Speaker Identification Based on Integrated Face Direction in a Group Conversation

Naoto Ienaga   Yuko Ozasa   Hideo Saito
Keio University
{ienaga, saito}@hvrl.ics.keio.ac.jp, yuko.ozasa@keio.jp

Abstract

We present a method for vision-based speaker identification in a group conversation. The group context in the conversation is modeled by the integrated face direction of group members. Experimental results show that integrated face direction of group members is effective for speaker identification in a group.

1. Introduction

In recent years, communication robots equipped with a spoken-dialog system have been widely used at public facilities such as shopping malls, restaurants, etc [3]. Some robots introduce information about the public facility in reaction to a conversation with customers. In such a situation, the robots are required to have a conversation with one customer or a group of customers [8]. This is particularly challenging when there are many customers in a group at a public facility, and so we focus on group conversation.

One of the inputs of a dialog system is speech recognition, what the customer said. For a dialog system for group conversation, speaker identification, which customer spoke in a group, is also required [9]. In this paper, we focus on speaker identification in a group for the dialog system for group conversation.

Speech information has been used in previous studies for speaker identification in a group. Many researchers estimate sound source position by gathering speech information and identifying a speaker [12]. When we assume that there is a lot of outside sound interference which make performing speaker identification more challenging in public facilities, it is worth trying to do vision-based speaker identification to get knowledge about how accurate it is. We then focus on vision-based speaker identification which differs from multimodal speaker diarization such as [10].

Some research uses image information for speaker identification [8]. Many works based on image information use face direction or gaze information as the image information [13]. The gaze information is obtained from the face direction. In a group conversation, the group members may talk with each other, and a speaker is watched by other group members. So, when the speaker changes, the face directions of the group members change. The face direction is therefore useful for speaker identification, and that is what is used to determine speaker identification in this paper.

There must be a context in a group conversation. However, the previous method treated the face direction of each member individually [1, 8]. They then identified speaker using each face direction. In this study, we model the group context by integrating face directions of the group members. We assume that the integration can model the relation of the face directions. Fathi et al. integrated each individual's role which was determined by using where people looking at [5]. Though they used first person view camera, we use cameras set in an environment because we assume that robots have cameras and customers do not have as we mentioned above.

In this paper, we propose vision-based speaker identification method for a group conversation which considers the group context. The group context is modeled by integrating the face directions of the group members. The face directions of group members are linearly combined, and these combined directions are used as a feature vector in speaker identification. Support Vector Machine (SVM) [6] is used as the discriminator in the identification [14]. In our experiments, a new dataset for speaker identification in a group is constructed, and used as an evaluation of the proposed method. Experimental results show that integrated face direction of group members is effective for speaker identification in a group.

2. Speaker identification based on integrated face direction of group members

In this paper, we suppose that group members do not speak at the same time and that the number of speakers is one in each frame in a group conversation. Considering this assumption, the speaker is identified by utterance detection in a group conversation. Utterance detection de-

978-1-5090-4941-7/17 $31.00 © 2017 IEEE
DOI 10.1109/WACVW.2017.16

IEEE computer society

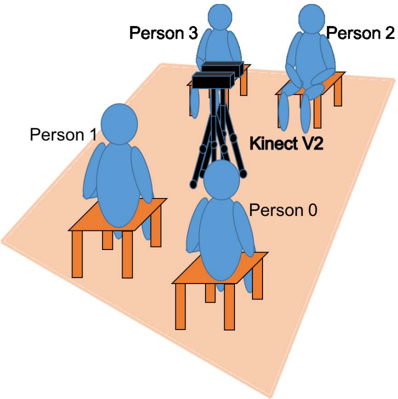

Figure 1. Illustration of our setting.

notes "whether a person is speaking or not." Speaker identification denotes "who is speaking." The result of speaker identification is the person with the highest confidence toward the model of utterance detection.

