# Evaluating Look-to-Talk: A Gaze-Aware Interface in a Collaborative Environment

**Alice Oh, Harold Fox, Max Van Kleek, Aaron Adler, Krzysztof Gajos, Louis-Philippe Morency, and Trevor Darrell**

MIT Artificial Intelligence Laboratory
Cambridge, MA, USA
{aoh, hfox, emax, cadlerun, kgajos, lmorency, trevor}@ai.mit.edu

## ABSTRACT

We present "look-to-talk", a gaze-aware interface for directing a spoken utterance to a software agent in a multi-user collaborative environment. Through a prototype and a Wizard-of-Oz (WOz) experiment, we show that "look-to-talk" is indeed a natural alternative to speech and other paradigms.

### Keywords

Multimodal user interface, gaze, intelligent environment.

## INTRODUCTION

An intelligent environment (IE) is a setting where multiple users interact with one another and with a multitude of software agents. In such a setting, knowing who is speaking to whom is an important and difficult question that cannot always be answered with speech alone. Gaze tracking has been identified as an effective cue to help disambiguate the addressee of a spoken utterance [4].

In a study of eye gaze patterns in multi-party (more than two people) conversations, Vertegaal, et al. [6] showed that people are much more likely to look at the people they are talking to, than any other people in the room. Also, in another study, Maglio, et al. [3] found that users in a room with multiple devices almost always look at the devices before talking to them. In conversational agents, the importance of nonverbal gestures has already been recognized [1]. These observations led us to hypothesize that using gaze as an interface to activate the automatic speech recognition (ASR) would enable natural human-computer interaction (HCI) in a collaborative environment.

To test our hypothesis, we have implemented "look-to-talk" (LTT), a gaze-driven paradigm, and "talk-to-talk" (TTT), a spoken keyword-driven paradigm. We have also implemented "push-to-talk" (PTT), where the user pushes a button to activate ASR. We present and discuss a user evaluation of our prototype system as well as a Wizard of Oz (WOz) setup.

## EXPERIMENTS

To compare the usability of LTT with the other modes, we ran two experiments in the MIT AI Lab's Intelligent Room [2] (from here on "the I-Room"). We ran the first experiment with a real vision- and speech-based system, and the second experiment with a WOz setup where gaze tracking and ASR were simulated by an experimenter behind the scenes. Each subject was asked to use all three modes to activate ASR and then to evaluate each mode.

### Set-up

We set up the experiment to simulate a collaboration activity among two subjects and a software agent. The first subject (subject A) sits facing the front wall displays, and a second "helper" subject (subject B) sits across from subject A. The task is displayed on the wall facing subject A. The camera is on the table in front of subject A, and Sam, an animated character representing the software agent, is displayed on the side wall (see Fig. 1). Subject A wears a wireless microphone and communicates with Sam via IBM ViaVoice. Subject B discusses the task with subject A and acts as a collaborator. The I-Room does not detect subject B's words and head-pose.

### Estimating Gaze

For a natural LTT interface, we need a fast and reliable computer vision system to accurately track the user's gaze. This has barred gaze-based interfaces from being widely implemented in IEs. However, fast and reliable gaze trackers using state-of-the-art vision technologies are now becoming available and are being used to estimate the focus of attention. For example, Stiefelhagen *et al.* [5] showed that the focus of attention can be predicted from the head position 74% of the time during a meeting scenario.

| Mode | Activate | Feedback | Deactivate | Feedback |
|---|---|---|---|---|
| PTT | Switch the microphone to "on" | Physical status of the switch | Switch the microphone to "mute" | Physical status of the switch |
| LTT | Turn head toward Sam | Sam shows listening expression | Turn head away from Sam | Sam shows normal expression |
| TTT | Say "computer" | Special beep | Automatic (after 5 sec) | None |

**Table 1: How to activate and deactivate the speech interface**

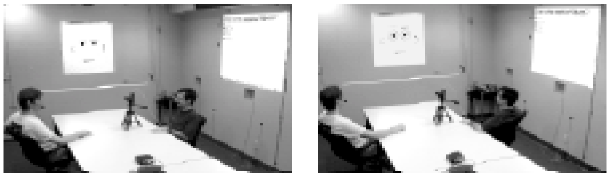

**Figure 1: LTT in non-listening (left) and listening (right) mode.**

In our prototype system, we estimate gaze with a 3-D gradient-based head-pose tracker [4] that uses shape and intensity of the moving object. The tracker provides a good non-intrusive approximation of the user's gaze.

