A Salience Driven Approach to Robust Input Interpretation in Multimodal Conversational Systems

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

Abstract

To improve the robustness in multimodal input interpretation, this paper presents a new salience driven approach. This approach is based on the observation that, during multimodal conversation, information from deictic gestures (e.g., point or circle) on a graphical display can signal a part of the physical world (i.e., representation of the domain and task) of the application which is salient during the communication. This salient part of the physical world will prime what users tend to communicate in speech and in turn can be used to constrain hypotheses for spoken language understanding, thus improving overall input interpretation. Our experimental results have indicated the potential of this approach in reducing word error rate and improving concept identification in multimodal conversation.

1 Introduction

Multimodal conversational systems promote more natural and effective human machine communication by allowing users to interact with systems through multiple modalities such as speech and gesture (Cohen et al., 1996; Johnston et al., 2002; Pieraccini et al., 2004). Despite recent advances, interpreting what users communicate to the system is still a significant challenge due to insufficient recognition (e.g., speech recognition) and understanding (e.g., language understanding) performance. Significant improvement in the robustness of multimodal interpretation is crucial if multimodal systems are to be effective and practical for real world applications.

Previous studies have shown that, in multimodal conversation, multiple modalities tend to complement each other (Cassell et al. 1994). Fusing two or more modalities can be an effective means of reducing recognition uncertainties, for example, through mutual disambiguation (Oviatt 1999). For semantically-rich modalities such as speech and pen-based gesture, mutual disambiguation usually happens at the fusion stage where partial semantic representations from individual modalities are disambiguated and combined into an overall interpretation (Johnston 1998, Chai et al., 2004a). One problem is that some critical but low probability information from individual modalities (e.g., recognized alternatives with low probabilities) may never reach the fusion stage. Therefore, this paper addresses how to use information from one modality (e.g., deictic gesture) to directly influence the semantic processing of another modality (e.g., spoken language understanding) even before the fusion stage.

In particular we present a new salience driven approach that uses gesture to influence spoken language understanding. This approach is based on the observation that, during multimodal conversation, information from deictic gestures (e.g., point or circle) on a graphical interface can signal a part of the physical world (i.e., representation of the domain and task) of the application which is salient during the communication. This salient part of the physical world will prime what users tend to communicate in speech and thus in turn can be used to constrain hypotheses for spoken language understanding. In particular, this approach incorporates a notion of salience from deictic gestures into language models for spoken language processing. Our experimental results indicate the potential of this approach in reducing word error rate and improving concept identification from spoken utterances.

Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 217–224, Vancouver, October 2005. ©2005 Association for Computational Linguistics

In the following sections, we first introduce the current architecture for multimodal interpretation. Then we describe our salience driven approach and present empirical results.

2 Input Interpretation

Input interpretation is the identification of semantic meanings in user inputs. In multimodal conversation, user inputs can come from multiple channels (e.g., speech and gesture). Thus, most work on input interpretation is based on semantic fusion that includes individual recognizers and a sequential integration processes as shown in Figure 1. In this approach, a system first creates possible partial meaning representations from recognized hypotheses (e.g., N-best lists) independently of other modalities. For example, suppose a user says "what is the price of this painting" and at the same time points to a position on the screen. The partial meaning representations from the speech input and the gesture input are shown in (a-b) in Figure 1. The system uses the partial meaning representations to disambiguate each other and combines compatible partial representations together into an overall semantic representation as in Figure1(c).

In this architecture, the partial semantic representations from individual modalities are crucial for mutual disambiguation during multimodal fusion. The quality of partial semantic representations depends on how individual modalities are processed. For example, if the speech input is recognized as "what is the prize of this pant", then the partial representation from the speech input will not be created in the first place. Without a candidate partial representation, it is not likely for multimodal fusion to reach an overall meaning of the input given this late fusion architecture.

