Tablet System for Sensing and Visualizing Statistical Profiles of Multi-Party Conversation

Hiroyuki Adachi
Ritsumeikan University
Email: adachi@i.ci.ritsumei.ac.jp

Seiko Myojin
Ritsumeikan University
Email: seiko@i.ci.ritsumei.ac.jp

Nobutaka Shimada
Ritsumeikan University
Email: shimada@ci.ritsumei.ac.jp

***Abstract*—In this paper, we present a tablet system that measures and visualizes who speaks to whom, who looks to whom, and their cumulative time in face-to-face multi-party conversation. The system measures where each participant is and when he/she speaks by using the front and back cameras and microphone of tablets. The evaluation result suggests that the system can measure such information with good accuracy. Our study aims to support the motivation of participants and enhance communication.**

I. Introduction

In multi-party conversation, someone speaks more often or less than others. We cannot obtain enough information from those who speaks less than others. In order to enhance the conversation, various ways for supporting communication have been researched. TableTalkPlus [1] is a system that visualizes the dynamics of communication like a change of atmosphere generated through the participants' relationship on a projector. It motivates participant to talk, changes the direction of conversation, and designs the field of conversation. Terken *et al.* suggest a system that provides visual feedback about speaking time and gaze behavior in small group meetings [2]. On the system, each participant is wearing headbands with two pieces of reflective tape to detect gaze behavior and a microphone to detect speaking time. They showed that the feedback influenced the amount of speaking of the participants. The systems that display feedbacks for participants on a projector like these are popular [3], [4], [5]. There is another visual feedback system during mealtime communication [6]. It does not direct a user to a specific action, but affects conversation implicitly by visualizing user's behavioral tendency. On the other hand, Schiavo *et al.* present a system that consists of four Microsoft Kinect sensors and four tablets. It acts as an automatic facilitator by supporting the flow of communication in conversation [7].

The systems described above are difficult to set up, because these systems need special things like having or wearing a microphone [2], [4], [5], a room equipped with a projector [1], [2], [3], [4], [5], and so on [6], [7]. In our system, tablets process both sensing and visualizing who speaks/looks to whom by using the front and back cameras of the tablets. Therefore, the system requires only the tablets, and has the advantage of easy to use.

II. Overview of our System

Our tablet system can obtain the information about *who*, *when*, and *where* individual utterance occurs, and also obtain a history of this information in multi-party conversation. *The statistical profiles of utterance* is measured from such information. In addition, the system obtains *the statistical profiles of conversation* from assembling the statistical profiles of utterance of each participant. It means the information between two people that who speaks to whom, who looks to whom, and their history.

The information about 1) *who* spoke is obtained from the ID of the tablet each participant has. The information about 2) *when* participant spoke is obtained by picking up the voice from the microphone of the tablet. The information about 3) *where* means that a place of the participant and the participant's face direction. Moreover, the information about *who spoke to whom* is estimated from the above information.

Fig. 1 shows the system structure as an example on three-person-conversation. Each participant has a tablet and talks around a table which has a marker on it. The tablet has front and back cameras; the front camera takes a picture of the participant's face and the back camera takes a picture of the marker. The tablet also has a microphone and picks up the participant's voice. Each tablet is connected to a server via wireless network, and sends information about the statistical profiles of utterance. The server integrates this information as the statistical profiles of conversation, and then sends back to each tablet. Each participant talks with others face-to-face with glancing the visualized information of the tablet.

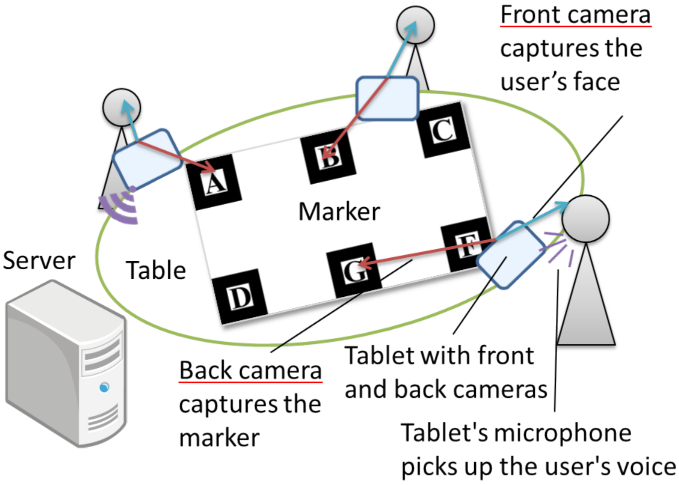


Fig. 1. System structure.

