Influence of Gestural Salience on the Interpretation of Spoken Requests

Gideon Kowadlo, Patrick Ye, Ingrid Zukerman

Faculty of Information Technology, Monash University, Clayton, VICTORIA 3800, Australia

gkowadlo@gmail.com, ye.patrick@gmail.com, Ingrid.Zukerman@monash.edu

Abstract

We present a probabilistic, salience-based mechanism for the interpretation of pointing gestures together with spoken utterances. Our formulation models dependencies between spatial and temporal aspects of gestures and features of objects. The results from our corpus-based evaluation show that the incorporation of pointing information improves interpretation accuracy.
Index Terms: understanding spoken language and gesture, probabilistic approach

1. Introduction

In [1], we described *Scusi?* — a spoken language interpretation module which considers multiple interpretations, and employs a probabilistic formalism to estimate their goodness (Section 2). This module is part of a system called *DORIS* (Dialogue Oriented Roaming Interactive System) — a spoken dialogue system designed to be mounted on a household robot. In this paper, we extend *Scusi?*'s probabilistic formalism to integrate pointing gestures with spoken language. We adopt a salience-based approach where we take into account spatial and temporal information to estimate the probability that a pointing gesture refers to an object. Specifically, we consider (1) the location of each object relative to the spatial range of the pointing gesture, and (2) the timing of the gesture relative to the timing of the terms in a user's utterance.

To evaluate our formalism, we collected a corpus of requests where people were allowed to point (Section 4). Our results show that when people point, our mechanism yields significant improvements in interpretation performance. However, when pointing was artificially added to utterances where people did not point, it had a modest effect on performance.

This paper is organized as follows. Section 2 outlines the interpretation process for a spoken request. Section 3 describes the estimation of the probability of a pointing gesture. Our evaluation is detailed in Section 4. Related research and concluding remarks are given in Section 5 and 6 respectively.

2. Interpreting Spoken Requests

This section summarizes our previous work on the interpretation of single-sentence utterances [1]. *Scusi?* processes spoken input in three stages: speech recognition, parsing and semantic interpretation. First, it runs Automatic Speech Recognition (ASR) software (Microsoft Speech SDK 5.3) to generate candidate hypotheses (texts) from a speech signal, where each text is associated with a probability. In the second stage, Charniak's probabilistic parser (`ftp://ftp.cs.brown.edu/pub/nlparser/`) is applied to the texts in descending order of probability, associating each resultant parse tree with a probability.

During semantic interpretation, parse trees are successively mapped into two representations based on Concept Graphs [2]. First *Uninstantiated Concept Graphs (UCGs)*, and then *Instantiated Concept Graphs (ICGs)*. UCGs, which represent syntactic information, are obtained from parse trees deterministically — one parse tree generates one UCG. Each UCG can generate many ICGs. This is done by nominating different instantiated concepts and relations from the system's knowledge base as potential realizations for each concept and relation in a UCG. Instantiated concepts are objects and actions in the domain (e.g., `mug01`, `mug02` and `cup01` are possible instantiations of the uninstantiated concept "mug"). The interpretation process continues until a preset number of sub-interpretations (including texts, parse trees, UCGs and ICGs) has been generated or all options have been exhausted.

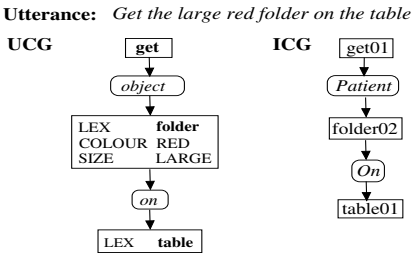


Figure 1: UCG and ICG for a sample utterance

Figure 1 illustrates a UCG and an ICG for the request "get the large red folder on the table". The *intrinsic* features of an object (lexical item, colour and size) are stored in the UCG node for this object. *Structural* features, which involve two objects (e.g., "*folder* on the *table*"), are represented as sub-graphs of the UCG (and the ICG).

