New applications of gaze tracking in speech science

Mattias Bystedt[0000-1111-2222-3333] and Jens Edlund[0000-0001-9327-9482]

KTH Royal Institute of Technology, Speech, Music & Hearing, Stockholm
[2] Springer Heidelberg, Tiergartenstr. 17, 69121 Heidelberg, Germany
mbystedt@kth.se, edlund@speech.kth.se

Abstract. We present an overview of speech research applications of gaze tracking technology, where gaze behaviours are exploited as a tool for analysis rather than as a primary object of study. The methods presented are all in their infancy, but can greatly assist the analysis of digital audio and video as well as unlock the relationship between writing and other encodings on the one hand, and natural language, such as speech, on the other.

We discuss three directions in this type of gaze tracking application: modelling of text that is read aloud, evaluation and annotation with naïve informants, and evaluation and annotation with expert annotators. In each of these areas, we use gaze tracking information to gauge the behaviour of people when working with speech and conversation, rather than when reading text aloud or partaking in conversations, in order to learn something about how the speech may be analysed from a human perspective.

Keywords: gaze tracking, speech technology, label acquisition, annotation

1 Introduction

Gaze tracking is used in a number of applications that aim to improve interaction between humans or between humans and machines. Examples include interfaces that utilize gaze tracking to allow hands-free pointing, to track the addressee of speech, or that utilize pupil dilation to track cognitive load, for example in drivers. There is a wide range of such application areas and more. Gaze has been associated with for example cognitive state, cognitive load, direction of visual attention and turn taking in conversation (Eckstein, Guerra-Carrillo, Miller Singley, & Bunge 2017; Rayner 1998).

More recently, a number of new applications of gaze tracking have surfaced in which gaze tracking is used to model the behaviour of people who are somehow working with or analysing actions (e.g. speech), rather than people who are in the process of performing them (e.g. partaking in a conversation).

In the following, we describe three main areas where gaze tracking is exploited in this relatively new manner, as a tool to assist in the analysis of speech and language, rather than as an object of research in itself:

- Modelling of text to be read aloud
- Evaluation and annotation with naïve informants
- Evaluation and annotation with expert annotators

We discuss the potential benefits and the risks involved, and highlight the particular requirements placed on gaze tracking technology by these specific application areas, as compared to other areas where gaze is used, such as in experimental psychology and in interaction design.

2 Modelling of text to be read aloud

Reading texts aloud is an important application area of speech technology. Apart from its use in commercial applications, speech synthesis, or text-to-speech (TTS), is used by government authorities to produce talking books as a means of making texts accessible to those who are for some reason not able to read in the traditional sense of the word. As an example, the Swedish Agency for Accessible Media has produced thousands of talking university text books and continuously produces around a hundred talking newspapers using TTS. The impact of less-than-perfect TTS, then, is great, but we still struggle with finding viable ways of modelling how text should best be read aloud.

A key characteristic of speech is that it exhibits variation in prominence. Speakers make some words more prominent by lengthening and by expansion of spectral characteristics, whereas other words are more backgrounded, that is, reduced. A number of models have been suggested to help assess which words in a text that is to be read aloud should be made prominent, including measures based on the probabilistic aspects of the text (Malisz, Brandt, Möbius, Oh, & Andreeva 2018). Namely, high contextual probability of a word correlates with its lower prominence and vice versa (Aylett & Turk 2004), consequently assessing the probability might help identify words to be made prominent.

Behavioural research has long been using gaze data to study reading, which has illuminated salient linguistic cues used by skilled readers. Word frequency, neighbourhood frequency, and syllable length have all been associated with our reading behaviour and with our processing of textual information. It is therefore likely that gaze data can refine, compliment or substitute statistical models of text.

A key difference between gaze behaviour and language models is that the latter, while efficient in capturing predictability from a within-text point of view, they do not capture the processing steps as they occur online while a person is reading. Gaze data has also been shown to correlate with subjective measures of word predictability in sentences (Bystedt 2016; Schwanenflugel & LaCount 1988).

In ongoing studies, we are attempting to model prominence via gaze, by collating gaze tracking data from multiple readers of the same text to determine to which words readers pay particular attention. We hypothesize that quantified gaze data can accompany and strengthen models based on sample distribution of contextual predictability of words in the text – statistics which frequently form the basis for prosodic models. Others have also pointed out that future prosodic modelling with gaze or other similarly inspired applications could emerge as alternative methods for modelling prosody in TTS (Vainio, Suni, & Aalto 2015).