Our method integrates face directions of group members and uses the integrated directions as a feature vector for speaker identification of the group members. Three types of face direction are used in our method, such as roll, pitch, and yaw. Roll, pitch, and yaw are denoted as r, p, and y, and $0° \leq r, p, y \leq 360°$. Face directions, and r, p, y are obtained by Microsoft Kinect V2 [15].

When N persons have a group conversation, the person ID is denoted as i, and the roll, pitch, and yaw of i-th person are denoted as r_i, p_i, and y_i where $i = \{0, 1, \cdots, N-1\}$, $0 \leq (r_i, p_i, y_i) < 360$. The feature vector of t-th frame $\mathbf{f}^t$ is as follows:

$$\mathbf{f}^t = \{r^t_0, p^t_0, y^t_0, \cdots, r^t_{N-1}, p^t_{N-1}, y^t_{N-1}\} \tag{1}$$

$\mathbf{f}^t$ is $3N$-dimensional vector. The speaker is identified by SVM [11] using the feature vector $\mathbf{f}^t$. In this method, the SVM is multiclass SVM, the number of the classes is $N-1$, and teaching signal is a speaker label j^t ($j^t = \{0, 1, \cdots, N-1\}$).

3. Experiments

First, there is no dataset which includes the face directions. A new dataset was therefore constructed in our experiments. Second, accuracy of the utterance detection was evaluated for the preliminary experiment of speaker identification. Third, the identification using integrated face direction was evaluated. Finally, the accuracy of the identification was evaluated while changing the amount of training data.

3.1. Dataset

Four persons ($N = 4$) cooperated with our experiment, and they had an approximately 30 minute conversation. The situation of the conversation was captured by two Kinect V2s. The data was annotated that each person was speaking or not.

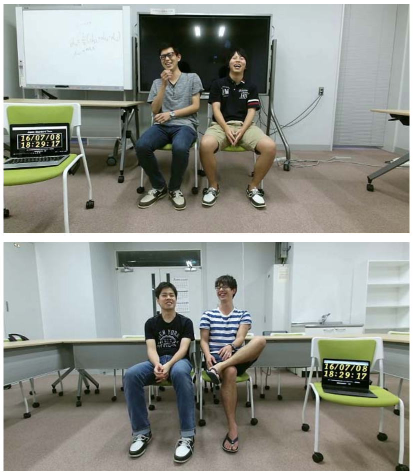

Figure 2. Conversation scene of four research subjects.

Figure 1 shows the setting. Two Kinect V2s were set back to back, and two persons were captured by each Kinect V2. An example scene of the conversation is shown in Figure 2. Three types of face directions (r, p, y) of each person were captured by each frame. The face directions obtained by Kinect V2s denote how tilted it was when looking at Kinect V2 in front. Therefore, the directions of positive and negative of roll and yaw were reversed between persons 0 and 1, and 2 and 3. The roll and yaw of persons 0 and 1 were changed as $\acute{r}$ and $\acute{y}$, where

$$\acute{r}^t_{0,1} = 360 - r^t_{0,1} \tag{2}$$

$$\acute{y}^t_{0,1} = 360 - y^t_{0,1} \tag{3}$$

The number of frames of data used in our experiments is 6835. For the internal division of data, the number of the frames of persons 0, 1, 2, and 3 are 1507, 1590, 1700, and 2083, respectively.

3.2. Evaluation of utterance detection

In this section, utterance detection using face direction is evaluated as the preliminary experiment of speaker identification. The result of the identification is the person with the highest confidence toward the model of utterance detection. Therefore, if the accuracy of the utterance detection becomes higher, the accuracy of the identification becomes higher.

The result of the utterance detection is whether a person is speaking or not. In this experiment, two-class SVM was used for detection and face direction is used for a feature vector. In training SVM phase, if a person is speaking,

Table 1. Accuracy of utterance detection for each person.

