Audio-visual Technology for Conversation Scene Analysis

Kazuhiro Otsuka[†] *and Shoko Araki*

Abstract

This article overviews recent research progress toward understanding human communication scenes with multimodal information including audio and video in the Media Processing Laboratory of NTT Communication Science Laboratories. First, it briefly introduces our motivation and goal and the background of the research field. Second, it describes the latest advances in audio technology including voice activity detection and in computer vision technology including visual head pose tracking. Third, it presents our realtime multimodal system for analyzing human conversation scenes that combines audio and vision technologies. Finally, it discusses future work and potential application areas.

1. Introduction

Face-to-face conversation is one of the most basic forms of communication in daily life and group meetings are used for conveying/sharing information, understanding others' intentions/emotions, and making decisions. To develop information and communications technology that can support our communication in a face-to-face setting and/or remote meetings, it is important to understand how we communicate with each other and what kinds of behavior must be conveyed in messages for communications to be successful. To answer these questions, we have focused on nonverbal messages/behaviors that appear in face-to-face conversations because psychologists have suggested that they play important roles in human communications [1]. Nonverbal messages are expressed by nonverbal behaviors in multimodal channels such as eye gaze, facial expressions, head motion, hand gestures, body posture, and prosody. Therefore, it is expected that conversation scenes can be largely understood by observing people's nonverbal behaviors with sensing devices such as cameras and microphones.

From the abovementioned viewpoint, we have been working on audio and vision technologies and their integration in order to understand human-human conversation scenes. We call the task here conversation scene analysis; its goal is to provide automatic description of conversation scenes in terms of 6W1H, i.e., Who, When, Where, Whom, What, Why, and How. By combining some 6W1H questions, we can define a number of problems from low-level (close to physical behavior) to high-level (contextual and social level) ones. For example, *who is speaking* is the most essential question and is called the speaker detection problem. The combination of *Who* and *When* yields problems called speaker diarization, i.e., *who* is speaking *when*. So far, speaker diarization has been a central problem in the field of audio-based communication analysis. Furthermore, the question of *who is talking to whom and when* requires multimodal information because the direction of addressing cannot be detected from only the audio signal, but is indicated by visual behaviors such as gaze, face direction, and body pose. Moreover, *What* involves the content of the conversation, which is mainly carried by the verbal modality. *How* questions are related to the emotions and attitudes of people, and *Why* questions require total understanding of the context of the conversations.

As the initial step to conversation scene analysis, our research groups have focused on lower-level problems such as *who is speaking when, who is talk-*

[†] NTT Communication Science Laboratories
Atsugi-shi, 243-0198 Japan

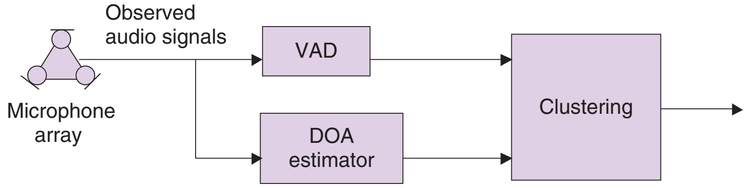

Fig. 1. System flow of the audio part, which achieves speaker diarization.

ing to whom, and *who responds to whom, when, and how*. This article introduces some of our recent research progress. Section 2 overviews the component technologies of audio, visual, and multimodal processing. Section 3 describes a demonstration system of realtime conversation scene analysis based on multimodal integration. Section 4 discuss future works and conclude this article.

2. Audio-visual component technology

To capture nonverbal behaviors appearing in face-to-face meeting situations, our groups have been developing component technologies for analyzing audio, visual, and multimodal data. Audio technology is mainly being developed in the signal processing group in the Media Information Laboratory in NTT Communication Science Labs. Vision and multimodal technologies are mainly being developed in the Media Recognition Group.

2.1 Audio technology

Audio-related nonverbal behaviors include the presence/absence of utterances, prosody, stressing, and rate of speech. Among them we have achieved automatic detection of the presence/absence of utterances in a meeting situation. Currently, we aim to estimate who is speaking when in a meeting. As mentioned in the introduction, the estimation of who is speaking when is an essential technique for conversation scene analysis, and it is termed speaker diarization.

