Detecting Address Estimation Errors from Users' Reactions in Multi-user Agent Conversation

Ryo Hotta, Hung-Hsuan Huang*, Shochi Otogi, and Kyoji Kawagoe

Graduate School of Information Science & Engineering,
Ritsumeikan University, Japan
hhhuang@acm.org

Abstract. Nowadays, embodied conversational agents are gradually getting deployed in real-world applications like the guides in museums or exhibitions. In these applications, it is necessary for the agent to identify the addressee of each user utterance to deliberate appropriate responses in interacting with visitor groups. However, as long as the addressee identification mechanism is not completely correct, the agent makes error in its responses. Once there is an error, the agent's hypothesis collapses and the following decision-making path may go to a totally different direction. We are working on developing the mechanism to detect the error from the users' reactions and the mechanism to recover the error. This paper presents the first step, a method to detect laughing, surprises, and confused facial expressions after the agent's wrong responses. This method is machine learning base with the data (user reactions) collected in a WOZ (Wizard of Oz) experiment and reached an accuracy over 90%.

Keywords: Multi-party conversation, human-agent interaction, Gaze.

1 Introduction

Various kinds of kiosk information systems are used in public places, such as shopping malls, museums, and visitor centers. The typical situation in which such systems are used is that a group of people stand in front of the kiosk and operate it in order to retrieve the information they request while talking with one another. Therefore, in order to implement conversational agents that can serve as an information kiosk in public places, multi-party conversation functionality for simultaneous interaction with multiple users is indispensable. In dyadic dialogues where only two participants are involved, it can be assumed that in most cases, one participant is the addressee when the other one participant is speaking. In multi-party dialogues, however, distinctions must be made among the roles of the participants, such as speaker, addressee, and listeners standing by. When a person is involved in a human-human-agent triadic conversation,

* Corresponding Author.

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 245–253, 2014.

he/she may speak to the agent or may talk to the other person (his/her partner). When the person speaks to the agent, the agent needs to respond to that utterance. However, when the person speaks to the partner, the agent should not mistakenly respond to that utterance. Therefore, one of the basic functionalities a conversational agent needs in order to engage in multi-party conversation is the ability to identify the addressee of each user utterance.

Based on this need, this paper presents a work that aims to determine the addressee of user utterances in triadic conversations among two users and an agent. In the literature, [1–3], it has been reported that in addition to their explicit verbal utterances, humans use nonverbal signals such as their gaze, nods, and postures to regulate their conversation, e.g., to show their intention to yield or willingness to take turns in speaking. If there are specific patterns in the user's gaze behavior that depend upon whom he/she is talking to, it would be possible for the kiosk agent to automatically identify the addressee. Thus, as regards eye-gaze approximation, this study will exploit head direction information obtained from a face-tracking system. It has also been found that in human multi-party conversations and human-robot communication, not just visual cues such as eye gaze and head direction are useful in predicting the addressee, but prosodic cues of the voice as well [4, 5].

In previous stages of this project, we have developed a fully autonomous kiosk agent who can engage with two users at the same time by utilizing non-verbal information only [6–8]. The accuracy of the addressee estimation component was 80%, that means the estimation mechanism has a 20% error rate. Even a human can make such a mistake in multi-party conversation, it can not be expected that the error rate can be reduced to 0%. If the agent decides its actions in responding to the users according to an assumption that all of its perceptions are correct, the conversation afterward will crash and proceed to an unexpected path in the state transition model. Therefore, for further improvement on the system, the mechanisms for detecting and recovering the errors are required. This paper presents the analysis results of users' facial expressions after a wrong estimation of addressee by the agent.

2 Related Work

Research on human communication showed that the eye gaze is an important communication signal in face-to-face conversations. The speaker looks at the addressee to monitor her/his understanding or attitude, and, the addressee looks at the speaker in order to be able to offer positive feedback in return [1, 9]. Eye gaze also plays an important role in turn taking. When yielding his/her turns to speak, the speaker looks at the next speaker at the end of his/her utterances [2]. Vertegaal [10] reported that the gaze is a reliable predictor of addressee-hood. Likewise, Takemae [11] provided evidence that the speaker's gaze indicates addressee-hood and plays a regulatory role in turn management.