Person		0	1	2	3
Accuracy	Non-integrated	50.1	50.0	49.9	48.9
	Integrated	51.6	48.2	50.2	**54.8**

Table 2. Accuracy of speaker identification based on integrated face direction (%).

NN	RF	Linear SVM	Non-linear SVM(RBF)
28.1	32.8	31.3	**33.6**

the teaching signal is set to 1; otherwise, the teaching signal is set to 0. Two types of feature vectors are compared in this experiment; the face direction of a person, and the integrated face direction of person.

Table 1 shows the accuracy of utterance detection in each person. "Non-integrated" denotes the result using a 3-dimensional feature vector (r^t, p^t, y^t) of one person. "Integrated" denotes the result using a 12-dimensional feature vector $\mathbf{f}^t$, which consists of face directions of group members. The discriminator is SVM with RBF kernel. 80% of the 6835 frames, from the beginning when all the frames were arranged in chronological order, were set as training data, and the remaining frames were set as test data. The accuracy of "Integrated" is higher than that of "Non-integrated," and it shows that the integration of the face directions of group members is effective. However, each accuracy is around 50%, and this shows that it is difficult to detect utterance only by the face direction.

3.3. Evaluation of speaker identification based on integrated face direction

In this experiment, the methods using integrated face direction with different discriminators were compared. The discriminators prepared as the comparison targets were Nearest Neighbor [4] and Random Forests [2]. The discriminators of our method; linear SVM [11], and non-linear SVM with RBF kernel, were prepared. The selection of training data and test data was as same as the experiment in the previous section. The number of the trees in random forests and the parameter of SVMs were optimized in the experiments. To determine the randomness of the random forest, the accuracy is obtained 10 times, and the average of the accuracy is shown in the result when the random forest was used.

Table 2 shows the accuracy of speaker identification based on integrated face direction with several discriminators. The proposed method using SVM with RBF kernel is most effective of all.

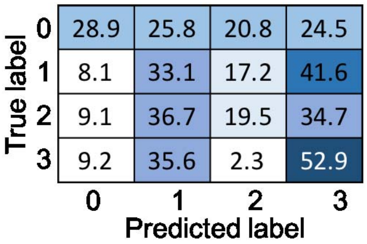

Figure 3. Confusion matrix of SVM with RBF kernel

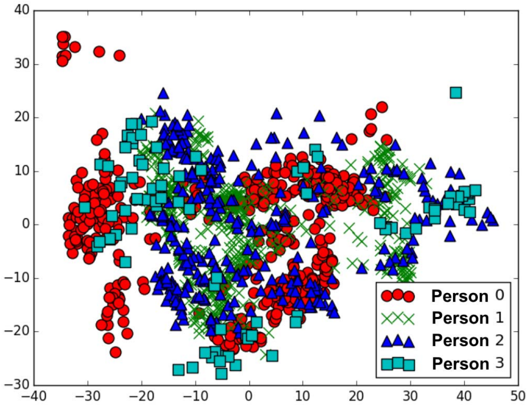

Figure 4. Visualization of test data by principal component analysis.

Table 3. Accuracy for each amount of training data.

Amount of the data	50	60	70	80	90
Accuracy	20.5	24.8	31.0	**33.6**	23.2

Figure 3 shows the confusion matrix of the proposed method using SVM with RBF kernel. The accuracy of each person was 28.9%, 33.1%, 19.5%, and 52.0%, respectively. The accuracy varied widely between persons.

Figure 4 is a visualization of the test data using principal component analysis (PCA). The feature vector of integrated face direction $\mathbf{f}^t$ is a 12-dimensional vector, but the dimensions of the vector were reduced into 2-dimension using PCA. The data of each person is plotted in Figure 4. The area of each data overlaps widely, implying that it is difficult to discriminate speakers only by the integrated face direction.

