A Speaker Diarization Method based on the Probabilistic Fusion of Audio-Visual Location Information

Kentaro Ishizuka
NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho
Kyoto 619-0237 Japan
ishizuka@cslab.kecl.ntt.co.jp

Shoko Araki
NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho
Kyoto 619-0237 Japan
shoko@cslab.kecl.ntt.co.jp

Kazuhiro Otsuka
NTT Communication Science Laboratories, NTT Corporation
3-1 Morinosato-Wakamiya
Atsugi 247-0198 Japan
otsuka@eye.brl.ntt.co.jp

Tomohiro Nakatani
NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho
Kyoto 619-0237 Japan
nak@cslab.kecl.ntt.co.jp

Masakiyo Fujimoto
NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho
Kyoto 619-0237 Japan
masakiyo@cslab.kecl.ntt.co.jp

ABSTRACT
This paper proposes a speaker diarization method for determining "who spoke when" in multi-party conversations, based on the probabilistic fusion of audio and visual location information. The audio and visual information is obtained from a compact system designed to analyze round table multi-party conversations. The system consists of two cameras and a triangular microphone array with three microphones, and can cover a spherical region. Speaker locations are estimated from audio and visual observations in terms of azimuths from this recording system. Unlike conventional speech diarization methods, our proposed method estimates the probability of the presence of multiple simultaneous speakers in a physical space with a small microphone setup instead of using a cascade consisting of speech activity detection, direction of arrival estimation, acoustic feature extraction, and information criteria based speaker segmentation. To estimate the speaker presence more correctly, the speech presence probabilities in a physical space are integrated with the probabilities estimated from participants' face locations obtained with a robust particle filtering based face tracker with two cameras equipped with fisheye lenses. The locations in a physical space with highly integrated probabilities are then classified into a certain number of speaker classes by using on-line classification to realize speaker diarization. The probability calculations and speaker classifications are conducted on-line, making it unnecessary to observe all the conversation data. An experiment using real casual conversations, which include more overlaps and short speech segments than formal meetings, showed the advantages of the proposed method.

ICMI-MLMI'09, November 2–4, 2009, Cambridge, MA, USA.

Categories and Subject Descriptors
H.1.2 [**Models and Principles**]: User/Machine System – *Human Information Processing.*

General Terms
Algorithms, Human Factors

Keywords
Multi-party conversation analysis, Speaker diarization, Multi-modal systems

1. INTRODUCTION
Many multi-party conversations have been recorded in relation to, for example, Augmented Multi-party Interaction (AMI) and AMI with Distant Access (AMIDA) [1], Computers in the Human Interaction Loop (CHIL) [2], and the NIST Rich Transcription Meeting Recognition Project [3]. These corpora have triggered both an analysis of multi-party interactions to determine human behavior in these interactions, and the development of automatic indexing systems for multi-party conversations that can process multi-modal information. The automatic indexing of multi-party conversations such as group meetings can allow their rapid retrieval from archives, automatic minute taking, and automatic summarization of conversations [4]–[7]. One essential function for indexing such recorded data is to answer the question, "who spoke when?" during the conversations, and this can be determined by using speaker diarization methods [8]–[14].

Typical conventional speaker diarization systems [8]–[14] assume the use of all the recorded conversation data, and have realized speaker diarization by the cascade combination of speech activity detection (SAD) and information criterion based speaker segmentation and classification. The SAD methods are usually based on a Gaussian mixture model (GMM) [15]–[17], and its acoustic feature is generally the mel-frequency cepstral

coefficient (MFCC), which is widely used in automatic speech recognition. Speaker segmentation and classification are generally based on the Bayesian information criterion (BIC), and classify GMMs trained with MFCCs obtained from speech segments into a certain number of speaker classes [8]–[14]. The classification part iteratively trains GMMs for each speech segment, and merges the most similar pair of GMMs based on the BIC criterion. Since conventional recordings of multi-party conversations also include recordings obtained with multiple distant microphones [1]–[3], some speaker diarization systems utilize spatial information to detect speech activity [18][19], enhance speech signals [20], and cluster the speech segments [21][22]. According to the evaluation results of the recent NIST Rich Transcription Project, speaker diarization performance for meeting recordings can be significantly improved by employing information corresponding to the direction of arrival (DOA) of speech as a feature vector in addition to MFCC features [3][9][22]. This can be considered because the location of a speaker in a physical space correlates closely with the speaker identity in a meeting, that is, the participants move very little. Most methods employ the time difference of arrival (TDOA) estimated with the generalized cross correlation using the phase transform (GCC-PHAT) [23]. Although TDOA is a very strong cue as regards detecting coherent signals such as speech produced from certain positions corresponding to speaker locations, TDOA obtained with the GCC-PHAT [20]–[22] cannot deal with multiple speakers within one audio signal analysis window that has a certain temporal length such as 32 ms. Therefore, a mechanism that can detect overlapping speech is also needed to deal with conversations containing a lot of overlapping speech.

