GAZE-ENHANCED SPEECH RECOGNITION

Malcolm Slaney, Rahul Rajan, Andreas Stolcke, and Partha Parthasarathy

Microsoft Corporation, 1065 La Avenida, Mountain View, CA 94043

malcolm@ieee.org, andreas.stolcke@microsoft.com, sarangp@microsoft.com, rahul.rajan@gmail.com

ABSTRACT

This work demonstrates through simulations and experimental work the potential of eye-gaze data to improve speech-recognition results. Multimodal interfaces, where users see information on a display and use their voice to control an interaction, are of growing importance as mobile phones and tablets grow in popularity. We demonstrate an improvement in speech-recognition performance, as measured by word error rate, by rescoring the output from a large-vocabulary speech-recognition system. We use eye-gaze data as a spotlight and collect bigram word statistics near to where the user looks in time and space. We see a 25% relative reduction in the word-error rate over a generic language model, and approximately a 10% reduction in errors over a strong, page-specific baseline language model.

Index Terms— Speech Recognition, Eye Gaze, Pointing

1. INTRODUCTION

In this paper we show that eye-gaze information can make automatic speech recognition (ASR) easier, and thus improves performance.

In many multimodal situations, users gather information visually and would like to issue spoken commands. This is a powerful interaction modality because the eyes can select and gather data quickly, and spoken natural language is a natural, high-bandwidth control channel. This type of interaction is becoming more important as displays of all size become more ubiquitious. A typical interaction might be to issue a search query to a restaurant review site, selecting by voice one of many options, asking for more information such as address and reservation times, until the task is completed.

Speech recognition, as it is often practiced now, is still a difficult problem. Detailed language models that dramatically reduce the perplexity and thus the size of the problem were the first and most important breakthrough for successful ASR. Constrained language models, appropriate for a specific task, such as travel or reminders, further reduce the language model (LM) perplexity and increase performance. We want to use pointing or eye gaze to further constrain the problem.

It is now easier than ever to acquire eye-gaze and pointer information [1]. We can use infrared light to illuminate the eye and a camera that looks for the glint. Commercial eye-gaze hardware costs as little as $100 and less expensive solutions have been proposed for mobile phones. One disadvantage of eye gaze, however, is that it is more difficult to get the required resolution in an image of the eye at more than .6 to .8 meters from the screen. Gathering gesture and pointing data is even easier because of the larger target, and will have similar advantages and uses as eye-gaze data.

Our work is motivated by two previous studies. Zhai and his IBM colleagues note that "... it is unnatural to overload a perceptual channel such as vision with a motor-control task" [2]. While we can direct our gaze anywhere we want, it is very unnatural to do this for longer periods of time because we primarily use our eyes to gather information about the environment. Thus eye-gaze data is good for providing contextual information about what a user has seen. We can then leverage this information to predict what the user might say next, and further constrain the ASR LM. This additional information about the user's needs and context will make the speech-recognition system's job easier.

The second motivational study comes from prior work in our lab on using natural pointing gestures to constrain a conversational web browser [3]. In this study users' gestures were used to constrain the recognizer to valid web links, as indicated by their anchor text (usually underlined with a blue line.) They demonstrated an improvement in the word error rate (WER) from 28.2% to 23.7% (a relative improvement of 16%) when recognizing utterances corresponding to links, without a loss of accuracy on the rest of the tasks. However, this was only using an explicit gesture to help figure out which of the active links to select, as opposed to full speech recognition and implicit eye gaze as we study in this paper.

We postulate that eye-gaze data can help a multimodal conversational system in three ways: constraining ASR, disambiguating dialog intentions or natural language processing (by identifying ambiguous references in a spoken utterance), and predicting user-interface failures (by noting characteristic user behaviors when the system does not behave as the user expects). In this paper we only address the first issue, but the other advantages are orthogonal to ASR, provide additional performance, and further justify gathering eye-gaze data.

Section 2 describes an experiment we performed to gather eye-gaze data while the user read phrases from the screen. We collected eye-gaze information before and while our users were reading text. We then adapted the LM of an ASR system to recognize the spoken text, as described in Section 3. Finally, we analyze the connection between eye-gaze and speech-recognition results in Section 4.

