Toward Natural Turn-Taking in a Virtual Human Negotiation Agent

David DeVault and **Johnathan Mell** and **Jonathan Gratch**
USC Institute for Creative Technologies
Playa Vista, CA 90094

Abstract

In this paper we assess our progress toward creating a virtual human negotiation agent with fluid turn-taking skills. To facilitate the design of this agent, we have collected a corpus of human-human negotiation roleplays as well as a corpus of Wizard-controlled human-agent negotiations in the same roleplay scenario. We compare the natural turn-taking behavior in our human-human corpus with that achieved in our Wizard-of-Oz corpus, and quantify our virtual human's turn-taking skills using a combination of subjective and objective metrics. We also discuss our design for a Wizard user interface to support real-time control of the virtual human's turn-taking and dialogue behavior, and analyze our wizard's usage of this interface.

1 Introduction

In this paper we explore the turn-taking behavior of a virtual human negotiation agent under wizard control. Wizard-of-Oz studies provide several important kinds of methodological value. They enable researchers to explore hypotheses about how people will interact with computers or virtual humans as opposed to another person (Dahlbäck, Jönsson, and Ahrenberg 1998). They enable design alternatives for future automated systems to be explored in a more economical way than building out the systems themselves. And they can provide training data that helps system builders to bootstrap an automated system to follow. In our prior work on virtual human systems, we have found Wizard-of-Oz studies valuable for all of these reasons (DeVault et al. 2013; Gratch et al. 2014; DeVault et al. 2014).

The work presented here is an initial investigation into the prospects for using Wizard control to achieve fluid turn-taking in negotiation roleplay dialogues between a human user and a virtual human. The natural speed of human turn-taking behavior poses special challenges for Wizard control. For example, while it is common for current spoken dialogue systems and virtual humans to have response latencies on the order of a second or more, human speakers tend to understand and respond to speech much more quickly (Sacks, Schegloff, and Jefferson 1974). It is an interesting question to what extent human-like response latencies can be achieved in Wizard-mediated interactions, where system responses must be triggered through a user interface.

Copyright © 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Figure 1: Sam, a virtual human negotiation agent (male).

Negotiation dialogues are a rich domain in which to explore turn-taking behavior. During a negotiation, interlocutors can be under substantial cognitive load, and as we will illustrate, natural silent pauses and speaker switch times can be substantial (occasionally more than ten seconds in our data). Additionally, turn-taking skills can be important to establishing rapport and solidarity during a negotiation, or to expressing a position forcefully during a dispute. In future work, we will be exploring the role of turn-taking factors in achieving positive negotiation outcomes.

We are ultimately interested in advancing computational models that support the automation of low-latency turn-taking decisions, the use of filled pauses to coordinate the turn-taking process, and the use of backchannels and overlapping speech in implemented systems. In all of these areas, human speakers have much more nuanced turn-taking skills than the current generation of spoken dialogue systems and virtual humans. But there has been much recent progress in computational models of these skills, for example in models of multi-party turn-taking (Bohus and Horvitz 2011), the use of filled pauses during system processing latencies (Skantze and Hjalmarsson 2013; Baumann and Schlangen 2013), detecting the end of a user's turn (Raux and Eskenazi 2012), predictive models of when to initiate the next turn (Ward, Fuentes, and Vega 2010; Laskowski, Edlund, and Heldner 2011), or predicting the timing of a verbal or non-verbal backchannel (Solorio et al. 2006;

Morency, Kok, and Gratch 2010). These models help predict and identify upcoming points in time which are natural opportunities for systems to take the floor and act. We intend to use the negotiation data sets we are creating to explore the use of similar computational models to achieve natural and effective turn-taking behavior in virtual human negotiations.

We have created two prototype virtual humans, a male (pictured in Figure 1) and a female (not pictured). These agents are designed to serve as virtual roleplayers for standard negotiation and conflict-resolution exercises used in leadership development courses (Murnighan 1991).

In this paper, we report on a pilot version of our agents in which two human Wizards worked together to control the agent's verbal and non-verbal behavior, and did so with an aim not only to engage in a coherent negotiation roleplay with the human user, but also to achieve relatively natural and fluid turn-taking between the human and virtual human. In addition to a corpus of 30 such Wizard-controlled virtual human dialogues, we have also collected a corpus of 89 face-to-face human-human negotiations in the same negotiation roleplay. We discuss the design of our Wizard control interface, which aims to support low latency, fluid control of the virtual human's turn-taking and dialogue behavior, and also analyze our wizards' use of this interface as it relates to turn-taking. We contrast the natural turn-taking behavior in our face-to-face corpus with that achieved in our Wizard-of-Oz corpus, and evaluate our virtual human system using a combination of subjective questionnaire data as well as objective metrics related to speaker switch time, pause durations, and use of overlapping speech.

