Eye Movements in Speech Technologies: an overview of current research

Mattias Nilsson

Department of linguistics and Philology, Uppsala University
Box 635, SE-751 26 Uppsala, Sweden
Graduate School of Language Technology (GSLT)

Abstract

We present a summary overview of recent work using eye movement data to improve speech technologies. We summarize the experimental psycholinguistic evidence motivating these applications and provide an overview of a number of gaze-speech studies in the areas of multimodal human-computer interaction, synthesized speech evaluation and automatic speech recognition.

1 Introduction

When listeners follow spoken instructions to manipulate real objects or objects in a visual display, their eye-movements to the objects are closely time-locked to the spoken words referring to those objects (Eberhard et al., 1995). In other words, listeners naturally make saccadic [1] eye movements to objects as they recognize the spoken words referring to them. For the last fifteen years this central observation in psycholinguistic research has provided a wealth of insights into the time course of spoken language processing. More recently, a growing number of researchers in speech technology and human-computer interaction has drawn on the experimental evidence and are now using eye tracking to address diverse issues such as dialog system design, synthesized speech evaluation and automatic speech recognition. Currently, however, there is no designated forum for research on the ways in which eye movements may inform speech technologies, and papers addressing these questions are spread out and often published in quite different journals. Hence it is decidedly hard to get a general overview of the problems addressed, the methods used and the

Email address: mattias.nilsson@lingfil.uu.se (Mattias Nilsson).

[1] Saccades are very rapid eye movements that transport the eyes from one fixation point to another.

results obtained in this line of research. In this paper, therefore, we attempt to provide a brief summary of recent studies using eye tracking to advance speech technologies. This survey will be rather selective and we have no hope in covering all relevant studies. If some work is not mentioned, it does not mean it's not important. Moreover, let us be clear from the outset that we do not intend to present any research of our own. In section 2 we provide the relevant psycholinguistic backgound on what is known about the coordination of eye movements and spoken language processing. The subsequent three sections present applications of eye tracking in speech technology. We report on studies in the areas of *multimodal human computer interaction*, *synthesized speech evaluation*, and *automatic speech recognition* respectively. Section 6 concludes the summary.

2 Eye movements in spoken language processing

2.1 Eye movements in spoken language comprehension

By recording the eye movements of a person following spoken requests to move visually presented objects it is possible to monitor the on-line comprehension process on a millisecond time scale. In psycholinguistic research this experimental methodology is generally known as *the visual world paradigm*. In a typical visual world experiment subjects follow simple instructions such as *look at*, *pick up* or *move* a small number of objects displayed on a computer screen while their eye movements are being monitored by an eye tracker system which records the locations and durations of individual fixations. The eye movements are monitored using a light-weight head-mounted eye tracker which does not require the subject to retain his or her head in a fixed position. Therefore, head mounted eye tracking is generally considered relatively comfortable for the subject. A large number of studies using this experimental set-up have shown that subjects eye movement response to a particular object is closely time-locked to the input speech stream. Tanenhaus et al. (1995) provide an early influential report of a number of studies carried out at their lab using the visual world paradigm. In one experiment which investigated the time course of definite refererence resolution subjects were instructed to touch one of four objects that differed in marking (plain or starred), colour (pink, yellow, blue and red) and shape (square or rectangle). The processing latency was measured from the beginning of the spoken noun phrase until the onset of the eye movement which fixated the target object. The results showed that subjects initiated an eye movement to the target object on average 250 ms after the onset of the spoken word that uniquely determined the target object. For example, when listening to an instruction such as "touch the starred yellow square" under the condition when there was only one starred object in the display, subjects made an eye movement to the target object 250 ms after the end of the word "starred". Under the condition when there was two

starred yellow objects in the display, subjects fixated the target object 250 ms after the disambiguating word "square". Since it is known that it takes about 200 milliseconds to plan a saccade before the eyes actually begin to move, this implies that subjects actually identified the referent approximately somewhere near the middle of the word which uniquely identified the target. Another experiment investigated the time course of ambiguity resolution in word recognition. In this experiment subjects were presented with a visual display depicting everyday objects that sometimes included two objects with similar onsets, such as "candy" and "candle". The subjects were then given instructions to move the objects around (e.g., "Pick up the candy. Now put it above the fork"). In the case when all the names of the visual objects had different onsets, the average time to initiate an eye movement to the target object was 145 ms from the end of the spoken word. When there was an object present with a similar onset as another object in the display, the average time to launch an eye movement to the target object was 230 ms. Again, because it takes about 200 milliseconds to plan the execution of a saccade, the results demonstrate that the referent is actually identified near the middle of the spoken word in the case when all objects had different onsets.