2. DATA

We collected data by asking subjects to interact using only their voice with web pages displayed on a large monitor. A wizard had access to the real-time eye-gaze information, interpreted the user's utterances, and performed all necessary actions on behalf of the user. We recorded all interactions, segmented the audio [4], transcribed each utterance, and labeled each utterance with its purpose (navigation, reading, text input, or selection).

We asked users to perform two kinds of tasks: natural web browsing and reading. In the natural task we asked subjects to perform web-based tasks such as finding a restaurant and making a specific dinner reservation, or buying some shoes. But this only gives a few utterances per task, so we also asked the user to perform additional readings. In this task—using the front pages from the New York Times, Bing News and Yahoo News—subjects were

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

asked to choose random text on the page and read it aloud. This semi-natural task gave us about 60 utterances per subject, with which we can study the connection between speech recognition and eye gaze. We collected data from 27 subjects who were fluent, but mostly non-native, speakers of English.

We collected real-time eye-gaze data using a Tobii REX. Users were seated at slightly more than arm's distance from a 24" display. We used the standard Tobii calibration process. This system provide eye-gaze information at approximately 30 Hz.

Our eyes process information during short fixation times when the eye is not moving. The eyes reorient during quick ballistic movements known as saccades, but we are not processing information during these times. To know what has been read we need to identify the fixation points. We used an algorithm by Salvucci [5] to identify each fixation point from our eye-gaze data. For each recorded eye-gaze location, we look for a set of points extending over at least 100ms that are clustered together. A cluster is defined by a Manhattan distance of less than 40 pixels. Thus with an average sampling rate of 30Hz, we need at least three points in close proximity to determine that there is a fixation point. In this work, we use the centroid of this cluster as the fixation point, and we assume that the subject is only reading text in close proximity to this point.

Eye-gaze data only gives us a hint about the user's intentions. Most importantly, eye-gaze locations are a time-varying signal, and the time and locations of fixation point tells us what the user might have seen, and thus what they might speak next. Unfortunately, both the temporal and spatial extent of the information-gathering process are unknown.

Figure 1 shows a typical example of the eye-gaze data, superimposed on the text from a news web page. Words were scraped from each web page's DOM (document object model) and the word's location recorded at the center of its bounding box on the original page, regardless of the font size. We fix the temporal axis to the end of the utterance because we can assume by this time the user has gathered all necessary information. In this example, the user's eye gaze has settled close to the words that were spoken, and then already moved on to the next task before the end of the spoken utterance.

Most of the results in this paper are functions of the radius in pixels of the visual "spotlight," a circular region around the estimated fixation point. It is important to note that the amount of text captured by the spotlight is a function of the font and screen typography. Figure 2 shows how pixels translate into words for our displays. The left plot shows the distance in pixels from one word to its nearest word. The mode of this distribution, at around 16 pixels, probably corresponds to a bit more than the average inter-line spacing (since words are often displaced horizontally on different lines). The right side of Figure 2 shows a more complete picture. The inflection points are significant: an inter-line distance of 10 pixels, horizontal inter-word distance of 100 pixels, and a full page of words within 2000 pixels. We say that a word is seen if the word's position (center of the bounding box) is within r pixels of the fixation point.

We studied two versions of our experimental data. The full data set consisted of 608 utterances, but this dataset also contained errors due to web-scraping problems because of dynamic content and parsing mistakes. Thus we also considered a conservative cleaning of the data to remove utterances where the eye-gaze data had little to do with the spoken words. We did this by removing any utterances where the closest distance between the eye-gaze tracks and any word in the spoken utterance was more than 100 pixels. This left us with 365 utterances in the clean dataset. We call these two variations the full and cleaned experiments.