Otsuka *et al.* have recently developed a system for analyzing multi-party conversations [24]. This system can automatically detect and visualize "who is speaking" and "who is looking at whom" in real time by using a noise robust SAD [25], a speaker location detection method based on the DOA estimations for each time-frequency bin to handle simultaneous speech [26], and a face location and pose tracking method that employs Graphics Processing Unit (GPU) based particle filtering [27]. Although this system can visualize "who is speaking" based on the combination of the two audio processing components, visual information obtained with the face tracker has not been utilized in audio signal processing because the components have been simply connected in a cascade manner.

This study aims to realize speaker diarization based on the closer fusion of audio and visual information in a probabilistic manner. This study also focuses on the use of location information in a physical space because it is a particularly effective cue for speaker diarization for multi-party conversations as described above. In addition, this study aims to utilize the estimations of speech presence in a physical space for speaker diarization [35] instead of using the cascade of SAD, GCC-PHAT based DOA estimation, and BIC based speaker segmentation, which is widely employed by typical conventional speaker diarization systems [8]–[14]. By estimating speech presence probabilities in a physical space, we can deal with the overlapping speech. This is unlike GCC-PHAT based DOA estimation, which can estimate only one DOA within each audio analysis window.

The fusion of audio and visual information with a view to analyzing multi-party conversations has also been widely studied

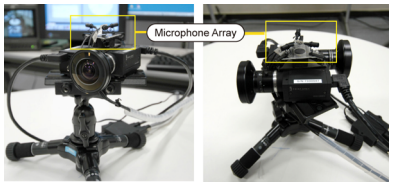

Figure 1. Omnidirectional camera-microphone system.

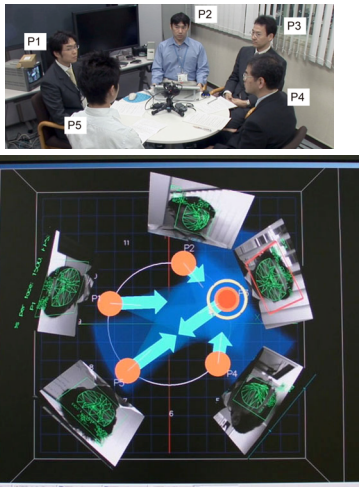

Figure 2. (Top) A round table meeting that can be analyzed with the recording system [24]. (Bottom) An example visualization result obtained with this system. The arrows show "who is looking at whom", and the red circle shows "who is speaking now".

in recent years. For example, speaker tracking [28][29], group action detection [30], speech event detection [31], attention recognition [32], and participants' interaction detection [33][34] have been realized by fusing audio and visual information. The fusion proposed in this study is designed to improve the speaker diarization performance obtained with the conventional audio information based system [14] by introducing visual information. A conventional multi-modal speaker diarization study combined MFCC with video features [36], while this study employs location information obtained from audio signals. It should be noted that the performance of the location information based speaker diarization method dealt with in this study is expected to be improved by combining it with conventional acoustic features that reflect speaker characteristics, such as MFCC. Moreover, this study aims to realize on-line speaker diarization, that is, unlike conventional speaker diarization, it does not require knowledge of all the conversation data.

The remainder of this paper is organized as follows. Section 2 describes the recording system setup. Section 3 provides a detailed explanation of the proposed speaker diarization method. Section 4 reports an evaluation experiment that employed real casual multi-party conversations. Section 5 concludes this study.

2. RECORDING SYSTEM

The recording system used in this study consists of two cameras with fisheye lenses and a triangular microphone array with three microphones [24]. This system has been developed for analyzing round table meetings attended by up to about eight people, and it is positioned in the center of the round table during the recording.

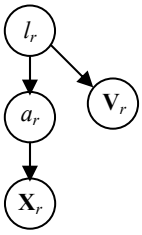

Figure 3. Graphical model that captures speaker locations and speech activities in conversations.

This system can both record the meetings, and process audio and visual information, and thus visualize "who is speaking" and "who is looking at whom" in real time. The distinctive characteristics of this system are its compact sensor setup and its use of real-time processing. With a view to developing a highly portable system this recording setup is rather compact unlike the equipment used in conventional recording projects [1]–[3], where multi-party interactions are observed with multiple sensors incorporated in a smart room as an ambient intelligence. The system setup and an example visualization result are shown in Figs. 1 and 2, respectively. This compact recording system setup can be used almost anywhere. That is, the setup has high portability, and so the system can be used for recording small meetings or conversations in small meeting spaces. If such recordings can be analyzed automatically, we can easily mine important data from our own daily conversations. Such applications cannot be realized with conventional large-scale recording setups.

The hardware settings are as follows (details are provided in [24]). The system consists of two personal computers (PCs) for processing audio and visual information, respectively. The two PCs are connected via a gigabit ether network for information sharing. The PC used for visual information processing has an Intel Core 2 Extreme QX9650 3.0GHz central processing unit (CPU), and its GPU is an NVIDIA GeForce9800GX2 (two GPU cores are installed in one package). Its operating system (OS) is Windows XP SP2. The cameras are Point Grey Research's Grasshoppers (B/W 5.0 Mega pixel model, 2/3" CCD). The fisheye lens is Fujinon's FE185C086HA-1 (f=2.7mm). The camera and PC are connected by IEEE1394b links. The audio processing PC uses an AMD Athlon 64, 2.4 GHz as CPU, and its OS is Linux.

