Should Agents Speak Like, um, Humans? The Use of Conversational Fillers by Virtual Agents

Laura M. Pfeifer and Timothy Bickmore

Northeastern University College of Computer and Information Science
202 WVH, 360 Huntington Avenue, Boston, MA 02115
{laurap,bickmore}@ccs.neu.edu

Abstract. We describe the design and evaluation of an agent that uses the fillers *um* and *uh* in its speech. We describe an empirical study of human-human dialogue, analyzing gaze behavior during the production of fillers and use this data to develop a model of agent-based gaze behavior. We find that speakers are significantly more likely to gaze away from their dialogue partner while uttering fillers, especially if the filler occurs at the beginning of a speaking turn. This model is evaluated in a preliminary experiment. Results indicate mixed attitudes towards an agent that uses conversational fillers in its speech.

Keywords: embodied conversational agent, fillers, filled pause, gaze.

1 Introduction

Embodied Conversational Agents (ECAs) have existed for over a decade and have been used for a variety of purposes, such as education, counseling and social engagement [1]. A goal of these agents is often to emulate human behavior through both verbal and non-verbal strategies. However, agents have traditionally used perfectly fluent language in their speech, despite the fact that human dialogue consistently contains disfluencies such as restarts, rephrases, and filled pauses [2].

While the use of filled pauses, or conversational fillers, such as *um*, *uh*, *like*, *you know*, etc., may be seen as a type of disfluency, many believe that fillers indeed carry linguistic signals, and are in the same class as English interjections, similar to *ah*, *well*, and *oh*, in which the speaker gives the listener cues about the dialogue structure [3]. For this paper, we will follow Clark and Fox Tree's definition of a conversational filler which, contrary to common belief, does not exist simply to "fill a pause" in speech. Rather, fillers are considered as actual words, and are used by speakers as collateral signals, in effect, to manage the on-going performance of the dialogue [4].

In this paper we explore verbal and non-verbal conversational strategies that agents can use to more closely match human-human dialogue. We present an empirical study of face-to-face dialogue by humans, which indicates a strong relationship between gaze patterns and conversational fillers in speech. We then evaluate this model in a preliminary randomized experiment with an ECA that uses fillers, and explore human reactions to the use of these fillers by agents.

Zs. Ruttkay et al. (Eds.): IVA 2009, LNAI 5773, pp. 460–466, 2009.

2 Background and Related Work

Conversational fillers are a common form of grounding in dialogue and serve a variety of purposes. A speaker might use the fillers *um* and *uh* to indicate to the listener that they are searching their memory for a word [5], that they want to hold or cede their turn in the conversation [4], or to signal hesitation, doubt or uncertainty [6].

The fillers *um* and *uh* can also signal new, upcoming information. In one study, listeners were more successfully able to identify words from speech recordings if those words were preceded by the word *uh* [7]. In another study listeners selected a picture on a computer screen more quickly if the spoken description was preceded by the words "*theee, um*" [8].

Analysis of human-computer speech has found that people speak simply and directly when talking to a computer, and use significantly fewer disfluencies than normal [9]. However, no studies thus far have examined the use of fillers by an ECA, and the effects those fillers might have on a human dialogue partner.

3 The Use of Fillers and Gaze Behavior by Human Speakers

We conducted an empirical study to develop a model of behavior involving conversational fillers in face-to-face conversation. Since we planned to implement this behavior in an ECA, we also modeled some of the non-verbal behavior that accompanied the delivery of fillers. Preliminary analyses indicated that gaze behavior frequently co-occurred with the use of fillers, so we focused our initial non-verbal behavior modeling efforts on the gaze behavior of the speaker.

Five people participated in the study recruited via flyers posted around the Northeastern University campus, and were compensated for their time. Participants had to be at least 18 years of age, with English as their native language. Ages ranged from 19 to 57 years old (mean=32.6) and 60% were female. The study took place in the Human-Computer Interaction laboratory at Northeastern University. Participants were consented, completed a demographic questionnaire, and were then told they would be having a conversation with a professional exercise trainer about their exercise behavior. The trainer and her "client" were introduced, were seated facing each other, and the experimenter left the room. All conversations were videotaped for later analysis. The trainer and clients were blind to the purpose of the study.