The processing in the audio part of our system is illustrated in **Fig. 1**. We utilize audio signals captured by a microphone array to determine who is speaking at each time step. Our diarization technique uses the methods proposed in [2]; the key parts are a noise-robust voice activity detector (VAD), a direction of arrival (DOA) estimator, and a DOA clustering part. That is, our diarization system relies on the speaker positions to estimate who is speaking when. The advantage of our system is that we do not require any prior knowledge about the number of people and noise sources.

2.1.1 Voice activity detection

First, our VAD technique discriminates whether or not the current audio observations include speech signals. This step is very important, especially for a noisy meeting. For a speech/non-speech discriminator, we have developed a VAD method called "Multi Stream Combination of Likelihood Evolution of VAD" (MUSCLE-VAD) [3]. It integrates two speech features to improve the robustness over various noises. A block diagram of MUSCLE-VAD is shown in **Fig. 2**. It is constructed using two stream speech/non-speech discriminators, i.e., periodic to aperiodic component ratio-based detection (PARADE) [4], and an approach based on the switching Kalman filter (SKF) [5]. PARADE is robust against burst noise and SKF is robust against stationary and non-stationary noises. The combination of these two methods makes MUSCLE-VAD robust against a wide variety of noises that could conceivably occur in a meeting.

2.1.2 Direction of arrival estimation

To estimate the DOA for each speaker in a meeting, the GCC-PHAT technique [6] has been widely used. However, this technique sets the constraint of just one DOA in each timeframe, and it often fails to detect speakers correctly in the case of overlapping speech. To avoid this problem, we use time-frequency domain DOA (TFDOA) estimation, which was recently proposed in [2]. TFDOA estimates the DOA for each time-frequency slot, instead of for each timeframe. As a result of the sparseness attribute of speech signals in the time-frequency domain, we can estimate multiple DOAs even if there are several speakers' utterances in a timeframe.

2.1.3 Direction of arrival clustering

Finally, we cluster the TFDOA information for speech-containing timeframes, which are estimated by DOA, to determine who is speaking when. To enable clustering even when the number of people in a meeting is unknown, we use an online clustering algorithm known as leader-follower clustering [7].

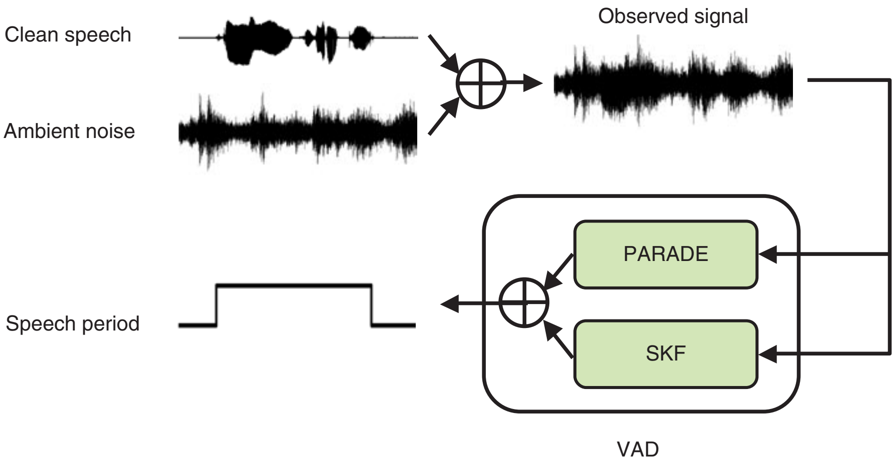

Fig. 2. Block diagram of VAD.

Each cluster corresponds to one sound source, which should be a potential speaker. Probabilistic integration of the VAD result and TFDOA estimation result was also proposed in [2].

2.2 Vision technology

Vision-related nonverbal cues include the positions of people, face directions, head gestures, facial expressions, hand gestures, and postures. Among these, we have been working on automatic face pose tracking, head gesture recognition, and facial expression recognition. Our face pose tracker and its extension to facial expression recognition are introduced below.

2.2.1 Face pose tracking

The importance of measuring face pose arises from the fact that it is a reasonable indicator of people's gaze and direction of visual attention. Among the possible nonverbal messages/behaviors, eye gaze is especially important because it has various roles such as monitoring others, expressing one's attitude/interest, and regulating conversation flow [8]. However, gaze direction during natural conversation is difficult to measure directly. Therefore, face direction is often used as a reasonable alternative. However, it is more than just an alternative; by itself it is a useful indicator of people's attention to others during meetings. In addition, the temporal changes in face direction indicate head gestures such as nodding. Therefore, face direction is an important cue in analyzing meetings.