2.2 Eye movements in spoken language production

Influenced by the many crucial insights about the coordination of eye movements and comprehension, psycholinguistic research has more recently began to address questions concerning the relation between eye movements and spoken language production. Although there are not nearly as many studies on eye movements in production compared to comprehension, initial results suggest that eye movements and language production are closely coupled. That is, when describing visual scenes, speakers typically gaze at objects while preparing to speak their names. In typical experiments, speakers view scenes on a computer screen and are asked to describe them. According to Griffin (2004), the latency between fixating an object and beginning to say its name is relatively consistent across subjects in spontaneouos scene description. Furthermore, even if speakers have previously fixated an object, they tend to return their gaze to it roughly a second before mentioning it (Griffin & Bock, 2000). One way to measure the gaze and speech latency is to compute the *eye-voice span* in speaking. The eye-voice span is the time between the onset of the gaze to an object in the scene and the subsequent onset of the spoken word referring to that object. In the first study of eye movements in spontaneous scene descriptions, eye-voice spans for fluently spoken nouns in subject position were 902 ms on average and 932 ms on average for nouns in object position (Griffin & Bock, 2000).

3 Speech and gaze in human computer interaction

Motivated by psycholinguistic findings such as those reviewed in the previous section, Campana et al. (2001) describe a dialogue system which uses eye movement data in order to determine underspecified referents. Campana et al. argue that underspecification is a natural and pervasive characteristic of human communication and that most dialogue systems are unable to provide full support for underspecified definite descriptions. Given the time-locked characteristic of eye movements and speaking, however, they suggest that eye tracking data can be used both to infer which referent the user is referring to, and furthermore to gain information about whether the user has understood the utterance produced by the system. Hence, they argue that by monitoring the eye movements of the user, it should be possible to provide a more natural and effective interaction. The eye tracking information is integrated into a simulated version of a personal satellite assistant (PSA), which is a robot developed at NASA (National Aeronautics and Space Administration). The eye-tracking based reference resolution scheme is deployed in the case where there are multiple possible referents to a noun phrase spoken by the user, and the noun phrase is underspecified to such an extent that it can not be safely determined by the default anaphor resolution algorithm. Drawing on the experimental evidence that people tend to visually fixate the object they are about to mention 900 ms before the onset of the utterance, gaze information is used to identify the target referent. In their system, an underspecified referent is resolved by selecting the object fixated by the user the second before the noun phrase is pronounced. For example, if the user looks at the crew hatch (in a space shuttle) just before pronouncing "door" in the command "open that door", then the deictic expression "that" will be identified as referring to the crew hatch. The assumption expressed by Campana et al. is that this behaviour will reduce the number of turn-takings required to complete tasks in the PSA environment. Unfortunately though, they do not present any evaluation of their system. Of course, then it is very hard to tell if the eye-tracking based resolution scheme works and to what extent turn-takings are reduced, if at all.

Kaur et al. (2003) explore the relation between gaze and speech in a precise and well-defined task in an multimodal system. While their general goal is to investigate the possibility of integrating gaze and speech into a natural input device replacing the mouse, the study focuses on the simplified task of using these modalities to move an object from a set of objects to a new location on the screen by speaking the phrase "Move it there". They argue that this constrained problem setting will allow them to determine precisely to what extent it is possible to predict which object the person wants to move ("it"). They further argue that gaze input systems are appealing for a number of reasons. Most importantly, gaze manipulation of screen objects is expected to be significantly faster than hand-eye coordination. Moreover, gaze allows for

hands-free interaction. They also claim that gaze can be used as a "natural" mode of input that does not require learning of coordinated motor control movements. Kaur et al. further define three questions about gaze-speech multimodal systems that they consider particularly pertinent:

(1) What is the time relationship between a deictic reference and accompanying gaze patterns?
(2) How robust is this relationship, i.e., can it be used in software algorithms to accurately predict the intended screen location?
(3) Does the relationship hold across users or is it unique to each user, i.e., is a user required to train a speech gaze system to his or her eye-speech patterns?