3 Evaluation and annotation with naïve informants

Speech science and speech technology research are increasingly data driven, and today, the greatest bottleneck for machine learning of speech and human interaction behaviours is no longer limited access to speech data, of which there is a lot, but rather access to useful annotations of speech data. Acquiring good annotations, or labels, on which to train models is expensive and time consuming, and developing efficient methods for these purposes is becoming increasingly important.

In addition, speech technology research is in most cases an iterative process, where a method is developed and trained, then evaluated, and then retrained using more or better data. The results of many evaluation results can be viewed as a series of examples of what went well and what did not work. Seen from another perspective, this is training data, and it is often the case that the results of an evaluation that ends an iteration become the input for the training of the next iteration. Evaluation and annotation, then, are closely related.

Gaze tracking has been used both for evaluation and annotation in speech science. Here, we discuss two methods that operate on naïve informants in the sense that the informants are not professionals, nor have they received any special training to complete their tasks. Instead, these methods tap into more or less subconscious human communicative behaviours. The first method evaluates TTS quality by measuring gaze patterns of informants responding to instructions during an audio-visual task, and the second tracks the gaze target of people observing recorded interactions to learn more about speaker changes in conversation.

Most TTS evaluation methods fall short when it comes to pinpoint *where* in the evaluated speech a problem occurs. When informants fill in a questionnaire after listening to TTS, they are not able to point to particular instances of the speech to which they listened, but rather make general judgements. Gaze data, on the other hand, is synchronous, meaning that it may give insights about a person's perception at the same moment it takes place.

Swift, Campana, Allen and Tanenhaus (2002) first employed gaze data as an objective measure for TTS evaluation. Using a temporal resolution of 60 Hz, they could capture incremental recognition of single words, i.e. phoneme-by-phoneme processing. The experiment was quite specific: informants responded to an audio instruction to fixate one of many items situated in front of them while gaze movement was recorded. Subsequent researchers manipulated prosody in the audio instructions and

looked for facilitation; that is, when the informant completes the task faster than normal, and also compared TTS to a gold standard of human speech.

Van Hooijdonk, Commandeur, Cozijn, Krahmer and Marsi (2007) used the same experimental paradigm to determine that two consecutive instructions to select the same object facilitated object localization when prosodic marking was present. They also found an interaction between object and speech condition in TTS as opposed to human speech, indicating that anticipation and not prosody aided audio recognition of TTS with low intelligibility.

More recently, White, Rajkumar, Ito and Speer (2014) tested prosody on two levels. A target word and its adjective were accented in one of two different ways: Condition 1: adj + noun = L*H* + no-acc; Condition 2: adj + noun = H* + H*, and words with L*H accent were hypothesized to be acoustically salient and therefore most likely to attract attention. Gaze tracking showed that the L*H* marking facilitated object localization when it occurred in human speech, but not in TTS. After acoustic and various metric analyses, they concluded that a combination of data from offline subjective measures (e.g. ratings) and online objective measures (e.g. gaze) can reveal differences between how people perceive and process synthetic as compared to human speech.

When it comes to speaker changes, a recurring problem with data from real conversations is that one cannot be sure that a place where a speaker change occurred was actually a suitable place just because it occurred, as people sometimes flaunt conventions, and similarly one cannot be certain that a place where no speaker change occurred was an inappropriate place for a speaker change.

Based on the observations that lookers-on of a conversation fixate the speaking person and redirect their gaze in expectation of a speaker change (Edlund et al. 2012), gaze tracking of 3rd party observers of conversations has been used to provide insights into speaker changes that might have occurred but did not, and those that occurred where it may not have been expected. Tice and Henetz (2011) introduced the method, and others have since then used it in different experimental settings with similar results (Edlund et al. 2012). In short, the paradigm consists of analysing data from informants viewing a recorded dyadic conversation, which is presented split-screen with one conversant on each side of the screen. The visual attention of the observer is used to assess the predictability of speaker changes that occur, and to point at times where a speaker change could well have occurred but did not. Compared to standard means of annotation, this measure is continuous and depends on real-time perception of a human observer on very small time frames, which creates a signal with strong potential as a machine learning feature.

4 Annotation with expert annotators

Trained professionals are paid to annotate data. The standard way of getting transcriptions of speech, for example, is to pay transcribers to write down what is said, usually painstakingly following a detailed transcription manual. Other tasks are performed similarly. Phonetic segmentation, for example, is the task of splitting utterances into their phonetic units. Phonetically segmented speech is used for a wide range of purposes in

speech technology development, including the training of ASR and TTS. Automatic segmentation – or *forced alignment* – does this job very well for some recordings, but in other cases manual labour is still required to reach acceptable quality.