To estimate the position and pose of people in face-to-face meetings, we have developed a face pose tracker called STCTracker (sparse template condensation tracker) [9]. This tracker was originally proposed by Matsubara and Shakunaga [10] and we first applied it to conversation scene analysis [11]. Recently, we enhanced it into a more robust, accurate, and faster tracker for following multiple faces [9]. Furthermore, we verified its performance in conversation scene analysis [12]. As reported in [12], the advantages of STCTracker are its robustness against large head rotation, up to ±60º in the horizontal direction, and its speed: it can track multiple faces simultaneously in real time by utilizing a modern graphics processing unit. Furthermore, it can automatically build three-dimensional (3-D) face templates after initialization.

A diagram of STCTracker is shown in **Fig. 3**. It consists of particle filtering and initialization. The particle filter used in STCTracker combines template matching with particle filtering. In contrast to traditional template matching, which assesses all pixels in a rectangular region, sparse template matching focuses on a sparse set of feature points within a template region, as shown in **Fig. 4**. The state of a template, which represents the position and pose of the face, is defined as a seven-dimensional vector consisting of 2-DOF (degrees of freedom) translation on the image plane, 3-DOF rotation, a scale (we assume weak-perspective projection), and an illumination coefficient. The particle filter is used to sequentially estimate the posterior density of the template state, which is represented as a particle set. The weight of each particle is calculated based on the matching error between input images and the template, whose state is assigned by that particle—a higher weight is given to a particle

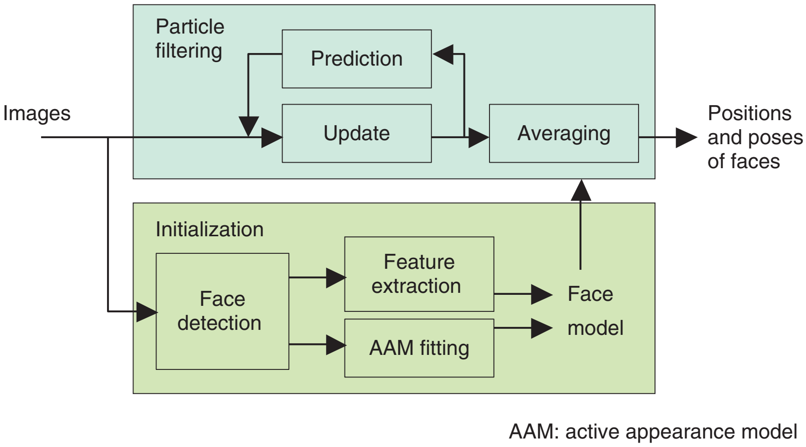

Fig. 3. Diagram of STCTracker.

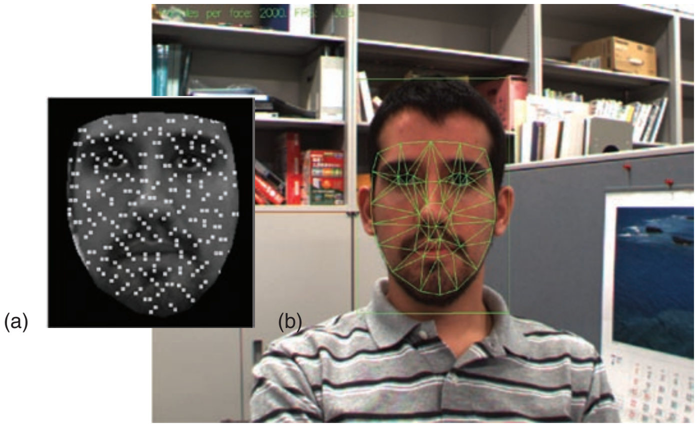

Fig. 4. Face model (sparse template) used in STCTracker. (a) Feature points and (b) face shape obtained by AAM fitting.

with a smaller matching error. The particles with higher weights indicate the plausible position and pose of the target face. STCTracker has reasonable speed owing to the sparseness of the feature points and robustness owing to robust template matching combined with multiple-hypothesis generation/testing by the particle filter framework.

The initialization part creates a template for each person before tracking starts. First, the frontal face is detected. Next, feature points are detected over the detected face region, and active appearance model (AAM) fitting is conducted to create the shape model of the detected face, as shown in Fig. 4. Finally, a face model (i.e., face template) is built upon the extracted feature points with shape information.