2 Negotiation Scenario

Negotiations are dialogues aimed at reaching an agreement between parties when there is a perceived divergence of interest (Pruitt and Carnevale 1993). Although this definition is broad, researchers have sought to abstract essential elements of negotiations into more structured tasks that are suitable for both teaching and scientific enquiry. In this paper we focus on one useful and common abstraction known as the multi-issue bargaining task (Kelley and Schenitzki 1972). In multi-issue bargaining, each party is attempting to reach a single agreement spanning multiple issues. Each issue is formalized as having a set of possible levels and parties obtain different rewards depending on which levels they mutually agree upon for each issue. The rewards are typically unknown but bargaining tasks are often crafted so that parties believe their interests are divergent, even though they may partially align. Only through conversation can they discover each other's interests, but parties often withhold or misrepresent their true interests to avoid exploitation or gain a strategic advantage. This means parties often fail to discover mutually-beneficial solutions, and makes bargaining dialogues an especially rich tool for study (Van Kleef, De Dreu, and Manstead 2004) and teaching (Murnighan 1991) of human social skills, as well as a tool for advancing artificially intelligent agents (Baarslag et al. 2013).

Figure 1 illustrates an instance of the three-issue bargaining task that we use in this paper. Participants are told that they must negotiate with another party how to divide the contents of a storage locker filled with three classes of valuable items (antique paintings, art deco lamps and vintage record albums). Each of these items corresponds to a separate "issue" and each issue has two or more "levels" (the painting can be given to one party or the other - two levels; the lamps can be given to one or the other party, but also split - three levels; and the records can be portioned in four possible ways - four levels). Each party in the negotiation receives a private payoff-matrix that defines how much money they can earn from each level of each issue. Depending on the combination of payoff matrices, it is possible to manipulate the actual divergence of interests between the two players. For example, if the agent Sam in Figure 1 most prefers records and the human player most prefers lamps, a win-win solution is possible.

For the corpora described in this article we use two different sets of payoff matrices. Side A (always played by the agent in the Wizard-of-Oz study) always prefers records the most (one record is worth all the lamps), prefers the lamps next, and only slightly prefers the painting. Side B is randomly assigned one of two possible payoff matrices. The divergent matrix assigns Side B the identical preferences to Side A (with the exception that Side B assigns zero value to the painting). Thus, parties have divergent interests and must fight over the items. The convergent matrix gives Side B a complementary set of preferences: lamps are preferred the most (one lamp is worth all the records) and the painting holds no value. Participants are motivated to reach a deal that conforms with their preferences. The value of their solution is converted into lottery tickets that gives them a chance at winning $100 after the experiment concludes. If they fail to reach an agreement in fifteen minutes they receive a small number of tickets equal to one of their most preferred items.

3 System Design

3.1 Virtual human architecture

Our system is implemented using a modular virtual human toolkit (Hartholt et al. 2013). In our Wizard-of-Oz setup, the virtual human is semi-automated, with many low-level functions carried out automatically, while two wizards make high-level decisions about the agent's verbal and non-verbal behavior. The agent's speech is synthesized by the NeoSpeech text-to-speech system (neospeech.com). Gestures and expressions associated with speech are selected automatically by NVBG (Lee and Marsella 2006) and realized using the SmartBody character animation system (Thiebaux et al. 2008). This low-level automation complements and facilitates the decision-making of the wizards.[1]

3.2 Utterance set

As part of developing the utterance set for our agents, we segmented, transcribed, and semantically annotated all 89 face-to-face dialogues. The Wizard-of-Oz UI is informed by

[1] In this paper we focus on the verbal wizard and omit analysis of the non-verbal wizard's role. To summarize, the non-verbal wizard uses a similar but much simplified interface that enables control of agent posture shifts, movement of items on the table, deictic gestures, gaze direction, hand gestures, and facial expressions.

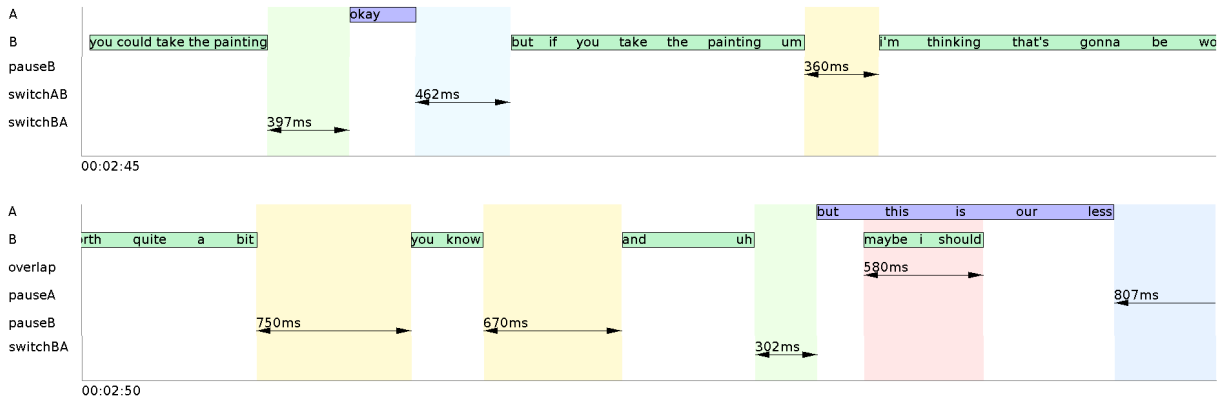

Figure 2: Example Face-to-Face dialogue excerpt, in which two human participants in roles A and B negotiate.

analysis and annotation of the resulting 11,745 speech segments. Each speech segment is an inter-pausal unit that has been manually identified by a human transcriber and is separated from other segments by a silence period of 300ms or greater. Figure 2 illustrates a face-to-face dialogue fragment.

