Speaker Diarization using Eye-gaze Information in Multi-party Conversations

Koji Inoue[1]*, Yukoh Wakabayashi*[2]*, Hiromasa Yoshimoto*[2]*, and Tatsuya Kawahara*[1,2]

[1]School of Informatics, Kyoto University
[2]Academic Center for Computing and Media Studies, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan

Abstract

We present a novel speaker diarization method by using eye-gaze information in multi-party conversations. In real environments, speaker diarization or speech activity detection of each participant of the conversation is challenging because of distant talking and ambient noise. In contrast, eye-gaze information is robust against acoustic degradation, and it is presumed that eye-gaze behavior plays an important role in turn-taking and thus in predicting utterances. The proposed method stochastically integrates eye-gaze information with acoustic information for speaker diarization. Specifically, three models are investigated for multi-modal integration in this paper. Experimental evaluations in real poster sessions demonstrate that the proposed method improves accuracy of speaker diarization from the baseline acoustic method.

Index Terms: speaker diarization, multi-modal interaction, eye-gaze

10.21437/Interspeech.2014-137

1. Introduction

Analysis of multi-party interaction such as meetings and conversations has been studied in recent years [1, 2]. In multi-party conversations, participants convey various information via not only verbal communications but also nonverbal channels. Nonverbal channels include back-channels, nodding, and eye-gaze behavior. Taking account of these factors makes conversation structures complicated, but these behaviors of participants provide an important cue in analyzing conversations.

We have been conducting a project focusing on conversations in poster sessions (= poster conversations) in which a presenter makes an interactive presentation to a small audience. This conversation form is the norm in academic conventions including InterSpeech conferences. To analyze poster conversations, we have designed a multi-modal recording environment equipped with a microphone array and cameras, called smart posterboard [3]. We have investigated turn-taking behavior [4] and also interest and comprehension levels of audience [5] in poster conversations by combining multi-modal information.

This work addresses speaker diarization, that is to identify "who spoke when" in multi-party conversations. Until now, a number of methods have been investigated [6, 7], but they are mainly based on acoustic information input from single or multiple microphones. However, the performance of speaker diarization is drastically degraded by distant talking and ambient noise in real environments. In addition, participants of poster conversations do not sit still unlike meetings, which means that the participants can make utterances as they move. This makes it difficult to localize the speakers. In natural conversations, moreover, utterances are spontaneous, sometimes ambiguous or overlapping with others. Especially in poster conversations, utterances of the audience are fewer in frequency, which means that it is difficult to constitute separation filters like independent component analysis [8] while the audience's utterances are more important and should not be missed.

In this paper, we propose a novel approach to speaker diarization which integrates eye-gaze information with acoustic information. Eye-gaze behavior plays an important role of turn-taking in multi-party conversations. For example, it is often observed that a speaker ends utterance by looking at an audience and a person looks back as he takes a turn to speak, by which the speaker and the listener exchange their roles [9]. Thus, it is presumed that eye-gaze behavior is related with utterance prediction [5, 10], but the effect of eye-gaze information on speaker diarization has not been seriously investigated. Since eye-gaze information is free from the above-mentioned adversary acoustic condition, it is expected to complement acoustic processing.

The rest of this paper is organized as follows. Section 2 introduces the corpus of poster conversations and its annotation. Section 3 gives a baseline method and features of the acoustic and eye-gaze information. Section 4 presents three models for multi-modal integration of the acoustic and eye-gaze features. Experimental evaluations of the proposed method are presented in Section 5. Section 6 concludes this paper with discussions of the proposed method.

2. Multi-modal corpus of poster conversations

The smart posterboard system we are developing consists of a 19-channel microphone array, Kinect sensors, and HD cameras, which are attached to the top or the side of a large LCD [3]. With this setting, four poster sessions were recorded, in which the presenters and audiences are different from each other. In each session, one presenter made a poster presentation on his/her academic research, and there was an audience of two persons, standing in front of the poster and listening to the presentation. The duration of each session is 20 to 30 minutes. Speaker diarization is to be conducted with the sensors (the microphone array and Kinect) attached to the posterboard, so the participants do not have to wear any devices.

For the ground truth annotation of the data used in this work, speech data were also recorded with wireless head-set microphones, and eye-gaze information was captured by magnetometric sensors.