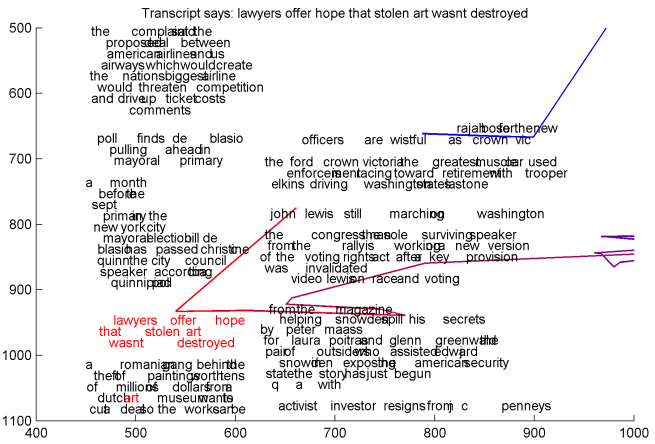

Fig. 1. A subject's eye-gaze locations while choosing text for reading from a portion of the front page of a web news site. Words are shown in their original position, without regard to font size or character spacing. Each block of text is a teaser for a longer article on the site. The words in the spoken utterance are shown in red. The line shows eye-tracking data and the line changes from blue, a full 20 seconds before this utterance, to the color red at the end of the utterance.

3. RECOGNITION EXPERIMENTS

We use a state-of-the-art large vocabulary speech recognizer in our experiments [6]. The acoustic models incorporate the latest advances in context-dependent deep neural networks (DNN) for estimating senone likelihoods. The language model (LM) is a general-purpose backoff 4-gram model with a vocabulary of about 400K words. This generic LM (GLM) was trained on a wide variety of sources ranging from transcribed speech from deployed ASR applications, such as voice search, to text from a diverse set of web resources. The GLM was not tailored or adapted to the tasks of our study.

To study the potential benefit of eye-gaze information for speech recognition we performed LM adaptation experiments in an N-best rescoring framework [7]. We generated lists of the 100 best hypotheses for each utterances, using the GLM. The baseline word error rate was 43.8%. The best achievable (oracle) error rate, by rescoring the 100 best hypothesis, was 22.5%.

Besides the generic LM, we also investigated a second, stronger baseline system in which we derive an utterance-specific bigram LM from the full screen contents, irrespective of eye-gaze information. This LM is restrictive since there are only roughly one thousand words on a single page. The utterance-specific whole-page LM was combined with the GLM via log-linear score combination at the utterance level. This corresponds to a log-linear interpolation of the two LMs [8], but without normalizing the combined probability distribution. We estimated the linear weights for GLM and utterance-specific LM log probabilities on one half of the test speakers and applied to the other half, in a jack-knifing experiment. The N-best hypotheses were rescored with the combined LM and the new 1-best hypotheses extracted.

Finally, we built gaze-conditioned utterance-specific LMs, based on the estimated location of the user's gaze before and during the time of each utterance. To build the gaze-conditioned LM, we collected words appearing on the screen at the appropriate times and locations, as described in the next section. We then found bigrams by sorting the word locations into reading order, and combining words

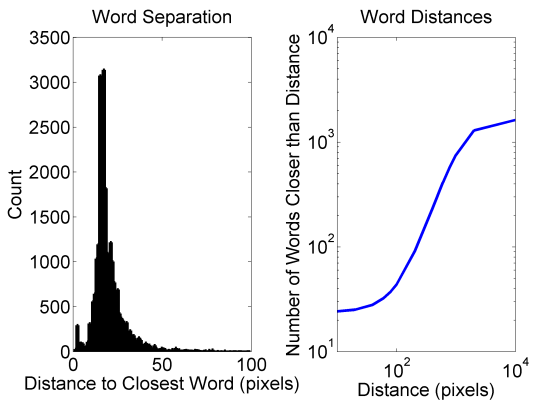

Fig. 2. Distances between words (left) and word counts as a function of distance (right). Eye-gaze performance, especially on a speech-reading task, depends on the inter-word spacing. These plots summarize our experimental displays. A visual angle of 1° corresponds to 50 pixels for the user position and display used in this study.

into bigrams if they are on the same line and adjacent to each other. From the bigrams thus collected, another utterance-specific LM was estimated, and combined with both baseline LMs (GLM and whole-page LM) via log-linear score combination, again using jack-knifing for weight estimation.

4. RESULTS

We evaluate the performance of eye gaze using three types of measures: sensor noise, information retrieval, and ASR.

