An Incremental Response Policy in an Automatic Word-Game

Eli Pincus & David Traum

USC Institute for Creative Technologies
12015 Waterfront Drive
Playa Vista, Ca 90094, USA

1 Introduction

Turn-taking is an important aspect of human-human and human-computer interaction. Rapid turn-taking is a feature of human-human interaction that is difficult for today's dialogue systems to emulate. For example, typical human-human interactions can involve an original sending interlocutor changing or stopping their speech mid-utterance as a result of overlapping speech from the other interlocutor. The overlapping utterances from the other interlocutor are typically called *barge-in* utterances. An example of this phenomena is seen in the two turns of dialogue in the top half of Figure 1. In this dialogue segment Student A first reveals his test score in the **original utterance**. Student A then begins to tell student B that he had heard Student B got a perfect score. Student B interrupts Student A with a **barge-in utterance** that contains new information (that actually he had not performed well on the test) causing Student A to halt his speech and not finish his original utterance. We call the unspoken part of student A's original utterance Student A's **originally intended utterance**. Student A then makes a decision based on the new information to not say his originally intended utterance. This is likely due to the originally intended utterance no longer being appropriate considering the new information made available to Student A. Student A then makes an intelligent next choice of what to say which can be seen in Student A's **updated utterance** which takes into account the new information contained in Student B's barge-in utterance. In this work we refer to Student A's dialogue move as an **intelligent update**.

In this work we present updates to Mr. Clue, a fully automated embodied dialogue system that plays the role of clue-giver in a word-guessing game. We discuss results from an evaluation that compares a version of the system that employs a **user-initiative barge-in policy**, a policy that allows Mr. Clue to handle user utterances that barge-in on system speech, with a version of the system that flushes all user speech while Mr. Clue is speaking. The results show that a system that can process user-initiative barge-in utterances in this domain results in marginally significant higher game scores and leads players to"skip" or "move on" in the game significantly more often than a system that can not process barge-in utterances. We describe how Mr. Clue's user-initiative barge-in policy moves beyond classical barge-in policies as it can perform intelligent updating that typically takes place in rapid turn-taking in human-human dialogue. An example of this capability can be seen in the bottom half of Figure 1. Here, Mr. Clue halts his clue (for the target-word *"Mountain"*) in his original utterance and performs an intelligent update to confirm to the user that the user's barge-in utterance was a correct guess.

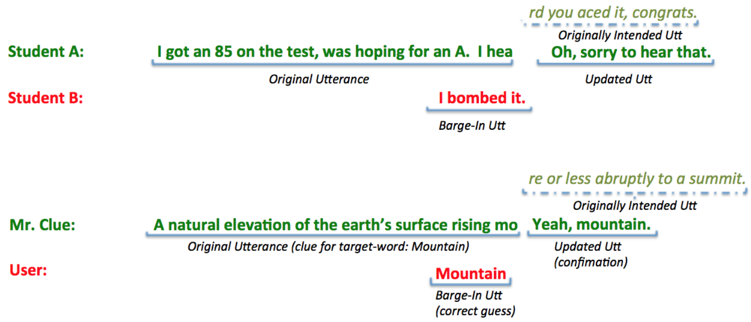


Fig. 1: Example Human-Human & Mr. Clue-Human Intelligent Update Dialogue Move

The rest of the paper is organized as follows. In the next section we discuss foundational work that that our system directly builds on. In Section 3 we outline Mr. Clue's architecture and explore issues of different user initiative barge-in policies that were considered during system design. Section 4 discusses an experiment we conducted to test the subjective and objective impact of endowing Mr. Clue with the ability to process user barge-in utterances compared to a version of Mr. Clue that flushes all user speech while Mr. Clue is speaking. Section 5 discusses the results from this experiment. In Section 6 we briefly put this work in context with other user-initiative barge-in policy models. Finally, in Section 7 we conclude.