In this study, audio signals are recorded at a sampling rate of 16 kHz and with 16-bit quantization. As regards visual processing, the resolution of the recorded image is 4896 pixels wide for 360-degree coverage on the horizontal plane and 512 pixels high, and the grabbing frame rate is up to 30.0 frames per second (fps).

3. METHOD

Speaker diarization is defined as the problem of detecting "who spoke when" from observed signals. Depending on the system setup, it can be considered that speaker location (azimuth) corresponds to speaker identification by assuming that the participants do not move a great deal. Therefore, this study realizes speaker diarization by detecting speaker locations and speech activities.

With the above as a basis, the proposed method considers local observations from azimuth r, and uses the generative graphical model shown in Fig. 3. In this model, l_r, a_r, $\mathbf{X}_r$, and $\mathbf{V}_r$ represent the proposition of speaker presence, the proposition of speech activities, audio observations, and visual observations for azimuth r, respectively. The proposed method estimates the posterior probability $p(a_r, l_r | \mathbf{X}_r, \mathbf{V}_r)$ for each r, and compares it with a threshold to realize speaker diarization. The posterior probability is given as follows.

$$p(a_r, l_r | \mathbf{X}_r, \mathbf{V}_r) = p(a_r, l_r, \mathbf{X}_r, \mathbf{V}_r) / p(\mathbf{X}_r) p(\mathbf{V}_r) \quad (1)$$

where it is assumed that $\mathbf{X}_r$ and $\mathbf{V}_r$ are mutually independent. The right side of equation (1) can be written as follows.

$$p(a_r, l_r, \mathbf{X}_r, \mathbf{V}_r) / p(\mathbf{X}_r) p(\mathbf{V}_r)$$

$$= p(\mathbf{V}_r | a_r, l_r, \mathbf{X}_r) p(a_r, l_r, \mathbf{X}_r) / p(\mathbf{X}_r) p(\mathbf{V}_r) \quad (2)$$

$$= p(\mathbf{V}_r | l_r) p(\mathbf{X}_r | a_r, l_r) p(a_r, l_r) / p(\mathbf{X}_r) p(\mathbf{V}_r) \quad (3)$$

$$= p(\mathbf{V}_r | l_r) p(a_r, l_r | \mathbf{X}_r) / p(\mathbf{V}_r) \quad (4)$$

$$\propto p(\mathbf{V}_r | l_r) p(a_r, l_r | \mathbf{X}_r) \quad (5)$$

where $p(\mathbf{V}_r | a_r, l_r, \mathbf{X}_r) = p(\mathbf{V}_r | l_r)$ is assumed to obtain equation (3). Equation (4) is obtained by applying the Bayes' theorem to the second terms of the denominator of equation (3). In addition, equation (5) is obtained assuming the prior probability $p(\mathbf{V}_r)$ is a constant.

Based on equation (5), audio and visual information can be integrated for speaker diarization by multiplying the probabilities obtained with two components as follows. First, $p(a_r, l_r | \mathbf{X}_r)$ can be considered a joint probability of speaker location and speech presence that can be estimated from acoustic signal spectra. Second, $p(\mathbf{V}_r | l_r)$ can be considered a participant's location probability that can be estimated from visual observations.

The following sections describe how to estimate the above probabilities from the audio and visual observations. Audio and visual information is also analyzed with analysis frames of a certain temporal length in this study. The above posterior probabilities are estimated frame by frame.

3.1 Speech presence probability estimation

Speech presence probabilities in a physical space are estimated by utilizing the framework of a likelihood ratio test (LRT) based SAD approach for multi-party conversation analysis [35]. This approach is robust as regards environmental noise without *a priori* knowledge of the acoustic conditions, speaker locations, or number of speakers. In addition, this method can deal with simultaneous speech even if the speakers outnumber the microphones. This method adopts the LRT approach [37][38] for the spatial distributions of the magnitude of observed signals estimated from multi-microphone recordings. Unlike the conventional LRT based SAD method with a microphone array [38], which utilizes *a priori* signal-to-noise ratios (SNRs) for the spectrum obtained from the estimated speaker location, this method utilizes only *a priori* SNRs obtained from the spatial distributions of the magnitude of the observed signals. This method detects speech activities by utilizing the changes in

