Development of Acoustic Nonverbal Information Estimation System for Unconstrained Long-Term Monitoring of Daily Office Activity*

Hitomi YOKOYAMA†(a), Masano NAKAYAMA†, Hiroaki MURATA†, Nonmembers, and Kinya FUJITA(b), Member

SUMMARY  Aimed at long-term monitoring of daily office conversations without recording the conversational content, a system is presented for estimating acoustic nonverbal information such as utterance duration, utterance frequency, and turn-taking. The system combines a sound localization technique based on the sound energy distribution with 16 beam-forming microphone-array modules mounted in the ceiling for reducing the influence of multiple sound reflection. Furthermore, human detection using a wide field of view camera is integrated to the system for more robust speaker estimation. The system estimates the speaker for each utterance and calculates nonverbal information based on it. An evaluation analyzing data collected over ten 12-hour workdays in an office with three assigned workers showed that the system had 72% speech segmentation detection accuracy and 86% speaker identification accuracy when utterances were correctly detected. Even with false voice detection and incorrect speaker identification and even in cases where the participants frequently made noise or where seven participants had gathered together for a discussion, the order of the amount of calculated acoustic nonverbal information uttered by the participants coincided with that based on human-coded acoustic nonverbal information. Continuous analysis of communication dynamics such as dominance and conversation participation roles through nonverbal information will reveal the dynamics of a group. The main contribution of this study is to demonstrate the feasibility of unconstrained long-term monitoring of daily office activity through acoustic nonverbal information.

key words: acoustic nonverbal information system, speaker estimation, long-term monitoring, daily office conversations

1. Introduction

Members of an organization typically work collaboratively towards a common goal [1], while team members typically work together as a social unit to effectively perform an assigned task [2]. The relationships among team members are significant factors in task performance [3], [4]-better relationships enhance information and knowledge sharing, which in turn enhances operation efficiency and adaptation to environmental changes [5]. Moreover, effective interactions among team members result in the creation of new knowledge, which leads to higher team performance [6]. In the unstable and complex business environment of today, more dynamic interaction is needed [7]. It is thus incumbent upon managers to understand the dynamics of the groups they manage at all times so that they can give them appropriate feedback or reorganize a team when necessary to maximize its performance.

Communication, which is a process of information exchange among people [8], is indispensable for establishing and maintaining common ground in collaborative tasks [9] and to maintaining interpersonal relationships in the group. In other words, effective communication is a key to successful team performance. An attractive way of understanding and managing the dynamics of a team is to continuously and automatically perform quantitative analysis of communications among team members.

Since communication is an exchange of information by using various media, identifying the information sender and receiver(s) in a conversation provides a better understanding of the group. Goffman’s participation framework [10] enables the roles of conversation participants to be understood. He classified the participants as a “speaker,” who sends the information, a “ratified addressee,” who is the person to whom the speaker speaks, and “side-participants,” who also receive the information. A number of researchers have suggested that utterance duration is strongly and positively correlated with dominance and leadership [11]–[13]. This, in turn, is related to the concept of “convensional floor,” i.e., conversational turn-taking. Hayashi classified several types of conversational floor; for instance, a “single floor” is when one person dominates the conversation, and a “cooperative floor” is when several people participate equally in the conversation and thus “share the floor” [14].

Both verbal information, i.e., conversational content, and nonverbal information, such as utterance duration and gaze, enable the person who is sending information to others to be identified [12], [15]–[17]. Although verbal information enables detailed analysis of conversation structures and conversational patterns, analyzing the conversational content has potential risks of leaking information belonging to the organization and violating the privacies of workers. Fortunately, acoustic nonverbal information, such as on the utterance duration, turn-taking, and loudness, is a powerful instrument of conversation [18], [19], so it is useful for analyzing conversations. This study focuses on acoustic nonverbal information in conversations.

The amount of acoustic nonverbal information is positively related to dominance and leadership [11]–[13]. Thus, long-term analysis of nonverbal information in an office environment should reveal the dynamics of a group. Since research has shown that analysis of utterance duration [12], ut-
terance frequency [20], and loudness [21], for example, can reveal who dominated a conversation, long-term analysis of dominance based on such nonverbal information should reveal dynamic changes in conversational dominance. Such changes would likely reflect changes in group leadership.

In most of the previous studies on conversation analysis, especially psychological ones, the conversational situations were controlled, the data were recorded in a laboratory setting, and only particular conversational segments were analyzed. While this approach enables in-depth analyses of various types of conversation, the longitudinal changes in group communication remain hidden because controlled and clipped-out conversations in a laboratory setting differ from actual ones in many ways.

Aimed at understanding communication and social interactions in daily office work scenarios, various systems have been developed for wearable devices. Choudhury and Pentland analyzed social networks by having participants wear devices equipped with an infrared transceiver, a microphone, and two accelerometers for detecting the proximity of other persons, conversations, and motions [22]. Ara et al. used badge-shaped devices and investigated the social network in an organization [23]. Although these studies led to a better understanding of social networks, the roles of team members such as conversational dominance and its long-term changes were outside of their scope. To make such a detailed analysis through nonverbal information, the system should detect the utterances of each person. Furthermore, a non-worn system is preferable because it reduces the users’ effort and the risk of data loss if the device becomes detached or someone forgets to wear one.

One method of detecting individual utterances is speaker diarization [24], which detects utterances and identifies the speaker of each utterance. Although numerous studies on speaker diarization have been carried out, practical diarization in actual office work settings only on the basis of acoustic features is still a challenge because workers speak in various places. On the other hand, integration of acoustic speaker estimation and image-based human detection is another promising way of identifying the speaker in a room.

Speaker diarization on the basis of audio-video integration equals to specify the speaking person among the persons captured in the video image. Several studies adopted synchrony between the voice and the video image capturing faces for identifying the speaker [25], [26]. The limitation of this method is the requirement for face image. Other studies estimated the direction of arrival (DOA) of sound using array of microphones, and combined it with video-based human detection [27], [28]. Although use of DOA removes the need for face images, it makes difficult to distinguish speakers in adjacent directions. Thus, there is a system, which is composed of multiple microphone-arrays and a video module, has been developed [29]. However, the influence of multiple reflection and various noises, which is unavoidable in actual office work scenario, was unclear. One study applied audio-video speaker estimation system to a classroom, which resembles offices, by installing multiple microphones at the ceiling and calculating the time difference among the microphones [30]. However, its time resolution was 0.5 s and higher resolution has been desired for analyzing acoustic nonverbal information [31].