2 Foundational Work

The current Mr. Clue is an iteration of an earlier version of the system [1]. The early version of the system was built on components of the Virtual Human Toolkit [4] as well as four additional components: a clue generator that queries databases and scrapes web content, a dialogue manager that governs game flow, a database that stores game actions, and an auxiliary game interface with game information. The virtual human toolkit components include modules for text-to-speech services, automatic speech recognition, and non-verbal behavior generation [5]. The current version of the system still uses the same components as the earlier version of the system as well as the same resources (*wikipedia.com* webpages, *dictionary.com* webpages and *WordNet* [7] database for clue generation. Updates made for this version of the system include generation of a much larger database ($\approx$ twice the size) of clues for a different target-word list. Further, the earlier version of the system did not process user-initiative barge-in utterances as does this version of the system. We were motivated to process user-initiative barge-in utterances via an analysis of the Rapid Dialogue Corpus [2]. This corpus contains video and audio recordings of pairs of humans playing word-guessing games. We examined recordings for 4 different pairs of players who played 24 rounds of the RDG-Phrase game. We calculated that 1 minute of speech from a total of 27 minutes of speech was overlapping (3.7%). While at first glance this seems like a relatively small amount of speech, subjective analysis indicated that the overlapping speech came at critical points that helped forward progress in the game.

3 Mr. Clue Architecture

In this section we discuss the modular architecture of Mr. Clue as it existed for the evaluation described in Section 4. Figure 2 shows the architecture in relation to a user. A user's speech is transformed by an out-of-the-box Automatic Speech Recognition module, the Apple OS X El Capitan's dictation ASR[1] into an ASR Hypothesis. Wrapper code was written to allow the ASR to communicate via the virtual human took-kit messaging protocol, a variant of Active MQ messaging [2]. Partial and final ASR hypotheses are then sent via the toolkit's Active MQ messaging server to the system's dialogue manager which can send response behavior messages back to different components of the virtual human toolkit based on the game-state and interpretation of the user utterance.

We briefly discuss Mr. Clue's non-verbal behavior and text-to-speech modules. The off-the-shelf toolkit non-verbal generator is used with no changes which produces behaviors such as head nodding for affirmative utterances, e.g.- ”yes”, and pointing to self when self referencing, e.g. - ”I”. Based on the TTS evaluation in [6] we use NeoSpeech James' voice [3] which had significantly higher objective and subjective scores than the other synthetic voices. Mr. Clue also has an auxiliary graphical user interface. The GUI displays game information to a user including round #, time left, total score and round score which are all updated in real-time as the game progresses. A screen shot of the GUI and Mr. Clue along with another female avatar that plays the game judge (who serves the role of giving instructions and notifying the user when the end of a round has been reached) can be found in Figure 3. The next sub-section discusses the system's dialogue manager.

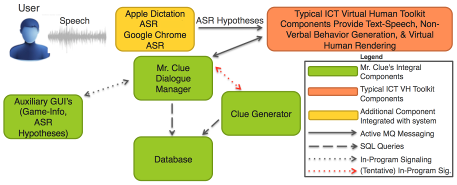
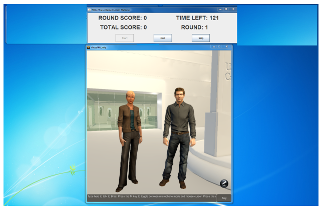

User
Speech
Apple Dictation ASR Google Chrome ASR
ASR Hypotheses
Typical ICT Virtual Human Toolkit Components Provide Text-Speech, Non-Verbal Behavior Generation, & Virtual Human Rendering
Mr. Clue Dialogue Manager
Auxiliary GUI's (Game-Info, ASR Hypotheses)
Clue Generator
Database
Legend
Mr. Clue's Integral Components
Typical ICT VH Toolkit Components
Additional Component Integrated with system
Active MQ Messaging
SQL Queries
In-Program Signaling
(Tentative) In-Program Sig.

