Speaker Diarization based on Audio-Visual Integration for Smart Posterboard

Yukoh Wakabayashi*, Koji Inoue*, Hiromasa Yoshimoto*, and Tatsuya Kawahara*
*Academic Center for Computing and Media Studies, Kyoto University

***Abstract*—We present a speaker diarization method based on an audio-visual integration approach. We deal with poster conversations which are more challenging than general meetings, because participants are moving freely and the audience utter infrequently. In this case, it is difficult to detect "who spoke when" by only using acoustic information. Therefore we incorporate visual information to improve diarization accuracy. We propose two integration methods: rule-based and stochastic method. Experiments in real poster conversations show that the integration methods significantly outperform the baseline method which uses acoustic information only.**

I. Introduction

In recent years, analysis of multi-party conversations such as meetings and discussions has been studied [1], [2], [3]. Speaker diarization, i.e. detecting who spoke when, plays an important role in not only analysis of such conversations but also speech enhancement and blind source separation. We are conducting a project called "smart posterboard" [4] focusing on poster sessions in which one participant (=presenter) makes a presentation and the others (=audience) ask questions about and comment on the presentation. Speaker diarization is useful for reviewing the audience's feedbacks and detecting their interest level.

In poster conversations, utterances of the audience are infrequent and they move freely in contrast to general meetings. This infrequency makes it difficult to generate a separation filter like independent component analysis [6] and to realize speaker diarization based on blind source separation. The participant's move makes it difficult to conduct sound source localization and tracking by using acoustic information only. Moreover, in poster conversations, ambient noise such as diffuse noise degrades the performance of diarization.

In this paper, we investigate incorporation of visual information for improvement of speaker diarization. Speaker tracking based on visual information is reliable even when people are moving and robust under ambient noise. Specifically, we propose two diarization methods based on audio-visual integration. One is a rule-based method and the other is a stochastic method

The remainder of this paper is organized as follows. Section Ⅱ reviews the baseline MUSIC method for speaker diarization. We describe two integration methods of audio and visual information in Section Ⅲ, and evaluation of these methods in Section Ⅳ . Finally, conclusions are drawn in Section V.

II. Speaker Diarization Based On Acoustic Information

Conventional speaker diarization methods are composed of feature extraction, voice activity detection (VAD) and speaker clustering steps. In addition, when a microphone array is used, spatial information such as Time Difference Of Arrival (TDOA) and Direction Of Arrival (DOA) of speech are also utilized for diarization. The Generalized Cross Correlation with Phase transform (GCC-Phat) method [5] is often used to estimate DOA in previous works. This method can detect only one direction in a time-frame and cannot be applied to the case when multiple participants utter simultaneously. This overlapping occurs when discussion is lively and thus is important in analyzing conversations.

A. MUSIC Method

In this study, we adopt MUltiple SIgnal Classification (MUSIC) [7], which is a well-known DOA estimation method using a microphone array and can detect simultaneous utterances. The MUSIC method estimates DOA based on the orthogonality of the signal subspace. MUSIC spectrum $P_{MU}(\theta)$ is given by

$$P_{MU}(\theta) = \frac{\|\boldsymbol{a}^H(\theta)\boldsymbol{a}(\theta)\|}{\sum_{i=N+1}^{M} \|\boldsymbol{a}^H(\theta)\boldsymbol{e}_i\|}, \tag{1}$$

where $\boldsymbol{a}^H$ denotes the conjugate transpose of vector $\boldsymbol{a}$, N and M are the number of sound sources and microphones respectively, $\boldsymbol{e}_i$ $(i = 1, \cdots, M)$ is an eigen vector of a spatial correlation matrix $\mathbf{R}_{\boldsymbol{x}}$ of an observed signal $\boldsymbol{x}$, and eigen values corresponding to these vectors satisfy the following condition,

$$\lambda_1 \geq \cdots \geq \lambda_M. \tag{2}$$

The spatial correlation matrix $\mathbf{R}_{\boldsymbol{x}} \in \mathbb{C}^{M\times M}$ at frame t is estimated as

$$\mathbf{R}_{\boldsymbol{x}} = \frac{1}{2\triangle+1}\sum_{j=t-\triangle}^{t+\triangle} \boldsymbol{x}_j\boldsymbol{x}_j^H, \tag{3}$$

where $\boldsymbol{x}_j \in \mathbb{C}^M$ is an observed signal vector at frame j, and $\triangle$ is the number of averaging frames.

