# Implementation and Evaluation of a Multimodal Addressee Identification Mechanism for Multiparty Conversation Systems

Yukiko I. Nakano[1] Naoya Baba[1] Hung-Hsuan Huang[2] Yuki Hayashi[1]

[1] Seikei University
3-3-1 Kichijoji-kitamachi
Musashino-shi, Tokyo 180-8633, Japan
+81 422 37 3748
{y.nakano, hayashi}@st.seikei.ac.jp

[2] Ritsumeikan University
1-1-1 Noji-higashi
Kusatsu-shi, Shiga 525-8577, Japan
+81 77 561 4927
hhhuang@acm.org

## ABSTRACT
In conversational agents with multiparty communication functionality, a system needs to be able to identify the addressee for the current floor and respond to the user when the utterance is addressed to the agent. This study proposes some addressee identification models based on speech and gaze information, and tests whether the models can be applied to different proxemics. We build an addressee identification mechanism by implementing the models and incorporate it into a fully autonomous multiparty conversational agent. The system identifies the addressee from online multimodal data and uses this information in language understanding and dialogue management. Finally, an evaluation experiment shows that the proposed addressee identification mechanism works well in a real-time system, with an F-measure for addressee estimation of 0.8 for agent-addressed utterances. We also found that our system more successfully avoided disturbing the conversation by mistakenly taking a turn when the agent is not addressed.

## Categories and Subject Descriptors
H.1.2 [User/Machine Systems]: Human factors; H.5.2 [User Interfaces]: Evaluation/methodology; I.2.1 [Applications and Expert Systems]: Games, Natural language interfaces

## General Terms
Design, Experimentation, Human Factors.

## Keywords
Addressee identification, multiparty conversation systems, autonomous virtual agent, evaluation.

## 1. INTRODUCTION
In information kiosk agents that can communicate with a group of users, one of the most important functions is to respond to the user with proper timing by judging whether an utterance is addressed to the agent or to another user. Unless the system correctly recognizes addressee-hood, it does not respond to the user even when the user asks the agent a question. More seriously, if the agent speaks when it should not take a turn, such behaviors annoy the users and disturb the communication among the users. This paper aims to implement an addressee identification mechanism in a fully autonomous conversational agent, and investigates whether the proposed mechanism contributes to improving the effectiveness of human-agent multiparty communication.

However, addressee identification cannot be easily implemented. The system needs to have a speech processing mechanism that processes the speech signal and extracts prosodic and/or language information, along with motion sensors or computer-vision-based systems that track the user's motions. All of these types of multimodal information are fused in determining the addressee, but it sometimes happens that these input modules fail to recognize part of the user's speech and motions. In order to build an addressee identification mechanism that can effectively work in a real-world application, it is necessary to implement the mechanism in a practical manner and test whether its performance is good enough for practical usage. Moreover, there is a possibility that addressee identification models and their parameter settings need to be changed depending on the proxemics between the agent and the users. To address this issue, we will collect human-agent multiparty conversations in different proxemics and, based on the data, we will propose a general model that is applicable regardless of the distance from the agent.

In the following sections, first, related work for addressee identification will be discussed in section 2. Section 3 describes data analysis and proposes addressee identification models. In section 4, we implement an addressee identification mechanism based on the models, and incorporate the mechanism into a fully autonomous conversational agent. In section 5, we conduct an evaluation experiment to compare our system with a conversational agent with a naïve addressee identification mechanism, and show that our system properly communicates with novice users who have never used a conversational humanoid.

## 2. RELATED WORK
Floor control is a scheme that organizes who is taking the turns, and this is also important in human-computer dialogue systems. Schrangen [17] used lexical and prosodic information at the end of utterances to predict turn changes. Not only speech information, but also nonverbal bodily behaviors play an important role in coordinating turn-taking. Duncan [10] proposed that nonverbal cues, such as gestures and gaze, signal turn-taking. In multiparty


*ICMI'13*, December 9–13, 2013, Sydney, Australia.

http://dx.doi.org/10.1145/2522848.2522872

conversations, floor control becomes more complex compared to dyadic conversations because there are multiple possibilities about who takes the next turn. Using a group conversation corpus, Chen et al. [8] proposed a model for detecting a floor control shift based on verbal and nonverbal information, and showed that visual information such as gesture and gaze are useful in detecting turn-taking. Addressee identification is closely related to floor control because the addressee in the current turn may be the next turn holder with high probability. Thus, factors discussed in floor control are also useful in addressee identification. For example, many studies in addressee identification found that gaze information is useful in addressing [22, 18, 3]. Jovanovic et al. [13] proposed a model for addressee identification by applying Bayesian Network and Naive Bayes classifiers to a corpus annotated with gaze information, utterance, and conversational context.