Similar results were found in mixed human-human and human-computer conversations in [4, 12, 13]. In these works, perception experiments were conducted

in which subjects guessed who the addressee of a given utterance was. It was found that prosodic and visual cues were about equally effective, and that the combination of auditory and visual cues resulted in better performance. Moreover, the motivation of [5] was quite similar to that of this study, in that researchers proposed a method of identifying the addressee in a human-human-robot interaction by combining prosodic and visual cues. As a visual cue, they used the horizontal head orientation to distinguish addressees. They reported that in 35% of the cases, a person talked to the other human while looking at the robot. They then addressed the fact that visual cues alone might not be sufficient, and proposed the further incorporation of prosodic cues. As prosodic cues, they identified a number of linguistic features obtained from the speech recognition system, such as sentence length, typical phrases, and language models, and used them to distinguish the addressee from the other human. They reported that the speech addressed to the robot was detected with an F-measure of 0.72.

In this study, we share a similar motive but tackle the problem using a different approach. First, to estimate the head direction, we add more parameters, namely the position and rotation of the head. We also focus on shifts in head direction during an utterance. As for the prosodic cues, while [5]focused on linguistic features, we assume that the user's tone of voice may be different depending on whether he/she is speaking to the agent or to the partner. Thus, to measure the user's tone of voice, this study focuses on pitch, intensity (volume), and the rate of speech as the most important prosodic features [14]. It has already been found that prosodic features are useful in the recognition of emotion, and can thus be expected to be useful in characterizing the tone of voice. Considering all these aspects discussed in previous studies, this study employs machine learning techniques to create an automatic classifier for estimating the addressee via the integration of visual and prosodic information.

However, in our current prototype, after the decision making component receive the output of the addressee estimation component, it can not confirm whether the estimation is correct or not. It can only assume that the estimation is correct and go proceed the decision making. We then go forward to the next step of addressee estimation to deal with the situation when there was an error in this estimation. The idea is to get the reactions (feedbacks) from the users in a short period after the agent takes its action in responding to a user utterance. Follow the basic ideas of addressee estimation, we focus in using nonverbal feedbacks from the user, facial expressions and prosodic information.

3 Corpus Collecting Experiment

Regarding to addressee estimation, the possible errors are defined as the following situations:

Unexpected Response (UR): the addressee of a user utterance should be another user, but the agent mistakenly responded.

No Response (NR): the addressee should be the agent, but the agent did not response.

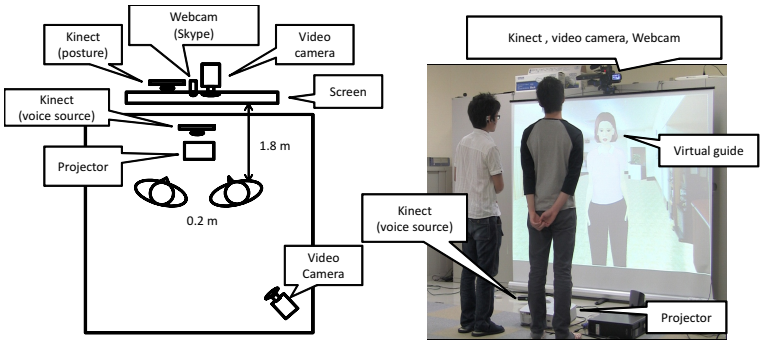

Fig. 1. Setup of the WOZ experiment to collect data for the interaction corpus

In order to collect users' reactions on wrong addressee estimations, a WOZ experiment on three collaborative decision making tasks was conducted. We expect that the subjects' reactions toward the agent may differ to how they talk with a human information provider. To observe the natural interaction with humans and agents, we chose the WOZ experiment setting instead of a human-human one.

Pairs of experiment participants were instructed to interact with a life-size female character on a screen. They had to retrieve information from the character in order to make a decision regarding the given tasks until the agreement between them achieved. As shown in Figure 1, the subjects stood about 1.8 m away from the screen where the character was projected. Two video cameras were used to record the whole experiment, one from the front to take the upper bodies of the participants and the other one takes the whole scene including the participants and the character from the rear. One Webcam was used for the telecommunication software Skype to connect the WOZ operator to the experiment room. The microphone array of one Microsoft Kinect sensor was to identify the voice source (left or right user) of user utterances. The other Kinect sensor was used to record body postures with depth images. The conversation experiment was conducted with the following premises:

- The participants want to make a decision base on their agreement from multiple candidates with the help of the agent who is knowledgeable about that task domain.
- The participants have a rough image of what they want, but they do not have idea about particular candidates in advance.
- The participants discuss on their own and acquire new information from the agent.
- The conversation ends when the participants made the final decision.

A total of 15 pairs of college students were recruited as the participants in the experiment, all of whom were native Japanese speakers. The students came from various departments ranging from economics, life science to engineering at

average age, 19.2. 11 of the all 15 pairs were male ones and the other four were female ones. Each pair was instructed to complete three decision-making tasks: travel planing, lecture registration, and part-time job finding.

