INTEGRATION OF EYE FIXATION INFORMATION WITH SPEECH RECOGNITION SYSTEMS

Ramesh R. Sarukkai & Craig Hunter

Dept. of Computer Science
University of Rochester
Rochester, NY-14627
e-mail: ramesh@kurzweil.com ; craigh@isn.com

ABSTRACT

In this paper, a semi-tight coupling between visual and auditory modalities is proposed: in particular, eye fixation information is used to enhance the output of speech recognition systems. This is achieved by treating natural human eye fixations as *diectic references* to symbolic objects, and passing on this information to the speech recognizer. The speech recognizer biases its search towards these set of symbols/words during the best word sequence search process. As an illustrative example, the TRAINS interactive planning assistant system has been used as a test-bed; eye-fixations provide important cues to city names which the user sees on the map. Experimental results indicate that eye fixations help reduce speech recognition errors. This work suggests that integrating information from different interfaces to bootstrap each other would enable the development of reliable and robust interactive multi-modal human-computer systems[1].

1. Introduction

The notion of multi-modal interfaces has been around for a long time. The integration of the visual and auditory modality has been primarily been examined in the context of lipreading[2]. Other recent applications of multi-modal interfaces include HearingAid[9], EagleEyes[8], and so on. In [9], the programming by demonstration paradigm is enhanced by allowing simultaneous speech inputs to help resolve ambiguities.

[1]Dr. Ramesh Sarukkai is now with Kurzweil A.I.

Researchers have noted that gaze information is an integral part of face-to-face communication. Typically, if objects of discussion are in someone's vicinity, humans talking about the objects tend to gaze in that direction. In [10], an eye-based interface has been described where the deictic behaviour of the eye has been explored. [5] exploits the eyes as a consious interface, which can be intertwined with mouse and keyboard inputs. The role of the eyes however is just that of a pointer similar to a mouse. EagleEyes[8] is another system that allows the computer to sense the user's eye and head movements in order to adjust the display or perform corresponding operations (such as playing a digitized video or sound). Eye fixation information also provide useful cues in a face-to-face communication, and this has been demonstrated in [11] where an on-screen face is animated according to the user's hand gestures, gaze and intonation. More recently, in [9], the programming by demonstration paradigm is enhanced by allowing simultaneous speech inputs to help resolve ambiguities.

2. Motivation

Most of the actual integration of multi-modal signals has been explored mainly in the context of lip-reading (for example [2]). However, [14] have explored the integration of simultaneous inputs from speech, gaze, and hand gestures. The approach taken in [14] work is as follows: 3 streams of time-stamped data are generated using the different modalities, namely words from the speech recognizer, position and postures of the hands, and

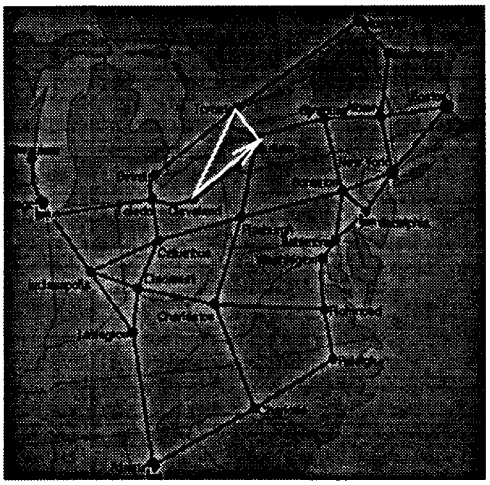

Figure 1: An example eye-fixation sequence durring the speech utterance: "TRAIN FROM CLEVELAND TO BUFFALO"

point of gaze. These input streams are parsed *seperately* to produce frame-like descriptions. Finally, these frames are interconnected and evaluated together to resolve the references in the users utterance. Thus, multi-modal information is used to interpret collective information.

The goals in this work is quite different from the above work. In the real world, each of the individual modalities have errors associated with them: sensing errors, algorithm approximation errors, and so on. It is clear that perception of different modalities occur in a complex combined manner. For example, [4] have demonstrated how eye-movements can be used to study spoken language comprehension. But from the point of view of a human-computer multi-modal interface, the task of integrating different modalities is an arduous task. The dimensionality of the problem increases (along with the sparseness of observations) with added modalities. One mechanism for simplifying the integration problem is presented and explored in the context of using eye-fixation cues to enhance the speech recognition system.