These findings for addressing in multiparty conversations have been re-confirmed and/or reconsidered in multiparty conversations between human users and a computer (or a robot). Lunsford et al. [15] showed videos of multiparty conversations to subjects and asked them to judge to whom the utterance was addressed. They analyzed the human judgments of addressee-hood and reported that, in their judgment, gaze and prosodic information were useful. Lunsford et al. [16] analyzed triadic conversations between two human subjects and a computer system that asked them about basic problems in mathematics and gave explanations about technical terms of mathematics when subjects asked a question. They found that the subjects' speech power increased two to three dB in talking to the computer, and suggested that the change of speech power was a means for indicating the addressee that the speaker intended. Bakx et al. [4] analyzed pairs of users' behaviors in interacting with an information kiosk using speech input and a touch screen. They found that, unlike human conversations, the subjects looked at the screen the majority of the time, both when talking to the system and when talking to the other person. Terken et al. [19] analyzed three-party conversations in which two subjects played the role of customers and the third person played the role of a clerk at a travel agency. Based on their analysis, they discussed what type of information would be useful in implementing information kiosk conversational agents that can work as clerks and suggested that both gaze and speech information would be useful in addressee identification.

However, all these are analytical studies, and they did not propose addressee identification methods or models that can be implemented into a computer system. Moreover, these studies analyzed multiparty conversations collected in different experimental settings. For instance, the distance between the agent and the subjects, proxemics, and the size of the conversational agent were different depending on the study. Thus, there is no assurance that suggestions discussed in one study are applicable to different situations. Aiming at establishing more generic addressee identification models, this study examines whether the speech and gaze information discussed in the previous studies is useful even if the distance and the size of the agent are changed.

By referring to the empirical analyses shown above, some addressee identification models were proposed using machine-learning techniques. Turnhout et al. [21] proposed a model that judges whether an utterance is addressed to the system by applying a Naive Bayes classifier to a set of multimodal features: gaze, dialogue history, and utterance length information. Katzenmaier et al. [14] used gaze and linguistic information to identify the addressee in multiparty conversations with two human users and a robot. In their study, the gaze information was obtained based on automatic face tracking which determined whether a participant was looking at a robot or another participant. The linguistic information was obtained from an automatic speech recognition system. More recently, Vinyals et al. [23] proposed a method for addressee identification as well as speaker and overlap detection in multiparty conversations by directly reasoning about temporally streaming features.

In one study sharing a goal with this study in terms of building multiparty conversational agents, Traum et al. [20] built a system in which a user negotiates with two virtual agents. The agents were implemented as different agent systems with different standpoints, and the user communicates with the agents to persuade them. A series of studies by Bohus et al. [5, 7], which are closely related to this study, implemented multiparty conversation systems with two users interacting with one conversational agent. The main goal of their research was floor management in multiparty conversational systems, and as a part of this functionality, they implemented an addressee identification mechanism based on attentional focus information. They also implemented a mechanism that produced verbal and nonverbal agent's behaviors based on their turn-taking model [6]. On the other hand, this study solely focuses on addressee identification itself, and tries to establish a more general model that can be applicable to different proxemics, and to implement a more robust mechanism that is useful for practical usage where loss of sensing data frequently happens. Through implementation and evaluation of our addressee identification mechanism, we will show how well our model and mechanism work in practical multiparty conversational systems.

# 3. CORPUS ANALYSIS AND MODELS

This section analyzes multiparty conversation corpora in different proxemics between two users and an agent, and establishes addressee identification models based on speech and head-direction parameters.

## 3.1 Corpus Collection

Employing the Wizard-of-Oz (WOZ) method, we collected human-human-agent multiparty conversations in two different experimental settings. Pairs of subjects were instructed to interact with an animated agent on a screen/display, and to retrieve information in order to make a joint decision regarding given tasks.

Experiment 1 (Exp 1): the subjects stood about 1.5 m away from a 120-inch rear-projection type screen and interacted with a life-sized female animated character on the screen. The distance between the subjects was 20cm. The experimental setting is shown in Figure 1(a).

Experiment 2 (Exp 2): each subject sat on a chair with a 20-inch computer display on the table in front of them. A half-body female animated character was shown on the display. The subjects sit 90cm away from each other with the distance between the display and each subject also being 90cm. Thus, two subjects and the display formed an equilateral triangle. The experimental setting is shown in Figure 1(b).


(a) Experiment 1


(b) Experiment 2

**Figure 1. Proxemics between subjects and the agent**

In both experiments, each pair of subjects was instructed to complete two decision-making tasks, such as choosing lectures to register for together and choosing places to visit in Kyoto. Sentences that the agent could speak were pre-defined in a GUI menu, and the WOZ operators selected the agent utterances from the menu. There was also a text field that allowed them to type arbitrary utterances. More detailed description of the experimental procedure is given in [11].

The data for 17 pairs (10 male and 7 female pairs, or 34 subjects) collected in Exp 1 were analyzed. In Exp 2, ten pairs of university students (five male pairs and five female pairs) participated in the experiment and we analyzed the collected data.

The participants' interactions with the agent were recorded by two video cameras. In addition, a USB webcam (960 x 720 pixels, 30 fps) was set on the top of the screen to collect video data to be processed by FaceAPI [1], a vision-based face-tracking system.

