Salience Modeling based on Non-Verbal Modalities for Spoken Language Understanding

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

ABSTRACT

Previous studies have shown that, in multimodal conversational systems, fusing information from multiple modalities together can improve the overall input interpretation through mutual disambiguation. Inspired by these findings, this paper investigates non-verbal modalities, in particular deictic gesture, in spoken language processing. Our assumption is that during multimodal conversation, user's deictic gestures on the graphic display can signal the underlying domain model that is salient at that particular point of interaction. This salient domain model can be used to constrain hypotheses for spoken language processing. Based on this assumption, this paper examines different configurations of salience driven language models (e.g., n-gram and probabilistic context free grammar) for spoken language processing across different stages. Our empirical results have shown the potential of integrating salience models based on non-verbal modalities in spoken language understanding.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Theory and methods, Natural language

General Terms

Algorithms, Design, Experimentation

Keywords

Multimodal Interfaces, Salience Modeling, Language Modeling, Spoken Language Understanding

1. INTRODUCTION

Multimodal interfaces allow users to interact with systems through multiple modalities such as speech, gesture, and eye gaze. These types of systems can promote more natural human machine interaction [3, 8], and they can cope with the limitations of the speech technology in spoken language interfaces [13]. Recent studies have shown that, in multimodal conversational systems, fusing information from multiple modalities together can improve the overall input interpretation through mutual disambiguation [14]. Inspired by earlier work, our research investigates how non-verbal modalities can be used to facilitate spoken language processing in a multimodal conversational system.

Non-verbal modalities can provide important information about a user's intent. For example, a deictic gesture on the graphic display usually indicates a user's attention. Based on this observation, earlier work has incorporated deictic gestures to resolve speech referring expressions (e.g., using gesture information to resolve what *this* refers to in the utterance "*how much does this cost*") [10, 4]. In our view, a deictic gesture not only indicates attention, but also activates the underlying domain model that is associated with the selected objects. This domain model contributes to *domain context* that is relevant to spoken language communication, and thus can be used to enable *context-aware* language processing. Therefore, our goal is to go beyond reference resolution and use non-verbal modalities such as deictic gestures to facilitate overall spoken language understanding.

Towards this direction, this paper presents an empirical investigation that incorporates deictic gestures for spoken language processing across different stages. Our assumption is that during multimodal conversation, a user's deictic gestures on the graphic display can signal the domain context that is salient at that particular point of interaction. This salient domain context can be used to tailor language modeling and constrain speech hypotheses. Based on this assumption, this paper examines different configurations of salience driven language models using n-gram and probabilistic context free grammar. Our empirical results have shown the potential of integrating salience models based on deictic gestures in spoken language understanding.

In the following sections, we first give a brief introduction of our multimodal conversational system, then describe the salience-driven language modeling in detail, and finally present results from empirical evaluations.

2. RELATED WORK

The work reported here is inspired by previous research in multimodal fusion, context-aware language processing, and salience modeling.

2.1 Multimodal Fusion

Studies have shown that multimodal systems can achieve more robust input interpretation due to mutual disambigua-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ICMI'06, November 2–4, 2006, Banff, Alberta, Canada.
Copyright 2006 ACM 1-59593-541-X/06/0011 ...$5.00.

tion of multiple modalities [14]. In most multimodal systems, input interpretation is based on a semantic fusion approach. The system first creates all possible partial meaning representations individually for each modality, and then uses these partial meaning representations to disambiguate each other and to form a complete semantic representation. For example, a finite state approach [8], a unification-based approach [7], and a graph-matching approach [2] have been developed to fuse the gestural and speech inputs.

In this semantics-based fusion approach, information from multiple modalities is only used at the fusion stage. One potential problem is that some low probability information (e.g., recognized alternatives with low probabilities) that could turn out to be very crucial in terms of the overall interpretation may never reach the fusion stage. Therefore, it is desirable to use information from multiple sources at an earlier stage, for example, using one modality to facilitate semantic processing of another modality. Therefore, this paper focuses on how to use contextual information indicated by deictic gestures to facilitate speech recognition before semantic processing of the recognized results.

