Mark ter Maat*
Khiet P. Truong
Dirk Heylen
Human Media Interaction
University of Twente
The Netherlands

How Agents' Turn-Taking Strategies Influence Impressions and Response Behaviors

Abstract

Different turn-taking strategies of an agent influence the impression that people have of it and the behaviors that they display in response. To study these influences, we carried out several studies. In the first study, subjects listened as bystanders to computer-generated, unintelligible conversations between two speakers. In the second study, subjects talked to an artificial interviewer which was controlled by a human in a Wizard of Oz setting. Questionnaires with semantic differential scales concerning personality, emotion, social skill, and interviewing skills were used in both studies to assess the impressions that the subjects have of the agents that carried out different turn-taking strategies. In addition, in order to assess the effects of these strategies on the subjects' behavior, we measured several aspects in the subjects' speech, such as speaking rate and turn length. We found that different turn-taking strategies indeed influence the user's perception. Starting too early (interrupting the user) is mostly associated with negative and strong personality attributes and is perceived as less agreeable and more assertive. Leaving pauses between turns is perceived as more agreeable, less assertive, and creates the feeling of having more rapport. Finally, we found that turn-taking strategies also influence the subjects' speaking behavior.

1 Introduction

When creating an embodied conversational agent (or an ECA, see Cassell, Sullivan, Prevost, & Churchill, 2000), everything the ECA does has a certain effect on the user, every aspect of the ECA changes the user's perception of the agent. This includes what the ECA says and how, its appearance, gestures, facial expressions, head movements, and the timing and other properties of all these aspects. In several studies, ECAs have been investigated and developed. For example, in the SEMAINE project (http://www.semaine-project.eu), four different ECAs are created, each with a different character. There is Poppy, who is cheerful and optimistic; Obadiah, who is gloomy and depressed; Spike, who is aggressive and negative, and Prudence, who is always pragmatic. These ECAs have different personalities (throughout this paper, we will continue to use the term personality with a meaning as described by Mcrorie, Sneddon, Sevin, Bevacqua, & Pelachaud, 2009, and as used in the Semaine project) to draw

Presence, *Vol. 20, No. 5, October 2011,* 000–000

*Correspondence to m.termaat@utwente.nl.

the user into the same emotional state, the main motivations for this project being to induce emotions from users.

These ECAs intentionally have a certain personality, but whether intentionally or not, every ECA will probably evoke the idea of a particular personality as it shows in the way the agent responds to conversational events. For example, Kopp, Gesellensetter, Krämer, and Wachsmuth (2005) describe Max, an ECA that acts as a museum guide in a museum. By default, Max behaves politely. Max greets participants and encourages them to start a conversation, and it stops speaking when it detects that the user wants to say something. Ochs, Pelachaud, and Sadek (2008) describe an ECA with an empathic personality; the agent expresses an empathic emotion when it detects an emotion of the user. Van Deemter et al. (2008) describe the NECA project, in which a system was built that can automatically generate scripts for virtual actors. The personality of the agents—as the authors call it, which can be polite or impolite, and good humored or ill tempered—affect dialogue act selection, text generation, gesture alignment, and speech synthesis. Changing utterances and nonverbal behaviors, or the way the agent looks, have been the most prominent markers of personality.

One class of behaviors that has not been exploited much to create a certain appearance is turn-taking, which is one of the fundamental and universal aspects of conversation. Researchers and developers dealing with turn-taking often assume one standard default model of turn-taking; hence, they have not explored the use of variations of turn-taking strategies as a tool to display different personalities, stances, emotions, and social behavior. Within the computational literature, researchers have mainly tried to ensure clean turn-taking, where the agent waits for its turn, which is not the rule in human-human conversations. For example, Atterer, Baumann, and Schlangen (2008) and Schlangen (2006) developed algorithms that predict turn-endings as soon as possible such that the system can behave quick enough to simulate human-like behavior. Raux & Eskenazi (2008) explain how they use audio features to detect an end of turn as soon as possible, so an agent can start speaking as soon as possible. Jonsdottir, Thorisson, and Nivel (2008) and Jonsdottir and Thorisson (2009) developed a real-time turn-taking model that is optimized to minimize the silence gap between the human's speech turn and the system's speech turn. All these models seem to assume that there is some kind of perfect turn-taking behavior in which the duration of overlapping speech and pauses is as short as possible. But as we said before, this may not be true to what happens in the real world.

Studies have shown that turn-taking behavior can be indicative of speakers' attitudes, stances, and social roles in conversation. For example, Trimboli and Walker (1984) found that cooperative conversations have significantly fewer interruptions than competitive conversations. Beattie (1981) shows that the interruption behavior of high-status and low-status participants in a conversation differs. In conversations between tutors and students he found that the high-status participants interrupt significantly less, but their interruptions are significantly more often barge-in interruptions, in which the current speaker has not reached a possible completion point yet. When studying conversations more closely, one can see that overlapping speech is not wrong by default, and a lot can be learned from overlapping speech and interruptions. Schegloff (2000) argued that certain types of overlapping speech are not problematic at all; for example, when a next speaker starts just before the prior speaker finishes. Observing turn-taking behavior can be very informative about the kind of person one is dealing with, how a person is feeling, or how the person wants to be perceived: "A clash of opinions also means a clash of turn-taking" (Oreström, 1983). According to Goldberg (1990), interruptions are affected by many variables including "... obligations and wants of the speaker and listener; the personality traits of the speaker and the listener; and their respective moods, relational dispositions and levels of topical involvement." Robinson and Reis (1989) found that people who interrupt often are seen as less sociable, more assertive, more masculine, and less feminine than people who did not interrupt.

These previous findings lead to the focus of this research: how can we use different turn-taking strategies (i.e., the management of when to speak) as a mechanism

for ECAs to convey different emotional, interpersonal, and other states to a user, and how do these different strategies change the user's behavior and perception of the ECA? Some previous studies have partly addressed related issues. In a study by Robinson and Reis (1989), an offline study was performed in which participants listened to recorded human–human conversations and judged the participants on several personality and sociability scales. Fukayama, Ohno, Mukawa, Sawaki, and Hagita (2002) evaluated the impressions conveyed with different gaze models for an embodied agent. Although their research is not related to turn-taking, their study on how certain behavior can influence perception served as an important source of inspiration for our studies. Baumann (2008) created a setup in which two artificial agents communicate with each other in real time by exchanging audio streams. The agents analyze this audio and use simple turn-taking strategies to determine whether they start or stop speaking. The generated conversations are analyzed on the percentage of overlaps and silences. Hirasawa, Nakano, Kawabata, and Aikawa (1999) carried out a study about the user impressions of different timings of backchannels. In the first condition (the immediate condition), the system responded with a backchannel immediately after the acceptance of important information. In the second condition (the orderly condition), the system waited with a backchannel until the end of a user utterance. The authors found that the users had a more negative impression about the first system, and they found it harder to speak to it. The studies described demonstrate that certain agent behavior influences user perception; however, none of these studies have considered turn management as a tool to create different ECA personalities and have investigated the effects of agents' turn-taking strategies on user impressions.

In this paper we will perform an explorative study into the effects of different turn-taking strategies on the impression that a user has of an agent using a particular strategy: *How do different turn-taking strategies, used by the agent, influence the user's behavior and perception of the agent?* We will address this research question by carrying out two perception studies. In the first study, called the passive study, the users are bystanders in a computer-generated, unintelligible conversation between two agents, in which one agent uses the turn-taking strategy we want to test. This study was originally described by Ter Maat and Heylen (2009). In the second study, called the active study (originally described by Ter Maat, Truong, and Heylen, 2010), the user is actively involved in the conversation. Here, the computer (controlled by a human in a Wizard of Oz setting) takes the role of an interviewer, and asks a series of questions to the user. In the second study, where the user takes an active role, we have also looked at how the different turn-taking strategies change the behavior of the users themselves. This paper extends the previous publications by adding more detail on the analysis of the experimental data and offering a comparison between the two studies. In addition, we look at how agents' turn-taking behaviors can affect users' speaking behaviors.

