An Exploration of Eye Gaze in Spoken Language Processing for Multimodal Conversational Interfaces

Shaolin Qu **Joyce Y. Chai**
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824
{qushaoli,jchai}@cse.msu.edu

Abstract

Motivated by psycholinguistic findings, we are currently investigating the role of eye gaze in spoken language understanding for multimodal conversational systems. Our assumption is that, during human machine conversation, a user's eye gaze on the graphical display indicates salient entities on which the user's attention is focused. The specific domain information about the salient entities is likely to be the content of communication and therefore can be used to constrain speech hypotheses and help language understanding. Based on this assumption, this paper describes an exploratory study that incorporates eye gaze in salience modeling for spoken language processing. Our empirical results show that eye gaze has a potential in improving automated language processing. Eye gaze is subconscious and involuntary during human machine conversation. Our work motivates more in-depth investigation on eye gaze in attention prediction and its implication in automated language processing.

1 Introduction

Psycholinguistic experiments have shown that eye gaze is tightly linked to human language processing. Eye gaze is one of the reliable indicators of what a person is "thinking about" (Henderson and Ferreira, 2004). The direction of gaze carries information about the focus of the users attention (Just and Carpenter, 1976). The perceived visual context influences spoken word recognition and mediates syntactic processing (Tanenhaus et al., 1995; Roy and Mukherjee, 2005). In addition, directly before speaking a word, the eyes move to the mentioned object (Griffin and Bock, 2000).

Motivated by these psycholinguistic findings, we are currently investigating the role of eye gaze in spoken language understanding during human machine conversation. Through multimodal interfaces, a user can look at a graphic display and converse with the system at the same time. Our assumption is that, during human machine conversation, a user's eye gaze on the graphical display can indicate salient entities on which the user's attention is focused. The specific domain information about the salient entities is likely linked to the content of communication and therefore can be used to constrain speech hypotheses and influence language understanding.

Based on this assumption, we carried out an exploration study where eye gaze information is incorporated in a salience model to tailor a language model for spoken language processing. Our preliminary results show that eye gaze can be useful in improving spoken language processing and the effect of eye gaze varies among different users. Because eye gaze is subconscious and involuntary in human machine conversation, our work also motivates systematic investigations on how eye gaze contributes to attention prediction and its implications in automated language processing.

2 Related Work

Eye gaze has been mainly used in human machine interaction as a pointing mechanism in direct manipulation interfaces (Jacob, 1990; Jacob, 1995; Zhai et al., 1999), as a facilitator in computer supported human human communication (Velichkovsky, 1995; Vertegaal, 1999); or as an additional modality during speech or multimodal communication (Starker and Bolt, 1990; Campana et al., 2001; Kaur et al.,

Proceedings of NAACL HLT 2007, pages 284–291,
Rochester, NY, April 2007.

2003; Qvarfordt and Zhai, 2005). This last area of investigation is more related to our work.

In the context of speech and multimodal communication, studies have shown that speech and eye gaze integration patterns can be modeled reliably for users. For example, by studying patterns of eye gaze and speech in the phrase "move it there", researchers found that the gaze fixation closest to the intended object begins, with high probability, before the beginning of the word "move" (Kaur et al., 2003). Recent work has also shown that eye gaze has a potential to improve reference resolution in a spoken dialog system (Campana et al., 2001). Furthermore, eye gaze also plays an important role in managing dialog in conversational systems (Qvarfordt and Zhai, 2005).

Salience modeling has been used in both natural language and multimodal language processing. Linguistic salience describes entities with their accessibility in a hearer's memory and their implications in language production and interpretation. Linguistic salience modeling has been used for language interpretations such as reference resolution (Huls et al., 1995; Eisenstein and Christoudias, 2004). Visual salience measures how much attention an entity attracts from a user based on its visual properties. Visual salience can tailor users' referring expressions and thus can be used for multimodal reference resolution (Kehler, 2000). Our recent work has also investigated salience modeling based on deictic gestures to improve spoken language understanding (Chai and Qu, 2005; Qu and Chai, 2006).