Fig. 2: Mr. Clue Architecture


RDG-Phrase Game Current Statistics

ROUND SCORE: 0 TIME LEFT: 121
TOTAL SCORE: 0 ROUND: 1

vhtoolkitUnity

Type here to talk to Brad. Press the M key to toggle between microphone mode and mouse cursor. Press the

Fig. 3: Game Judge (left) Mr. Clue (right)

3.1 Dialogue Manager

An enumeration of Mr. Clue's dialogue policy for the two versions of the system tested in the evaluation described in the next section can be found in Table 1. The first column indicates which mode Mr. Clue is being run in (either the mode that allows for user-initiative barge-in, barge-in mode, or the mode that flushes all user speech during system speech, non-barge-in mode) and which dialogue state Mr. Clue is in (i.e.- whether or not the game has started and whether or not Mr. Clue is talking). The second column indicates a possible utterance the user might say in the corresponding state and mode. The

[1] http://www.apple.com/osx/whatsnew/
[2] http://activemq.apache.org/
[3] http://www.neospeech.com

last column shows how Mr. Clue would respond given the corresponding system mode, game state and user action. During a game-round user utterances are classified as a $< CORRECT\_GUESS >$ (which we refer to as *cg*), an $< INCORRECT\_GUESS >$ which we refer to as *ig*, or a <SKIP>, which we refer to as *sk*. For a more complete list of the types of dialogue moves made in word-guessing games see [3].

For this evaluation, the dialogue acts that are considered user-initiative barge-in acts, interrupt actions, are *cg* and *sk*. These dialogue acts (which are recognized by one key-word) all trigger Mr. Clue (in interrupt mode) to halt all verbal and non-verbal behaviors (assuming he is talking) and take the appropriate next action. Note also that the end of round signal (a signal triggered when round time is over) is a barge-in for both user and the system speech. This signal can be considered a **domain-initiative** barge-in.

Dialogue acts were selected as interrupt actions if they are *game-critical actions* which we define to be actions that would likely cause a human clue-giver to interrupt their current speech and take an updated appropriate action. The algorithm that dictates which partial/final ASR hypothesis message Mr. Clue will respond to follows: If the ASR hypothesis is a word that represents an interrupt action then Mr. Clue processes that partial message immediately otherwise he waits for a final ASR hypothesis message before taking his next action. The assumption in this policy decision was that if an interrupt keyword such as "start" or "skip" is recognized by the ASR (even as a partial message) game performance & perception is improved more by assuming that the partial message reflects the user's intended dialogue act as opposed to classifying the dialogue act based on a higher confidence final ASR hypothesis. To understand how this policy fits into the context of the policies of other dialogue systems capable of doing intelligent updating see Section 6. When Mr. Clue's non-verbal behavior is interrupted all animation stops and he immediately returns to his neutral pose.

Table 1: Dialogue Policy

States	**User Utterance**	**Giver Response**
Non-Game Mode Not Playing	1. Contains "Start" 2. Does not contain "Start"	1. Gives Clue. 2. Asks user to say "start"
Game Mode Non-Interrupt (Mr. Clue not talking)	1. Incorrect Guess 2. Correct Guess 3. "Skip"	1. [Disabuse] then gives a new clue. 2. Gives a confirmation then a clue for a new target. 3. Gives clue for new target.
Game Mode Interrupt (Mr. Clue Talking)	1. Incorrect Guess 2. Correct Guess 3. "Skip"	1. If at least 60% of words said [Disabuse] then gives next clue; else user utterance flushed 2. Interrupts, confirmation, next clue. 3. Interrupts, "new target", next clue.
Game Mode Silence Threshold	1. <Silence >	1. Once reached says next clue.

Analysis of the RDG corpus showed that human clue givers frequently gave additional clues if the guesser was silent for a certain amount of time (most likely assuming no guess indicated the guesser needed more information) as well as if an incorrect guess was said near the end of a clue. This led us to define two parameters in the dialogue manager's policy, a silence threshold s and a give next clue immediately threshold i (which is only used in interrupt mode). If a user says nothing for s seconds after Mr. Clue finishes a clue the next clue for that target-word is given. If at least one *ig* is made after i percent of the current clue has been said and no *cg* is made before the end of the current clue Mr. Clue gives the next clue for the current target-word immediately. s was set to 6 seconds and i to 60% for this evaluation based on rough observations of the timing of these behaviors when performed by human clue-givers that were recorded in the RDG-Corpus. The next sub-section discusses issues of three user-initiative barge-in policies that were considered during system design.