2.2 Context-aware Language Processing

Several previous attempts have been made to improve spoken language processing by taking into account the context of communication. Different types of context, such as conversation context and visual context, have been used in different applications.

In [17], the system first detects the type of a task-oriented dialog based on intonation. It then chooses a bigram language model specific to the detected dialog type to process speech input. The results have shown that using the language model specifically targeted to a particular context (in this case dialog type) achieves higher word accuracy than a basic language model. In a visual scene description domain [15], the visual features of objects extracted by image analysis, such as color and shape, are used to tailor the language model for recognizing users' utterances describing objects in a visual scene. The results have shown that the incorporation of visual context into speech recognition reduces both word error rate and language understanding error rate.

To address the context-aware language processing, this paper focuses on incorporating domain context into speech processing. The domain context is signaled by users' deictic gestures through salience modeling during communication.

2.3 Salience Modeling

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 both language generation [16] and language interpretation. Most salience-based language interpretations have focused on reference resolution [12, 5, 6].

Visual salience measures how much attention an entity attracts from a user. An entity is more salient when it attracts a user's attention more than other entities. Visual salience can also be useful in multimodal language interpretation. Studies have shown that the user's perceived salience of entities on the graphical interfaces can tailor user's referring expressions and thus can be used for multimodal reference resolution [10].

A salience-driven language model based on gestures for post-processing speech inputs was developed in [3], where salience modeling was incorporated into a class-based bigram model. Extending this earlier work, in this paper, we systematically investigate a range of salience driven language models including n-grams and probabilistic context free grammar at different stages of processing. This systematic investigation will provide insight on the potential and tradeoff of salience modeling using non-verbal modalities for spoken language processing.

3. A MULTIMODAL CONVERSATIONAL SYSTEM

Our multimodal conversational system allows users to interact with the system through both speech and deictic gesture. The system was built on a client/server architecture. Users interact with the client while the server processes users' inputs and gives responses. The system consists of the following components:

- **Speech component** – speech recording, recognition and synthesis
- **Gesture component** – gesture analysis and recording
- **Graphics component** – display and manipulation of 3D visual scenes (with Microsoft Direct 3D)
- **Network component** – communication between client and server (via TCP/IP)
- **Log component** – logging the interactions between system and user
- **Main component** – consisting of three modules: multimodal interpreter, dialog manager, and presentation manager. The multimodal interpreter identifies semantic meaning from the user's multimodal inputs. The Dialog manager controls the interaction flow and decides what the system should do next based on the interpretation of user inputs. The presentation manager is responsible for presenting the system's responses.

Figure 1: A 3D bedroom domain

The domain we are working on is a 3D bedroom decoration domain as shown in Fig.1. There are 13 types of entities (3D objects) in the bedroom scene. Users can interact with the system using both speech and deictic gestures to query information about the entities or arrange the room by adding, removing, moving, and coloring the entities. For example, the user may say "*remove this lamp*" or ask "*what's the power of this lamp?*" while pointing at a lamp in the scene.

Based on this system and the domain, this paper specifically investigates the role of deictic gestures in speech processing.

4. SALIENCE DRIVEN LANGUAGE MODELING

In this section, we describe the use of salience driven language models in processing spoken language. First we give a brief review of speech recognition.

The task of speech recognition is to, given an observed spoken utterance O, find the word sequence W^* such that

$$W^* = \arg\max_W p(O|W)p(W) \tag{1}$$

where $p(O|W)$ is the acoustic model and $p(W)$ is the language model.

In speech recognition systems, the acoustic model provides 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 represented as follows:

$$p(W) = p(w_1^n) = p(w_1)p(w_2|w_1)p(w_3|w_1^2)\dots p(w_n|w_1^{n-1}) \tag{2}$$

The language model can be approximated by a bigram model using first-order Markov assumption:

$$p(w_1^n) = \prod_{k=1}^{n} p(w_k|w_{k-1}) \tag{3}$$

or by a trigram model using second-order Markov assumption:

$$p(w_1^n) = \prod_{k=1}^{n} p(w_k|w_{k-1}, w_{k-2}) \tag{4}$$

By clustering words into classes, the class-based n-gram model reduces the training data requirement and improves the robustness of probability estimates compared to word n-gram model. The class-based bigram model is given by [1]:

$$p(w_i|w_{i-1}) = p(w_i|c_i)p(c_i|c_{i-1}) \tag{5}$$

where c_i and c_{i-1} are the classes of word w_i and w_{i-1} respectively.