The paper is structured as follows. Section 2 presents the passive study in which the subjects are passive bystanders and have to rate the impressions that they have of one of the agents. The active study is described in Section 3. In Section 4, we describe a comparison between the passive and active studies. Finally, in Section 5, we conclude with a discussion and general conclusions of the results obtained.

2 Rating of Unintelligible Conversations

We present the setup and the results of the passive study. This study aims to investigate the effects of a certain turn-taking strategy on the impression one has of this person using said strategy. Several artificial agent–agent conversations were generated in which one agent talks to another. The ways in which one agent deals with overlap (overlap resolution strategy) and when it starts speaking (start-up strategy) were systematically varied. These conversations were generated by a conversation simulator which we built to allow us to define different turn-management rules. The computer-generated, unintelligible conversations were presented to human raters who judged one of the agents on various semantic scales. We describe the turn-management strategies

(Section 2.1), the setup of the study (Section 2.2), and the results of the analysis of the subjects' ratings (Section 2.3).

2.1 Turn-Taking Strategies

We investigated two types of turn management strategies: start-up strategies and overlap resolution strategies. A start-up strategy determines when to start a new utterance (i.e., speech turn). We considered the following three strategies:

1. Early. A speaker will start its turn just before the end of the interlocutor's turn.
2. Direct. A speaker will start its turn immediately after the interlocutor's turn has finished.
3. Late. A speaker will leave a pause (of a few seconds) before starting its turn after the interlocutor's turn has finished.

An overlap resolution strategy determines what to do when there is overlap; that is, when two persons are speaking at the same time. When there is overlap, a speaker can decide to stop speaking (stop), a speaker can continue normally with speaking (normal), or a speaker can raise his or her voice (raised). Combinations of start-up and overlap resolution strategies can be made; for example, an early + raised strategy means that a speaker will start his or her turn early during the interlocutor's speech turn with a raised voice. All these strategies were realized by the conversation simulator (which will be described below) in which the strategies can be scripted. The strategy early + stop was not considered because a speaker would start while the other speaker was speaking and would immediately stop again, which resulted in rather weird conversations.

2.2 Experimental Setup

We will now describe how the study was set up and how the different turn-taking strategies were realized in the computer-generated conversations. We also present the semantic differential scales that were used in the questionnaire to measure the users' impressions of the agent.

2.2.1 Participants. Ten people participated in the perception study. The subjects, mainly students, were all between 20 and 30 years old. There were six male and four female subjects.

2.2.2 Stimuli: Unintelligible Conversations. The participants listened to unintelligible agent–agent conversations, which were generated by the conversation simulator (Ter Maat and Heylen, 2010). We specifically used unintelligible conversations for several reasons. First, it allows us to generate conversations and to control the start times of the speech turns and the prosodic characteristics (i.e., raised voice) without having to take into account the content of the speech turns. Second, the content of the sentences may influence the perception of the users; by masking the content, the users are forced to focus specifically on the timing and nonverbal characteristics of the speech turns. An example conversation can be found on the publisher's webpage of this paper.

The conversation simulator allows one to program the behavior of two agents that can communicate with one another by sending each other information on their communicative actions. They indicate whether they are silent, speaking normally, or using elevated speech. By adding noise to the agents' communication, we make sure that occasionally things go wrong; for example, an agent starting too early or two agents starting at the same time, since this also happens in real human–human conversations. For more information about this simulator, see Ter Maat & Heylen (2010).

The behavior of the two agents can be scripted. In these scripts one can define how an agent reacts to different situations. The core conversation simulator runs the scripts of the agents in parallel and takes care that the rules are executed and that the variables are updated accordingly. The conversations of the agents can be visualized (see Figure 1) and made audible. **Figure 1**

For the speech rendition that was used in the study we wanted to output natural but incomprehensible speech. To this end, we extracted several sentences from the AMI corpus[1] with a clear start and end point. These fragments were then passed through a Pass Hann Band **Fn 1**

1. http://www.amiproject.org

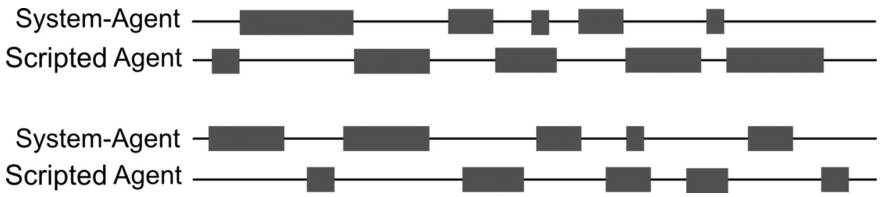

Figure 1. *Examples of conversations generated with the conversation simulator (Ter Maat and Heylen, 2009). top pane: direct + raised, bottom pane: late + continue.*

Filter, from 500 Hz to 500 Hz with a smoothing of 100 Hz. With this method, the fragments kept their prosodic information but lost their content.

Using the conversation simulator, eight different agents—all combinations of start-up strategies and overlap resolution strategies minus the early + stop strategy—were scripted using the described turn-taking strategies. This resulted in conversations that differed noticeably from each other. Examples of the conversations can be seen in Figure 1. The contributions of the scripted agent that varies its strategy are shown on the lower tier. This agent is scripted to use a certain start-up and overlap resolution strategy. The fixed system-agent was programmed to use random turn-taking behavior each turn and is shown on the top tier in each case. Because the system-agent uses random behavior each turn, all scripted agents encounter all possible turn-taking strategies during a conversation, which makes the comparison of the agents more fair. In order to make it clear who is speaking, the system-agent or the scripted agent, the scripted agent's pitch was made a bit higher than the system-agent's pitch, and we made sure that the system-agent's voice would be heard from the loudspeaker on the left of the user and the scripted agent's voice from the loudspeaker on the right of the user. We can actually see in Figure 1 that the conversations differ from each other. The question is whether these interactions lead to different perceptions of the scripted agent.

2.2.3 Procedure. The subjects listened to the computer-generated conversations and filled in a questionnaire afterward. They were seated in front of a PC which ran a powerpoint presentation. On each slide they could click on an audio file that would then play. The audio of the system agent came from the left loudspeaker and the audio from the agent which they had to rate from the right loudspeaker. We made sure that each rater knew which loudspeaker they had to rate.

The conversations were ordered such that conversations in which the system was more talkative (faster turn changes, caused by the strategy used) than the agent alternated with conversations in which the agent was more talkative. We had five raters listen to this order (A) and five raters listened to an order in which the first four conversations of A changed position with the last four conversations of A. These results were combined for the analysis. The raters were asked to fill in the questionnaire on how they perceived the person from the right loudspeaker after each conversation.

2.2.4 Measures. In order to measure the perceived impression the users got from the agent, we adopted semantic differential scales: pairs of bipolar adjectives were placed at the extremes of 5-point Likert items. The selection of adjectives was based on previous studies by Fukayama et al. (2002) and Goldberg (1993). Because of the different characters of the Semaine system (Mcrorie et al., 2009), we also added some semantic adjectives that are related to the different characters. In general, our goal was to have a set of items that captures users' impressions of emotional and interpersonal related attributes (see Table 1). We chose these items mostly because they are useful for the Semaine characters or because we think turn-taking affects them.

Table 1

2.3 Results

In this section we present the results of the ratings of the first study. Table 2 shows the average ratings on

Table 2

Table 1. *Semantic Differential Adjectives Used in Questionnaire*

Unfriendly–friendly[a]	Disagreeable–agreeable[b]
Cold–warm[a]	Passive–active[c]
Undependable–responsible[a]	Negative–positive[c]
Rude–respectful[a]	Not aroused–aroused[c]
Distant–pleasant[a]	Unattentive–attentive[c]
Unpredictable–stable[b]	Submissive–dominant[c]
Negligent–conscientious[b]	

[a]Adjectives based on Fukayama et al. (2002).
[b]Adjectives based on Goldberg (1993).
[c]Adjectives based on the Semaine characters.