2.1. Estimating the probability of an ICG

Scusi? ranks candidate ICGs according to their probability of being the intended meaning of a spoken utterance. Given a speech signal W and a context $\mathcal{C}$, the probability of an ICG I, $\Pr(I|W,\mathcal{C})$, is proportional to

$$\sum_{\Lambda} \Pr(T|W)\cdot\Pr(P|T)\cdot\Pr(U|P)\cdot\Pr(I|U,\mathcal{C}) \qquad (1)$$

where T, P and U denote text, parse tree and UCG respectively. The summation is taken over all possible paths $\Lambda = \{T, P, U\}$ from a speech wave to the ICG, because a UCG and an ICG can have more than one ancestor. As mentioned above, the ASR and the parser return an estimate of $\Pr(T|W)$ and $\Pr(P|T)$ respectively; and $\Pr(U|P) = 1$, since the process of generating a UCG from a parse tree is deterministic. The estimation of $\Pr(I|U,\mathcal{C})$ is described in [1]. Here we present the final equation obtained for $\Pr(I|U,\mathcal{C})$, and outline the ideas involved in its calculation.

$$\Pr(I|U,\mathcal{C}) \approx \prod_{k\in I} \Pr(u|k,\mathcal{C})\,\Pr(k|k_p,k_{gp})\,\Pr(k|\mathcal{C}) \qquad (2)$$

where u is a node in UCG U, k is the corresponding instantiated node in ICG I, k_p is k's parent node, and k_{gp} is k's grand-

parent node. For example, *On* is the parent of `table01`, and `folder02` the grandparent in the ICG in Figure 1.

- $\Pr(u|k)$ is the "match probability" between the specifications for node u in UCG U and the intrinsic features of the corresponding node k in ICG I, i.e., the probability that a speaker who intended a particular object k (e.g., `mug01`) gave the specifications in u (e.g., "the big red mug").
- $\Pr(k|k_p, k_{gp})$ represents the structural probability of ICG I, where structural information is simplified to node trigrams, e.g., whether `folder02` is *On* `table01`.
- $\Pr(k|\mathcal{C})$ is the probability of a concept given the context.

Scusi? handles three intrinsic features: lexical item, colour and size; and two structural features: ownership and several locative relations (e.g., on, under, near). The match probability $\Pr(u|k)$ and the structural probability $\Pr(k|k_p, k_{gp})$ are estimated using distance functions between the requirements specified by the user and what is found in reality [1]).

3. Incorporating Pointing Gestures

Pointing affects the salience of objects and the language used to refer to objects: objects in the temporal and spatial vicinity of a pointing gesture are more salient than objects that are farther away, and pointing is often associated with demonstrative determiners. Here we focus on the effect of pointing on salience, i.e., its effect on $\Pr(k|\mathcal{C})$ in Equation 2. Our calculations are based on information returned by Li and Jarvis's gesture recognition system [3]: gesture type, time, probability and relevant parameters (e.g., a vector for a pointing gesture). Owing to our focus on pointing gestures, we convert the probabilities expected from Li and Jarvis's system into the probability of Pointing and that of Not Pointing, which comprises all other gestures and no gesture (all these hypotheses are returned at the same time).[1] This yields the following probability for object k

$$\begin{aligned}\Pr(k|\mathcal{C}) \quad = \quad & \Pr(k|\mathcal{P},\mathcal{C}) \cdot \Pr(\mathcal{P}|\mathcal{C}) + \\ & \Pr(k|\neg\mathcal{P},\mathcal{C}) \cdot \Pr(\neg\mathcal{P}|\mathcal{C})\end{aligned} \tag{3}$$

where $\mathcal{P}$ designates Pointing, $\Pr(\mathcal{P}|\mathcal{C})$ and its complement are returned by the gesture system, and $\Pr(k|\neg\mathcal{P},\mathcal{C}) = \frac{1}{N}$ (N is the number of objects in the room, i.e., in the absence of pointing, we assume that all the objects in the room are equiprobable).

Pointing is spatially correlated with objects, and temporally correlated with words that refer to objects. Hence, we separate a pointing gesture $\mathcal{P}$ into two components, spatial ($\mathcal{P}_s$) and temporal ($\mathcal{P}_t$).

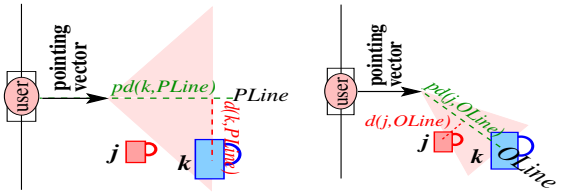
$$\Pr(k|\mathcal{P},\mathcal{C}) = \Pr(k|\mathcal{P}_s,\mathcal{P}_t,\mathcal{C})$$

The influence of each component on the probability of object k is modeled by a sigmoid function, yielding the following formulation.