Next we first introduce the gesture-based salience modeling, then present different salience driven language models based on these basic models described above and PCFG.

4.1 Gesture-based Salience Modeling

As mentioned earlier, a deictic gesture on the graphical display can signal the underlying domain modal that is salient at that particular point of communication. In other words, the deictic gesture will activate a salience distribution for entities represented in the domain model. More specifically, for each entity e in the domain, a gesture g at time t can activate its salience value as the following:

$$p(e) = \begin{cases} \frac{\sum_g \alpha_g(t)p(e|g)}{\sum_{e,g} \alpha_g(t)p(e|g)} & \sum_{e,g} \alpha_g(t)p(e|g) \neq 0 \\ 0 & \sum_{e,g} \alpha_g(t)p(e|g) = 0 \end{cases} \tag{6}$$

where $p(e|g)$ is the probability of entity e being selected by gesture g (calculated based on the distance from the gesture point to the center of the entity), $\alpha_g(t)$ is the weight of gesture g contributing to the salience distribution at time t and is defined as follows:

$$\alpha_g(t) = \begin{cases} e^{-\frac{t-t_g}{2000}} & t \geq t_g \\ 0 & t < t_g \end{cases} \tag{7}$$

In Equation (7), t_g stands for the beginning time (in milliseconds) of gesture g. Weight $\alpha_g(t)$ says that gesture g has more impact on the salience distribution at a time closer to the gesture's occurrence. Note that given time t, we only consider gestures that occur before that time (i.e., $t \geq t_g$).

Using the gesture-based salience modeling, we can tailor a language model to make it favor the domain context indicated by the salience model. Next, we describe different ways of incorporating salience modeling in language models.

4.2 Salience Driven N-gram Models

4.2.1 Salience Driven Bigram Model

The salience driven bigram probability $p_s(w_i|w_{i-1})$ is given by:

$$p_s(w_i|w_{i-1}) = \frac{p(w_i|w_{i-1}) + \lambda \sum_e p(w_i|w_{i-1}, e)p(e)}{1+\lambda} \tag{8}$$

where $p(e)$ is the salience distribution, λ is the priming weight. The priming weight λ decides how much the original bigram probability will be tailored by the salient entities that are indicated by gestures. The priming weight was decided by regression test in our experiments. Bigram probabilities $p(w_i|w_{i-1})$ were estimated by the maximum likelihood estimation using Katz's backoff method [9] with 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.

4.2.2 Salience Driven Class-based Bigram Model

Following the idea in [3], the salience driven bigram probability $p_s(w_i|w_{i-1})$ is given by:

$$p_s(w_i|w_{i-1}) = \begin{cases} p(c_i|c_{i-1}) \sum_e p(w_i|c_i, e)p(e) & \sum_e p(e) \neq 0 \\ p(w_i|w_{i-1}) & \sum_e p(e) = 0 \end{cases} \tag{9}$$

where $p(e)$ is the salience distribution, c_i and c_{i-1} are the semantic classes of word w_i and w_{i-1} respectively.

4.3 Salience Driven PCFG

Probabilistic context free grammar (PCFG) can also be used as a language model in speech recognition by constraining the speech recognizer to generate only grammatical sentences as defined by the grammar.

4.3.1 Domain CFG

Based on the domain knowledge, we first define a domain-specific context free grammar (CFG) as shown in Fig.2. This CFG covers all the language that is "legal" in the interior decoration domain. An utterance is said to be"legal" in the domain if a semantic representation specific to the domain can be built from the utterance. The defined grammar covers the "legal" commands like "*this table*", "*remove this chair*", "*move this plant on this table*", and query questions like "*how much is this table?*", "*who is the artist of this painting?*", "*what is the wattage of this lamp?*".