Table 1 summarizes the statistics of the utterances in the four sessions. It is observed that the presenter holds a majority of the turns. In contrast, the utterances by the audience are not frequent, which means it would be difficult to detect these utterances accurately.

Table 1: Statistics of utterance duration [sec.]

	presenter	audience		total
session 1	1,343	165	124	1,632
session 2	1,229	134	118	1,481
session 3	1,205	106	268	1,579
session 4	1,208	216	136	1,560
total	4,985	621	646	6,252

3. Baseline acoustic method and eye-gaze features

First, we design acoustic and eye-gaze features for speaker diarization.

3.1. MUSIC spectrum for acoustic features

The speaker diarization method is implemented based on a Direction Of Arrival (DOA) estimator [11]. Here, we adopt the MUltiple SIgnal Classification (MUSIC) method [12]. The MUSIC method is able to detect multiple DOA simultaneously, and it has been applied to sound source localization in real environments [13].

The MUSIC spectrum is given by

$$P_{MU}(\theta) = \frac{||\boldsymbol{t}(\theta)||^2}{\sum_{i=N+1}^{M} |\boldsymbol{t}^H(\theta)\ \boldsymbol{e}_i|^2}. \quad (1)$$

Note that θ is an angle between the microphone array and the target of estimation, $\boldsymbol{t}(\theta)$ is the steering vector, $\boldsymbol{e}_i$ is the eigen vector obtained by the eigen-value decomposition of the correlation matrix of the observation signals, N and M are the number of sound sources and microphones respectively, and $\boldsymbol{t}^H$ denotes the conjugate transpose of the vector $\boldsymbol{t}$. The MUSIC spectrum $P_{MU}(\theta^*)$ has a peak at an angle θ^* where a sound source is located or a participant makes an utterance. The baseline method conducts speaker diarization by peak tracking of the MUSIC spectrum.

We can make use of visual information on the participants' location ($\hat{\theta}$) tracked by the cameras or the magnetometric sensors. The possible location of speakers is constrained within a certain range ($\pm\theta_B$) from the visually-detected location. We parameterize the acoustic features which consist of the MUSIC spectrum in neighboring angles of the detected participant's location. It is represented as

$$\boldsymbol{a} = \left(P_{MU}(\hat{\theta}-\theta_B)\ \cdots\ P_{MU}(\hat{\theta})\ \cdots\ P_{MU}(\hat{\theta}+\theta_B)\right)^T. \quad (2)$$

3.2. Eye-gaze features

Detection of eye-gaze is approximated by detection of head orientation, and it is done by using images (color and time-of-flight) captured by the Kinect sensors. First, the face of the participant is detected and then head model is applied and adjusted. Tracking of the head orientation is done by a particle filter [14]. The object of eye-gaze is determined by the head-orientation vector and the location of the objects. In this study, the eye-gaze object is limited to the poster and other participants.

The eye-gaze features for speaker diarization are designed based on the object of the eye-gaze and joint eye-gaze events by the presenter and the audience. The eye-gaze feature vector $\boldsymbol{g}$ consists of the followings based on the previous work [4] :

1. Eye-gaze object
   This feature represents which object the participant looks at in the time frame. For the presenter, (P) poster or (I) audience; For (anybody in) the audience, (p) poster, (i) presenter, or (o) other person in the audience.
2. Joint eye-gaze event: "Ii", "Ip", "Pi", "Pp"
   Combination of the eye-gaze objects by the presenter and the audience. These events are defined for each person in the audience.
3. Bigram of joint eye-gaze events:
   This feature represents the transition of the joint eye-gaze events.
4. Duration of the above 1. (for (I) and (i))
5. Duration of the above 2. (except for "Pp")

These features are calculated for each time frame and the duration is measured during the preceding period (the preceding C seconds).

4. Combination models of acoustic and eye-gaze information

This section presents three models to combine eye-gaze information with acoustic information. The acoustic and eye-gaze features are calculated for each participant, and regarded as stochastic variables. Then, speaker diarization is conducted based on a probabilistic framework, which predicts utterances of each participant independently.

Let a, g, and v be stochastic variables of the acoustic features, the eye-gaze features, and the event of utterance, respectively. Note that the two variables, a and g, are represented by the feature vectors described in Section 3, and the utterance variable v is binary, speaking ($v = 1$) or not-speaking ($v = 0$).