Therefore, we have also developed an audio-visual integrating speaker estimation system using distributed microphones at the ceiling of an office. Our system uses sixteen beam-forming microphone-arrays to reduce the influence of multiple reflection in actual office, and estimates the sound source location using the spatial distribution of sound power [32]–[34]. Furthermore, the system estimates nonverbal information every 0.1 s based on the estimated speaker for use in long-term analysis of conversations. The system was installed in an actual office with three assigned workers. Evaluations in terms of speech segmentation accuracy, speaker identification accuracy, and comparisons of estimated and human-coded nonverbal information have been performed. The main contribution of this study is to demonstrate the feasibility of acoustic nonverbal information estimation in actual office work scenario towards unconstrained long-term monitoring of daily office activity.

2. Acoustic Nonverbal Information Estimation System

2.1 Related Works and Design Approach

To estimate the nonverbal information of each group member, the system needs to detect utterances and identify the speaker of each utterance. Furthermore, subsecond time resolution is needed because a certain fraction of utterances is shorter than a second [31]. In addition, multiple reflection due to walls and furniture, the changing positions of workers, and the typically low sound intensities of voices in actual offices make speech detection especially challenging.

One method of identifying a speaker on the basis of only acoustic information is to use individual variations in voice [35] such as in Mel-frequency cepstral coefficients (MFCC) [36]–[38] and linear predictive coding [37], [39] along with vector quantization [36], [39] or a Gaussian mixture model [36]–[39]. However, reliable speaker identification requires a certain speaking duration. Thus, this method is not suitable for detecting backchannel feedback, which is frequently observed in daily office conversations.

Another approach is to estimate the DOA of sound based on the difference in time delay between microphones [40]. Methods using cross correlation [41], cross-power spectrum phase (CSP) analysis [42], [43], and methods to steer the formed beam and find the direction giving maximum power [44]–[46] have been proposed. However, these methods provide only the direction of the arriving sound, not the position of the sound source. Therefore, integration of multiple microphone arrays and direction information derived from them are needed to identify the position of the speaker in a room.

Yet another sound localization method is to use the ratio of the detected energies by using a set of coarsely-located
microphones, because acoustic energy attenuates as a function of distance from the sound source [47]. The location of a sound source was estimated from the energy ratios for the multiple microphone pairs by using the unconstrained least squares algorithm [32]. Other studies estimated the location of a sound by solving a maximum likelihood problem using an expectation-maximization-like iterative algorithm [33] or a projection-onto-convex-sets method [34]. However, their application to actual office scenarios remains a challenge because of multiple reflection.

Aiming for more robust speaker detection, the use of visual information in combination with acoustic information has been examined [24]. Nock et al. proposed a method to identify the speaker on the basis of the synchrony between the acoustic feature vector derived from MFCC and the video feature vector provided by the discrete cosine transform of lip image [25]. Similarly, Fisher et al. also calculated the mutual information between the voice and the head images and estimated the speaker as the person whose mutual information is maximum [26]. Integration of two pieces of information would enable more robust and accurate speaker detection. However, these methods require having face images, and such requirement limits the applicable scenes to those where people do not move much such as in a meeting.

Another line of studies combined acoustic DOA estimation with video-based human detection. Ishizuka et al. estimated the speaker by combining the speech presence probability calculated for each direction using a microphone-array at the center of a round table and the human presence probability [28]. Wakabayashi et al. enabled to be applied for environments where speech overlap occurs by using MULTiple SIgnal Classification (MUSIC) method. Since these methods estimate the speaker based on DOA [27], further improvement is required to discriminate multiple speakers in a same direction. Thus, D’Arca used multiple microphone-arrays [29]. They enabled tracking of occluded speaker by employing audio tracker based on extended Kalman filter and video-based human tracker with particle filter. However, similar to the other studies, its decrease in robustness and accuracy in actual offices, where multiple reflection and various noises are unavoidable, has not been studied. In addition, if a method on the basis of DOA in horizontal plane is applied to an office, installing the microphones may physically disturb the activities of the workers. Nishiguchi et al. developed another audio-video integrated system for being used in a classroom where severe multiple reflection occurs similar to actual offices [30]. They installed omnidirectional microphones at the ceiling and calculated the time difference among them using CSP analysis for robust sound source localization. However, the time resolution of the system was 0.5 s while higher resolution is desired for analyzing acoustic nonverbal information including shorter utterances [31].

Thus, we chose a policy to follow the path of the Nishiguchi’s study but with the aim of improving the time and spatial resolution to facilitate analysis of the nonverbal information for each person in the room. Our system uses sixteen beam-forming microphone-array modules installed close to the ceiling facing downward and calculates the spatial distribution of the sound pressure [32]–[34] for reducing the influence of multiple reflection on speaker estimation, with the aim of enabling long-term monitoring of a group’s dynamics through analysis of the acoustic nonverbal information in its conversations without restricting group members.

2.2 System Overview

The developed system estimates the speaker by integrating two types of position information: sound source position estimated using multiple beam-forming microphone-array modules and human position detected using a ceiling-installed camera.

Figure 1 (a) shows the configuration of the sensors, and Fig. 1 (b) shows a view of the system setup in an uninsulated laboratory room (7 m x 6 m). To detect voices in conversation and calculate their energy distribution for sound localization, 16 beam-forming microphone-array modules were mounted facing downwards on a 2.4-m-high frame. The spatial intervals between the microphones were set to 1 to 1.5 m because the distance-sound pressure characteristic of the microphone-array module is sharper within 1.5 m, as shown in Fig. 3. Each module has four microphones in a 50 mm x 50 mm square configuration [48], as shown in Fig. 1 (c). The sample rate and the bit depth of the mod-

---

**Fig. 1** (a) Configuration of sensors: circle, triangles, and square represent omnidirectional microphone, directional microphone-array modules, and wide field of view camera, respectively. (b) View of system setup: solid line shows camera, and dashed line represents microphone-array modules. (c) Microphone-array module. Sensors were mounted on 2.4-m-high frame.
ule were 16kHz and 16 bits. A camera (Point Grey Research, CMLN-13S2C-CS) for human detection was set on the frame at the center of the room. The sample rate, bit depth, and resolution of the camera were 0.5 fps, 16 bits, and 640 x 480 pixels. A low-distortion wide-field-of-view lens (Nittoh Kogaku K.K., SY125M) with a 125-degree horizontal viewing angle was attached to the camera for recording images of the whole room. An omnidirectional microphone (Yamaha, JP-25-UR) was set on the frame to manually code utterances for use in assessing system accuracy. The sample rate and the bit depth of the microphone were 44.1kHz and 16 bits. The system was set to automatically record data from 7:00 am to 7:00 pm every weekday. The system synchronized the signals from the microphone-array modules by generating an impulsive sound once every hour.