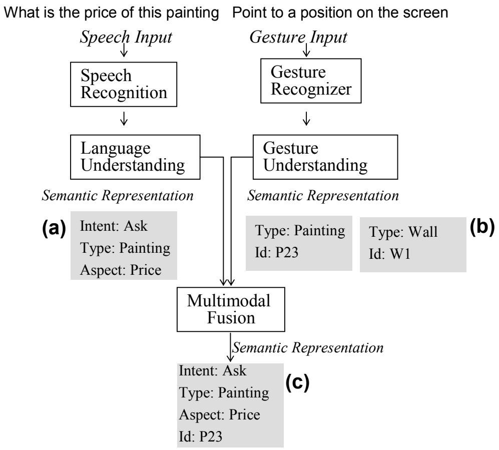

Figure 1: Semantics-based multimodal interpretation

Thus, a problem with the semantics-based fusion approach is that information from multiple modalities is only used during the fusion stage to disambiguate or combine partial semantic representations. This late use of information from other sources in the pipelined process can cause the loss of some low probability information (e.g., recognized alternatives with low probabilities which did not make it to the N-best list) which could be very crucial in terms of the overall interpretation. It is desirable to use information from multiple sources at an earlier stage before partial representations are created from individual modalities. For example, in ((Bangalore and Johnston 2000), a finite-state approach was applied to tightly couple multimodal language processing (e.g., gesture and speech) and speech recognition to improve recognition hypotheses. To further address this issue, in this paper, we present a salience driven approach that particularly applies gesture information (e.g., pen-based deictic gestures) to robust spoken language understanding before multimodal fusion.

3 Related Work on Salience Modeling

We first give a brief overview on the notion of salience and how salience modeling is applied in earlier work on natural language and multimodal language processing.

Linguistic salience describes the accessibility of entities in a speaker/hearer's memory and its implication in language production and interpretation. Many theories on linguistic salience have been developed, including how the salience of entities affects the form of referring expressions as in the Givenness Hierarchy (Gundel et al., 1993) and the local coherence of discourse as in the Centering Theory (Grosz et al., 1995). Salience modeling is used for both language generation and language interpretation; the latter is more relevant to our work. Most salience-based interpretation has focused on reference resolution for both linguistic referring expressions (e.g., pronouns) (Lappin and Leass 1995) and multimodal expressions (Hul et al. 1995; Eisenstein and Christoudias 2004).

Visual salience considers an object salient when it attracts a user's visual attention more than others. The cause of such attention depends on many factors including user intention, familiarity, and physical characteristics of objects. For example, an object may be salient when it has some properties the others do not have, such as it is the only one that is highlighted, or the only one of a certain size, category, or color

(Landragin et al., 2001). Visual salience can also be useful in input interpretation, for example, for multimodal reference resolution (Kehler 2000) and cross-modal coreference interpretation (Byron et al., 2005).

We believe that salience modeling should go beyond reference resolution. Our view is that the salience not only affects the use of referring expressions (and thus is useful for interpreting referring expressions), but also influences the linguistic context of the referring expressions. The spoken utterances that contain these expressions tend to describe information relating to the salient objects (e.g., properties or actions). Therefore, our goal in this paper is to take salience modeling a step further from reference resolution, towards overall language understanding.

4 A Salience Driven Approach

The new salience driven approach is based on the cognitive theory of Conversation Implicature (Grice 1975) and earlier empirical findings of user speech and gesture behavior in multimodal conversation (Oviatt 1999). The theory of Conversation Implicature (Grice 1975) states that speakers tend to make their contribution as informative as is required (for the current purpose of communication) and not make their contribution more informative than is required. In the context of multimodal conversation that involves speech and pen-based gesture, this theory indicates that users most likely will not make any unnecessary deictic gestures unless those gestures help in communicating users' intention. This is especially true since gestures usually take an extra effort from a user. When a pen-based gesture is intentionally delivered by a user, the information conveyed is often a crucial component in interpretation (Chai et al., 2005).

Speech and gesture also tend to complement each other. For example, when a speech utterance is accompanied by a deictic gesture (e.g., point or circle) on a graphical display, the speech input tends to issue commands or inquiries about properties of objects, and the deictic gestures tend to indicate the objects of interest. In addition, as shown in (Oviatt 1999), the deictic gestures often occur before spoken utterances. Our previous work (Chai et al., 2004b) also showed that 85% of time gestures occurred before corresponding speech units. Therefore, gestures can be used as an earlier indicator to anticipate the content of communication in the subsequent spoken utterances.