3.1 Use of Fillers in Dialogue

Conversational fillers used by clients were coded throughout the entire conversation using Anvil [10]. Table 1 presents a summary of the conversations and use of fillers by the clients. The number of fillers within a dialogue turn was highly correlated with the number of words in the turn, r=.81, p<.01, and with the length (seconds) of the turn r=.73, p<.01. Throughout the conversations, *um* and *like* were the most commonly used fillers, with complete frequencies shown in Table 2. Use of fillers among clients showed near, but not significant differences by Friedman's Test, $\chi^2(6)=11.02$, p=.088, indicating a high amount of inter-subject variation regarding which fillers

were commonly spoken, and how often they were used throughout the dialogue. Thirty percent of dialogue turns by clients contained a filler uttered as the first word. *Um* was the most common example (63%) of a filler used at the beginning of a turn, and 64% of all occurrences of *um* were located at the start of a dialogue turn.

Table 1. Client behavior in conversations analyzed

Client	Time Speaking (Seconds)	Num Turns	Time Uttering Fillers (Seconds)	% of Time Uttering Fillers	% of Turns with Fillers
1	238	36	29	12.33%	47.22%
2	491	88	41	8.33%	54.55%
3	642	82	35	5.39%	39.02%
4	895	64	72	8.07%	62.50%
5	535	56	104	19.40%	71.43%
Mean	560	65	56	10.70%	54.94%

Table 2. Number of fillers spoken by clients, per conversation turn

Client	*um*	*uh*	*like*	*So*	*just*	*you know*	*kind of*	Mean
1	0.81	0.03	1.25	0.25	0.47	0.08	0.11	0.43
2	0.76	0.00	0.26	0.18	0.41	0.11	0.00	0.25
3	0.16	0.28	0.16	0.16	0.10	0.28	0.04	0.17
4	2.02	0.31	0.25	0.38	0.23	0.58	0.11	0.55
5	0.29	0.05	5.52	1.04	1.45	0.07	0.13	1.22
Mean	0.81	0.13	1.49	0.40	0.53	0.23	0.08	0.52

3.2 Model of Gaze Behavior for Speech Containing Fillers

We coded client gaze patterns throughout the entire conversation. Gazes were divided into nine categories, according to the client's perspective: up, up and to the left, up and to the right, down, down and to the left, down and to the right, left, right, and at the trainer. Friedman's Test shows significant differences of gaze patterns among the clients both while speaking and uttering fillers, $\chi^2(8)=21.61$, $p<.01$, and while speaking without uttering fillers, $\chi^2(8)=21.71$, $p=.005$, indicating a high amount of inter-subject variation in gaze patterns.

Clients spent significantly more time looking *at* the trainer while speaking without using fillers (61%), and *away* from the trainer while uttering fillers (57%), *paired-t* (4) = -6.45, $p<.01$. We also analyzed client gaze patterns according to their location within a conversational turn (Table 3). When a conversational filler was uttered at the beginning of a turn, client gaze shifts were more likely to be directed *away* from the trainer, *paired-t* (4) = 3.33, $p<.05$. During the middle or end of a turn, clients were equally likely to shift their gaze *away* or *towards* the trainer.

Table 3. Percent of client gaze changes that are directed towards or away from the trainer, based on the location within a turn

Location Within a Turn	Speaking Filler		Not Speaking Filler	
	At Trainer	Away	At Trainer	Away
Beginning	30.69%	69.31%	63.59%	36.41%
Mid or End	49.53%	50.47%	49.53%	50.47%

4 Conversational Fillers and Gaze Model Implementation

An existing virtual agent framework was modified to provide appropriate co-occurring gaze and speech behavior during the presence of conversational fillers [11]. Co-verbal behavior is determined for each utterance using the BEAT text-to-embodied-speech system [12]. User contributions to the conversation are made by selecting an item from a multiple-choice menu of utterance options, updated at each turn of the conversation.

We observed gazing in the left direction to be the most common gaze-away pattern while uttering fillers, so we extended the agent animations to allow for speaking while gazing to the left (Fig. 1). We also extended the system to be able to speak while gazing at a document artifact held in the agent's hands. In previous work, we found that when explaining a document, humans gazed at the document between 65-83% of the time [11]. Upon re-analysis and consideration of fillers, we found that time gazing at the document was not affected by use of fillers.