Figure 2 shows a diagram of the speaker estimation and acoustic nonverbal information estimation system, which is an offline system. First, the system determines speech segments by detecting human voices using both threshold processing on the sound pressure and voice activity detection (VAD) [49], [50]. Then, the speaker for each detected voice sample is identified. To unify the acoustic and image-based information for speaker estimation, we defined two 2-dimensional potentials. The first is the “sound source potential,” which is calculated from the sound pressures of the microphone modules and expresses the probability of a sound source existing at each position on the floor. Similarly, human position cannot be strictly determined from a camera image due to individual differences in body height and changes in posture. Thus, the probability of human existence for each position on the floor is described as a two-dimensional “human potential”. Since the speaker is the person who is speaking, our system estimates the speaker position as the position where the product of the two potentials takes a maximum value every 0.1 s.

The detected voice samples, which were labeled for each speaker, included false voice detections and missed detections due to low sound pressure. The system thus applied noise reduction and interpolation so that the speaker estimation was similar to that of human judgment. Details of the calculation are described below.

2.3 Speech Segmentation

First, the voice candidate samples were detected on the basis of the total sound pressure by the 16 beam-forming microphone-array modules. The threshold was set to 2.5 times the background noise level by trial-and-error adjustment. Although setting the threshold to 3.5 times the background level provided higher F-measure, we set lower threshold to reduce false rejections in expectation that VAD filter removes the increased false detections to some extent. Here, in actual office environments, the background noise level changes depending on the situation such as device operation. Therefore, the background noise level was dynamically adjusted by calculating the minimum sound pressure level for the preceding 60 s to ensure detection of the “background” noise level without occasional sounds from people and devices in the room. Although various methods of adaptive background noise subtraction have been proposed such as ones using median filter [51] and autoregressive Gaussian model [52], we chose a modest method that did not increase the risk of failure in detecting utterances with less acoustic power.

Then, a VAD filter that detects human voices on the basis of harmonicity and pitch deviation was applied. It was tuned especially not to miss conversational voice in daily office work scene [53]. Although the VAD filter reduced the number of false voice detections, it sometimes eliminated actual voice samples in error. However, it rarely eliminated all voice samples in an utterance. Therefore, the voice candidate samples within 5 s of the VAD-detected voice samples were also included in the speech segmentation.

2.4 Sound Source Potential

Although a previous study used the energy ratios of distributed microphones [32]–[34], we defined a potential that represents the probability of a sound source existing at each position on the floor for image-sound speaker estimation. Since the 16 directional microphone-array modules provide different sound pressure levels depending on their position, we defined a 16-dimensional normalized vector $f$, which denotes the ratio of the output levels of the modules. Here, if the position of the sound source is given (assume the magnitude of sound source is normalized), the sound pressure for each microphone module can be estimated because the position and the distance-sound pressure characteristic of the mi-
Fig. 3  Distance-sound pressure characteristic of beam-forming directional microphone-array module.

crophone are known. In other words, the theoretical vector, which is composed of the estimated sound pressure levels, for each floor position can be calculated by using the premeasured distance-sound pressure characteristic. The characteristic shown in Fig. 3 was obtained by repeatedly measuring the sound pressure derived from a speaker set at 1 m high. We varied the horizontal distance between the speaker and the beam-forming microphone-array module at the center of the room from 0 to 4 m.

By using this theoretical vector, we defined the sound source potential $S_{ij}$ for each position $(i, j)$ in the floor plane coordinate system. $S_{ij}$ is the reciprocal of the Euclidean distance between the theoretical vector $f_{ij}$ for the position and the vector $f$ measure by 16 microphone modules, as shown in Eq. (1). To cancel out the influence of the sound source level, we normalized both $f_{ij}$ and $f$ before the calculation. Since the calculated potential is inversely proportional to the difference between the theoretical and measured vectors, it approximates the probability of sound source existing at each point.

$$S_{ij} = \frac{1}{|f - f_{ij}|} \quad (|f - f_{ij}| \neq 0) \quad (1)$$

A shorter calculation interval of $S_{ij}$ is preferable, because it determines the time resolution of utterances. Previous studies in psychology reported that utterances shorter than 1 s account for a certain fraction of utterances, even after removing back channel feedbacks such as “yeah” [31]. Thus, to adequately analyze the lengths of the utterances, the time resolution of analysis needs to be sufficiently shorter than 1 s. On the other hand, multiple reflection, which produces pseudo-sound sources at various positions following the actual sound for several hundred milliseconds, are unavoidable in actual office spaces. Thus, the longer interval is preferred to reduce its influence. To balance these conflicting requirements with emphasis on time resolution, we set the calculation interval to 0.1 s.

2.5 Human Detection and Human Potential

The human area in the camera image was detected using the background image subtraction method after applying local binary pattern (LBP) processing [54], which is robust to lighting condition changes in both current and background images. However, the use of a fixed background image causes detection errors due to object movement, i.e., when objects are moved to different positions. Thus, the background image should be dynamically updated. Furthermore, small motions of persons working at their desks are difficult to detect by using a single subtraction. For that reason, image subtraction was done using dynamically updated images taken 0.5, 10, and 60 s before detection. The image sampling times were experimentally decided. If no difference was detected with the dynamically updated background images at a position where the LBP-filtered process recognized a human, that position was excluded from the human position candidates. This process enabled to reduce false detections due to object movement except for continuous ones. The persons were tracked on the basis of the proximity of the detected positions in the previous and the current frames.