3 Data Collection

We conducted user studies to collect speech and eye gaze data. In the experiments, a static 3D bedroom scene was shown to the user. The system verbally asked a user a list of questions one at a time about the bedroom and the user answered the questions by speaking to the system. Fig.1 shows the 14 questions in the experiments. The user's speech was recorded through an open microphone and the user's eye gaze was captured by an Eye Link II eye tracker. From 7 users' experiments, we collected 554 utterances with a vocabulary of 489 words. Each utterance was transcribed and annotated with entities that were being talked about in the utterance.

1	Describe this room.
2	What do you like/dislike about the arrangement?
3	Describe anything in the room that seems strange to you.
4	Is there a bed in this room?
5	How big is the bed?
6	Describe the area around the bed.
7	Would you make any changes to the area around the bed?
8	Describe the left wall.
9	How many paintings are there in this room?
10	Which is your favorite painting?
11	Which is your least favorite painting?
12	What is your favorite piece of furniture in the room?
13	What is your least favorite piece of furniture in the room?
14	How would you change this piece of furniture to make it better?

Figure 1: Questions for users in experiments

The collected raw gaze data consists of the screen coordinates of each gaze point sampled at 4 ms. As shown in Fig.2a, this raw data is not very useful for identifying fixated entities. The raw gaze data are processed to eliminate invalid and saccadic gaze points, leaving only pertinent eye fixations. Invalid gaze points occur when users look off the screen. Saccadic gaze points occur during ballistic eye movements between fixations. Vision studies have shown that no visual processing occurs during saccades (i.e., saccadic suppression). It is well known that eyes do not stay still, but rather make small, frequent jerky movements. In order to best determine fixation locations, nearby gaze points are averaged together to identify fixations. The processed eye gaze fixations can be seen in Fig.2b.

Fig.3 shows an excerpt of the collected speech and gaze fixation with fixated entities. In the speech stream, each word starts at a particular timestamp. In the gaze stream, each gaze fixation f has a starting timestamp t_f and a duration T_f. Gaze fixations can have different durations. An entity e on the graphical display is fixated by gaze fixation f if the area of e contains the fixation point of f. One gaze fixation can fall on multiple entities or no entity.

4 Salience Driven Language Modeling

Our goal is to use the domain specific information about the salient entities on a graphical display, as indicated by the user's eye gaze, to help recognition of the user's utterances. In particular, we incorporate this salient domain information in speech recognition via salience driven language modeling.

(a) Raw gaze points (b) Processed gaze fixations

Figure 2: Gaze fixations on a scene

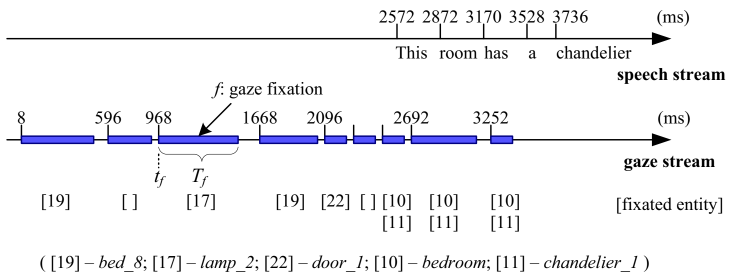


Figure 3: An excerpt of speech and gaze stream data

We first briefly introduce speech recognition. The task of speech recognition is to, given an observed spoken utterance O, find the word sequence W^* such that $W^* = \arg\max_W p(O|W)p(W)$, where $p(O|W)$ is the acoustic model and $p(W)$ is the language model. The acoustic model provides the probability of observing the acoustic features given hypothesized word sequences while the language model provides the probability of a word sequence. The language model is represented as:

$$p(W) = p(w_1^n) = \prod_{k=1}^{n} p(w_k|w_1^{k-1}) \tag{1}$$

Using first-order Markov assumption, the above language model can be approximated by a bigram model:

$$p(w_1^n) = \prod_{k=1}^{n} p(w_k|w_{k-1}) \tag{2}$$

In the following sections, we first introduce the salience modeling based on eye gaze, then present how the gaze-based salience models can be used to tailor language models.