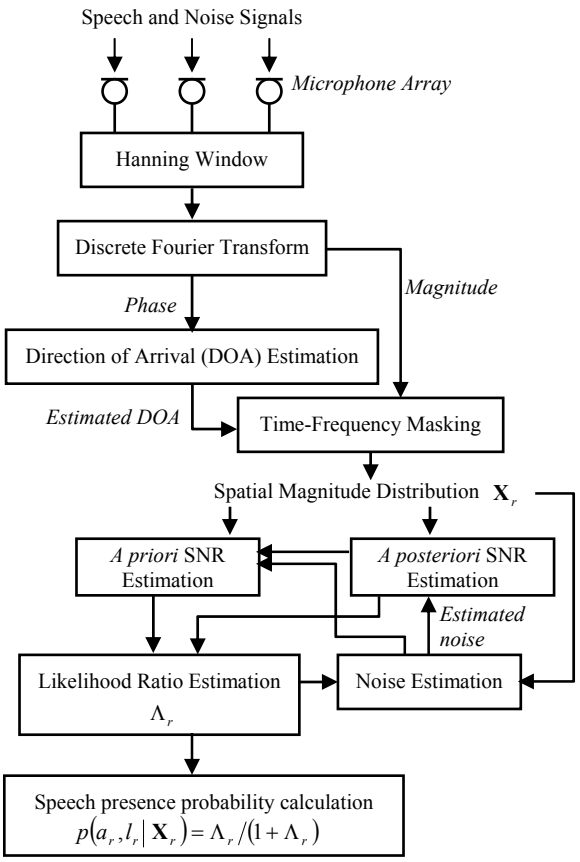

Figure 4. Block diagram of estimation of speech presence probability.

magnitude that come from a certain spatial range, and mitigates the effect of environmental noise incorporating stationary directional noise by employing *a priori* SNRs obtained from spatial magnitude distributions.

Figure 4 shows a block diagram of the estimation of speech presence probability. Signals observed with three microphones are analyzed using a Hanning window with a certain amount of overlapping. Then, the spatial magnitude distributions of the observed signals are estimated by applying time-frequency masking [39] to the frequency spectra based on the estimated DOA for each time-frequency bin of the discrete Fourier representation of the speech signals [26]. It should be noted that this method performs well only when the distance between microphones is small enough to avoid spatial aliasing because the DOA estimation uses phase differences. The space observed by the microphone array is split into R discrete spatial regions, and the spatial magnitude is estimated for each region. R is given before the processing. An example of the estimated spatial magnitude distribution is shown in Fig. 5(a). This study considers this estimated magnitude distribution as the audio observation in equation (1).

To estimate the speech presence probabilities for each region r (r = 1 .. R), the two following hypotheses are considered for each analyzed frame assuming that the speech signal is uncorrelated with additive noise signals.

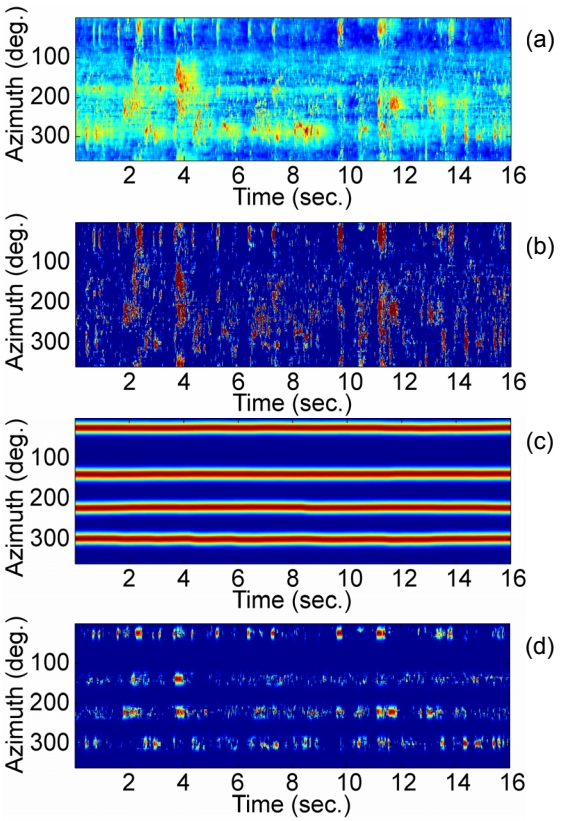

Figure 5. Example (a) spatial magnitude distribution $\mathbf{X}_r$, (b) speech presence probabilities $p(a_r,l_r|\mathbf{X}_r)$, (c) face presence probabilities $p(\mathbf{V}_r|l_r)$, and (d) posterior probabilities $p(a_r, l_r | \mathbf{X}_r, \mathbf{V}_r)$ obtained with (b) and (c). Red and blue indicate highest lowest values, respectively.

H_0: Speech absent; $\mathbf{X}_r = \mathbf{N}_r$

H_1: Speech present; $\mathbf{X}_r = \mathbf{S}_r + \mathbf{N}_r$

where $\mathbf{X}_r$, $\mathbf{N}_r$, and $\mathbf{S}_r$ indicate the estimated spatial power distribution, the noise power distribution, and the speech power distribution for region r, respectively. Based on the above two hypotheses, the log likelihood ratio is estimated as follows in this LRT based SAD method.

$$\Lambda_r = \frac{p(\mathbf{X}_r | H_1)}{p(\mathbf{X}_r | H_0)} \tag{6}$$

where $p(\mathbf{X}_r | H_i)$, i = 0, 1, is the probability density function (PDF) of the observed data under hypothesis H_i. The PDFs can also be modeled by Gaussian distributions because the magnitude distributions are estimated by the summation of discrete Fourier coefficients if we can assume that the coefficients follow a zero-mean Gaussian process. This also allows us to adopt conventional clean speech magnitude estimation methods such as the minimum-mean squared error based method [40] to estimate the *a priori* SNR of a certain spatial region. The log likelihood ratio Λ_r can be calculated by estimating the parameters of the Gaussian distributions. Once the likelihood ratio is obtained, the speech presence probability $p(a_r,l_r|\mathbf{X}_r)$ can be estimated as follows.

$$p(a_r, l_r | \mathbf{X}_r) = \Lambda_r / (1 + \Lambda_r) = p(\mathbf{X}_r | H_1) / (p(\mathbf{X}_r | H_1) + p(\mathbf{X}_r | H_0)) \quad (7)$$

An example of the estimated speech presence probabilities is shown in Fig. 5(b). This study uses this estimation as the second term of equation (5).