In order to provide an initial answer to these questions they set up a study in which subjects move objects on a computer screen while their speech and eye movements are being recorded. The results demonstrate that the gaze fixation closest to the intended object begins, with high probability, before the beginning of the word "Move". Hence selecting the object fixated at the onset of the word "Move" is shown to give an accuracy of 95%. This can be contrasted with choosing the object fixated at the onset of the word "it" which only gives 60% accuracy. A relatively small and stable user variability is observed within subjects, while the user variablility across subjects is considerably larger. Kaur et al. conclude that the experimental results show that speech and gaze coordination patterns can be modeled reliably for individual users.

Similar work investigating gaze as an additional modality to speech can be found in Starker & Bolt (1990), and Qvarfordt & Zhai (2005).

4 Eye movements in synthesized speech evaluation

Swift et al. (2002) present a new approach to synthesized speech evaluation based on the monitoring of subjects eye movements as they respond to synthesized speech instructions in a visual workspace. In effect, this is the visual world paradigm but with synthesized speech instructions instead of human speech input. The authors recognize the need for more objective and fine grained evaluation methods than the ones most often used. It is further argued that if people process synthesized speech in much the same way they process human speech, then eye-tracking can provide a detailed on-line processing metric of synthesized speech processing. Furthermore, the feasability of this approach will be substantiated if the eye-movement data is detailed enough to reveal subtle differences between (1) the processing of synthesized speech and human speech, and (2) the processing of different speech synthesizers. Two experiments are carried out investigating the time course of lexical access and referential domain circumscription in synthesized speech process-

ing. In both experiments, the spoken instructions were given by two different text-to-speech synthesizers, and also a human voice for comparison. The experimental data demonstrates that synthesized speech processing is immediate and incremental just like human speech processing. However, it is also shown that there are important differences between synthesized and human speech processing. For example, disambiguation of an ambiguous word occurs somewhat later for both synthesized voices compared to the human voice. This implies that listeners require more time to process and interpret synthesized speech than natural speech. Furthermore, it is shown that the eye movement patterns also differs with respect to the two different synthesized voices. Swift concludes that monitoring the eye movements of listeners in a visual world setting can provide an objective and detailed measure of the quality and naturalness of synthesized speech.

5 Speech and gaze in automatic speech recognition

Another investigation of using speech and gaze in a conversational dialogue system is presented by Zhang et al. (2003, 2004). In contrast to other gaze-based dialogue systems such as that described by Campana et al. (2001), however, this study does not directly concern gaze-based reference resolution. Instead, they use eye movements in order to automatically resolve speech recognition errors. The authors note that most gaze-based multimodal systems make the simplifying assumption that the user's speech input is error-free and hence these systems do not generally deal with with speech recognition errors. This applies to the dialog system described by Campana et al. (2001) but also to earlier presented systems such as Neal et al. (1991) and Bolt (1980). Zhang et al. (2003, 2004) further note that while both speech and gaze modalities are error-prone, they can be combined in such a way as to minimize the recognition errors. This combination of the individual modalities will then provide more robust multimodal systems. So, the general assumption is that one mode of communication (e.g., gaze) can help to improve the performance of the other (e.g., speech). In their implementation they use n-best lists from both gaze and speech in order to correct potential speech recognition errors. The candidates in the gaze n-best list are ranked according to the distance from the gaze fixation to the objects. The object closest to the fixation ranks first. The candidates in the speech n-best list are ranked according to the speech recognition score and the one with the highest score ranks first. The integration of these information sources then works as follows. First, candidates that are not in the intersection of the speech n-best list and the gaze n-best list are discarded from consideration. Next, the candidate with the highest speech recognition score in the intersection of the n-best lists are chosen as the result. This integration strategy is shown to have a positive effect on the correction of speech recognition errors. The same approach is applied to the problem of resolving ambiguous speech input. According to the authors, nine in ten am-

biguous verbal commands can be resolved with the help of gaze information.