4.1. Model 1 (Joint discriminative model)

A joint discriminative model is designed to predict the following posterior probability

$$p(v|a, g). \quad (3)$$

In this work, we adopt Logistic Regression (LR) model to directly compute this probability.

4.2. Model 2 (Independent discriminative models)

Independent discriminative models are designed by assuming that the two variables (a and g) are independent. The posterior probabilities computed by the two models are linearly interpolated,

$$\alpha\, p(v|a) + (1-\alpha)\, p(v|g). \quad (4)$$

The parameter α is the weighting coefficient. Each discriminative model is implemented by LR.

4.3. Model 3 (Noisy channel model)

The posterior probability $p(v|a, g)$ can be developed by the Bayes' theorem as

$$p(v|a,g) = \frac{p(a|v,g)\, p(v|g)}{p(a|g)} \quad (5)$$

$$= \frac{p(a|v)\, p(v|g)}{p(a)} \quad (6)$$

when we assume that the two variables (a and g) are independent. The denominator can be ignored for the decision, so we can get a likelihood as

$$l(v|a,g) = p(a|v)\, p(v|g). \tag{7}$$

The generative model to compute $p(a|v)$ is realized by Gaussian Mixture Model (GMM). The model to compute $p(v|g)$ is same as the discriminative model (LR) in Model 2, but in this case, it is regarded as a predictive model just like language model in automatic speech recognition (ASR). Just like ASR, we take the logarithm of the above likelihood, and an additional weight parameter β is introduced to control the difference in their dynamic range,

$$ll(v|a,g) = \log p(a|v) + \beta \log p(v|g). \tag{8}$$

Speech activity detection is based on the difference of log likelihood $ll(v|a,g)$ as

$$ll(v=1|a,g) - ll(v=0|a,g) \begin{cases} \geq \Theta_{ll} & \text{speaking} \\ < \Theta_{ll} & \text{not-speaking} \end{cases} \tag{9}$$

where the parameter Θ_{ll} represents the threshold.

5. Experimental evaluations

We compared the proposed method with the baseline method and other methods by using the corpus of poster conversations.

5.1. Setup

The sampling rate of speech was 16 kHz. The frame size for the MUSIC spectrum was 32 ms, and the frame shift was 16 ms. The block size of MUSIC was 5. In this experiment, the MUSIC spectrum was computed by using the 19-channel audio signals.

The constrained range of the MUSIC spectrum in the acoustic features was 10 degrees ($\theta_B = 10^\circ$). This setting was intended to prevent the ranges overlapping between the participants. The MUSIC spectrum was calculated every 1 degree, thus the dimension of the acoustic features was 21. The scope to calculate the duration in the eye-gaze features was 10 seconds ($C = 10$).

The GMM and LR were trained by cross-validation of the four sessions, while the mixture size of GMM was fixed to 8. The weight coefficients in Model 2 and Model 3 (α and β) were subsequently determined in the cross-validation manner. To evaluate one session, other three sessions were used to train the GMM and LR and to determine the weight coefficients.

We made comparison of the three models described in Section 4 as well as other methods listed below:

1. Baseline MUSIC
   The baseline method conducts peak tracking of the MUSIC spectrum and GMM-based clustering in the angle domain. Each cluster corresponds to each participant. This method does not use any cue from visual information.
2. Baseline + location constraint
   This method also performs peak tracking of the MUSIC spectrum and incorporates constraint on the participants' location obtained by the image processing, which was described in Section 3. The detected peaks are adopted if they are consistent with the estimated location within the θ_B range.

Table 2: 11-point average precision for presenter

Method	SNR			
	clean	20 dB	15 dB	10 dB
Baseline MUSIC	0.887	0.938	0.886	0.883
Baseline + location	0.870	0.870	0.871	0.868
Acoustic-only LR	0.943	0.942	0.942	0.943
Model 1	0.943	0.941	0.942	0.943
Model 2	0.944	0.940	0.942	0.944
Model 3	0.952	0.950	0.949	0.949

Table 3: 11-point average precision for audience

Method	SNR			
	clean	20 dB	15 dB	10 dB
Baseline MUSIC	0.199	0.175	0.169	0.167
Baseline + location	0.276	0.277	0.276	0.264
Acoustic-only LR	0.419	0.424	0.403	0.365
Model 1	0.350	0.342	0.327	0.307
Model 2	0.414	0.430	0.426	0.412
Model 3	0.476	0.485	0.472	0.452

3. Acoustic-only LR model
   This model does not use the eye-gaze information in Model 1 and Model 2, and only uses the acoustic information of the MUSIC spectrum to compute $p(v|a)$ with LR.