## 3.2 Corpus Analysis

We analyzed the interaction for one experimental session for each pair. First, we divided the corpus into utterance units. When more than 200 ms of silence was observed, it was automatically identified as the end of the current utterance, and the subsequent speech was regarded as a new utterance. We also annotated the speaker and the addressee for each utterance, and used the annotation as the ground truth. The number of utterances addressed to the agent was 863 in Exp 1 and 518 in Exp 2, and that for utterances addressed to the other subject of the pair (hereafter referred to as the "partner") was 967 in Exp 1 and 533 in Exp 2.

### *3.2.1 Analysis of speech*

We analyzed pitch ($F_0$), intensity, speech rate, and duration of each user utterance. These prosodic features were extracted from each utterance using a speech analysis tool, Praat, at 0.01 sec interval. Since the precise speech rate, the number of phonemes per second, cannot be obtained automatically by Praat, we approximately calculated it from syllable counts [9].

The average values for pitch ($F_0$), intensity, speech rate, and duration for Exp 1 and 2 are shown in Table 1. The results of paired t-test are also shown in the table. In both experimental settings, the $F_0$ and the intensity were higher and the speech duration was longer when a person was speaking to the agent than when speaking to the partner. These results were statistically significant except for the $F_0$ in female pairs in Exp 1, where a statistical trend was found. The speech rate was slower when the person was speaking to the agent, but the results were not clear for male pairs in Exp 2.

In order to examine whether these findings were consistent regardless of the proxemics (experimental settings) and gender, a two-way ANOVA was used to test two between-subject factors: proxemics (Exp 1 or 2) and gender (male or female). We did not find statistically significant results for $F_0$, speech rate, and duration. This suggests that proxemics and gender did not affect the manner of speaking (i.e., the difference between speaking to the agent and speaking to the partner was consistently observed through the two experiments with both male and female pairs). However, these factors affected the intensity (proxemics: $F(1, 50) = 11.37$, $p < 0.01$, gender: $F(1, 50) = 8.84$, $p < 0.01$). The reason for the difference between Exp 1 and Exp 2 was that the microphones used in Exp 2 were different from those used in Exp 1, and the average audio amplitude was different between the experiments. The gender difference indicated that the speech intensity in male pairs was more clearly distinguished between the speech to the agent and that to the partner. Note that the interaction between these two factors was not statistically significant, indicating that we found consistent results that the speech intensity to the agent was greater than that to the partner regardless of proxemics and gender.

These results suggest that speech features are different depending on to whom the person talks. When people talk to the agent, they speak with a higher tone of voice and also speak more loudly and slowly. This tendency of speech is consistent regardless of proxemics and gender. Thus, we consider that these prosodic features are generic and useful in estimating the addressee.

**Table 1. Results of speech analysis**

| | | $F_0$ (Hz) | | Intensity (dB) | | | Speech Rate (syllable/s) | | | Duration (s) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Male | Female | Male | Female | Combined | Male | Female | Combined | Male | Female | Combined |
| Exp 1 | Agent | 131.6 | 241.3 | 63.8 | 63.6 | 63.7 | 4.59 | 4.56 | 4.61 | 1.99 | 2.22 | 2.08 |
| | Partner | 124.3 | 233.0 | 60.2 | 60.7 | 60.4 | 4.91 | 5.10 | 4.99 | 0.96 | 1.06 | 1.00 |
| | t(df) = t-value P | t(19) = 4.44 <.01 | t(13) = 1.46 <.10 | t(19) = 9.38 <.01 | t(13) = 7.84 <.01 | t(33) = 12.04 <.01 | t(19) = -3.00 <.01 | t(13) = -3.64 <.01 | t(33) = -4.64 <.01 | t(19) = 13.38 <.01 | t(13) = 10.05 <.01 | t(33) = 16.53 <.01 |
| Exp 2 | Agent | 123.3 | 232.6 | 56.3 | 59.6 | 57.9 | 4.89 | 4.90 | 4.89 | 2.12 | 2.29 | 2.20 |
| | Partner | 116.8 | 221.0 | 50.3 | 55.9 | 53.1 | 4.97 | 5.23 | 5.11 | 0.85 | 1.19 | 1.02 |
| | t(df) = t-value P | t(9) = 2.26 <.05 | t(9) = 4.37 <.01 | t(9) = 8.09 <.01 | t(9) = 9.84 <.01 | t(19) = 10.27 <.01 | t(8) = -0.19 <.50 | t(9) = -3.29 <.01 | t(18) = -1.11 <.20 | t(9) = 6.54 <.01 | t(9) = 9.57 <.01 | t(19) = 10.65 <.01 |

### *3.2.2 Analysis of head direction*

The video data collected in the experiments were analyzed using face-tracking software, FaceAPI [1], which can measure the head position and rotation in x, y, and z coordinates.