S → NP | VP | WRB JJ VBZ NP | WRB JJ NN VBZ NP VB | WP VBZ NP PP | WRB VBZ NP VBN | VBZ NP NP
VP → VB NP | VB NP PP | VB NP JJ | VB NP RB
NP → NN | DT NN | PRP
PP → IN DT NN | TO DT NN
WP → what | who
WRB → how | where
JJ → big | black | blue | dark | expensive | gray | green | ...
VBZ → does | is
VB → add | align | bring | buy | change | delete | ...
RB → back | backward | backwards | down | forward | here | ...
NN → age | alternative | artist | artwork | back | bar | bed ...
DT → a | an | that | the | these | this | those
PRP → it | them
IN → about | above | against | among | around | at | behind ...
TO → to
VBN → made | produced

Figure 2: Domain specific context free grammar

<S> = <NP> | <VP> | <WRB> <JJ> <VBZ> <NP> | ...;
<VP> = <VB> <NP> | <VB> <NP> <PP> | <VB> <NP> <JJ> | <VB> <NP> <RB>;
<NP> = <NN> | <DT> <NN> | <PRP>;
<PP> = <IN> <DT> <NN> | <TO> <DT> <NN>;
<DT> = /117/ this | /59/ the | /16/ that | /3/ these | /1/ those | /1/ a | /1/ an;
<IN> = /34/ of | /17/ on | /10/ about | /7/ with | /4/ in | /2/ behind | ...;
<JJ> = /8/ many | /2/ much | /1/ small | /1/ left | /1/ expensive | ...;
<NN> = /144/ lamp | /24/ wattage | /7/ place | /7/ information | /6/ table | /5/ power ...;
<PRP> = /3/ it | /1/ them;
<RB> = /9/ here | /2/ back | /2/ up | /2/ there;
<TO> = to;
<VB> = /27/ remove | /18/ move | /7/ show | /6/ put | /6/ change | /6/ replace ...;
<VBN> = /2/ made | /1/ produced;
<VBZ> = /30/ is | /3/ does;
<WP> = /26/ what | /4/ who;
<WRB> = /9/ how | /5/ where;

Figure 3: Trained PCFG for entity *lamp* (in Java Speech Grammar Format)

4.3.2 Training of Entity Specific PCFG

To build the grammar, the Stanford Parser [11] was used for parsing the transcribed utterances. For each entity e in the 3D room scene, a PCFG was trained by maximum likelihood estimation on the transcripts annotated with entity e. In the trained PCFG, only the lexicon-part rules were associated with probabilities.

An example of trained PCFG for entity *lamp* is shown in Fig.3. The PCFG in Fig.3 is in the Java Speech Grammar Format (JSGF) and the numbers in the "/ /" are the weights of the rules. When normalized, the weights are the rule probabilities. As we can see in Fig.3, the words closely related to entity *lamp* such as "lamp" and "wattage" achieve higher weights in the trained PCFG. It means that those words closely related to *lamp* will be more likely chosen during the speech recognition process when the entity *lamp* is selected by the gesture.

4.3.3 Salience Driven PCFG

Gesture-based salience was used in creating a new salience driven PCFG by combining the PCFGs associated with the salient entities begin gestured. The weight of a rule r in the combined PCFG is given by:

$$w(r) = \sum_e w_e(r)p(e) \tag{10}$$

where $p(e)$ is the salience distribution, $w_e(r)$ is the weight of rule r in PCFG of entity e.

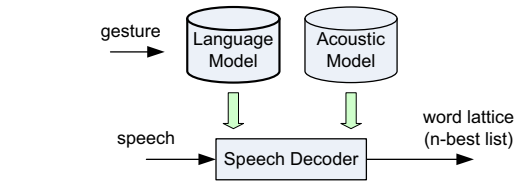
5. APPLICATION OF SALIENCE MODELS

The salience driven language models can be integrated into speech recognition in two stages: an early stage before word lattice (n-best list) is generated, or in a later stage where the word lattice (n-best list) is post-processed (Fig.4). We only focus on the language model component in speech recognition.