III. Methods

A. Measurements of individual utterance

In this section, we describe the methods for measurements of information about *who*, *when*, and *where* of individual utterance.

1) *Who*: Each tablet is connected to the server and has a connection ID. A current speaker is specified by this ID.
2) *When*: The voice of each participant is picked up by the microphone of his/her tablet. When the voice signal level of utterance exceeds the pre-determined threshold, the system recognizes that the participant is speaking.
3) *Where*: A participant's position and the face direction, in the world coordinate determined by markers on the table, are calculated from the geometric relation of user-tablet-marker and captured images by the front and back cameras of the tablet (Fig. 2). Tomioka *et al.* [8] proposed a pseudo see-through tablet by employing the front and back cameras in the similar framework.
The homogeneous transformation matrix ${}^{m}\mathrm{T}_{f}$ represents the above information. It is obtained by multiplying the following three matrices as;

$$ {}^{m}\mathrm{T}_{f} = ({}^{bc}\mathrm{T}_{m})^{-1}\, {}^{bc}\mathrm{T}_{fc}\, {}^{fc}\mathrm{T}_{f}. \qquad (1) $$

First, ${}^{bc}\mathrm{T}_{m}$ is the transformation matrix from the back camera to the marker and is measured from the back camera image by using ARToolKit [9]. Fig. 3 shows the result of the marker detection and the axes of the world coordinate as an example. Second, ${}^{bc}\mathrm{T}_{fc}$ is the transformation matrix from the back camera to the front camera. It can be calibrated in advance because the relative position of the two cameras on the particular tablet is fixed. Last, ${}^{fc}\mathrm{T}_{f}$ is the transformation matrix from the front camera to the participant's face. The matrix is composed a face rotation matrix and translations of the face. These elements are detected from a front camera image by using OKAO® Vision which is OMRON's face sensing technology [10]. Fig. 4 shows an example image of the face detection and the face rotation detection.

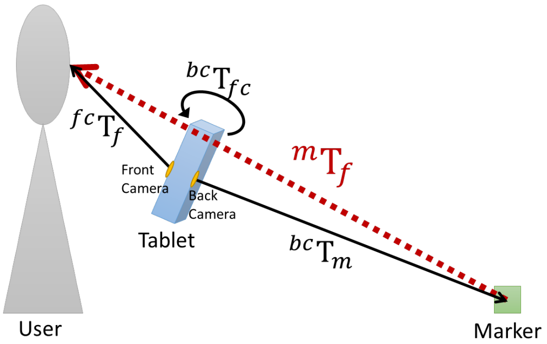


Fig. 2. Geometric relation of user-tablet-marker.

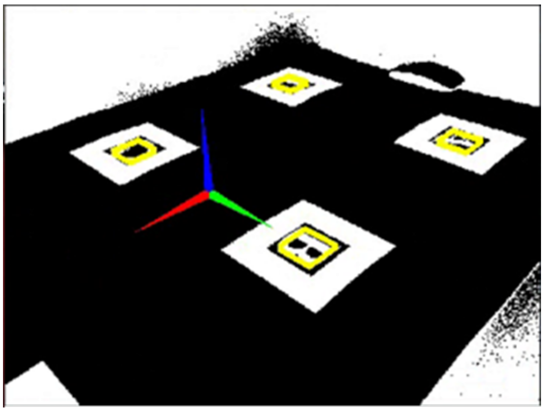


Fig. 3. Marker detection through the back camera.

