# AUTOMATIC ADDRESSEE IDENTIFICATION BASED ON PARTICIPANTS' HEAD ORIENTATION AND UTTERANCES FOR MULTIPARTY CONVERSATIONS

*Yoshinao Takemae*[†*]

NTT Cyber Solutions Laboratories[†],
NTT Corporation
1–1 Hikari-no-oka, Yokosuka-Shi,
Kanagawa 239–0847 Japan
takemae.yoshinao@lab.ntt.co.jp

*Shinji Ozawa*[*]

Department of Information and Computer Science[*],
Keio University
3–14–1 Hiyoshi, Kohoku-Ku, Yokohama-Shi,
Kanagawa 223–8522 Japan
ozawa@ozawa.ics.keio.ac.jp

## ABSTRACT

We propose a method that uses the participants' head orientation and utterances for automatically identifying the addressee of each utterance in face-to-face multiparty conversations, such as meetings. First, each participant's head orientation is determined through vision-based detection and the presence/absence of utterances is extracted using the power of voices captured by microphones. Second, gaze direction (whom each participant is looking at) is estimated from just detected head orientation using the Support Vector Machine. Third, several related features such as *amount and frequency of gaze and eye contact* are calculated in each utterance interval. Finally, a Bayesian Network is used to classify each utterance into one of two types of utterances: *(a) the speaker is addressing a single participant* and *(b) the speaker is addressing all participants*. Experiments on addressee estimation with 3-person conversations confirm the usefulness of our method.

## 1. INTRODUCTION

Meetings are one of the most important activities in many workgroups. Often, due to scheduling conflicts or travel constraints, some cannot attend their scheduled meetings. We can overcome these problems by archiving the meetings and teleconferences. The need for systems that can effectively archive such sessions is increasing.

While this study focuses on archiving meetings for later review, a considerable overlap exists between this domain and teleconferencing. Most conventional systems for archiving meetings use a fixed-viewpoint camera or a fixed panoramic view camera [1, 2]. In large multiparty situations, participant face size is small. Hence these systems cannot sufficiently convey nonverbal information such as changes in facial expressions and gaze. These visual cues greatly contribute to the viewers' understanding of the participants' response. Moreover, other conventional systems based on participants' utterances [1] cannot adequately convey who the addressee is or her/his response, to the viewers, because only selected speakers are shown.

To solve this problem, a first and essential step is to automatically identify the addressee of each utterance during the conversation. We propose a novel method that uses the participants' head orientation and utterances for automatically identifying the addressee of each utterance in face-to-face multiparty conversations. The rest of the paper is organized as follows: Section 2 overviews the problem of addressee identification. Section 3 presents our approach. Section 4 details our method. Section 5 describes the experiments conducted. We summarize this paper in Section 6.

## 2. ADDRESSEE IDENTIFICATION -PROBLEM OVERVIEW-

Clark [3] or Goffman [4] proposed a taxonomy of conversational roles such as speaker and addressee. In more than 2-participant multiparty conversations, the participants dynamically change roles such as speaker, addressee (currently being talked to by the speaker), and side participant (not currently being addressed). These dynamic changes are signaled and managed through the use of various verbal and nonverbal cues such as gaze, posture, and gestures among the participants. In this paper, we adopt the above taxonomy.

Identifying the addressee is an important and urgent task in multiparty conversations and is the subject of this paper. Figure 1 shows an example of the two types of utterances assumed in this paper for 3-participant conversations. Figure 1 (a) shows the single-addressee utterance, the speaker is addressing just one participant. Figure 1 (b) shows the multi-addressee utterance, the speaker is addressing all participants. Hence, the problem of addressee identification amounts to the problem of distinguishing the addressee from the side participants.

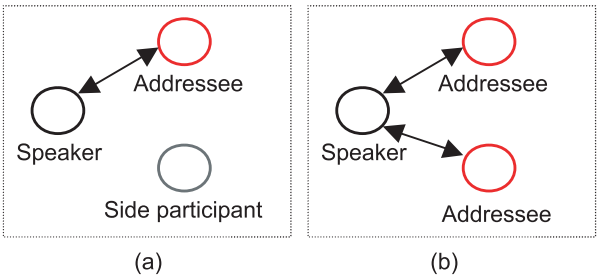


**Fig. 1**. An example of two types of utterances in a 3-person conversation. (a) shows the single-addressee utterance. (b) shows the multi-addressee utterance.

