The Relation of Eye Gaze and Face Pose: Potential Impact on Speech Recognition

Malcolm Slaney
Microsoft Research
Mountain View, CA USA
malcolm@ieee.org

Andreas Stolcke
Microsoft Research
Mountain View, CA USA
Andreas.Stolcke@microsoft.com

Dilek Hakkani-Tür
Microsoft Research
Mountain View, CA USA
dilek@ieee.org

ABSTRACT

We are interested in using context to improve speech recognition and speech understanding. Knowing what the user is attending to visually helps us predict their utterances and thus makes speech recognition easier. Eye gaze is one way to access this signal, but is often unavailable (or expensive to gather) at longer distances. In this paper we look at joint eye-gaze and facial-pose information while users perform a speech reading task. We hypothesize, and verify experimentally, that the eyes lead, and then the face follows. Face pose might not be as fast, or as accurate a signal of visual attention as eye gaze, but based on experiments correlating eye gaze with speech recognition, we conclude that face pose provides useful information to bias a recognizer toward higher accuracy.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Speech recognition and synthesis

General Terms

Evaluation

Keywords

Speech recognition; eye gaze; face pose; language models

1. MOTIVATION

Several research groups have demonstrated improved speech recognition with access to a user's eye-gaze information [1, 6]. But eye-gaze information is difficult to obtain at a distance. In this paper we wish to demonstrate the utility of face-pose information as a proxy for eye-gaze information.

Eye gaze information will always be the better source of information. The fovea is highly specialized for gathering information, and is the portion of the eye with the highest spatial resolution. It is certainly possible to glimpse information from the corner of one's eye, but that is probably not how most people read information.

The system we envision combines a screen with speech recognition. Explicit pointing will always be valuable, using gestures, touch or a pointing device. This paper deals with eye-gaze and face-pose information because one must look at an object before you can point at it. This information makes the recognizer's job easier because intent and visual history are both important contextual information for a recognizer.

Our hypothesis is that the orientation of the face, or its pose, is a proxy or an approximation of eye gaze. While one can certainly gaze in a wide range of directions without moving one's head, the natural action appears to be that the eyes move first, and then the head follows. We wish to characterize the temporal course of the head-pose signal, and its accuracy. Both the eye-gaze and the face-pose signals function as a spotlight, effectively selecting certain words on the screen and potentially biasing the speech recognizer's language model (LM).

2. EXPERIMENTAL SETUP

We collected joint data using a Tobii REX for eye-gaze data, and a Microsoft Kinect for face-pose data. Users were asked to read text at random locations on a large screen, thus insuring that the user was focused on the task at hand.

Our two sensors have conflicting requirements. The eye-gaze hardware works best with close distances so it can capture an image of the user's eyes with high resolution, while the face-pose software wants to see more of the body so it can reliably detect the user and identify the face. We wanted to collect simultaneous eye-gaze and face-pose data, so we had a narrow range of user-to-screen distances so as to satisfy the requirements of both devices.

Figure 1 depicts our overall experiment setup. A large (132cm diagonal) display shows text to the user, who stands at a distance of 74cm from a 40 dots per inch (dpi) screen. A Kinect is mounted above the display and looks down on the user, and a REX is mounted at the bottom of the display and looks up at the user's eyes. The user's speech utterances were collected with a headset-mounted noise-cancelling microphone but are not analyzed in this paper. Instead, we compare the relation between eye gaze and face pose in our data with data from a previous experiment that examined the relationship between eye-gaze and speech-recognition performance.

ICMI'14, November 12–16, 2014, Istanbul, Turkey.

http://dx.doi.org/10.1145/2663204.2663251.

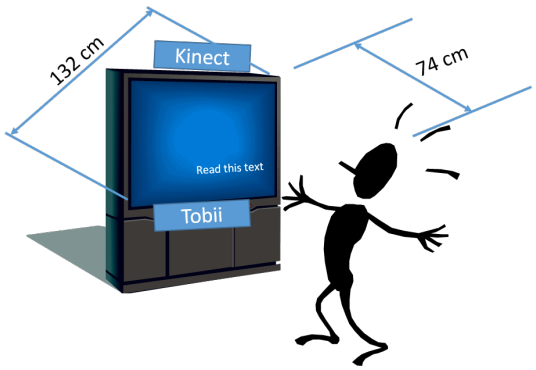

Figure 1: Our basic experimental setup. A user sees a blank screen, and then seeks and reads text placed at random on the screen.