3.2 Face location probability estimation

The probability $p(\mathbf{V}_r|l_r)$ is estimated with a face location detection and tracking algorithm called sparse template condensation tracker (STCTracker) [27], which is built into the recording system [24]. When the tracker is initialized it can automatically build 3-D face templates, and it is robust as regards large head rotations of up to ±60 degrees in the horizontal direction. In addition, it can track multiple faces simultaneously in real time by particle filtering with a GPU.

The two fisheye omnidirectional cameras shown in Fig. 1 can cover a near spherical region, and the distortion caused by the fisheye projection is compensated by a panoramic transformation. The initialization part of STCTracker runs sparse template matching that focuses on a sparse set of feature points within a template region. The state of a template, which represents the location and pose of the face, is defined as a seven-dimensional vector consisting of a 2-degree of freedom (DOF) translation on the image plane, a 3-DOF rotation, a scale, and an illumination coefficient. The particle filtering part of STCTracker sequentially estimates the posterior density of the template state, which is represented as a particle set. The weight of each particle is calculated based on matching errors between the input images and the template. STCTracker works robustly in real time owing to the sparseness of the feature points and robust template matching combined with multiple-hypothesis generation/testing by the particle filter framework. Although the face model is rigid, it can accept a certain amount of facial deformation caused by events such as the production of utterances and facial expressions. Figure 2 also shows meshes that simultaneously track all the participants' faces.

This study considers the face azimuth θ^{V} estimated with STCTracker for each participant to be the centroids of that participant's location. It is assumed that the probability $p(\mathbf{V}_r|l_r)$ can be modeled by the estimation error distribution of STCTracker, and the error distribution is modeled by a Gaussian distribution whose mean and standard deviation are θ_n^{V}, which is the face azimuth of the participant n, and σ, which has a certain value (e.g. 15 degrees), as follows.

$$p(\mathbf{V}_r | l_r) = N(r ; \theta_n^{\mathrm{V}}, \sigma^2) \quad (8)$$

An example is shown in Fig. 5(c). The participant index n that minimizes the distance between r and θ_n^{V} is selected. It should be noted that the probability is set at $1/R$ when no face is detected in the visual observation. Although this study only utilizes the face tracking result of STCTracker, it should also be noted that these probabilities can be obtained with the particles used in STCTracker. This will constitute future work.

3.3 Speaker clustering

Based on equations (5), (7), and (8), the posterior probability $p(a_r, l_r | \mathbf{X}_r, \mathbf{V}_r)$ can be obtained for each r. An example is shown in Fig. 5 (d). To determine the speech segment, the presence of speech in region r is first determined by thresholding the posterior probability. Then, speech-present regions in which speech is present are classified into speaker clusters using θ_n^{V} as the cluster centroids. Because STCTracker can track identical faces precisely, the speech-present regions are classified as the nearest centroids. Finally, whether participant n speaks or not is determined by thresholding the numbers of the n-th cluster members for each analysis frame.

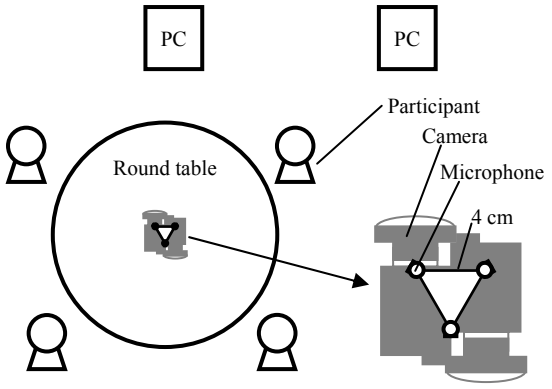

Figure 6. Experimental setup. Participants were all seated, and moved little during conversations.

4. EXPERIMENT

4.1 Recording setup

Experiments were performed in a room with a reverberation time of around 350 ms. Figure 6 shows the recording setup. Two casual conversations engaged in by four participants were recorded in the room. The recording system was placed in the center of a round table, and the participants sat around the table. Each item of recorded data was ten minutes long. The participants were about 1 m from the microphone array. Table 1 shows the properties of the recorded conversations. The recorded casual conversations contained more speaker turn-taking, speaker overlaps, and laughs than usual group meeting recordings, which made speaker diarization more difficult.