Cooke & Russell (2006) present a method for integrating eye movement information into automatic speech recognition systems for decoding spontaneous, conversational speech in a visually constrained environment. This work relies on the assumption that the fixation of an object in the visual field increases the probability that a subsequent utterance will refer to that object. To implement this assumption Cooke develops *gaze-contingent language models* which provide a probabilistic measure of word likelihood from n-gram models incorporating gaze direction information. These language models shift probability mass continuously depending on the current focus of the speaker's visual attention. The results show that the integration of gaze has little effect on Word Error Rate (WER) but improves Figure Of Merit (FOM) which is based on the number of keywords that are correctly recognized. Cooke argues that the FOM metric is more appropriate for evaluating gaze-contingent ASR performance than WER since it is directly related to the meaning and identification of referents in the visual context. The ASR system is not aided by eye movements in recognizing short and frequent words, e.g., function words, since eye movements do not provide any information about such words. Cooke further argues that the modest increase in recognition performance is explained by the fact that people tend to clearly speak the content words associated with the objects in their visual focus. Since the speech recognizer generally performs well at recognizing these words, there is not much room for improvement using gaze-direction information. However, this is not likely to be the case in a noisy environment and Cooke concludes that it is reasonable to assume that the recognition performance of the gaze-contingent ASR system will increase in such settings.

6 Conclusion

As the present summary demonstrates, the use of eye tracking and eye movement information in speech technology and human computer interaction is currently an active field of research. We believe this research will continue to expand and mature, not least because of the fast growing availability of increasingly advanced, robust and portable eye tracking systems on the market. We also believe that this research is fundamentally necessary given what is known about human language processing and communication. Many important features of human communication rely on extra-linguistic cues such as gaze and gestures. In order to build computational systems designed to interact naturally with humans, these aspects of communication must be taken into account. While most previous research on the integration of gaze and speech has been concerned primarily with designing multimodal input devices able to replace traditional devices such as the mouse and keyboard, we have shown that a broader range of applications are now being considered, including

gaze-based automatic speech recognition and synthesized speech evaluation. We have also shown that much of this research rely on central findings in experimental psycholinguistics. It is clear, therefore, that such research can serve to inform and advance research on speech technologies.

References

Bolt, R.A. (1980). Put-that-there: Voice and gesture in the graphic interface. In *Proceedings of ACM conf. on computer graphics*, New York, 1980, 262–270.

Campana, E. Baldridge, J., Dowding, J., Hockey, B.A., Remington, R.W., Stone, L.S. (2001). Using eye movements to determine referents in a spoken dialogue system. In *Proceedings of Perceptive User Interface* (Orland, FL, 2001).

Cooke, N., and Russell, M. (2005). Using the focus of visual attention to improve automatic speech recognition. In *Proceedings of INTERSPEECH 2005 - 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal.

Eberhard, K.M., Spivey-Knowlton, M.J., Sedivy, J.C., and Tanenhaus, M.K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research* 24, 409–436.

Griffin, Z.M., and Bock, K. (2000). What the eyes say about speaking. *Psychological Science* 11, 274–279.

Griffin, Z.M. (2004). Why look? Reasons for eye movements related to language production. In: J.M. Henderson and F. Ferreira (Eds.), *The integration of language, vision, and action: Eye movements and the visual world.* New York: Psychology Press.

Kaur, M., Tremaine, M., Huang, Ning. (2003). Where is "it"? Event synchronization in gaze-speech input systems. In *Proceedings of ICMI* (Vancouver, Canada, 2003).

Neal, J.G., Thielman, C.Y., Dobes, Z., Haller, S.M. and Shapiro, S.C. (1991). Natural language with integrated deictic and graphic gestures. In *Readings in intelligent user interfaces*. M.T. Maybury (Ed.) Morgan Kaufmann Publishers, 1991, 38–51.

Qvarfordt, P. and Zhai, S. (2005). Conversing with the user based on eye-gaze patterns. In *Proceedings of CHI*, 2005.

Starker, I. and Bolt, R.A. (1990). A gaze-responsive self-disclosing display. In *Proceedings of CHI*, 1990.

Swift, M.D., Campana, E., Allen, J.F., and Tanenhaus, M.K. (2002). Monitoring eye movements as an evaluation of synthesized speech. In *Proceedings of IEEE* (2002).

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., and Sedivy, J.C.

(1995). Integration of visual and linguistic information in language comprehension. *Science* 268, 1632–1634.

Zhang, Q., Go, K., Imamiya, A., Mao, X. (2003). Designing a robust speech and gaze multimodal system for diverse users. In *proceedings of IEEE* (2003).

Zhang, Q., Go, K., Imamiya, A., Mao, X. (2004). Overriding errors in a speech and gaze multimodal architecture. In *proceedings of IUI* (Madeira, Portugal 2004).