4.1 Gaze-based Salience Modeling

We first define a gaze fixation set $F_{t_0}^{t_0+T}(e)$, which contains all gaze fixations that fall on entity e within a time window $t_0 \sim (t_0 + T)$:

$$F_{t_0}^{t_0+T}(e) = \{f | f \text{ falls on } e \text{ within } t_0 \sim (t_0 + T)\}$$

We model gaze-based salience in two ways.

4.1.1 Gaze Salience Model 1

Salience model 1 is based on the assumption that when an entity has more gaze fixations on it than other entities, this entity is more likely attended by the user and thus has higher salience:

$$p_{t_0,T}(e) = \frac{\#\text{elements in } F_{t_0}^{t_0+T}(e)}{\sum_e(\#\text{elements in } F_{t_0}^{t_0+T}(e))} \tag{3}$$

Here, $p_{t_0,T}(e)$ tells how likely it is that the user is focusing on entity e within time period $t_0 \sim (t_0+T)$ based on how many gaze fixations are on e among all gaze fixations that fall on entities within $t_0 \sim (t_0 + T)$.

4.1.2 Gaze Salience Model 2

Salience model 2 is based on the assumption that when an entity has longer gaze fixations on it than other entities, this entity is more likely attended by the user and thus has higher salience:

$$p_{t_0,T}(e) = \frac{D_{t_0}^{t_0+T}(e)}{\sum_e D_{t_0}^{t_0+T}(e)} \tag{4}$$

where

$$D_{t_0}^{t_0+T}(e) = \sum_{f \in F_{t_0}^{t_0+T}(e)} T_f \tag{5}$$

Here, $p_{t_0,T}(e)$ tells how likely it is that the user is focusing on entity e within time period $t_0 \sim (t_0 + t)$

based on how long e has been fixated by gaze fixations among the overall time length of all gaze fixations that fall on entities within $t_0 \sim (t_0 + T)$.

4.2 Salience Driven N-gram Model

Salience models can be incorporated in different language models, such as bigram models, class-based bigram models, and probabilistic context free grammar. Among these language models, the salience driven bigram model based on deictic gesture has been shown to achieve best performance on speech recognition (Qu and Chai, 2006). In our initial investigation of gaze-based salience, we incorporate the gaze-based salience in a bigram model.

The salience driven bigram probability is given by:

$$p_s(w_i|w_{i-1}) = (1-\lambda)p(w_i|w_{i-1}) + \lambda \textstyle\sum_e p(w_i|w_{i-1}, e)p_{t_0,T}(e) \quad (6)$$

where $p_{t_0,T}(e)$ is the salience distribution as modeled in equations (3) and (4). In applying the salience driven bigram model for speech recognition, we set t_0 as the starting timestamp of the utterance and T as the duration of the utterance. The priming weight λ decides how much the original bigram probability will be tailored by the salient entities indicated by eye gaze. Currently, we set $\lambda = 0.67$ empirically. We also tried learning the priming weight with an EM algorithm. However, we found out that the learned priming weight performed worse than the empirical one in our experiments. This is probably due to insufficient development data. Bigram probabilities $p(w_i|w_{i-1})$ were estimated by the maximum likelihood estimation using Katz's backoff method (Katz, 1987) with a frequency cutoff of 1. The same method was used to estimate $p(w_i|w_{i-1}, e)$ from the users' utterance transcripts with entity annotation of e.

5 Application of Salience Driven LMs

The salience driven language models can be integrated into speech processing in two stages: an **early stage** before a word lattice (n-best list) is generated (Fig.4a), or in a **late stage** where the word lattice (n-best list) is post-processed (Fig.4b).