4.2 Evaluation and discussion

Speaker diarization experiments were performed with the recorded conversation data described in Section 4.1. The observed signals were analyzed with a frame that was 32 ms long with 16 ms overlaps. The threshold for the audio speaker clustering was 15 degrees. STCTracker ran at around 24 fps, and the face

Table 1. Properties of recorded multi-party conversations.

	Speaker ID	# utterances	Speech time (sec)	Overlap speech time (sec)
Conversation 1	1	170	200.7	171.1
	2	104	170.4	
	3	88	140.2	
	4	133	195.7	
Conversation 2	1	107	155.9	157.0
	2	113	195.6	
	3	60	102.0	
	4	175	237.2	

Table 2: Experimental results obtained with our proposed and previous methods.

	Measures			
	DER	MST	FST	SET
Audio-visual (Proposed)	35.4	18.1	14.4	2.9
Audio-visual (Cascade)	40.4	23.2	13.8	3.4
Audio only	40.4	22.7	14.5	3.3

(%)

locations were updated for each observation. The processing times of the visual and audio information were aligned by synchronizing the times of the two PCs. Reference diarization labels were generated by employing hand-labeled transcription, which included the speech onsets and speech offsets of each speaker.

The Diarization Error Rate (DER), which is used for NIST Rich Transcription Meeting Recognition [3], was employed as a performance measure for this evaluation. The DER can completely account for False-alarm Speech Time (FST), Missed Speech Time (MST), and Speaker Error Time (SET). The DER is calculated as follows.

$$DER = \frac{FST + MST + SET}{Total\ length\ of\ data} \times 100 \quad (\%) \tag{9}$$

The evaluation criteria also follow those provided by NIST [7], that is, the speech segments were split with non-speech periods of more than 300 ms in length, and vocalizations such as laughing and coughing were considered to be non-speech. The allowed tolerance for the difference between the system outputs and the correct labels was 250 ms. The correct speech onset and offset labels for the recorded data were generated manually.

To evaluate the effect of the proposed fusion of the audio and visual information, we compared the results obtained with the proposed method and another speaker diarization method that employs only audio information [41]. This method works based on $p(a_r,l_r|\mathbf{X}_r)$ obtained with a probabilistic integration of SAD results [25] and DOA estimations [26]. The estimation can be interpreted as follows.

$$p(a_r, l_r|\mathbf{X}_r) \approx p(a_r|\mathbf{X}_r)p(l_r|\mathbf{X}_r) \tag{10}$$

The first term is estimated by SAD, and the second term is estimated by DOA estimations. To compare the effectiveness of the estimations of probability $p(a_r,l_r|\mathbf{X}_r)$ in this study, we also compared the proposed method with a method in which $p(a_r,l_r|\mathbf{X}_r)$ obtained by equation (10) [41] is integrated with equation (8).

Table 2 shows the experimental results, which are the average values for two recorded conversations. The result obtained with an integration of SAD results and DOA estimations is shown as 'Audio only'. The result obtained with simple combinations of this approach and visual probabilities is shown as 'Audio-visual (cascade)'. The result obtained with the proposed method is shown as 'Audio-visual (Proposed)'. The reason for the high MST is that the lengths of most utterances were very short, thus it was difficult for the speaker diarization methods to detect them accurately. The reason for the high FST is that the speaker diarization methods sometimes detect simultaneous speech when one speaker spoke. Both phenomena resulted with the fact that the test data were casual conversations as mentioned above. The results confirmed that the proposed integration of the audio and visual information could improve the speaker diarization performance by 5.0 %. The improvements in DER and MST are mainly due to the fact that the proposed method could detect speech onset more robustly, because face tracking works all the time even if there is no acoustic event, i.e. the participants do not speak. Robust onset detection is particularly important when analyzing group conversations. On the other hand, the simple integration of SAD results, DOA estimations, and visual observations cannot improve performance. Since a GCC-PHAT based method cannot handle overlapping speech for one analysis frame, the conventional method could not confident the speech activity when there are multiple speakers because the DOA of the speech signal is not stable. Because it sometimes failed to detect speech onset correctly, the MST became high. The experimental result suggests that the estimation of speech presence probability in this study perform better than the combination of SAD results and DOA estimations.

5. CONCLUSION

This study proposed a probabilistic approach to audio and visual information integration designed to improve speaker diarization performance in a compact system for analyzing group conversations. Experiments employing casual conversations, which constitute a difficult task for speaker diarization, confirmed that the proposed method improved the performance by using visual information in addition to audio information. Although the experiment described in this paper used data in which the participants did not move greatly, the proposed method can be applied to data in which participants move more because STCTracker can robustly track the participants' faces [42], and the SAD of the proposed method is independent of the participants' locations. Such an evaluation will be future work. Future work will also include finding a way to integrate audio and visual information more effectively to detect both "who spoke when" and to analyze aspects of the participants' behavior such as backchannels, dialog acts, and state changes. The automatic speaker identification / speech recognition of recorded speech will also be studied.

6. REFERENCES

[1] Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaj, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., and Wellner, P., 2006. The AMI Meeting Corpus: A pre-announcement. Machine Learning for Multimodal Interaction, Renals, S. and Bengio, S. (Eds.), LNCS 3869, Springer-Verlag, 28–39.