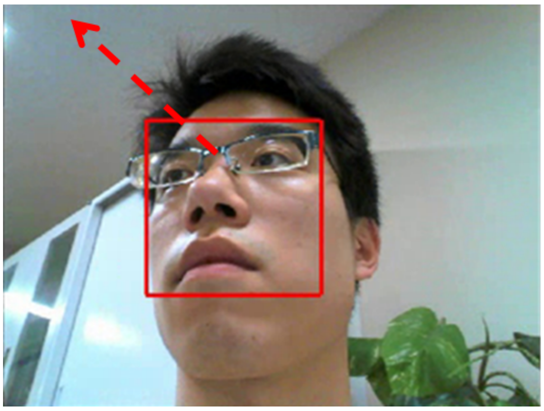


Fig. 4. Face detection through the front camera.

B. Measurements of statistical profiles of conversation

In the previous sections we described the way how the tablet system specifies who exists where (tablet ID and position estimated by marker), where faces to (facial direction), and when he/she speaks (auditory sensing). By assembling these observations obtained from individual participant's tablet, the conversational partner (information about who speaks/looks to whom) is estimated. Fig. 5 shows the positions and face directions of participants. The conversational partner of each participant is estimated through the following steps.

1) Calculate a vector U_i as a face direction of the participant i.
2) Calculate a vector V_{ij} which directs from the participant i to the participant j.
3) Calculate a similarity of U_i and V_{ij} as;

$$ \mathrm{Sim}_{ij} = \begin{cases} \dfrac{U_i \cdot V_{ij}}{\|U_i\|\|V_{ij}\|}, & \text{if} -\dfrac{4}{\pi} < \theta < \dfrac{4}{\pi} \\ 0, & \text{otherwise} \end{cases} \qquad (2) $$

4) Select the participant j with maximum Sim_{ij} as a conversational partner of the participant i. If $\mathrm{Sim}_{ij} = 0$ the participant i has no conversational partner.

In addition, storing of this data, we calculate the cumulative time of who looks/speaks to whom as the statistical profiles of conversation.

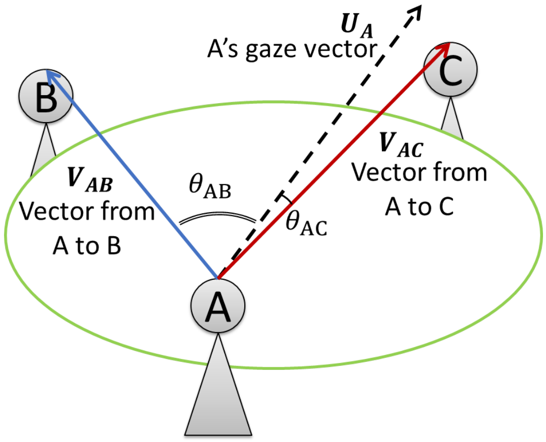


Fig. 5. Conversational partner estimation.

C. *Visualization of statistical profiles of conversation*

Fig. 6 shows an example of a situation of multi-party conversation using our system. Fig. 7 shows a visualization example of the statistical profiles of conversation on a tablet's screen. It represents a situation of conversation where user A talks to user B viewed from user A's eye. This example is on table-centric view. There is another way of visualization like a user-centric view.

The positions of participants and facial directions are represented as circles and dotted arrows respectively. The pink bar besides a face circle represents the amount of conversation from user A to user B and the orange bar represents the cumulative time that user A was looking at user B.

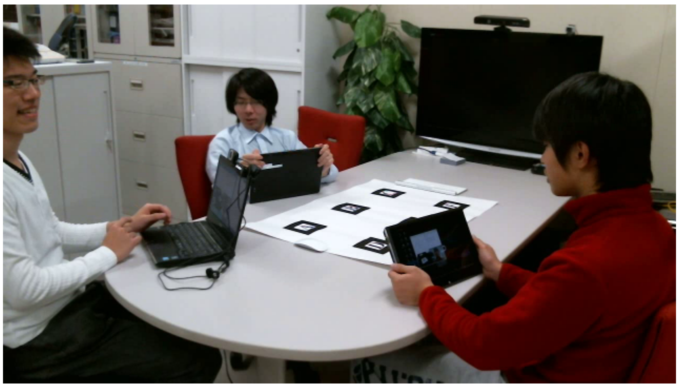


Fig. 6. Situation example of multii-party conversation using our system.

