Gesture Features for Coreference Resolution

Jacob Eisenstein and Randall Davis

Computer Science and Artificial Intelligence Laboratory,
Massachusetts Institute of Technology, Cambridge MA 02139, USA
{jacobe,davis}@csail.mit.edu

Abstract. If gesture communicates semantics, as argued by many psychologists, then it should be relevant to bridging the gap between syntax and semantics in natural language processing. One benchmark problem for computational semantics is *coreference resolution*: determining whether two noun phrases refer to the same semantic entity. Focusing on coreference allows us to conduct a quantitative analysis of the relationship between gesture and semantics, without having to explicitly formalize semantics through an ontology. We introduce a new, small-scale video corpus of spontaneous spoken-language dialogues, from which we have used computer vision to automatically derive a set of gesture features. The relevance of these features to coreference resolution is then discussed. An analysis of the timing of these features also enables us to present new findings on gesture-speech synchronization.

1 Introduction

Although the natural-language processing community has traditionally focused mainly on text, the actual usage of natural language between people is primarily oral and face-to-face. Extension of robust NLP to face-to-face communication offers the potential for breakthrough applications in domains such as meetings, lectures, and presentations. We believe that in face-to-face discourse, it is important to consider the possibility that non-verbal communication may offer information that is crucial to language understanding. However, due to the long-standing emphasis on text datasets, there has been little work on non-textual features.

In this paper, we investigate the relationship between gesture and semantics. We use machine vision to extract hand positions from a corpus of sixteen videos. We present a set of features that are derived from these hand positions, and use statistical methods to characterize the relationship between the gesture features and the linguistic semantics. Semantics is captured concretely in the context of *coreference*, which occurs when two noun phrases refer to the same entity. If gesture features can predict whether two noun phrases corefer, then they can contribute to the semantic analysis of speech.

2 Corpus

To conduct this research, we have begun to gather a corpus of multimodal dialogues. This work is preliminary, and the size of the corpus is relatively small; as we will describe in more detail at the end of Section 3, our corpus is roughly half the size of

S. Renals, S. Bengio, and J. Fiscus (Eds.): MLMI 2006, LNCS 4299, pp. 154–165, 2006.
© Springer-Verlag Berlin Heidelberg 2006

the MUC-6 coreference evaluation formal corpus [1]. We hope that interest in multimodal natural language processing will increase, leading to the development of better and broader corpora.

2.1 Procedure

Thirty college students and staff, aged 18-32, joined the study by responding to posters on our university campus. A subset of nine pairs of participants was selected on the basis of recording quality, and their speech was transcribed and annotated. The corpus is composed of two videos from each of the nine pairs; technical problems forced us to exclude two videos, yielding 16 annotated documents, each between two and three minutes in duration.

McNeill [2] and others have long advocated studying dialogues in which the speaker and listener already know each other. This eliminates a known confound in which the speaker and listener increase the rate of gestures as they become acquainted over the course of the experiment. For this reason, we recruited participants in pairs; 78% of participants described themselves as "close friends" or spouses; 20% as "friends", and 3% as "acquaintances".

One participant was randomly selected from each pair to be the "speaker," and the other to be the "listener." The speaker's role was to explain the behavior of a mechanical device to the listener. The listener's role was to understand the speaker's explanations well enough to take a quiz given later. The listener was allowed to ask questions of the speaker; however the listener's speech has not yet been transcribed, and is not considered in this study.

Prior to each discussion, the speaker either privately viewed a simulation of the device in operation, or left the room and examined the actual physical object. In explaining the device, the speaker was provided with either a whiteboard marker with which to create a sketch, a pre-printed diagram of the device, or no visual aids at all. In this paper, we will consider only data from the condition with the pre-printed visual aid. The interpretation of gestures in this condition is thought to be more straightforward; many, if not most of the gestures involve pointing at locations on the diagram. While the (presumably) more challenging problem of understanding gesture without visual aids is interesting future work, printed or projected diagrams are common in business presentations and lectures, so this restriction does not seem to be overly limiting to the applicability of our work.

The speaker was limited to two minutes to view the video or object and three minutes to explain it; the majority of speakers used all of the time allotted. This suggests that we could have obtained more natural data by not limiting the explanation time. However, we found in pilot studies that this led to problematic ordering effects, where participants devoted a long time to the early conditions, and then rushed through later conditions. With these time constraints, the total running time of the experiment was usually around 45 minutes. More details on the data-gathering portion of this research can be found in [3].

2.2 Speech and Vision Analysis

Speech was recorded using headset microphones. A homebrew Java system controlled the synchronization of the microphones and video cameras. Speech was transcribed

manually, and audio was hand-segmented into well-separated chunks with duration not longer than twenty seconds. The chunks were then force-aligned by the SPHINX-II speech recognition system [4].