4.1. Noise Measurements

To get a sense of the eye-tracking measurement noise, we asked a subject to fixate for a few seconds at one of several points on the screen. The resulting eye-gaze measurements had a standard deviation of 9.6 pixels. Additional noise sources, which we did not measure, include static biases due to effects such as inter-person differences and subject placement, and dynamic variations due to subject movements and lighting variations. Thus the 9.6 pixel noise that we measured is a lower bound on the quality of the eye-gaze data from the Tobii REX used in this experiment.

4.2. Measures of Information Retrieval

It is useful to think of the eye-gaze locations in terms of an information-retrieval (IR) problem [9]. The eye-gaze data provides a filter or a spotlight over time, from which we want to infer the user's spoken information (as transcribed by humans). The eye-gaze data has high precision if it reliably points to *only* the words that are spoken, while high recall means that the actual spoken words are *always* in the visual spotlight.

Figure 3 shows precision and recall as a function of the visual-spotlight size, for both idealized and measured eye-gaze data, for all our reading data. First start with the ideal case. For each spoken utterance we found the positions of the most likely cluster of these words on the page. We threw out any utterance where we could not find all the words in the DOM (indicating a system parsing error). Thus the idealized recall is always 1.0. The idealized precision is more interesting. As the spotlight radius increases, more and

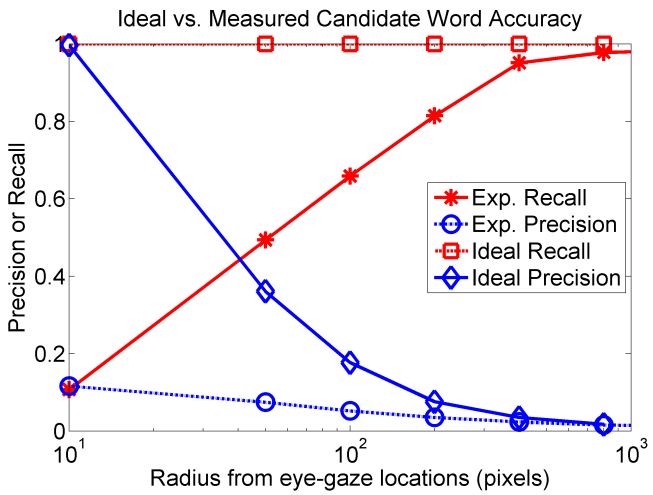

Fig. 3. Precision and recall vs. analysis radius for experimental (solid lines) and idealized (dashed) gaze measurements. In this plot we calculated precision and recall using words within the eye-gaze spotlight from 20 seconds before the start of the utterance and to the end of the utterance.

more extraneous words are included in the eye-gaze signal, decreasing precision, and providing less of a constraint for the recognizer.

We performed a similar precision–recall analysis using our experimental subjects' eye-gaze data. We measured recall based on whether the subject saw the spoken words over a period of time before and during the spoken utterance. Again, precision is calculated based on the total number of words seen by the subject, as indicated by the eye-tracking signal and the spotlight radius.

In our experimental setup, precision and recall show the amount of noise in the resulting spotlight signal. At the very smallest spotlight radius there is only a 10% chance of seeing one of the spoken words, and only about 10% of the words in the spotlight are part of the user's utterance. As the spotlight radius increases the recall goes up, reaching 50% at a radius of 350 pixels. Likewise precision goes down as the radius increases. At the same radius of 350 pixels, only 8% of the words within the spotlight are part of the user's utterance. (The cleaned data has slightly lower precision and recall because invariably we filter out some words that matter to the recognition task.)

The location of the eye gaze varies over time. We don't know the relationship between what and when the user sees some text and what they might say. But it is reasonable to assume that there is a small window of time, before the user speaks, that has the biggest effect on the upcoming speech request. The ultimate test to decide this question is via an ASR or application-specific metric, but we can use IR metrics to get initial insight.

Figure 4 shows the f-measure of the text identified by eye gaze, as a function of both temporal and spatial windows. The f-measure is a standard technique for combining the precision and recall of an IR system into a single metric. It is computed as the harmonic mean of the precision and recall measurements, and thus weights their contributions equally. The f-measure shows a peak for a temporal window starting 2 seconds before the speech utterance and a spatial radius of about 50 pixels. We used the 2 second window in the ASR studies.