3.2 User-Initiative Barge-in Policy Issues

We considered three different user-initiative barge-in policies for Mr. Clue:

- **Option 1** Flushing all user speech while system speaking (non-barge-in mode)
- **Option 2** Queue user speech while system speaking and make a decision on how to respond after system is finished speaking current utterance
- **Option 3** Interrupt current system utterance (original utterance) and take an intelligent next action which could be just the continuation of the current utterance or continuing with a new updated utterance. (barge-in mode)

The first option (non-barge-in mode), which flushes all user speech while system is speaking has two shortcomings. First, it suffers from the *Time-Waste Issue* - Mr. Clue does not interrupt himself for actions a human clue-giver would interrupt himself for (e.g. - a correct guess) and therefore wastes unnecessary time. Second, it creates a *User-Timing Burden Issue*, i.e. -the burden is on the user to time there speech so that it occurs after the system has finished speaking but before the the next clue is given (in our case when s seconds is reached).

The second policy we considered was to keep track of user speech while the system was talking but to only process certain interrupt actions such as a *cg* or a *sk* after Mr. Clue is finished speaking. This policy introduces a *Priority-Scheme Issue*- i.e. priorities need to be assigned to interrupt actions so that if multiple interrupt actions occur while the system is speaking; there is a decision process for which (or what order) the interrupt actions should be processed. This policy also suffers from the time-waste issue that the first policy suffers from but not the user-timing burden issue.

The third policy (Mr. Clue's current policy in barge-in mode), described in Section 3.1, seems to result in behavior closest to simulating what humans do in real game-play. This policy does not suffer from either the time-waste issue, the user-timing burden issue, or the priority-scheme issue. We next discuss an evaluation that compares the Option 1 and Option 3 user-initiative barge-in policies using the Mr. Clue system as a test-bed.

4 Evaluation Experiment

In this section we discuss an evaluation of Mr. Clue. We examined two main independent variables:

1. **Embodied**: whether or not participants could see Mr. Clue and his non-verbal behavior, including lip synch, gaze and some other gestures.
2. **Barge-in**: whether or not Mr. Clue was run in barge-in mode or non-barge-in mode.

We used a between subjects method for the two variables, since we felt it would be disconcerting to change the user interface without changing other aspects of the agent. We collected data from 52 players that played 2-8 150-second rounds of the word-guessing game with Mr. Clue and then filled out a post-survey containing questions on their subjective evaluations of aspects of Mr. Clue. Af ter each round the game-judge asked the participants to give two ratings on a 1-7 scale in response to the following 2 questions: "How effective did you find the clues in the last round?" & "Other than the clues he gave, how do you think Mr. Clue compares to a human clue-giver?" Players were recruited via Craig's List and paid $25 for their time. Participants saw a monitor displaying a screen similar to one shown in Figure 3. Participants spoke into a wireless Sennheiser microphone which did not pick up speech being output by the computer speakers. Audio files were recorded and all relevant game actions stored in a SQL database. Participants interacted with Mr. Clue in one of 4 conditions. Table 2 shows the conditions, the # of people who interacted with the system in these conditions and the average # of rounds played.

Table 2: Experiment Conditions

Condition	**# of Participants**	**Avg. # of Rounds Played in Condition**
Embodied Barge-In	17	5.9
Embodied Non-Barge-In	15	5.9
Disembodied Barge-In	14	6.3
Disembodied Non-Barge-In	6	8

We had 2 main hypotheses for this experiment. **Hypothesis 1**: The # of utterances recognized by the system for each interrupt dialogue act (i.e. - *cg* and *sk*) would be higher in the barge-in condition vs the non-barge-in condition. Note if evidence is found for hypothesis 1, in particular if there are more *cg* utterances recognized by the system for players in the barge-in condition then players will have higher scores in that condition since each *cg* earns players 1-point. Motivation for hypothesis 1 was found in the fact that in barge-in mode Mr. Clue is able to interrupt himself when he is speaking for interrupt dialogue acts and thus has more time recognize those dialogue acts than when in non-barge-in mode.

