Fluid Semantic Back-Channel Feedback in Dialogue: Challenges & Progress

Gudny Ragna Jonsdottir[1], Jonathan Gratch[2], Edward Fast[2], Kristinn R. Thórisson[1]

[1] CADIA / Department of Computer Science, Reykjavik University
Ofanleiti 2, IS-103 Reykjavik, Iceland
[2] University of Southern California, Institute for Creative Technologies,
12374 Fiji Way, Marina del Rey, CA 90292
{gudny04, thorisson}@ru.is,{gratch, fast}@ict.usc.edu

Abstract. Participation in natural, real-time dialogue calls for behaviors supported by perception-action cycles from around 100 msec and up. Generating certain kinds of such behaviors, namely envelope feedback, has been possible since the early 90s. Real-time backchannel feedback related to the content of a dialogue has been more difficult to achieve. In this paper we describe our progress in allowing virtual humans to give rapid within-utterance content-specific feedback in real-time dialogue. We present results from human-subject studies of content feedback, where results show that content feedback to a particular phrase or word in human-human dialogue comes 560-2500 msec from the phrase's onset, 1 second on average. We also describe a system that produces such feedback with an autonomous agent in limited topic domains, present performance data of this agent in human-agent interactions experiments and discuss technical challenges in light of the observed human-subject data.

Keywords: face-to-face dialogue, real-time, envelope feedback, content feedback, interactive virtual agent.

1 Introduction

The fluidity and expressiveness of human dialogue presents significant challenges to developers of embodied conversational agents. Partners in a conversation effortlessly exchange verbal and nonverbal signals that help regulate the interaction and provide key semantic and emotional feedback. The variety of communication channels involved (speech content and prosody, facial expressions, gestures, postures, respiration, etc.) and the rapidity with which people can produce and process such information, tax the technical capabilities of autonomous agents intended to capture natural human speech. Therefore, contemporary conversational systems typically focus on a small number of channels and enforce explicit, structured turn taking. The resulting interaction is more akin to conversations with astronauts on the moon than normal face-to-face interaction, in both structure and pacing.

The present work aims to understand better key perceptual and behavior mechanisms in people's communicative behavior and to move closer to autonomous agents capable of fluid, dynamic speech interaction.

Several systems have attempted to improve the fluidity of virtual human feedback by providing back-channel feedback to non-lexical features of human speakers [1],[2],[3],[4]. Due to technological limitations the content of the speech has been largely ignored, at least as a source of realtime feedback. Thórisson's autonomous agents J. Jr. [2] and Gandalf [3] produced believable gaze, back-channel feedback and turntaking in real-time, based on automatic analysis of prosody and gesture input, without attending to speech content. People, too, can give such feedback without attending to speech content, termed *envelope feedback* by Thórisson [3],[5] or *generic feedback* by Bevales et al. [6]. Bavelas and colleagues demonstrated that people can produce well-timed nods even while engaged in a demanding distraction task that prevented them from attending to the speaker's content, in support of the Thórisson's hypothesis that separate cognitive mechanisms are responsible for envelope and semantic feedback [5]. Envelope feedback plays an important interaction function, signaling "everything is OK, please continue/I'm paying attention", and can contribute to a sense of mutual understanding and liking, factors associated with rapport [7]. Agents that provide such feedback can improve speaker engagement and speech fluency [1],[8].

Content feedback is back-channel feedback that makes reference to the content of the speech.[1] For example, Bavelas et al. [6] found that storytellers expected emotional feedback from their listeners to key events in the story and found it hard to construct effective narratives without it. In their study some listeners were required to perform a demanding distraction task while listening. The listeners were able to provide some envelope feedback (nods and vocalizations such as "mm-hmm") while listening but they were unable to produce responses related to content (such as wincing, looking surprised, etc.). Narrators found this lack of feedback disruptive, and generated less structured and less satisfying stories.

In this paper we describe our progress in allowing virtual humans to give within-utterance *content feedback* to user speech. As we will illustrate, this is a challenging problem in terms of the rapidity with which people expect such feedback. The next section describes the requirements such a system must satisfy. We describe a study that elicits content feedback and discuss the form and temporal dynamics of such behavior. Section 3 describes our results in achieving this performance through real-time speech recognition in an integrated system.

2 The character of content feedback

To better support the study of the general constraints that govern the response characteristics, timing and individual variability of listeners' behavior we constructed a database of naturalistic listener feedback. The database serves as a reference point for judging the effectiveness of automated techniques.

[1] Bavelas et al [6] use the term "specific feedback" to refer to feedback produced in response to the content/meaning of speech. We prefer the more descriptive term "content feedback".