The speaker was identified by integrating the detected human area and the estimated sound source position. The detected human area thus needed to be converted from the camera coordinate system into the floor-plane coordinate system. However, a single-camera image does not provide three-dimensional positions of the body parts. The head, which was the highest body part in most cases, was recorded closer to the image center. We thus assumed that the most peripheral part of the detected human area was the head. After applying this image recognition process, we manually corrected errors in the head position and human tracking using a GUI-based verification system.

As shown in Fig. 4, if the height of the camera position $H$ and the view-angle of the camera are known, one point in a camera image provides a corresponding point $P$ in the floor-plane coordinate system. Then, if the height of the human $h$ is known, the foot position $Q$ is given by

$$Q = P - \frac{h}{H}P \quad (2)$$

However, human height differs between individuals and varies with posture. Therefore, we introduced a probability function $p(h)$ that represents the probability of a human being at each position. $p(h)$ was determined for a range from 500 mm to 2000 mm high. We set the peak at 1300 mm in consideration of the height of a typical Japanese adult in a sitting position. This process provides a rough approximation of the human position as a probabilistic human exis-
tence potential $M$ at each point around the projected point $P$. If we assume that an 1800-mm-tall person is standing where the viewing angle from the camera is 45 degrees downward, the expected error due to the use of a prefixed probability function is 500 mm. Since it is longer than the intimate distance [55], we expected that it will not cause any serious misestimations of the speaker. The sample interval for the human potential calculation was set to 0.5 s because the time resolution requirement is looser than the utterance detection requirement.

2.6 Potential Integration and Speaker Identification

To identify the speaker, the integrated potential $E$ was defined as the product of the sound source potential $S$ and human potential $M$. The position at which $E$ reached a maximum was determined to be the speaker position.

$$E = S \cdot M$$

(3)

Figure 5 shows an example in which one of the three workers spoke. The distribution of sound source potentials around the speaker and the three distributions of human potentials around the room members are illustrated. Similar to this example, the system identified the speaker for each voice sample on the basis of these sound and human potentials.

2.7 Noise Reduction and Gap Interpolation

Figure 6 (a) shows an example of detected utterances, which are labeled by speaker ID. The isolated voice samples, i.e., ones with an utterance length of 0.1 s, were considered misdetections and removed. However, most utterances involve sound level decreases and a short blank period due to intonation or breathing. Thus, to avoid mistaken removals, such isolated samples were not removed if the same speaker’s voice was detected within 0.4 s before or after. After that, to remove the sill remained misdetections, the system removed the utterances shorter than 0.3 s. Figure 6 (b) shows the results of this process.

Next, the utterance gaps were interpolated. While people generally recognize the gaps due to intonation or breathing, i.e., short silences, as part of an utterance, speech segmentation based on sound pressure divides the utterance at such gaps. The system interpolated the gaps of up to 0.3 s, as shown in Fig. 6 (c). This duration was set in accordance with the finding that people generally recognize gaps of more than 0.4 s as a pause [56]. The results are similar to those of human judgment, which are shown in Fig. 6 (d).

2.8 Nonverbal Information Calculation

Aiming to analyze the conversations to obtain an understanding of the group’s dynamics, we calculated the nonverbal information, i.e., utterance duration, utterance frequency, and turn-taking frequency, for each worker in the office on the basis of the estimated speaker.

A “turn” was defined as the time from taking the right to talk or make a statement to losing that right. Backchannel feedback utterances are not regarded as a turn [17] and are generally too short to broaden the conversation. Utterances that broaden the conversation or maintain a conversation, such as ones expressing an opinion or providing information, are longer than other types of utterances [57]. Thus, the number of turns was calculated only for utterances longer than 1 s. A turn was calculated as the time from the beginning of the utterance longer than 1 s to the beginning of an utterance delivered by another person. Intervals longer than 6 s were regarded as the end of the turn, because people recognize utterance intervals longer than several seconds as unnatural turn-taking [58].
3. Evaluation

We conducted an experiment to evaluate the ability of our system to estimate acoustic nonverbal information during daily office activities.

3.1 Conditions

3.1.1 Environment

The system was installed in an office in which the three workers had been engaged in research for two years. Figure 7 shows the layout of the office furniture. The potential noise sources were a printer, servers, and a beam projector. The ambient noise level varied day by day from 41 to 43 dB(A). The equivalent continuous sound pressure level reached 52 dB(A) on Day 4, when renovation work was done in the next building. The reverberation time of the room was 0.2seconds at 1kHz. The room was air-conditioned at 26 degrees Celsius.

3.1.2 Participants

Table 1 shows the attributes of the participants (A, B, and C), who were key members of this study. They were faculty members (one man and two women) in their 30s and at the same academic level. Two of them specialized in information science, and the other specialized in social science. There were 11 (eight men and three women) occasional participants (visitors to the room), with ages ranging from in their 20s to in their 50s. They comprised three faculty members, two clerical workers, two technical experts, and four students. None of the participants had a language impairment or hearing disability.

3.1.3 Procedure

This study was approved by the ethical review committee of the university. The research objective, the information to be recorded, the intended data use, and the privacy protection policy were explained to the participants in writing. All the participants signed an informed consent form.

The experiment was conducted after the system development period, which took more than a year. As the main participants had worked in the room during that period, they had become accustomed to the experimental environment. Interviews showed that the system had already faded from the participants’ consciousness and did not affect their behavior.

3.1.4 Manual Coding of Individual Utterances

Two psychologists, who are experienced in behavior analysis, manually coded the recorded acoustic data in an offline environment for use in evaluating the system accuracy. One of them was also a main participant; the other did not know the purpose of this study.

The coders used audio player software (Audacity 2.0.5) to judge utterance existence and the speaker with 0.1 seconds resolution by listening to the recorded data while visually checking the speech waveforms. In the coding, backchannel feedback, such as “yes/no,” “uh,” “hmmm,” and laughing were judged as utterances. Coughs and throat clearings were not considered utterances.

Consistency between the coders was examined by calculating the kappa coefficient for 120 minutes of coded data; it showed high consistency (κ = 0.96). The coders then coded different parts of the 118 hours of the remaining data.

3.2 Data Analysis

3.2.1 Data Collection and Selection

The system recorded data for 12 hours a day for 25 days from April 16 to May 26, 2015, except for weekends.

As detailed at the bottom of Table 2, we selected data for ten of those days that would enable us to examine system performance under various conditions. We first eliminated those days when the three participants were not in the room together for more than 6 hours. From the remaining days, we selected ten days representing typical office scenarios such as desk work, discussion, physical task performance, and visitor reception.