Since it is well known that the speaker frequently looks at the addressee during his or her speech, it is assumed that the subjects were looking at the agent when speaking to the agent, and looking at the partner when speaking to the partner. However, in previous work analyzing interaction between two users and an information kiosk system [4], it was reported that the subjects looked mainly at the computer screen, both when talking to the system and when talking to the other person. In our head-direction data for Exp 1, the subjects looked at the agent 93.2% of the time while they were talking to the agent. On the other hand, the speaker looked at his or her partner only 33.5% of the time while they were talking to one another. In Exp 2, the subjects looked at the agent 68.6% of the time while talking to the agent, and looked at the partner 40.4% of the time while talking to each other. In both experiments, when the subjects were talking to the partner, they looked at the agent the majority of the time. These results are consistent with the previous study [4]. Therefore, it is difficult to estimate the addressee solely based on head-direction information, suggesting that combining that information with speech information is necessary to get better estimation of addressee-hood.

## 3.3 Addressee Identification Models

Analyses in previous sections revealed that speech and head-direction information is useful in identifying the addressee regardless of gender and proxemics between the subjects and the agent. In this section, we will create addressee identification models for different user-agent proxemics, and show that the accuracy of these models is good enough for further implementation.

Based on the data analysis described in section 3.2, we determined six speech features, including average $F_0$, intensity, speech rate, and duration of the utterance, and seven head-direction features for each participant (14 features in all), consisting of time proportion of looking at the partner/agent/elsewhere and head-direction transition, such as agent to partner, agent to elsewhere, and partner to agent. A list of parameters is shown in the Appendix.

We employed an SVM classifier with a linear kernel and C parameter set to 1.0. We used the SMO implementation in Weka. Table 2 shows the results of 10-fold cross-validation for estimating the addressee; agent or partner. We created three models for each proxemics: Speech, Head_direction, and Speech+Head_direction. A Speech model was trained only using speech features, a Head_direction model was trained only using head-direction features, and a Speech+Head_direction model was trained using a full set of features. Since the utterances whose face tracking data were missing were eliminated in the Head_direction model, the number of cases used in models exploiting head-direction information is less than in the Speech model.

Note that in both experimental settings, the Speech+Head_direction model always performed better than the other two models. The accuracy of the Speech+Head_direction model for Exp 1 was 80.3%, and the F-measure values were 0.80 and 0.81 for the agent and the partner, respectively. In Exp 2, the accuracy of the Speech+Head_direction model was 85.1%, and the F-measure values were 0.86 and 0.85 for the agent and the partner, respectively. We also created general models by combining the data in Exp 1 and 2. Among the general models, the Speech+Head_direction model still performs the best; its accuracy was 83.4%, and the F-measure values were 0.84 and 0.83 for the agent and the partner, respectively. Thus, both the results for addressee identification shown here and the data analysis in section 3.2 suggest that both speech and head-direction information contributes to identifying the addressee, and models learned from the full set of features performed the best.

**Table 2. Model evaluation using cross-validation**

| | | | Speech | Head direction | Speech + Head direction |
|---|---|---|---|---|---|
| Exp1 | # of utterances | | 1,830 | 1,237 | 1,237 |
| | F-measure | Agent | 0.717 | 0.759 | 0.799 |
| | | Partner | 0.781 | 0.656 | 0.806 |
| | Accuracy | | 75.3% | 71.6% | 80.3% |
| Exp2 | # of utterances | | 1071 | 953 | 953 |
| | F-measure | Agent | 0.787 | 0.805 | 0.857 |
| | | Partner | 0.823 | 0.744 | 0.845 |
| | Accuracy | | 80.7% | 77.9% | 85.1% |
| General | # of utterances | | | | |
| | F-measure | Agent | 0.736 | 0.769 | 0.836 |
| | | Partner | 0.789 | 0.691 | 0.832 |
| | Accuracy | | 76.5% | 73.5% | 83.4% |

**Table 3. Evaluation of Speech+Head_direction models using test data**

| | | Exp1 model tested by Exp2 data | Exp2 model tested by Exp1 data | General model (split) |
|---|---|---|---|---|
| F-measure | Agent | 0.732 | 0.795 | 0.829 |
| | Partner | 0.779 | 0.71 | 0.826 |
| Accuracy | | 75.8% | 76.0% | 82.8% |

To test the compatibility between the models for Exp 1 and 2, we conducted cross evaluation by applying the Speech+Head_direction model in Exp 1 to Exp 2 data and vice versa. The accuracy dropped from 80.3% to 75.8% in applying Exp 1 model to Exp 2 data, and it dropped from 85.1% to 76.0% in applying Exp 2 model to Exp 1 data. On the other hand, when we created a general model using the first half of the Exp 1 and Exp 2 data and tested the model using the second half of the data, the model worked well with the test data. There was no degradation in accuracy (the accuracy remained around 83.0%). These results suggest that the speech and head direction features exploited in this study are commonly useful in both proxemics, but the models are not always compatible with each other. Therefore, it would be better to create models that are applicable to both proxemics by combining the data from both experiments. Note that this does not mean that the models for different proxemics are not useful at all. Their accuracy was still over 75% in the other experimental settings, and in some cases, they were better than the original models that use only speech features or head-direction features. Thus, they are still useful as substitute models.