Our semantic annotation scheme uses a frame representation to represent the meaning expressed in regions of speech we call utterance segments. These utterance segments are composed of contiguous sequences of words, and may span partial or multiple speech segments. They generally correspond to the granularity of individual dialogue acts in our frame representation. Each utterance segment is annotated using values for up to 8 different keys, including generic dialogue act (GDA), negotiation-specific dialogue act (NDA), propositional content templates (CONT), item type (ITEM), divisions of items (DIV), valence (VAL), topic (TOP), meta-information (MET), and lexical framing of offers using words like *take* or *get* (FRA). For example, the utterance *i'm most interested in the records* is annotated with frame 1, while *how about if i take two records and you take both lamps?* is annotated with frame 2:

Frame 1	
key	*value*
GDA	Statement
CONT	i-like-ITEM-best
ITEM	records

Frame 2	
key	*value*
GDA	Open-question
GDA	Wh-question
NDA	offer-DIV
DIV	S: R2 U: L2
FRA	take

We studied and generalized our corpus of human-human annotated data in order to design the output utterance set for our agent. This process yielded a set of 11,203 frames and 11,487 potential system utterances. 342 of these frames do not contain a DIV key-value (similar to frame 1), and a total of 805 system utterances are available to express these frames. There is thus an average of 2.35 utterance texts per non-DIV frame, providing the wizards with some variety of expression in case the same frame must be used repeatedly. The remaining 10,861 DIV-containing frames are permutations of frames such as frame 2 and contain various DIV key-values.[2] The system generally includes one unique utterance text for each DIV-containing frame; there are a total of 10,682 distinct utterances for these DIV-containing frames. There is such a large number of these frames mainly due to the many different ways of dividing up the 6 items on the table, including partial divisions that do not settle all the items. Some variety of expression is available through different values of the FRA key-value, which allows an offer like *i'll take two records* to be expressed in other ways such as *i'll get two records* or *i'll have two records*.

3.3 Wizard verbal control interface

A particular challenge in this Wizard-of-Oz system lies in the large number of utterances (11,487) to be made available to the wizards, and the desire at the same time to achieve low latency control and fluid turn-taking.[3] Because it is impractical to display thousands of utterances simultaneously for the wizard to choose from, the central problem is one of enabling *rapid navigation*. After exploring and pilot testing several alternatives, we decided to use a set of "quick buttons" for a set of high-frequency utterances, and to use our semantic annotation scheme to structure the navigation problem for all remaining utterances.

The user interface (UI) we created, and used by our verbal wizard in this study, is pictured in Figure 3. For space reasons, we show only the upper left portion of the UI. At the top left are several panels (Quick Words, Quick Pause, and Quick Acknowledge) that allow one-click activation of several types of high-frequency utterances, as follows.

Acknowledgments, agreements, acceptances, and yes answers. Based on analysis of human-human data, the agent utterance set includes 110 utterances that serve a positive response function such as acknowledgments, agreements, ac-

[2]The DIV key-values can assign a level for up to three different items (with records denoted by R, lamps by L, and paintings by P) to each of the system S and user U. For example, DIV S: R2 U: L2 assigns two records to the system and two lamps to the user.

[3]For comparison, in a recent Wizard-controlled virtual human system developed at our institute (DeVault et al. 2014), the number of available utterances was 191, and so this is a substantial increase in the expressive range of the system.