The data for each day were categorized into three scenes.

1. Desk work: Each participant was sitting at his or her desk, basically performing desk work, but with occasional conversations. The distances between them
### 3.2.2 Fundamental Results

Table 2 shows the maximum number of participants, the sum of activity durations (SA), which represents how long the participants spent their time for each scene, and the ratio of utterance duration to activity duration (RU), which is the proportion of the summed length of utterances to each SA. RU was calculated on the basis of the manually coded utterance durations.

The average ratio of utterance duration to activity duration for all scenes was 0.25. Even when the participants performed individual tasks, the average ratio for the desk work scene was 0.17. These results suggest that conversation is indispensable to group task performance, one of the main activities in the office.

The average activity duration for the desk work scene was the longest, accounting for more than 50% of the total activity duration for most of the ten days. This indicates that the accuracy for the desk work scene dominated total accuracy.

### 3.3 Evaluation Results

System accuracy was evaluated by comparing the estimation results with the manually coded results. First, speech segmentation accuracy and speaker identification accuracy were examined individually. Then, the integrated estimation results, i.e., the speaker-labeled utterances, were evaluated.

#### 3.3.1 Speech Segmentation

Speech segmentation accuracy was calculated by dividing the number of correctly detected utterances by the number of detected utterances. The results are shown in the left columns of Table 3.

The average accuracy for all the scenes was 0.72. The average accuracies for desk work, gathering, and walking scenes were 0.70, 0.77, and 0.67, respectively. Since the algorithm and threshold were set to minimize detection failure, the number of false detections appears to be high. Despite the use of the VAD filter, the accuracy for Day 4, when there was noise due to renovation work in the building next door for most of the day, fell to 0.45. The larger drop in accuracy for the desk work scene than for the gathering scene is attributed to the smaller number of utterances, as shown in Table 2, which amplified the effect of false voice detection.

We also calculated the average rate of false rejections, which is the number of undetected utterances divided by the correct number of utterances. The rate including all scenes was 0.08. This implies that the group's communication activities and the member contributions were basically reflected in the estimation results.

#### 3.3.2 Speaker Identification

Speaker identification accuracy (independent of speech segmentation) was calculated by dividing the number of correctly identified utterances by the number of detected utterances. The results are shown in the middle columns of Table 3.

The average accuracy for all the scenes was 0.72. The average accuracies for desk work, gathering, and walking scenes were 0.70, 0.77, and 0.67, respectively. Since the algorithm and threshold were set to minimize detection failure, the number of false detections appears to be high. Despite the use of the VAD filter, the accuracy for Day 4, when there was noise due to renovation work in the building next door for most of the day, fell to 0.45. The larger drop in accuracy for the desk work scene than for the gathering scene is attributed to the smaller number of utterances, as shown in Table 2, which amplified the effect of false voice detection.

We also calculated the average rate of false rejections, which is the number of undetected utterances divided by the correct number of utterances. The rate including all scenes was 0.08. This implies that the group's communication activities and the member contributions were basically reflected in the estimation results.
The lower accuracy for the gathering scene is attributed to the difficulty of identifying the speaker when the participants were close together.

The average accuracy for all scenes was 0.86. The accuracy for Day 7, when seven participants conversed for more than 2 hours in a gathering scene, was 0.80, which is not greatly different from the accuracy for the other days. The relatively high accuracy for Day 4, when there was construction noise for most of the day, suggests that noise originating from other sources than the participants did not seriously affect speaker identification. These results indicate that the speaker can be identified for most correctly detected utterances.

3.3.3 Integrated Estimation Accuracy

Integrated estimation, i.e., accuracy including both speech segmentation and speaker identification, was calculated by dividing the number of correctly identified speakers by the number of detected utterances. If multiple participants spoke at the same time, only one of them was counted as the correctly identified speaker. The results are shown in the right columns of Table 3.

The average accuracy for all the scenes was 0.62. Exclusion of the data for Days 4 and 10, when there was noise for most of the day, improved accuracy to 0.67. The average accuracy for the desk work scene, which accounted for more than half the recorded data, was a bit higher, 0.71. As shown in Table 3, the main cause of the relatively low integrated estimation accuracy was lower speech segmentation accuracy due to false voice detection. Therefore, excluding utterances shorter than 1 s, which presumably include more noise, should improve the integrated estimation accuracy.

3.3.4 Acoustic Nonverbal Information

The system calculated the acoustic nonverbal information for each speaker on the basis of the detected utterances longer than 1 s. Figure 8 shows the accumulated duration and frequency of the detected utterances longer than 1 s and the turn-taking frequency of the main participants, who were in the room for most of the ten days.

The average estimated acoustic nonverbal information for each participant, i.e., the duration and frequency of utterances longer than 1 s, and the turn-taking frequency, except for Day 4, were 1.51, 1.33, and 1.54 times higher than the human-coded values for all ten days. The higher estimation results are attributed to false voice detections, as shown in Table 3.

Nevertheless, even when seven participants were gathered close together in a discussion (Day 7) and when the participants made more noise (Day 10), the order of the participants (A, B, C) by type of calculated acoustic nonverbal information coincided with that of human coding for all three types. These results suggest the feasibility of automatically quantifying the contribution of each member to group communication over time. Additionally, the successful estimation of the turn-taking frequency, which reflects interactions among participants, suggests the feasibility of analyzing the dynamics of a group by calculating the entire set of conversational features, such as the conversational floor, from the estimated nonverbal information.

4. Discussion

4.1 Further Improvement of Acoustic Nonverbal Information Estimation System

In the image-based human detection process, the dynamic update of the background images allowed us to reject moved objects by persons after 60 s at longest, because no autonomously moving objects were in the target office. To apply for environments with autonomously moving objects such as robots, more complex rejecting process is needed. Moreover, manual correction of errors in the head position and human tracking were needed. Since this process prevents fully automated analysis, we are developing a human tracking system employing multiple depth cameras, which enables more accurate human tracking through active mea-
Fig. 8 Acoustic nonverbal information for main participants based on estimated and human-coded utterances: (a) accumulated duration of utterances longer than 1 s, (b) frequency of utterances longer than 1 s, and (c) turn-taking frequency.