Video recording was performed using standard digital camcorders. Participants were given different colored gloves to facilitate hand tracking. Despite the use of gloves, a post-study questionnaire indicated that only one of the thirty participants guessed that the study was related to gesture. The study was deliberately designed so that participants had very little free time to think; when not actually conducting the dialogue, the speaker was busy viewing the next mechanical system, and the participant was busy being tested on the previous conversation. We also presented consent forms immediately after the gloves, which may have diverted attention from the gloves' purpose.

An articulated upper-body tracker was used to model the position of the speaker's torso, arms, and hands. By building a complete upper-body tracker, rather than simply tracking the individual hands, we were able to model occlusion directly. Search across configurations of the tracker was performed using an annealed particle filter, implemented using the OpenCV library.[1] Essentially, the system performed a randomized beam search to simultaneously achieve three objectives: a) maximize coverage of the foreground area, b) match the known glove color to the color observed at the hypothesized hand positions, c) respect physiological constraints and temporal continuity.

The tracker was based largely on the work of [5]; the main differences were that Deutscher et al. did not use color cues such as gloves, but had access to multiple cameras to facilitate 3D tracking. We used only a single monocular camera and a 2.5D model (with just one degree of freedom in the depth plane). From inspection, the lack of depth information was the cause of many of our system's errors; bending of the arm joints in the depth dimension caused the arm length to appear to change in ways that were confusing to our model. Nonetheless, we estimate that both hands were tracked accurately and smoothly over 90% of the time. It is difficult to assess the tracker performance more precisely, as that would require ground truth data in which the actual hand positions were annotated manually at each time step.

3 Annotation

Coreference annotation is a two-step process [6]. First, all noun phrases (NPs) that may participate in coreference relations are selected; these are sometimes called "markables." All third-person noun phrases were included as markables; in addition, nested markables were annotated (e.g., "[one of [these walls]]" [2]). First and second person NPs were excluded as markables, as they were thought to be irrelevant to the issue of understanding the explanatory narratives about physical devices; furthermore, it seemed unlikely that gesture could play much of a role in disambiguating coreferences among such entities. It would be easy to automatically separate out these NPs in most cases. Manual annotation also disambiguated different senses of words like "that," which can be a referential pronoun (e.g. "that is rotating") or a relative pronoun "the one that rotates." The word "there" and "it" were also not included as markables when they did not indicate entities, as in "there's no way out of here," and "it's hard to tell."

[1] http://www.intel.com/technology/computing/opencv/

[2] In this notation, noun phrases are set off by brackets.

The corpus consists of spontaneous speech, so disfluencies abound. Repeated word disfluencies were automatically eliminated when the repetitions were adjacent, but other disfluencies were left uncorrected; rather than performing coreference resolution on an error-free text, we include the markables that occur inside disfluencies without prejudice. A frequent type of disfluency involves restatement of a noun phrase, e.g., "so this pushes [these] [all these things] up." This is a *substitution* disfluency, where "all these things" substitutes for "these." However, both are treated as markables, with a coreference relation between them, since they refer to the same set of objects.

A total of 1141 markables were found in the corpus, an average of 71 per video sequence ($\sigma = 27$). After the markables were annotated, the second step is to specify the coreference relations between them. As with the selection of markables, coreference annotation was performed by the first author, following the MUC-7 task definition [6]. A coreference relation was annotated whenever two noun phrases were judged to have an identical reference. 74.5% of markables participated in coreference relations, and there were a total of 474 entities, yielding a markable-to-entity ratio of 2.4.

4 Features

To assess the relationship between gestures and coreference, we computed a set of features describing the position and motion of the tracked hands. Two of the these features are *comparative*, in that they can be used to measure the similarity of gestures during different points in time. These can be applied directly to coreference, comparing the gestures observed during the two candidate noun phrases. Five other features are not comparative, meaning that they describe only a single gesture. These features can be used to assess the likelihood that an individual noun phrase relates to *any* previously defined entity, or the likelihood that the comparative features will be applicable to determining coreference.

Some of the features invoke the idea of "focus": which hand, if any, is gesturing during the utterance of the noun phrase. There are four logical possibilities: both, left, right, or neither. For the moment we ignore bimanual gestures, which were not frequent in our data. The determination of which hand is in focus is governed by the following heuristic: select the hand farthest from the body in the x-dimension, as long as the hand is not occluded and its y-position is not below the midsection of the speaker's body. If neither hand meets these criteria, than no hand is said to be in focus.

In the notation that follows, $x_{start_j,L}$ refers to the x-position of the left hand at the start of the noun phrase j. For the non-comparative features, there is only one noun phrase, and so no need to index them. $y_{start,F}$ refers to the y-position of whichever hand is in focus at the start of the noun phrase.