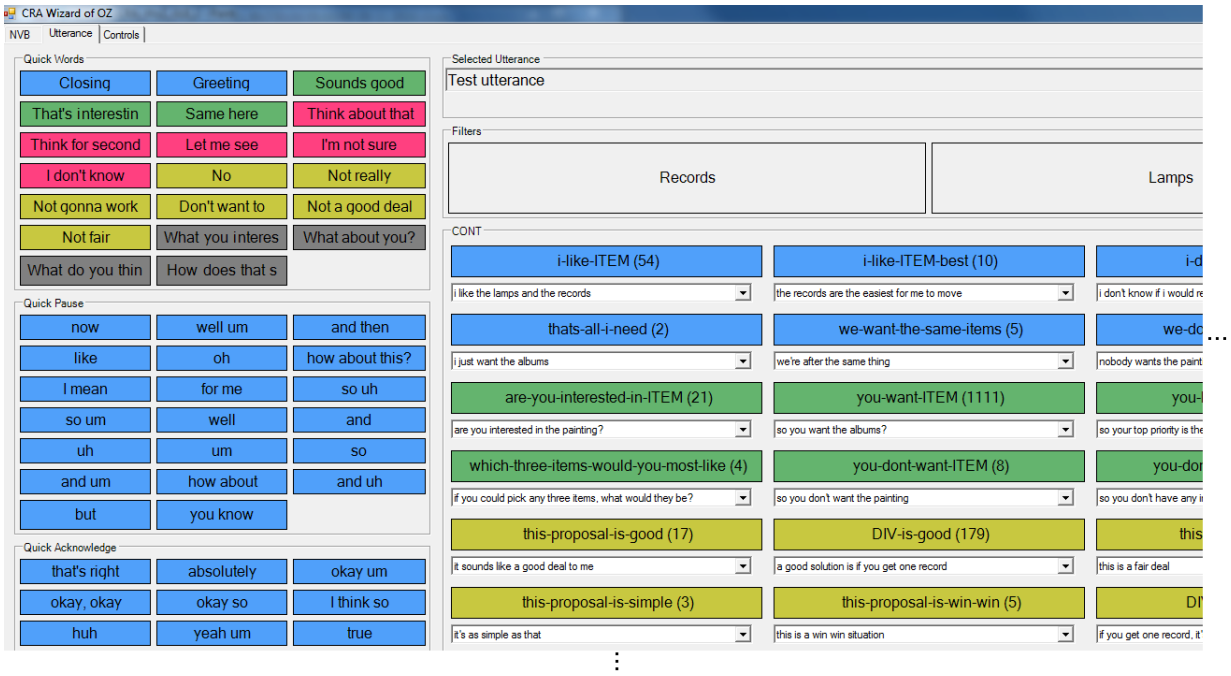

Figure 3: Partial screenshot of verbal wizard UI.

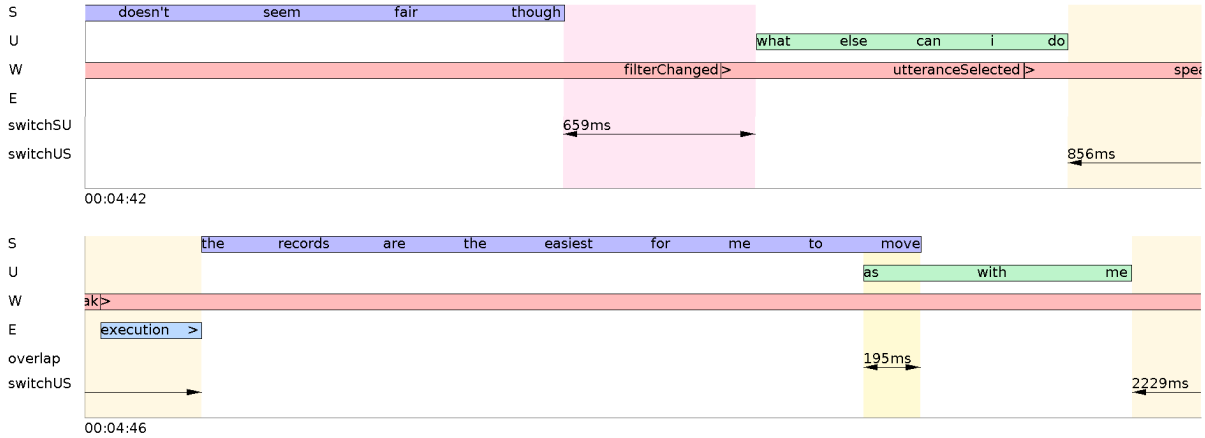

Figure 4: Example Wizard-of-Oz dialogue excerpt, in which the user U talks to the system S.

ceptances, and yes answers. Utterance segments that served only one of these positive response functions constituted 19.1% of our human-human utterance segments. Among the most frequent of these are *yeah, okay, right, alright, yes,* and *sure,* which were all made accessible in the Quick Acknowledge panel in the UI. Additionally, the 'a' key on the wizard's keyboard randomly generates *okay* with 86% chance or *alright* with 14% (based on corpus frequencies).

Filled pauses, discourse connectives and conjunctions, interjections. In our human-human data, 4.7% of annotated utterance segments are one of these phrases: *um, so, uh, well, and, but,* and *you know*. These are all made accessible through the Quick Pause panel in the wizard UI. Additionally, the 'f' key on the wizard's keyboard randomly generates *uh* with 61% chance, *um* with 24%, or *well* with 15%.

Backchannels. Backchannels also occurred often in our human-human data. While our current annotation scheme does not distinguish backchannels that occur as *yeah* or *right* or *okay* from positive response functions (described above), we observe that backchannels in the form of *mhm, mm, uh huh, mm k,* or *hmm* constitute 3.8% of our annotated utterance segments. Based on observed limitations in speech synthesis, we made *uh huh* (along with *yeah* and *right*) accessible to the verbal wizard through the Quick Acknowledge panel in the wizard UI. Additionally, the 'b' key on the wizard's keyboard was made to randomly generate *uh huh* with 47% chance, *yeah* with 42%, and *right* with 11%.

Other utterances. For other utterances, two or more UI

selection events are usually required, and the wizard uses individual key-values from the frame representation to narrow down possibilities and identify their desired utterance. The rest of the UI consists of widgets that allow specific key-values to be selected, and that allow utterances compatible with all current selections to be viewed in drop-down lists. For example, by clicking buttons for key-values ITEM records and CONT i-like-ITEM-best, a desired frame such as frame 1 can be distinguished quickly. A drop-down list is always available showing all compatible utterances, and this list becomes shorter as more key-value constraints are applied. Panels are generally organized by key (CONT, GDA, NDA, etc.) and expose one button for each key-value. At the bottom left (not pictured), we provide a DIV-construction panel that allows the wizard to click numbers of items for each person, and thereby construct an arbitrary DIV key-value to suit a desired negotiation move. For example, the wizard can click "2 records" for the agent, click "2 lamps" for the user, and this suffices to create the key-value DIV S: R2 U: L2. This DIV can be combined with other selections to find offers, confirmations, and clarifications.