As shown in Fig. 8, the order of the amount of calculated acoustic nonverbal information among the participants coincided with that for the human-coded utterance results. This suggests the feasibility of automatic long-term analysis of the group dynamics in an office without restricting the members by using the estimated acoustic nonverbal information.

As shown in the left columns of Table 3, 28% of the utterances, including ones shorter than 1 s, detected by the system were false detections. In this study, acoustic nonverbal information was calculated on the basis of detected utterances longer than 1 s. Introduction of additional microphones located closer to the participants for voice detection and VAD improvement should reduce the effect of the background noise on speech segmentation and improve accuracy. By doing so, more detailed analyses of conversations, such as backchannel feedback, which has less verbal information but nevertheless plays a certain role in conversation turn-control, will be possible.

The relatively lower accuracy of speaker identification in the gathering scene is mainly attributed to the shorter distances among the participants. However, improving spatial resolution is a big challenge in sound source position estimation. One promising approach is to use directional information from the microphone-array modules (this study used only the distribution of localized sound pressures). To do that, the modules would need to be installed closer to the participants because sound source direction estimation is sensitive to multiple reflection.

Another potential method is to integrate multiple cameras located on desktops and detect the lip motion of the users by using face recognition technology [25], [26], [59]. This may complement VAD in distinguishing actual voices from environmental noise, especially in deskwork scenes. Face recognition may also enable a misidentified speaker to be correctly identified as the person who spoke. Further study is needed in this regard.

4.2 Case Study of Estimating Conversation Participation Styles towards Understanding of Group Dynamics

Conversation participant styles are proposed and analyzed, as detailed in Sects. 4.2.1 and 4.2.2. The possibility of estimating conversation participation styles on the basis of acoustic nonverbal information estimated by the system was examined for the purpose of achieving unrestricted and long-term analysis of group dynamics.

The data for Days 1 and 5 were analyzed for various conversation scenarios, such as report, discussion, and sug-
The conversation participation styles were calculated using tentative calculation rules, the utterances longer than 1 s, and the turn-taking frequency. The duration of an utterance reflects the amount of information provided, and the turn-taking frequency reflects the degree of engaging in a conversation to develop or maintain the conversation.

4.2.1 Conversation Participation Styles

Conversation participation styles reflect how people engage in or contribute to a conversation. According to previous studies, conversation participation styles can be classified into three types: information provider, cooperative commentator, and information receiver.

Leadership and dominance are positively and strongly correlated with utterance duration [11], [13]. Therefore, a person who demonstrates leadership or dominance in a conversation tends to proactively develop and control the conversation, such as by expressing his or her own opinions, providing new or given information, and speaking longer than others. Such a person is thus defined as an information provider. The tentative calculation rule we used for identifying the information provider was to define the person with the longest utterance duration as the information provider.

Followers are people who actively take part in helping their organization succeed [60]. In a conversation, followers tend to play a vital role by helping the leader develop and maintain the conversation. They do this by asking questions and/or making helpful comments. Therefore, followers are defined as cooperative commentators. Since cooperative commentators support the information provider, their turn-taking frequency might be similar or slightly lower than that of the information provider. The tentative calculation rule we used for identifying the cooperative commentators was to define participants with a turn-taking frequency that was more than 70% of that of the information provider as cooperative commentators.

In contrast to the information provider and cooperative commentators, an information receiver tends not to help a conversation but rather to react to it passively by speaking short phrases, i.e., giving backchannel feedback, or by responding with either “yes” or “no” or a short response to questions. Therefore, an information receiver may have less opportunity to take a turn. The tentative calculation rule we used for identifying the information receivers was to define participants with a turn-taking frequency that was less than 70% of that of the information provider as information receivers.

4.2.2 Preliminary Estimation of Conversation Participation Styles

Figure 9 shows the estimated conversation participation styles based on the total of the estimated utterance duraton...
tions longer than 1 s and the turn-taking frequency every ten minutes and on the actual dynamics as judged by an experimenter using the recorded data.

According to the results for the two days, the estimated conversation participation styles for each participant were the same as the judged ones, except for two situations caused by false voice detections (i.e., 12:10–12:30 on Day 1 and 9:50–10:10 on Day 5). These results indicate the possibility of estimating conversation participation styles by using acoustic nonverbal data.

On Day 5, participant A dominated the conversations by expressing her opinions and/or providing new information and demonstrated leadership; therefore, she is considered to have been the leader on Day 5. The estimation results indicate that she acted as the information provider most of the time. These results suggest the possibility of estimating the dynamics of leadership by using acoustic nonverbal data. If so, long-term analysis of conversation participation styles might enable identification of the group leader.

When participants A and B had a discussion on Day 5 from 8:30 to 9:30, they alternatively acted as the information provider and the cooperative commentator. In contrast, when participant A proposed a method for improving the system and explained it to participant C on Day 1 from 15:50 to 17:30, participant C asked questions and made comments about the proposal and thus acted only as a cooperative commentator. These results suggest that the dynamic changes in the conversation participation styles reflect the type of conversation, e.g., discussion or proposal. They also suggest that the cooperative commentator demonstrates followership by helping the information provider develop or maintain the conversation.

This case study was performed to examine the potential for long-term analysis of acoustic nonverbal information estimated by the system. Further studies are needed that focus on the calculation rules and/or the resolution time of the conversation styles.

4.3 Expected Effect on Social Sciences

Non-restrictive, automatic, long-term analysis of daily office conversations on the basis of calculated acoustic nonverbal information rather than analyzing controlled and clipped-out conversations in a laboratory, as was done in previous work, would enable provision of natural data reflecting real-world situations and social context to better understand group dynamics and to expand the foundation of psychological knowledge.

In addition to the conversation participation styles discussed in Sect. 4.2, dominance, leadership, and floor type could be analyzed over a long period on the basis of utterance duration [12], [61] and utterance frequency [20], leading to a better understanding of group dynamics.

While participant A, who spoke the most during the ten days, was the consistent group leader, a longer analysis period might reveal dynamic changes in leadership.

Members of organizations these days are expected to work more autonomously, and organizational structures have become more fluid and dynamic [62]. Our approach should provide managers with clues that would enable them to better understand individual and organizational changes. For example, a manager could monitor the growth in a person’s leadership skills and see the effects of personnel changes.