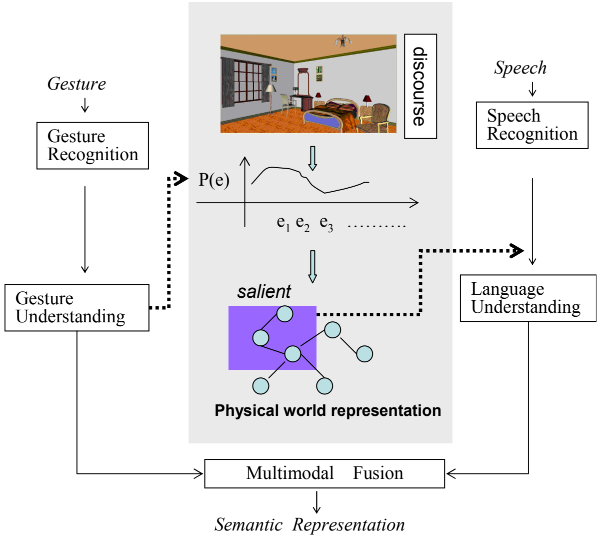

Figure 2: The salience driven approach: the salience distribution calculated from gesture is used to tailor language models for spoken language understanding

4.1 Overview

The general idea of the salience based approach is shown in Figure 2. For each application domain, there is a physical world representation that captures domain knowledge (details are described later). A deictic gesture can activate several objects on the graphical display. This activation will signal a distribution of objects that are salient. The salient objects are mapped to the physical world representation to indicate a salient part of representation that includes relevant properties or tasks related to the salient objects. This salient part of the physical world is likely to be the potential content of the spoken communication, and thus can be used to tailor language models for spoken language understanding. This process is shown in the middle shaded box of Figure 2. It bridges gesture understanding and language understanding at a stage before multimodal fusion. Note that the use of gesture information can be applied at different stages: during speech recognition to generate hypotheses or post processing of recognized hypotheses before language understanding. In this paper, we focus on the latter.

The physical world representation includes the following components:

- Domain Model. This component captures the relevant knowledge about the domain including domain objects, properties of the objects, relations between objects, and task models related to objects. Previous studies have shown that domain knowledge can be used to improve spoken language understanding (Wai et al, 2001). Currently, we apply a frame-based representation where a frame represents an object (or a type of object) in the domain and frame elements represent attributes and tasks related to the objects. Each frame element is associated with a semantic tag which indicates the semantic content of that element. In the future, the domain model might also include knowledge about the interface, for example, visual properties and spatial relations between objects on the interface.

- <u>Domain Grammar</u>. This component specifies grammar and vocabularies used to process language inputs. There are two types of representation. The first type is a semantics-based context free grammar where each non-terminal symbol represents a semantic tag (indicating semantic information such as the semantic type of an object, etc). Each word (i.e., the terminal symbol) in the lexicon relates to one or more semantic tags. Some of these semantic tags are directly linked to the frame elements in the domain model since they represent certain properties or tasks. This grammar was manually developed.

The second type of representation is based on annotated user spoken utterances. The data are annotated in terms of relevant semantic information (i.e., using semantic tags) in the utterance and the intended objects of interest (which are directly linked to the domain model). Based on the annotated data, N-grams can be learned to represent the dependency of language in our domain.

Based on the physical world representation, our approach supports the following operations:

<u>Salience modeling</u>. This operation calculates a salience distribution of entities in the physical world. In our current investigation, we limit the scope of entities to a closed set of objects from our physical world representation since the system has knowledge about those objects. These entities could have different salience values depending on whether they are visible on the graphical display, gestured by a user, or mentioned in the prior conversation. In this paper, we focus on the salience modeling using gesture information only.