All the buttons have been grouped and color-coded by our wizard. As each key-value is selected, the UI dynamically grays out and disables buttons and widgets for incompatible key-values, streamlining the search process. The wizard can select an utterance to be spoken later. This enables planning an agent response to a user utterance in progress. Utterances are actually triggered when the wizard presses the space key.

To enable later analysis of how the wizard's usage of this UI relates to the turn-taking achieved, the UI logs all wizard selection events with a millisecond precision timestamp. Section 6 presents an initial analysis using this UI log data.

4 Data sets

4.1 Face-to-Face data

We recruited 178 participants (89 pairs) from Craigslist to engage in a face-to-face version of the negotiation described in Section 2. For these negotiations, participants were seated across a table and were provided with wooden blocks representing the painting, lamps, and records, analogous to the virtual negotiation tabletop seen in Figure 1. Participants were recruited in the Los Angeles area, and were gender matched.[4] 62 pairs were male-male and 27 were female-female. All dialogues were segmented, transcribed, and annotated as described in Section 3.2. An example excerpt from one of these negotiations is shown in Figure 2.

4.2 Wizard-of-Oz data

We recruited 30 participants from Craigslist in the Los Angeles area to engage in a negotiation roleplay with our Wizard-controlled virtual human. Male participants interacted with our male virtual human, pictured in Figure 1, while females interacted with a female virtual human controlled by the same wizard interface.[5] A short fragment of a Wizard-controlled interaction is provided in Figure 4.

Participants were instructed that they were interacting with a computer-controlled virtual agent, and were not informed of the wizard control. The same two wizards controlled all interactions from a separate room. They observed a live video feed of the participant during each interaction.

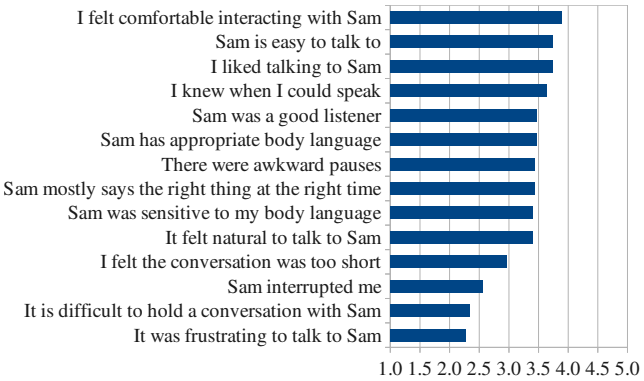

Figure 5: Subjective questionnaire assessment of turn-taking

5 Evaluation of virtual human turn-taking

In this section, we use a combination of subjective and objective measures to assess our progress in achieving natural turn-taking in our virtual human.

5.1 Subjective assessment

All Wizard-of-Oz participants completed a post-negotiation questionnaire in which they provided ratings on a 5-point Likert scale to a set of statements that relate to various aspects of the virtual human's interaction and turn-taking skills. Figure 5 provides mean ratings for these statements. For each statement, the ratings range from 1, representing strong disagreement, to 5, representing strong agreement. We observe favorable ratings for many positively framed statements about the interaction, such as *I liked talking to Sam* and *Sam is easy to talk to*, and lower ratings for negatively framed statements such as *It is difficult to hold a conversation with Sam*. The most critical assessment was for *There were awkward pauses*, where participants tended to express agreement. We explore further the pause structure of the dialogues with our objective metrics and UI analysis.

5.2 Objective metrics

We have a set of 89 face-to-face dialogues, and 30 Wizard-of-Oz dialogues. The human utterances in all dialogues were segmented and transcribed using ELAN (Wittenburg et al. 2006), using the segmentation protocol described in Section 3.2. For each speaker and dialogue, this process resulted in a sequence of human speech segments of the form

[4]To control for gender effects, we gender-matched the participants. Same-sex dyads were chosen over opposite-sex dyads because there is empirical evidence for perceived gender differences in negotiation ability (Kray, Kennedy, and Zant 2014). Using same-sex dyads also allowed us to remove, for heterosexual participants, the possible element of physical attraction from the interaction.

[5]The female uses the same utterance set and general dialogue policy, but does differ in appearance, voice, and certain animation details for her gestures and non-verbal behavior.

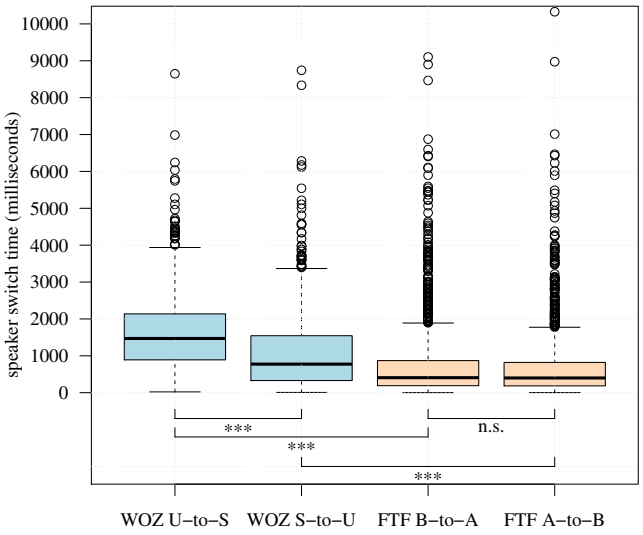

Figure 6: Speaker switch time across conditions and roles. *** $p < 0.0005$ (t-test).