For the early stage integration, the gaze-based salience driven language model is used together with

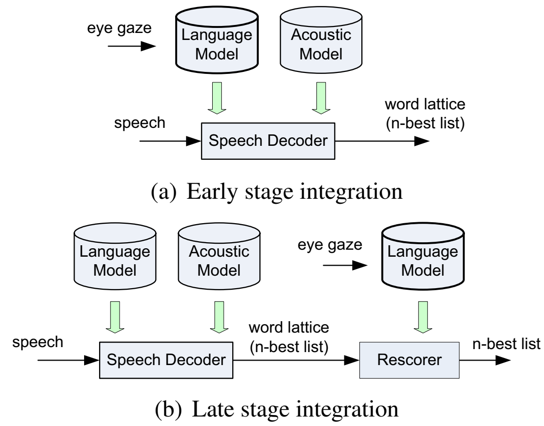


Figure 4: Integration of gaze-based salience driven language model in speech processing

the acoustic model to generate the word lattice, typically by Viterbi search.

For the late stage integration, the gaze-based salience driven language model is used to rescore the word lattice generated by a speech recognizer with a basic language model not involving salience modeling. A* search can be applied to find the n-best paths in the word lattice.

6 Evaluation

The evaluations were conducted on data collected from user studies (Sec. 3). We evaluated the gaze-based salience driven bigram models when applied for speech recognition at early and late stages.

6.1 Evaluation Results

Users' speech was first segmented, then recognized by the CMU Sphinx-4 speech recognizer using different language models. Evaluation was done by a 14-fold cross validation. We compare the performances of the early and late applications of two gaze-based salience driven language models:

- S-Bigram1 – salience driven language model based on salience modeling 1 (Sec. 4.1.1)
- S-Bigram2 – salience driven language model based on salience modeling 2 (Sec. 4.1.2)

Table 1 and Table 2 show the results of early and late application of the salience driven language models based on eye gaze. We can see that all word error rates (WERs) are high. In the experiments, users were instructed to only answer systems questions one by one. There was no flow of a real conversation. In this setting, users were more free to express themselves than in the situation where users believed they were conversing with a machine. Thus, we observe much longer sentences that often contain disfluencies. Here is one example:

> System: "*How big is the bed?*"
> User: "*I would to have to offer a guess that the bed, if I look the chair that's beside it [pause] in a relative angle to the bed, it's probably six feet long, possibly, or shorter, slightly shorter.*"

The high WER was mainly caused by the complexity and disfluencies of users' speech. Poor speech recording quality is another reason for the bad recognition performance. It was found that the trigram model performed worse than the bigram model in the experiment. This is probably due to the sparseness of trigrams in the corpus. The amount of data available is too small considering the vocabulary size.

Language Model	Lattice-WER	WER
Bigram	0.613	0.707
Trigram	0.643	0.719
S-Bigram 1	0.605	0.690
S-Bigram 2	0.604	0.689

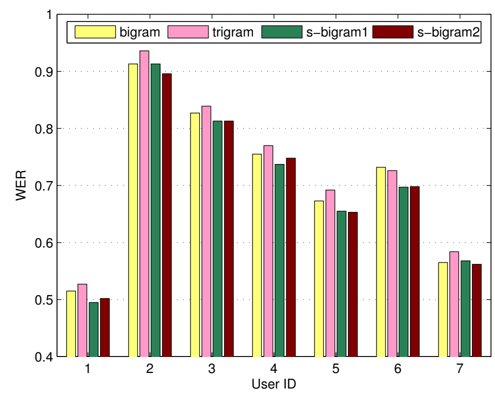
Table 1: WER of early application of LMs