$$\begin{aligned}&\Pr(k|\mathcal{P}_s,\mathcal{P}_t,\mathcal{C}) \\ &= \frac{N-1}{N}\frac{1}{1+e^{-\frac{\Pr(k|\mathcal{P}_s,\mathcal{C})-\mu}{\nu}}}\frac{1}{1+e^{-\frac{\Pr(k|\mathcal{P}_t,\mathcal{C})-\mu}{\nu}}}+\frac{1}{N}\end{aligned} \tag{4}$$

where μ and ν are parameters that ensure that a sigmoid functions yields a value close to 0 when the probability of k given $\mathcal{P}_s$ (or $\mathcal{P}_t$) is 0, and a value of 0.99 when this probability is 1 ($\mu = 0.5$ and $\nu = -\frac{0.5}{\ln(1/0.99-1)} = 0.1088$). $\Pr(k|\mathcal{P}_s,\mathcal{C})$ and $\Pr(\mathcal{P}_t|k,\mathcal{C})$ are estimated as described in Section 3.1 and 3.2 respectively.

[1] Owing to timing limitations of the gesture recognition system, we simulate its output (Section 4).



(a) Pointing (b) Occlusion

Figure 2: Spatial pointing and occlusion

The combination of sigmoids yields a high probability of intending an object k only when k is both spatially and temporally salient. The scaling factor $\frac{N}{N-1}$ and the offset $\frac{1}{N}$ ensure that $\Pr(k|\mathcal{P},\mathcal{C}) \approx \frac{1}{N}$ when the pointing gesture yields probabilities below the uninformed prior ($\frac{1}{N}$).

3.1. Estimating $\Pr(k|\mathcal{P}_s,\mathcal{C})$

$\Pr(k|\mathcal{P}_s,\mathcal{C})$, the probability that the user intended object k when pointing to a location in space, is estimated using a conic Gaussian density function around *PLine*, the *Pointing Line* created by extending the pointing vector returned by the gesture identification system (Figure 2(a)).

$$\Pr(k|\mathcal{P}_s,\mathcal{C}) = \frac{\alpha\theta_k}{\sqrt{2\pi}\sigma_{P_s}(pd)} e^{-\frac{d(k,PLine)^2}{2\sigma^2_{P_s}(pd)}} \tag{5}$$

where α is a scaling constant;[2] $\sigma^2_{P_s}(pd)$ is the variance of the Gaussian cone as a function of $pd(k, PLine)$, the *projected distance* between the user's pointing hand and the projection of object k on *PLine*; $d(k, PLine)$ is the shortest distance between the center of object k and *PLine*; and θ_k is a factor that reduces the probability of object k if it is (partially) *occluded*.

The ***projected distance*** pd is employed to take into account the imprecision of pointing — a problem that is exacerbated by the uncertainty associated with sensing the pointing vector. A small angular error in the detected pointing vector yields a discrepancy in the distance between the pointing line and candidate objects. This discrepancy increases as $pd(k, PLine)$ increases. To compensate for this situation, we increase the variance of the Gaussian distribution linearly with the projected distance from the user's hand (we start with a small standard deviation of $\sigma_0 = 5$ mm at the user's fingers, attributed to sensor error). This allows farther objects with a relatively high displacement from the pointing vector to be encompassed in a pointing gesture (e.g., the larger mug in Figure 2(a)), while closer objects with the same displacement are excluded (e.g., the smaller mug). This yields the following equation for the variance.

$$\sigma^2_{P_s}(pd) = \sigma^2_0 + K \cdot pd(k, PLine)$$

where $K = 2.5$ mm is an empirically determined increase rate.

The ***occlusion factor*** θ_k reduces the probability of objects as they become more occluded. We approximate θ_k by considering the objects that are closer to the user than k, and estimating the extent to which these objects occlude k (Figure 2(b)). This estimate is a function of the position of these objects and their size — the larger an intervening object, the lower the probability that the user is pointing at k. These factors are taken into account as follows.

$$\Pr_\theta(j|k) = \frac{\gamma}{\sqrt{2\pi}\sigma_\theta(pd)} e^{-\frac{(d(j,OLine)-\frac{1}{2}\dim_{\min}(j))^2}{2\sigma^2_\theta(pd)}} \tag{6}$$

[2] Since this is a continuous density function, it does not directly yield a point probability. Hence, it is scaled on the basis of the largest possible returned value.