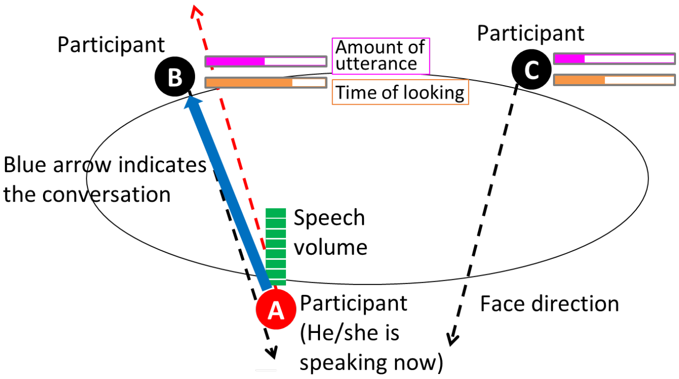


Fig. 7. Visualization example of the statistical profiles on the screen.

Therefore, the participants obtain their conversation amounts, and we consider this visualization may provide motivation for them, for example, to speak to someone who has never talked with them so much.

IV. Accuracy Evaluation

We evaluated the accuracy of the measurement of facial direction through the experimental situation by using the implemented system. Arrange targets in a quarter of a circle in increments of 15 degrees, a user (an author) turns to look at each target at 30 seconds. Fig. 8 shows the result of the measurement of an error of the target direction and the user's face direction. As a result of this evaluation, the measurement of face direction in horizontal has a margin of error of 2 degrees one way or the other.

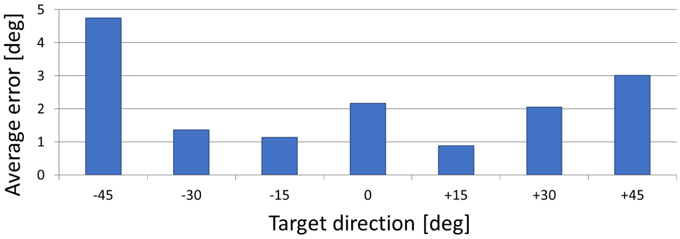


Fig. 8. Average error of the measurement of face direction.

V. Experiment and Performance Evaluation

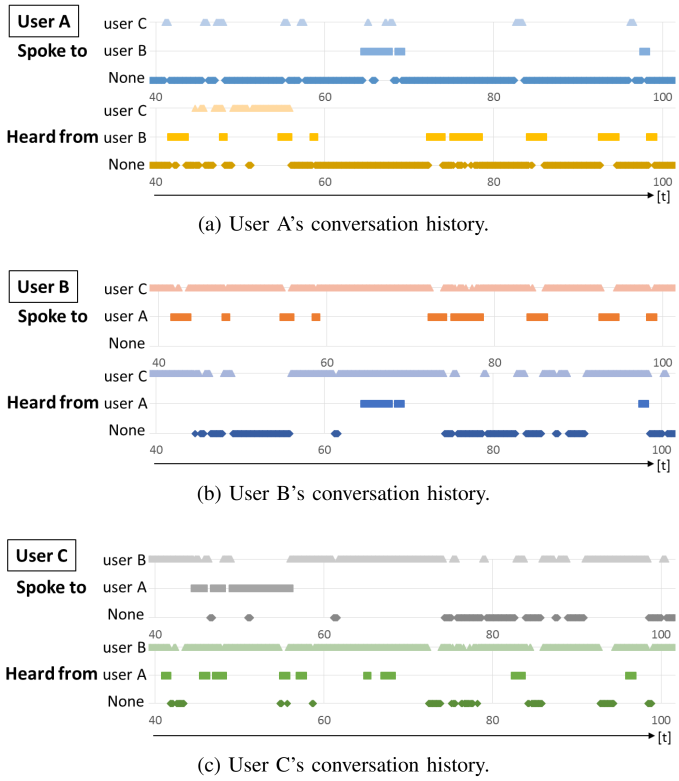
The system was evaluated in the two minutes of three-person-conversation for confirmation of how well sensing and visualizing conversation (Fig. 6). One of the participant is an author and the others are students. In this evaluation, we use the two tablets (Sony VAIO Duo 11) and a laptop with two webcams (Logicool HD Webcam C615) as substitute for a tablet.