Table 2. *Results Study 1: Average Scores of 5-Point Semantic Differential Scales*

	Early		Direct		Late	
	Average	(*SD*)	Average	(*SD*)	Average	(*SD*)
Negative–positive	2.4	(1.0)	2.9	(0.9)	2.7	(0.7)
Not aroused–aroused	3.9	(1.3)	3.8	(0.8)	3.3	(1.1)
Unfriendly–friendly	2.8[a]	(1.1)	3.5	(1.0)	3.4	(0.6)
Disagreeable–agreeable	2.1[a]	(1.1)	2.7	(1.1)	2.9	(1.1)
Negligent–conscientious	3.6	(1.2)	3.8	(0.9)	3.2	(0.9)
Rude–respectful	2.6[a]	(1.3)	3.4	(1.1)	3.6	(0.9)
Distant–pleasant	2.6	(0.8)	3.3	(1.1)	2.7	(0.9)
Unpredictable–stable	2.9	(1.1)	3.3[a]	(1.0)	3.3	(1.1)
Unattentive–attentive	3.2	(1.2)	3.8	(0.9)	3.5	(0.6)
Cold–warm	2.5[a]	(0.9)	3.2	(0.9)	2.9	(0.7)
Passive–active	4.6[a]	(0.5)	4.1[b]	(0.8)	3.0[a]	(1.2)
Submissive–dominant	4.4	(1.0)	3.8	(1.1)	2.4[a]	(1.2)
Undependable–responsible	3.0	(1.1)	3.5	(0.8)	2.9	(0.8)

[a]$p < .05$.
[b]$p < .001$.

each item for the three different turn-taking strategies. In these tables, the statistical significance is indicated with one or more asterisks. The significance was calculated for every adjective by performing a two-paired t-test for all combinations (early–direct, direct–late, early–late). The type of t-test (equal variance or unequal variance) was determined by performing an f-test first. A group was said to be significantly different when both t-tests with the other groups scored a $p < .05$. So, for example, the passive value for starting late is significant because both the t-test with the direct and the early strategy resulted in a $p < .05$.

The results show that starting early is seen as more unfriendly, disagreeable, rude, cold, and more active, compared to starting at the end or after the end. Starting late is perceived as more passive, submissive, and respectful. The most pleasant person to talk with would probably be a speaker who starts

Table 3. *Results of the Factor Analysis Applied to All Scores of the Scales (Factors and Scales Ordered by Cronbachs α Value and Correlation Strength)*

	Low value–high value	Correlation
Factor 1, agreeableness α = 0.886	Distant–pleasant	0.85
	Cold–warm	0.81
	Negative–positive	0.78
	Unfriendly–friendly	0.76
	Disagreeable–agreeable	0.63
	Rude–respectful	0.58
Factor 2, assertiveness α = 0.766	Submissive–dominant	0.90
	Passive–active	0.84
Factor 3, neuroticism α = 0.703	Aroused–not aroused	0.81
	Unpredictable–stable	0.78
Factor 4, conversational skill α = 0.679	Negligent–conscientious	0.87
	Unattentive–attentive	0.72
	Undependable–responsible	0.63

slightly later than the interlocutor's turn but not too late.

When looking at the overlap resolution strategies, stopping when the agent detects overlap is perceived as significantly warmer, more passive, and more submissive than continuing normally. Continuing and talking louder is perceived as significantly more negative, more aroused, less friendly, less agreeable, more rude, and more unpredictable than stopping or continuing normally.

In order to allow for a better comparison with the active study, the items were reduced to a smaller number of factors. Using a factor analysis (principal component analysis with a Varimax rotation with Kaiser normalization), we grouped the items with a correlation > 0.5 and created four factors (see Table 3).

Table 3

The five main personality traits, such as reported by Goldberg (1993), can be used to describe the first three factors. Factor 1 can be described by agreeableness, with high values corresponding to someone who is cooperative, compassionate, friendly, and optimistic. Factor 2 is similar to the trait assertiveness (previously used by Robinson & Reis, 1989), and someone with a high assertiveness is usually extraverted, strong, and pushy.

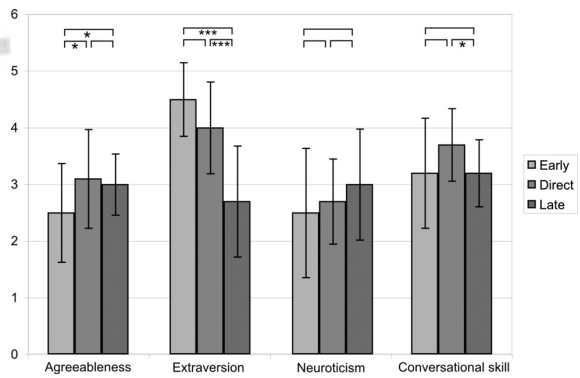

Figure 2. *The results of different startup strategies.* ∗ = $p < .05$, ∗∗ = $p < .01$, ∗∗∗ = $p < .001$.

Factor 3 is best covered with the trait neuroticism, and people with low values of neuroticism are usually more calm and emotionally stable. Factor 4 cannot be covered with a personality trait, but can be described by the term conversational skill, meaning how adept a person is in a conversation.

Figure 2 shows the mean values of the factors for the different strategies and their significant differences. It shows that starting early is seen as significantly less

Figure 2

agreeable and more extraverted, and starting late is the opposite. Interestingly, starting direct is perceived as having more conversational skill.

2.4 Conclusions

In this section we have already described a basic conversation simulator that can generate artificial conversations that closely resemble human face-to-face conversations. We have used this simulator to generate a number of conversations where strategies for timing the beginning of a turn were varied. We showed through an agent perception study how these variations in turn-management changed the impressions that people received from the agent as they listened to the various conversations. The study shows that, in general, the manipulation of turn-taking strategies can lead to different perceptions of an agent on emotional and interpersonal scales. However, in this study the users never actively participated in the conversations but were merely bystanders. What would happen if the different turn-taking strategies were used by an agent that was talking to them directly?

3 "Active" Participation and Rating of Interviews

We present the second study (called the active study) in which, in contrast with the passive study, (1) the subjects are no longer passive bystanders, and (2) the conversations are placed in a context, namely, that of an interview. In the active Wizard of Oz study, the subjects are now themselves participating in the human–agent interaction, acting as interviewees. The interviewer is an agent controlled by a human wizard who applies a certain turn-taking strategy. Participants converse with this interviewing agent and complete a questionnaire afterward. Similar to the passive study, we aim to investigate how agents' turn-taking strategies influence users' impressions of these agents. Additionally, we would like to know how these turn-taking strategies influence the users' response behaviors. We describe the turn-taking strategies used (Section 3.1), the setup of the study (Section 3.2), and how we validated whether the turn-taking strategies were correctly applied since these are error-prone (due to the human wizard who has to apply these strategies, Section 3.3). Subsequently, the results of the analysis of the subjects' ratings and their speech behavior with respect to the strategies applied are discussed in Section 3.4.

3.1 Turn-Taking Strategies

The turn-taking strategies in the active study used are almost similar to the start-up strategies early, direct, and late, described in Section 2.1. In this case, however, the overlap resolution strategies are not taken into consideration because during a pilot test, we noticed that the human subjects usually stopped speaking as soon as the agent started speaking. This means that the overlap resolution strategies could only marginally come into effect. Note that this does not say anything about whether the overlapping speech was perceived as correct or not, because the human subject could also stop speaking because he or she was almost finished. Another important difference with the passive study is that here, the strategies are applied by a human wizard rather than a computer agent that is programmed to start early, late, or direct. As humans are not machines, this means in practice that the human wizard will start its turns (i.e., questions) at variable time intervals before or after the interlocutor's turn, and that not all turns start at the exact same times with respect to the strategy. Hence, prior to the analysis of the subjects' ratings, we analyzed the speech recordings of the study and verified that the timing of the questions was in accordance with the intended strategy (see Section 3.3).