Fig. 1. Co-occurring gaze and speech by the agent

5 Preliminary Evaluation

We conducted a preliminary evaluation to test the efficacy of an agent that uses conversational fillers in its speech. For this experiment, we used the fillers *um* and *uh*, as they have been shown to have comprehension effects [7]. In order to examine the effects of fillers in various conversational styles, each participant had two conversations with the agent: a social conversation and an educational conversation. The *social*

conversation used a between-subjects (FILLERS vs. NO-FILLERS) experimental design and the *educational* conversation used a within-subjects (FILLER-PHRASE vs. NO-FILLER-PHRASE) experimental design. During the *educational* conversation, all participants interacted with an agent that used conversational fillers, with half of the key concepts preceded in speech by a filler. We hypothesize that ratings of satisfaction and naturalness will be higher for participants in the FILLERS condition of the *social* conversation, and that participants will perform better on test questions regarding content of the *educational* conversation, if the content was preceded in speech by a filler (FILLER-PHRASE).

The Loquendo text-to-speech engine was chosen for the experiment, and the intonation and timing of each filler was adjusted to sound as natural as possible. In most cases, this consisted of lowering the pitch, reducing the voice speed, and following the filler by a short pause (50-100 ms). The social script consisted of 14 turns of dialogue and lasted approximately one and a half minutes, and the educational script consisted of 34 turns of dialogue and lasted approximately four minutes. The educational dialogue contained a shortened, simplified version of an agent-based explanation of a hospital discharge pamphlet [11]. The participants talked to the agent using a Wizard-Of-Oz setup, and were instructed to say one of the allowed utterances displayed on a menu during each turn.

5.1 Measures

Along with basic demographics, we assessed computer attitudes with the question "How do you feel about using computers?" We also created a knowledge test based on the educational dialogue, containing eight questions on the content of the dialogue. Evaluation questionnaires were also developed, assessing satisfaction, trust, likability, perceived knowledge of the agent, along with naturalness of the dialogue and naturalness of the agent's eye-gaze behavior, all evaluated on 7-point scales. All measures were administered via paper-and-pencil.

5.2 Procedure

Twenty-three people participated in the evaluation study, aged 19-66 years, 70% female. Fifty-seven percent of participants indicated neutral attitudes towards computers, with the rest indicating positive attitudes towards computers. After informed consent was obtained, participants were randomized into conditions, and demographic questionnaires were administered. The experimenter left the room, the agent and participant conducted the social dialogue, and attitudinal measures on the agent and the conversation were administered. Participants were then asked to role-play that they were in the hospital, and the agent would explain a hospital discharge booklet to them. The experimenter left the room, and the agent and participant conducted the educational conversation. Afterwards, participants completed the knowledge test and attitudinal measures, followed by a semi-structured interview to obtain impressions of the experiment and agent.

5.3 Results

Attitudinal measures towards the agent after the social and educational conversations are shown in Table 4. There were no significant effects of study conditions on attitudes towards the agent. Trends indicate that participants in the NO-FILLERS condition *liked* the agent more than those in the FILLERS condition, $t(21)=1.69$, $p=.10$.

Table 4. Attitudinal measures towards the agent (mean (SD)). Measures are on a scale from 1 (not at all) to 7 (very much).

Question	Social Conversation			Educational Conversation
	Fillers	No Fillers	*p*	
How *satisfied* are you with the conversation experience?	6.00 (1.18)	5.58 (.996)	0.37	5.61 (1.20)
How much do you *trust* Elizabeth?	5.68 (1.10)	5.50 (.905)	0.67	5.72 (1.01)
How much do you *like* Elizabeth?	5.32 (.783)	5.92 (.900)	0.10	5.65 (1.03)
How *knowledgeable* was Elizabeth?	5.36 (.924)	5.17 (1.47)	0.76	5.96 (0.98)
How *natural* was the speaking style of Elizabeth?	4.09 (1.76)	4.33 (1.97)	0.71	4.22 (1.62)
How *natural* was the eye-gaze behavior of Elizabeth?	4.45 (1.80)	4.58 (1.50)	0.85	4.87 (1.66)

Knowledge test scores of the educational conversation were coded on the following scale: 0=incorrect, 0.5=partially correct, 1=correct. A comparison of the scores of test questions on dialogue content preceded in the conversation by a filler (FILLER-PHRASE), vs. questions on dialogue content not preceded by a filler (NO-FILLER-PHRASE) found no significant differences, *paired-t* (22) = .789, $p=.44$.

