# Speech-Gaze Temporal Alignment for Automatic Word Acquisition in Multimodal Conversational Systems

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

## Abstract

One major bottleneck in conversational systems is their incapability in interpreting unexpected user language inputs such as out-of-vocabulary words. To overcome this problem, conversational systems must be able to learn new words automatically during human machine conversation. Motivated by psycholinguistic findings on eye gaze and human language processing, we are developing techniques to incorporate human eye gaze for automatic word acquisition. Our results indicate that eye gaze provides a potential channel for automatically acquiring new words. Modeling temporal alignment between speech and gaze significantly improves acquisition performance.

## 1 Introduction

Interpreting human language is a challenging problem in human machine conversational systems due to the flexibility of human language behavior. When unexpected inputs (e.g., user vocabulary is outside of system's knowledge) are encountered, conversational systems tend to fail. It is desirable that conversational systems can learn new words automatically during human machine conversation.

Psycholinguistic studies 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). Gaze direction 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). In addition, directly before speaking a word, the eyes move to the mentioned object (Griffin and Bock, 2000).

Motivated by the psycholinguistic findings, we are investigating the use of eye gaze for automatic word acquisition in multimodal conversation. This paper investigates the role of temporal alignment of speech and gaze in automated word acquisition.

## 2 Related Work

Word acquisition by grounding words in representations of entities in the physical world has been studied in many language grounding systems. In these systems, word acquisition has been achieved by associating words with physical objects (Yu and Ballard, 2004; Roy and Pentland, 2002), colors (Mojsilovic, 2005), image regions (Barnard et al., 2003), and event logic expressions defining motions (Siskind, 2001). Unlike the reliable visual attention foci in previous work, the visual attention foci we are working with is indicated by eye gaze. Eye gaze is an implicit and subconscious input, which brings additional challenges in word acquisition.

An initial investigation of word acquisition from speech and eye gaze in human machine conversation was reported in (Liu et al., 2007), in which a translation model was developed to associate words with visual objects on a graphical display. Extending this work, we investigate how temporal information about eye gaze fixation and spoken words can facilitate word acquisition.

## 3 Data Collection

We conducted user studies to collect speech and eye gaze data. In the experiments, a 3D room scene, as shown in Figure 1, was shown to the user. The system verbally asked the user a question or issued a request about the room. The user provided responses by speaking to the system. The user's speech was recorded and the user's eye gaze was captured by an Eye Link II eye tracker.

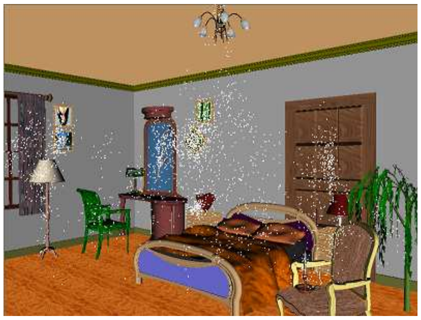


Figure 1: Domain scene with a user's gaze fixations

Users' speech was transcribed. The collected raw gaze data consists of the screen coordinates of each gaze point sampled every 4 ms. These raw gaze points are very noisy. They can not be used directly for identifying fixated entities in the scene. We processed the raw gaze data to eliminate invalid and saccadic gaze points. Since eyes do not stay still but rather make small, frequent jerky movements, we average nearby gaze points to better identify gaze fixations.

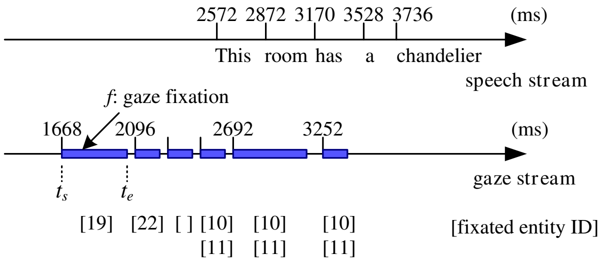


Figure 2: Parallel speech and gaze streams

Figure 2 shows an excerpt of the collected speech and gaze fixation with fixated entities in one experiment. In the speech stream, each word starts at a particular timestamp. In the gaze stream, each gaze fixation has a starting timestamp $t_s$ and an ending timestamp $t_e$. An entity $e$ on the graphical display is fixated by gaze fixation $f$ if the area of $e$ contains fixation point of $f$. Multiple entities or no entity can be fixated by one gaze fixation.