3.2 Experimental Setup

We describe how the interview sessions were set up and arranged, and what type of topics and questions were used for the interviews. Furthermore, we describe the semantic differential scales used in the questionnaire.

3.2.1 Participants. Twenty-two people participated in the study, and most of them were PhD

students. There were 16 male and six female subjects with an average age of 27.55 ($SD = 3.41$). Each participant signed a form giving consent to the use of their speech recordings for research purposes.

3.2.2 Stimuli: Scenarios of Interviews. In contrast to the passive study, the conversations had a particular content and context. It should be noted that the interview setting with the agent in the role of an interviewer constrains the flow of the conversation as the initiative lies mainly with the agent. This allows us to limit the number of utterances the agent should be able to say. The agent will ask a question, and independent of the content of the answer of the user, the agent (or human wizard rather) anticipates the user's turn-end and then asks the next question using one of the three start-up strategies.

In such a setup, the agent's questions are very important. We designed the questions such that they would be easy to answer, since a complex question can disrupt the flow of the conversation. Also, the questions asked by the agent were designed so as not to be answerable with one or two words only, since it is hard to apply a certain start-up strategy when the user only says "Yes." Examples of the questions used are "Can you tell me what movies you like?", or "Which restaurants do you recommend and why?"

Another possible problem is that certain questions can influence the results because each question has certain connotations that are perceived differently by each user. For example, person A could really like a certain topic, while person B absolutely hates it. Therefore, to improve the generalizability, we decided to create three sets of questions, each on a different topic (food and drinks, media, and school and study). By making three different groups it was possible to interchange the questions used in each session (a single conversation of a user with the system). This decreases the influence of the questions on the results. Also, by making three sets of related questions, the questions fit in the same context and are not expected to disrupt the flow of the conversation.

Another factor to consider is the voice to use. The difference between a male and a female voice can greatly influence the perception of the user. One voice may sound friendlier than another, or male and female participants may react differently to male and female voices. To control for this variable we introduced two agents: one with a male and another with a female voice. The spoken questions were synthesized with the Loquendo TTS software.[2] Each session (i.e., a single conversation of the user with the agent) thus followed a scenario that consists of a certain startup strategy (early, direct, or late), a certain voice (male or female), and a certain topic (food and drinks, media, as school and study). The exact properties—start-up strategy, voice, and topic—of each session were randomized and counterbalanced. An example interview can be found on the publisher's webpage of this paper. Fn 2

3.2.3 Recordings. Speech recordings (mono, 44.1 kHz) were made of each session with a microphone that was placed on a desk near the subject. The microphone captured both the voices of the interviewing agent and the subject. The speech recordings were made (1) to validate the human-enforced turn-taking strategies, and (2) to investigate the effects of turn-taking strategies on the subject's speech behavior.

3.2.4 Procedure. The participants were told that they would talk to a speech-to-speech dialogue system, with the agent in the role of an interviewer. They were told that we implemented different personalities in different parts of the dialogue system, and that their role was to write down, in a questionnaire, how they perceived each agent. After this introduction they talked with the agent three times (three sessions), each session with a different scenario. These scenarios were created in such a way that every possible combination and order was used at least once.

During each session the participant sat in front of a microphone and a set of loudspeakers. The human wizard (we had one wizard, who knew the purpose of the study) sat behind the participant, out of sight but in the same room. During the interview, the wizard would follow the current startup strategy by clicking on the

2. http://www.loquendo.com

Table 4. *Semantic Differential Adjectives Used in the Active Study, in Addition to the Original Set in Table 1 that was Used*

Disengaged–engaged	Competitive–cooperative
Aggressive–calm	Impolite–polite
Closed–open	Introvert–extravert
Weak–strong	Inexperienced–experienced
Pushy–laid back	Shy–bold
Arrogant–modest	Insecure–confident
Not socially skilled–socially skilled	Tensed–relaxed
Distant–close[a]	Careless[b]–responsible

[a]Was called pleasant in the passive study.
[b]Was called undependable in the passive study.

button to start the next question at the right time (as intended by the strategy).

Although the wizard in this study is not blind, that is, he is aware of the aims of the study, we expect and presume this will not affect the results of this study. The only aspect that the wizard could control was the start time of the next question—the wizard's task was to start questions at starting times as intended by the current turn-taking strategy. A check (see Section 3.3) showed that the human wizard was able to apply the strategies correctly. Our analyses are carried out on the users' responses which are only influenced by the starting times of the questions.

3.2.5 Measures: Questionnaire Design. After the interview, the subjects completed a questionnaire about how they perceived the interviewer, and rated 7-point Likert items (see Tables 1 and 4) to capture the perceived impressions of the users. The semantic differential scales used were the same as presented in Table 1. This set was extended with new items tuned to the specific interviewing setting to capture more social-skills related attributes and the interviewer's interviewing capabilities (see Table 4). For two items adopted from the passive study, the adjectives were changed because we felt that the new adjectives would better describe the interpersonal attitude that we intended to measure.

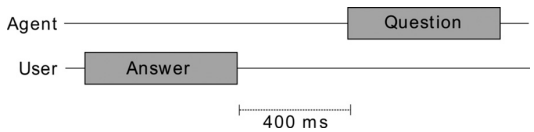

Figure 3. *An example of a gap length of 400 ms.*

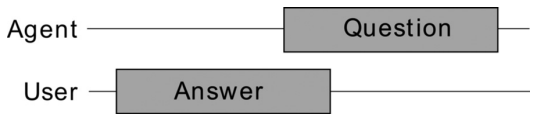

Figure 4. *An example of an instance of overlap.*

3.3 Validation of the Human-Enforced Turn-Taking Strategies

Since applying the correct strategy by a human wizard is error-prone, one requires an objective measure to see how consistently each start-up strategy was applied. We therefore annotated the speech recordings of the interviews on who was speaking when and we looked at the two objective measures of gap length and number of overlaps. The gap length is the duration of silence between the end of the user's turn and the start of the agent's next question (see Figure 3). When the turn-taking strategies are applied correctly, the gap length should be shortest for the early strategy and longest for the late strategy, and should be significantly different from each other for each strategy. The number of overlaps is the average number of overlaps per session, where an overlap is defined as an agent that starts the next question while the user is still speaking (see Figure 4). The number of overlaps should be highest for the early strategy, and lowest for the late strategy.

Table 5 shows the average gap length between the user's current turn and the following interviewer's question, grouped by the start-up strategy that was used. Note that the gap length is shortest for the early strategy and longest for the late strategy. Note that the gap length for the early strategy is not negative because the strategy was not applied correctly all the times, making the agent start too late in a number of cases. Also, when overlapping speech occurred, a gap length of 0 ms was used because it is impossible to measure how early the

Table 5. *Gap Lengths Between the User's Turn and the Following Interviewer's Question*

	Gap length	
	Mean	(*SD*)
Early	0.72	(0.69)
Direct	1.07	(0.58)
Late	1.97	(0.57)

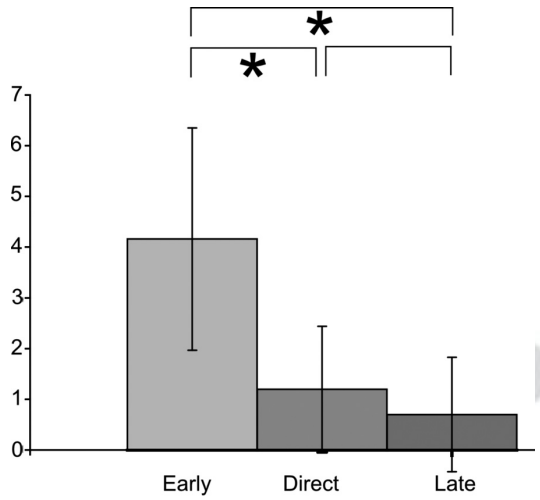

Figure 5. *Number of overlaps (user and interviewer speaking at the same time. ∗ = $p < .05$.*

wizard was. The differences between the gap lengths of the three strategies are highly statistically significant ($p < .001$).