To evaluate performance under ambient noise, we prepared audio data with ambient noise. This was done by superimposing the 19-channel audio signals on a diffusive noise actually recorded in a crowded place. Signal-to-Noise Ratios (SNRs) were 20, 15, and 10 dB.

5.2. Results and discussion

To evaluate the performance of speaker diarization, precision and recall are computed as

$$\text{Precision} = \frac{\#\text{TP}}{\#\text{TP} + \#\text{FP}},$$
$$\text{Recall} = \frac{\#\text{TP}}{\#\text{TP} + \#\text{FN}},$$

where TP, FP, and FN stand for true positive, false positive, and false negative, respectively. In addition, 11-point average precision was also calculated. This is the mean value of interpolated precision on a fixed 11 recall points from 0 to 1: $\{0, 0.1, 0.2, \cdots, 0.9, 1.0\}$.

Figure 1 and Figure 2 show the precision-recall curves for the presenter and audience, respectively, under the clean speech data. The results under ambient noise (SNR = 10 dB) are shown in Figure 3 and Figure 4. The curves were obtained by varying the thresholds in (3), (4), and (9). Table 2 and Table 3 list 11-point average precision for the presenter and audience.

As shown in these results, detection of the presenter's speech is easy because the presenter stands close to the microphones and speaks in most of the session. All the methods show similar high performance, and there are no significant differences among them, expect baseline+location lowers recall due to the location errors.

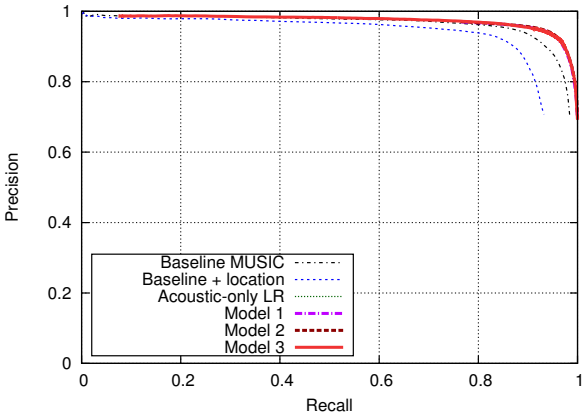

Figure 1: Precision-recall curve for presenter (Clean speech)

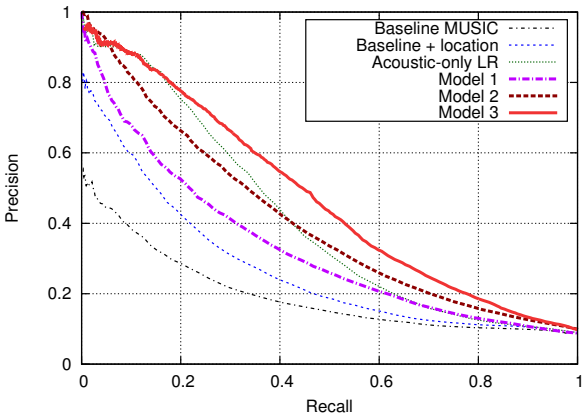

Figure 2: Precision-recall curve for audience (Clean speech)

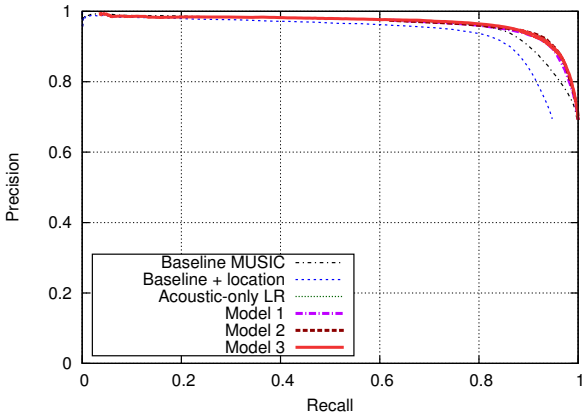

Figure 3: Precision-recall curve for presenter (SNR = 10 dB)

