# A multimodal multiparty human-robot dialogue corpus for real world interaction

# A Multimodal Multiparty Human-Robot Dialogue Corpus for Real World Interaction

**Kotaro Funakoshi**
Graduate School of Informatics, Kyoto University, funakoshi.k@i.kyoto-u.ac.jp
Honda Research Institute Japan Co., Ltd., funakoshi@jp.honda-ri.com

**Abstract**

We have developed the MPR multimodal dialogue corpus and describe research activities using the corpus aimed for enabling multiparty human-robot verbal communication in real-world settings. While aiming for that as the final goal, the immediate focus of our project and the corpus is non-verbal communication, especially social signal processing by machines as the foundation of human-machine verbal communication. The MPR corpus stores annotated audio-visual recordings of dialogues between one robot and one or multiple (up to tree) participants. The annotations include speech segment, addressee of speech, transcript, interaction state, and, dialogue act types. Our research on multiparty dialogue management, boredom recognition, response obligation recognition, surprise detection and repair detection using the corpus is briefly introduced, and an analysis on repair in multiuser situations is presented. It exhibits richer repair behaviors and demands more sophisticated repair handling by machines.

**Keywords:** human-robot interaction, verbal and non-verbal communication, social signal processing

## 1. Introduction

Although most conventional (spoken) dialogue system research has assumed one-to-one conversations between a user and a machine, one-to-many situations between multiple users and a servicer (a machine) are also common and even predominant in the real world. Therefore, there has been much recent research on how to handle such situations (Bennewitz et al., 2005; Bohus and Horvitz, 2009; Al Moubayed et al., 2012; Matsuyama et al., 2015).

Machines assuming one-to-one conversation can get by with just being reactive to input speech and indifferent to the user's status. However, those assuming one-to-many situations must be more proactive and attentive to users' statuses in order to provide meaningful interactions. For example, a conversational system must identify the addressee of a speaking user, i.e., the system or the other human participant (Nakano et al., 2014). Handling social signals (Vinciarelli et al., 2009) and non-verbal information thus has greater significance in multiparty dialogues.

In light of this background, we designed HALOGEN, a software framework for enabling multimodal human-robot interaction (Funakoshi and Nakano, 2017) (shown in Figure 1). HALOGEN primarily handles non-verbal information (mostly audio and visual) from sensor inputs. The information is handled by multiple sub-modules (detectors/recognizers) and integrated by the core module, which oversees not only the final information integration but also mutual communication with the dialogue manager such as (Nakano et al., 2011). It also manages user profile information collected from both the non-verbal and verbal information. Social information such as name, gender, and occupation can be obtained or confirmed through dialogue, and the dialogue manager passes such information to HALOGEN. The dialogue manager can query HALOGEN about both user profiles and user statuses to achieve better verbal communication. HALOGEN can use the confirmed information from the dialogue manager in order to refine user status estimation.

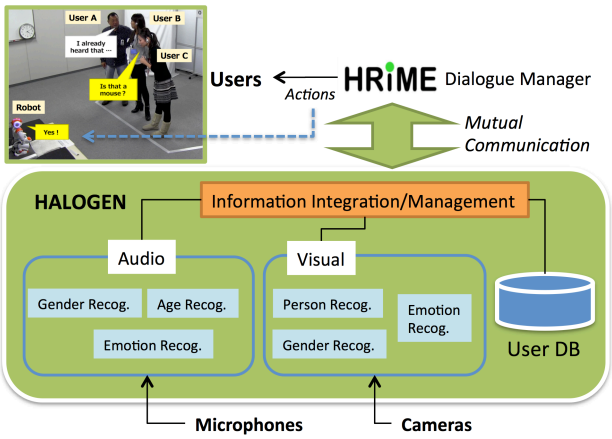


Figure 1: HALOGEN framework.

To implement and evaluate HALOGEN, we have collected sessions of human-robot interaction between a robot and multiple users. The data was collected in a situation between the public and laboratory settings described in (Al Moubayed et al., 2012), although the participants were somewhat controlled in the laboratory by the experimenter. Sections 2. and 3. of this paper describe the recording of the audio-visual data and the annotation, respectively, which form the core of the MPR (Multi-Party Robot) corpus. In Section 4., we introduce several research topics using the corpus. We conclude in Section 5. with a brief summary.

## 2. Data Recording

We recorded human-robot interaction data twice, once in 2012 and once in 2016. These two recordings were done in mostly similar settings but with a few differences in terms of participants, recording environment, and interaction design. We refer to them as MPR2012 and MPR2016.