Figure 5 Figure 5 shows the average number of overlaps grouped by the start-up strategy that was used. As shown, the number of overlaps is highest in the early strategy and lowest in the late strategy. The difference between the early strategy and the other two strategies is highly significant ($p < .001$), but the difference between the direct and the late strategy is not.

These results show that there is indeed a significant difference between the start-up strategies in accordance with the desired effect, which means that the turn-taking strategies were correctly applied by the human wizard.

3.4 Results

Here, we present the results of the active study. First, the number of semantic differential scales used in the questionnaire is reduced by applying a factor analysis that will group items into a smaller number of so-called factors. Subsequently, using these factors, we analyze the users' ratings with respect to the various turn-taking strategies to see the effect of turn-taking strategy on user impressions. In addition, we look at how the turn-taking strategies influence the user's response behavior by analyzing the user's speech from the recordings.

3.4.1 Grouping Items in the Questionnaire by Factor Analysis. As a way to reduce the number of items used (we used 27 items, see Tables 1 and 4), a factor analysis was performed to see whether some of the items could be grouped together. We used a principal component analysis, with the rotation method Varimax with Kaiser normalization. From the results we used the items with a correlation >0.5, which resulted in four different factors. These four factors, the corresponding scales, and the corresponding correlations can be found in Table 6. **Table 6**

Intuitively, the grouping of the items makes sense. Factor 1 can be described as agreeableness, one of the five main personality traits as reported by Goldberg (1993). A high level of agreeableness corresponds to someone who is cooperative, compassionate, friendly, and optimistic. Next, the adjectives strong, dominant, extravert, bold, arrogant, and pushy, can be covered under the term assertiveness (which was previously used in a similar context by Robinson and Reis, 1989). The third factor is grouped by items that have more to do with conversational skills of the agent; that is, the interviewer. The last factor seems to be related to rapport. High rapport means that the participants are in sync or on the same wavelength, which in turn means that the participants are very close and engaged during the interaction. There are four items which show too low correlations with any of the factors; these are: rude–respectful, not aroused–aroused, insecure–confident, and passive–active.

Table 6. *Results of the Factor Analysis Applied to All Scores of the Items (Factors and Items Ordered by Strength)*

	Low value–high value	Correlation
Factor 1 (agreeableness)	Cold–warm	0.86
	Unfriendly–friendly	0.78
	Tensed–relaxed	0.72
	Disagreeable–agreeable	0.70
	Aggressive–calm	0.63
	Competitive–cooperative	0.60
	Negative–positive	0.60
	Impolite–polite	0.52
Factor 2 (assertiveness)	Strong–weak	0.85
	Dominant–submissive	0.79
	Extravert–introvert	0.76
	Bold–shy	0.73
	Arrogant–modest	0.57
	Pushy–laid back	0.53
Factor 3 (conversational skill)	Inexperienced–experienced	0.72
	Not socially skilled–socially skilled	0.72
	Unpredictable–stable	0.69
	Careless–responsible	0.60
	Unattentive–attentive	0.50
Factor 4 (rapport)	Closed–open	0.82
	Disengaged–engaged	0.71
	Distant–close	0.62
	Negligent–conscientious	0.58

3.4.2 Analysis of Subjects' Ratings. In order to see the effects of the strategies on the ratings in the factors and items, an ANOVA test was performed on the data with a Bonferroni posthoc test. We used the ratings from the four factors found in the previous section, and the ratings from the four items that did not fit in these factors.

Figure 6 Figure 6 shows the results of the four factors and the rude–respectful item (the other three items that did not fit the factor analysis did not provide any significant results). All factors indicate a significant difference between the early and late strategy. The strongest factor is Factor 1, agreeableness, where the ratings for all three strategies are significantly different. Starting early is seen as more unfriendly, tensed, aggressive, competitive, and negative, and starting late is perceived as more friendly, relaxed, agreeable, cooperative and positive. For the factor assertiveness, the early strategy differs significantly from early and late, but there is no significant difference between direct and late. Starting early was rated as more strong, dominant, extravert, and bold. Similarly, for conversational skill, the direct and late strategies do not differ significantly from each other, but the early and late do. People perceived agents starting early as more inexperienced, less socially skilled, more unpredictable, more careless, and unattentive than agents starting late. Finally, rapport was perceived significantly differently in the early and late strategy.

3.4.3 Analysis of the Subjects' Speech Behavior. The previous section shows that different turn-taking strategies change the perception that the user has

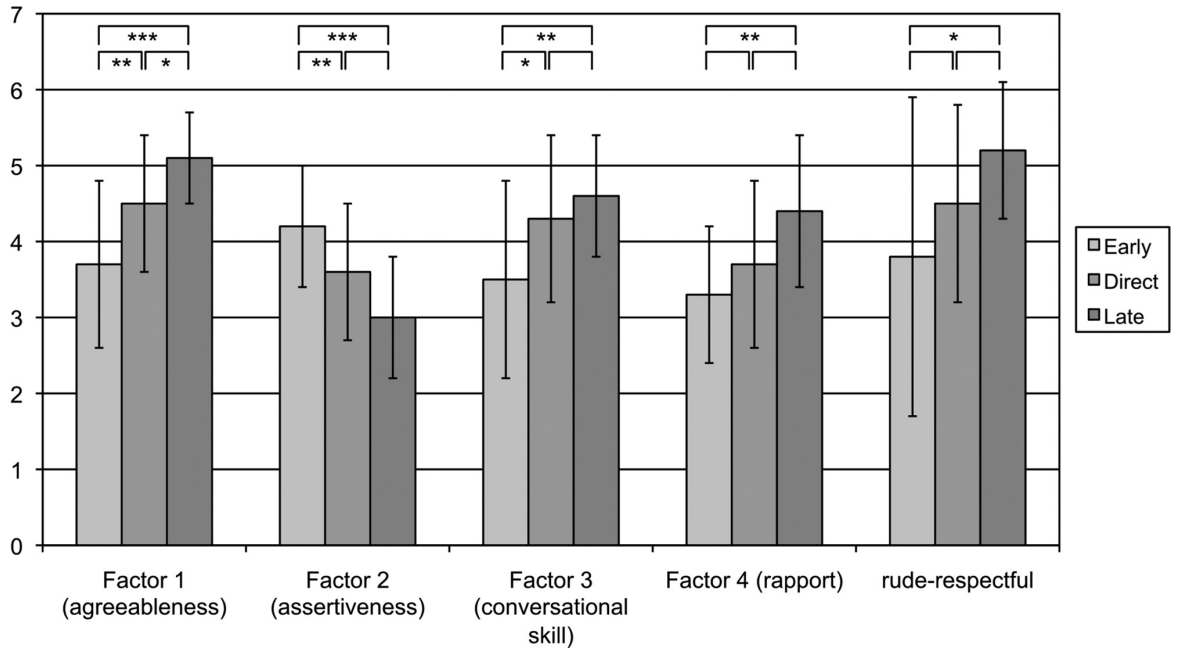

Figure 6. *The results of the different startup strategies.* $* = p < .05$, $** = p < .01$, $*** = p < .005$.

of the agent. In this section we will investigate whether different turn-taking strategies affect the behavior of the user as well. Under the assumption that people, in general, accommodate their speech behavior to the speech behavior of their conversation partner (Giles, Taylor, & Bourhis, 1973; Staum Casasanto, Jasmin, & Casasanto, 2010), we expect that certain turn-taking strategies will influence the subjects' speech behavior. Listening to the recordings of the interviews, we observed that some of the subjects started to speak faster and make shorter turns when the early strategy was applied. This observation makes sense in that people may want to avoid interruptions by speaking faster and by making shorter turns. Hence, we measured speech rate and the lengths of the subjects' turns (i.e., their answers) and analyzed these with respect to the turn-taking strategy applied.