## 3. OUR APPROACH

### 3.1. Cues for identifying addressee

We focus on participants' gaze behavior as cues in developing a method for identifying addressee for the following reasons. In face-to-face multiparty conversations, the dynamic conversational roles such as speaker and addressee are determined through the use of various nonverbal cues such as gaze, facial expression, posture, and gestures in addition to verbal information. Among various nonverbal cues, it was pointed out in 1967 that gaze is closely related to the speech act (exchange of conversation roles) [5, 6]. In more recent work, Vertegaal et al. experimentally indicated that gaze direction (whom the participant is looking at during a conversation) is a very important resource in identifying the addressee [7].

### 3.2. Measuring gaze direction

We track the participants' head orientation rather than their gaze direction for the following reasons: 1) It has been shown that there is a high correlation between gaze direction and head orientation in meeting situations [8]. 2) While many automatic vision-based gaze tracking techniques been developed [9, 10], most fail to meet the requirement of not interfering with natural conversation. A more practical solution is to use one of the recent vision-based face tracking techniques that can robustly estimate head orientation without hindering the conversation.

## 4. PROPOSED METHOD

### 4.1. Overview

Figure 2 overviews our method. It consists of five modules: data gathering, head tracking, gaze estimation, feature extraction, and classification. In data gathering, stereo images of each participant's head are captured from the stereo camera assigned to each participant. Utterance intervals are extracted from the audio data captured by the clip-on microphone attached to each participant. In head tracking, the head orientation of each participant is automatically detected from the sequences output by each stereo camera. In gaze estimation, gaze direction is estimated by a trained Support Vector Machine [12]. In feature extraction, several features such as *amount of gaze and eye contact* are extracted in each utterance interval. Classification uses a trained Bayesian Network to label each utterance as one of two types: single-addressee utterance or multiple-addressee utterance.

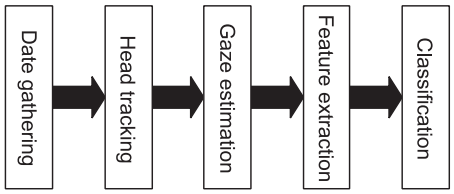


**Fig. 2**. Overview of our method

### 4.2. Head tracking

We implemented the novel approach proposed by Morency et al. [11] for detecting head orientation. This approach has several advantages: 1) This technique requires no expert knowledge. It does not need the registration of a training template of each participant's face. As soon as the participant appears in front of the stereo camera, the technique can track head orientation online with high accuracy (rotational RMS error is smaller than 3°). 2) This technique is robust against variations in illumination and large head motion for long periods of time. Figure 3 shows the results of head tracking in a 3-participant conversation. Figure 4 shows the histograms of horizontal head orientation (azimuth) of one subject during a conversation. This histogram exhibits two peaks. These correspond to the directions of the other two participants.

### 4.3. Gaze estimation

Each participant's gaze direction (whom each participant is looking at) is extracted from both azimuth and elevation of estimated head orientation using a trained Support Vector Machine [12]; the training is done using captured conversation data. Concretely, for each frame of the captured video, whether the participant's gaze is directed toward another's participant's face area or somewhere else is estimated.

### 4.4. Feature extraction

In each utterance, the following features based on gaze behavior are extracted.

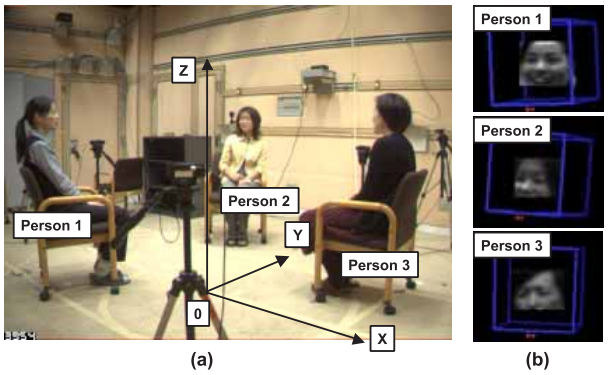


**Fig. 3**. Head tracking results from a 3-participant conversation. (a) shows overall view of participants and world coordinates (X, Y, Z). (b) shows head tracking for each participant. Cubes represent participants' head orientation and position.