The two main differences between the two editions are (1) the specifications of the visual recording devices and (2) the robot operations (full manual operation only in 2012; both

full manual and full automatic operations in 2016).

## 2.1. MPR2012

We recruited 30 trios of individuals through a research-support agency. The three participants in each trio were friends or family and ranged in age from their 20s to 60s. The genders of the participants were balanced.

### 2.1.1. Situation and recording settings

Each trio participated in two 25-minute interaction sessions with a robot in which they repeatedly engaged in a conversational game. They were instructed that the autonomous robot was under development and that they should be tolerant of errors. After the sessions, they were informed that the robot was controlled by a human operator.

The robot spoke English only, as it was explained that the robot was designed for English learning purposes. The participants were allowed to speak either English or Japanese. The interaction setting is shown Figure 2, and the upper-left picture in Figure 1 shows a recording scene in this setting. The participants started the sessions from the waiting spaces. They came into the interaction field and went back to one of the two waiting spaces in accordance with the instructions from the director (experimenter), who stayed outside the laboratory. The interaction field was indicated by lines so that they would stay in the proper shooting area. The instructions included *participating in the game, observing the game, passing through the field*, and *returning to one of the waiting spaces*. Each participant in a session stayed in the field for about 15 minutes in total.

In the waiting spaces, the participants stayed quiet while listening to music provided through headphones so that they could not sense what was going on in the field. Throughout each session, they were equipped with a transceiver and an earphone. The instructions from the director were sent only to the earphone of the relevant participant. The idea here was to create information imbalance among the human participants that would result in frequent communications among them.

The operator watched the situation and the participants by means of a NAO robot[1] with a camera installed in its head and a static birds-eye-view camcorder, which was also used to record the sessions for annotation. The operator controlled the direction of the robot's head and hand gestures according to the interaction. The possible utterances of the robot were fixed and prepared as buttons in the GUI interface for the operator. The sessions were recorded with Microsoft Kinect v1 and four omni-directional distant microphones behind the robot.[2]

### 2.1.2. Interaction games

Each trio engaged in the 20 Questions game for their first session. In this game, the robot as game master secretly chooses a target object *tiger, sushi, cell phone, etc.* The participants can ask the robot about an attribute of the object as a yes-no question or make a guess up to 20 times. If they make a correct guess within 20 tries, they win; otherwise the robot wins. Although they were instructed about

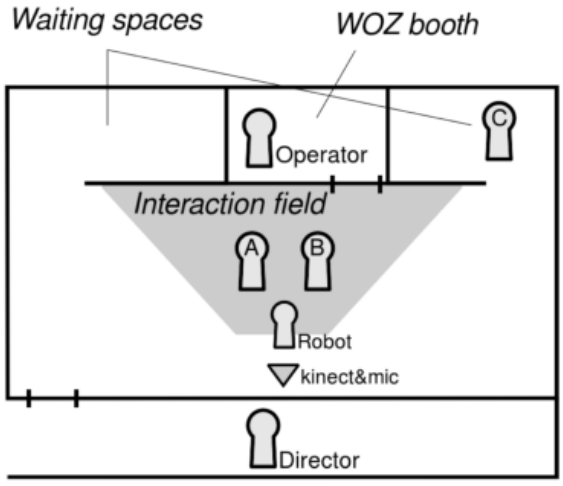


Figure 2: Data recording settings used for MPR2012.

the game before starting the session, the robot explained the rules of the game after initial greetings and engagement phases (inviting participants to the game). When one game ended, the robot immediately started another game.

In their second session, participants played a gesture mimicking game. First, each participant was taught gestures corresponding to various English words. When more than one participant was ready in the field, the robot started a mimicking speed competition in which it said a word and participants had to make the corresponding gesture as fast as they could. The robot judged the answers and declared as the winner whoever made the correct gesture fastest.

## 2.2. MPR2016

The participant population and statistics are nearly identical to MPR2012, with gender-balanced 90 participants divided into 30 trios.

### 2.2.1. Situation and recording settings

Figure 3 shows a data collection scene in MPR2016, where the setting was almost equal to that of MPR2012 except that the waiting spaces and the operator booth were not behind the interaction field. In this edition, we adopted a prototype spoken dialogue system for the second sessions. A manually operated robot was used for the first sessions, the same as MPR2012.