2.2.2 Face expression recognition

We have been working on facial expression recognition based on STCTracker in cooperation with the University of Tokyo [13], [14]. The key idea is an extension of the sparse template, called a variable intensity template (VIT), which can model the variation in image intensity depending on the facial expression by using the mixture of image intensity distributions depending on each facial expression category. For now, the target expressions are neutral, happy, angry, sad, and surprised. The main advantage of the VIT-based method is in its robustness against head pose variations. In a meeting, people often turn their heads to pay attention to others. Such motion yields significant changes in head pose. Therefore, our method is suitable for such a situation.

2.3 Multimodal technology

So far, we have proposed a probabilistic framework

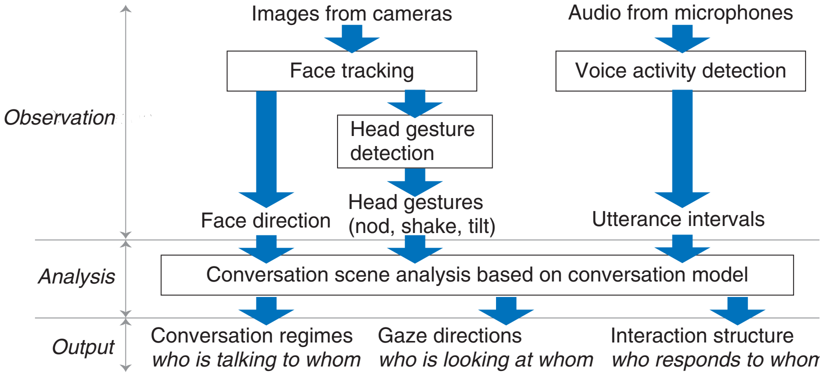

Fig. 5. Flow of multimodal meeting analysis.

for estimating the structure of conversations from pre-recorded multimodal data [15], as shown in **Fig. 5**. The goal is to automatically discover *who is responding to whom*, *when*, *and how*, as well as *who is talking to whom and who is listening to whom.*

First, the facial image of each participant is captured with a different camera, and the voice of each participant is recorded with a separate lapel microphone. From the video, the head pose of each person is estimated with an offline version of STCTracker. Head gestures are also detected using the temporal sequence of head pose. From the audio data, the utterance interval of each person is determined by a VAD technique that was developed especially for lapel microphones. Each participant is equipped with a label microphone so that his/her voice can be recorded separately from the others' voices. Using these behavior measurements, we have developed a probabilistic conversation model based on a dynamic Bayesian network that can represent the relationship among measured behaviors (head pose, gestures, and utterances), gaze directions, interaction structure, and conversation structures. Here, the gaze direction indicates *who is looking at whom or averting gazes from everyone*. The interaction structure indicates *who is responding* to whom. The conversation structure indicates typical patterns of message exchange among participants, including convergence (monologue), dyad-link (dialogue between two), and divergence (others). The conversation structures and gaze directions are estimated offline using a Markov chain Monte Carlo method.

From the estimated conversation structures and gaze directions, we have proposed a novel measure for automatically quantifying the amount of interpersonal influence present in face-to-face conversations [16]. The basic idea is that the amount of influence is defined on the basis of the amount of attention paid to speakers in monologues and to persons with whom the participants interact with during dialogues. Experiments confirm that this measure reveals some aspects of interpersonal influence in conversations.

3. Realtime multimodal system for conversation scene analysis

3.1 System overview

We have developed a realtime system for analyzing group meetings that uses a novel omnidirectional camera-microphone system [17]. The goal is to automatically discover the visual focus of attention, i.e. *who is looking at whom*, in addition to speaker diarization, i.e., *who is speaking and when*. These are essential to describe the structure of conversations, such as *who is talking to whom and who is listening to whom*. This system features a novel table-top sensing device, which consists of two cameras, each having a pair of fisheye lenses and a microphone array. It can capture omnidirectional images and audio at the same time. From the omnidirectional images captured with the cameras, the positions and poses of people's faces are estimated by STCTracker. Realtime tracking is achieved by utilizing graphics processing units. The face position/pose data is used to estimate the focus of attention in the group. Using the microphone array, robust speaker diarization is carried out by VAD and by DOA estimation. We have also developed new 3-D visualization schemes for analyzing the results. Using two personal computers (PCs), one for vision and one for audio processing, the system runs at 27.1 frames per second on average for a five-person meeting.