2.1 Human subject study: Listener feedback elicitation

The main goal of this study was to identify features of a storyteller's behavior that are correlated with content-related back-channel feedback, and that a computational system might reasonably be able to identify and react to in real-time. Eighty people (60% women, 40% men) from the general Los Angeles area participated in this study. They were recruited using Craig's List and were compensated $20 for one hour of their participation. We used a video clip taken from the *Edge Training Systems Sexual Harassment Awareness* video. The video clip was merged from two segments: The first is about a woman at work who receives unwanted attention through the Internet from a colleague at work, and the second is about a man who is confronted by a female business associate, who asks him for a foot massage in return for her business.

There were two experimental conditions: the *Face-to-face* condition (n=40) and the *Mediated* condition (n=40), to which participants were randomly assigned. In each condition, two participants were randomly assigned the role of storyteller (*Speaker*) or story listener (*Listener).*[2] In both Face-to-face and Mediated conditions the Speaker viewed the video while the Listener waited in another room. Face-to-face condition: When the Speaker had finished viewing the video, the Listener entered and the Speaker told him/her about the video. Mediated condition: When the Speaker had finished viewing the video, the Listener entered and sat across from the Speaker but separated by a physical barrier; the Listener saw a live video image of a human Speaker displayed on a large monitor; the Speaker saw a computer generated avatar that matched the Listener's head movements via a vision-based tracking system and told him/her about the video. In all conditions the Speaker and Listener were on opposite sides of a table separated by 2 meters.

2.2 Analysis & Results

No significant differences were found between the Mediated and Face-to-face conditions on the dependent variables reported below. We collapsed data across conditions for the purpose of analysis. One listener was excluded from the analysis due to a failure of the recording equipment yielding a final sample of thirty-nine Listeners.

Fig. 1. Subjects showed a range of facial upon hearing the term "foot massage." These included expressions of disgust, lowered brows, raised brows, gaze shifts, and various expressions of amusement. Subjects responded rapidly, within 350 milliseconds on average after "foot massage" was spoken. A quarter of the subjects showed no obvious response.

Lexical Feedback Markers. There was considerable similarity in the words Speakers used to describe events. Facial expressions of Listeners would often immediately follow certain Speaker phrases; 36 of 39 Speakers mentioned the exact

[2] The two conditions were created to address a secondary goal: to tease apart what aspects of listener feedback are crucial for speakers. For example, does the speaker need to see the listener or could a graphical representation of the listener be just as effective? This paper only focuses on the first goal. For the purpose of this article, the two conditions simply represent different methods to elicit feedback.

phrase "foot massage" and in 26 of these Listeners rapidly thereafter displayed a visible facial expression. We refer to these key phrases as *lexical feedback markers (LFMs).*[3]

Listener Facial Feedback: Subjects produced a variety of facial feedback during the narratives. We explored peoples' responses associated with the LFM "foot massage" and examined Listeners' facial responses to the first mention of the phrase; most conveyed some notion of surprise but the specific facial response varied widely in its form and intensity (Figure 1). Responses included raised brows (8 subjects), smiles (6 subjects), grimaces (6 subjects), gaze shifts (4 subjects), and laughter (3 subjects). In some cases there was a complex unfolding of expressions (e.g., a brow raise shortly followed by a smile) as predicted by theories of facial expressions [9].

Listener Feedback Delays: Table 1 summarizes the listener feedback delays (average for 22 subjects[4]) for the "foot massage" LFM. (Since people might understand the phrase before it is fully completed, we report reaction times from both the beginning and completion of the phrase.) Subjects showed significant variability in the timing of their responses. In general, feedback was quite rapid, within 400 msec of the completion of the LFM. [5] In two cases back occurred in midst of the LFM.

Table 1. Subject reaction times.

	Time to Produce Lexical Feedback Marker	Delay between Lexical Feedback Marker and Listener Expression	
		From its start	From its completion
Avg time	775 msec (σ=153)	1038 msec (σ=418)	344 msec (σ=444)
Min time	550 msec	560 msec	-220 msec
Max time	1080 msec	2510 msec	1630 msec

3 System Design & Setup

We constructed a multi-module system to produce content feedback by incorporating continuous speech recognition into the Rapport Agent of Gratch et. al [1], an architecture for exploring the social impact of nonverbal behavior (see Figure 2). The agent was set up to recognize and react to the LFMs identified in the elicitation study. The agent can produce natural envelope feedback in responses to body movements and speech prosody, as well as content-related facial feedback, as seen in the human subject study. In its standard configuration, the Rapport Agent generates envelope feedback by real-time analysis solely from a narrator's speech and body movements using a prosody detector, named LAUN, including backchannel opportunity points [2],[10], disfluencies, questions, and loudness. Using the Watson vision-based gesture detector [11], it detects speaker gestures including head nods, shakes, gaze shifts and posture shifts. Neither Watson nor LAUN can extract content or meaning from communicative behavior.