The Kinect device was upgraded to version 2. This brought us (1) higher image resolution (HD), (2) better skeletal tracking performance, and (3) synchronized recording of video and audio in one Kinect data file. The other equipment used in the recording was the same.

The instructions given to the participants and the director's role were also almost the same as in MPR2012. This time, however, the participants were informed in advance that the robot was operated by a human in the first session and by a system in the second, as the difference in interaction quality between a human operator and the system was significant. The director also had another task, namely, to insert attention-drawing events into the recording sessions in order to collect preliminary data for the surprise detection

[1] http://www.aldebaran.com/en/cool-robots/nao

[2] Kinect v1 could not record audio by itself.

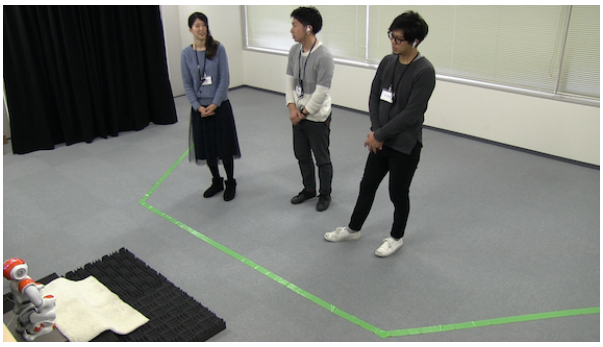

Figure 3: Data recording scene in MPR2016.

research (discussed in Section 4.4.)[3]
For the first 10 trios, the robot spoke English, the same as it did for MPR2012. For the remaining 20 trios, the dialogue system was brushed up based on the experiences and data from the first 10 trios, and it was also modified to speak Japanese, as some participants could not quite catch it when the robot suddenly uttered an irrelevant or irrational message in English to induce surprised reactions.

#### 2.2.2. Interaction games

The game used for the first sessions was the same as in MPR2012, i.e., 20 Questions. In the second sessions we used a simplified version of 20 Questions, as it was hard for the system to correctly answer participants' arbitrary yes-no questions with regard to the chosen object. In the simplified version, participants could only make a guess or ask for a hint.

## 3. Annotation

### 3.1. MPR2012

The recorded data was annotated using ELAN (Brugman and Russel, 2004) with regard to the following information for each participant.

**Participation status** One of the following labels was placed over the timeline whenever a given participant was visible: *participating* (maintaining interaction with the robot), *observing* (staying in the field without interacting with the robot), and *passing* (leaving from the field or transiting behind the field).

**Gaze** Based on the head directions, the targets of attention were labeled as the combination of the three participants and the robot, That is, each label was a combination of the names of the participants (A, B, and C) and the robot (NAO). If the gaze at the moment could not be estimated, an *invalid* label was given.

**Utterance segment and addressee** The speech segments of the participants and the robot were annotated by segmenting speech with pauses over 400 ms or sentential boundaries. Each speech segment was labeled with its addressee(s) in the same way as gaze was. Otherwise, it was labeled as *laugher* or *monologue*.

**Transcript** Each speech segment was transcribed. These transcripts contain a mixture of English and Japanese.

**Dialogue act** Each speech segment of 39 sessions out of 60 (9 first sessions and 30 second sessions) was labeled with one of the dialogue act types based on DIT++ (Bunt, 2009). The dialogue types contain some domain-specific labels such as *Quiz-Challenge* (making a guess) and *Quiz-Judge* (evaluating the guess).

### 3.2. MPR2016

The recorded data was annotated using ELAN with regard to participation status, utterance segment and addressee, and transcript in the same manner as MPR2012. Gaze and dialogue act have not been annotated thus far.

## 4. Research Using the MPR Corpus

In this section, our studies using the MPR corpus are introduced.

### 4.1. Multiparty Dialogue Management

Dialogue management entails tracking the the current dialogue state based on the given context, i.e., present situation and past discourse, and deciding the system's next action, that is, *what to speak*. In addition to *what to speak*, a dialogue system in a multiparty situation has to consider seriously *who to speak to* and *when to speak*.
Using the dialogue act annotation in MPR2012 as a basis, (Kennington et al., 2014) proposed and preliminarily evaluated a model for multiparty dialogue management that manages a multiparty situation as a bundle of one-to-one dialogues while suppressing conflicting or redundant actions at a pre-output action manager. The action manager is also responsible for *who to speak to* and *when to speak*. By this means, the model can flexibly handle an arbitrary number of participants.