where γ is a scaling constant; the numerator of the exponent is the maximum distance from the edge of object j to the line between the user's hand and object k, denoted *Object Line* (*OLine*); and

$$\sigma_\theta^2(pd) = \frac{1}{2}\left(\sigma_0^2 + K \cdot pd(j, OLine)\right)$$

is the variance of a cone from the user's hand to object k as a function of distance. We employ a thin "occlusion cone" (Figure 2(b)), which has half the variance of that used for the "pointing cone", to represent the idea that object j must be close to *OLine* in order to occlude object k. θ_k is then estimated as 1 minus the maximum occlusion caused by the objects that are closer to the user than k.

$$\theta_k = 1 - \max_{\forall j \ d(j,\text{hand}) < d(k,\text{hand})} \{\Pr_\theta(j|k)\} \quad (7)$$

3.2. Estimating $\Pr(k|\mathcal{P}_t, \mathcal{C})$

$\Pr(k|\mathcal{P}_t, \mathcal{C})$, the probability that the user intended object k when pointing at a particular time, is estimated on the basis of $T(u_{\text{lex}}(k))$, the start time of the word designating k in the parent UCG node of the ICG node containing k. That is,

$$\Pr(k|\mathcal{P}_t, \mathcal{C}) = \Pr(T(u_{\text{lex}}(k))|\mathcal{P}_t, \mathcal{C})$$

$\Pr(T(u_{\text{lex}}(k))|\mathcal{P}_t, \mathcal{C})$ is obtained from a Gaussian time distribution for pointing.

$$\Pr(T(u_{\text{lex}}(k))|\mathcal{P}_t, \mathcal{C}) = \frac{\beta}{\sqrt{2\pi}\sigma_{P_t}} e^{-\frac{(T(u_{\text{lex}}(k)) - \mathcal{P}_t)^2}{2\sigma_{P_t}^2}} \quad (8)$$

where β is a scaling constant, and σ_{P_t} is the standard deviation of the Gaussian density function, which is currently set to 647 msec (based on our corpus).

4. Evaluation

To obtain a corpus, we conducted a user study where we set up a room with labeled objects (Figure 3), and asked trial participants to ask *DORIS* for 12 specific items. The room contained 33 items in total, including distractors, and one of the authors pretended to be *DORIS*. We designated the items to be requested using labels, and the participants chose the wording and gestures (if any) for their requests. The objects in the room were selected and laid out in the room to reflect a variety of conditions, e.g., common and rare objects (e.g., vacuum tube); unique, non-unique and similar objects (e.g., white cups); and objects placed near each other and far from each other.

We divided our corpus into two parts: with and without pointing gestures. *Scusi?*'s performance was tested on input obtained from the ASR and on text (perfect ASR). We considered two scenarios for each sub-corpus: *Scusi?*-Pointing, where our pointing mechanism was activated on the basis of a simulated pointing gesture,[3] and *Scusi?*-NoPointing, where our pointing mechanism was not activated. This was done in order to test two hypotheses: (1) when people point, pointing information improves interpretation performance; and (2) when people do not point, even perfect pointing has little effect on performance.

Scusi? was set to generate at most 300 sub-interpretations in total (including texts, parse trees, UCGs and ICGs) for each spoken request, and at most 200 sub-interpretations for each textual request. An interpretation was deemed successful if it correctly represented the speaker's intention, which was encoded in one or more *Gold ICG*s. These ICGs were manually constructed on the basis of the requested objects and the participants' utterances. Multiple Gold ICGs were allowed if there were several suitable actions in the knowledge base.

[3] At present we assume accurate pointing. In the near future, we will study the sensitivity of our mechanism by incorporating pointing error.

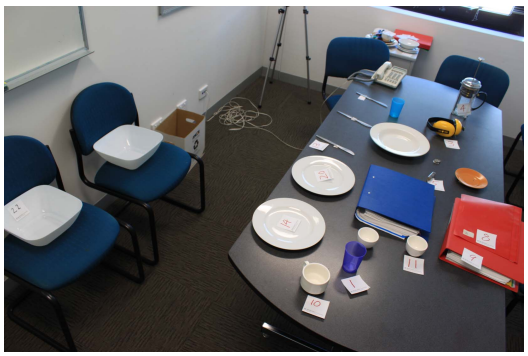


Figure 3: Experimental Setup

4.1. The Corpus

19 people participated in the trial, generating a total of 267 requests, of which 136 involved pointing gestures. We filtered out 64 requests, which included concepts our system cannot yet handle, e.g., projective modifiers (e.g., "behind/left"), ordinals ("first/second"), references to groups of things (e.g., "the six blue pens"), and zero- and one-anaphora. This yielded 212 requests, of which 105 involved pointing gestures.

