To be presented at the Spoken Language Technology Workshop, December 2014.

EYE GAZE FOR SPEECH RECOGNITION AND UNDERSTANDING

Malcolm Slaney and Dilek Hakkani-Tür

Microsoft Research, Mountain View, CA, USA

ABSTRACT

We show how eye-gaze information can improve several different aspects of the speech-processing pipeline. We see significant improvements in speech recognition and speech understanding, and demonstrate these advantages in a simple screen-based application.

***Index Terms*—**Eye gaze, attention, speech recognition, speech understanding

1. INTRODUCTION

We want to show how your eyes can improve speech recognition and speech understanding. Many of the scenarios in which we are interested combine a screen with voice input (See Figure 1). Screens of all sizes are great ways to present a lot of information to a user. And speech is a natural and high-bandwidth input signal. Yet speech recognition remains a challenging problem, especially in the natural (noisy) environments where we often want to communicate with our devices.

Thus, attention is the key. What we are attending to is probably a good clue about what we might say next. Were you just looking at the Italian restaurant listing, or the Indian? We use the eyes (and by proxy face pose) as an important cue for better recognizing and understanding speech. Sensors for eye-gaze and face-pose information are inexpensive, as many of our devices already include one or more cameras.

We have a simple demo, using a Tobii eye-tracker as input, that demonstrates the potential of eye tracking to improve speech recognition and speech understanding. While the demo is contrived to make the speech problem as difficult as possible, it provides a good vehicle to see the effect, and to discuss the ramifications. Figure 2 shows a screen shot.

2. BENEFITS

We have three studies that show the benefit of eye gaze (and face pose) to different parts of the speech-processing pipeline. These studies are summarized in Figure 3.

Addressee detection, deciding whether the user is addressing the computer, or somebody else, seems like it would be an easy task for face pose. Our initial study in this area showed the difficulty of acquiring and using the signal [1]. Users participated in a trivia contest, where the computer asked a question and two or more people stood in front of the screen, collaborated on an answer, asked questions of the system, and then responded with an answer. We found it difficult using a single camera to acquire the face-pose information over the entire range of angles. And more importantly, the pose information is time varying and rather subtle. It's not as easy as push to talk.

The first positive results occurred with automatic speech recognition (ASR). In our work [2], and that by others [3], we showed that eye-gaze information can bias the language model and give a 10% improvement in word-error rate. In our work we first recognized the utterance with a generic language model, then we used the eye-gaze data to reweight the language model, and then rescored the first-pass recognition lattice. We saw a similar potential for improvement using face-pose data to approximate the eye-gaze data [4].

Finally, we investigated two different approaches to use the ASR output, along with eye-gaze data, to improve spoken-language understanding (SLU) in a web-browsing application. The first approach uses heuristics such as distance between the eye-gaze data and the web-page link [5], while the second approach builds a heatmap to model the probability that the user saw each particular word on the page

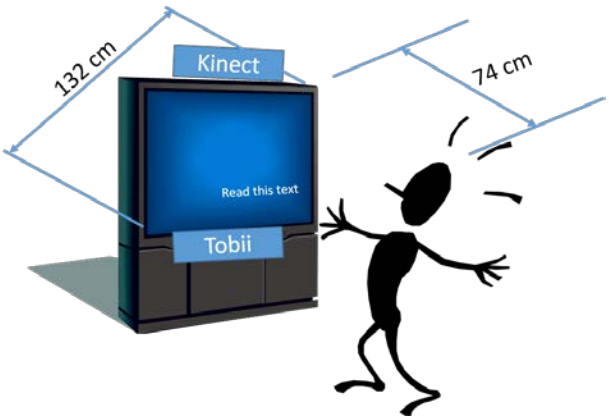

Figure 1: One scenario when eye-gaze and face-pose information can be used to enhance speech recognition.