<u>Salience driven language understanding</u>. This operation maps the salience distribution to the physical world representation and uses the salient world to influence spoken language understanding. Note that, in this paper, we are not concerned with acoustic models for speech recognition, but rather we are interested in the use of the salience distribution to prime language models and facilitate language understanding.

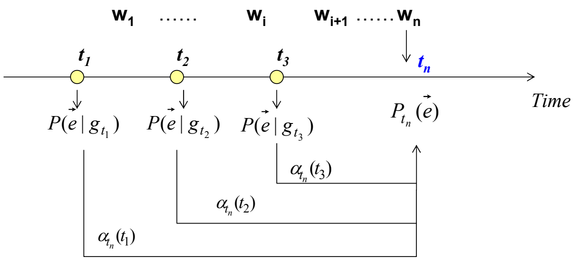

Figure 3: Salience modeling: the salience distribution at time t_n is calculated by a joint effect of gestures that happen before t_n.

4.2 Salience Modeling

We use a vector $\vec{e}$ to represent entities in the physical world representation. For each entity $e_k \in \vec{e}$, we use $P_{t_n}(e_k)$ to represent its salience value at time t_n. For all the entities, we use $P_{t_n}(\vec{e})$ to represent a salience distribution at time t_n. Figure 3 shows a sequence of words with corresponding gestures that occur at t_1, t_2, and t_3. As shown in Figure 3, the salience distribution at any given time t_n is influenced by a joint effect from this sequence of gestures that happen before t_n etc. Depending on its time of occurrence, each gesture may have a different impact on the salience distribution at time t_n. Note that although each gesture may have a short duration, here we only consider the beginning time of a gesture. Therefore, for an entity e_k, its salience value at time t_n is computed as follows:

$$P_{t_n}(e_k) = \frac{\sum_{i=1}^{m} \alpha_{t_n}(g_{t_i}) P(e_k | g_{t_i})}{\sum_{e \in \vec{e}} \sum_{i=1}^{m} \alpha_{t_n}(g_{t_i}) P(e | g_{t_i})} \qquad (1)$$

In Equation (1), m ($m \geq 1$) is the number of gestures that have occurred before t_n. The different impact of a gesture g_{t_i} at time t_i that contributes to the salience distribution at time t_n is represented as the weight $\alpha_{t_n}(g_{t_i})$ in Equation (1). Currently, we calculate the weight depending on the temporal distance as follows:

$$\alpha_{t_n}(g_{t_i}) = \exp[\frac{-(t_n - t_i)}{2000}] \quad (t_n \geq t_i) \qquad (2)$$

Equation (2) indicates that at a given time t_n (measured in milliseconds), the closer a gesture (at t_i) is to the time t_n, the higher impact this gesture has on the salience distribution (Chai et al., 2004b).

It is worth mentioning that a deictic gesture on the graphic display (e.g., pointing and circling) could have ambiguous interpretation by itself. For example, given an interface, a point or a circle on the screen could result in selection of different entities with different probabilities. Therefore, in Equation (1), $P(e \mid g_{t_i})$ is the selection probability which indicates the likelihood of selecting an entity *e* given a gesture at time t_i. This selection probability is calculated by a function of the distance between the location of the entity and the focus point of the recognized gesture on the display (Chai et al., 2004a). A normalization factor is incorporated to ensure that the summation of selection probabilities over all possible entities adds up to one.

When no gesture is involved in a given input, the salience distribution at any given time is a uniform distribution. If one or more gestures are involved, then Equation (1) is used to calculate the salience distribution.

4.3 Salience Driven Spoken Language Understanding

The salience distribution of entities identified based on the gesture information (as described above) is used to constrain hypotheses for language understanding. More specifically, for each onset of a spoken word at time *t* (i.e., the beginning time stamp of a spoken word), the salience distribution at *t* can be calculated based on a sequence of gestures that happen before *t* by Equation (1). This salience distribution can then be used to prime language models for spoken language processing.