3.4. Evaluation of speaker identification while changing the amount of training data

The accuracy changes when the amount of the training data is changed. In this section, the accuracy of the method using integrated face direction was evaluated while the amount of the training data was changed. The SVM with RBF kernel was used for evaluation.

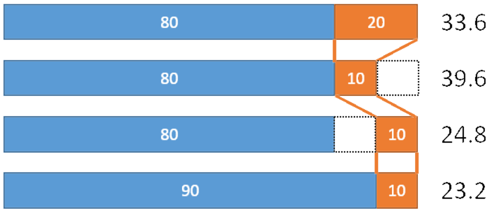

Figure 5. Changes of accuracy while the ratio of training data changed.

The result is shown in Table 3. 50% of the training data denotes that 50% of the 6835 frames from the beginning, when all the frames were arranged in chronological order, was used as training data. In this case, the test data remaining was 50%. The accuracy became higher when the amount of training data became larger, from 50% to 80%. When the amount became 90%, the accuracy became lower. The reason is shown in Figure 5. The blue area denotes the training data, and the orange area denotes the test data. When the training data was 80% of all frames and the test data was the remaining 20%, the accuracy of the identification was 33.6%. When 20% from the beginning of the data was split in half and the former and the latter ware used for the test data, the accuracy of the identification was 39.6% and 24.8% , respectively. The accuracy of the method using the former data is higher than that of the method using the latter data. This means that the end of 10% data was peculiar data, and hence the accuracy became lower when the test data is the end of 10% data, and the remaining is the training data.

4. Discussion

We presented vision-based speaker identification which did not use sound information because we assumed the situation that there is a lot of noise. In this section, assuming that we can use not accurate but rough sound information, sound direction, we present a method using sound direction.

4.1. Speaker identification based on sound direction

Both confidence c of the estimation [7] and estimated sound direction a are obtained by Kinect V2 where $0 \leq c \leq 1$, $-50^\circ \leq a \leq 50^\circ$ by 1°. When there are N persons in front of one Kinect V2, they have to be in the range of -50° to 50° of Kinect V2, and we identify the speaker as follows. First, we evenly divide the range in which Kinect V2 got sound by N.

$$\theta_{sd} = (max_{sd} + min_{sd})/N \tag{4}$$

max_{sd} is maximum a of data and min_{sd} is minimum. Next, we identify the speaker by assuming persons are sitting equal distances away from each other. Figure 6 illustrates this method. When a is between the angle denoted by the red arc, the identified speaker will be the person sitting second from the right end.

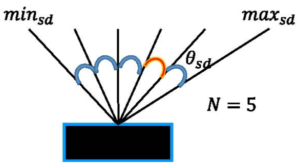

Figure 6. Method based on sound direction when $N = 5$.

Table 4. Accuracy of the identification using sound direction.

	Face + Sound (4.2.)	Face (3.3.)	Sound (4.1.)
NN	27.9	28.1	32.8
RF	34.2	32.8	
SVM	36.7	31.3	
RBF	**37.3**	33.6	

The reason we use such a method is that it is difficult to measure angles between persons and Kinect V2. In other words, we cannot know where persons are sitting.

When we use K Kinect V2s, we identify the speaker by using a of Kinect V2, which has the highest c.

4.2. Speaker identification based on integrated face direction and sound direction

A method using both integrated face direction and sound direction is presented in this section. The integrated face direction of group members, the confidence of sound direction c, and sound direction a are linearly combined, and it is used as a feature vector in our method. The feature vector of t-th frame $\mathbf{F}^t$ is as follows:

$$\begin{aligned} \mathbf{F}^t = \{ & r_0^t, p_0^t, y_0^t, \cdots, r_{N-1}^t, p_{N-1}^t, y_{N-1}^t, \\ & a_0^t, c_0^t, \cdots, a_K^t, c_K^t \} \end{aligned} \tag{5}$$

$\mathbf{F}^t$ is $3N + 2K$-dimensional vector. Using this feature vector, the speaker is identified in the same way as the method in Sec. 2.