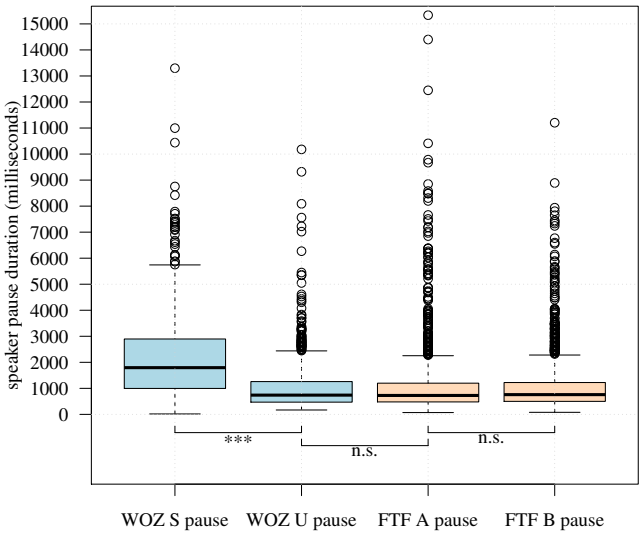

Figure 7: Speaker pause durations across conditions and roles. *** $p < 0.0005$ (t-test).

$g_i = \langle s_i, b_i, e_i, t_i \rangle$ where speaker s_i speaks for a time period starting at b_i and ending at e_i, and where t_i is the manual text transcription of the corresponding audio segment. The agent's speech segments for WOZ dialogues, including starting and ending timestamps and verbatim transcripts of system utterances, were recovered from the system log files.

Total duration of speech We observed no significant differences in the total duration of speech for any of the individual interlocutors: WOZ system (mean 105.1s, std 57.6), WOZ user (mean 121.7s, std 88.6), FTF A role (mean 123.3s, std 110.9), and FTF B role (mean 125.0s, std 115.4). The total amount of speech in the virtual human negotiations appears comparable to that in human-human negotiations.

Speaker switch times We quantify differences in speaker switch times as a coarse, high-level measure of the speed of turn-taking. For the purposes of this paper, we restrict our attention to changes from one speaker to another in which there is no overlapping speech. This is in comparison to metrics such as *floor transfer offset,* in which there may be overlapping speech during a floor transition, and the floor transfer offset time can therefore be negative (de Ruiter, Mitterer, and Enfield 2006). For our analysis, we sort all speech segments for both speakers into ascending order by start time, and define a speaker switch as occurring between segments g_i and g_{i+k} if $k \geq 1$ and $s_i \neq s_{i+k}$ and $b_{i+k} > e_i$ and no other speech segment overlaps the temporal region from e_i to b_{i+k}. We illustrate several such speaker switches in Figure 2, both from speaker A to B (switchAB) and from speaker B to A (switchBA). Figure 4 likewise provides examples of switches from system to user (switchSU) and user to system (switchUS).

We report aggregate results for speaker switch times in Figure 6. As our agent always plays role A, we generally compare the system S to role A in the FTF data, and the WOZ user U to role B in the FTF data. We find that speaker switches for WOZ U-to-S (mean 1719ms, std 2005) are significantly longer ($p < 0.0005$, t-test) than WOZ S-to-U (mean 1121ms, std 1151). Not only is our agent slower to initiate speech than human speakers in WOZ, but interestingly, speaker switch times for WOZ S-to-U are significantly longer ($p < 0.0005$, t-test) than FTF A-to-B (mean 713ms, std 1142). This suggests that in terms of speaker transitions, users are adapting somewhat and not interacting with our virtual human exactly as they would with a human speaker. Note also the long tails of long speaker switches in these negotiations; we observed speaker switches of up to 26 seconds in FTF (not depicted in the figure for space reasons).

Pause durations We also observed differences in pause durations between conditions. We define a pause as occurring whenever there are two consecutive speech segments by the same speaker. Figure 2 illustrates several pauses by speaker B. Aggregate results for pause durations are shown in Figure 7. We find that our virtual human's pauses (mean 2225ms, std 1772) are significantly longer ($p < 0.0005$, $t-$test) than those of WOZ users (mean 1034ms, std 962). No significant differences in pause durations were observed between WOZ users and FTF speakers.

Overlapping speech We also observed significant differences in the rate of overlapping speech between WOZ and FTF conditions, as illustrated in Figure 8. The figure shows the fraction of each interlocutor's speech that is overlapped by speech from the other interlocutor. We find the fraction of overlapping speech for WOZ system speech (dialogue mean 0.079, std 0.052) is significantly lower ($p < 0.0005$, t-test) than the FTF A role (dialogue mean 0.15, std 0.090). Similarly, the fraction of the WOZ user's speech that is overlapped (dialogue mean 0.068, std 0.029) is significantly less ($p < 0.0005$, $t-$test) than the FTF B role (dialogue mean 0.15, std 0.097). No significant difference was observed between the two roles in WOZ or FTF. Further investigation