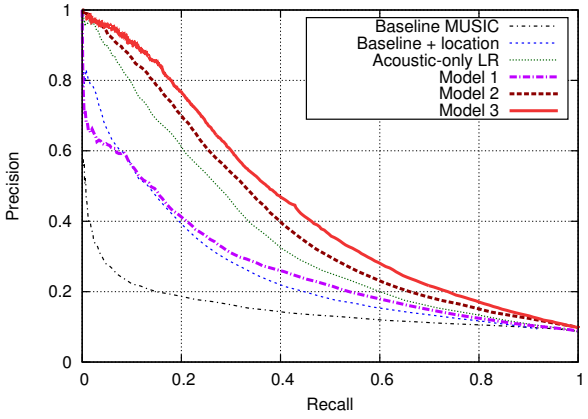

Figure 4: Precision-recall curve for audience (SNR = 10 dB)

Detection of speech by the audience is more difficult and the performance is much lower because of distant talking and infrequency of the utterances. In this case, the proposed Model 2 and Model 3 outperform the other methods. Especially under ambient noise, superiority of the models drastically increases. This result demonstrates that eye-gaze information contributes to improvement of diarization accuracy in real-world conversations.

Among the proposed three models, Model 1 has lower performance than the other models and the acoustic-only LR model. This suggests that the two variables (a and g) should be dealt independently. Model 3 shows higher performance than Model 2. It is easier to estimate the generative model of the acoustic features than the discriminative model with a limited size of training data.

Among the three reference methods, the acoustic-only LR model shows higher performance than the baseline MUSIC and baseline+location. This result indicates the effectiveness of the machine learning approach on this diarization problem. However, the robustness of the machine learning including the proposed methods needs to be investigated.

All the results shown here are based on automatic measurement of the eye-gaze features or head orientations. The mean errors of the estimated participants' locations and head orientations were 12.2 millimeters and 5.21 degrees respectively. The degradation by using the automatically detected eye-gaze information from the oracle information measured by the magnetometric sensors is about 1-3%.

6. Conclusions

This paper has proposed a novel approach to speaker diarization in multi-party conversations by combining acoustic and eye-gaze information. We presented the three models to combine the eye-gaze features with the acoustic features. From the experimental results, the proposed method achieved significant improvement of diarization accuracy especially for the audience's utterances.

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

7. References

[1] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image and Vision Computing*, vol. 27, no. 12, pp. 1775–1787, 2009.

[2] K. Otsuka, "Conversation scene analysis," *Signal Processing Magazine, IEEE*, vol. 28, no. 4, pp. 127–131, 2011.

[3] T. Kawahara, "Smart posterboard: Multi-modal sensing and analysis of poster conversations," in *Proc. APSIPA ASC*, 2013, pp. 1–5.

[4] T. Kawahara, T. Iwatate, and K. Takanashi, "Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations," in *Proc. INTERSPEECH*, 2012, pp. 727–730.

[5] T. Kawahara, S. Hayashi, and K. Takanashi, "Estimation of interest and comprehension level of audience through multi-modal behaviors in poster conversations," in *Proc. INTERSPEECH*, 2013, pp. 25–29.

[6] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[7] G. Friedland, A. Janin, D. Imseng, X. A. Miro, L. Gottlieb, M. Huijbregts, M. T. Knox, and O. Vinyals, "The ICSI RT-09 speaker diarization system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 371–381, 2012.

[8] 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.

[9] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta psychologica*, vol. 26, no. 1, pp. 22–63, 1967.

[10] K. Jokinen, K. Harada, M. Nishida, and S. Yamamoto, "Turn-alignment using eye-gaze and speech in conversational interaction." in *Proc. INTERSPEECH*, 2010, pp. 2018–2021.

[11] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, "A DOA based speaker diarization system for real meetings," in *Proc. HSCMA*, 2008, pp. 29–32.

[12] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[13] F. Asano, M. Goto, K. Itou, and H. Asoh, "Real-time sound source localization and separation system and its application to automatic speech recognition." in *Proc. EUROSPEECH*, 2001, pp. 1013–1016.

[14] H. Yoshimoto and Y. Nakamura, "Cubistic representation for real-time 3D shape and pose estimation of unknown rigid object," in *Proc. ICCV Workshop*, 2013, pp. 522–529.