Table I shows the parcentage of cumulative time for watching and speaking to another in the conversation time. In this conversation, participant A speaks about 17 seconds (13% + 7.1% of 2 minutes) and looks to participant B and participant C very little, participant B speaks about 1 minute and 47 seconds, and participant C speaks 1 minute and 12 seconds.

TABLE I. Parcentage of cumulative watching time/cumulative speaking time in the conversation time.

whom who	A	B	C
A	/	28% / 6.0%	57% / 8.1%
B	1.4% / 13%	/	13% / 76%
C	1.0% / 7.1%	42% / 53%	/

Fig. 9 shows a part of conversation histories, who spoke to whom and who heard from whom. Fig. 9(a) shows user A's conversation history, Fig. 9(b) shows user B's conversation history, and Fig. 9(c) shows user C's conversation history. We describe Fig. 9(a) as an example of the conversation history.



(a) User A's conversation history.

(b) User B's conversation history.

(c) User C's conversation history.

Fig. 9. Conversation histories.

The first line is the timelime of speaking to user C, and the second line is the timeline of speaking to user B. The third line is the timeline when user A did not speak to anyone. Besides the speaking timelines, there are the listening timelines. The forth line is the timelines of listening to user C (user C is speaking to user A), and the fifth line is the timeline of listening to user B too. The last line is the timeline when no one spoke to user A. In these timelines, the symbols such as triangles, squares, and diamond shapes represent timing when user A spoke to someone or when user A heard from someone. Fig. 9(b) and 9(c) are similar format.

Fig. 10 shows the users' conversation histories after a lapse of 50 seconds from starting the covnersation. These represent that user A was not speaking, user B was speaking to user A, and user C was speaking to user B. This situation is visualized on the users' statistical profiles as shown in Fig. 11; user B's statistical profiles at the time as an example. Fig. 12 shows the each user's cumulative time for watching and speaking to another at the time, for example, user B had been speaking to user C about 30 seconds until then. Fig. 13 shows the users' conversation histories after a lapse of 100 seconds. These represent that user A was not watching and speaking to anyone, and user B and user C were speaking with each other. This situation is visualized on the users' statistical profiles as shown in Fig. 14 too. Fig. 15 shows the each user's cumulative time for watching and speaking to another at the time, user B had been speaking to user C about 70 seconds. User B's speaking time for user C has increased to 70 seconds from 30 seconds in 50 seconds. The system measures the progress of conversation and visualizes on the statistical profiles for each user.

These history records well explain the actual conversation

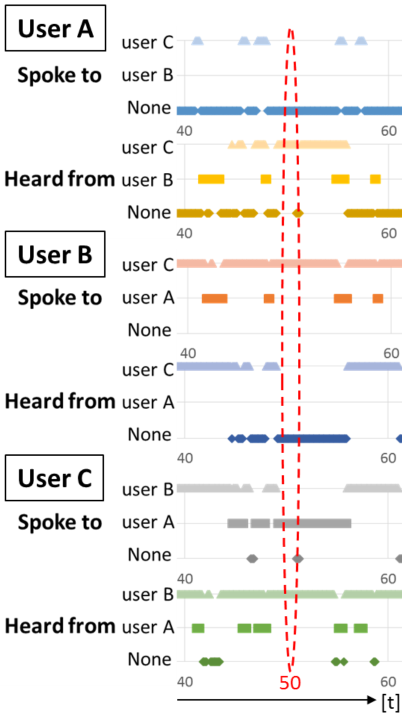


Fig. 10. Conversation histories after a lapse of 50 seconds.