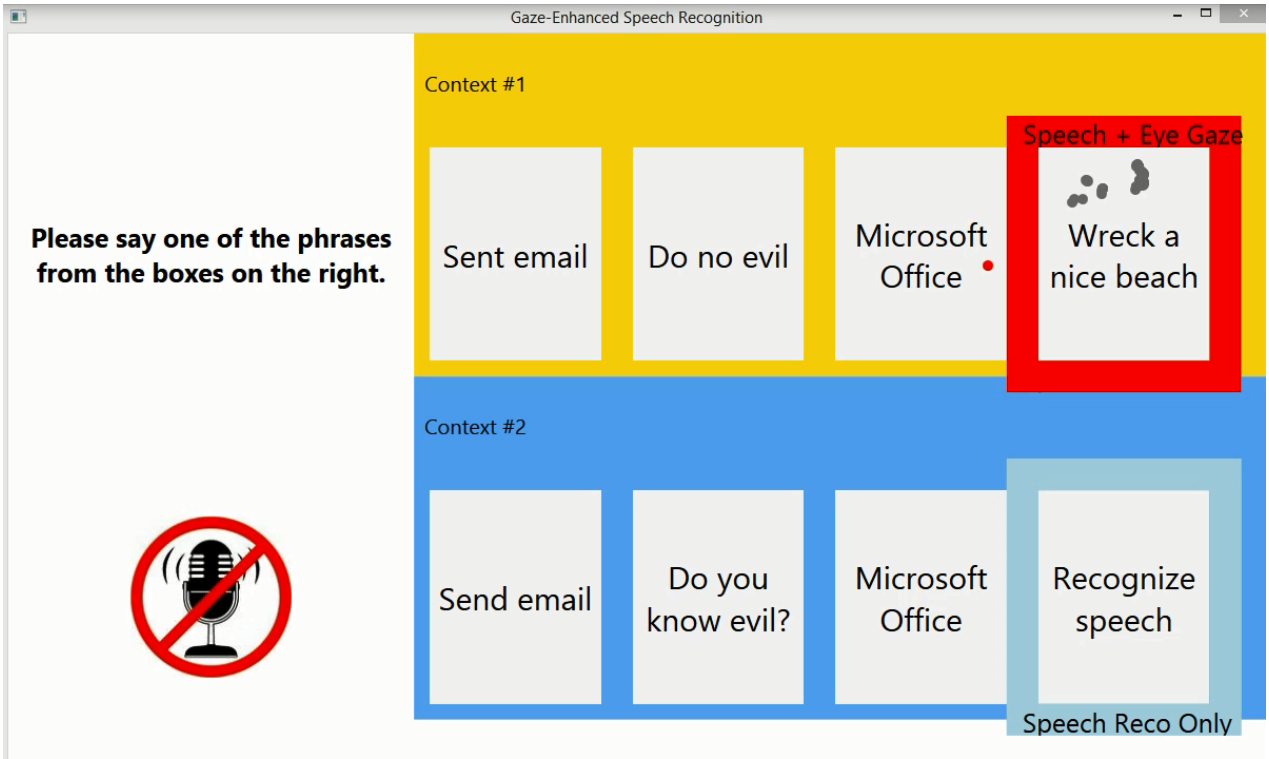

Figure 2: A screen shot showing the user's options for this demonstration showing eye gaze driving speech recognition. The user clearly said "Recognize speech." This was the result of the first-pass recognition effort. The second pass adds the eye-gaze information (shown as gray dots) and correctly recognizes the desired utterance "Wreck a nice beach."

[6]. The two approaches are complementary, and in total give a 17% improvement in the f-score when determining the correct action for the user. (Both of these approaches use lexical measures such as longest-common-substring to match the words in the utterance to the link's anchor text.)

3. CONCLUSIONS

We have shown how eye-gaze data can improve both ASR and SLU, and we are optimistic that eye-gaze data will help with addressee detection. With large screens, where eye-gaze data is harder to obtain because the user is standing further from the camera, we believe that face-pose data is a good, albeit noisy, approximation to eye-gaze data.

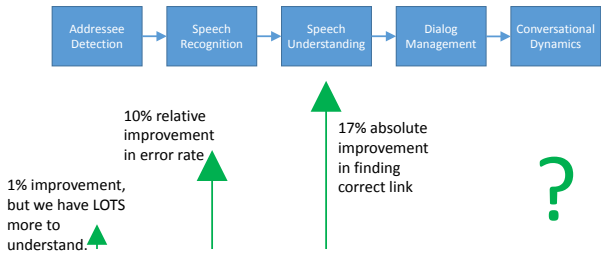

Figure 3: A summary of eye-gaze results for different parts of the speech-processing pipeline.

4. REFERENCES

[1] TJ Tsai, Andreas Stolcke, Malcolm Slaney. Multimodal addressee detection in multiparty dialogue systems. Submitted to *IEEE ICASSP*, 2015.

[2] Malcolm Slaney, Rahul Rajan, Andreas Stolcke, and Partha Parthasarathy. Gaze-enhanced speech recognition. In *Proc. IEEE ICASSP*, IEEE SPS, Florence, Italy. May 2014.

[3]Neil Cooke and Martin Russell. Gaze-Contingent ASR for Spontaneous, Conversational Speech: An Evaluation. *Proc. IEEE ICASSP*, 2008.

[4] Malcolm Slaney, Andreas Stolcke, Dilek Hakkani-Tür. The relation of eye gaze and face pose: Potential impact on speech recognition. *ACM International Conference on Multimodal Interactions (ICMI)*, Istanbul, Turkey, November 2014.

[5] Dilek Hakkani-Tür, Malcolm Slaney, Asli Celikyilmaz, Larry Heck. Eye gaze for spoken language understanding in multi-modal conversational interactions. *ACM International Conference on Multimodal Interactions (ICMI)*, Istanbul, Turkey, November 2014.

[6] Anna Prokofieva, Malcolm Slaney, Dilek Hakkani-Tür Probabilistic features for connecting eye gaze to spoken language understanding. Submitted to *IEEE ICASSP*, 2015.