### *Sam: An Animated Character*

Sam is the I-Room's emotive user interface agent. Sam consists of simple shapes forming a face, which animate to continually reflect the I-Room's state (see Fig. 1). During this experiment, Sam read quiz questions through a text-to-speech synthesizer, and was constrained to two facial expressions: non-listening and listening.

### *Task*

Each pair of subjects was posed three sets of six trivia questions, each set using a different mode of interaction in counterbalanced order. In the WOz setup, we ran a fourth set in which all three modes were available, and the subjects were told to use any one of them for each question. Table 1 illustrates how users activate and deactivate ASR using the three modes, and what feedback the system provides for each mode.

### *Subjects*

There were 13 subjects, 6 for the first experiment and 7 for the WOz setup. They were students in computer science, some of whom had prior experience with TTT in the I-Room.

### *Usability Survey*

After the experiment, the subjects rated each of the three modes on a scale of one to five on three dimensions: ease-of-use, naturalness, and future use. We also asked the subjects to tell us which mode they liked best and why.

## RESULTS AND DISCUSSIONS

For the first experiment, there was no significant difference (using anova at $\alpha$=0.05) between the three modes for any of the surveyed dimensions. However, most users preferred TTT to the other two. They reported that TTT seemed more accurate than LTT and more convenient than PTT.

For the WOz experiment, there was a significant difference in the naturalness rating between PTT and the other two (p=0.01). This shows that, with better perception technologies, both LTT and TTT will be better choices for natural HCI. Between LTT and TTT, there was no significant difference on any of the dimensions. Five out of the seven subjects reported, however, that they liked TTT best, compared to two subjects who preferred LTT. One reason for preferring TTT to LTT was that there seemed to be a shorter latency in TTT than LTT. Also, a few subjects remarked that Sam seemed disconnected from the task, and thus it felt awkward to look at Sam.

Despite the subjects' survey answers, for the fourth set, 19 out of 30 questions were answered using LTT, compared with 9 using TTT (we have this data for five out of the seven subjects; the other two chose a mode before beginning fourth set to use for the entire set, and they each picked LTT and TTT). When asked why he chose to use LTT even though he liked TTT better, one subject answered "I just turned my head to answer and noticed that the Room was already in listening mode." This confirms the findings in [2] that users naturally look at agents before talking to them.

## CONCLUSION AND FUTURE WORK

Under ideal conditions (i.e., WOz), users preferred perceptual interfaces to push-to-talk. In addition, they used look-to-talk more often for interacting with agents in the environment. This has led us to believe that look-to-talk is a promising interface. However, it is clear that having all three modalities available for users provides convenience and efficiency for different contexts and user preferences. We are currently working to incorporate look-to-talk with the other modalities.

We are also investigating ways to improve gaze tracking accuracy and speed. As the prototype tracker performance approaches that of the WOz system, we expect the look-to-talk user experience to improve significantly.

## ACKNOWLEDGMENTS

Thanks to all of the anonymous subjects in our user study.

This work was funded by Acer Inc., Delta Electronics Inc., HP Corp., NTT Inc., Nokia Research Center, and Philips Research under the MIT Project Oxygen partnership, and by DARPA through the Office of Naval Research under contract number N66001-99-2-891702.

## REFERENCES

1. Cassell, J. Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. *Embodied Conversational Agents.* MIT Press.

2. Coen, M. Design principles for intelligent environments. *Fifteenth National Conference on Artificial Intelligence*. Madison, Wisconsin, 1998.

3. Maglio, P., Matlock, T., Campbell, C., Zhai, S., Smith, B.A. Gaze and speech in attentive user interface. *Proc. of the Third Int'l Conf. on Multimodal Interfaces*. Beijing, China, Oct. 2000.

4. Rahimi, A., Morency, L., Darrell, T. Reducing drift in parametric motion tracking, *The Eighth IEEE Int'l Conf. on Computer Vision*. Vancouver, Canada, 2001.

5. Stiefelhagen, R., Yang, J., Waibel, A. Estimating focus of attention based on gaze and sound. *Workshop on Perceptive User Interfaces (PUI '01).* Orlando, Florida, Nov. 2001.

6. Vertegaal, R., Slagter, R., Van der Veer, G.C., Nijholt, A. Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. *Proc of ACM Conf. on Human Factors in Computing Systems*, 2000.