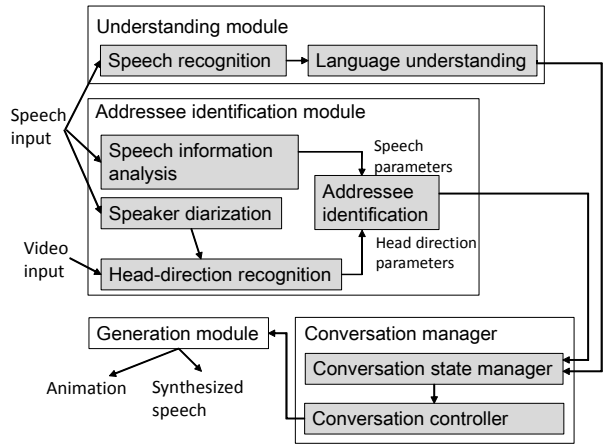


**Figure 2. System architecture**

**Table 4. Models implemented in the system**

| Model name | Number of parameters | Applicable condition |
|---|---|---|
| Speech + Head_direction | Speech: 6<br>Head: 14 | Obtain all the parameter values |
| Speech | Speech:6<br>Head:0 | Fail in measuring head direction |
| Power + Speech_duration + Head_direction | Speech: 3<br>Head: 14 | Fail in measuring $F_0$ and speech rate |
| Power + Speech_duration | Speech: 3<br>Head: 0 | Fail in measuring $F_0$, speech rate, and head direction |

# 4. MULTIPARTY CONVERSATIONAL AGENT

We implemented a multiparty conversation system, including an addressee identification mechanism based on the model proposed in the previous section. The system architecture is shown in Figure 2. The details of each component will be described in the following subsections.

## 4.1 Addressee Identification Module

This module consists of the following four sub-modules.

**(1) Speaker diarization**: In order to identify who is speaking, we use a microphone array installed in Microsoft Kinect. If the microphone array detects speech sound from the right direction, then the system judges that the right user is speaking and vice versa.

**(2) Speech information analysis**: Speech information analysis calculates speech feature values. We use Praat scripts for this purpose, and compute the average $F_0$, intensity, and speech rate for a given utterance. The parameter values are then sent to the addressee identification module.

**(3) Head direction recognition**: When this module receives outputs from face-tracking software (FaceAPI), it estimates the head direction: front, right, left. For this purpose, we created a decision tree using a J48 program, which is an implementation of the C4.5 algorithm in Weka. The decision tree estimates the head direction from the head position and rotation data from the face-tracking software. The details of head-direction recognition were described in our previous study [11]. Once the head direction is estimated, the head-direction feature values for addressee identification are calculated for a given speech interval.

**(4) Addressee identification**: This module receives speech parameter values from the speech information analysis component, and head-direction parameter values from the head-direction recognition component. Then, it applies these data to the addressee identification models to identify the addressee.

However, in real-time processing, input devices do not always work perfectly, and part of the speech and/or head-direction data cannot be measured. In such cases, some of the parameter values are missing. For example, once face-tracking software loses the face, when the user moves his or her head rapidly, it stops tracking, and it takes some time to grab the face again. During the time that the face-tracking data are not being measured, the head-direction parameters cannot be calculated. For speech parameters, pitch ($F_0$) and speech rate sometimes cannot be obtained, but speech power is more stable information. To keep the addressee identification mechanism working even if some of the parameter values are missing, we set up four models and switched the models depending on the parameter values obtained for a given speech interval. The models and the applicable conditions are shown in Table 4. All these models were created by SVM, as described in section 3.3 (the Speech+Head_direction model and Speech model are identical to the general models shown in Table 2). When $F_0$ and speech rate cannot be measured, (s1), (s3) (The definition of each parameter is shown in Appendix), and (s5) parameters cannot be obtained. In such a case, the Power+Speech_duration+Head_direction model is applied, using parameter values for the speech power ((s2) and (s6)), speech duration (s4), and head direction (f1)-(f7). This switching mechanism enables stable addressee identification, contributes to better performance and improvement in robustness of a multiparty conversation system.

## 4.2 Understanding Module

**(1) Speech recognition**: The speech interval obtained by the speech information analysis module is sent to an automatic speech recognition (ASR) system. We employ the Google speech recognition server as our ASR engine. The speech audio file is sent to the server, and the server returns the recognition results in JASON (JavaScript Object Notation) format. The system extracts keywords from the recognition results. The keywords were determined using the following process. First, from the conversation corpus collected in section 3.1, we picked up utterances addressed to the agent. Then, the transcriptions of these utterances were analyzed using a part-of-speech (POS) tagger to assign POS tags to each word. From the POS-tagging results, noun, verb, adjective, and interjection were determined as keywords.