Language Model	Lattice-WER	WER
S-Bigram 1	0.643	0.709
S-Bigram 2	0.643	0.710

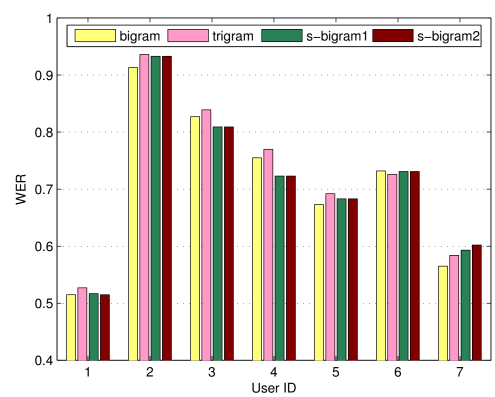
Table 2: WER of late application of LMs

The S-Bigram1 and S-Bigram2 achieved similar results in both early application (Table 1) and late application (Table 2). In early application, the S-Bigram1 model performed better than the trigram model ($t = 5.24$, $p < 0.001$, one-tailed) and the bigram model ($t = 3.31$, $p < 0.001$, one-tailed). The S-Bigram2 model also performed better than the trigram model ($t = 5.15$, $p < 0.001$, one-tailed) and the bigram model ($t = 3.33$, $p < 0.001$, one-tailed) in early application. In late application, the S-Bigram1 model performed better than the trigram model ($t = 2.11$, $p < 0.02$, one-tailed), so did the S-Bigram2 model ($t = 1.99$, $p < 0.025$, one-tailed). However, compared to the bigram model, the S-Bigram1 model did not change the recognition performance significantly ($t = 0.38$, N.S., two-tailed) in late application, neither did the S-Bigram2 model ($t = 0.50$, N.S., two-tailed).

We also compare performances of the salience driven language models for individual users. In early application (Fig.5a), both the S-Bigram1 and the S-Bigram2 model performed better than the baselines of the bigram and trigram models for all users except user 2 and user 7. T-tests have shown that these are significant improvements. For user 2, the S-Bigram1 model achieved the same WER as the bigram model. For user 7, neither of the salience driven language models improved recognition compared to the bigram model. In late application (Fig.5b), only for user 3 and user 4, both salience driven language models performed better than the baselines of the bigram and trigram models. These improvements have also been confirmed by t-tests as significant.



(a) WER of early application



(b) WER of Late application

Figure 5: WERs of LMs for individual users

Comparing early and late application of the salience driven language models, it is observed that early application performed better than late application for all users except user 3 and user 4. T-tests have confirmed that these differences are significant.

It is interesting to see that the effect of gaze-based salience modeling is different among users. For two users (i.e., user 3 and user 4), the gaze-based salience driven language models consistently outperformed the bigram and trigram models in both early application and late application. However, for some other users (e.g., user 7), this is not the case. In fact, the gaze-based salience driven language models performed worse than the bigram model. This observation indicates that during language production, a user's eye gaze is voluntary and unconscious. This is different from deictic gesture, which is more intentionally delivered by a user. Therefore, incorporating this "unconscious" mode of modality in salience modeling requires more in-depth research on the role of eye gaze in attention prediction during multimodal human computer interaction.

6.2 Discussion

Gaze-based salience driven language models are built on the assumption that when a user is fixating on an entity, the user is saying something related to the entity. With this assumption, gaze-based salience driven language models have the potential to improve speech recognition by biasing the speech decoder to favor the words that are consistent with the entity indicated by the user's eye gaze, especially when the user's utterance contains words describing unique characteristics of the entity. These particular characteristics could be the entity's name or physical properties (e.g., color, material, size).