In addition, the software we used has the following limitations: the gesture recognition system [3] requires users to hold a gesture for 2 seconds, and the ASR system is speaker dependent and cannot recognize certain words (e.g., "mug", "bowl" and "pen"). To circumvent these problems, each pointing gesture was manually encoded into a time-stamped vector on the basis of video recordings of the participants; and one of the authors read slightly sanitized versions of participants' utterances into the ASR, specifically "can you", "please" and "*DORIS*" were omitted, and words that were problematic for the ASR were replaced (e.g., "pencil" was used instead of "pen").

Requests with a pointing gesture were somewhat shorter than those without pointing (5.84 versus 6.27 words on average). ASR performance was worse for requests with pointing: the top ASR interpretation was correct for 72% of (sanitized) requests with pointing, compared to 79.5% for requests without pointing. This difference may be attributed to the language model of the ASR not coping well with constructs associated with pointing. Overall the ASR returned the correct interpretation, at any rank, for 90.6% of the requests.

4.2. Results

Table 1 summarizes our results. Column 1 displays the test condition (sub-corpus with/without pointing gesture, Text/ASR, and *Scusi?* with/without pointing module). Columns 2-3 show the percentage of utterances that had Gold ICGs whose probability was among the top 1 and top 3, e.g., in the sub-corpus with pointing, when *Scusi?*-Pointing was run on Text, it yielded Gold ICGs with the highest probability (top 1) 87.6% of the time, and within the top 3 probabilities 91.4% of the time. The average *adjusted rank (AR)* and *rank* of the Gold ICG appear in Column 4. The rank of an ICG I is its position in a list sorted in descending order of probability (starting from position 0), such that all equiprobable ICGs are deemed to have the same position. The AR of an ICG I is the mean of the positions of all ICGs that have the same probability as I, e.g., if we have 4 equiprobable ICGs in positions 0-3, each has a rank of 0, but an adjusted rank of $\frac{r_{\text{best}} + r_{\text{worst}}}{2} = 1.5$. Column 5 shows the percentage of utterances that didn't yield a Gold ICG (% Not Found).

Our results confirm that the main role of pointing is in ref-

Table 1: *Scusi?*'s interpretation performance

	% Gold ICGs top 1	% Gold ICGs top 3	Avg adjusted rank (rank)	% Not found
Sub-corpus without pointing gesture				
Text, *Scusi?*-NoPointing	86.9	95.3	1.56 (0.48)	1.9
Text, *Scusi?*-Pointing	86.0	93.5	**0.99 (0.74)**	0.9
ASR, *Scusi?*-NoPointing	82.2	91.6	4.73 (0.68)	4.7
ASR, *Scusi?*-Pointing	81.3	88.8	**3.12 (1.06)**	4.7
Sub-corpus with pointing gesture				
Text, *Scusi?*-NoPointing	86.7	94.3	3.31 (0.26)	1.9
Text, *Scusi?*-Pointing	87.6	91.4	**1.90 (0.43)**	1.0
ASR, *Scusi?*-NoPointing	74.3	86.7	10.01 (0.56)	8.6
ASR, *Scusi?*-Pointing	76.2	81.9	**7.45 (1.09)**	6.7

erent disambiguation. This is evident from the reduction in AR (Column 4) for the sub-corpus with pointing gesture, which for ASR goes from 10.01 under *Scusi?*-NoPointing to 7.45 under *Scusi?*-Pointing, and for Text goes from 3.31 to 1.90. Both differences are statistically significant with $p < 0.01$.[4] As expected, the improvements obtained by artificially introducing pointing gestures in the sub-corpus without pointing are smaller: from 4.73 to 3.12 for ASR ($p < 0.05$), and from 1.56 to 0.99 for Text ($p < 0.01$). Comparing across language modalities, the impact of pointing on *Scusi?*'s performance with ASR input is larger than its impact on *Scusi?*'s performance with Text. We posit that this happens because the information obtained from pointing overcomes ASR error.

As seen in Columns 1-2, *Scusi?*-Pointing yields an apparent reduction in the percentage of interpretations with top ranks. This is because under *Scusi?*-NoPointing, there are often several equiprobable interpretations, which have the same rank. This happens less often under *Scusi?*-Pointing, owing to the disambiguating effect of pointing. It is worth noting that normally there is a trade-off between the number of Not Found Gold ICGs and average AR. ICGs that are not found by one approach but are found by another approach typically have a high (bad) rank when they are eventually found [1]. Thus, an approach that fails to find such "difficult" ICGs yields artificially lower ranks than an approach that finds these ICGs. An increase in the number of found Gold ICGs coupled with a reduction in average AR therefore demonstrate substantial performance improvements.