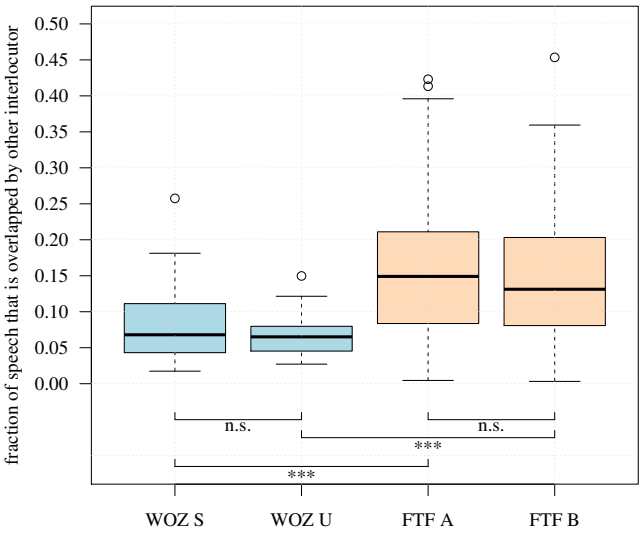

Figure 8: Rate of overlapping speech across conditions and roles. *** $p < 0.0005$ (t-test).

events	1	2	3	4	5	6	7	8	9	10
count	792	589	282	81	50	32	21	9	3	2

Table 1: Number of UI selection events per utterance

revealed that individual instances of overlapping speech, examples of which are provided in Figures 2 and 4, are significantly shorter in WOZ (mean 443.5ms, std 362.2) than in FTF (mean 530.6ms, std 494.9) ($p < 0.0005$, $t-$test).

6 Evaluation of Wizard-of-Oz UI

In order to better understand our findings for the objective metrics, we analyzed the verbal wizard's UI event log (described in Section 3.3). We found that the verbal wizard took from 1 to 10 UI selection actions in the UI per utterance. The frequency distribution is depicted in Table 1. 42.5% of the agent's utterances were associated with a single UI button click; these are the "quick buttons" described in Section 3.3.

Due to our findings of high latency in WOZ U-to-S speaker switches, as well as long WOZ system pauses, we analyzed in detail the UI activities of the verbal wizard, and associated system execution latencies, during only these specific speaker switches and pauses. During these analyzed silence periods, our UI and system log data allow us to infer to a large extent what was happening in the wizard UI and in the system architecture. Figure 4 visualizes the wizard's UI activities in the row marked W, and also the system's execution latency in the row marked E, for a short WOZ fragment.

The wizard's activities include a filterChanged event that occurs during a switch from the system to the user (highlighted region marked switchSU). This event occurs at the moment the vertical bar appears after 'filterChanged' in the diagram. This event indicates the wizard clicked one of the buttons that filters other options, such as the ITEM or DIV panel buttons in the UI. The next event is an utteranceSelected event, indicating that the wizard has identified the system's next utterance in a drop down list (*the records are the easiest for me to move*). Note that this selection occurs during the user's utterance; this type of parallelism can help reduce system response latency. The wizard then chooses to trigger the agent to actually speak this utterance at the moment marked by a vertical bar after 'speak' appears in the diagram. Following this, there is an execution period (marked by a blue bar in the E row) while several system modules are invoked to decide non-verbal behavior, synthesize speech, and animate the virtual human to perform this utterance. The system's speech begins after this execution latency.

UI event	percentage of analyzed time	mean latency per instance (ms)
quickButtonPressed	28.2%	1364.5
execution time	27.7%	548.7
utteranceSelected	24.7%	1371.1
filterChanged	6.3%	698.7
speak	6.1%	313.7
randomFiller	5.0%	1247.6
proceduralButton	1.3%	924.7
clearSelection	0.2%	679.0
randomAcknowledge	0.2%	845.6

Table 2: UI events during analyzed VH silence periods

Table 2 summarizes our findings. For each analyzed event type, we present the percentage of time during the analyzed silence periods that can be attributed to that event type, as well as the mean latency in milliseconds associated with each event instance. We find that execution time accounts for 27.7% of system latency during these periods, and on average adds 548.7 milliseconds of latency to each utterance. In the context of mean FTF speaker switch times of around 700ms, our current system execution latency is itself a substantial impediment to achieving human-like turn-taking with wizard control. The utteranceSelected events account for 24.7% of the analyzed silence periods. In fact, much of the 28.2% of silence time associated with quickButtonPressed and randomFiller (invoked via the keyboard key 'f') in this analysis can be explained by floor-holding uses of filled pauses such as *uh* and *um* during difficult utterance selections. The wizard often initiated these when finding the next utterance was taking so long that a filled pause became appropriate. Typically these appear as part of a subdialogue such as *System: how about this?...uh...well...i could give you two lamps and the painting*. Selecting individual filters such as clicking the ITEM records button in the UI took 698.7ms on average per instance. The choice to trigger the agent's actual speech (after selecting an utterance) took an additional 313.7ms per instance. Other UI events explain a relatively small amount of the analyzed silence periods.