The steering vector of direction θ is defined as

$$\boldsymbol{a}(\theta) = [\exp(-j\omega\tau_1), \cdots, \exp(-j\omega\tau_M)]^T, \tag{4}$$

978-616-361-823-8 © 2014 APSIPA

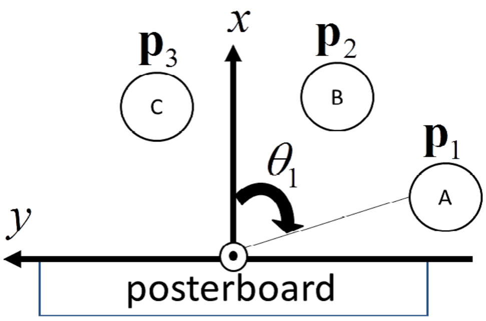

Fig. 1. Recording setting of poster conversations

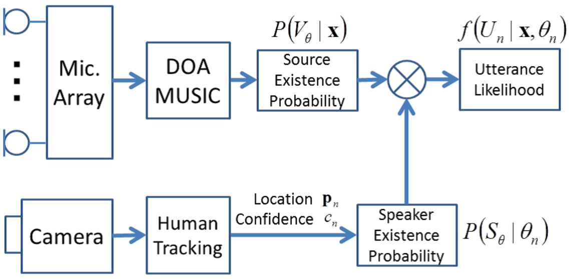

Fig. 2. Overview of the proposed method

where ω is a frequency index, and τ_i $(i = 1, \cdots, M)$ is a relative delay in the propagation of signals from direction θ between the reference point and the i-th microphone. When a DOA vector is represented as $\boldsymbol{u} = [-\cos\theta, \sin\theta, 0]^T$, τ_i is calculated as

$$\tau_i = \boldsymbol{u}^T \boldsymbol{d}_i / v_c, \tag{5}$$

where $\boldsymbol{d}_i$ is a position coordinate of the i-th microphone, v_c is the propagation velocity of the speech signals.

The MUSIC spectra $P_{MU}(\theta)$ has a peak at $\theta = \theta'$ when an utterance occurs in direction θ'. In this paper, $\theta \in [-90, 90]$ is defined as a discrete value by 1° in the horizontal angle.

B. Baseline Method by Using MUSIC Spectrum

We introduce speaker diarization based on acoustic information for a baseline method which conducts DOA estimation and clustering [3]. In this paper, we use peaks of the MUSIC spectrum as DOA estimates and Gaussian Mixture Model (GMM) for clustering. The baseline method tracks peaks of the MUSIC spectrum which are above a threshold. Then, GMM-based clustering whose mixture size is defined as the number of participants in the angle domain is conducted. Each cluster corresponds to each participant and the peak location θ_l $(l = 1, \cdots, L(t))$ which belongs to the n-th GMM cluster represents the n-th participant's voice activity, where $L(t)$ is the number of peaks in the MUSIC spectrum at time-frame t.

III. Audio-Visual Integration

In this section, we propose two integration methods of audio and visual information: rule-based method and stochastic method. Human's lip motion is often used for voice activity detection, especially in the field of human-robot dialogue [8], [9]. However, it requires a front image in a good resolution. The assumption does not always hold in poster conversations, thus we cannot use it for our task.

In this study, we use participants' head location estimated with a computer vision technique. Audio-visual integration is operated in the horizontal angle domain shown in Fig. 1, that is, when a location coordinate of the n-th participant's head $\boldsymbol{p}_n = [x_n, y_n, z_n]$ is given, the angle location θ_n is defined as

$$\theta_n = \tan^{-1}(\frac{y_n}{x_n}).$$

In this paper, tracking head locations is conducted by model fitting and particle filtering [10]. This method enables to estimate head locations without a front image.

A. Rule-Based Integration

First, we propose a rule-based integration method. This method conducts peak tracking and thresholding of the MUSIC spectrum just like the baseline method, but selects peaks with a constraint

$$\theta_n - \epsilon < \theta_l < \theta_n + \epsilon \ (l = 1, \cdots, L(t)), \tag{6}$$

where $L(t)$ is the number of searched and thresholded peaks at time-frame t, θ_n is the n-th participant's location computed via visual information, θ_l is the peak location of the MUSIC spectrum, and ϵ is a tolerable error between θ_l and θ_n. The integration replaces GMM-clustering of the baseline method.