4.4 Feasibility of Other Applications

Progress in artificial intelligence and robot technology has made possible service robots and agents for the home as well as office. Furthermore, robots are acquiring both verbal and nonverbal communication functions [63], [64]. Beyond the original scope of this study, the system might be able to be used for understanding communications in mixed human-non-human groups [65], [66].

Furthermore, real-time analysis of communication may allow us to optimize automated services. For instance, it might help service robots to provide suitable services at suitable times. Automatic recognition of conversation status will let smart-speakers, which is rapidly growing commercially, actively speak to their target client without disrupting the ongoing conversation. To realize such applications, the design of the system, especially its hardware, should be simplified.

4.5 Limitations

The developed system estimates acoustic nonverbal information on the basis of individual utterances estimated by assuming a single sound source. Overlapping utterances and simultaneous conversations were beyond the scope of this work. Future work will hence include improving the position estimation algorithm so that it can be applied to multiple sound sources.

Moreover, this study targeted office workers who mainly talk to and discuss things with colleagues in the same room. Future work will thus include modification of the algorithm for calculating nonverbal information, especially turn-taking, so that it can handle cases in which the workers frequently talk with coworkers over the phone. Furthermore, a comprehensive analysis covering various non-vocal communication media over the Internet such as e-mail and text chat in wider ranging scenarios would allow us to understand team communication in depth.

Lastly, this study assumed an environment in which the workers mainly work in an office. Analysis of conversations outside the office was outside its scope. Future work will thus include investigation of the use of wearable devices such as sociometer badges [22] in combination with the developed system to enable unified analysis of the conversations both inside and outside the office.

5. Conclusion

To understand the group dynamics in an office with three
workers through long-term monitoring of daily communication, we developed an acoustic nonverbal information estimation system. The system combined an energy ratio-based sound-source localization technique with beam-forming microphone-array modules as well as image-based human detection for robust speaker estimation for utterances of individuals. We used data collected from 7:00 am to 7:00 pm over the course of ten days to evaluate the system. The results showed that the order of the amount of estimated acoustic nonverbal information among the participants was the same as that for the human-coded information. These results demonstrate the feasibility of estimating acoustic nonverbal information over a long term.

Acknowledgments

We thank Professors Takafumi Saito, Toshiyuki Kondo, Hironori Nakajo, Kaori Fujinami, Seiji Hotta, Ikuko Shimizu, Katsuhide Fujita, Yuichiro Fujimoto, and Tomoaki Tsutsumi for their insightful comments and suggestions. We used data collected from 7:00 am to 7:00 pm over the course of ten days to evaluate the system. The results showed that the order of the amount of estimated acoustic nonverbal information among the participants was the same as that for the human-coded information. These results demonstrate the feasibility of estimating acoustic nonverbal information over a long term.

References

[1] C. Barnard, The function of the executive, Harvard University Press, Cambridge, 1938.
[2] M.A. West, Effective teamwork, practical lessons from organizational research, 3rd ed., John Wiley & Sons, Chichester, 2012.
[3] C.R. Leana and H.J. Van Buren, “Organizational social capital and employment practice,” Academy of Management Review, vol.24, no.3, pp.538–555, July 1999.
[4] C. Yu and D. Junshu, “A literature review of the effect of social capital: From the personal network perspective,” International Journal of Business and Social Science, vol.4, no.12, pp.251–259, Sept. 2013.
[5] J.-P. Hatala and J.G. Lutta, “Managing information sharing within an organizational setting: A social network perspective,” Performance Improvement Quarterly, vol.21, no.4, pp.5–33, Jan. 2009.
[6] I. Nonaka and H. Takeuchi, The knowledge-creating company: How Japanese companies create the dynamics of innovation, Oxford University Press, New York, 1995.
[7] R.L. Daft, Organization theory and design, 10th ed., South-Western College Pub., Florence, 2009.
[8] D. Katz and R.L. Kahn, The social psychology of organizations, 2nd ed., Wiley, New York, 1978.
[9] F. Gevirts, K. Eberhard, and M. Scheutz, “Team communication as a collaborative process,” Frontiers in Robotics and AI, vol.3, Oct. 2016.
[10] E. Goffman, Forms of talk, University of Pennsylvania Press, Philadelphia, 1981.
[11] B. Mullen, E. Salas, and J.E. Driskell, “Saliency, motivation, and artifact as contributions to the relation between participation rate and leadership,” Journal of Experimental Social Psychology, vol.25, no.6, pp.545–559, Nov. 1989.
[12] M.S. Mast, “Dominance as expressed and inferred through speaking time: A meta-analysis,” Human Communication Research, vol.28, no.3, pp.420–450, July 2002.
[13] R.T. Stein and T. Heller, “An empirical analysis of the correlations between leadership status and participation rates reported in the literature,” Journal of Personality and Social Psychology, vol.37, no.11, pp.1993–2002, Nov. 1979.
[14] R. Hayashi, “Floor structure of English and Japanese conversation,” Journal of Pragmatics, vol.16, no.1, pp.1–30, July 1991.
[15] M. Argyle, Bodily communication, 2nd ed., Routledge, New York, 1988.
[16] S. Duncan, “Some signals and rules for taking speaking turns in conversations,” Journal of Personality and Social Psychology, vol.23, no.2, pp.283–292, Aug. 1972.
[17] H. Sacks, E.A. Schegloff, and G. Jefferson, “A simplest systematics for the organization of turn-taking for conversation,” Language, vol.50, no.4, pp.696–735, Dec. 1974.
[18] A.P. Andersen, Nonverbal communication: Forms and function, 2nd ed., Waveland Press, Long Grove, 2008.
[19] J.K. Burgoon, L.K. Guerrero, and K. Floyd, “Relational messages: Power, dominance, and influence,” in Nonverbal communication, eds. J.K. Burgoon, pp.334–372, Allyn & Bacon, Boston, 2010.
[20] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez, “Estimating dominance in multi-party meetings using speaker diarization,” IEEE Trans. Audio, Speech, Language Process., vol.19, no.4, pp.847–860, May 2011.
[21] K.J. Tusing and J.P. Dillard, “The sounds of dominance: Vocal precursors of perceived dominance during interpersonal influence,” Human Communication Research, vol.26, no.1, pp.148–171, Jan. 2000.
[22] T. Choudhury and A. Pentland, “Modeling face-to-face communication using the sociometer,” Proc. ACM Conference on Ubiquitous Computing (Workshop: Supporting Social Interaction and Face-to-face Communication in Public Spaces), Seattle, USA, pp.3–8, Oct. 2003.
[23] K. Ara, N. Kanchira, D.O. Olguin, B.N. Waver, T. Kim, A. Mohan, P. Gloo, R. Laubacher, D. Oster, A.S. Pentland, and K. Yano, “Sensible organizations: Changing our businesses and work styles through sensor data,” Journal of Information Processing, vol.16, pp.1–12, 2008.
[24] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, “Speaker diarization: A review of recent research,” IEEE Trans. Audio, Speech, Language Process., vol.20, no.2, pp.356–370, Feb. 2012.
[25] H.J. Nock, G. Iyengar, and C. Neti, “Speaker localisation using audio-visual synchrony: an empirical study,” Proc. 2nd International Conference on Image and Video Retrieval, Illinois, USA, Lecture Notes in Computer Science, vol.2728, pp.488–499, Springer, Berlin, Heidelberg, July 2003.
[26] J.W. Fisher and T. Darrell, “Speaker association with signal-level audio-visual fusion,” IEEE Trans. Multimed., vol.6, no.3, pp.406–413, June 2004.
[27] Y. Wakabayashi, K. Inoue, H. Yoshimoto, and T. Kawahara, “Speaker diarization based on audio-visual integration for smart posterboard,” Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Chiang Mai, Thailand, pp.1–4, Dec. 2014.
[28] K. Ishizuka, S. Araki, K. Nakatani, and M. Fujimoto, “A speaker diarization method based on the probabilistic fusion of audio-visual location information,” Proc. 11th International Conference on Multimodal Interfaces, Cambridge, USA, pp.55–62, Nov. 2009.
[29] E. D’Arca, N.M. Robertson, and J. Hopgood, “Using the voice spectrum for improved tracking of people in a joint audio-video scheme,” Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, Canada, pp.3622–3626, May 2013.
[30] S. Nishiguchi, K. Higashi, Y. Kameda, K. Kakusho, and M. Mi-noh, “Audio and visual information integration for speaker’s localization in automatic shooting of lecture,” IEEE Trans. Electronics, Information and Systems, vol.124, no.3, pp.729–739, June 2004 (in Japanese).
[31] J. Edlund, M. Heldner, S. Al Moubayed, A. Gravano, and J. Hirschberg, “Very short utterances in conversation,” Proc. Fonetik, Lund, Sweden, pp.11–16, June. 2010.
[32] D. Li and Y.H. Hu, “Energy-based collaborative source localization using acoustic microsensor array,” EURASIP Journal on Advances in Signal Processing, vol.2003, no.4, pp.321–337, Dec. 2003.