Manual phonetic segmentation is an example where data sets on expert annotator gaze behaviour can add new layers of information to training data for developing automatic methods. Khan, Steiner, Sugano, Bulling and Macdonald (2018) captured gaze data and other behaviours from annotators required to draw exact time boundaries between segments in spectrograms. Preliminary results on their data show improvements on automatic segmentation using the behaviour data. At KTH Speech, Music and Hearing, researchers are investigating a similar method where the gaze behaviours of a so-called Wizard-of-Oz, a person controlling a spoken dialogue system behind the scenes for data collection purposes are collected to be used as a feature for training.

5 Requirements as compared to other application areas

In most fields where gaze tracking has been employed, there is a need to pay great attention to aspects such as control, exactness (which requires precise calibration), and clear instructions. Often, the goal of the tracking makes it necessary to know, for each moment in the experiment, exactly where the gaze rests. Conversely, the applications we discuss here generally require large amounts of statistical data, where ill effects of a small proportion of errors may be less damaging. On the other hand, ease of use is important (or the professionals will not agree), and calibration must be unobtrusive or even hidden (lest the presence of the instrument endangers the ecological validity of the experiment). In general, the requirements of these applications are more concerned with usability issues and less with experimental control.

6 Conclusions and next steps

In summary, we believe that tracked gaze behaviours can boost a wide range of speech technology and spoken interaction research methods. However, gaze applications of this nature place different requirements on the gaze tracking technology, and the methods are still in their infancy.

Our work for the near future focusses on modelling of text to be read aloud with TTS and on further exploring the possibilities afforded by registering 3rd party observers' gaze behaviours. We predict, however, that as gaze tracking hardware becomes increasingly accessible, the range of applications of the type discussed here will quickly grow.

References

[1] M. K. Eckstein, B. Guerra-Carrillo, A. T. Miller Singley, and S. A. Bunge, "Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development?," *Dev. Cogn. Neurosci.*, vol. 25, pp. 69–91, Jun. 2017.

[2] K. Rayner, "Eye movements in Reading and Information Processing: 20 Years of Research.," *Psychol. Bull.*, 1998.

[3] Z. Malisz, E. Brandt, B. Möbius, Y. M. Oh, and B. Andreeva, "Dimensions of Segmental Variability: Interaction of Prosody and Surprisal in Six Languages," *Front. Commun.*, vol. 3, p. 25, Jul. 2018.

[4] M. Aylett and A. Turk, "The Smooth Signal Redundancy Hypothesis: A Functional Explanation for Relationships between Redundancy, Prosodic Prominence and Duration in Spontaneous Speech," *Lang. Speech*, vol. 47, no. 1, pp. 31–56, 2004.

[5] P. J. Schwanenflugel and K. L. LaCount, "Semantic Relatedness and the Scope of Facilitation for Upcoming Words in Sentences," *J. Exp. Psychol. Learn. Mem. Cogn.*, 1988.

[6] M. Kurnik, "Bilingual lexical access in Reading," *Cent. Res. Biling. Dep. Swedish Lang. Multiling.*, 2016.

[7] M. Vainio, A. Suni, and D. Aalto, "Emphasis, Word Prominence, and Continuous Wavelet Transform in the Control of HMM-Based Synthesis," Springer, Berlin, Heidelberg, 2015, pp. 173–188.

[8] M. D. Swift, E. Campana, J. F. Allen, and M. K. Tanenhaus, "Monitoring eye movements as an evaluation of synthesized speech," *Proc. 2002 IEEE Work. Speech Synth.*, no. November, pp. 19–22, 2002.

[9] C. Van Hooijdonk, E. Commandeur, R. Cozijn, E. Krahmer, and E. Marsi, "Using eye movements for online evaluation of speech synthesis," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 1, no. August 2015, pp. 217–220, 2007.

[10] M. White, R. Rajkumar, K. Ito, and S. R. Speer, "Eye tracking for the online evaluation of prosody in speech synthesis," in *Natural Language Generation in Interactive Systems*, A. Stent and S. Bangalore, Eds. Cambridge: Cambridge University Press, 2014, pp. 281–301.

[11] J. Edlund *et al.*, "3rd party observer gaze as a continuous measure of dialogue flow," in *LREC 2012*, 2012.

[12] M. Tice and T. Henetz, "The eye gaze of 3rd party observers reflects turn-end boundary projectio," 2011.

[13] A. Khan, I. Steiner, Y. Sugano, A. Bulling, and R. Macdonald, "A Multimodal Corpus of Expert Gaze and Behavior during Phonetic Segmentation Tasks," in *LREC 2018*, 2018.