B. Stochastic Integration

The second method is formulated in a stochastic manner. The flow of the system is shown in Fig. 2. We assume that a source existence probability $P(V_\theta|\boldsymbol{x})$ is proportional to the MUSIC spectrum $P_{MU}(\theta)$, given observed acoustic signals $\boldsymbol{x}$, where V_θ is a stochastic variable which represents existence of a source. Thus, $P(V_\theta|\boldsymbol{x})$ is calculated as

$$P(V_\theta|\boldsymbol{x}) \equiv \frac{1}{\sum_\phi P_{MU}(\phi)} P_{MU}(\theta). \tag{7}$$

In addition, a speaker existence probability $P(S_\theta|\theta_n)$ at direction θ, given the n-th participant's location θ_n, is assumed as

$$\begin{aligned} P(S_\theta|\theta_n) &\equiv N(\theta_n, \sigma(c_n)), && (8)\\ \sigma(c_n) &\equiv \alpha\{\beta_1 c_n + \beta_2\}. && (9)\end{aligned}$$

where S_θ is a stochastic variable which represents existence of a speaker, $N(\mu, \sigma)$ represents a normal distribution with mean μCvariance σ, and $c_n \in [0, 1]$ is the confidence score of the n-th participant's head location [10]. Here $\sigma(c_n)$ is assumed to be proportional to c_n, and α, β_1 and β_2 are parameters.

By using (7) and (8), an utterance likelihood of the n-th participant $f(U_n|\boldsymbol{x}, \theta_n)$ is defined as

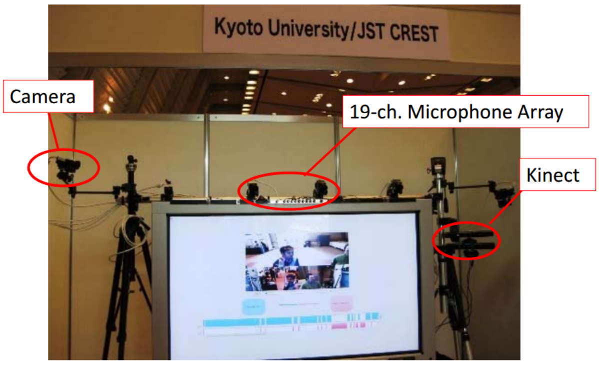

Fig. 3. Smart Posterboard

$$f(U_n|\boldsymbol{x},\theta_n) \equiv \sum_{\theta=\theta_n-\varphi}^{\theta_n+\varphi} P(V_\theta|\boldsymbol{x})P(S_\theta|\theta_n), \qquad (10)$$

where U_n is a stochastic variable which represents the n-th participant's utterance. With this operation, the spatial spectrum is masked by the human location information and used for speaker diarization. After a thresholding process on $f(U_n|\boldsymbol{x},\theta_n)$, the diarization results are smoothed over adjacent time-frames to make final detection.

IV. Experimental Evaluation

We conducted experiments to evaluate the audio-visual integration methods compared with the baseline method. We used 12 poster sessions recorded with the smart posterboard, which consists of a 19-channel microphone array, kinect sensors, and HD cameras at the top or the side of a large LCD as shown in Fig. 3. Participants of each session are one presenter and two persons in the audience. The duration of each session is about 30 min. and the total duration of presenter's and audience's utterances per session is 20-25 min. and 2-6 min. respectively as shown in Table I.

A. Experimental Condition

To evaluate the performance of speaker diarization, we employ a Recievwe Operating Characteristic (ROC) curve and the Diarization Error Rate (DER) [11] .

The ROC curve plots False Acceptance Rate (FAR) and False Rejection Rate (FRR). The closer the curve is to the origin, the better the performance of the method is. FAR and FRR are defined as

$$\mathrm{FAR} = \frac{\text{\# non-speech frames detected as speech frames}}{\text{\# non-speech frames}},$$
$$\mathrm{FRR} = \frac{\text{\# speech frames detected as non-speech frames}}{\text{\# speech frames}},$$

where the ground truth data i.e. the time stamp of the utterances is annotated manually. The ROC curve is plotted for the presenter and the audience separately because the performance is much different between them.