[2] Mostefa, D., Moreau, N., Chuoukri, K., Potamianos, G., Chu, S. M., Tyagi, A., Casas, J. R., Turmo, J., Cristoforetti, L., Tobia, F., Pnevmatikakis, A., Mylonakis, V., Tlantzis, F., Burger, S., Stiefelhagen, R., Bernardin, K., and Rochet, C., 2007. The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms, J. Language Resources and Evaluation 41, 389–407.

[3] Fiscus, J. G., Ajot, J., and Garofolo, J. S., 2008. The Rich Transcription 2007 meeting recognition evaluation. Multimodal Technologies for Perception of Humans,

Stifelhagen, R., Bowers, R., and Fiscus, J. (Eds.), LNCS 4625, 373–389.

[4] Waibel, A., Bett, M., Metze, F., Ries, K., Schaaf, T., Schultz, T., Soltau, H., Yu, H., and Zechner, K., 2001. Advances in automatic meeting record creation and access. Proc. Int. Conf. Acoust., Speech, Signal Process., 597–600.

[5] Cutler, R. and Davis, L., 2002. Distributed meetings: A meeting capture and broadcasting system. Proc. ACM Int. Conf. Multimedia, 503–512.

[6] Morgan, N., Baron, D., Bhagat, S., Carvey, H., Dhillon, R., Edwards, J., Gelbart, D., Janin, A., Krupski, A., Peskin, B., Pfau, T., Shriberg, E., Stolke, A., and Wooters, C., 2003. Meetings about meetings: Research at ICSI on speech multiparty conversations. Proc. Int. Conf. Acoust., Speech, Signal Process. 4, 740–743.

[7] Yu, Z., Ozeki, M., Fujii, Y., and Nakamura, Y., 2007. Towards smart meeting: Enabling technologies and a real-world application. Proc. ACM Int. Conf. Multimodal Interfaces, 86–93.

[8] Tranter, S. E. and Reynolds, D. A., 2006. An overview of automatic speaker diarization systems. IEEE Trans. Audio, Speech, Language Process. 14, 1557–1565.

[9] Wooters, C. and Huijbregts, M., 2008. The ICSI RT07s speaker diarization system. Multimodal Technologies for Perception of Humans, Stiefelhagen, R., Bowers, R., and Fiscus, J. (Eds.), LNCS 4625, 509–519.

[10] van Leeuwen, D. A. and Konečný, M., 2008. Progress in the AMIDA speaker diarization system for meeting data. Multimodal Technologies for Perception of Humans, Stiefelhagen, R., Bowers, R., and Fiscus, J. (Eds.), LNCS 4625, 475–483.

[11] Huang, J., Marcheret, E., Visweswariah, K., and Potamianos, G., 2008. The IBM RT07 evaluation systems for speaker diarization on lecture meetings. Multimodal Technologies for Perception of Humans, Stiefelhagen, R., Bowers, R., and Fiscus, J. (Eds.), LNCS 4625, 497–508.

[12] Luque, J., Anguera, X., Temko, A., and Hernando, J., 2008. Speaker diarization for conference room: The UPC RT07s evaluation system. Multimodal Technologies for Perception of Humans, Stiefelhagen, R., Bowers, R., and Fiscus, J. (Eds.), LNCS 4625, 543–553.

[13] Zhu, X., Barras, C., Lamel, L., and Gauvain, J.-L., 2006. Speaker diarization: From broadcast news to lectures. Machine Learning for Multimodal Interaction, Renals, S., Bengio, S., and Fiscus, J. (Eds.), LNCS 4299, 396–406.

[14] Fredouille, C. and Evans, N., 2008. The LIA RT'07 Speaker Diarization System. Multimodal Technologies for Perception of Humans, Stiefelhagen, R., Bowers, R., and Fiscus, J. (Eds.), LNCS 4625, 520–532.

[15] Chu, S. M., Marcheret, E., and Potamianos, G., 2006. Automatic speech recognition and speech activity detection in the CHIL smart room. Machine Learning for Multimodal Interaction, Renals, S., Bengio, S., and Fiscus, J. (Eds.), LNCS 4299, 332–343.

[16] Martin, A., Charlet, D., and Mauuary, L., 2001. Robust speech/non-speech detection using LDA applied to MFCC. Proc. Int. Conf. Acoust., Speech, Signal Process., 237–240.

[17] Padrell, J., Macho, D., and Nadeu, C., 2005. Robust speech activity detection using LDA applied to FF parameters. Proc. Int. Conf. Acoust., Speech, Signal Process. 1, 557–560.

[18] Armani, L., Matassoni, M., Omologo, M., and Svaizer, P., 2003. Use of a CSP-based voice activity detector for distant-talking ASR. Proc. INTERSPEECH, 501–504.

[19] Omologo, M., Svaizer, P., Brutti, A., and Cristoforetti, L., 2006. Speaker localization in CHIL lectures: Evaluation criteria and results. Machine Learning for Multimodal Interaction, Renals, S. and Bengio, S. (Eds.), LNCS 3869, 476–487.