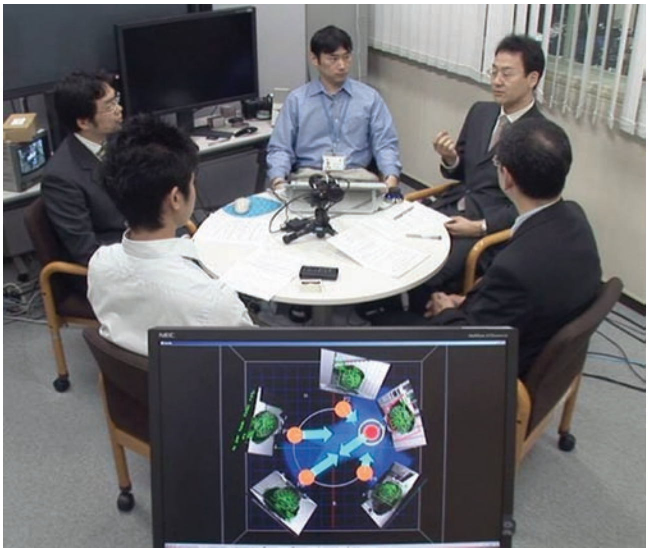

Fig. 6. Meeting scene. The display at the front shows the result of realtime processing.

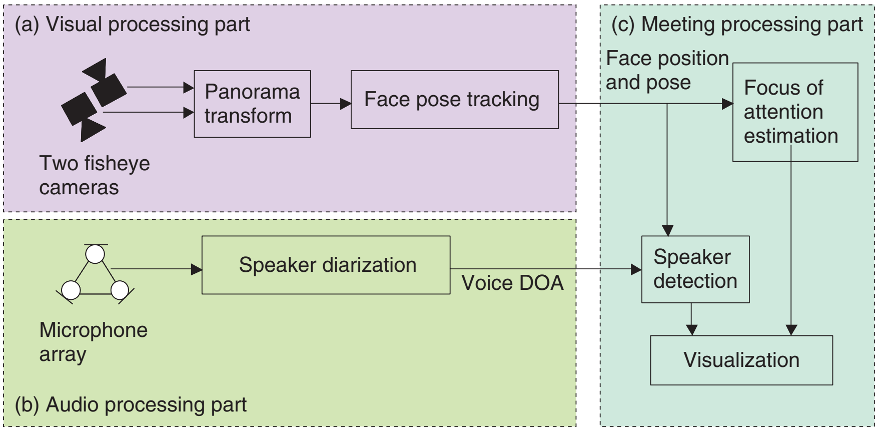

Fig. 7. System diagram.

To the best of our knowledge, this system is the first multimodal system to visually track not only face position, but also face pose in real time for group meeting analysis.

3.2 System configuration

This system targets meeting scenes, as shown in **Fig. 6**, and displays realtime output on a PC display, as described in section 3.1. As shown in **Fig. 7**, this system consists of three parts: (a) visual processing, (b) audio processing, and (c) meeting processing. The visual processing part consists of our new omnidirectional camera system (**Fig. 8**) and face tracking system using the image output by the camera system. For face tracking, we use STCTracker (described in 2.2.1). The audio processing part uses a microphone array to capture the voices of the participants. Robust speaker diarization estimation is performed. Diarization is achieved by our method described in section 2.1. Finally, the meeting processing part determines the utterance status (speaking or silent) of each meeting participant by cross-referencing the visual and audio information obtained in parts (a) and (b). Moreover, gaze direction (focus of attention) is estimated based on the positions and directions of faces. This information is displayed on a monitor using our new visualization schemes.

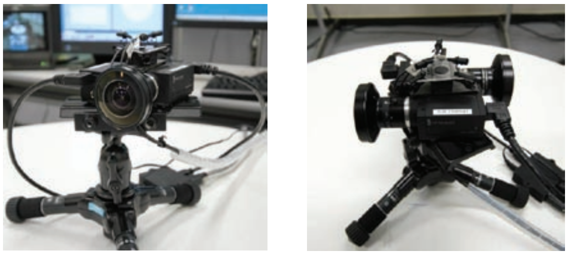

Fig. 8. Omnidirectional camera-microphone system.