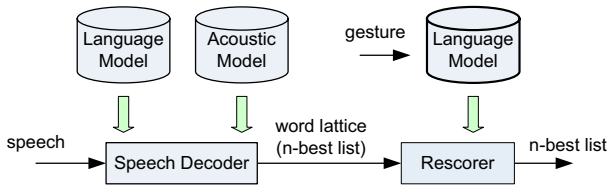
5.1 Early Application

For the early stage integration, as Fig.4(a) shows, the gesture-based salience driven language model is used together with the acoustic model to generate the word lattice, typically by Viterbi search.

Compared to n-gram models, CFG-based language models put more strict constraint on the speech recognition process, specifically on choosing the next set of possible words following a path during the searching process. When an n-gram



(a) Early application



(b) Late application

Figure 4: Integration of gesture-based salience driven language model in speech recognition

model is used, the next set of possible words includes any words in the vocabulary with non-zero transition probabilities (as specified by the n-gram model) from the previous n-1 words along the path. When a CFG-based language model is used, the next set of possible words only includes those allowable words as defined by the grammar.

5.2 Late Application

For the late stage integration, as shown in Fig.4(b), the gesture-based salience driven n-gram language model is used to rescore the word lattice generated by a speech recognizer with a basic language model not involving salience. A* search is applied to find the n-best paths in the word lattice.

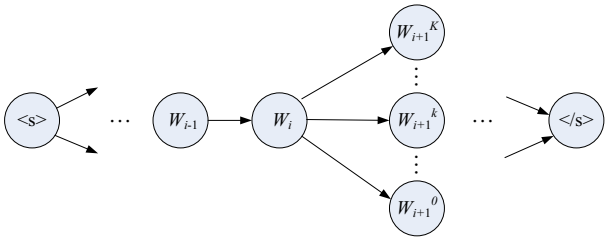


Figure 5: A* search in word lattice

A* search finds in a graph the optimal path from a given initial node to a given goal node. Specifically, in the word lattice shown in Fig.5, the task of A* search is to find a path from sentence start node "<s>" to sentence end node "</s>" that has the highest score. The score of a path $L = (w_0, w_1, \ldots, w_n)$ is defined as

$$f(L) = \sum_{i=0}^{n} \big(\log p_a(w_i) + \log p(w_i|w_{i-1})\big) \qquad (11)$$

where $p_a(w_i)$ is the acoustic model probability and $p(w_i|w_{i-1})$ is the language model probability. The language model probabilities can be tailored by the salience-driven language modeling described in Section 4.2.

In the word lattice, each node (i.e., a word hypothesis) is associated with a score. The score of a word w_i depends on two parts: the *true score* $g(w_i)$ which measures the actual score of the path from the start node to the current node and the *heuristic cost* $h(w_i)$ which measures the expected score of the path from the current node to the goal node. Depending on the score of a node, the system decides which node to expand during the search.

Before A* search begins, the heuristics at each node w_i are first calculated:

$$h(w_i) = \max_k \big\{ h(w_{i+1}^k) + \log p_a(w_{i+1}^k) + \log p(w_{i+1}^k|w_i) \big\} \qquad (12)$$

where $h(</s>) = 0$.

During the A* searching process, the score of the path up to node w_i is calculated as:

$$g(w_i) = g(w_{i-1}) + \log p_a(w_i) + \log p(w_i|w_{i-1}) \qquad (13)$$

where $g(<s>) = 0$.

6. EVALUATION

We empirically evaluated the different salience driven language models at the two stages using our multimodal conversational system.

6.1 Data collection

We conducted a wizard-of-oz study to collect data for our evaluation using the system described in Section 3. In the study, users were asked to accomplish two tasks. One task was to clean up and redecorate a messy room. Another task was to arrange and decorate the room so that it looks like the room in the pictures provided to user. Each of these tasks put the user into a specific role (e.g., college student, professor, merchant, etc), and the task had to be completed with a set of constraints (e.g. budget of furnishing, bed size, number of domestic products, etc).