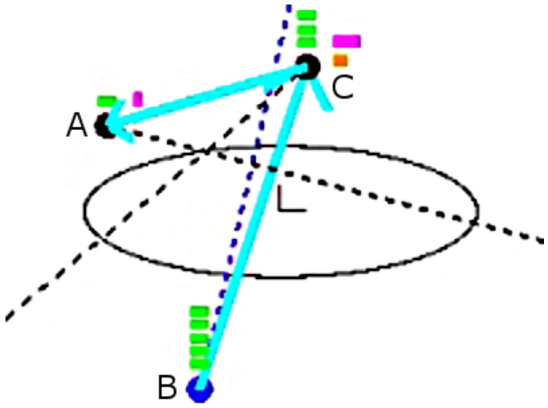


Fig. 11. Statistical profiles of user B after a lapse of 50 seconds.

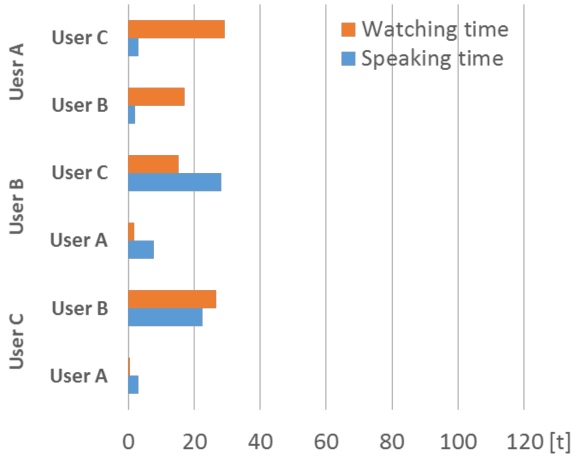


Fig. 12. Cumulative time for watching and speaking to another after a lapse of 50 seconds.

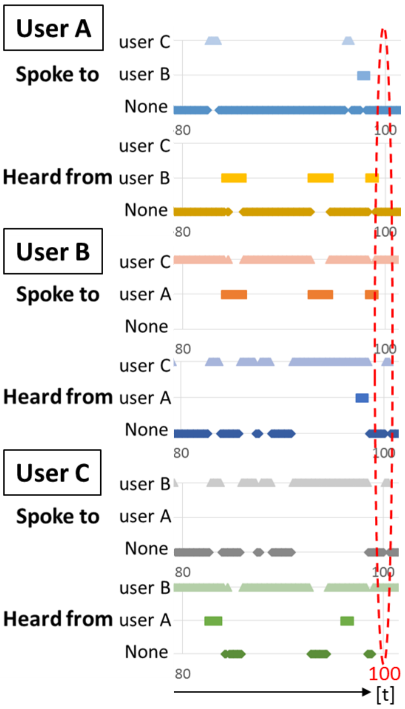


Fig. 13. Conversation histories after a lapse of 100 seconds.

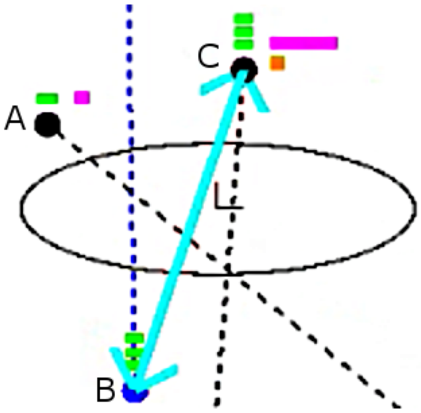


Fig. 14. Statistical profiles of user B after a lapse of 100 seconds.