2.1 Tobii REX Eye Tracker

The Tobii REX is an inexpensive eye-tracking sensor that detects infrared glints using a small camera. The specifications for the REX state that it is useful for screen-to-eye distances between 40 and 90cm. In our setup, we got position updates at about 30Hz in pixel coordinates. We used the vendor-supplied software to calibrate the eye-gaze calculations for each user. Our display was larger than their specification allowed, and we were at the limit of their depth range. Thus the calibration software often complained that it was not able to get a good look at all of the points used for calibration. But since we are mostly interested in temporal information, as opposed to precise pixel locations, this was judged to be sufficient for our purposes.

2.2 Microsoft Kinect

The Microsoft Kinect, on the other hand, is a general device for collecting body skeleton information. The Face-Tracking toolkit that is part of the Developer's Toolkit (version 1.8) uses depth and color image data to track one or more faces. Kinect includes a full body tracker, and a special mode in the Kinect for Windows device that only needs the user's upper body for tracking. In this near-range mode the Kinect has a practical range from 80cm to 250cm. The Kinect software returns the location of a face, in meters relative to the camera location, and the 3-D pose of the face as angles relative to the camera coordinate system. The camera has a vertical field of view of 43° We also received face-pose information at a rate of about 30Hz.

The Kinect provides head positions and angles, and does not include any provision for calibration. We needed to transform these positions and angles into screen coordinates. The ultimate solution is a two-camera approach suggested by Huang [3]. Instead, we implemented a simpler solution by putting the camera on the same plane as the screen, thus reducing parallax effects, and then using an affine transform to perform the final mapping into screen coordinates. To effect this transformation we asked the user to turn their head towards 8 different points around the outskirts of the display. We used the face-pose information provided by the Kinect, and simple geometric transformations to transform the raw Kinect data into the camera's imaging plane (which was slightly tilted with respect to the display.) We then found an optimum affine transform that transformed the points in the camera plane to pixels in the computer's world. Again, this transformation is not general, but was deemed sufficiently accurate for this paper's purposes.

2.3 Automatic Speech Recognition

We use a state-of-the-art large vocabulary speech recognizer in our experiments [2, 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 context information for speech recognition we performed LM adaptation experiments in an N-best rescoring framework [5]. 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 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 [4], 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 context-dependent utterance-specific LMs, based on the estimated location of the user's attention before and during the time of each utterance. To build the context-conditioned LM, we collected words appearing on the screen at the appropriate times and locations. We then found bigrams by sorting the word locations into reading order, and combining words 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.

2.4 Display Experiment

Before collecting speech we asked 6 users to calibrate themselves for both the eye gaze and face trackers. In addition to helping us map angles to pixels on the screen, this calibration procedure allowed us to get basic information about the static performance of the system and our users. Users were aware that we were tracking both their head pose and eye gaze, but were not aware of our specific hypothesis.

After calibration we asked each user to perform 20 to 30 speech-reading trials. Before each trial, users were told to look at the center of the screen, where a circle was fixed. Then after a few seconds a short text utterance (a few words from a news headline) with a 6mm high font was displayed

somewhere at random on the screen. This appeared suddenly so there was an orienting response by the user. We also added a short 2cm arrow to the center circle to indicate the direction of the text. We did not enforce a specific gaze location at the start of an experiment. And there was likely both head and eye movement as the user prepared for each trial. The utterance starts some number of seconds after the text appears, and we are interested in the time till the eye gaze and face pose estimates are stable.

3. RESULTS

We would like to relate eye gaze and face pose to speech-recognition performance. This is difficult for a number of reasons, including task and cognitive issues, but also due to simple physical effects. We assume a user absorbs information from within a visual spotlight that moves over time. In addition the sensors have their own physical limitations, which we can model as (Gaussian) noise added to each measurement. Finally, there is some function that relates the probability of a user's comfortable head positions to their desired eye-pose direction. This probability function is certainly related to physical considerations like maintaining comfortable positions, while not shaking the head too much, or too fast. We model the spotlight size as a sum of independent factors: fovea size, eye-tracking error, and if necessary comfortable head-eye orientations.

We compare eye-gaze and face-pose information in three ways. Most simply, we look at the basic sensor error and can quantify the "noise." Secondarily, we look at the time delay between eye gaze and face pose information. Finally, we translate these numbers into a perplexity measure by which we can characterize speech-recognition performance.