From 5 users' interactions with the system, we collected 878 utterances and 469 of them have accompanying gestures. The vocabulary size of the collected utterances is 250 words.

Each utterance was transcribed and annotated with referred entities if applicable. For example, an utterance like "*remove this lamp*" accompanied by a deictic gesture was annotated with the true entity *lamp1* being gestured, while an utterance like "*move this lamp to this table*" accompanied by two deictic gestures were annotated with the entities *lamp1* and *table1* being gestured respectively. Here the concept name followed by a numerical number indicates a specific object.

Each gesture resulted in a set of possibly selected entities. The selection probabilities of the entities were calculated based on the distances from the gesture point to the center of the entities. Users' gesture points were captured by a touch screen.

All the collected data, together with utterance transcripts and entity annotation, were saved in XML format. Fig. 6 shows an excerpt form one of the XML data files. The excerpt is the record of one turn in the conversation between the system and one user. In this turn, the user pointed to entity *picture_girl* and said "*flip this picture one hundred eighty degrees*". The pointing gesture resulted in an ambiguous selection of three entities (*bedroom*, *picture_girl*, *table_pc*) with different probabilities.

```
<turn>
 <user_input>
  <gesture>
   <curve start="2153" end="2309">
    <point>613 183</point>
    <point>613 183</point>
   </curve>
   <selection>
    <entity text="bedroom">0.458000</entity>
    <entity text="picture_girl">0.530700</entity>
    <entity text="table_pc">0.011300</entity>
   </selection>
  </gesture>
  <speech>
   <entity_annotation>
    picture_girl
   </entity_annotation>
   <transcription>
    flip this picture one hundred eighty degrees
   </transcription>
   <waveform>2005916-144311-707.wav</waveform>
  </speech>
 </user_input>
</turn>
```

Figure 6: An excerpt of XML data file

6.2 Evaluation Metrics

We compare the performances of the following different language models trained in our domain:

- General bigram model (Bigram)
- General trigram model (Trigram)
- General class-based bigram model (C-Bigram)
- Salience driven bigram model (S-Bigram)
- Salience driven class-based bigram model (S-C-Bigram)
- General PCFG (PCFG)
- Salience-driven PCFG (S-PCFG)

The evaluation metrics include the following aspects related to recognition results:

- Word error rate of the best hypothesis (WER)
- Word lattice WER (Lattice-WER)
  The minimal WER of all possible paths though the word lattice (output of speech recognition).

Since we are building a conversational system, we are also interested in the following metrics related to semantic interpretation:

- Concept identification precision (CI-Precision)
  The percentage of correctly identified concepts out of the total number of concepts in the top hypothesis of the n-best list.
- Concept identification recall (CI-Recall)
  The percentage of correctly identified concepts out of the total number of concepts in a user's utterance.
- F-measurement (F-score)

$$F = \frac{(\beta^2 + 1) \times \text{CI-Precision} \times \text{CI-Recall}}{\beta^2 \times \text{CI-Precision} + \text{CI-Recall}} \qquad (14)$$

where $\beta = 1$ in this experiment.

6.3 Evaluation Results

The CMU Sphinx-4 speech recognizer [18] was used in all the experiments. The experiments were done by an eight-fold cross validation. We compare the performances of the salience driven language models for both early and late applications.

6.3.1 *Results from Early Application*

Table 1 shows the experiment results on the utterances with accompanying gestures. Overall, all n-gram models except the S-C-Bigram model performed better than the PCFG-based models on WER, and the S-Bigram model performed the best. One possible reason for bad performance of the PCFG-based models is due to the less flexibility of the grammar-based approaches. In terms of the language understanding metrics, all salience driven models (S-Bigram, S-C-Bigram, and S-PCFG) achieved roughly the same results on concept identification precision, while the S-PCFG model achieved the highest concept identification recall and F-measurement. Overall, the S-Bigram model appears to be the best one for the early application in that it not only achieved the lowest WER but also achieved a high F-score on concept identification (close to the highest one).