4.3. Evaluation of speaker identification based on integrated face direction and sound direction

Based on the results shown in Table 1, integrated face direction is effective for speaker identification. In this section, we compare the accuracy between the methods using integrated face direction only, sound direction only, and both integrated face direction and sound direction and its confidence.

Table 4 shows the result of the comparison. $N = 4$ same as Sec. 3. and $K = 2$. “Face + Sound (4.2.)” denotes the method using the feature vector $\mathbf{F}^t$, which consists of both sound information and integrated face direction described in Section 4.2.; "Face (3.3.)" denotes the method using the feature vector $\mathbf{f}^t$, which consists of face directions described in Section 3.3.; and "Sound (4.1.)" denotes the method using the sound direction described in Section 4.1. The highest accuracy was 37.3%, when the vector $\mathbf{F}^t$ was used. The second highest was the method using $\mathbf{f}^t$, and the accuracy of the method using sound direction only was the lowest. This result shows that using integrated face direction is effective for speaker identification in a group, and using both face direction and sound direction is the most effective for the identification if we can utilize sound direction.

5. Conclusion

The integrated face direction of group members was used for speaker identification in a group conversation in this paper. Based on the experimental results, the integrated face direction was effective for speaker identification.

The proposed method still has low versatility. In future works, first we construct a dataset which captures various people in various situations. We use test data which is unseen data obtained from different subject groups from that of training data. It is ideal when we consider realistic situation. Next, we make a speaker identification method robust to various aspects. For example, if the height of a member changes greatly, the face direction also changes greatly, so we consider the speaker's own characteristics and the position of the speaker. In this research, we did not consider the change in the time axis direction. We present a speaker identification method in consideration of changes in the time axis direction.

Acknowledgment

This work was supported in part by JSPS Grant-in-Aid for Scientific Research(S) 24220004, and JST CREST.

References

[1] R. Böck, S. Glüge, I. Siegert, and A. Wendemuth. Annotation and classification of changes of involvement in group conversation. In *Affective Computing and Intelligent Interaction*, pages 803–808, 2013.

[2] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[3] P. Chiluiza, D. Fabian, C. Angulo Bahón, and M. Díaz Boladeras. An exploratory study of group-robot social interactions in a cultural center. 2015.

[4] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[5] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *Computer Vision and Pattern Recognition*, pages 1226–1233, 2012.

[6] C.-W. Hsu and C.-J. Lin. A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, 13(2):415–425, 2002.

[7] H. Jiang. Confidence measures for speech recognition: A survey. *Speech communication*, 45(4):455–470, 2005.

[8] Y. Matsusaka, S. Fujie, and T. Kobayashi. Modeling of conversational strategy for the robot participating in the group conversation. In *Interspeech*, pages 2173–2176, 2001.

[9] Y. Matsuyama, I. Akiba, S. Fujie, and T. Kobayashi. Four-participant group conversation: A facilitation robot controlling engagement density as the fourth participant. *Computer Speech and Language*, 33(1):1–24, 2015.

[10] K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich, and J. Yamato. A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization. In *the 10th international conference on Multimodal interfaces*, pages 257–264, 2008.

[11] J. A. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[12] O. Thyes, R. Kuhn, P. Nguyen, and J.-C. Junqua. Speaker identification and verification using eigenvoices. In *Interspeech*, pages 242–245, 2000.

[13] H. Vrzakova, R. Bednarik, Y. I. Nakano, and F. Nihei. Speakers' head and gaze dynamics weakly correlate in group conversation. In *the Ninth Biennial ACM Symposium on Eye Tracking Research and Applications*, pages 77–84, 2016.

[14] V. Wan and W. M. Campbell. Support vector machines for speaker verification and identification. In *IEEE Signal Processing Society Workshop, Neural Networks for Signal Processing X*, pages 775–784, 2000.

[15] Z. Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012.