3.2.1 Vision processing part

Our omnidirectional camera-microphone system is shown in Fig. 8. The camera part of the system consists of two cameras with fisheye lenses, which are facing in (180°) opposite directions. Since each fisheye lens covers a hemispherical region, the camera system could capture a nearly spherical region. However, our system captures only a horizontal strip, as shown in **Fig. 9**, so meeting participants are just covered by the image; this minimizes the transmission rate and allows high processing rates.

Using the images obtained with the omnidirectional cameras, the system estimates the face position and pose of each meeting participant in real time using STCTracker. Figure 9 shows an example of tracking results in an actual screenshot of the PC display during a meeting. The green meshes illustrate the face tracking results.

3.2.2 Audio processing part

Speaker diarization is done based on the audio signals observed by a microphone array, as shown in Fig. 8. The array consists of three tiny omnidirectional microphones placed at the vertices of a triangle with 4-cm sides and is located atop the camera unit. The VAD part is written in C and the TFDOA estimation and DOA clustering parts are implemented in MATLAB6.5. An example of diarization is shown in Fig. 9. The red dots along the axes indicate the DOA of the voice.

3.2.3 Meeting processing part

The meeting processing part uses the outputs of the visual processing and audio processing parts. Currently, our system implements utterance detection and focus of attention estimation. The presence/absence of utterances by each person is determined by combining the DOAs of speech from the audio processing part and the face positions from the vision processing part. This process is a data association problem; it aims to find the visual source responsible for utterances or noise. Here, we simply tackle this problem by the nearest neighbor rule with thresholding. The face position/pose from the face tracker is used to estimate the visual focus of attention in the group. More specifically, this article focuses on the discretized gaze direction of each person, i.e., looking at one person among all the people, or not looking at any one.

3.3 Visualization

To visualize the conversation scenes for meeting observers, who could be remote meeting participants in a teleconferencing or users of meeting archive systems, we implemented two visualization schemes with a manually configurable interface, as shown in

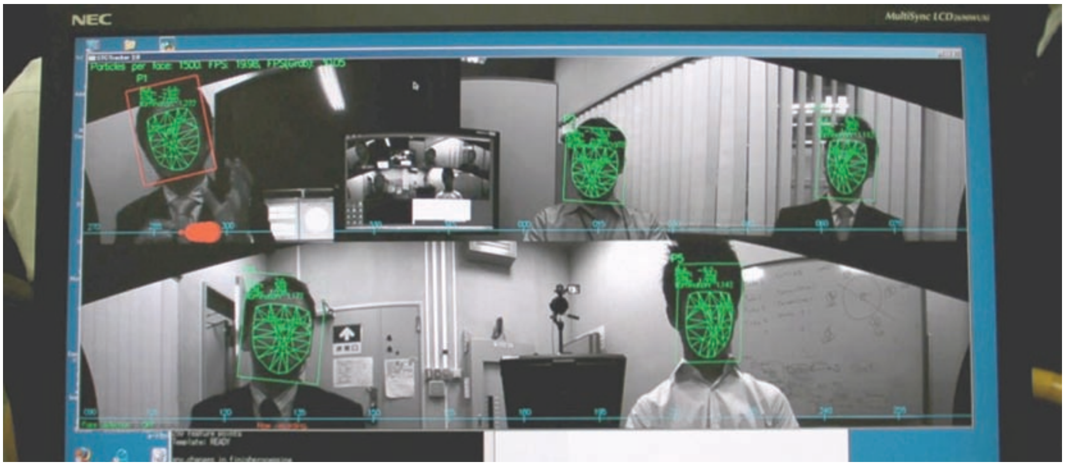

Fig. 9. Screenshot of system monitor displaying face tracking and VAD results.

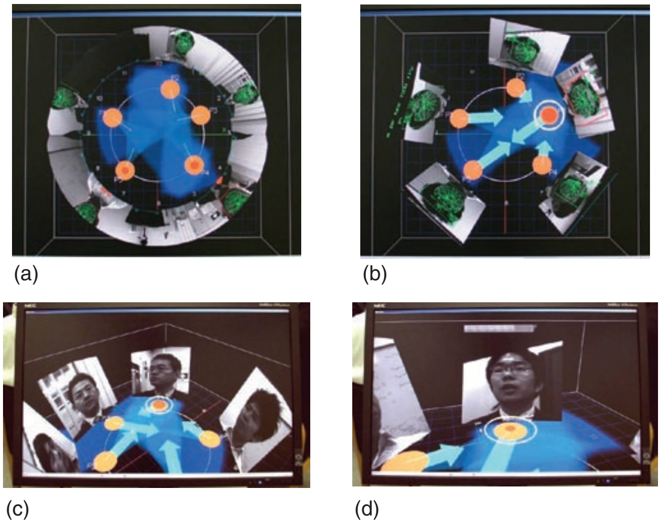
(a) (b) (c) (d)

Fig. 10. Visualization of conversation scenes. (a) Cylindrical visualization, (b) piecewise planar visualization, (c) viewpoint maneuver to middle range, and (d) viewpoint maneuver to close-up range.