Hypothesis 2: Subjective evaluations of the game would be highest in terms of enjoyment (**Hypothesis 2.1**) for the barge-in condition vs the non-barge-in condition. We believed players would be frustrated when their speech was ignored during agent speech decreasing their enjoyment of the game. Subjective evaluations in terms of naturalness for the voice (**Hypothesis 2.2**) would be higher in the embodied condition vs the non-embodied condition. We thought a disembodied voice would be disconcerting and non-human like for players. Finally, subjective evaluations in terms of naturalness for the clue-giver (**Hypothesis 2.3**) would be higher in the barge-in/embodied condition than the non-embodied/non-barge-in condition. We believed the barge-in/embodied condition comes closest to simulating playing the game with a human clue-giver and therefore would be perceived as more natural by players.

5 Results

This section contains the main findings from the evaluation.[4]

[4] All statistical tests discussed in this section are two-tailed un-paired independent t-tests.

5.1 Objective Results

We find evidence to support hypothesis 1 for both *cg* and *sk* dialogue acts. We found a (trending) significant difference for the # of *cg* utterances recognized in the barge-in (n=31) (M=1.4, SD=0.92) vs the non-barge-in (n=21) (M=1.0, SD=0.65) conditions (approaching significance (t(49.88)=-1.89, p= .064) We found a significant difference between average # of *sk* utterances recognized by Mr. Clue per round for the barge-in (M=1.28,SD=1.22) vs the non-barge-in (M=0.58,SD=0.73) conditions (t(49.39)=2.32, p= .024). We note there are two possible reasons more *cg* utterances were recognized in the barge-in condition. First, as mentioned, *cg* is an interrupt action and therefore Mr. Clue can recognize it while speaking in barge-in mode. Second, players in the barge-in condition skipped or "moved on" significantly more than players in the non-barge-in condition. It is likely this "moving on" came at times players felt they could not make a correct guess in a reasonable amount of time which afforded Mr. Clue more time in which to give clues for new target-words that players might have had a better chance of guessing correctly quicker.

5.2 Subjective Results

Subjective results were mainly calculated based on answers to a post-survey filled out by participants. We did not find evidence for hypothesis 2.1. We did find evidence to support hypothesis 2.2 for the embodied condition. A significant difference was found between the embodied and non-embodied conditions for the question "How natural did you find the voice?" on the post survey for the embodied condition (n=32) (M= 2.8, SD=1.36) vs the non-embodied condition (n= 20)(M:=2.0 SD=1.3)(t(41.87)=2.09, p= .04). This provides evidence that synthetic voices are found to be more natural when spoken by a realistic human avatar (with some basic non-verbal behavior) compared to a disembodied voice in the game context. We did not find evidence to support hypothesis 2.3. However, the difference for participants in the embodied condition (M= 2.3, SD=1.21) compared to ones in the non-embodied condition(M= 1.75 , SD=1.30) in response to the question "How natural did you find the clue-giver?" approached significance (t(42.69)=1.80, p= 0.07) indicating embodiment might be more important than barge-in for designing a game that felt more similar to the experience from a human-human game. This requires further investigation.

6 Previous Work

Now we briefly put our work in context with other dialogue systems that allow for user-initiative barge-in. As mentioned in Section 1, Mr. Clue's user-initiative barge-in policy moves beyond standard user-initiative barge-in models. As described in [12] standard user-initiative barge-in models stop a system prompt if user voice activity is detected and wait for a final ASR hypothesis before proceeding to take the next dialogue action. Mr. Clue's user-initiative barge-in policy moves beyond this model it is continuously listening for partial ASR hypotheses of the user's speech while speaking (rather than simply halting on voice activity). Only if a partial ASR hypothesis is classified as an interrupt action does Mr. Clue halt his speech and then proceed to take the next dialogue action (an intelligent update that takes into account the user's barge-in utterance).