Among n-gram models, the performance of the trigram model is roughly the same as the bigram model. The S-Bigram model improved speech recognition and understanding compared to the three baselines (Bigram, Trigram, and C-Bigram). Compared to the trigram model, the S-Bigram model reduced the WER by 7%. A t-test showed that this was a significant change: $t = 3.38$, $p < 0.004$, one-tailed. The precision and recall of concept identification gained an increase of 3% and 4% respectively. The overall F-measurement was increased by 3%. A t-test showed that this was also a significant improvement: $t = 3.01$, $p < 0.0015$, one-tailed. The S-C-Bigram model achieved the best result on the precision of concept identification, but had the worst results on all other metrics.

Compared to the general PCFG model, the S-PCFG model increased the precision and recall of concept identification by 5% and 3.5% respectively. The overall F-measurement was increased by 4%. A t-test confirmed that this was a significant improvement: $t = 3.30$, $p < 0.001$, one-tailed. The S-PCFG model did not change the WER much compared to the general PCFG model. A t-test confirmed that the change in WER was insignificant: $t = 0.49$, N.S., two-tailed.

When compared to the trigram model, the S-PCFG model did not improve the WER but improved the language understanding. The F-measurement was increased by 4%. A t-test showed that this was a significant improvement: $t = 2.77$, $p < 0.003$, one-tailed.

6.3.2 *Results from Late Application*

We further compared different n-gram models: C-Bigram, S-Bigram, and S-C-Bigram during the late application. In these experiments, the general trigram model trained on our domain was first used to generate word lattices, then the salience driven models were used in A* search (Section 5.2) to find the best paths in the word lattices.

Table 2 shows the results of the three models on the 469 utterances with accompanying gestures. During the late application, the S-Bigram model performed the best with the exception of concept identification precision. Compared to the trigram model, the S-Bigram in late application de-

Table 1: Performance of the early application of language models

Language Model	Lattice-WER	WER	CI-Precision	CI-Recall	F-score
Bigram	0.250	0.321	0.830	0.793	0.811
Trigram	0.258	0.312	0.838	0.797	0.817
C-Bigram	0.292	0.371	0.856	0.748	0.798
S-Bigram	**0.243**	**0.291**	0.861	0.830	0.845
S-C-Bigram	0.412	0.448	**0.863**	0.623	0.724
PCFG	0.323	0.360	0.819	0.816	0.817
S-PCFG	0.319	0.355	0.862	**0.845**	**0.853**

Table 2: Performance of the late application of n-gram models

Language Model	Lattice-WER	WER	CI-Precision	CI-Recall	F-score
C-Bigram	0.258	0.334	0.831	0.784	0.807
S-Bigram	0.258	**0.294**	0.854	**0.834**	**0.844**
S-C-Bigram	0.258	0.316	**0.858**	0.786	0.821

creased the WER by 6%. A t-test showed that this was a significant change: $t = 2.66$, $p < 0.005$, one-tailed. On language understanding, the S-Bigram model increased the F-measurement by 3% compared to the trigram model. A t-test confirmed that this was a significant improvement: $t = 2.92$, $p < 0.002$, one-tailed.

Compared to Table 1, Table 2 shows that there is no difference in performance whether the S-Bigram model is applied early or later. However, a significant difference is observed for the S-C-Bigram model. The S-C-Bigram model performed much better when it was applied in a later stage. However, its performance was close to the baseline (trigram model). The WER change achieved by the S-C-Bigram model was not statistically significant from the t-test ($t = 0.94$, N.S., two-tailed), neither was the F-measurement ($t = 0.22$, N.S., two-tailed).

Our initial assumption is that the early application should have an advantage over the late application on bringing the *good* hypothesized words with low acoustic probabilities into the word lattice. This is particularly important when using the Sphinx-4 speech recognizer, because the current release of Sphinx-4 does not provide a full word lattice. When the correct words are not in the word lattice output, a late application of salience driven language models will never succeed in retrieving those correct words by rescoring the word lattice. Fig.7 shows one example that demonstrates the difference between the early application and the late application. Here the correct word "lamp" did not appear in the word lattice generated by the trigam model, and thus could not be retrieved by the late application of the salience driven bigram model. When the salience driven bigram model was applied in an early stage, the top one in the generated n-best list turned out to be the correct recognition result.