In order to further motivate the use of eye-fixation information for speech recognition, consider an illustrative eye-movement sequence when an user uttered "(take the) train from Cleveland to Buffalo". This is a snapshot from the TRAINS map [12]: users interact with the system to plan travel plans from different cities. It can be noted from figure 1 that the user makes saccades from the city Buffalo to Toronto to Cleveland to Buffalo. Such eye-fixations are typical. A variety of eye-movement sequences have been observed:

- Cyclic eye-movements: The user looks at the path of cities back and forth.
- Sequential eye-movements: The users eye-fixations match the sequence of city visits as specified by the speech utterance.
- Mixed eye-movements: The user looks at some of the cities in the speech being uttered, but also makes fixations to other cities that may or may not have been in discussion in the previous utterances.
- Non-utility eye-movements: The user does not need to look at any of the cities while articulating the speech. This typically occurs when the user is certain of the start and goal cities, or when he/she is giving other commands.

It is important to note that the above eye-movements are non-intrusive in the sense that the user *is not forced* to look at these cities/objects, but does so naturally. The key idea that is being exploited is that humans use eye fixations for deictic references: in the following sections we demostrate a scheme whereby the eye-fixation information is utilized to enhance/correct certain speech recognition errors. Furthermore, the ideas and notions presented are general and can be applied to various other domains which utilize visual and auditory cues.

3. Using Eye Fixations to Enhance Speech Reognition

In this section, we discuss the method of integrating eye-fixation information into existing speech recognition systems. There are various approaches to the problem of multi-modal integration:

- Loose Coupling: By loose coupling methods such as ones proposed in [14] where information from different modalities are gathered and recognized seperately; multi-modal integration of these recognized tokens enables reference resolutions.
- Semi-tight coupling: In this approach, information gathered from various modalities are then

used for bootstrapping each other and the rescored recognition is performed. This is the approach taken in this paper.

• tight coupling: The various modalities are time-stamped and synchronized, and the reconiser runs on the combined information from all modalities together. Examples of such a coupling include various neural network architectures proposed for lipreading.

After examining the eye-fixations, it is clear that while temporal sequence of eye-fixations correlates with the sequence of occurence of those referred objects in the speech data in some cases, accurate temporal correlation is absent a lot of the time. This suggests that rather than incorporating temporal sequential information from the visual modality, it would be more useful to incorporate "symbolic" information: by this we refer to symbols such as city names or objects that are being discussed which appear in the field of view of the speakers.

This notion of non-sequential information has been researched upon in speech recognition systems in the form of triggers[13], and more recently word sets[1]. The idea behind the *word set probablity boosting* algorithm is to enhance the probability of occurence of "predicted words" in a speech utterance. In the present case, the predicted words are provided by the visual eye-fixation information. Figure 2 summarizes the ideas succinctly.

4. Summary of Experimental Results

Experiments have been performed using data recorded from 8 speakers constitued a total of 95 utterances from a single dialogue per speaker with the system planning train travel routes. The Sphinx-II system from CMU was used as the underlying speech recognition system. Experiments were performed on the TRAINS domain; users wore an Applied Science Laboratories series 4000 head-mounted eye-tracking system, headphones and a microphone. The cities that were fixated upon during a particular utterance were manually extracted and stored at 30 frames per second. These words(city names) were then boosted using the word set probability boosting algorithm[1].

The overall correct city recognition in the decoded word sequences improved from 74.2% to 82.3%. Surprisingly there were no incorrect cities additionally introduced despite boosting an average of 4.6 cities per utterance (the average number of cities actually spoken in an utterance was 1.6). We further experimented by boosting all 26 cities on the TRAINS map, and found similar improvements in city detection; what this suggests is that eye-fixation is just a means of extracting information relevant to the current domain of dialogue. Using deictic "eye-triggered" word sets in dynamically varying scenes is potentially useful information for improving speech recognition performance. The visual modality can sometimes provide useful cues as to what the speaker is talking about (such as a specific cities/objects), and this work suggests that it can be useful to bias speech recognition systems towards such dynamic domain-specific information. The improvement in