Travel Planning: the participants were instructed to pretend to be in a situation where they had a coupon from a travel agency that allows them to visit three of 14 sightseeing spots in Kyushu for free. The information, which includes a brief history, highlights, nearby restaurants, for each location was defined in advance. The sightseeing spots were selected from four of all seven prefectures inside Kyushu. The participants were instructed to complete their task by freely retrieving information from the travel agent, and to discuss their decisions on their own.

Lecture Registration: the participants were instructed to choose three out of 12 lectures to attend together in the next semester. The information about the lecturer, textbook, course difficulty, prerequisites, etc. for each lecture was defined in advance. The lectures were divided into four categories: information science, engineering, languages and communication, and social science. The subjects could freely ask the "tutor" agent for any information about the lectures or the agent itself, and then discuss this on their own in order to make the final decision.

Part-Time Job Hunting: the participants were instructed to request help in choosing three out of 14 part-time jobs to work together near the university. The information about the salary, location, workload, work type, etc. for each job was defined in advance. The jobs were divided into four categories: convenient stores, book shops, restaurants, and gas stations. The subjects could freely ask the agent for any information about the part-time jobs or the agent itself, and then discuss this on their own in order to make the final decision.

These tasks were chosen because the student participants are supposed to be familiar with these issues. In order to stimulate more active discussion, the participants were instructed to make rankings on the three final choices. All participant pairs were assigned to take all of the three tasks in three separate sessions, one task for one session. The sessions were conducted in all possible orders to cancel order effects. One student who major in computer science was recruited to operate the WOZ agent. He was chosen due to his familiarity with operating a GUI-based WOZ application, which ensured that there would be smooth interaction. The operator was asked to practice on the WOZ user interface for two hours prior to the experiment to further ensure that the agent's response time was quick enough. All the sentences that the agent could speak during the experiment were listed in a menu where relevant sentences were grouped for the WOZ operator to select from more easily. There was also a text field that allowed the operator to type arbitrary utterances, in the cases when they were needed but were not defined. The WOZ operator was instructed to try to end the interaction sessions in ten minutes, if possible.

The order of the task for each session was changed to achieve counter-balance, but the wrong responses were intentionally inputed according to the following rules:

- Errors are inputed around every three minutes
- Intentionally make mistakes in the situation when the agent was able to response in the interaction so for
- In the UR situations, the WOZ operator responded as keyword matching manner in simulating the autonomous agent

The wrong responses were only inputed in the third session to allow the user to get used to the conversation with a CG agent, to have time to approximate the agent's ability (100% accurate in the first two sessions), and to notice the errors more easily.

4 Features for Detecting Addressee Estimation Errors

The assumption of the error detection is, users should have some emotional reactions to the agent's errors. For example, the two cases: the users may feel surprised or funny if the agent responded to an utterance that is should not do; the users may feel confused if the agent should answer a user question but it did not, can be considered. The preliminary analysis on four groups was focused on the facial expressions of the users' reactions: laughed, confused, surprised comparing to neutral. Table 2 shows the results of facial expression annotation. The results showed that the users had high possibility to change their facial expressions after an error within five seconds (29 times among 40 error instances). Table 1 shows the relationship between each facial expression and the errors.

Table 1. Relationship between each facial expression and intentionally triggered errors in the experiment

	Neutral	Laughed	Confused	Surprised
Agent did not respond when user speak to it	8	7	13	0
Agent mistakenly responded to an utterance issued to the other user	3	5	1	3
Facial expression changes when there is no error	214	141	12	31
Percentage when there is an error	4.8%	7.8%	53.8%	8.8%

From the annotation process, we had the following three findings:

- The facial expression, confused appears at higher frequency than the other expressions when there are errors

- When one of the users showed surprised facial expression and the last speaker is the other user, there is relatively high possibility that the agent's response was not an error
- When one of the users showed confused facial expression and the last speaker is the agent, there is relatively high possibility that the agent's response was an error

Table 2. Summary of label instances of each subject

Expressions	Max.	Min.	Avg.	Std. Dev.
Neutral	35	9	21.62	7.06
Laughed	24	7	15.37	4.87
Confused	5	1	2.50	1.58
Surprised	7	0	2.75	2.04

We then used hand labeled data and FACS Action Units [15] recognized by visage|SDK [16] as the feature set (Table 3) to train a random forest with Weka [17]. The 10-fold cross-validation accuracy was over 90%. The results of the classification for each facial expression base on the proposed feature set is shown in Table 4. The detection of related facial expression itself is possible, but it is difficult to detect the agent's error merely by facial expressions. This is due to the fact that the users have frequent facial expression changes even when there is no error. The facial expression detection itself works well and should contribute to the detection of errors. However, since the users also make these facial expressions when there is no error, errors can not be detected merely by facial expressions, other features are required.