Language Modeling

We first give a brief background of language modeling. Given an observed speech utterance O, the goal of speech recognition is to find a sequence of words W* so that $W^* = \arg\max P(O \mid W)P(W)$, where *P(O|W)* is the acoustic model and *P(W)* is the language model. In traditional speech recognition systems, the acoustic model provides the probability of observing the acoustic features given hypothesized word sequences and the language model provides the probability of a sequence of words. The language model is computed as follows:

$$P(w_1^n) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1 w_2)...P(w_n \mid w_1^{n-1})$$

Using the Markov assumption, the language model can be approximated by a bigram model as in:

$$P(w_1^n) = \prod_{i=1}^{n} P(w_i \mid w_{i-1})$$

To improve the speech understanding results for spoken language interfaces, many systems have applied a loosely-integrated approach which decouples the language model from the acoustic model (Zue et al., 1991, Harper et al., 2000). This allows the development of powerful language models independent of the acoustic model, for example, utilizing topics of the utterances (Gildea and Hofmann 1999), syntactic or semantic labels (Heeman 1999), and linguistic structures (Chelba and Jelinek 2000, Wang and Harper 2002). Recently, we have seen work on language understanding based on environment (Schuler 2003) and language modeling using visual context (Roy and Mukherjee 2005). Our salience driven approach is inspired by this earlier work. Here, we do not address the acoustic model of speech recognition, but rather incorporate the salience distribution for language modeling. In particular, our focus is on investigating the effect of incorporating additional information from other modalities (e.g., gesture) with traditional language models.

Primed Language Model

The calculated salience distribution is used to prime the language model. More specifically, we use a class-based bigram model from (Brown et al, 1992):

$$P(w_i \mid w_{i-1}) = P(w_i \mid c_i)P(c_i \mid c_{i-1}) \quad (3)$$

In Equation (3), c_i is the class of the word w_i, which could be a syntactic class or a semantic class. $P(c_i \mid c_{i-1})$ is the class transition probability, which reflects the grammatical formation of utterances. $P(w_i \mid c_i)$ is the word class probability which measures the probability of seeing a word w_i given a class c_i. The class-based N-gram model can make better use of limited training data by clustering words into classes. A number of researchers have shown that the class-based N-gram model can successfully improve the performance of speech recognition (Jelinek 1990, Heeman 1999, Kneser and Ney 1993, Samuelsson and Reichl, 1999).

In our approach, the "class" used in the class-based bigram model comes from combined semantic and functional classes designed for our domain. For example, "this" is tagged as Demonstrative, and "price" is tagged as AttrPrice. As shown in Equation (3), there are two types of parameter estimation. In terms of the class transition probability, as in earlier work, we directly use the annotated data. In terms of the word class distribution, we incorporate the notion of salience. We use the salience distribution to dynamically adjust the world class probability $P(w_i \mid c_i)$ as follows:

$$P(w_i \mid c_i) = \sum_{e_k \in \vec{e}} \frac{P(w_i, c_i \mid e_k)}{P(c_i \mid e_k)} P_{t_i}(e_k) \qquad (4)$$

In Equation (4), $P_{t_i}(e_k)$ is the salience value for an entity e_k at time t_i (the onset of the spoken word w_i), which can be calculated by Equation (1). Equation (4) indicates that only information associated with the salient entities is used to estimate the word class distribution. In other words, the word class probability favors the salient physical world as indicated by the salience distribution $P_{t_i}(\vec{e})$. More specifically, at time t_i, given a semantic class c_i, the choice of word "w_i" is dependent on the salient physical world at the moment, which is represented as the salience distribution $P_{t_i}(\vec{e})$ at time t_i. For all w_i, the summation of this word class probability is equal to one. Furthermore, given an entity e_k, $P(w_i, c_i \mid e_k)$ and $P(c_i \mid e_k)$ are not dependent on time t_i, but rather on the domain and the use of language expressions. Therefore they can be estimated based on the training data that are annotated in terms of semantic information and the intended objects of interest (as discussed in Section 4.1). Since the annotated data is very limited, the sparse data can become a problem for the maximum likelihood estimation. Therefore, a smoothing technique based on the Katz backoff model (Katz, 1987) is applied. For example, to calculate $P(w_i, c_i \mid e_k)$, if a word w_i has one or more occurrences in the training data associated with the class c_i and the entity e_k, then its count is discounted by a fraction in the maximum likelihood estimation. If w_i does not occur, then this approach backs off to the domain grammar and redistributes the remaining probability mass uniformly among words in the lexicon that are linked with class c_i and entity e_k.