Utterance: "a tree growing from the floor"

Gaze salience:
p(bedroom) = 0.2414 p(plant_willow) = 0.2414
p(chair_soft) = 0.2414 p(door_1) = 0.1378
p(bed_8) = 0.1378

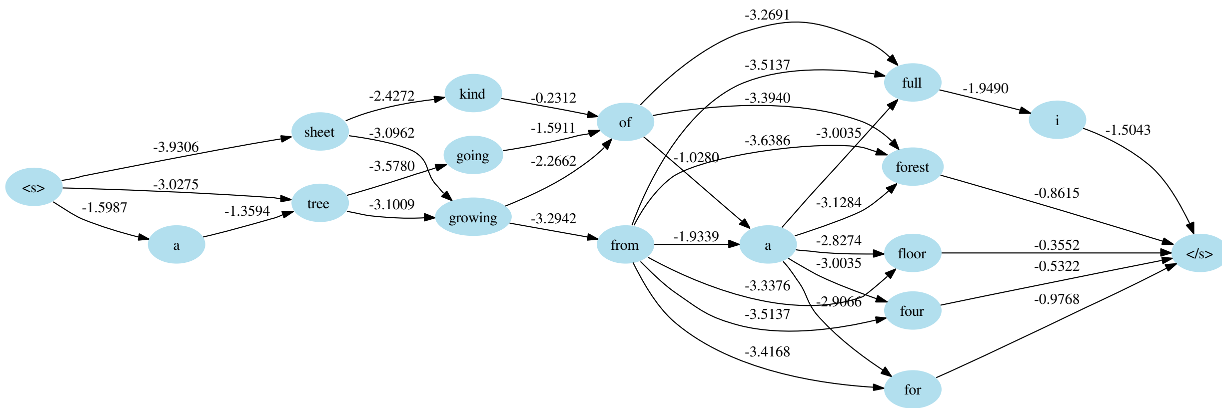
Bigram n-best list:
sheet growing from a four
sheet growing from a for
sheet growing from a floor
...

S-Bigram2 n-best list:
a tree growing from the floor
a tree growing from the for
a tree growing from the floor a
...

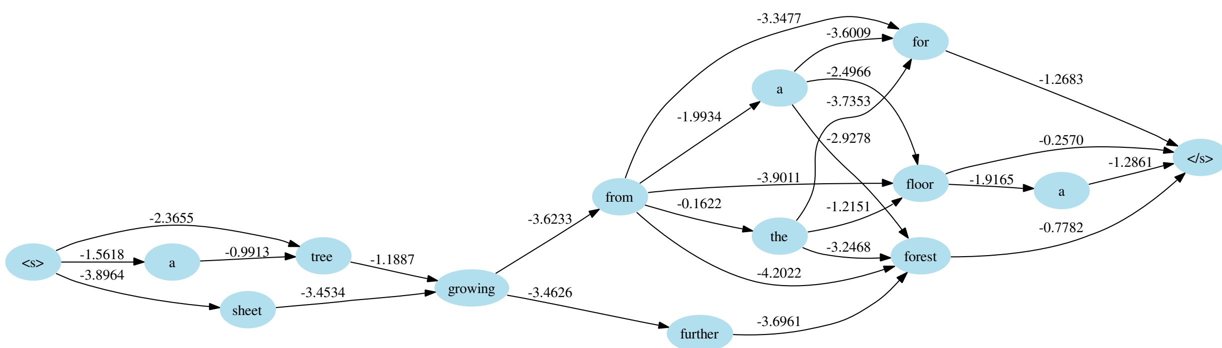
Figure 6: N-best lists of utterance "*a tree growing from the floor*"

Fig.6 shows an example where the S-Bigram2 model in early application improved recognition of the utterance "*a tree growing from the floor*". In this example, the user's gaze fixations accompanying the utterance resulted in a list of candidate entities with fixating probabilities (cf. Eqn. (4)), among which entities *bedroom* and *plant_willow* were assigned higher probabilities. Two n-best lists, the Bigram n-best list and the S-Bigram2 n-best list, were generated by the speech recognizer when the bigram model and the S-Bigram2 model were applied separately. The speech recognizer did not get the correct recognition when the bigram model was used, but got the correct result when the S-Bigram2 model was used.