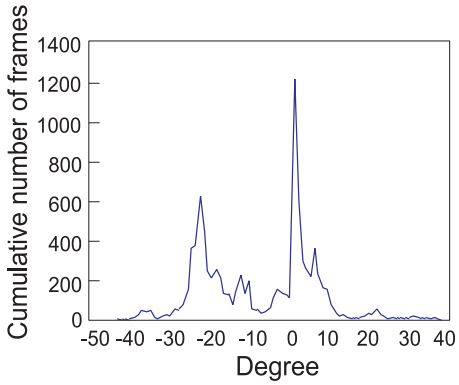


**Fig. 4**. Histogram of horizontal head orientation (azimuth) of one subject during a conversation.

#### 4.4.1. *Features based on speaker's gaze*

Based on past findings that a speaker uses gaze primarily to indicate whom the speaker is addressing in multiparty situations [5], we hypothesize that the speaker' gaze direction is more focused (relative duration and frequency) on the addressee than participants in single-addressee utterances.

*(1) C1: Relative gaze duration.* $T_i$ is the total amount of time (duration) the speaker directs his/her gaze direction at other participants in utterance interval $i$. We select the person receiving the longest gaze duration, $T_{pi}$, as the single-addressee candidate P. $C1_i$ is defined as the percentage of $T_{pi}$ to $T_i$ as follows:

$$C1_i = \frac{T_{pi}}{T_i}\,(\%) \tag{1}$$

*(2) C2: Relative gaze frequency.* $F_i$ is the total number of times the speaker gazed at other participants in utterance interval $i$. We select the person receiving the highest number of gazes, $F_{pi}$, as addressee candidate P. $C2_i$ is defined as the percentage of $F_{pi}$ to $F_i$ as follows:

$$C2_i = \frac{F_{pi}}{F_i}\,(\%) \tag{2}$$

#### 4.4.2. *Features based on eye contact*

Based on past findings that eye contact between two people is used as a signal for regulating conversation flow (turn taking) [5], we hypothesize that a message exchange can be limited, through eye contact, to just two people in single-addressee utterances (the speaker and one addressee); other participants are side-participants. Hence, the relative amount and frequency of eye contact between a speaker and one participant is expected to be larger in single-addressee utterance than in multiple-addressee utterance.

*(1) C3: Relative amount of eye contact.* $S_i$ is the total amount of time (duration) of eye contact between the speaker and other participants in utterance interval $i$. We select the person receiving longest duration of eye contact, $S_{pi}$, as the single-addressee candidate P. $C3_i$ is defined as the percentage of $S_{pi}$ to $S_i$ as follows:

$$C3_i = \frac{S_{pi}}{S_i}\,(\%) \tag{3}$$

*(2) C4: Relative frequency of eye contact.* $U_i$ is the total number of times eye contact is made between the speaker and other participants in utterance interval $i$. We select the person receiving the highest number of eye contacts, $U_{pi}$, as addressee candidate P. $C4_i$ is defined as the percentage of $U_{pi}$ to $U_i$ as follows:

$$C4_i = \frac{U_{pi}}{U_i}\,(\%) \tag{4}$$

### 4.5. Classification

We use naive Bayesian predictors [13], which are widely used for human modeling and pattern recognition. Based on the extracted features mentioned in Section 4.4, each utterance is classified as one of two types of utterances: single-addressee utterance and multiple-addressee utterance. For this we use naive Bayesian predictors that are trained using captured conversation data. Note that if the utterance is found to be a single-addressee utterance, the person receiving the largest amount of gaze is identified as the single-addressee candidate.

## 5. EXPERIMENTS

We conducted experiments to verify the effectiveness of our method in estimating the addressee of each utterance.

### 5.1. Data collection

We focused on face-to-face 3-participant debates. Two groups participated in the debates. The participants in each group were Japanese females (average age was 34.3). Each group debated topics such as *"whether we should legally recognize the death penalty in Japan"*. The two debates took about 404 and 518 seconds. Video sequences from three stereo cameras were captured. Voice data was captured by clip-on microphones. The logarithmic powers of recorded voice in 100 [ms] intervals were calculated. By determining the threshold of these powers, silent intervals were extracted. An utterance interval was extracted as a temporal subsection bounded by prior/subsequent silent intervals longer than 500 [ms]. One evaluator (male, age 27), who did not participate in the debates, subjectively determined the "correct" addressee of each utterance by viewing the above debates videos with voice, since unfortunately there is still no criterion for objectively determining addressee. This evaluation was done based on a synthetic determination including speech, the context of conversation, the relations between participants, and nonverbal cues. Utterances that included only extraneous information such as back-channel comments and the sound of laughter were removed manually, and the remaining 106 utterance intervals were used in the evaluations. The total number of single-addressee utterances and multiple-addressee utterances was 40 and 66, respectively.