[3] We do not claim that these terms necessarily elicited the listener feedback, but they closely preceded it and would serve as reasonable makers for a computer system to attempt to recognize and respond to as a proxy for true understanding. (See list of terms in Table 2).

[4] Accurate timing statistics for the remaining 24 subjects was not yet complete at the time of submission but results appear comparable.

[5] The machine running Dragon, on which all measurements were made, is a 2x dual core 2.61GHz Intel Pentium-class processor running Windows XP with 3.37GB RAM.

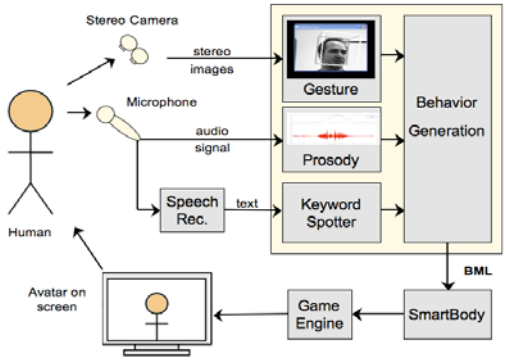

Fig. 2. Our approach to semantic feedback incorporates continuous speech recognition to the existing Rapport Agent approach to providing envelope feedback via gesture and prosody detection.

Speech Recognition. We use a continuous, large vocabulary, general-purpose dictation speech recognizer. Dragon has proven to have relatively good accuracy, and user-independence, as shown in our own test (even without the individual training recommended by the manufacturer). Normal operation of Dragon is to wait for silences before starting to process; however, between pauses (larger than 100 msecs) it produces hypotheses about what has said so far. These are less accurate than the final response, it may mean the difference of several seconds to wait for the final output. We need timestamps and confidence of words, as well as *n*-best hypothesis of the uttered phrase, which Dragon provides through their API. However, the time to produce hypotheses varies considerably (see Figure 3). When the final hypothesis is released, a final timestamp estimation for each word is produced, representing the actual time (in the past) that the words were spoken by the user.

Pattern Matching. We use a continuous speech recognizer to extract text from the Speakers' speech; specialized pattern matchers extract meaning from the recognized text. Since the time to search the text is a linear function of the number of words/phrases we are looking for, we run multiple matchers in parallel, each matching a limited set of phrases with relatively simple techniques. Data collected in the human subject experiment was used to construct the patterns for detecting the LFMs.

Listener feedback. All perception modules (prosody, behavior and lexical) communicate with a reactive Behavior Generation system which probabilistically selects a single appropriate feedback response given the recognized input and internal state information (details described in Gratch et al., [1]). Behaviors represented in the Behavior Markup Language (BML) [12] are passed to an animation sys-

Table 2. Accuracy of recognition per LFM category.

Category	Lexical Feedback Markers	Accuracy
Foot massage	Foot massage, Foot rub, Rub her feet	63%
Harassment	Harass	94%
Sexual	Sexual	93%
Legal	Legal, Law department	92%
Quit	Quit her job, Quitting, To quit	48%
Start with	Start with that, Take it from there, Where we'll start, Going to start	47%
Stalking	Stalking	33%
Sweet	How sweet	25%

Table 3. Accuracy of system modules

	Occurrences	Percentage
Recognized topics	100	66%
Missed by Speech recognizer	37	25%
Missed by pattern matchers	14	9%
False positive rate	0	0%
Total occurrences	151	100%

tem that seamlessly blends animations and procedural behaviors. Finally, these animations are rendered in the Unreal Tournament™ game engine and displayed to the Speaker.

3.1 System accuracy evaluation

To test the system's performance we ran 36 recordings of speakers telling the foot massage story through the system. Average accuracy was 66% for recognizing LFM categories (see table 3). The substantial differences in accuracy between LFM categories (e.g. foot massage vs. harassment) have two causes. First, the general vocabulary is biased – some words are more difficult to recognize than others (e.g. "stalking" misrecognized as "stocking"). A high number of variations contribute to low scores in categories such as "quit" (e.g. choice of 1st person and 3rd person) were not in our set of selected LFMs and thus missed by the pattern matchers. Speed varies quite a bit between runs (Dragon's response time is fairly non-deterministic), so to test time performance we did 3 runs on the same dataset of 36 interviews: Average response time was just over 2 sec; 15% occured under 1 sec. The system is thus still far from reliably achieving natural response times of 1 sec on average.