However, our current experimental results have not shown the particular advantage of the early application. Given these somewhat surprising results, we are currently investigating what have caused this behavior to reach a better understanding.

We also tested the effect of the accuracy of gesture recognition on the performance of salience driven language models. It is expected that the more accurate the gesture recognition, the better the performance salience driven language models should achieve. In another round of evaluation, we used the true entities pointed by the gestures in the salience driven

Utterance: "remove this lamp"

Gesture selection:
p(bedroom) = 0.0995
p(lamp_bank) = 0.5288
p(table_dresser) = 0.3604
p(table_pc) = 0.0114

N-best list with general trigram model:
remove this stand
remove this them
remove this left

N-best list with early integration of S-Bigram model:
remove this lamp
remove this lamp a

N-best list with late integration of S-Bigram model:
remove this left
remove this stand
remove this them

Figure 7: N-best list of an utterance: early stage integration v.s. late stage integration

language models. The experimental results have not shown significant difference from the results obtained by automated gesture recognition. The reason is that although one gesture could result in multiple possible selection, it turned out in our data that the true entities were usually associated with highest selection probabilities and could be easily identified.

7. CONCLUSIONS

This paper presents the results from a systematic investigation of incorporating domain context into speech recognition via gesture-based salience driven language modeling. Three salience driven language models based on the bigram model, the class-based bigram model, and the PCFG are compared. Our experimental results have shown that the salience driven bigram model can improve spoken language understanding in both early and late applications, while the salience driven class-based bigram model seems only useful for the late application. In the early application, the salience driven PCFG model has also shown a potential advantage in improving spoken language understanding. Given our re-

stricted domain, the potential of the salience modeling using non-verbal information may seem limited based on our evaluation. We are currently collecting more data from more complex scenarios to further evaluate the potential of these approaches.

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 anonymous reviewers for their valuable comments and suggestions.

9. REFERENCES

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

[2] J. Chai, P. Hong, M. Zhou, and Z. Prasov. Optimization in multimodal interpretation. In *Proceedings of 42nd Annual Meeting of Association for Computational Linguistics (ACL)*, 2004.

[3] J. Chai and S. Qu. A salience driven approach to robust input interpretation in multimodal conversational systems. In *Conferences on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, 2005.

[4] J. Y. Chai, P. Hong, and M. X. Zhou. A probabilistic approach to reference resolution in multimodal user interfaces. In *Proceedings of IUI'04*, pages 70–77, 2004.

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

[6] E. J. and C. C. A salience-based approach to gesture-speech alignment. In *Proceedings of HLT/NAACL'04*, 2004.

[7] M. Johnston. Unification-based multimodal parsing. In *Proceedings of COLING-ACL'98*, 1998.

[8] M. Johnston and S. Bangalore. Finite-state multimodal parsing and understanding. In *Proceedings of COLING'00*, 2000.

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

[10] A. Kehler. Cognitive status and form of reference in multimodal human-computer interaction. In *Proceedings of AAAI'00*, pages 685–689, 2000.

[11] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 2003.

[12] S. Lappin and H. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, (4):535–561, 1994.

[13] S. Oviatt. Breaking the robustness barrier: Recent progress on the design of robust multimodal systems. *Advances in Computers*, 56:305–325, 2002.

[14] S. L. Oviatt. Mutual disambiguation of recognition errors in a multimodal architecture. In *Proceedings of Conference on Human Factors in Computing Systems: CHI'99*, 1999.

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

[16] R. Stevenson. The role of salience in the production of referring expressions: A psycholinguistic perspective. In K. van Deemter and R. Kibble, editors, *Information Sharing*. CSLI Publ., 2002.

[17] P. Taylor, S. King, S. Isard, H. Wright, and J. Kowtko. Using intonation to constrain language models in speech recognition. In *Proceedings of ICASSP*, 1997.

[18] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel. Sphinx-4: A flexible open source framework for speech recognition. Technical Report TR-2004-139, Sun Microsystems Laboratories, 2004.