Fig.7a and 7b show the word lattices of the utterance generated by the speech recognizer using the bigram model and the S-Bigram2 model respectively. The n-best lists in Fig.6 were generated from those word lattices. In the word lattices, each path going from the start node <s> to the end node </s> forms a recognition hypothesis. The bigram probabilities along the edges are in the logarithm of base 10. In the bigram case, the path "<s> a tree" has a higher language score (summation of bigram probabilities along the path) than "<s> sheet", and "a floor" has a higher language score than "a full". However, these correct paths "<s> a tree" and "a floor" (not exactly correct, but better than "a full") do not appear in the best hypothesis in the resulting n-best list. This is because the system tries to find an overall best hypothesis by considering both language and acoustic score. Because of the noisy speech, the incorrect hypotheses may happen to have higher acoustic confidence than the correct ones. After tailoring the bigram model with gaze salience, the salient entity *plant_willow* significantly increases the probability of "a tree" (from -1.3594 to -0.9913) and "tree growing" (from -3.1009 to -1.1887), while it decreases the probability of "sheet growing" (from -3.0962 to -3.4534). This probability change is made by the entity conditional probability $p(w_i|w_{i-1}, e)$ in tailoring of bigram by salience (cf. Eqn. (6)). Probability $p(w_i|w_{i-1}, e)$, trained from the annotated utterances, reflects what words are more likely to be spoken by a user while talking about an entity e. The increased probabilities of "a tree" and "tree growing" show that word "tree" appears more likely than "sheet" when the user is talking about entity



(a) Word lattice with bigram model



(b) Word lattice with S-Bigram 2

Figure 7: Word lattices of utterance "*a tree growing from the floor*"

"*plant_willow*. This is in accordance with our common sense. Likewise, the salient entity *bedroom*, of which *floor* is a component, makes the probability of the correct hypothesis "the floor" much higher than other hypotheses ("the for" and "the forest"). These enlarged language score differences make the correct hypotheses "a tree" and "the floor" win out in the searching procedure despite the noisy speech.

Utterance: "I like the picture with like a forest in it"

Gaze salience:
p(bedroom) = 0.5960 p(chandelier_1) = 0.4040

Bigram n-best list:
and i eight that picture rid like got five
and i eight that picture rid identifiable
and i eight that picture rid like got forest
...

S-Bigram2 n-best list:
and i that bedroom it like upside
and i that bedroom it like a five
and i that bedroom it like a forest
...

Figure 8: N-best lists of utterance "*I like the picture with like a forest in it*"

Unlike the active input mode of deictic gesture, eye gaze is a passive input mode. The salience information indicated by eye gaze is not as reliable as the one indicated by deictic gesture. When the salient entities indicated by eye gaze are not the true entities the user is referring to, the salience driven language model can worsen speech recognition. Fig.8 shows an example where the S-Bigram2 model in early application worsened the recognition of a user's utterance "*I like the picture with like a forest in it*" because of wrong salience information. In this example, the user was talking about a picture entity *picture_bamboo*. However, this entity was not salient, only entities *bedroom* and *chandelier_1* were salient. As a result, the recognition with the S-Bigram2 model becomes worse than the baseline. The correct word "picture" is missing and the wrong word "bedroom" appears in the result.

The failure to identify the actual referred entity *picture_bamboo* as salient in the above example can also be caused by the visual properties of entities. Smaller entities on the screen are harder to be fixated by eye gaze than larger entities. To address this issue, more reliable salience modeling that takes into account the visual features is needed.

7 Conclusion

This paper presents an empirical exploration of incorporating eye gaze in spoken language processing via salience driven language modeling. Our preliminary results have shown the potential of eye gaze in improving spoken language processing. Nevertheless, this exploratory study is only the first step in our investigation. Many interesting research questions remain. During human machine conversation, how is eye gaze aligned with speech production? How reliable is eye gaze for attention prediction? Are there any other factors such as interface design and visual properties that will affect eye gaze behavior and therefore attention prediction? The answers to these questions will affect how eye gaze should be appropriately modeled and used for language processing.