3.4.3.1 Speech Rate. First we extracted all the user turns. By using PRAAT (Boersma & Weenink, 2001) and a script that automatically detects syllable nuclei (De Jong & Wempe, 2009) we extracted the speech rates of all turns. We define speech rate as the number of syllables per second. To compensate for people having different speech rates, we normalized the speech rates for each user by subtracting the average speech rate of that user and dividing the result by the standard deviation of the user's speech rate.

We wanted to verify whether people (consciously or unconsciously) change their behavior when they notice that the agent uses a certain turn-taking strategy. We assume that it takes some time for people to adjust to the interviewer's behavior, so we split all the conversations into two parts and compared the normalized speech rates of the first half with the speech rates of the second half to see whether accommodation took place. Table 7 contains the results of this comparison.

Table 7

Table 7 shows that especially in the early strategy people change their behavior. In the second half of the conversations with an agent using the early strategy, users increase their speech rate compared to the first half, probably because they want to finish their sentence before the agent interrupts them. So, only in

Table 7. *Comparing Mean Speech Rates of the First Half Against the Second Half of an Interview Within a Strategy*

Strategy	Mean	(*SD*)	Significance
Early half 1	3.05	(0.89)	.011[a]
Early half 2	3.33	(0.95)	
Direct half 1	3.18	(0.86)	.169
Direct half 2	3.28	(0.90)	
Late half 1	3.35	(0.64)	.202
Late half 2	3.18	(0.75)	

[a] $p < .05$.

Table 8. *Speech Rate Comparisons Across Strategies*

Strategy 1	Comparison of Mean	Strategy 2	Significance
Early half 1	<	Direct half 1	.578
Direct half 1	<	Late half 1	.200
Early half 1	<	Late half 1	.052
Early half 2	>	Direct half 2	.413
Direct half 2	>	Late half 2	.154
Early half 2	>	Late half 2	.045[a]

[a] $p < .05$.

the early strategy did the subjects accommodate to the interviewer's behavior by speaking faster.

Next, we compared the speech rates found in the different strategies with each other to find out whether a subject speaks faster in one strategy than another. The results of this comparison can be found in Table 8.

Table 8 shows that in the first part of the early strategy, people talked slower than in the first part of the late strategy (almost significant, $p = .052$). However, in the second part of the early strategy people talked faster than in the second part of the late strategy ($p < .05$). It is no surprise that the speech rates in the first halves of the different strategies are not significantly different from each other since speakers need some time to adapt to the interlocutor's behavior. Thus, when people are adapted to a certain strategy, they talk faster when they get interrupted than when the agent leaves silences between the turns. Table 7 and 8 show that users start to speak faster when the early strategy is applied; and in doing so, they speak faster than when the late strategy is applied.

Table 9. *Non-Interrupted Turn Durations*

	Early	Direct	Late	Mean	(*SD*)
Early	—	[b]	[b]	4.49	(3.87)
Direct	[b]	—	[a]	6.83	(5.06)
Late	[b]	[a]	—	12.53	(14.84)

[a] $p < .05$.
[b] $p < .001$.

3.4.3.2 Turn Durations. We also observed that when people are interrupted often, they not only speak faster, they also have shorter speech turns. To verify this observation, we extracted from the data all user turns that were not interrupted by the agent, so we only consider user turns that are completed. After measuring the duration of these turns, we compared them to each other. The results are shown in Table 9.

This table shows that the durations of the user turns in the different scenarios are all significantly different. When interrupted often (in the early strategy), people take much shorter turns, and when the agent is very slow in taking a turn, the users take much longer turns.

3.5 Conclusions

Based on the results we found, we conclude that an agent that uses a certain turn-taking strategy can indeed influence the impression that a user has of this agent, and the strategy also influences the speaking behavior of the user. Starting too early (i.e., interrupting) is mostly associated with negative and strong personality attributes: agents are perceived as less agreeable and more assertive. Leaving pauses between turns has contrary associations: this behavior is perceived as more agreeable, less assertive, and creates the feeling of having more rapport. In addition, users start to speak significantly faster in the early than in the late strategy. Users also have significantly different turn lengths when different turn-taking strategies are applied. We keep in

mind that the results are specific to this interviewing domain, and that some findings might not generalize to a free-talk conversation in which dialogue partners can talk about anything they like, or in a setting in which both dialogue partners can ask questions of each other. In Section 4, we discuss the generalizability of the results: Are the results of the passive study transferable to the results of the active study (and vice versa)?

4 Comparisons

Both studies described tried to measure similar user perceptions, although with a different method and in a different context. Eleven of the semantic differential items were used in both studies, and in this section we compare the results for those items. We do this in a qualitative way, looking mostly at the trends in the data. We also look at the results of the factor analysis of both studies.

To make a fair comparison, we only used the data from the passive study in which the continue normally overlap resolution strategy was used. Because in the Wizard of Oz setup in the active study a question from the system could not be stopped once started, the data from the passive study with the continue normally strategy is most similar to this data.

We compared the following 11 items: negative–positive, rude–respectful, unfriendly–friendly, disagreeable–agreeable, unattentive–attentive, passive–active, submissive–dominant, cold–warm, unpredictable–stable, negligent–conscientious, and not_aroused–aroused. Because the items in the passive study were ranked on 5-point scales, and the items in the active study on 7-point scales, we recoded the 5-point scales to 7-point scales linearly (Colman, Norris, & Preston, 1997; Dawes, 2008).

Figure 7 Figure 7 shows some results of this comparison.

The most apparent result is where the items in each turn-taking strategy follow the same pattern (a rising trend or a falling trend). An example can be seen in Figure 7(a), with the negative–positive item. The results from both studies show that the direct strategy is perceived as more positive than the early strategy, and the late strategy is perceived as even more positive. This similar trend can be seen in the majority of the items: negative–positive, unfriendly–friendly, disagreeable–agreeable, rude–respectful, unattentive–attentive, and cold–warm.

The negligent–conscientious item (see Figure 7[b]) follows a different trend. As can be seen in the figure, the trends from the studies are inverted. In the passive study, a later starting time of the sentence is seen as more negligent, while in the active study a later starting time is seen as more conscientious. This difference could be due to the different contexts that were used, but after the studies we sometimes received questions or remarks about this item. Therefore, we think that the difference can be explained by the fact that the users did not fully understand the item.

Figure 7(c) shows another pattern in which the results of the passive study follow the general trend of the results of the active study, with the exception that the second bar (from the direct strategy) is higher than the first bar (from the early strategy). This can be seen in the following items: unpredictable–stable, passive–active, and submissive–dominant. This might look strange at first glance, since one would expect the early and the late strategy to have an extreme value, but not the direct strategy. However, the differences between the early and the direct strategy in the passive study are very small and not significant ($p > .05$). Therefore, it is likely that the higher rating for the direct strategy in the passive study was caused by chance.

The not_aroused–aroused item follows yet another pattern (see Figure 7[d]). This figure shows that in both studies the direct strategy had the most extreme score. However, in the passive study the direct strategy scored higher than the other two strategies, and in the active study the direct strategy scored lower than the other two.