These findings suggest a number of improvements to the UI and system to enable more fluid turn-taking under wizard control. The utterance lists in the UI are currently somewhat cluttered due the presence of multiple ways to express system frames (see Section 3.3). Since identifying utterances in these lists is creating so much latency, many such varia-

tions can either be eliminated or removed from wizard control through randomization. We also observed that certain offers recur repeatedly, and are associated with high latency due to the need to click up to six buttons in the DIV selection panel. (This time appears under filterChanged in the table.) The most commonly recurring DIVs can be represented in the top-level UI using single buttons, enabling common offers to be invoked with fewer clicks. Finally, system execution time is a major factor, and we have begun to analyze this latency more closely, with an eye toward optimization.

7 Conclusion & Future work

In this paper we have presented the results of an experiment in wizard control of a virtual human system designed to participate in negotiate roleplays with fluid turn-taking skills. We have observed encouraging impressions by participants of the ease of interaction in subjective survey results, but at the same time have identified a number of differences in interaction latency and overlapping speech when compared with face-to-face data. In future work, we will use this analysis to further improve our UI for wizard control, reduce system execution latency, and aim to achieve a more human-like turn-taking capability in wizard-based research.

8 Acknowledgments

We thank Gale Lucas. The project or effort described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

Baarslag, T.; Fujita, K.; Gerding, E. H.; Hindriks, K.; Ito, T.; Jennings, N. R.; Jonker, C.; Kraus, S.; Lin, R.; and Robu, V. 2013. Evaluating practical negotiating agents: Results and analysis of the 2011 international competition. *Artificial Intelligence* 198:73–103.

Baumann, T., and Schlangen, D. 2013. Open-ended, extensible system utterances are preferred, even if they require filled pauses. In *SigDial*, 280–283.

Bohus, D., and Horvitz, E. 2011. Multiparty turn taking in situated dialog: Study, lessons, and directions. In *SigDial*.

Dahlbäck, N.; Jönsson, A.; and Ahrenberg, L. 1998. Wizard of Oz studies – why and how. In Maybury, M. T., and Wahlster, W., eds., *Readings in Intelligent User Interfaces*.

de Ruiter, J.; Mitterer, H.; and Enfield, N. J. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. 82(3):515–535.

DeVault, D.; Georgila, K.; Artstein, R.; Morbini, F.; Traum, D.; Scherer, S.; Rizzo, A. S.; and Morency, L.-P. 2013. Verbal indicators of psychological distress in interactive dialogue with a virtual human. In *Proceedings of SIGdial*.

DeVault, D.; Artstein, R.; Benn, G.; and et al. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of AAMAS*.

Gratch, J.; Lucas, G.; King, A.; and Morency, L.-P. 2014. It's only a computer: The impact of human-agent interaction in clinical interviews. In *Proceedings of AAMAS*.

Hartholt, A.; Traum, D.; Marsella, S.; Shapiro, A.; Stratou, G.; Leuski, A.; Morency, L.-P.; and Gratch, J. 2013. All together now, introducing the virtual human toolkit. In *IVA*.

Kelley, H. H., and Schenitzki, D. P. 1972. Bargaining. *Experimental Social Psychology. New York: Holt, Rinehart, and Winston* 298–337.

Kray, L. J.; Kennedy, J. A.; and Zant, A. B. V. 2014. Not competent enough to know the difference? Gender stereotypes about women's ease of being misled predict negotiator deception. *Organizational Behavior and Human Decision Processes* 125(2):61 – 72.

Laskowski, K.; Edlund, J.; and Heldner, M. 2011. Incremental learning and forgetting in stochastic turn-taking models. In *Interspeech*.

Lee, J., and Marsella, S. 2006. Nonverbal behavior generator for embodied conversational agents. In Gratch, J.; Young, M.; Aylett, R.; Ballin, D.; and Olivier, P., eds., *Intelligent Virtual Agents*, volume 4133 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 243–255.

Morency, L.-P.; Kok, I.; and Gratch, J. 2010. A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems* 20.

Murnighan, J. K. 1991. *The dynamics of bargaining games*. Prentice Hall Englewood Cliffs, NJ.

Pruitt, D., and Carnevale, P. 1993. *Negotiation in social conflict*. Thomson Brooks/Cole Publishing Co.

Raux, A., and Eskenazi, M. 2012. Optimizing the turn-taking behavior of task-oriented spoken dialog systems. *ACM Trans. Speech Lang. Process.* 9(1):1:1–1:23.

Sacks, H.; Schegloff, E.; and Jefferson, G. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 696–735.

Skantze, G., and Hjalmarsson, A. 2013. Towards incremental speech generation in conversational systems. *Computer Speech & Language* 27(1):243 – 262.

Solorio, T.; Fuentes, O.; Ward, N. G.; and Bayyari, Y. A. 2006. Prosodic feature generation for back-channel prediction. In *Interspeech*.

Thiebaux, M.; Marsella, S.; Marshall, A. N.; and Kallmann, M. 2008. Smartbody: behavior realization for embodied conversational agents. In *AAMAS*, 151–158.

Van Kleef, G. A.; De Dreu, C. K. W.; and Manstead, A. S. R. 2004. The interpersonal effects of anger and happiness in negotiations. *Journal of Personality and Social Psychology* 86(1):57–76.

Ward, N. G.; Fuentes, O.; and Vega, A. 2010. Dialog prediction for a general model of turn-taking. In *Interspeech*.

Wittenburg, P.; Brugman, H.; Russel, A.; Klassmann, A.; and Sloetjes, H. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of LREC*.