Once the speech recognition component receives 5-best recognition results from the ASR server, it picks up keywords from each recognition result by referring to the keyword list


(a) (b) (c)

**Figure 3. System running in real time. (a) interaction between subject pair and agent, (b) face tracking for each subject, (c) snapshot of part of information state**

created above, and represents each result as a keyword vector. Each keyword in the vectors has a weight that is determined based on the likelihood of recognition results. The input keyword vector is sent to the language understanding component.

**(2) Language understanding**: We employ similarity-based language understanding. First, user utterances collected in section 3.1 are used as possible user utterances. Each utterance candidate is represented as a keyword vector. Keywords in the candidate vectors are also weighted. We calculated the tf-idf value of each keyword and used it as the weight of the keyword. The similarity between each candidate and the input keyword vector obtained from the speech recognition component is calculated using the vector-space model, and a candidate with the highest cosine similarity is chosen as the result of language understanding. We also set a threshold for the similarity judgment. If the input vector does not have enough similarity to any of the possible user utterances (no candidate exceeds the threshold), the language understanding fails. Finally, once a candidate with the best similarity score is determined, a semantic representation for that utterance is produced and sent to the conversation manager.

## 4.3 Conversation Manager

The conversation manager consists of two components: the conversation state manager and the conversation controller. The conversation state manager maintains and updates the conversation state, and the conversation controller determines the agent's next action to respond to the user.

**(1) Conversation state manager**: As the information state of the conversation, the conversation state manager maintains ten kinds of information, including time stamp, current and previous states of the conversation controller, number of users, utterance ID, result of the understanding module, speaker, addressee, and agent utterance as system output. When a user utterance is detected or the agent utterance is produced, information in the conversation state is updated.

**(2) Conversation controller**: The conversation controller determines the next agent action using a state transition model. Our state transition model was implanted using an extended version of GECA Scenario Markup Language (GSML) [12], which shares the basic ideas of Artificial Intelligence Markup Language (AIML).

The conversation controller refers the semantic representation of the user's utterance and the addressee identification result stored in the information state. When the addressee of the current user speech is estimated to be the agent, the state in the transition model is shifted according to the meaning of the user's utterance. Then, the next agent utterance is determined based on the state. In the current implementation of the conversation controller, the state transition is triggered by the user's utterance to the agent, so that the agent responds to the user only when the utterance is addressed to the agent. Once the next agent utterance is determined, it is sent to the generation module.

## 4.4 Generation Module

In the generation module, the utterance content determined in the conversation controller is realized as surface language expressions using a simple template generation technique, and then a sequence of words is sent to the animation module. In the animation module, the agent character animation is produced using visageSDK [2]. The word string is also sent to a TTS to produce synthesized voice.

Figure 3 shows the system running in real time. Figure 3(a) shows a pair of users interacting with the agent, and the two pictures in (b) are the monitoring windows for face tracking. Figure 3(c) is a snapshot of the information state that is updated whenever a user utterance or the result of addressee identification is perceived.

# 5. EVALUATION EXPERIMENT

We conducted an evaluation experiment to examine the effectiveness of the addressee identification mechanism in our autonomous multiparty conversation agent.

## 5.1 Procedure and Conditions

We compared the following two systems, in which different addressee identification mechanisms were installed.

**Proposed addressee identification mechanism (Proposed system)**: This is the proposed system implemented in section 4. In this system, the model for judging the addressee is switched according to which information is obtained from the input devices, as described in section 4.1.

**Simple head-orientation-based addressee identification mechanism (Baseline system)**: As a naïve addressee identification mechanism, we implemented a system that determined the addressee based on the time ratio of the head-direction. By comparing the time ratios of looking at the agent and looking at the partner for a given utterance, the gaze target with the largest time ratio was determined as the addressee.

The experiment was conducted using the within-subject design, with eight pairs of subjects participating in the experiment. We used the same travel-planning task as in the data collection

**Table 5. Evaluation of system performance**

| | Speech to the agent | | Speech to the partner | |
|---|---|---|---|---|
| | Proposed | Baseline | Proposed | Baseline |
| Precision | 0.88 | 0.84 | 0.56 | 0.51 |
| Recall | 0.76 | 0.85 | 0.70 | 0.46 |
| F-measure | 0.80 | 0.83 | 0.56 | 0.44 |

experiment in section 3.1. The task of the subjects was also the same except for the number of places to choose from. In this experiment, the subjects chose one place from eight historical places. The subjects were instructed to ask the agent about the places if necessary. In addition to the contents for the Kyoto area, we developed new contents for Nara, another historical city in Japan. Thus, we had four combinations in this experiment (2 experimental conditions (Proposed system or Baseline system) × 2 conversation contents (Kyoto or Nara)). Each subject pair was randomly assigned to one of these combinations to cancel out the bias caused by the order and the contents.

As shown in Figure 3, the agent was displayed on a 57-inch display, and the subjects were standing about 1.2m away from the display. A web-cam was mounted on top of the display for face tracking, and a microphone for addressee identification and speech recognition was set up between the subjects 50cm away from them.