During semi-structured interviews, 22 of the participants were asked if they recognized whether or not the agent used the fillers *um* and *uh* during the dialogue (all participants heard fillers in at least one conversation). Eight participants (36%) thought that the agent did not use fillers, 4 (18%) were not sure, and 10 (46%) recognized the use of fillers.

Also during interviews, 15 participants volunteered an opinion, negative, positive or neutral, towards agents using fillers in their speech, or were asked by the interviewer, "How do you feel about a computer character using the words *um* and *uh* in its speech?" Participants with positive attitudes towards computers were significantly more likely to indicate that the usage of fillers by agents was a positive aspect of the conversation, compared to participants with neutral attitudes towards computers, $\chi^2(2)=8.89$, $p=.01$.

Overall, participants reported mixed feelings about interacting with an agent that uses fillers. Five participants indicated that the use of fillers by a conversational agent seemed inappropriate, given that computers have the ability to speak perfectly, and another five participants indicated that the usage of fillers by the agent was a positive aspect of the conversation and "humanized" the experience.

6 Discussion

Should agents speak like humans? It appears to be open for additional research. At this time we do not see significant differences on satisfaction and naturalness, and we did not observe the recall effects associated with fillers in human-human dialogue. Previous work showed these recall effects after participants listened to audio recordings, and it is possible that the addition of an ECA - providing an audio *and* visual signal - mitigates the effects. Another possible reason is that our agent's production of fillers and accompanying nonverbal behavior need further refinement in order to match human behavior. Although this preliminary evaluation is limited, it provides us with a good overview of attitudes towards the use of fillers by agents.

Our future work is focused on extending the evaluation study with a broader range of participants having various levels of computer attitudes, age, and personality. We also intend to evaluate a wider range of fillers by agents, such as *like*, *you know*, etc. Finally, we plan to evaluate the use of fillers by agents that speak with a text-to-speech engine to agents that speak with a human-recorded voice.

Acknowledgments. Many thanks to Jenna Zaffini and Donna Byron for their assistance with the study. This work was supported by NSF CAREER IIS-0545932.

References

1. Cassell, J.: Embodied Conversational Agents. MIT Press, Cambridge (2000)
2. Fox Tree, J.E.: The Effects of False Starts and Repetitions on the Processing of Subsequent Words in Spontaneous Speech. J. Mem. Lang. 34, 709–738 (1995)
3. Swerts, M.: Filled Pauses as Markers of Discourse Structure. J. Pragmat. 30, 485–496 (1998)
4. Clark, H., Fox Tree, J.E.: Using Uh and Um in Spontaneous Speaking. Cognition 84, 73–111 (2002)
5. Goodwin, C.: Forgetfulness as an Interactive Resource. Soc. Psychol. Q. 50, 115–130 (1987)
6. The American Heritage dictionary of the English language. Houghton Mifflin, Boston (2006)
7. Fox Tree, J.E.: Listeners' Uses of Um and Uh in Speech Comprehension. J. Mem. Cognit. 29, 320–326 (2001)
8. Arnold, J.E., Fagnano, M., Tanenhaus, M.K.: Disfluencies Signal theee, um, New Information. J. Psycholinguist Res. 32, 25–36 (2003)
9. Oviatt, S.: Predicting Spoken Disfluencies During Human-Computer Interaction. Comp. Speech Lang. 9, 19–35 (1995)
10. Kipp, M.: ANVIL – A Generic Annotation Tool for Multimodal Dialogue. In: 7th European Conference on Speech Communication and Technology, pp. 1367–1370 (2001)
11. Bickmore, T.W., Pfeifer, L.M., Paasche-Orlow, M.K.: Health document explanation by virtual agents. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 183–196. Springer, Heidelberg (2007)
12. Cassell, J., Vilhjalmsson, H., Bickmore, T.: BEAT: The Behavior Expression Animation Toolkit. In: SIGGRAPH 2001: Proceedings of the 28th annual conference on computer graphics and interactive techniques, pp. 477–486 (2001)