A Proposal of Objective Evaluation Measures Based on Eye-Contact and Face to Face Conversation for Videophone

Keiko Masuda, Ryuhei Hishiki, and Seiichiro Hangai(✉)

Department of Electrical Engineering, Faculty of Engineering,
Tokyo University of Science, 6-3-1 Niijuku, Katsushika, Tokyo 1258585, Japan
{masuda,hangai}@ee.kagu.tus.ac.jp
http://www.tus.ac.jp/en/fac/p/index.php?678

Abstract. In order to realize eye-contact and face to face communication, videophone or virtual conversational system which takes gaze line into consideration. Although many systems have been studied and reported, there is no objective measure for evaluating the quality of conversations. In this paper, we propose two objective measures such as the eye-contact conversation ratio (ECCR) and the face to face conversation ratio (FFCR) for evaluating the communication quality. By changing the position of camera from the above to the center of display, the ECCR increases from 24.3% to 25.3% in talking and decreases from 33.1% to 25.6% in listening. It is also found that the FFCR improved from 74.7% to 88.0% by centering a camera.

1 Introduction

In popular personal videophone system using PCs and Tablets, listening and talking with downcast eyes is inevitable. This is because a camera is installed above a display and there is no gaze line matching between two persons. In the teleconference system for multiple persons, half mirrors and cameras were used for realizing eye contact talks [1]. In another study, the picture plane was rotated for compensating the gaze direction, and the improvement of subjective perception was reported by the number of votes by 52 subjects [2]. Gaze correction method [3] and multi-viewpoint videos merging method [4] have been reported the improvement of eye-contact communication, too. However, there is no objective evaluation result in those studies.

Generally, in a natural conversation, eye-contact and face to face communication can be observed frequently, and those human behaviors should be taken into account by evaluating a system. In e-learning applications, eye mark recorder which recorded the fixation point movement on the view was applied to analyze the effectiveness of presentation methods [5].

In this paper, we define two objective measures, i.e., the ECCR and the FFCR, and show the experimental results using eye mark recorder [6] with changing the position of camera from the above to the center of display. We also discuss what makes a conversation natural using videophone and virtual conversational system.

S. Battiato et al. (Eds.): ICIAP 2017, Part II, LNCS 10485, pp. 343–350, 2017.
https://doi.org/10.1007/978-3-319-68548-9_32

2 Human Behaviors in Conversation

Mutual gaze during natural conversation is one of important interactions [7]. However, in personal videophone system, inconsistent gaze behavior, e.g., gaze at partner's clothes or out of display, is frequently observed.

Figure 1 shows the relationship between the gaze at partner's eye $G_{eye}(t)$, the gaze at partner's face $G_{face}(t)$, the Talk by subject $T_s(t)$, and the Talk by partner $T_p(t)$, and behavioral states. As each feature is represented by ON(1) or OFF(0), there are 16 behavioral states. In this figure, the duration of $G_{eye}(t) = 1$ and $G_{face}(t) = 1$, when $T_s(t) = 1$ or $T_p(t) = 1$, is the most important state. As shown in Fig. 1, we define Eye-Contact Conversation (ECC) state in which $G_{eye}(t) = 1$ and $T_s(t) = 1$ or $T_p(t) = 1$, and Face to Face Conversation (FFC) state in which $G_{face}(t) = 1$ and $T_s(t) = 1$ or $T_p(t) = 1$. The former is marked up by black bar and the latter is by gray bar in the bottom of figure.

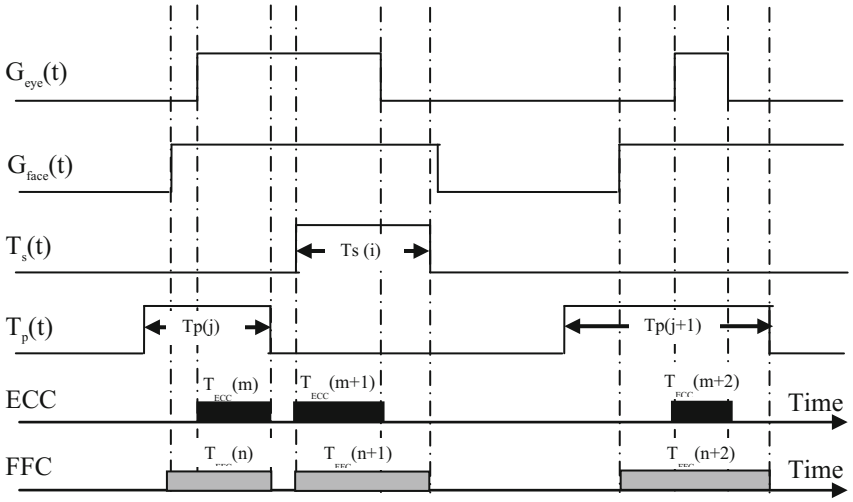

Fig. 1. Relationship between 4 features and 2 states

2.1 Eye-Contact Conversation Ratio (ECCR)

In order to estimate the eye-contact conversation objectively, we add up ECC durations and calculate ECC ratio by the following equation,