## 5.2 Results

We collected 349 user utterances for the proposed system and 308 utterances for the baseline system. Addressee-hood was manually judged for these utterances, and was used as the ground truth in evaluating the systems. The evaluation results are shown in Table 5.

For utterances addressed to the agent, the precision rate for the proposed system was 0.88, which was higher than that for the baseline system (0.84). However, the recall rate for the proposed system was lower than the baseline system. Thus, the F-measure of the proposed system (0.80) was slightly worse than the baseline system (0.83).

For utterances addressed to the partner, both the precision and recall rates of the proposed system were higher than those of the baseline system. Especially for the recall rate, the proposed system was 0.70, but in the baseline system, it dropped to 0.46. Moreover, the F-measure for the proposed system was also higher than that for the baseline system. The analysis in section 3.2.2 can account for this result. In both Exp 1 and 2, the time ratio of looking at the partner was lower than 50% when talking to the partner. Similarly, in the evaluation experiment, the subjects did not always look at the partner when talking to the partner, thus, the baseline system that relied only on the head-direction information was more likely to misjudge the addressee.

Compared to the baseline system, the proposed system was more careful in judging the agent as the addressee, and more likely to judge the partner as the addressee. Thus, the precision rate for utterances to the agent was very high (0.88), but the recall rate was lower (0.76). As the proposed system used prosodic features in addressee-hood judgment, the addressee was mistakenly judged as the partner when the speech power was not large enough.

In contrast, the baseline system was more likely to judge the agent as the addressee. This caused serious problems in human-agent interaction because the baseline system interrupted the conversation between the subjects by mistakenly judging an utterance addressed to the partner as one to the agent. In such cases, the conversation was disrupted by the agent's abrupt interruption.

Overall, for the utterances addressed to the agent, the F-measure of the proposed system was 0.80, which is close to the model evaluation result shown in Table 2, indicating that the system performance is good enough. For utterances addressed to the partner, the F-measure was 0.56, which was much worse than the model evaluation result, but this is still much better than the baseline system, especially in recall rate. Thus, the evaluation results indicate that with our addressee identification method installed in an autonomous conversational agent, the system can avoid annoying the users, and makes human-agent multiparty conversations more stable.

# 6. CONCLUSIONS AND FUTURE DIRECTIONS

This study proposed a fully autonomous multiparty conversational agent with an addressee identification mechanism using speech and head-direction information. First, we conducted two WOZ experiments to collect multiparty conversations under two different proxemics between a subject pair and an agent. The results of the data analysis showed that speech and head-direction information is useful in addressee identification regardless of the proxemics. Based on the results, we exploited 20 features, and by applying SVM to all the collected data, we created general models for addressee identification. The accuracy of the models was over 80% in model evaluation, suggesting that the models are good enough to be used in system implementation. Based on the model, we developed an addressee identification mechanism that applied the models to speech information and head direction estimated from face tracking data, and then, incorporated the mechanism into a multiparty conversation system. Finally, we conducted a system evaluation experiment to test whether the proposed addressee identification mechanism performs well in a real-time system. The F-measure for speech addressed to the agent was 0.8 and that addressed to the partner was 0.56. We also found that our system successfully avoided disturbing the conversation by mistakenly taking a turn when the agent is not addressed.

Our addressee identification mechanism needs to be improved, especially for speech addressed to the partner. One promising approach is to consider the meaning of the user's utterance by checking the consistency between the utterance content interpreted by the understanding module and the addressee identification result. Suppose that the user asks about details of a temple after the agent's overview explanation about that place. In such a case, the system should infer that the utterance is addressed to the agent. If the utterance content and the addressee are not consistent, the system should clarify the user's utterance to avoid conversation failure. In addition, it is also necessary to measure more parameters automatically and add them to the models.

The generation side also needs to be improved. In the current system, the agent does not display any gaze or facial expression. It would be preferable if the agent could display floor

management nonverbal signals according to the participation roles: speaker, listener, and side-participant.


## Acknowledgements

This work is partially supported by Grant-in-Aid for Scientific Research (B) 25280076 and Grant-in-Aid for Young Scientists (B) (23700183).


## REFERENCES


[1] *FaceAPI*. http://www.seeingmachines.com/product/faceapi/.

[2] *visageSDK*. http://www.visagetechnologies.com/#&panel1-1.

[3] Akker, R.O.D. and Traum, D. *A Comparison of Addressee Detection Methods for Multiparty Conversations*. in *13th Workshop on the Semantics and Pragmatics of Dialogue*. 2009.

[4] Bakx, I., Turnhout, K.V., and Terken, J. *Facial Orientation During Multi-party Interaction with Information Kiosks*. in *INTERACT'03*. p. 701-704. 2003.

[5] Bohus, D. and Horvitz, E. *Dialog in the Open World: Platform and Applications*. in *ICMI-MLMI'09*. 2009.