[20] Anguera, X., Wooters, C., and Hernando, J., 2007. Acoustic beamforming for speaker diarization of meetings. IEEE Trans. Audio, Speech, Language Process. 15, 2011–2022.

[21] Pardo, J. M., Anguera, X., and Wooters, C., 2006a. Speaker diarization for multi-microphone meetings using only between-channel differences. Machine Learning for Multimodal Interaction, Renals, S. and Bengio, S. (Eds.), LNCS 3869, Springer-Verlag, 257–264.

[22] Pardo, J. M., Anguera, X., and Wooters, C., 2006b. Speaker diarization for multiple distant microphone meetings: Mixing acoustic features and inter-channel time differences. Proc. INTERSPEECH, 2194–2197.

[23] Knapp, C. H. and Carter, G. C., 1976. The generalized correlation method for estimation of time delay. IEEE Trans. Acoust., Speech, Signal Process. ASSP-24, 320–327.

[24] Otsuka, K., Araki, S., Ishizuka, K., Fujimoto, M., Heinrich, M., and Yamato, J., 2008. A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization. Proc. ACM Int. Conf. Multimodal Interfaces, 257–262.

[25] Fujimoto, M., Ishizuka, K., and Nakatani, T., 2008. A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme. Proc. Int. Conf. Acoust., Speech, Signal Process., 4441–4444.

[26] Araki, S., Sawada, H., Mukai, R., and Makino, S., 2006. DOA estimation for multiple sparse sources with normalized observation vector clustering. Proc. Int. Conf. Acoust., Speech, Signal Process. 5, 33–36.

[27] Mateo Lozano, O. and Otsuka, K., in press. Simultaneous and fast 3D tracking of multiple faces in video sequences by using a particle filter. J. Signal Process. Systems, DOI 10.1007/s11265-008-0250-2.

[28] Khalidov, V., Forbes, F., Hansard, M., Arnaud, E., and Horaud, R., 2008. Audio-visual clustering for 3D speaker localization. Machine Learning for Multimodal Interaction, Popescu-Belis, A. and Stiefelhagen, R. (Eds.), LNCS 5237, 86–97.

[29] Gatica-Perez, D., Lathoud, G., Odobez, J. M., and McCowan, I., 2007. Audiovisual probabilistic tracking of multiple speakers in meetings. IEEE Trans. Audio, Speech, Language Process. 15, 601–616.

[30] McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M., and Zhang, D., 2005. Automatic analysis of multimodal group actions in meetings. IEEE Trans. Pattern Anal. Mach. Intell. 27, 305–317.

[31] Asano, F., Yamamoto, K., Hara, I., Ogata, J., Yoshimura, T., Motomura, Y., Ichimura, N., and Asoh, H., 2004. Detection and separation of speech event using audio and video information fusion and its application to robust speech interface. EURASIP J. Applied Signal Process. 11, 1727–1738.

[32] Ba, S. O. and Odobez, J. M., 2008. Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues. Proc. Int. Conf. Acoust., Speech, Signal Process., 2221–2224.

[33] Busso, C., Georgiou, P. G., and Narayanan, S. S., 2007. Real-time monitoring of participants' interaction in a meeting using audio-visual sensors. Proc. Int. Conf. Acoust., Speech, Signal Process. 2, 685–688.

[34] Potamianos, G., Huang, J., Marcheret, E., Libal, V., Balchandran, R., Epstein, M., Seredi, L., Labsky, M., Ures, L., Black, M., and Lucey, P., 2008. Far-field multimodal speech processing and conversational interaction in smart spaces. Proc. Joint Workshop Hands-free Speech Commun. Microphone Arrays, 119–123.

[35] Ishizuka, K., Araki, S. and Kawahara, T., 2008. Statistical speech activity detection based on spatial power distribution for analyses of poster presentations," Proc. INTERSPEECH, 99–102.

[36] Friedland, G., Hung, H., and Yeo, C., 2009. Multi-modal speaker diarization of real-world meetings using compressed-domain video features. Proc. Int. Conf. Acoust., Speech, Signal Process., 4069–4072.

[37] Sohn, J., Kim, N.-S., and Sung, W., 1999. A statistical model-based voice activity detection. IEEE Signal Process. Letters 6, 1–3.

[38] Potamitis, I. and Fishler, E., 2004. Speech activity detection and enhancement of a moving speaker based on the wideband generalized likelihood ratio and microphone arrays. J. Acoust. Soc. Am. 116, 2406–2415.

[39] Yılmaz, Ö and Rickard, S., 2004. Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Signal Process. 52, 1830-1847.

[40] Ephraim, Y. and Malah, D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust., Speech, Signal Process. ASSP-32, 1109–1121.

[41] Araki, S., Fujimoto, M., Ishizuka, K., Sawada, H., and Makino, S., 2008. A DOA based speaker diarization system for real meetings. Proc. Joint Workshop Hands-free Speech Commun. Microphone Arrays, 29–32.

[42] Mikami, D., Otsuka, K., and Yamato, J., 2009. Memory-based particle filter for face pose tracking robust under complex dynamics. Proc. IEEE Conf. Computer Vision Pattern Recognition, 999–1006.