Fig. 10. Cylindrical visualization of panoramic images and the relative position of each meeting participant (indicated by a circle) are shown in Fig. 10(a), which also shows the approximate field of view as (blue) translucent triangles; overlapping fields of views indicate where people pay attention to each other. Moreover, the voice activity of each participant is displayed by the red dot in each person's circle. An example of the output of the second visualization scheme, called piecewise planar representation, is shown in Fig. 10(b); the face image of each person is mapped to a planar surface, which is arranged to indicate the relative positions of the participants. This visualization provides the viewers with larger face images, which enables better understanding of the individual's expressions, while still clearly indicating their interpersonal positioning and interactive behaviors. In addition to the field of view and voice activity included in Fig. 10(a), Fig. 10(b) shows the discretized gaze direction of each person by an arrow and the focus of attention, people who are attracting the gaze of more than a person, by one or more circles.

For both visualization schemes, our system offers a maneuverable interface controlled by a 3-D mouse. With this device, users can freely and intuitively manipulate their viewpoints, as shown in Figs. 10(c). The rotation operation can choose the person and the zooming operation can control the focus (performed by pushing/pulling the knob); from one-person (Fig. 10(d)) to everyone (Fig. 10(b)). As a result of the high-resolution imaging provided by our new system, zoom-up face images retain sufficient details.

4. Conclusion

This article overviewed audio-visual technologies for understanding human communication scenes developed in the Media Information Laboratory of NTT Communication Science Laboratories. In addition to the individual component technologies, we have been focusing on multimodal integration of audio and visual technologies for fully capturing human nonverbal behaviors, which have been used as cues for human-human communications. As the initial step, this article introduced our realtime system for meeting analysis. Although this system is currently at a primitive level, it has the potential to open up a new field, realtime multimodal meeting analysis, which is an important key to a wide range of applications, such as teleconferencing, multimodal meeting archiving and browsing, automatic creation of meeting minutes, and social robot/agent systems.

Future work includes the following. First, we need to increase the accuracy and robustness of each technique including face pose tracking and voice activity detection. Second, it is important to extend the range of recognizable nonverbal behaviors. Possible targets include spontaneous facial expression, direct measurement gaze directions from video, postures, gestures, and prosody. Third, using measured low-level human behavior, it is important to move forward to discover the high-level state of meetings such as the roles of participants, conversation structures, dialogue acts, and social relationships among people. The low-level behavior data and high-level states of the meeting and its participants are useful for studying human science including cognitive neuroscience, psychologies, and sociologies. Finally, developing applications such as telecommunication systems and social agents/robots is an important research area. We believe that such applications have the potential to overcome communication barriers caused by time and space and to change our life and business styles.

References

[1] M. Argyle, Bodily Communication—2nd ed. Routledge, London and New York, 1988.

[2] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, "A DOA based speaker diarization system for real meetings," HSCMA2008, pp. 29–32, May 2008.

[3] M. Fujimoto, K. Ishizuka, and T. Nakatani, "A voice activity detection based on adaptive integration of multiple speech feature and signal decision scheme," Proc. of ICASSP2008, pp. 4441–4444, Mar. 2008.

[4] K. Ishizuka, T. Nakatani, M. Fujimoto, and N. Miyazaki, "Noise robust front-end processing with voice activity detection based on periodic to aperiodic component ratio," Proc. of Interspeech2007, pp. 230–233, Aug. 2007.

[5] M. Fujimoto and K. Ishizuka, "Noise robust voice activity detection based on switching Kalman filter," Proc. of Interspeech2007, pp. 2933–2936, Aug. 2007.

[6] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. Acoust. Speech and Signal Processing, Vol. 24, No. 4, pp. 320–327, 1976.

[7] R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern Classification," Wiley Interscience, 2nd ed., 2000.

[8] A. Kendon, "Some functions of gaze-direction in social interaction," Acta Psychologica, Vol. 26, No. 1, pp. 22–63, 1967.