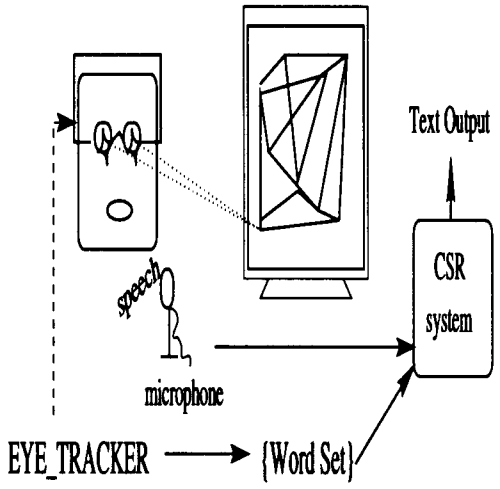

Figure 2: Overview of the integration of eye-fixation information to enhance speech recognition systems

Method	% Cities Correct
CSR-alone	74.2 %
CSR+Eye	82.3 %

Table 1: % Correct city identification with and without eye-fixation information

correct city identification is summarized in table 1.

5. Conclusion

In this paper, the concept of utilizing eye-fixations to enhance speech recognition has been explored and implemented. This is achieved by treating natural human eye fixations as *diectic references* to symbolic objects, and passing on this information to the speech recognizer. The speech recognizer biases its search towards these set of symbols/words during the best word sequence search process. As an illustrative example, the TRAINS interactive planning assistant system has been used as a testbed; eye-fixations provide important cues to city names which the user sees on the map. Experimental results indicate that eye fixations help reduce speech recognition errors. This work suggests that integrating information from different interfaces to bootstrap each other would enable the development of reliable and robust interactive human-computer systems

6. References

[1] Ramesh R. Sarukkai, and Dana H. Ballard, "Word Set Probability Boosting for Improved Spontaneous Dialogue Recognition", to appear in IEEE Trans. on Speech and Audio Proc.

[2] C. Bregler, H. Hild, S. Mahnke, and A. Waibel "Bimodal Sensor Integration on the Example of 'Speechreading'", Proc. of IEEE Int. Conf on Neural Networks, San Francisco, 1993

[3] Cooper, Roger M. "The Control of Eye Fixation by the Meaning of Spoken Language", Cognitive Psychology, 1974, pp. 84-107.

[4] K. Eberhard, M. Spivey-Knowlton, J. Sedivey, and M. Tanenhaus. "Eye Movements as a Window into RealTime Spoken Language", Journal of Psycholinguistic Research(in press), 1995.

[5] Jacob, R.J.K. "The Use of Eye Movements in Human-Computer Interaction Techniques", ACM Transactions on Information Systems, 1991, 9(3) pp. 152-169.

[6] Locke, John L. "A Child's Path to Spoken Language", Cambridge, Mass. : Harvard University Press, 1993.

[7] A. Kobsa, J. Allgayer, C. Reddig, N. Reithinger, D.Schmauks, K. Harbusch, and W. Wahlster "Combining Deictic Gestures and Natural Language for Referent Identification", Proceedings of the 11th Conference on Computational Linguistics, Bonn, Germany, 1986,pp. 356-361.

[8] P. Olivieri, J. Gips, J. McHugh, "EagleEyes: eye controlled multimedia", in Proc. of ACM Multi-Media'95, San Francisco, pp:537-538 (Video summary)

[9] E. Stoehr, and H. Lieberman, "Hearing Aid: Adding verbal hints to a learning interface", Proc. of ACM Multimedia'95, San Francisco, pp:223-230

[10] I. Starker and R. A. Bolt, "A Gaze-Responsive Self-Disclosing Display", Proc. ACM CHI'90 Human Factors in Computing Systems Conference, pp 3-9, Addison-Wesley/ACM Press(1990).

[11] Thorisson, K. R. "Face-to-Face Communication with Computer Agents", Working Notes, AAAI Spring Symposium on Believable Agents, Stanford University, California, Aug. 13-16, 1994, pp 86-90

[12] James F. Allen, George Ferguson, Brad Miller, and Eric Ringger, "Spoken Dialogue and Interactive Planning", in Proc. of ARPA Spoken Language Technology Workshop, Austin, TX, 1995.

[13] Ronald Rosenfeld, "Adaptive Statistical Language Modeling: A Maximum Entropy Approach", CMU-CS-94-138.

[14] Koons, D. B., Sparrel, C. J. and Thorisson, K. R. "Integrating Simultaneous Input from Speech, Gaze and Hand Gestures", In M. T. Maybury(Ed.), *Intelligent Multi-Media Interfaces.* AAAI/MIT Press.

Acknowledgments: We would like to thank Prof. James Allen, and Prof. Dana Ballard for useful discussions, and the members of the TRAINS group for their assistance.