$$ECCR = \frac{\sum_{m=1}^{M} T_{ECC}(m)}{\sum_{i=1}^{I} T_s(i) + \sum_{j=1}^{J} T_p(j)} \times 100 \quad (1)$$

where, $T_{ECC}(m)$ is the m-th duration of ECC state, $T_s(i)$ is the i-th duration of subject's talk, and $T_p(j)$ is the j-th duration of partner's talk.

From the equation, ECCR presents eye-contact conversation ratio in both talking period and listening period.

2.2 Face to Face Conversation Ratio (FFCR)

As same as ECCR, we sum up FFC durations and calculate FFC ratio by the following equation,

$$FFCR = \frac{\sum_{n=1}^{N} T_{FFC}(n)}{\sum_{i=1}^{I} T_s(i) + \sum_{j=1}^{J} T_p(j)} \times 100 \quad (2)$$

where, $T_{FFC}(n)$ is the n-th duration of FFC state.

From the equation, FFCR presents face to face conversation ratio in both talking period and listening period. As shown in Fig. 1, FFC duration includes ECC duration, because eye is a part of face.

3 Experimental System

In order to evaluate the eye-contact conversation and the face to face conversation using videophone with different camera position, we have developed a videophone system, in which the camera position can be changed. Gaze point is recorded by the eye mark recorder which uses the infrared reflection of pupil/cornea, and the decision whether the gaze position of subject is at face or at eye or at others is made by analyzing the recorded images. Conversations are also recorded and separated into subject's talk and partner's talk after noise reduction.

In this section, the developed videophone system and the flow of signal processing are described.

3.1 Videophone System

The developed videophone system is shown in Fig. 2. A half mirror is located in front of a subject with 45° angle for realizing face to face conversation with the image of a partner. The flipped horizontal image is displayed on the monitor for avoiding left and right being reversed.

The height of the camera can be changed in any position. In this experiment, we use two positions such as center position and above position which simulates PC's camera. Two sets of systems are used in the experiment. Specification of each system is as follows,

Display size: 24.1 in. LCD
Videophone application: Skype
Left and Right Reverse: ManyCam[1]
Camera: 640pixels (H) by 480pixels (W), 24bit color, 30fps
Audio: fs = 44.1 kHz, 16bit

[1] https://manycam.com/.

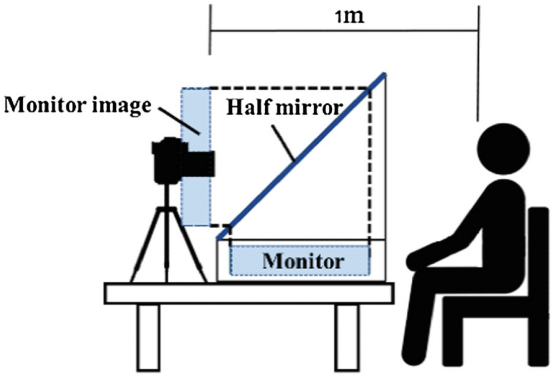

Fig. 2. Developed videophone system

A scene of the experiment using the developed videophone system is shown in Fig. 3.

Fig. 3. Experimental setup

3.2 Gaze Point Estimation

The gaze points and audio information of the subject wearing the eye mark recorder (EMR-9) [6] are recorded.

The recorder measures the sight angle of the subject based on the infrared reflection image in the cornea and the pupil movement. The detection range is ±40° in horizontal and ±20° in vertical. The gaze points (Left: + , Right: □) and a parallax corrected gaze point: ○ are displayed on the image (640 × 480 pixels) taken by field of view cameras installed at the brim of a cap as shown in Fig. 4. The image was being recorded while conversation and analyzed together with recorded voice after the experiment to look into the location of the parallax gaze point.

Figure 4 shows examples of (a) “Eye-Contact Conversation in talking” scene, and (b) “Face to Face Conversation in listening” scene.