Eye-tracking systems are no longer bulky, stationary systems that prevent natural human machine communication. Recently developed display mounted gaze-tracking systems (e.g., Tobii) are completely non-intrusive, can tolerate head motion, and provide high tracking quality. These features have been demonstrated in several successful applications (Duchowski, 2002). Integrating eye tracking with conversational interfaces is no longer beyond reach. We believe it is time to conduct systematic investigations and fully explore the additional channel provided by eye gaze in improving robustness of human machine conversation.

8 Acknowledgments

This work was supported by a Career Award IIS-0347548 and IIS-0535112 from the National Science Foundation. The authors would like to thank Zahar Prasov for his contribution on data collection and thank anonymous reviewers for their valuable comments and suggestions.

References

E. Campana, J. Baldridge, J. Dowding, B. Hockey, R. Remington, and L. Stone. 2001. Using eye movements to determine referents in a spoken dialogue system. In *Proceedings of the Workshop on Perceptive User Interface*.

J. Chai and S. Qu. 2005. A salience driven approach to robust input interpretation in multimodal conversational systems. In *Proceedings of HLT/EMNLP'05*.

A. T. Duchowski. 2002. A breath-first survey of eye tracking applications. *Behavior Research methods, Instruments, and Computers*, 33(4).

J. Eisenstein and C. M. Christoudias. 2004. A salience-based approach to gesture-speech alignment. In *Proceedings of HLT/NAACL'04*.

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

J. M. Henderson and F. Ferreira. 2004. *The interface of language, vision, and action: Eye movements and the visual world*. New York: Taylor & Francis.

C. Huls, E. Bos, and W. Classen. 1995. Automatic referent resolution of deictic and anaphoric expressions. *Computational Linguistics*, 21(1):59–79.

R. J. K. Jacob. 1990. What you look is what you get: Eye movement-based interaction techniques. In *Proceedings of CHI'90*.

R. J. K. Jacob. 1995. Eye tracking in advanced interface design. In W. Barfield and T. Furness, editors, *Advanced Interface Design and Virtual Environments*, pages 258–288. Oxford University Press.

M. Just and P. Carpenter. 1976. Eye fixations and cognitive processes. *Cognitive Psychology*, 8:441–480.

S. Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recogniser. *IEEE Trans. Acous., Speech and Sig. Processing*, 35(3):400–401.

M. Kaur, M. Termaine, N. Huang, J. Wilder, Z. Gacovski, F. Flippo, and C. S. Mantravadi. 2003. Where is "it"? event synchronization in gaze-speech input systems. In *Proceedings of ICMI'03*.

A. Kehler. 2000. Cognitive status and form of reference in multimodal human-computer interaction. In *Proceedings of AAAI'00*.

S. Qu and J. Chai. 2006. Salience modeling based on non-verbal modalities for spoken language understanding. In *Proceedings of ICMI'06*.

P. Qvarfordt and S. Zhai. 2005. Conversing with the user based on eye-gaze patterns. In *Proceedings of CHI'05*.

D. Roy and N. Mukherjee. 2005. Towards situated speech understanding: Visual context priming of language models. *Computer Speech and Language*, 19(2):227–248.

I. Starker and R. A. Bolt. 1990. A gaze-responsive self-disclosing display. In *Proceedings of CHI'90*.

M. K. Tanenhaus, M. J. Spivey-Knowlton, K. M. Eberhard, and J. E. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634.

B. M. Velichkovsky. 1995. Communicating attention-gaze position transfer in cooperative problem solving. *Pragmatics and Cognition*, 3:99–224.

R. Vertegaal. 1999. The gaze groupware system: Mediating joint attention in multiparty communication and collaboration. In *Proceedings of CHI'99*.

S. Zhai, C. Morimoto, and S. Ihde. 1999. Manual and gaze input cascaded (magic) pointing. In *Proceedings of CHI'99*.