[9] O. M. Lozano and K. Otsuka, "Real-time visual tracker by stream processing," Journal of Signal Processing Systems, 2008, DOI 10.1007/s11265-008-0250-2.

[10] Y. Matsubara and T. Shakunaga, "Sparse template matching and its application to real-time object tracking," IPSJ Trans. Computer Vision and Image Media 46(SIG9) pp. 60–71, 2005 (in Japanese).

[11] K. Otsuka, J. Yamato, Y. Takemae, and H. Murase, "Conversation Scene Analysis with Dynamic Bayesian Network Based on Visual Head Tracking," Proc. of IEEE ICME'06, Toronto, Canada, July 2006.

[12] K. Otsuka and J. Yamato, "Fast and Robust Face Tracking for Analyzing Multiparty Face-to-Face Meetings," Proc. of 5th Joint Workshop on Machine Learning and Multimodal Interaction (MLMI2008), Lecture Notes in Computer Science, Vol. 5237, pp. 14–25, Utrecht, Netherlands, Sept. 2008.

[13] S. Kumano, K. Otsuka, J. Yamato, E. Maeda, and Y. Sato, "Pose-Invariant Facial Expression Recognition Using Variable-Intensity Templates," Proc. of 8th Asian Conference on Computer Vision (ACCV2007), Part I, Lecture Notes in Computer Science Vol. 4843, pp. 324–334, Tokyo, Japan, Nov. 2007.

[14] S. Kumano, K. Otsuka, J. Yamato, E. Maeda, and Y. Sato, "Pose-Invariant Facial Expression Recognition Using Variable-Intensity Templates," International Journal of Computer Vision, Springer Netherlands, 2008, DOI 10.1007/s11263-008-0185-x.

[15] K. Otsuka, H. Sawada, and J. Yamato, "Automatic Inference of Cross-modal Nonverbal Interactions in Multiparty Conversations," Proc. of ACM 9th Int. Conf. Multimodal Interfaces (ICMI2007), pp. 255–262, Nagoya, Japan, Nov. 2007.

[16] K. Otsuka, J. Yamato, Y. Takemae, and H. Murase, "Quantifying Interpersonal Influence in Face-to-face Conversations based on Visual Attention Patterns," Proc. of ACM CHI Extended Abstract, pp. 1175–1180, Montreal, Canada, Apr. 2006.

[17] K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich, and J. Yamato, "A Realtime Multimodal System for Analyzing Group Meetings by Combining Face Pose Tracking and Speaker Diarization," Proc. of ACM 10th Int. Conf. Multimodal Interfaces (ICMI2008), pp. 257–264, Chania, Greece, Oct. 2008.

Kazuhiro Otsuka

Senior Research Scientist, Media Information Laboratory, NTT Communication Science Laboratories.

He received the B.E. and M.E. degrees in electrical and computer engineering from Yokohama National University, Kanagawa, and the Ph.D. degree in information science from Nagoya University, Aichi, in 1993, 1995, and 2007, respectively. He joined NTT Human Interface Laboratories in 1995. His current research interests include computer vision and communication scene analysis. He received the Best Paper Award of the Information Processing Society of Japan (IPSJ) National Convention in 1998, the Young Researcher's Award of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan in 1998, the Best Paper Award of the IAPR International Conference on Image Analysis and Processing in 1999, and the Outstanding Paper Award of the ACM International Conference on Multimodal Interfaces in 2007. He is a member of IEEE, IEICE , and IPSJ.

Shoko Araki

Research Scientist, Media Information Laboratory, NTT Communication Science Laboratories.

She received the B.E. and M.E. degrees from the University of Tokyo, Tokyo, and the Ph.D. degree in information science from Hokkaido University, Hokkaido, in 1998, 2000, and 2007, respectively. Since joining NTT in 2000, she has been engaged in research on acoustic signal processing, array signal processing, blind source separation applied to speech signals, meeting diarization, and auditory scene analysis. She received the 19th Awaya Prize from the Acoustical Society of Japan (ASJ) in 2001, the Best Paper Award of IWAENC in 2003, the TELECOM System Technology Award from the Telecommunications Advancement Foundation in 2004, the Academic Encouraging Prize from IEICE in 2006, and the Itakura Prize Innovative Young Researcher Award from ASJ in 2008. She is a member of IEEE, IEICE, and ASJ.