[6] Bohus, D. and Horvitz, E. *Facilitating Multiparty Dialog with Gaze, Gesture, and Speech*. in *ICMI-MLMI'10*. 2010.

[7] Bohus, D. and Horvitz, E. *Multiparty Turn Taking in Situated Dialog: Study, Lessons, and Directions*. in *SIGDIAL 2011 Conference (SIGDIAL '11)*. p. 98-109. 2011.

[8] Chen, L. and Harper, M.P. *Multimodal Floor Control Shift Detection*. in *ICMI-MLMI'09*. 2009.

[9] De Jong, N. and Wempe, T., *Praat script to detect syllable nuclei and measure speech rate automatically*. Behavior Research Methods. **41**(2): p. 385 - 390, 2009.

[10] Duncan, S., *Some signals and rules for taking speaking turns in conversations.* Journal of Personality and Social Psychology. **23**(2): p. 283-292, 1972.

[11] Huang, H.-H., Baba, N., and Nakano., Y. *Making a Virtual Conversational Agent be Aware of the Addressee of Users' Utterances in Multi-user Conversation from Nonverbal Information*. in *the 13th International Conference on Multimodal Interaction (ICMI2011)*. 2011.

[12] Huang, H.-H., et al. *The Design of a Generic Framework for Integrating ECA Components*. in *The 7th International Conference of Autonomous Agents and Multiagent Systems (AAMAS 2008)*. p. 128-135. 2008.

[13] Jovanovic, N., Akker, R.O.D., and Nijholt, A. *Addressee Identification in Face-to-Face Meetings*. in *The European Chapter of the ACL (EACL 2006)*. 2006.

[14] Katzenmaier, M., Stiefelhagen, R., and Schultz, T. *Identifying the Addressee in HumanHumanRobot Interactions based on Head Pose and Speech*. in *international Conference on Multimodal Interfaces (ICMI04)*. p. 144--151. 2004.

[15] Lunsford, R. and Oviatt, S. *Human perception of intended addressee during computer-assisted meetings*. in *the 8th international Conference on Multimodal Interfaces (ICMI06)*. p. 20-27. 2006.

[16] Lunsford, R., Oviatt, S., and Arthur, A.M. *Toward Open-Microphone Engagement for Multiparty Interactions*. in *the 8th international Conference on Multimodal Interfaces (ICMI06)*. 2006.

[17] Schlangen, D. *From Reaction to Prediction Experiments with Computational Models of Turn-taking*. in *International Conference on Spoken Language Processing (ICSLP)*. 2006.

[18] Takemae, Y., Otsuka, K., and Mukawa, N. *Video Cut Editing Rule based on Participants' Gaze in Multiparty Conversation*. in *the 11th ACM International Conference on Multimedia*. p. 303-306. 2003.

[19] Terken, J., Joris, I., and Valk, L.D. *Multimodal Cues for Addressee-hood in Triadic Communication with a Human Information Retrieval Agent*. in *International Conference on Multimodal Interfaces (ICMI2007)*. p. 94--101. 2007.

[20] Traum, D., et al. *Multi-party, Multi-issue, Multi-strategy Negotiation for Multi-modal Virtual Agents*. in *the 8th International Conference on Intelligent Virtual Agents (IVA08)*. 2008.

[21] Turnhout, K.V., et al. *Identifying the Intended Addressee in Mixed Human-Human and Human-Computer Interaction from Non-verbal Features*. in *International Conference on Multimodal Interaface (ICMI2005)*. p. 175--182. 2005.

[22] Vertegaal, R., et al. *Eye Gaze Patterns in Conversations: There is More the Conversational Agents than Meets the Eyes*. in *CHI 2001*. p. 301-308. 2001.

[23] Vinyals, O., Bohus, D., and Caruana, R. *Learning Models for Speaker, Addressee and Overlap Detection from Multimodal Streams*. in *14th ACM International Conference on Multimodal Interaction (ICMI'12)*. p. 417-424. 2012.


## Appendix: A list of parameters for addressee identification

| | |
|---|---|
| <Speech features><br>- s1: average $F_0$ of the utterance<br>- s2: average intensity of the utterance<br>- s3: speech rate of the utterance<br>- s4: duration of the utterance<br>- s5: difference between s1 and the average $F_0$ for all the subjects of the same gender<br>- s6: difference between s2 and the average intensity for all the subjects of the same gender<br><Head direction features><br>- f1: ratio of the time the speaker spends looking at the agent to the duration of the current utterance<br>- f2: ratio of the time the speaker spends looking at the partner to the duration of the current utterance. | - f3: ratio of the time the speaker spends looking elsewhere to the duration of the current utterance.<br>- f4: the number of head direction shifts from the agent to the partner in the current utterance (agent->partner).<br>- fh5: the number of head direction shifts from the agent to elsewhere in the current utterance (agent->elsewhere).<br>- f6: the number of head direction shifts from the partner to the agent in the current utterance (partner->agent).<br>- f7: the number of head direction shifts from elsewhere to the agent in the current utterance (elsewhere->agent). |