5 Evaluation

We evaluated the salience model during post processing recognized hypotheses. Given possible hypotheses from a speech recognizer, we use the salience-based language model to identify the most likely sequence of words. The salience distribution based on gesture was used to favor words that are consistent with the attention indicated by gestures.

The data collected from our previous user studies were used in our evaluation (Chai et al., 2004b). In these studies, users interacted with our multimodal interface using both speech and deictic gestures to find information about real estate properties. In particular, each user was asked to accomplish five tasks. Each of these tasks required the user to retrieve different types of information from our interface. For example, one task was to find the least expensive house in the most populated town. The data were recorded from eleven subjects including five non-native speakers and six native speakers. Each user's voice was individually trained before the study. Table 1 shows the relevant information about the data such as the total number of inputs (or turns) from each subject, the number of speech alone inputs without any gesture, and the baseline recognition results without using salience-based post processing in terms of the word error rate (WER). In total, we have collected 226 user inputs with an average of eight words per spoken utterance[1]. As shown in Table 1, the majority of inputs consisted of both speech and gesture. Since currently we only use gesture

User index	# of Inputs	# inputs w/o gesture	Baseline WER
1	21	0	0.287
2	31	0	0.335
3	27	0	0.399
4	10	0	0.680
5	8	1	0.200
6	36	0	0.387
7	18	0	0.250
8	25	1	0.278
9	23	0	0.482
10	11	0	0.117
11	16	3	0.255

Table 1: Related information about the evaluation data: user type, the number of turns per user, and the baseline word recognition rate.

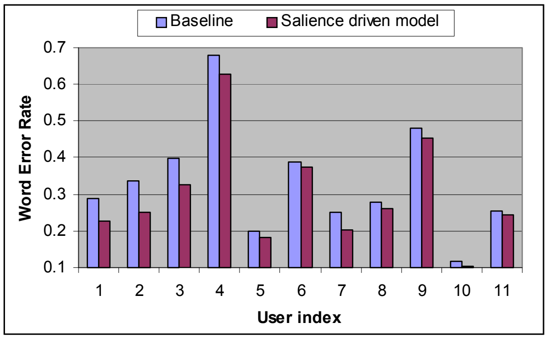

Figure 5: Comparison of the baseline and the result from post-processing in terms of WER

[1] The difference between the number of user inputs reported here and that in (Chai et al., 2004b) was caused by the situation where one intended user input (which was the unit for counting in our previous work) was split into a couple turns (which constitute the new counts here).

information in salience modeling, our approach will not affect speech only inputs.

To train the salience-based model, we applied the leave-one-out approach. The data from each user was held out as the testing data and the remaining users were used as the training data to acquire relevant probability estimations in Equation (3) and (4).

Figure 5 shows the comparison results between the baseline and the salience-based model in terms of word error rate (WER). The word error rate as a result of salience-based post processing is significantly better than that from the baseline recognizer (t = 4.75, p < 0.001). The average WER reduction is about 12%.

We further evaluated how the salience based model affects the final understanding results. This is because an improvement in WER may not directly lead to an improvement in understanding. We applied our semantic grammar on a sequence of words resulting from both the baseline and the salience-based post-processing to identify key concepts. In total, there were 686 concepts from the transcribed speech utterances. Table 2 shows the evaluation results. Precision measures the percentage of correctly identified concepts out of the total number of concepts identified based on a sequence of words. Recall measures the percentage of correctly identified concepts out of the total number of intended concepts from user's utterance. F-measurement combines precision and recall together as follows:

$$F = \frac{(\beta^2 + 1) \times \text{Precision} \times \text{Recall}}{\beta^2 \text{Precision} + \text{Recall}}, \quad where\, \beta = 1 .$$

Table 2 shows that, on average, the concept identification based on the word sequence resulting from the salience-based approach performs better than the baseline in terms of both precision and recall. Figure 6 provides two examples to show the difference between the baseline recognition and the salience-based post processing.