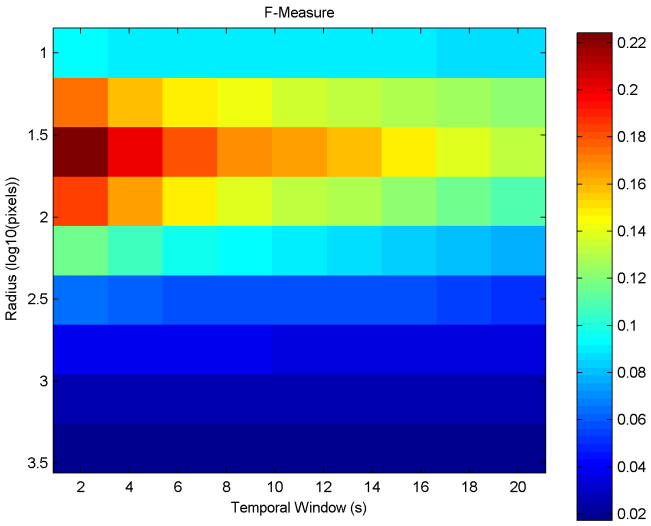

Fig. 4. The f-measure for experimental eye-gaze data. We calculated the match between the spoken utterance and the words seen around the subject's corresponding eye-gaze locations. We calculate precision and recall as a function of the spatial and temporal windows, and from this we calculated the f-measure.

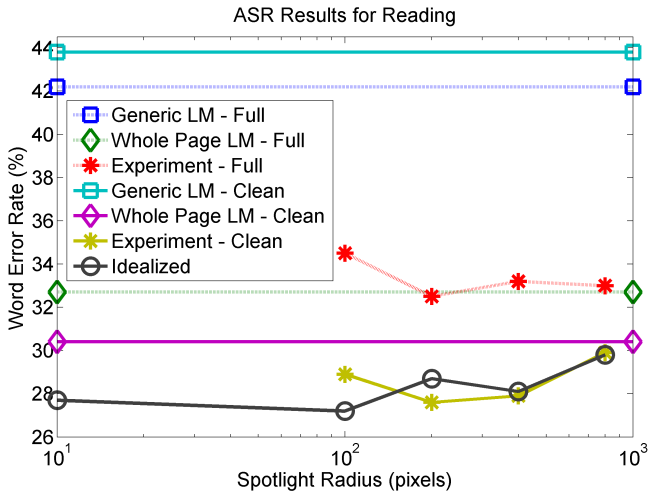

Fig. 5. Multimodal word error rates for ASR with and without eye-gaze information. Error rates with the eye-gaze data are plotted as a function of the spotlight radius (in pixels.) Recognition results assume a temporal window that extends from 2 second before the utterance to the end of the utterance.

4.3. ASR Results

Figure 5 summarizes the ASR results from our experiments. The dashed lines show the results from the full experiment, using all the utterances we collected. The solid lines show the results for the clean data. In each case we show the ASR results with the full 400k word generic LM (squares), the full-page LM (diamonds), and the experimental results (stars). Finally, the ideal results, assuming perfect eye-gaze data are shown with circles. The full and clean experiments have different baseline results because they represent two different sets of data (one a subset of the other).

The high error rate for the Generic LMs (as well as the perplexity of 1587) reflect the fact that the GLM is not well-matched to the user task in our experiment. Even with perfect (oracle) rescoring the best we could do with this two-pass process was a 22% error rate. The whole-page LMs are a tough baseline. This is information that is readily available to a multimodal interactive system, and represents a more practical baseline. We want to know if eye-gaze data can improve upon these two baseline measures.

Both the idealized result and the result with our clean data (no utterances where the eye-tracks were more than 100 pixels from the words) show approximately a 10% relative improvement over the results with the full-page LM. The lower line with stars shows ASR results using the clean experimental eye-gaze data, i.e., the GLM combined with the page-specific and gaze-based utterance-specific bigram LMs. The eye-gaze data improves upon the whole-page LM results and shows improved results for a visual spotlight of about 200 pixels. At very large radii the spotlight is seeing the entire page, so the results approach the whole-page LM results.