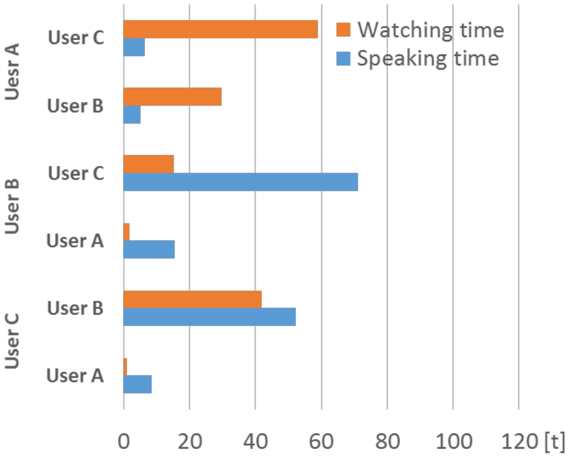


Fig. 15. Cumulative time for watching and speaking to another after a lapse of 100 seconds.

frequency for the conversations of the three persons, for example, the situation of user B spoke to user C often is represented by the light orange triangles on the top line of Fig. 9(b). However, we need to consider some noisy data like a tablet's fan noise, and another user's voice. Altough there may be some false recognition, the system measures the statistical profiles of conversation with good accuracy.

VI. Conclusion

In this research, we presented a tablet system that sensing and visualizing the statistical profiles of individual utterance and the statistical profiles of multi-party conversation like who speaks to whom and its cumulative time. Evaluation results showed that the system is able to measure individual utterance and visualize the statistical profiles at a reasonable level. However, there are some negative feedbacks, the system leaves a lot of room for improvement. Additionally, our goal is to enhance communication, a mechanism to perform it is necessary. Future work, we will be developing the marker less system to make it easier to use, and introducing a game element into the conversation like giving regards to a speaker and assessment system for users.

Acknowledgement

The authors would like to thank the OMRON Corporation let us use the OKAO Vision Library in their courtesy.

References

[1] N. Ohshima, K. Okazawa, H. Honda, and M. Okada, "Tabletalkplus: An artifact for promoting mutuality and social bonding among dialogue participants," *Human Interface Society*, vol. 11, no. 1, pp. 105–114, 2009, (in Japanese).

[2] J. Terken and J. Sturm, "Multimodal support for social dynamics in co-located meetings," *Personal and Ubiquitous Computing*, vol. 14, no. 8, pp. 703–714, 2010.

[3] T. Bergstrom and K. Karahalios, "Conversation clock: Visualizing audio patterns in co-located groups," in *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on.* IEEE, 2007, pp. 78–78.

[4] J. M. DiMicco, A. Pandolfo, and W. Bender, "Influencing group participation with a shared display," in *Proceedings of the 2004 ACM conference on Computer supported cooperative work.* ACM, 2004, pp. 614–623.

[5] K. Fujita, Y. Itoh, H. Ohsaki, N. Ono, K. Kagawa, K. Takashima, S. Tsugawa, K. Nakajima, Y. Hayashi, and F. Kishino, "Ambient suite: enhancing communication among multiple participants," in *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology.* ACM, 2011, p. 25.

[6] K. Ogawa, Y. Hori, T. Takeuchi, T. Narumi, T. Tanikawa, and M. Hirose, "Table talk enhancer: a tabletop system for enhancing and balancing mealtime conversations using utterance rates," in *Proceedings of the ACM multimedia 2012 workshop on Multimedia for cooking and eating activities.* ACM, 2012, pp. 25–30.

[7] G. Schiavo, A. Cappelletti, E. Mencarini, O. Stock, and M. Zancanaro, "Overt or subtle? supporting group conversations with automatically targeted directives," in *Proceedings of the 19th international conference on Intelligent User Interfaces.* ACM, 2014, pp. 225–234.

[8] M. Tomioka, S. Ikeda, and K. Sato, "Approximated user-perspective rendering in tablet-based augmented reality," in *Proceedings of IEEE International Symposium on Mixed and Augmented Reality (ISMAR2013).* IEEE, 2013, pp. 21–28.

[9] "ARToolKit Home Page," http://www.hitl.washington.edu/artoolkit/.

[10] OMRON Corporation, "OKAO Vision | OMRON Global," http://www.omron.com/r_d/coretech/vision/okao.html.