User #	Baseline	Salience-based
Precision	80.3%	84.6%
Recall	75.7%	83.8%
F-measure	77.9%	84.2%

Table2. Overall concept identification comparison between the baseline and the salience driven model.

Example 1:
Transcription: What is the population of this town
Baseline recognition: What is the publisher of this time
Salience-based processing: what is the **population** of this **town**

Example 2:
Transcription: How much is this gray house
Baseline recognition: How much is this great house
Salience-based processing: How much is this **gray** house

Figure 6: Examples of utterances with baseline recognition and improved recognition from the salience-based processing.

The evaluation reported here is only an initial step based on a limited domain. The small scale in the number of objects and the vocabulary size can only demonstrate the potential of the salience-based approach to a limited degree. To further understand the advantages and issues of this approach, we are currently working on a more complex domain with richer concepts and relations, as well as larger vocabularies.

It is worth mentioning that the goal of this work is to explore whether salience modeling based on other modalities (e.g., gesture) can be used to prime traditional language models to facilitate spoken language processing. The salience driven approach based on additional modalities can be combined with more sophisticated language modeling (e.g., better parameter estimation) in the future.

6 Conclusion and Future Work

This paper presents a new salience driven approach to robust input interpretation in multimodal conversational systems. This approach takes advantage of rich information from multiple modalities. Information from deictic gestures is used to identify a part of the physical world that is salient at a given point of communication. This salient part of the physical world is then used to prime language models for spoken language understanding. Our experimental results have shown the potential of this approach in reducing word error rate and improving concept identification from spoken utterances in our application. Although currently we have only investigated the use of gesture information in salience modeling, the salience driven approach can be extended to include other modalities (e.g., eye gaze) and information (e.g., conversation context). Our future work will specifically investigate how to combine information from multiple sources in salience modeling and how to apply the salience models in different early stages of processing.

Acknowledgement

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

References

Bangalore, S. and Johnston, M. 2000. Integrating Multimodal Language Processing with Speech Recognition. In *Proceedings of ICSLP*.

Brown, P., Della Pietra, V. J., deSouza, P. V., Lai, J. C, and Mercer, R. L. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467-479.

Byron, D., Mampilly, T., Sharma, V., and Xu, T. 2005. Utilizing Visual Attention for Cross-Modal Coreference Interpretation. *Spring Lecture Notes in Computer Science: Proceedings of Context-05*, page 83-96.

Cassell, J., Stone, M., Douville, B., Prevost, S., Achorn, B., Steedman, M., Badler, N., and Pelachaud, C. 1994. Modeling the interaction between speech and gesture. *Cognitive Science Society*.

Chai, J. Y., Prasov, Z., Blaim, J., and Jin, R. 2005. Linguistic Theories in Efficient Multimodal Reference Resolution: an Empirical Investigation. *The 10th International Conference on Intelligent User Interfaces (IUI-05),* pp. 43-50, San Diego, CA.

Chai, J. Y., Hong, P., Zhou, M. X, and Prasov, Z. 2004b. Optimization in Multimodal Interpretation. *In Proceedings of ACL*, pp. 1-8, Barcelona, Spain.

Chai, J. Y., Hong, P., and Zhou, M. 2004a. A Probabilistic Approach to Reference Resolution in Multimodal User Interfaces. *Proceedings of 9th International Conference on Intelligent User Interfaces (IUI-04)*, pp. 70-77, Madeira, Portugal.

Chelba, C. and Jelinek, F. 2000. Structured language modeling. *Computer Speech and Language*, 14(4):283–332.

Cohen, P., Johnston, M., McGee, D., Oviatt, S., Pittman, J.; Smith, I., Chen, L., and Clow, J. 1996. Quickset: Multimodal Interaction for Distributed Applications. *Proceedings of ACM Multimedia*, 31– 40.