The bottom curve (circles) shows the potential advantage of eye gaze to improve speech recognition. We used the idealized eye-gaze data described in Section 4.2 to form a fixed visual spotlight around each word in the utterance's transcript. Depending on the size of the spotlight, this becomes a very constrained language model; linear interpolation of the GLM with page-specific and utterance-specific LM lowers the perplexity to 7 for the smallest radius. The idealized eye-gaze data with a radius of 10 pixels reduces the relative error rate by 10–20% over the best baseline result.

Perhaps a better way to characterize the benefit of eye gaze is to look at the perplexity of the task. The perplexity of our reading task using the GLM is 1587, and tightening the language model to the optimal linear interpolation of the GLM and the screen contents reduces the perplexity to a more manageable 26. With our clean data, the GLM plus eye gaze data reduces the perplexity to 15–17, and combining GLM, whole screen and eye gaze gives a perplexity of 14. This is almost a factor of two advantage over the best perplexity without the eye-gaze data.

5. CONCLUSIONS

Our work shows the power of leveraging better ASR in multimodal system. In both an idealized experiment, and a real experiment with cleaned data we saw a 10% relative reduction in word-error rates. This improvement will only get better as eye-gaze measurements become more accurate and/or displays get larger.

With a high-perplexity language model, eye-gaze data has the potential to constrain the speech-recognition task, by better informing the ASR system of the user's intentions. But, as mentioned in the introduction, better ASR is only one use for the eye-gaze information. In our user study we also saw many occasions where ambiguous user requests, such as providing an address, were perfectly clear from the user's eye gaze. This suggests that dialog systems involving selection steps would benefit from gaze. Finally, we expect that users' confusion and user-interface failure will be all to obvious from the eye-gaze information.

6. ACKNOWLEDGMENTS

We are grateful for assistance we received from Lisa Stifelman, Susan Dumais, Zhengyou Zhang, Larry Heck, Dilek Hakkani-Tur, Ashley Fidler, Hugo Hernandez, and Adrian Medeiros.

7. REFERENCES

[1] Dan Witzner Hansen, Qiang Ji. In the Eye of the Beholder: A Survey of Models for Eyes and Gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, no. 3, pp. 478–500, March, 2010

[2] Shumin Zhai, Carlos Morimoto, Steven Ihde. Manual And Gaze Input Cascaded (MAGIC) Pointing. In *Proceedings of CHI99*, pp. 246–253, 1999.

[3] Larry P Heck, Dilek Hakkani-Tur, Madhu Chinthakunta, Gokhan Tur, Rukmini Iyer, Partha Parthasarathy, Lisa Stifelman, Elizabeth Shriberg, Ashley Fidler. Multi-Modal Conversational Search and Browse. *Proceedings of the SLAM workshop at INTERSPEECH,* pp. 96–101, 2013.

[4] Lori Lamel, Lawrence Rabiner, Aaron Rosenberg, J. Wilpon. An improved endpoint detector for isolated word recognition *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(4), pp. 777–785, 1981.

[5] Dario D. Salvucci, Joseph H.Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Eye Tracking Research & Application*, pp. 71–79, 2000.

[6] Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael Seltzer, Geoff Zweig, Xiaodong He, Jason Williams. Recent advances in deep learning for speech research at Microsoft. In the *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* 2013.

[7] M. Ostendorf, A. Kannan, S. Austin, O. Kimball, R. Schwartz, and J. R. Rohlicek. Integration of diverse recognition methodologies through reevaluation of N-best sentence hypotheses. In *Proceedings of the workshop on Speech and Natural Language (HLT '91).* Association for Computational Linguistics, Stroudsburg, PA, USA, 83–87, 1991.

[8] D. Klakow. Log-linear interpolation of language models. In *Proceedings of the International Conference on Spoken-Language Processing (ICSLP),* p. 1695, 1998.

[9] Ricardo Baeza-Yates, Berthier Ribeiro-Neto. *Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition).* Addison-Wesley Professional, 2011.