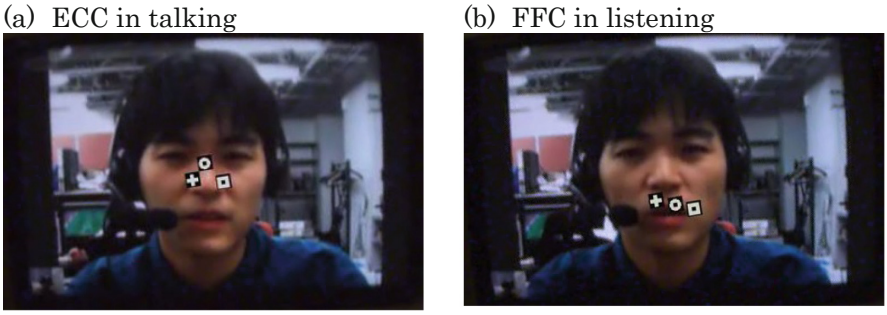

Fig. 4. Example of recorded image from EMR-9

Face area and eye area are manually determined by the face features such as skin color, eye-brow, eye, nose, mouth, and tin as shown in Fig. 5.

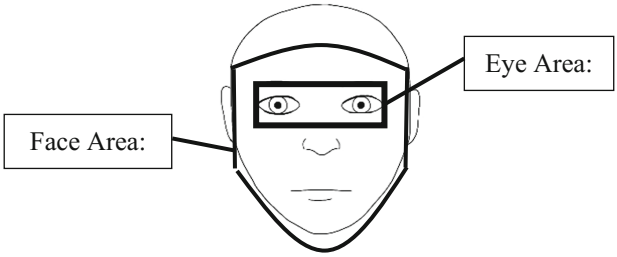

Fig. 5. Face area and eye area

3.3 Signal Processing Flow

Figure 6 shows the flow of signal processing to get 4 signals, i.e., $G_{eye}(t)$, $G_{face}(t)$, $T_s(t)$, and $T_p(t)$, and 4 states, i.e., T_{ECC} and T_{FFC} in talking/listening. In this study, $G_{eye}(t)$ and $G_{face}(t)$ are extracted manually, and the Talk by subject $T_s(t)$ and the Talk by partner $T_p(t)$ are extracted automatically based on the audio signal power.

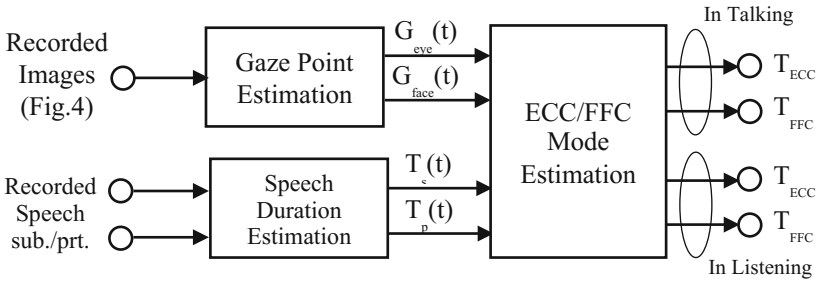

Fig. 6. Signal processing flow

4 Experimental Results and Discussion

10 male persons (Age: 22–24) were divided into 5 groups and made free conversation for 6 min. or more. After 1 min. passed, image including gaze point shown in Fig. 4 and speech were recorded for 5 min. and analyzed.

4.1 ECCR and FFCR in Higher Camera Position

It is expected that both a subject and their partner using PC based videophone are talking or listening with downcast eyes. This degrades the communication quality and leads to low ECCR and FFCR. Table 1 shows the ECCR and the FFCR in the conversation of 5 min. Total talking time of subject and Total talking time of partner are also indicated in seconds.

From Table 1, averaged ECCR of five subjects is 29.0% and the deviation is not large. On the other hand, averaged FFCR shows 74.7% even if a partner talks or listens with downcast eyes. Five FFCRs depend on subjects, and varies between 53.9% and 89.3%.

Table 1. ECCR, FFCR, and talking time in higher camera position

	ECCR	FFCR	$\sum_{m=1}^{M} Ts(m)$	$\sum_{n=1}^{N} Tp(n)$
Subject 1	31.5%	69.1%	94.5 s	192.8 s
Subject 2	30.3%	76.8%	84.5 s	169.9 s
Subject 3	28.1%	89.3%	168.6 s	103.3 s
Subject 4	26.1%	84.5%	124.1 s	135.5 s
Subject 5	29.0%	53.9%	165.0 s	104.4 s
Average	29.0%	74.7%	127.3 s	141.2 s

In order to inspect ECCR and FFCR in detail, we separate them into the ratios in talking and listening, and summarized as shown in Table 2. The suffix "T" and "L" shows "talking" and "listening" respectively.