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

Gildea, D. and Hofmann, T. 1999. Topic-based language models using EM. In *Proceedings of Eurospeech*.

Griffin, Z. M. 2001. Gaze durations during speech reflect word selection and phonological encoding. *Cognition* 82, B1-B14.

Grosz, B. J., Joshi, A. K., and Weinstein, S. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2).

Grice, H. P. Logic and Conversation. 1975. In Cole, P., and Morgan, J., eds. Speech Acts. New York, New York: Academic Press. 41-58.

Gundel, J. K., Hedberg, N., and Zacharski, R. 1993. Cognitive Status and the Form of Referring Expressions in Discourse. *Language* 69(2):274-307.

Harper, M.., White, C., Wang, W., Johnson, M., and Helzerman, R. 2000. The Effectiveness of Corpus-Induced Dependency Grammars for Post-processing Speech. *Proceedings of the North American Association for Computational Linguistics,* 102-109.

Heeman. P. 1999. POS tags and decision trees for language modeling. In *Proceedings of the Conference on Empirical Methods in Natural Language Process (EMNLP).*

Huls, C., Bos, E., and Classen, W. 1995. Automatic Referent Resolution of Deictic and Anaphoric Expressions. *Computational Linguistics*, 21(1):59-79.

Jelinek, F. 1990. Self-organized language modeling for speech recognition. In Waibel, A. and Lee, K. F. (Eds). *Readings in Speech Recognition*, pp. 450-506.

Johnston, M. 1998. Unification-based Multimodal parsing, *Proceedings of COLING-ACL*.

Johnston, M., Bangalore, S., Visireddy G., Stent, A., Ehlen, P., Walker, M., Whittaker, S., and Maloor, P. 2002. MATCH: An Architecture for Multimodal Dialog Systems, in *Proceedings of the 40th ACL*, Philadelphia, pp. 376-383.

Katz, S. M. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3).

Kehler, A. 2000. Cognitive Status and Form of Reference in Multimodal Human-Computer Interaction, *Proceedings of AAAI'01*.

Kneser, R. and Ney, H. 1993. Improved clustering techniques for class-based statistical language modeling. In *Eurospeech'93*, pp. 973-976.

Landragin, F., Bellalem, N., and Romary, L. 2001. Visual Salience and Perceptual Grouping in Multimodal Interactivity. In: *First International Workshop on Information Presentation and Natural Multimodal Dialogue*, Verona, Italy, pp. 151-155.

Lappin, S., and Leass, H. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4).

Oviatt, S. 1999. Mutual Disambiguation of Recognition Errors in a Multimodal Architecture. In *Proceedings of CHI.*

Pieraccini, R., Dayandhi, K., Bloom, J., Dahan, J.-G., Phillips, M., Goodman, B. R., Prasad, K. V., 2004. Multimodal Conversational Systems for Automobiles, *Communications of the ACM*, Vol. 47, No. 1, pp. 47-49

Roy, D. and Mukherjee, N. 2005. Towards Situated Speech Understanding: Visual Context Priming of Language Models. *Computer Speech and Language*, 19(2): 227-248.

Samuelsson, C. and Reichl, W. 1999. A class-based Language Model for Large Vocabulary Speech Recognition Extracted from Part-of-Speech Statistics. In *IEEE ICASSP'99*.

Schuler, W. 2003. Using model-theoretic semantic interpretation to guide statistical parsing and word recognition in a spoken language interface. *In Proceedings of ACL*, Sapporo, Japan.

Wai, C., Pierraccinni, R., and Meng, H. 2001. A Dynamic Semantic Model for Rescoring Recognition Hypothesis. *Proceedings of the ICASSP*.

Wang, W. and Harper. M. 2002. The superARV language model: In Investigating the effectiveness of tightly integrating multiple knowledge sources. In *Proceedings of EMNLP*, 238–247.

Zue, V., Glass, J., Goodine, D., Leung, H., Phillips, M., Polifroni, J., and Seneff, S. 1991. Integration of Speech Recognition and Natural Language Processing in the MIT Voyager System. *Proceedings of the ICASSP*.