3.1 Noise

As part of our calibration procedure, we asked users to stare at a moving dot on the screen. When the dot is not moving, and the user's head is still, we can use the eye-gaze and face-pose information to estimate the inherent noise in the sensor. For the eye gaze, at this screen size the variance was 65^2 pixels. While for the face pose, again given the geometry we used in this study, the variance of the sensor noise was 58^2 pixels. These numbers are important as we look at the optimum spotlight when adjusting the speech recognizer's language model.

3.2 Temporal Characteristics

The temporal patterns of eye gaze and face pose are certainly different. We hypothesize that eyes are faster to orient and then the face catches up. Here we only investigate the delay between eye-gaze and face-pose orientation; the full spectral–temporal relationship between these two signals is beyond the scope of this paper.

We asked subjects to read aloud text phrases we put at random locations onto an otherwise blank screen. Users were told that they could look at the center of the screen for an indication of where the string had appeared, but most users were actively scanning during the experiment.

Since the text was relatively small given the distance (0.5° visual height), users need to orient their eyes onto the text to perform the task. We also observed that users turned their heads toward the text. Thus we characterized the user's eye and facial orientation in terms of their average location when reading the text, and measured the distance to this average location over time. Before the spoken utterance, the distance should be much larger than it is during the utterance. Figure 2 shows an example of this behavior, quantified by the normalized distances from the final average fixation point. We can look for the time of orientation by correlation of the sensor signal with a unit step that goes from +1 to -1 at a variable point in time. With this simple correlation we could estimate the orientation time.[1]

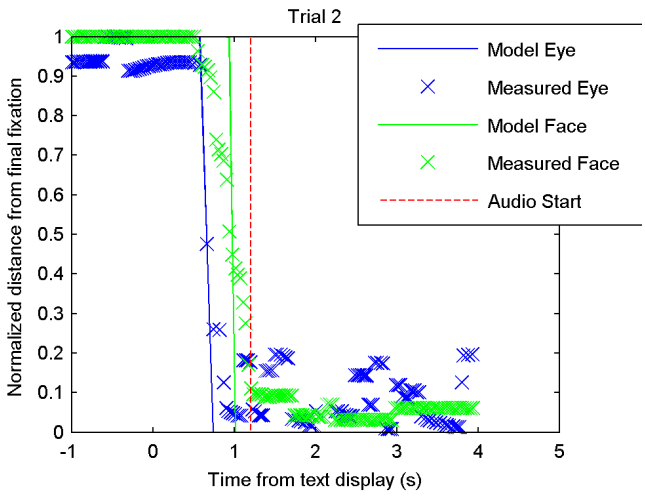

Figure 2: Normalized distance to the text for eye-gaze and face-pose signals during one trial.

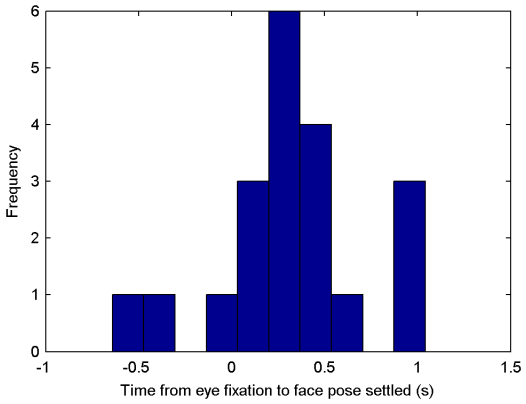

Figure 3: Distribution of the time delay between the eye-gaze fixation on the text and the eventual face orientation to the neutral position.

Because of the random nature of the task, and the user's eye-gaze and head orientation, we found 20 trials where we got a clear signal from both sensors. A histogram of the difference in orientation time between the two sensors is shown in Figure 3. In most cases, the head trailed the eyes by 0.3 seconds. But there were still cases where the head was pointed in the right location before the text appeared, and then eyes had to move to catch up to the head pose.

[1]We also looked at using logistic regression to model the data, but found that noise in the data made it hard to precisely estimate the transition time.