Table 2. ECCR and FFCR in talking and listening in higher camera position

	$ECCR_T$	$ECCR_L$	$FFCR_T$	$FFCR_L$
Subject 1	27.9%	33.3%	61.2%	73.1%
Subject 2	26.1%	32.4%	70.0%	80.1%
Subject 3	20.7%	40.2%	89.4%	89.1%
Subject 4	14.6%	36.5%	74.1%	94.1%
Subject 5	32.1%	22.9%	51.7%	57.4%
Average	24.3%	33.1%	69.3%	78.8%

Except for the ECCR of subject 5, both averaged $ECCR_T$ and $FFCR_T$ are less than averaged $ECCR_L$ and $FFCR_L$, respectively. This means that almost all subjects watch the partner's eye and face in listening rather than in talking. Also, it is found that the subjects watch the partner's face rather than in the eye while talking.

4.2 ECCR and FFCR in Center Camera Position

Table 3 shows the ECCR, the FFCR, Total talking time of subject, and Total talking time of partner, and Table 4 shows the detail of ECCR and FFCR.

By comparing Table 3 with Table 1, the following are found,

(1) Averaged ECCR decreases by locating a camera to the center. This trend can be seen except for subject 1.
(2) Averaged FFCR increases by locating a camera to the center. This trend can be seen true for all subjects.

Table 3. ECCR, FFCR, and talking time in center camera position

	ECCR	FFCR	$\sum_{m=1}^{M} Ts(m)$	$\sum_{n=1}^{N} Tp(n)$
Subject 1	37.7%	83.9%	105.3 s	157.8 s
Subject 2	27.1%	93.2%	112.7 s	140.9 s
Subject 3	14.7%	97.2%	141.2 s	138 s
Subject 4	21.6%	93.8%	97.73 s	181.4 s
Subject 5	25.9%	71.9%	155.7 s	117.8 s
Average	25.4%	88.0%	122.5 s	147.2 s

Table 4. ECCR and FFCR in talking and listening in center camera position

	$ECCR_T$	$ECCR_L$	$FFCR_T$	$FFCR_L$
Subject 1	45.3%	32.6%	87.0%	81.9%
Subject 2	33.3%	22.1%	91.5%	94.7%
Subject 3	9.2%	20.3%	96.9%	97.5%
Subject 4	14.0%	25.7%	88.5%	96.7%
Subject 5	24.6%	27.5%	67.0%	78.5%
Average	25.3%	25.6%	86.2%	89.9%

By comparing Table 4 with Table 2, the following are found,

(3) Averaged $ECCR_T$ slightly increases by centering a camera. But, this is not a remarkable trend.
(4) Averaged $ECCR_L$ slightly decreases by centering a camera. This trend can be seen except for subject 5.
(5) Both $FFCR_T$ and $FFCR_L$ increase by centering a camera. This trend can be seen true for all subjects.

5 Conclusion

In order to improve the naturality of a conversation using videophone or virtual conversational system, we have proposed two objective measures, ECCR and FFCR, and developed a videophone system with half mirror. By changing the position of camera

from above to center, FFCR increases from 74.7% to 88.0%. This means that the face to face conversation is affected by the gaze of the partner. Obviously, the conversation with mutual gaze increased by centering a camera, and the naturality of a conversation have been improved. However, because of wide eye area and no consideration of partner's gaze, ECCR in talking/listening does not change. For clarifying the true eye contact conversation ratio, the eye area and gaze of partner should be considered in future work. In addition, measures of affect or measure of emotion such as PANAS (Positive and Negative Affect Schedule) should be studied.

References

1. De Silva, L.C., et al.: A teleconferencing system capable of multiple person eye contact using half mirrors and cameras placed at common points of extended lines of gaze. IEEE Trans. Circ. Syst. Video Technol. **5**(4), 268–277 (1995)
2. Solina, F., Ravnik, R.: Fixing missing eye-contact in video conferencing system. In: Proceedings of ITI 2011, pp. 233–236 (2011)
3. Lu, J., Tao, X., Dong, L., Ge, N.: Chunk-wise face model based gaze correction in conversational videos with single camera. In: Proceedings of CITS 2016, pp. 1–5 (2016)
4. Ebara, Y., Nabuchi, T., Sakamoto, N., Koyamada, K.: Study on eye-to-eye contact by multi-viewpoint videos merging system for tele-immersive environment, In: Proceedings of AINA 2006, vol.2 (2006)
5. Ando, M., et al.: An analysis using eye-mark recorder of the effectiveness of presentation methods for e-learning. In: Proceedings of ICALT 2017, pp. 183–185 (2007)
6. http://www.eyemark.de/downloads/EMR9_Basic_Operations.pdf
7. Broz, F., et al.: Mutual gaze, personality, and familiarity: dual eye-tracking during conversation. In: Proceedings of IEEE RO-MAN 2012, pp. 858–864 (2012)