Finally, the rank of the request at the 75%-ile is 0 under all conditions, which indicates creditable performance. The larger number of Not Found Gold ICGs for the ASR condition, in particular for the pointing sub-corpus, is consistent with the above-mentioned ASR performance, which was significantly worse for the pointing sub-corpus. Other Not Found Gold ICGs were mainly due to parsing errors.

5. Related Research

Researchers in gesture and speech integration tend to favour one main modality, employing the other one for disambiguation. For instance, speech is the main input modality in [4, 5], while gesture is the main modality in [6, 7]. Different approaches are used for gesture detection, e.g., vision [4, 5] and sensor glove [6, 7, 8]; and for language interpretation, e.g., dedicated grammars [5], context-free grammars [4, 6], and keywords [7, 9]. Semantic fusion is often used to combine spoken input with pointing gestures, and is variously implemented using heuristics based on temporal overlap [10], or unification to determine which elements can be merged [4, 6, 8]. These are sometimes combined with search techniques coupled with penalties [5, 9]. With the exception of Bolt's system [10], these systems were tested on utterances that were quite short and constrained, whereas we can handle more complex utterances.

Like *Scusi?*, the systems described in [4, 6, 8] consider several hypotheses, but they do so using n-best lists. Cones have been used to model pointing gestures in [4, 6]. The system described in [6] is the most similar to *Scusi?* in its use of cones and its multiplication of probabilities obtained from speech and gesture. However, the cones are obtained using sensor gloves, and the probability of being "in the cone" is estimated using heuristics that combine different types of rankings.

Salience-based approaches are described in [9, 11]. They use salience to weigh the importance of factors pertaining to gesture-speech alignment, but they do not consider the uncertainty associated with pointing.

6. Conclusion

We have offered a formalism that takes into account relationships between spoken language and spatial and temporal aspects of gesture to integrate information about pointing gestures into the estimation of the probability of candidate interpretations of an utterance. Our empirical evaluation shows that our formalism significantly improves interpretation accuracy.

7. Acknowledgments

This research was supported in part by ARC grant DP0878195. The authors thank David Li and Ray Jarvis for their help with the gesture recognition system.

8. References

[1] E. Makalic, I. Zukerman, and M. Niemann, "A spoken language interpretation component for a robot dialogue system," in *Interspeech 2008*, 2008, pp. 195–198.

[2] J. Sowa, *Conceptual Structures: Information Processing in Mind and Machine*. Reading, MA: Addison-Wesley, 1984.

[3] Z. Li and R. Jarvis, "Real time hand gesture recognition using a range camera," in *Australasian Conf. on Robotics and Automation*, 2009.

[4] H. Holzapfel, K. Nickel, and R. Stiefelhagen, "Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3D pointing gestures," in *ICMI'04*, 2004, pp. 175–182.

[5] A. Brooks and C. Breazeal, "Working with robots and objects: Revisiting deictic reference for achieving spatial common ground," in *HRI2006*, 2006, pp. 297–304.

[6] E. Kaiser, A. Olwal, D. McGee, H. Benko, A. Corradini, X. Li, P. Cohen, and S. Feiner, "Mutual dissambiguation of 3D multimodal interaction in augmented and virtual reality," in *ICMI'03*, 2003, pp. 12–19.

[7] E. Tse, M. Hancock, and S. Greenberg, "Speech-filtered bubble ray: improving target acquisition on display walls," in *ICMI'03*, 2003, pp. 307–314.

[8] A. Corradini, R. Wesson, and P. Cohen, "A Map-Based system using speech and 3D gestures for pervasive computing," in *ICMI'02*, 2002, pp. 191–196.

[9] J. Einstein and C. Christoudias, "A salience-based approach to gesture-speech alignment," in *NAACL'2004*, 2004, pp. 25–32.

[10] R. Bolt, ""Put-that-there": Voice and gesture at the graphics interface," in *SIGGRAPH7*, 1980, pp. 262–270.

[11] C. Huls, W. Claassen, and E. Bos, "Automatic referent resolution of deictic and anaphoric expressions," *Computational Linguistics*, vol. 21, no. 1, pp. 59–79, 1995.

[4] Statistical significance was calculated using a paired t-test for the Gold ICGs that were found by *Scusi?*-Pointing and *Scusi?*-NoPointing.