3.3 Perplexity

In a previous study [6] we used a desktop display to measure the effect of eye gaze on speech-recognition performance and LM perplexity during a speech-reading task. The display had a diagonal of 24", 77 dpi, 17 pixels per line of text, eye-tracking noise with variance 9.6^2 pixels, and the user sat 30" from the screen. Figure 4 shows speech recognition difficulty for this reading task when using eye-gaze information to adjust the recognizer's language model. Difficulty is expressed in terms of perplexity, which is a measure of how good the language model is at predicting the next word, given the words it has already seen. Lower perplexity means the language model thinks fewer words are possible, thus reducing the complexity of the speech recognizer's task, and increasing performance. But too small a spotlight removes needed information. Thus the optimum spotlight size was 200 pixels, and reduced the speech-recognition error by about 20%. We would like to know how face-pose information might translate into this domain.

To make the speech-recognizer's job easier, our spotlight should be as small as possible, including all necessary words on the screen, in spite of any sensor errors. In the eye-tracking case there are two components to the spotlight: a cognitive/reading effect and the sensor noise. The same idea holds for face-pose data, but there is also a component that corresponds to the short-term discrepancy between face pose and eye-gaze as the eyes and head adjust to a new task. We do not have an estimate of this variable, except as shown in Section 3.2 that there is a 0.3 second delay.

We use the eye-tracking ASR experiment to guage the impact of using face pose to bias a recognizer. The noise due to the face sensor is higher, but overall has a small effect on the overall perplexity. The spotlight in the desktop display has a radius of 200 pixels, or approximately 5° of arc. The face-sensor noise was 58^2 pixels on a larger display, suggesting a noise of approximately 3° degrees. Overall, this represents less than a factor of two increase in potential spotlight size. This certainly affects perplexity, but as can be seen in Figure 4 the modified perplexity is higher, but still an improvement over the whole-screen perplexity. Thus face pose adds information to the speech recognizer, and, based on the prior study, has the potential to improve recognition accuracy. We hope to verify this in a future experiment that uses the face-pose signal directly for biasing the speech recognizer.

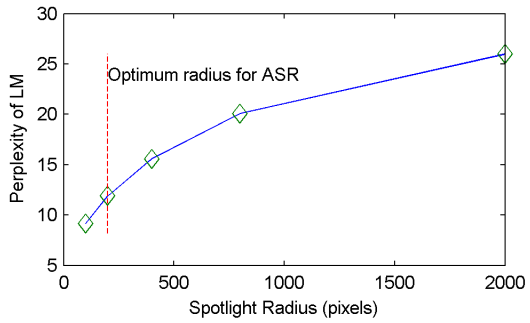

Figure 4: LM perplexity as a function of spotlight radius for the desktop screen, independent of modality.

4. CONCLUSIONS

We demonstrated the viability of face-pose information as a proxy for eye-gaze information. Eye-gaze information reduces the difficulty of the recognition task by a factor of two, in terms of language model perplexity. While eye gaze usually preceeds face orientation and current face sensors are not as accurate as eye trackers, face-pose information has the potential to also significantly reduce LM perplexity.

We have quantified eye-gaze and face-pose information in a joint experiment, where we jointly measure both signals from a single user. While face pose can not tell the whole story, it has similar errors, and a slight delay from the eye-gaze signal. This resulting perplexity reduction is important because it directly impacts speech-recognition performance. Speech recognition will be challenging in the large-display scenarios we envision because of multiple users, reverberent environments, and large microphone-to-user distances.

We need to perform further studies to quantify the effect that face pose information has on the visual spotlight needed for language modeling. This study shows that the errors are manageable, and suggest that there is a significant reduction in perplexity when using face-pose information. By this study we demonstrated the value of these new more ASR experiments using face-pose information.

5. REFERENCES

[1] N. J. Cooke and M. Russell. Gaze-contingent automatic speech recognition. In *Signal Processing*, pages 369–380, 2008.

[2] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, and J. Williams. Recent advances in deep learning for speech research at microsoft. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[3] J.-B. Huang, Q. Cai, Z. Liu, N. Ahuja, and Z. Zhang. Towards accurate and robust cross-ratio based gaze trackers through learning from simulation. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '14, pages 75–82, New York, NY, USA, 2014. ACM.

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

[5] 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)*, pages 83–87. Association for Computational Linguistics, 1991.

[6] M. Slaney, R. Rajen, A. Stolcke, and P. Parthasarathy. Gaze enhanced speech recognition. In *Proceedings of IEEE International Conference onAcoustics, Speech and Signal Processing (ICASSP)*, 2014.