Prior systems that have user-initiative barge-in policies include [8, 9, 11, 12]. We found one system that can handle **mixed-initiative barge-in** (i.e.- system can barge-in on user and user can barge-in on system) [10]. Only 2 of the systems from this list do intelligent updating based on partial ASR hypotheses [10, 12]. While the model proposed in [12] is capable of doing intelligent updating based on partial ASR hypotheses, their model halts speech for any stable partial ASR hypothesis (i.e. - a partial hypothesis that is not likely to be corrected in a later partial or final hypothesis) rather than only halting speech for partial ASR hypotheses that are expected to be a pre-defined interrupt action (in Mr. Clue those being *cg* or *sk*). [10] implemented a mixed-initiative barge-in dialogue system that makes intelligent updates based on user's partial ASR hypotheses but no evaluation was done.

7 Conclusions

This paper presents updates made to Mr. Clue, a fully-automated embodied dialogue agent who acts as a clue-giver in a collaborative word-guessing game. Mr. Clue's dialogue manager has been augmented with a user-initiative barge-policy that is capable of intelligent updating. We discuss results from an experiment designed to evaluate this new user-initiative barge-in policy in relation to a version of the system that flushed all user speech while the system is speaking. We show that game-scores are (trending) significantly higher and players "skip" or move on significantly more when the dialogue system is able to perform intelligent updating.

8 Acknowledgments

The effort described here is supported by the U.S. Army. Any opinion, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred. We would like to thank Anton Leuski and Ed Fast for help with this work.

References

1. Pincus, Eli; Devault, David; Traum, David Mr. Clue-A virtual agent that can play word-guessing games Tenth Artificial Intelligence and Interactive Digital Entertainment Conference (AIIDE), (2014)
2. Paetzel, Maike; Racca, David R; Devault, David A Multimodal Corpus of Rapid Dialogue Games Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), (2014)
3. Pincus, Eli; Traum, D Towards a multimodal taxonomy of dialogue moves for word-guessing games Proc. of the 10th Workshop on Multimodal Corpora (MMC) (2014)
4. Hartholt, Arno; Traum, David; Marsella, Stacy C; Shapiro, Ari; Stratou, Giota; Leuski, Anton; Morency, Louis-Philippe; Gratch, Jonathan All together now Intelligent Virtual Agents (IVA) (2013)
5. Lee, Jina and Marsella, Stacy Nonverbal behavior generator for embodied conversational agents Intelligent virtual agents (IVA), (2006)
6. Pincus, Eli; Georgila, Kallirroi; Traum, David Which Synthetic Voice Should I Choose for an Evocative Task? 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (2015)
7. Miller, George A. WordNet: A Lexical Database for English Commun. ACM, Vol 38 p. 39-41 (1995)
8. Dohsaka, Kohji; Shimazu, Akira System architecture for spoken utterance production in collaborative dialogue Working Notes of IJCAI 1997 Workshop on Collaboration, Cooperation and Conflict in Dialogue Systems (1997)
9. Kamm, Candace; Narayanan, Shrikanth; Dutton, Dawn; Ritenour, Russell Evaluating spoken dialog systems for telecommunication services Fifth European Conference on Speech Communication and Technology (1997)
10. Nakano, Mikio; Dohsaka, Kohji; Miyazaki, Noboru; Hirasawa, Jun-ichi; Tamoto, Masafumi; Kawamori, Masahito; Sugiyama, Akira; Kawabata, Takeshi, Handling rich turn-taking in spoken dialogue systems. EUROSPEECH (1999).
11. Strom, Nikko; Seneff, Stephanie Intelligent barge-in in conversational systems INTERSPEECH p. 652-655 (2000)
12. Selfridge, Ethan; Arizmendi, Iker; Heeman, Peter A; Williams, Jason D Continuously Predicting and Processing Barge-in During a Live Spoken Dialogue Task SIGDIAL Conference, p. 384–393 (2013)