Given the collected speech and gaze fixations, we build parallel speech-gaze data set as follows. For each spoken utterance and its accompanying gaze fixations, we construct a pair of word sequence and entity sequence $(\mathbf{w}, \mathbf{e})$. Since only nouns and adjectives are meaningful for the task of word acquisition in our domain. The word sequence $\mathbf{w}$ consists of only nouns and adjectives in the utterance. Each gaze fixation results in a fixated entity in the entity sequence $\mathbf{e}$. When multiple entities are fixated by one gaze fixation due to the overlapping of the entities, the forefront one is chosen.

## 4 Translation Model for Word Acquisition

The task of word acquisition is to associate spoken words with entities (3D objects) in the domain scene (Figure 1). Viewing this as a translation problem, we use translation models to acquire words for entities by estimating probabilities $p(w|e)$ from co-occurring speech and gaze fixation data set $\{(\mathbf{w}, \mathbf{e})\}$ and then choosing the most likely words for entity $e$.

### 4.1 Model Without Alignment

Using the translation model I (Brown et al., 1993), where each word is equally likely to be aligned with each entity, we have

$$p(\mathbf{w}|\mathbf{e}) = \frac{1}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} p(w_j|e_i) \qquad (1)$$

where $l(m)$ is the length of entity(word) sequence. This is the model used in (Liu et al., 2007). We refer this model as **model-1** in the rest of this paper.

### 4.2 Model With Positional Alignment

Using the translation model II (Brown et al., 1993), where alignments are dependent on word/entity positions and word/entity sequence lengths, we have

$$p(\mathbf{w}|\mathbf{e}) = \prod_{j=1}^{m} \sum_{i=0}^{l} p(a_j = i|j, m, l) p(w_j|e_i) \qquad (2)$$

where $a_j = i$ means $w_j$ is aligned with $e_i$. We refer this model as **model-2**.

Compared to model-1, model-2 considers the ordering of words and entities in word acquisition.

### 4.3 Model With Temporal Alignment

In model-2, word-entity alignments are estimated from co-occurring word and entity sequences in an unsupervised way. The estimated alignments are dependent on where the words/entities appear in the word/entity sequences, not on when those words and gaze fixated entities actually occur. Motivated by the finding that users move their eyes to the mentioned object directly before speaking a word (Griffin and Bock, 2000), we make the word-entity alignments dependent on their temporal relation in a new model (referred as **model-1t**):

$$p(\mathbf{w}|\mathbf{e}) = \prod_{j=1}^{m} \sum_{i=0}^{l} p(a_j = i|j, \mathbf{e}, \mathbf{w}) p(w_j|e_i) \quad (3)$$

where $p(a_j = i|j, \mathbf{e}, \mathbf{w})$ is the temporal alignment probability computed based on the temporal distance between entity $e_i$ and word $w_j$.

Define temporal distance between $e_i$ and $w_j$ as

$$d(e_i, w_j) = \begin{cases} 0 & t_s(e_i) \le t_s(w_j) \le t_e(e_i) \\ t_e(e_i) - t_s(w_j) & t_s(w_j) > t_e(e_i) \\ t_s(e_i) - t_s(w_j) & t_s(w_j) < t_s(e_i) \end{cases} \quad (4)$$

where $t_s(w_j)$ is the starting timestamp of word $w_j$, $t_s(e_i)$ and $t_e(e_i)$ are the starting and ending timestamps of gaze fixation on entity $e_i$.

The alignment of word $w_j$ and entity $e_i$ is decided by their temporal distance $d(e_i, w_j)$. Based on the psycholinguistic finding that eye gaze happens before spoken word, $w_j$ is not allowed to be aligned with $e_i$ when $w_j$ happens earlier than $e_i$ (i.e., $d(e_i, w_j) > 0$). When $w_j$ happens no earlier than $e_i$ (i.e., $d(e_i, w_j) \le 0$), the closer they are, the more likely they are aligned. Specifically, the temporal alignment probability of $w_j$ and $e_i$ in each co-occurring instance $(\mathbf{w}, \mathbf{e})$ is computed as

$$p(a_j = i|j, \mathbf{e}, \mathbf{w}) = \begin{cases} 0 & d(e_i, w_j) > 0 \\ \frac{\exp[\alpha \cdot d(e_i, w_j)]}{\sum_i \exp[\alpha \cdot d(e_i, w_j)]} & d(e_i, w_j) \le 0 \end{cases} \quad (5)$$

where $\alpha$ is a constant.