TABLE I
TOTAL DURATION OF POSTER SESSIONS [SEC]

ID	presenter	audience		total
121218-01	1,413	207	196	1,816
121218-02	1,268	122	126	1,516
121218-03	1,184	345	385	1,914
121218-04	1,299	400	328	2,027
140206-01	1,261	19	230	1,510
140206-02	1,417	285	166	1,868
140206-03	1,344	331	172	1,847
140206-04	1,507	131	104	1,742
140207-01	1,354	166	125	1,645
140207-02	1,239	135	119	1,493
140207-03	1,215	107	270	1,592
140207-04	1,218	218	137	1,573
total	15,719	4,824		20,543

TABLE II
Diarization Error Rate (DER)

Method	clean data	SNR = 10 dB
baseline	19.86 %	21.52 %
rule-based integration	14.60 %	17.91 %
stochastic integration	7.43 %	15.69 %

DER includes both false acceptance and false rejection errors and is defined as

$$\mathrm{DER} = \frac{\text{\# incorrectly labeled frames}}{\text{\# entire frame}}.$$

In this paper, according to [11] the tolerance of 250 ms from the ground truth is allowed.

In this experiment, the sampling rate of speech was 16 kHz, the analysis frame size was 32 ms, and the frame shift was 16 ms. The parameter $\triangle$ in Eq. (3) is 2 and ϵ in Eq. (6) is 10. The Parameters α, β_1, β_2 and φ in Eq. (9), (10) were 1/3, -5, 5, and $3\sigma(c_n)$, respectively. The values of these parameters were experimentally determined.

B. Results and Discussions

Figure 4 and 5 show the ROC curve for the presenter and the audience respectively. As shown in these figures, FAR and FRR are significantly improved by the two integration methods compared with the baseline method. These results suggest that visual information contributes improvement of diarization accuracy. In Fig. 4, there is little difference between the three methods, because the majority of utterances are made by the presenter and it is easy to detect them. On the other hand, in Fig. 5 the results of the proposed methods, especially the stochastic method, is much superior to the baseline method. It is because that GMM-clustering in the baseline method does not work well when the participants move i.e. DOA estimation scatters. The integration methods are effective in tracking the audience.

The diarization results under diffuse noise (SNR = 10 dB) are shown in Fig. 6 and 7. The accuracy of diarization is degraded from the case of the clean data, especially for the

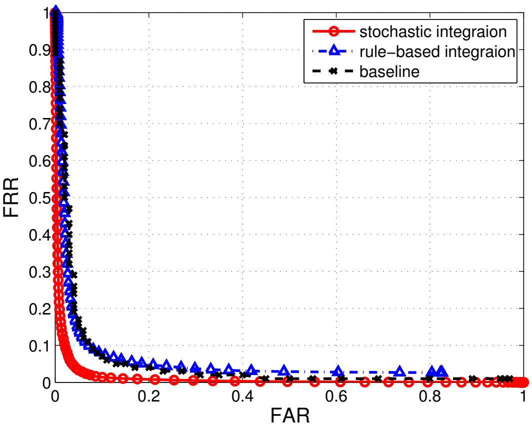

Fig. 4. ROC curve for presenter (clean data)

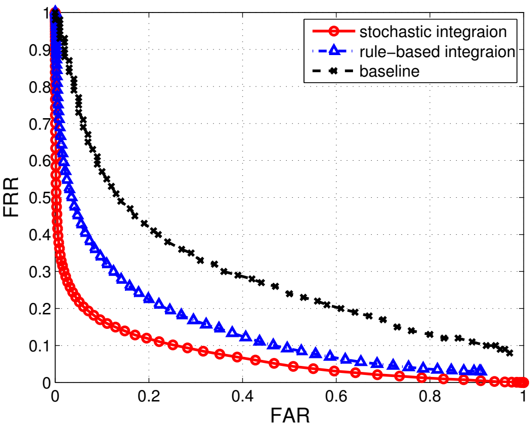

Fig. 5. ROC curve for audience (clean data)