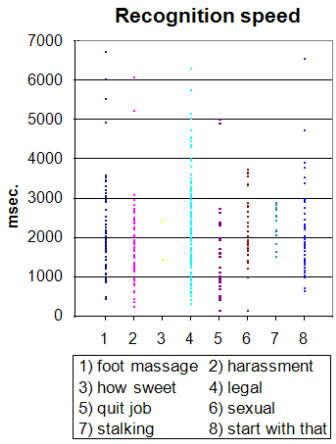

Fig. 3. Speed of recognition over 3 runs per audio file from the human-human interaction study.

4 Discussion & Future Work

We have described a framework intended to help with human subject dialogue experiments and in building autonomous agents and automatic dialog systems. We also described initial us of it in producing appropriate and timely content feedback. The results demonstrate progress in integrating behavioral, prosodic and lexical information to produce realtime listening feedback within a constrained setting. Although our findings indicate that significant advancements need to be made to reach human-level performance, they also highlight that embodied agents are inching towards the richness of natural conversational behavior by combining envelope and content feedback, and in the process opening up a host of new research questions, e.g. how to integrate such feedback with models of emotion [13].

An obvious next step is to improve the technology to match the speed and accuracy observed in the human-human condition. A more powerful pattern matching function would capture more surface variability in how people describe key narrative events. In the current system setup there were no false positives, but allowing for more general variations in the pattern matching will very likely raise their number; how much, however, is a function of the number of false positives in the speech recognition and the generality of the expressions allowed.

At present there is little data about how well people can adjust to delays or peculiarities of feedback produced by (virtual) humans. The role of non-lexical features in

the elicitation of feedback is also unclear (e.g. speaker prosody or facial expressions) and further work is needed to tease apart these factors. Embodied agents that react instantly and emotionally to human speech, albeit in simplified settings, have the potential to begin to address these questions, for the benefit of both autonomous agents and our understanding of human communication.

Acknowledgments. We are grateful for the substantive contributions of a number of individuals. Anya Okhmatovskaia and Alison Wiener contributed to our experimental design. Jillian Gerten, Ning Wang, and Robin Duffy assisted with the data elicitation and analysis. Jeremy Bailenson and Nicole Kraemer informed of relevant findings. This work was sponsored in part by the U.S. Army Research, Development, and Engineering Command (RDECOM), and the content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. This work was in part supported by a Marie Curie European Reintegration Grants within the 6th European Community Framework Programme, and by a research project grant from RANNÍS.

References

1. Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R., et al. (2006). *Virtual Rapport.* Paper presented at the 6th International Conference on Intelligent Virtual Agents, Marina del Rey, CA.
2. Thórisson, K. R. (1993). *Dialogue Control in Social Interface Agents.* Paper presented at the InterCHI Adjunct Proceedings, Conference on Human Factors in Computing Systems, Amsterdam.
3. Thórisson, K. R. (1996). *Communicative Humanoids: A Computational Model of Psycho-Social Dialogue Skills.* Unpublished Ph.D. thesis, Massachusetts Institute of Technology.
4. Tosa, N. (1993). Neurobaby. *ACM SIGGRAPH*, 212-213.
5. Thórisson, K. R. (2002). Natural Turntaking Needs No Manual: Computational Theory and Model, From Perception to Action. In B. Granström, D. House & I. Karlsson (Eds.), *Multimodality in Language and Speech Systems* (pp. 173-207). Dordrecht, The Netherlands: Kluwer Academic Publishers.
6. Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as Co-narrators. *Journal of Personality and Social Psychology, 79*(6), 941-952.
7. Tickle-Degnen, L., & Rosenthal, R. (1990). The Nature of Rapport and its Nonverbal Correlates. *Psychological Inquiry, 1*(4), 285-293.
8. Gratch, J., Wang, N., Okhmatovskaia, A., Lamothe, F., Morales, M., van der Werf, R., et al. (2007). *Can virtual humans be more engaging than real ones?* Paper presented at the 12th International Conference on Human-Computer Interaction, Beijing, China.
9. Scherer, K. R., & Ellgring, H. (2007). Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? *Emotion*.
10. Ward, N., & Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics, 23*, 1177-1207.
11. Morency, L.-P., Sidner, C., Lee, C., & Darrell, T. (2005). *Contextual Recognition of Head Gestures.* Paper presented at the 7th International Conference on Multimodal Interactions, Torento, Italy.
12. Kopp, S., Krenn, B., Marsella, S., Marshall, A., Pelachaud, C., Pirker, H., et al. (2006). *Towards a common framework for multimodal generation in ECAs: The behavior markup language.* Paper presented at the Intelligent Virtual Agents, Marina del Rey, CA.
13. Gratch, J., & Marsella, S. (2004). A domain independent framework for modeling emotion. *Journal of Cognitive Systems Research, 5*(4), 269-306.