### 4.2. Boredom Detection

To achieve a long-term relationship between users and a dialogue system, it is important to ensure that users maintain a willingness to continue using the system (Funakoshi et al., 2010). Detecting boredom in users is a key technology to maintain such willingness or motivation. (Shibasaki et al., 2017) annotated the first session's data of MPR2012 with regard to boredom based on the intuitive sense of two annotators, and proposed a detection model using the body motion of participants, the relationship between their face directions and standing positions, and the information obtained from participation statuses.

### 4.3. Response Obligation Estimation

Response obligation is whether the system has to respond to input sound or not. Even when a speech input from a user is directed to the system, it does not always mean that the system has to respond to it immediately. Moreover, in a multiparty situation, the system should not respond to a speech input that is directed from one user to another user.

[3] These events were mostly sounds such as breaking glass, the meow of a cat, a stamping sound made by a participant in the waiting space (he/she was instructed to do so by the director through the transceiver), etc. Although these sounds were made to draw the attentions of the participants, they were not nearly as loud as a sudden warning message made by the NAO robot when it was overheated.

(Sugiyama et al., 2015) proposed a response obligation estimation method using non-verbal information and evaluated it with MPR2012. The proposed method handles the response obligation estimation problem as the composition of noise rejection, addressee identification, end-of-turn detection, and speaker intention recognition.
The proposed method does not use speech recognition results so as to ensure domain versatility. This feature, however, and the diversity in user behaviors, limits the estimation performance, especially against unseen people in the model training data. To overcome this, currently we are working on error recovery from failures of response obligation estimation, and online adaptation to new users. In the next two subsections, two key elements of the error recovery are discussed.

### 4.4. Surprise Detection

Failures of response obligation estimation can happen in two ways. One is the false-positive case: the system wrongly responds to an irrelevant input sound. In response to this case, users often exhibit surprised reactions. It seems that such a reaction mostly appears as a sudden movement in the head or body, as a repair request such as "Huh?" (Dingemanse et al., 2013), or as an expression of confusion in the face or voice, typically as confused laughter.
To build an effective surprise detection method, we tried to artificially increase surprised reactions of participants in MPR2016. This data is to be examined in future.

### 4.5. Repair Detection

The other type of failure is the false-negative case: the system ignores a user's speech input that it should have responded to. In response to this case, users often try repair, i.e., repeating or rephrasing the previous utterance.
We are currently developing a repair detection method based on previous approaches such as (Cevik et al., 2008). However, all the previous approaches assume one-to-one clean communication. In a real multiparty situation, the problem is not so simple. Given an input sound, it is not obvious for the system to decide the target sound to be compared with the input for repair detection because occationaly noises, monologues, or conversation with other participants are interjected between a repairing utterance and the utterance to be repaired (in our case, the ignored utterance). Here, we discuss an analysis of repair activities on occasions of false-negative errors in response obligation estimation. Ten of the second sessions of MPR2016 were used. This data includes 2,032 utterances from the robot to participants, 2,506 from from participants to the robot, and 934 from a participant to other participant(s). Of the 2,032 robot-directed utterances, 613 are ignored.
Table 1 lists the distribution of user behaviors after a speech directed to the robot was ignored. In 31% of all cases, another participant talked first, and in 46%, the ignored participant made some action. In 22%, the participants did nothing until the robot made a prompting message.
After the robot's ignoring, in the 192 cases, another participant spoke 81 repairs instead of the ignored person. Table 2 shows the breakdown of these 81 cases along with the

Table 1: Participant responses after robot's ignoring.

| Response pattern | Count | Ratio |
|---|---|---|
| Another participant talks | 192 | 31% |
| Shifting to another topic | 150 | 24% |
| Repairing by repeating or rephrasing | 88 | 14% |
| Talking to another participant | 47 | 8% |
| Waiting until robot speaks | 136 | 22% |
| Total | 613 | 100% |

Table 2: Breakdown of repair behaviors by same participant (SP) and by different participant (DP).

| Repair behavior | SP | DP | Sum | Ratio |
|---|---|---|---|---|
| Rephrasing into different words | 49 | 34 | 83 | 49% |
| Repeating the original words | 16 | 33 | 49 | 29% |
| Repeating an extended expression | 13 | 9 | 22 | 13% |
| Repeating a part of the original | 10 | 5 | 15 | 9% |
| Total | 88 | 81 | 169 | 100% |