### 5.2. Evaluations

70 % of the 106 utterance intervals were randomly chosen and used as training data. The rest (30 %) were used as test data. We investigated how well our method could identify the correct utterance type and correct addressee in the test data. The results show that the correct rate of addressee estimation was 74%. A key reason for the error was that the speaker sometimes turned her gaze to one specific participant even though she was making a multiple-addressee utterance. Another basic problem is head tracking error. In the conversations observed, the subjects sometimes touched their face, which confused the head tracking method.

The above results indicate that our method is relatively effective for automatic addressee identification, since it dispenses with the need to analyze complex information such as the context of the conversation, verbal cues, and another nonverbal cues.

## 6. CONCLUSIONS AND FUTURE WORKS

Addressee identification is an interesting and important aspect of multiparty conversation. In this paper, we introduced a method that offers automatic addressee identification in multiparty conversations based on the participants' head orientation and utterances. Our system consists of five modules: data gathering, head tracking, gaze estimation, feature extraction, and classification. Experiments on addressee estimation with 3-person conversations confirm the usefulness of our method. In the future, as the next step to more accurately identifying the addressee, we will incorporate other human behaviors such as head gestures (like nodding and shaking). We then will develop more robust head and gaze tracking techniques. Furthermore, we will use a sequential statistical model such as Dynamic Bayesian Networks which internally store previous human behavior. This will allow the consideration of verbal and nonverbal cues dynamically exchanged among participants.

## 7. REFERENCES


[1] Cutler, R., Rui, Y., Gupta, A., Gadiz, J., Tashev, I., wei He, L., Colburn, A., Zhang, Z. and Silverberg, S. Distributed Meetings: A Meeting Capture and Broadcasting System, *Proc. ACM Multimedia*, pp.503–512, 2002.

[2] Lee, D.-S., Erol, B., Graham, J., Hull, J.J. and Murata, N. Portable Meeting Recorder, *Proc. ACM Multimedia*, pp. 493–502, 2002.

[3] Clark, H. H. Using Language, Cambridge University Press, 1996.

[4] Goffman, E. Replies and Responses, *Language in Society*, Vol. 5, pp.257–313, 1982.

[5] Kendon, A. Some Function of Gaze-Direction in Social Interaction, *Act. Psychologica*, Vol. 26, pp.22–63, 1967.

[6] Argyle, M. and Ingham, R. Gaze, Mutual Gaze and Proximity, *Semiotica 6(1)*, pp.32–49. 1972.

[7] Vertegaal, R., Slagter, R., Veer, van der G., and Nijholt, A. Eye Gaze Patterns in Conversations: There is More to Conversational Agents than Meets the Eyes, *Proc. CHI 2001*, pp.301–308, 2001

[8] Stiefelhagen, R., Zhu, J. Head Orientation and Gaze Direction in Meetings, *Ext. Abstracts CHI 2002*, pp.858–859, 2002.

[9] Matsumoto, Y., Ogasawara, T., and Zelinskey, A. Behavior Recognition Based on Head Pose and Gaze Direction Measurement, *Proc. IEEE International Conference on Intelligent Robots and Systems*, pp.262–267, 2000.

[10] Ohno, T., Mukawa, N., and Kawato, S. Just Blink Your Eyes: A Head-Free Gaze Tracking System, *Ext. Abstracts of CHI '03*, pp.950–951, 2003.

[11] Morency, L.-P., Rahimi, A., and Darrell, T. Adaptive View-based Appearance Model, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.803-810, 2003.

[12] Burges, C. J. C. A Tutorial on Support Vector Machines for Patter Recognition, *Data Mining and Knowledge Discovery*, 2, pp.121-167, 1998.

[13] Charniak, E. Bayesian Networks without Tears: Making Bayesian Networks more Accessible to the Probabilistically unsophisticated, *AI Magazine*, Vol.12, No.4, pp.50-63, 1991.