When looking at the results of the factor analysis, only the factor agreeableness is similar enough in both studies to make a comparison, because four of its items appear in both factors: disagreeable–agreeable, unfriendly–friendly, negative–positive, and cold–warm. When comparing the results, in the active study the lowest value was perceived with the early strategy, and the highest value with the

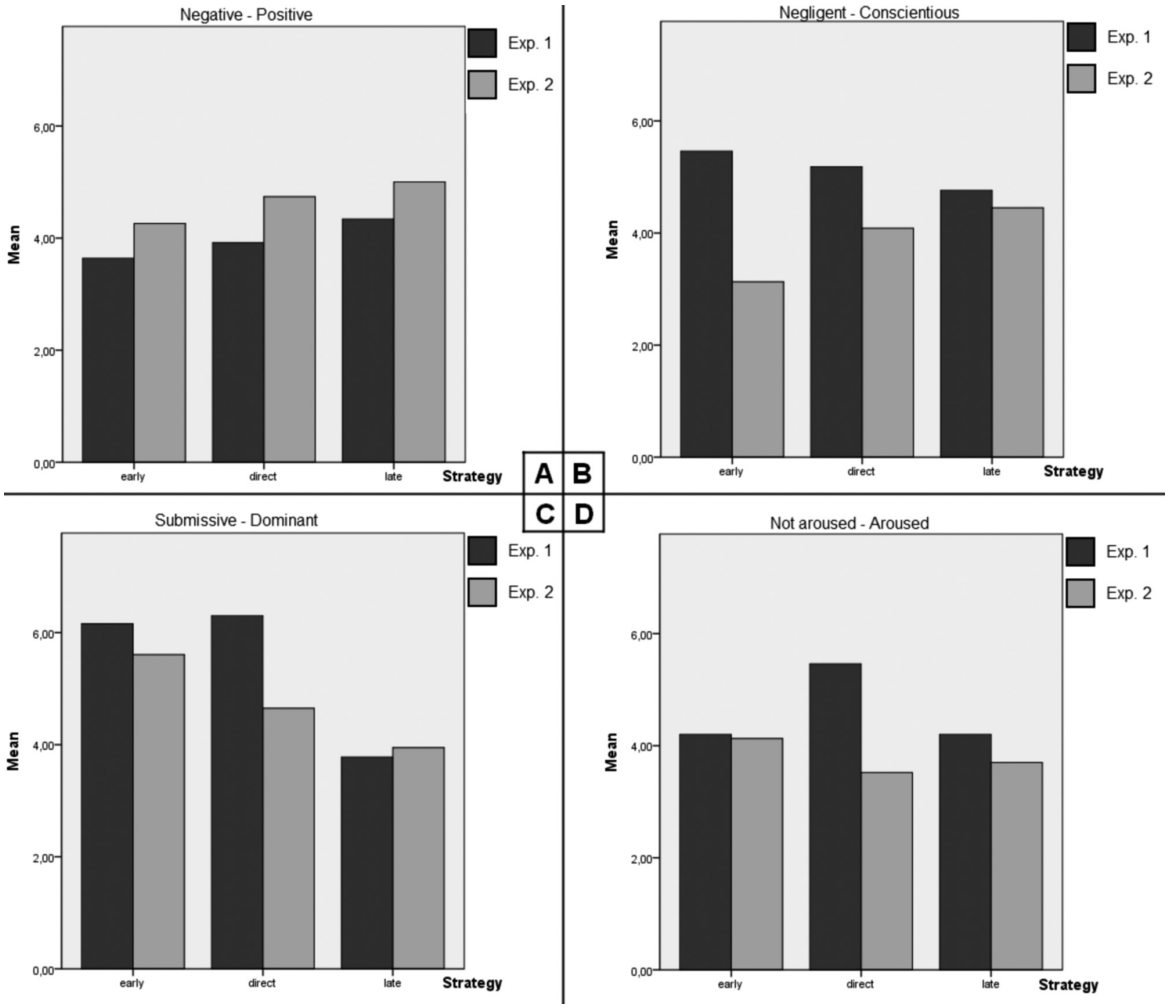

Figure 7. *Example comparisons of different scales.*

late strategy. However, in the passive study, the highest value was perceived with the direct strategy, while the late strategy has a slightly lower value (though not significantly lower). This is probably caused by the interview setting in the active study, in which it is more polite to wait longer before taking a turn.

The last thing we want to check is the difference between being a passive bystander and being actively engaged on the subjects' ratings in general. We already argued that we expect more extreme ratings in the active study because the user is actively involved. To verify this, we combined all items within each study and compared the averages of the items over the two studies, passive and active (see Figure 8). In this graph, the items were sorted such that lower scores are associated with negativity and passiveness while higher scores are associated with positivity and activity. The figure shows that the early strategy is perceived as significantly lower in the active study than in the passive study. On the other hand, the late strategy is rated as significantly higher in the active study. This is exactly the result we expected; when the participant is actively engaged in the conver-

Figure 8

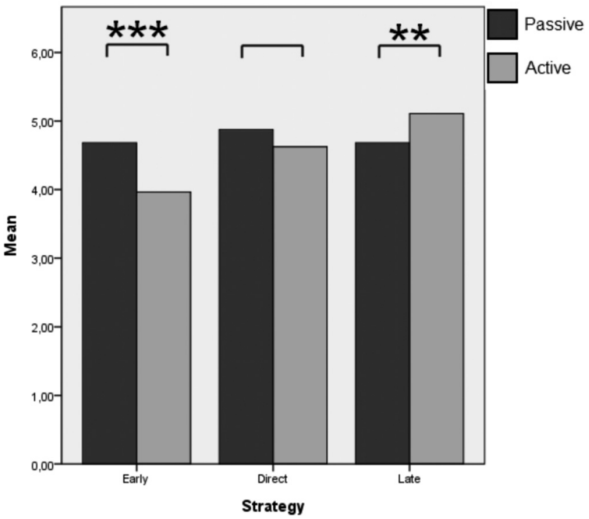

Figure 8. *The scales used in both studies combined.* $* = p < .05$, $** = p < .01$, $*** = p < .001$.

sation, then the ratings are more extreme (although we also have to keep in mind that the context is different for the two studies).

5 Conclusion and Discussion

We have described two perception studies of agents that have different turn-taking strategies, and we have evaluated how these strategies influence users' impressions of these agents, and how these strategies affect users' response behavior. How is an agent perceived when it starts speaking too early, directly after the interlocutor's turn, or when it starts speaking after a pause following the end of the interlocutor's turn? In the passive study, this question was addressed by letting human subjects listen to unintelligible, artificially generated conversations between two scripted agents. The different turn-taking strategies (i.e., early, direct, and late) were executed by scripts. In the active study, the human subjects were actively involved in the conversations; and hence, could experience and feel the effects of the turn-taking strategies themselves. In this Wizard of Oz study, the subject was being interviewed by an agent interviewer that was controlled by a human. The human wizard had to try to start each interview question at appropriate times, according to a certain turn-taking strategy. In both studies, after listening to a conversation or after being interviewed, the human subjects filled in questionnaires through which the subjects' impressions of the agents were measured. Based on these studies, we can draw several general conclusions.

Firstly, the results of both studies confirm that turn management can indeed be used to influence impressions that users have of an agent. In general, starting turns too early is associated with negative speaking behavior: agents are perceived as being dominant and highly active (see Table 2), and they are associated with negative interviewing behavior (see Figure 6). It seems that an optimal mean would be to start slightly later, but not too late, after the interlocutor's turn is finished.

Secondly, a comparison between the results of the studies shows that the impression one has of an agent does not radically change when a situational context (i.e., topics to talk about in an interview setting) is provided and when one is an active conversationalist in the study, as opposed to being a passive bystander listening to unintelligible conversations. In general, the results from the passive study can be transferred to the active study. A noticeable difference is that the ratings in general in the active studies are more extreme; users rated the agent in the active study lower in the early strategy and higher in the late strategy. Here, lower ratings correspond to a lower perceived level of assertiveness, less conversational skill, and less rapport. A possible explanation for these more extreme ratings is that the subjects who are in interaction with the interviewing agent are more engaged. They punish and reward the agent more because they are undergoing the effects of the strategies themselves rather than observing them.

Finally, an analysis of the vocal response utterances of the user in interaction with the interviewing agent showed that turn management not only influences the impression one has of an agent, but it also influences the response behavior of the user. The user seems to adapt to the interviewing agent's turn-taking strategy. Users speak significantly faster when the early strategy is applied than when the late strategy is applied (see Table 8). Furthermore, the results showed significant

differences in turn lengths between all strategies: the early strategy is associated with short turn lengths while the late strategy is associated with long turn lengths (see Table 9).