[33] X. Sheng and Y.-H. Hu, “Maximum likelihood multiple-source localization using acoustic energy measurements with wireless sensor networks,” IEEE Trans. Signal Process., vol.53, no.1, pp.44–53, Dec. 2005.

[34] D. Blatt and A.O. Hero, “Energy-based sensor network source localization via projection onto convex sets,” IEEE Trans. Signal Process., vol.54, no.9, pp.3614–3619, Sept. 2006.

[35] S. Furui, “Selected topics from 40 years of research on speech and speaker recognition,” Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Istanbul, Turkey, vol.2, pp.1788–1791, Oct. 1996.

[36] M.S. Brandstein and H.F. Silverman, “A practical methodology for speech source localization with microphone arrays,” Computer Speech and Language, vol.11, no.2, pp.91–126, April 1997.

[37] C.H. Knapp and G.C. Carter, “The generalized correlation method for estimation of time delay,” IEEE Trans. Acoustics, Speech, Signal Process., vol.24, no.4, pp.320–327, Aug. 1976.

[38] T. Nishiura, T. Yamada, S. Nakamura, and K. Shikano, “Lo- calization of multiple sound sources based on a CSP analysis with a microphone array,” Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, vol.2, pp.1053–1056, June 2000.

[39] M. Omologo and P. Svaizer, “Acoustic event localization using a T. Nishiura, T. Yamada, S. Nakamura, and K. Shikano, “Improvement of interruptibility estimation during PC work by reflecting conversation status,” IEICE Trans. Inf. & Syst., vol.E97-D, no.12, pp.3171–3180, Dec. 2014.

[40] M. Heikkila and M. Pietikainen, “A texture-based method for modeling the background and detecting moving objects,” IEEE Trans. Pattern Anal. Mach. Intell., vol.28, no.4, pp.657–662, Feb. 2006.

[41] J. Ramírez, J.M. Górriz, and J.C. Segura, “Voice activity detection, Fundamentals and speech recognition system robustness,” in Robust Speech Recognition and Understanding, eds. M. Grimm and K. Kroschel, pp.1–22, 2007.

[42] Hitomi Yokoyama received her M.D. and Ph.D. degrees in Human Sciences from Osaka University, Japan in 2008 and 2012, respectively. After working at Tohoku University and Tokyo University of Agriculture and Technology, she currently works as an associate professor in the Department of Management at Okayama University of Science.
Masano Nakayama received the B.Eng., M.Eng., and Ph.D. degrees from the Tohoku University, Japan, in 2007, quantum science and energy engineering and in 2009 and 2012 in aerospace engineering, respectively. After working at Graduate School of Media Design, Keio University, she joined Tokyo University of Agriculture and Technology in 2013. Her research interests included surgery simulator, mechanical engineering and haptics. Since 2016, she currently works at KATO MFG. CO., LTD. as an engineer.

Hiroaki Murata received his M.S. and Ph.D. degrees from Kanazawa University in 2008 and 2010, respectively. After working at Kanazawa University, he joined Tokyo University of Agriculture and Technology in 2013. His research interests include heuristics and swarm intelligence. He currently works as a system engineer.

Kinya Fujita received his Ph.D. degree from Keio University, Japan, in 1988. After working at Sagami Institute of Technology, Tohoku University, and Iwate University, he joined Tokyo University of Agriculture and Technology. He is currently working as a Professor in the Department of Computer and Information Sciences. His research interest is the development of smart human interface systems including interruption controls, remote communication systems, and haptics in virtual reality. He is a member of IEEE, ACM, and IPSJ, and a board member of HIS and VRSJ.