{J} a(j)R(i-j) + u(i) \tag{1}$$

$$R(i) = \frac{T(i)}{T(i) + S(i)} \tag{2}$$

$a(j)$: linear prediction coefficient
$T(i)$: talkspurt duration in the i th duration unit
$S(i)$: silence duration in the i th duration unit
$u(i)$: noise

$$M(i) = \sum_{j=1}^{K} b(j)V(i-j) + w(i) \tag{3}$$

$b(j)$: linear prediction coefficient
$V(i)$: voice
$w(i)$: noise

The body movements are related to the speech input in that the neck and one of the wrists, elbows, arms, or waist are operated when the body threshold is exceeded. The threshold is set lower than that of the nodding prediction of the MA model, which is expressed as the weighted sum of the binary speech signal to nodding. In other words,

when the InterActor functions as a listener for generating body movements, the relationship between nodding and other movements is dependent on the threshold values of the nodding estimation.

3.3 Developed System

The setup for this system is shown in Fig. 7. The virtual space is generated using Microsoft DirectX 9.0 SDK (June 2010) and a Windows 7 workstation (CPU: Corei7 2.93 GHz, Memory: 8 GB, Graphics: NVIDIA Geforce GTS250). The voice is sampled using 16 bits at 11 kHz with a headset (Logicool H330). The voice data is transmitted through the Ethernet in each system. The frame rate to represent CG characters is 30 frames per second.

In this system, the character was developed for expressing the pupil responses. Figure 8 shows the developed characters. The eyeball of this character was the focus of the experiment. The enhanced eyeball is shown in Fig. 9. The eyeball consists of the white of the eye, iris, and pupil by 3D model. It is developed with a sense of depth of iris by forcing the surface into the core of eyeball. The smooth pupil response is realized by generating the forward and back movement of the black 3D model (pupil) on the Z-axis at 0.05 *pixel/frame*. The size of enlarged pupil is a half as large as normal size on the basis of the experimental results (Fig. 10).