EM algorithms are used to estimate the probabilities $p(w|e)$ in the translation models.

## 5 Evaluation

We evaluate word acquisition performance of different translation models on the data collected from user studies.

### 5.1 Evaluation Metrics

The following metrics of word acquisition are evaluated.

- Precision

$$P = \frac{\sum_e \text{\# words correctly acquired for entity } e}{\sum_e \text{\# words acquired for entity } e}$$

- Recall

$$R = \frac{\sum_e \text{\# words correctly acquired for entity } e}{\sum_e \text{\# groundtruth words of entity } e}$$

- F-measure

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

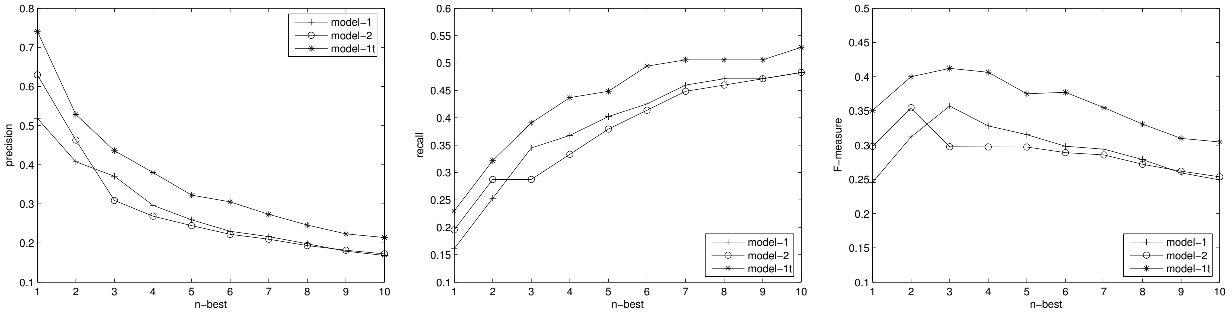
### 5.2 Evaluation Results

Figure 3 shows the precision, recall, and F-measure of word acquisition achieved different models. In the figure, *n-best* means the top *n* word candidates are chosen as acquired words for each entity.

The results show that model considering speech-gaze temporal relation (model-1t) achieves better performance than the models that do not consider this temporal information (model-1 and model-2). Compared to model-1, model-1t improves the F-measure by 15.4% $\sim$ 42.9% when different $n$-best word candidates are chosen. When compared to model-2, model-1t improves the F-measure by 12.8% $\sim$ 38.5%.

Compared to model-1, model-2 does not show a consistent improvement on the top $n$-best word candidates. This result shows that it is not very helpful to consider the positional alignment based on the index of word/entity in the word/entity sequence for word acquisition.

We also notice that the recall of the acquired words is not satisfying even when 10 best word candidates are chosen for each entity. This is mainly due



(a) precision (b) recall (c) F-measure

Figure 3: Performance of word acquisition

to the scarcity of those words that are not acquired in the data. Many of the words that are not acquired appear less than 3 times in the data, which makes them unlikely to be associated with any entity by the translation models. When more data is available, we expect to see a higher recall.

## 6 Conclusion

Motivated by the psycholinguistic findings, we investigate the use of eye gaze for automatic word acquisition in multimodal conversational systems. Particularly, we investigate the incorporation of temporal information about speech and eye gaze in word acquisition. Our experiments show that word acquisition performance is significantly improved when temporal information is considered, which is consistent with the previous psycholinguistic findings about speech and eye gaze.

## 7 Acknowledgments

This work was supported by IIS-0347548 and IIS-0535112 from the National Science Foundation. The authors would like to thank anonymous reviewers for their valuable comments and suggestions.

## References

Kobus Barnard, Pinar Duygulu, Nando de Freitas, David Forsyth, David Blei, and Michael I. Jordan. 2003. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135.

Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematic of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

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.

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

Y. Liu, J. Chai, and R. Jin. 2007. Automated vocabulary acquisition and interpretation in multimodal conversational systems. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*.

A. Mojsilovic. 2005. A computational model for color naming and describing color composition of images. *IEEE Trans. Image Processing*, 14:690–699.

D. Roy and A. Pentland. 2002. Learning words from sights and sounds, a computational model. *Cognitive Science*, 26(1):113–146.

J. Siskind. 2001. Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of Artificial Intelligence Research*, 15:31–90.

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.

Chen Yu and Dana H. Ballard. 2004. A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perceptions*, 1(1):57–80.