Table 3. Facial and head movement features used in classifying the facial expressions

Nose wrinkler (AU9)	Lip corner depressor (AU13/15)	Rotate eyes down (AU64)
Jaw drop (AU26)	Outer brow raiser (AU2)	Position distance
Lower lip drop (AU16)	Inner brows raiser (AU1)	Rotation distance
Upper lip raiser (AU10)	Brow lowerer (AU4)	Rotation (angle)
Lip stretcher (AU20)	Rotate eyes (AU61/62)	

5 Conclusion and Future Work

It is necessary for the agent to identify the addressee of each user utterance to deliberate appropriate responses in interacting with multiple users. However,

Table 4. Classification results of each facial expression

Expressions	Precision	Recall	F value
Neutral	0.866	0.866	0.866
Laughed	0.887	0.890	0.888
Confused	0.962	0.960	0.961
Surprised	0.947	0.944	0.945
10-fold cross validation	90.90%		

the agent can not confirm whether the estimation is correct or not. This paper presents a work on the mechanism to detect the error from the users' reactions. The first step, a method to detect laughing, surprises, and confused facial expressions after the agent's wrong responses. This method is machine learning base with the data (user reactions) collected in a WOZ (Wizard of Oz) experiment and reached an accuracy over 90%. From the analysis results, errors can not be detected merely by facial expressions, other features are required.

As future works, we are analyzing the situations when the addressee estimation component is prone to make errors. Also, we would like to consider other modalities like voice features or postures to improve the accuracy.

References

1. Kendon, A.: Some functions of gaze direction in social interaction. Acta Psychologica 26, 22–63 (1967)
2. Duncan, S.: Some signals and rules for taking speaking turns in conversations. Journal of Personality and Psychology 23(2), 283–292 (1972)
3. Sacks, H., Schegloff, E.A., Jefferson, G.: A simplest systematics for the organization of turn-taking for conversation. Language 50(4), 696–735 (1974)
4. Terken, J., Joris, I., Valk, L.D.: Multimodalcues for addressee-hood in triadic communication with a human information retrieval agent. In: Proceedings of the 9th International Conference on Multimodal Interfaces, ICMI 2007 (2007)
5. Katzenmaier, M., Stiefelhagen, R., Schultz, T.: Identifying the addressee in human-human-robot interactions based on head pose and speech. In: Proceedings of the 6th International Conference on Multimodal Interfaces, ICM 2004 (2004)
6. Huang, H.H., Baba, N., Nakano, Y.: Making virtual conversational agent aware of the addressee of users' utterances in multi-user conversation from nonverbal information. In: 13th International Conference on Multimodal Interaction (ICMI 2011), pp. 401–408 (2011)
7. Baba, N., Huang, H.H., Nakano, Y.: Addressee identification for human-human-agent multiparty conversations in different proxemics. In: 14th International Conference on Multimodal Interaction (ICMI 2012), 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction: Eye Gaze and Multimodality (2012)
8. Nakano, Y., Baba, N., Huang, H.H., Hayashi, Y.: Implementation and evaluation of multimodal addressee identification mechanism for multiparty conversation systems. In: 15th International Conference on Multimodal Interaction, ICMI 2013 (2013)

9. Argyle, M., Cook, M.: Gaze and Mutual Gaze. Cambridge University Press (1976)
10. Vertegaal, R., Slagter, R., van der Veer, G., Nijholt, A.: Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 301–308 (2001)
11. Takemae, Y., Otsuka, K., Mukawa, N.: Video cut editing rule based on participants' gaze in multiparty conversation. In: 11th ACM International Conference on Multimedia (2003)
12. Lunsford, R., Oviatt, S.: Human perception of intended addressee during computer-assisted meetings. In: Proceedings of the 8th International Conference on Multimodal Interfaces, ICMI 2006, pp. 20–27. ACM, New York (2006)
13. Dowding, J., Alena, R., Clancey, W.J., Sierhuis, M., Graham, J.: Are you talking to me? dialogue systems supporting mixed teams of humans and robots. In: AAAI Fall Symposium (2006)
14. Rodriguez, H., Beck, D., Lind, D., Lok, B.: Audio analysis of human/virtual-human interaction. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 154–161. Springer, Heidelberg (2008)
15. Ekman, P., Friesen, W.V., Hager, J.C.: Facial action coding system (facs). Website (2002)
16. Visage Technologies AB: Visage|SDK. Website (2008), `http://www.visagetechnologies.com`
17. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. ACM SIGKDD Explorations 11(1), 11–18 (2009)