88 cases where the repair was done by the ignored person. Rephrasing accounts for almost half of the cases. This indicates we should prepare for both repetition and rephrasing. Repetition detection based on Dynamic Time Warping between two speech sounds is expected to be robust against speech recognition errors. However, its performance would be degraded when the speakers are different (as in the DP case in Table 2). It is important to note that a repair utterance may not come immediately after the utterance to be repaired. Indeed, we found that three out of 11 repetition cases in one session contained interjections of one or two irrelevant utterances. We have to build a smarter repair handling that can manage all these issues.
Repair is considered a universal part of language use (Levinson, 2016), but handling repair in spoken dialogue systems is currently quite limited. As discussed above, it seems most previous approaches to repairing are oriented to verbal aspects. It is essential now that non-verbal approaches be studied, too.

## 5. Concluding Remarks

To expand the area in which dialogue systems and conversational machines can function, it is important to make systems capable of handling multiparty situations, where multimodal processing of non-verbal information or social signals is a key component.
We have designed and implemented the HALOGEN framework for multimodal multiparty interaction, and collected roughly 50 hours of audio-visual data on one-to-many human-robot interactions with 180 participants. The data is annotated with speech segment, addressee, transcript, etc. and has been used in several of the studies introduced in this paper. The corpus is not public but is available in research collaboration with Honda Research Institute Japan Co., Ltd.

## 6. Bibliographical References

Al Moubayed, S., Beskow, J., Granström, B., Gustafson, J., Mirnig, N., Skantze, G., and Tscheligi, M. (2012). Furhat goes to robotville: A large-scale multiparty

human-robot interaction data collection in a public space. In *Proc. LREC workshop on multimodal corpora for machine learning*.

Bennewitz, M., Faber, F., Joho, D., Schreiber, M., and Behnke, S. (2005). Towards a humanoid museum guide robot that interacts with multiple persons. In *Proc. Humanoids*, pages 418–424.

Bohus, D. and Horvitz, E. (2009). Models for multiparty engagement in open-world dialog. In *Proc. SIGDIAL*, pages 225–234.

Brugman, H. and Russel, A. (2004). Annotating multimedia/ multi-modal resources with elan. In *Proc. LREC*.

Bunt, H. (2009). The DIT++ taxonomy for functional dialogue markup. In *Proc. EDAML*.

Cevik, M., Weng, F., and Lee, C.-H. (2008). Detection of repetitions in spontaneous speech in dialogue sessions. In *Proc. Interspeech*, pages 471–474.

Dingemanse, M., Torreira, F., and Enfield, N. J. (2013). Is "huh?" a universal word? conversational infrastructure and the convergent evolution of linguistic items. *PLoS ONE*, 8(11).

Funakoshi, K. and Nakano, M. (2017). Online evaluation of response obligation estimation on the halogen multimodal interaction framework. In *Proc. HRI (LBR)*.

Funakoshi, K., Nakano, M., Kobayashi, K., Komatsu, T., and Yamada, S. (2010). Non-humanlike spoken dialogue: A design perspective. In *Proc. SIGDIAL*, pages 176–184.

Kennington, C., Funakoshi, K., Nakano, M., and Takahashi, Y. (2014). Probabilistic multiparty dialogue management for a game master robot. In *Proc. HRI (LBR)*.

Levinson, S. C. (2016). Turn-taking in human communication origins and implications for language processing. *Trends in Cognitive Sciences*, 20(1).

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

Nakano, M., Hasegawa, Y., Funakoshi, K., Takeuchi, J., Torii, T., Nakadai, K., Kanda, N., Komatani, K., Okuno, H. G., and Tsujino, H. (2011). A multi-expert model for dialogue and behavior control of conversational robots and agents. *Knowledge-Based Systems*, 24:248–256.

Nakano, Y., Baba, N., HUANG, H.-H., and Hayashi, Y. (2014). A multiparty conversation system with an addressee identification mechanism based on nonverbal information. *Transactions of the Japanese Society for Artificial Intelligence*, 29(1):69–79.

Shibasaki, Y., Funakoshi, K., and Shinoda, K. (2017). Boredom recognition based on users' spontaneous behaviors in multiparty human-robot interactions. In *Proc. MMM*.

Sugiyama, T., Funakoshi, K., Nakano, M., and Komatani, K. (2015). Estimating response obligation in multi-party human-robot dialogues. In *Proc. Humanoids*.

Vinciarelli, A., Pantic, M., and Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27:1743–1759.