We have seen that by applying different turn-taking strategies, one can evoke different impressions. Often, researchers employ other methods to evoke a certain agent impression; for example, by changing appearance features (both visually and vocally). Turn management has often been forgotten as a means to influence users' impressions of an agent. Moreover, besides the impression changing ability of turn management, we have shown that certain turn-taking strategies can also influence the speaking behavior of a user.

One important question is how the interview setting may have influenced the results, and how specific these results are for this domain. We argue that interviews are usually governed by very strict rules, and the most important aspect of the interview is the story of the interviewee. The interviewer wants as much of this story as possible, and the interviewee wants to tell as much as he or she can. This creates a certain imbalance in the conversation, where the interviewee is more important than the interviewer. Because of this, the interviewee's behavior—for example, interrupting the interviewer or ignoring attempts to take a turn—is more easily accepted than the interviewer's behavior. Basically, the interviewer is socially required to behave in a very polite manner, and impolite behavior is punished much more harshly. The results of the active experiment show this as well: Interrupting the interviewee was perceived as more rude, less agreeable, and showing less conversational skill than giving the interviewee more time to finish by starting late. This means that this setting is an extreme case, and we argue that the results of this study can at least be generalized to other polite conversations or settings, such as a virtual receptionist or guide. We have to mention, though, that we do not have any evidence for this claim, and leave this question open for further research.

We can make several other recommendations for future work. For example, we propose to look at more subtle differences between turn-taking styles and apply these as strategies. In our studies, each strategy was uniformly applied to all turn starts, that is, in the early strategy all turns start too early, in the late strategy all turns start too late. It would be interesting to identify further different turn-taking styles and apply these more locally. Also, we performed our study mostly with PhD students. Further research is needed to verify that our claims hold for other segments of the population, although we expect that the results will not be that different. Furthermore, the role of context and type of conversation in turn management deserves more attention. Although we have identified some results that are transferable between the passive and the active study, it remains an open issue how large the influence of context and type of conversation is in these types of perception studies. Related to this, the quality of the answers that the participants produced in the active study should be studied to verify how good the answers of the users are with different turn-taking strategies. Finally, we propose to implement the turn-taking strategies through a state-of-the-art real-time end-of-turn detector and evaluate the real effects of the strategies on perception through interactive studies. One of the challenges will lie in developing a real-time end-of-turn detector that can adopt the early and direct strategies.

References

Atterer, M., Baumann, T., & Schlangen, D. (2008). Towards incremental end-of-utterance detection in dialogue systems. *Proceedings of the 22nd International Conference on Computational Linguistics*, 11–14.

Baumann, T. (2008). Simulating spoken dialogue with a focus on realistic turn-taking. *Proceedings of the 13th ESSLLI Student Session*. Hamburg, Germany.

Beattie, G. (1981). Interruption in conversational interaction, and its relation to the sex and status of the interactants. *Linguistics. An Interdisciplinary Journal of the Language Sciences La Haye*, *19*(1–2), 15–35.

Boersma, P., & Weenink, D. (2001). PRAAT, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. (2000). *Embodied conversational agents*. Cambridge, MA: MIT Press.

Colman, A. M., Norris, C. E., & Preston, C. C. (1997). Comparing rating scales of different lengths: Equivalence

of scores from 5-point and 7-point scales. *Psychological Reports*, 80, 355–362.

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? *International Journal of Market Research*, *50*(1), 61–77.

De Jong, N., & Wempe, T. (2009). PRAAT script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390.

Fukayama, A., Ohno, T., Mukawa, N., Sawaki, M., & Hagita, N. (2002). Messages embedded in gaze of interface agents—Impression management with agent's gaze *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'02)* 41–48.

Giles, H., Taylor, D. M., & Bourhis, R. (1973). Towards a theory of interpersonal accommodation through language: Some Canadian data. *Language in Society*, *2*, 177–192.

Goldberg, J. A. (1990). Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power- and rapport-oriented acts. *Journal of Pragmatics*, *14*, 883–903.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, *48*(1), 26–34.

Hirasawa, J.-I., Nakano, M., Kawabata, T., & Aikawa, K. (1999). Effects of system barge-in responses on user impressions. *EUROSPEECH'99*, 1391–1394.

Jonsdottir, G. R., & Thorisson, K. R. (2009). Teaching computers to conduct spoken interviews: Breaking the realtime barrier with learning. *Proceedings of the International Conference on Interactive Virtual Agents (IVA 2009)*, 446–459.

Jonsdottir, G. R., Thorisson, K. R., & Nivel, E. (2008). Learning smooth, human-like turntaking in realtime dialogue. *Proceedings of the 8th International Conference on Intelligent Virtual Agents (IVA2008)*, 162–175.

Kopp, S., Gesellensetter, L., Krämer, N. C., & Wachsmuth, I. (2005). A conversational agent as museum guide – design and evaluation of a real-world application. In T. Panayiotopoulos, J. Gratch, R. Aylett, D. Ballin, P. Olivier, & T. Rist (Eds.), *Lecture Notes in Computer Science*, Vol. 3661. *Intelligent Virtual Agents* (pp. 329–343). Berlin: Springer-Verlag.

Mcrorie, M., Sneddon, I., Sevin, E., Bevacqua, E., & Pelachaud, C. (2009). A model of personality and emotional traits. *Proceedings of the 9th International Conference on Intelligent Virtual Agents*, IVA '09, 27–33.

Ochs, M., Pelachaud, C., & Sadek, D. (2008). An empathic virtual dialog agent to improve human–machine interaction. *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems, Vol. 1*, (pp. 89–96).

Oreström, B. (1983). *Turn-taking in English conversation.* Lund, Sweden: Liber.

Raux, A., & Eskenazi, M. (2008). Optimizing endpointing thresholds using dialogue features in a spoken dialogue system. *Proceedings of SIGdial 2008*, 1–10.

Robinson, L. F., & Reis, H. T. (1989). The effects of interruption, gender, and status on interpersonal perceptions. *Journal of Nonverbal Behavior*, *13*(3), 141–153.

Schegloff, E. (2000). Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, *29*(1), 1–63.

Schlangen, D. (2006). From reaction to prediction experiments with computational models of turn-taking. In *Proceedings of Interspeech 2006*, 2010–2013.

Staum Casasanto, L., Jasmin, K., & Casasanto, D. (2010). Virtually accommodating: Speech rate accommodation to a virtual interlocutor. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 127–132.

Ter Maat, M., & Heylen, D. (2009). Turn management or impression management? In Z. M. Ruttkay, M. Kipp, A. Nijholt, & H. H. Vilhjálmsson (Eds.), *Proceedings of the 9th International Conference on Intelligent Virtual Agents (IVA2009)*, 467–473.

Ter Maat, M., & Heylen, D. (2010). Generating simple conversations. In A. Esposito, N. Campbell, C. Vogel, A. Hussain, & A. Nijholt (Eds.), *Development of multimodal interfaces: Active listening and synchrony*, *Lecture Notes in Computer Science*, Vol. 5967 (pp. 92–101). Berlin: Springer-Verlag.

Ter Maat, M., Truong, K. P., & Heylen, D. K. J. (2010). How turn-taking strategies influence users' impressions of an agent. In J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, & A. Safonova (Eds.), *Proceedings of the International Conference on Intelligent Virtual Agents (IVA)*, *Lecture Notes in Computer Science*, Vol. 6356 (pp. 441–453). Berlin: Springer-Verlag.

Trimboli, C., & Walker, M. B. (1984). Switching pauses in cooperative and competitive conversations. *Journal of Experimental Social Psychology*, *20*(4), 297–311.

Van Deemter, K., Krenn, B., Piwek, P., Klesen, M., Schröder, M., & Baumann, S. (2008). Fully generated scripted dialogue for embodied agents. *Artificial Intelligence*, *172*(10), 1219–1244.