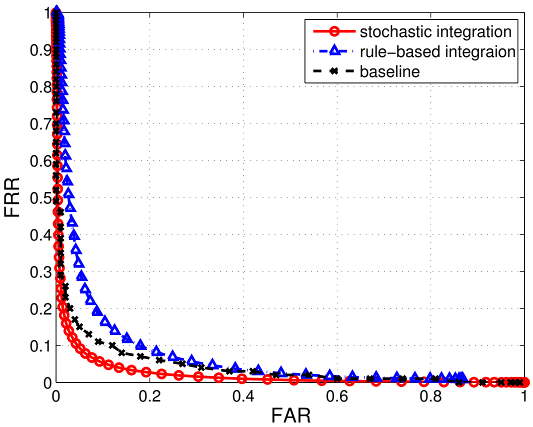

Fig. 6. ROC curve for presenter (SNR = 10 dB)

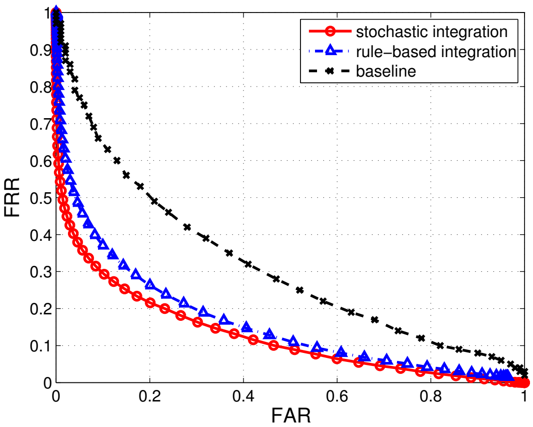

Fig. 7. ROC curve for audience (SNR = 10 dB)

audience. The degradation is caused by the change of the shape in the MUSIC spectrum such that pseudo-peaks occur and the shape becomes dull.

Table II shows DER in these two conditions. Under these two conditions, the performance of the stochastic integration method is the best in the three methods as observed in the ROC curve.

V. Conclusion

We have proposed speaker diarization methods based on the audio-visual integration approach. With 12 real poster sessions, we showed that these methods are effective compared with the baseline method. Especially, the stochastic integration method achieved significant improvement of accuracy.

VI. Acknowledgements

This work was supported by JST CREST and JSPS Grant-in-Aid for Scientic Research.

References

[1] D. A. V. Leeuwen and M. Huijbregts, “The AMI speaker diarization system for NIST RT06 meeting data,” in *Proc. NIST Rich Transcription 2006 Spring Meeting Recognition Evaluation Workshop,* Washington DC, pp.371-384, 2006.

[2] G. Friedland and A. Janin, D. Imseng, X. Anguera, L. Gottlieb, M. Huijbregts, M. T. Knox, O. Vinyals “The ICSI RT-09 Speaker Diarization System,” in *Transactions on Audio, Speech and Language Processing*, July 2011.

[3] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, “A DOA based speaker diarization system for real meeting,” in *Proc. HSCMA,* pp. 29-32. , 2008

[4] T. Kawahara, “Smart posterboard: Multi-modal sensing and analysis of poster conversation,” in *Proc. APSIPA ASC* (plenary overview) , 2013

[5] C. H. Knapp and G. C. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Trans. Acoust. Speech and Signal Processing,* vol. 24, no. 4, pp. 320-327, 1976.

[6] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, “A survey of convolutive blind source separation methods,” *MUltichannel Speech Processing Handbook,* pp. 1065-1084, 2007.

[7] R.O.Schmidt, “Multiple emitter location and signal parameter estimation,” *IEEE Trans. on Antennas and Propagation* vol. 34, no. 3, pp. 276-280, 1986.

[8] T. Yoshida and K. Nakadai, “Audio-visual voice activity detection based on an utterance state transition model,” *Advances Robotics,* Vol. 26, no. 10, pp. 1183-1201, 2012.

[9] H. D. Kim, J. S. Choi, and M. Kim, “Human-Robot Interaction in Real Environments by Audio-Visual Integration,” *International Journal of Control, Automatic and Systems,* vol. 5, no. 1, pp. 61-69, Feb. 2007.

[10] H. Yoshimoto and Y. Nakamura, “Cubistic Representation for Real-time 3D Shape and Pose Estimation of Unknown Rigid Object,” M. Young, *The IEEE ICCV Workshops*, Dec. 2013.

[11] J. G. Fiscus, J. Ajot, and J. S. Garofolo, “The Rich Transcription 2007 Meeting Recognition Evaluation,” http:www.nist.gov/speech/tests/rt/2007/workshop/RT07-SPKR-v7.pdf, *Lecture Notes in Computer Science,* vol. 4625, pp. 373-389, 2008.