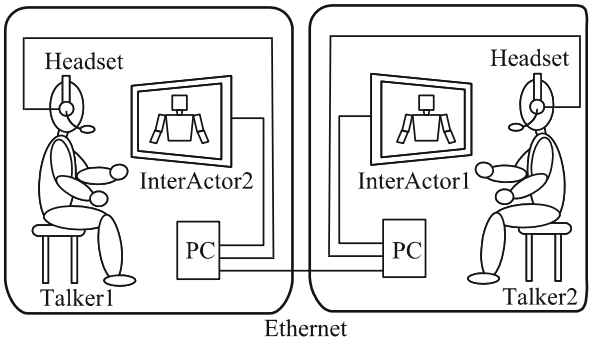

Fig. 7. Set up of the developed system

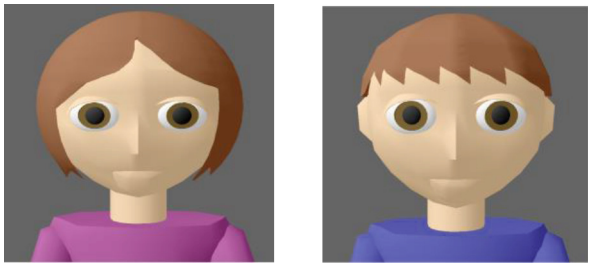

Fig. 8. Developed characters

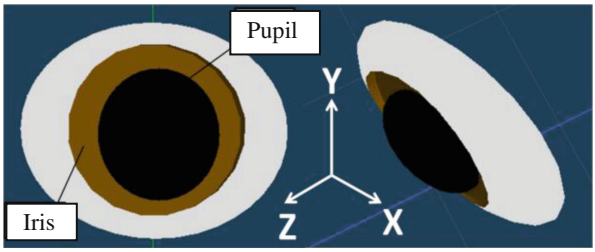

Fig. 9. 3D model of eyeball

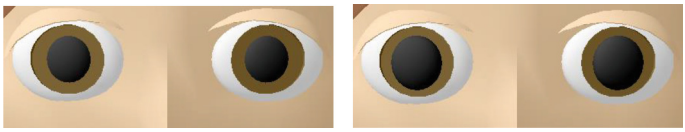

Fig. 10. Example of pupil response

With the developed eyeball, when talker1 speaks to InterActor2, InterActor2 responds to talker1's utterance with appropriate timing through its entire body motions, including nodding, blinking, and actions, in a manner similar to the body motions of a listener. The nodding movement of the developed character is the falling-rising motion of the head in the front-back direction at 0.1 rad/frame. The blinking movement is the same as the nodding movement at 0.5 rad/frame with an exponential distribution [11]. The body movement is defined as the backward and forward motion of the body at a speed of 0.025 rad/frame. In addition, InterActor1 generates pupil response based on the talker1's utterance. In this manner, two remote talkers can enjoy a conversation via InterActors within a communication environment in which the sense of unity is shared by embodied entrainment and empathy unconsciously, like typical human conversations.

4 Conclusion

In this paper, we analyzed the pupil response in the face-to-face and non-face-to-face communication. The results of the analysis are summarized as follows. Under the condition that the roles of speaker and listener are fixed, the speaker's pupil is enlarged about 1.5 times in face-to-face communication specially. The results demonstrated that the pupil response is related to speech in human embodied interaction and communication. On the basis of the analysis, we developed an advanced speech-driven embodied entrainment character system for expressing the pupil response. The system uses only speech input to generate a character's pupil response, but produces the outcome of typical human conversations.

Acknowledgments. This work was supported by JSPS KAKENHI Grant Number 26750223, 26280077, 25330239.

References

1. Condon, W.S., Sander, L.W.: Neonate movement is synchronized with adult speech. Science **183**, 99–101 (1974)
2. Watanabe, T.: Human-entrained embodied interaction and communication technology for advanced media society. In: Proceedings of 16th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN2007), pp. 31–36 (2007)
3. Sejima, Y., Watanabe, T., Jindai, M.: An embodied communication system using speech-driven embodied entrainment characters with an eyeball movement model. Trans. Jan. Soc. Mech. Eng. Ser. C **76**(762), 340–350 (2010). (in Japanese)
4. Sejima, Y., Watanabe, T., Jindai, M.: An avatar-mediated speech-driven embodied communication system with an eyeball movement model. In: Proceedings of IADIS International Conference e-Society 2013, pp. 291–298 (2013)
5. Matthew, L.: Area and brightness of stimulus related to the pupillary light reflex. J. Opt. Soc. Am. **24**(5), 130 (1934)
6. Hess, E.H.: Attitude and pupil size. Sci. Am. **212**(4), 46–54 (1965)
7. Iijima, A., Kosugi, T., Kiryu, T., Matsuki, K., Hasegawa, I., Bando, T.: Evaluation of stressed condition using pupillary responses. Trans. Japn Soc. Med. Biol. Eng. **49**(6), 946–951 (2011). (in Japanese)
8. Nagamatsu, T., Kamahara, J., Tanaka, N.: User-Calibration-Free Gaze Estimation Method Using Both Eyes Model. Industrial Publishing co. ltd., Japan, vol. 23, no. 6, pp. 29–34. (2012) (in Japanese)
9. Kikuchi, K., Takahira, H., Ishikawa, R.: Development of a device to measure movement of gaze and hand. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. **97**(2), 534–537 (2014)
10. Watanabe, T., Okubo, M., Nakashige, M., Danbara, R.: Interactor: speech-driven embodied interactive actor. Int. J. Hum. Comput. Interact. **17**, 43–60 (2004)
11. Watanabe, T., Yuuki, N.: A voice reaction system with a visualized response equivalent to nodding. Adv. Hum. Factors/Ergon. **12A**, 396–403 (1989)