4.1 Comparative Features

– FOCUS DISTANCE: The distance between the positions of the in-focus hand during the two candidate noun phrases. The focus distance is undefined if there is no focus hand during either candidate noun phrase. FOCUS DISTANCE is given by

$$\sqrt{(x_{midpoint_j,F_j} - x_{midpoint_i,F_i})^2 + (y_{midpoint_j,F_j} - y_{midpoint_i,F_i})^2}. \quad (1)$$

– WHICH HAND: Takes three values: *SAME*, if the same hand is in focus during both candidate NPs; *DIFFERENT*, if a different hand is in focus; *MISSING*, if no focus hand is found during at least one of the NPs.

4.2 Non-comparative Features

– FOCUS SPEED: The total displacement of the in-focus hand, divided by the duration of the word. FOCUS SPEED is undefined if there is no focus hand at either the beginning or end of the word, or if the focus hand is different at the end of the word from the focus hand at the beginning. Note that by this metric, the FOCUS SPEED is zero if a gesture ends in the same place that it begins, regardless of how much distance the hand traversed. FOCUS SPEED is given by

$$\sqrt{(x_{end,F} - x_{start,F})^2 + (y_{end,F} - y_{start,F})^2}/\text{duration}. \tag{2}$$

– JITTER: JITTER measures the average frame-by-frame speed of each hand, over the course of the NP. Since this feature does not require the determination of the focus hand, it is never undefined. However, contributions at instants when a hand is occluded are not counted in the sum. This feature was found to be more informative when smoothed by a Gaussian kernel. JITTER is given by

$$\sum_{t \geq start}^{end} \frac{\sqrt{(x_{t,L} - \bar{x}_L)^2 + (x_{t,R} - \bar{x}_R)^2 + (y_{t,L} - \bar{y}_L)^2 + (y_{t,R} - \bar{y}_R)^2}}{\text{duration}}. \tag{3}$$

– PURPOSE: This feature captures how much of the overall motion (the JITTER) is explained by purposeful motion between the start and end points of the gesture. PURPOSE is simply FOCUS SPEED / JITTER; it is defined to be zero whenever jitter is zero, and is undefined whenever FOCUS SPEED is undefined. The value is maximized for fast, directed motions that go linearly between the start point and the end point; it is minimized for erratic motions for which the end point is not far from the start point. As with FOCUS SPEED, this feature will give a low score to any periodic motion, even large circles.

– SYNCHRONIZATION: SYNCHRONIZATION measures the degree to which the two hands are moving on the same trajectory. Like JITTER, this feature does not use focus information, and is therefore never undefined. However, contributions at instants when a hand is occluded are not counted in the sum. The value of this feature is 1 if the two hands are perfectly synchronized, 0 if the hands are moving orthogonally, and -1 if the hands are antisynchronized (i.e., one hand is moving clockwise and the other counterclockwise). $atan2$ is the full circle arctan function, which returns values in the range $\{-\pi, \pi\}$.

$$\theta_{t,h} = atan2(y_{t,h} - y_{t-1,h}, x_{t,h} - x_{t-1,h}) \tag{4}$$

$$\text{synch}_t = cos(\theta_{t,L} - \theta_{t,R}) \tag{5}$$

$$\text{synch} = \sum_{t > start}^{end} \frac{\text{synch}_t}{\text{duration}} \tag{6}$$

– SCALED SYNCHRONIZATION: SCALED SYNCHRONIZATION scales the SYNCHRONIZATION score at each time step by the average of the speeds of the two hands at that time step. Thus, large synchronized movements are weighed more heavily than small ones. The velocities $v_{t,h}$ – referring to the velocity of hand h at time t – are smoothed using a Gaussian kernel.

$$v_{t,h} = \sqrt{(x_{t,h} - x_{t-1,h})^2 + (y_{t,h} - y_{t-1,h})^2} \tag{7}$$

$$\text{scaled\_synch} = \frac{\sum_{t>start}^{end}(v_{t,L} + v_{t,R})\text{synch}_t}{2 * \text{duration}} \tag{8}$$

5 Feature Relevance

Both of the comparative features were predictive of coreference. FOCUS DISTANCE, which measures the distance between the hand positions during the two noun phrases, varied significantly depending on whether the noun phrases coreferred (see Table 1). The average noun phrase has a FOCUS DISTANCE of 48.4 pixels to NPs with which it corefers ($\sigma = 32.4$), versus a distance of 74.8 to NPs that do not corefer ($\sigma = 27.1$); this difference is statistically significant ($p < .01, dof = 734$). We expected to find larger differences for pronouns or NPs starting with the word "this," expecting gesture to play a larger role is disambiguating such NPs. As shown in the table, this does not appear to be the case; there was no significant increase in the discriminability of the FOCUS DISTANCE feature for such linguistic phenomena. However, FOCUS DISTANCE does appear to play less of a discriminative role for definite NPs (e.g., "the ball") than for non-definite NPs ($t = 3.23, dof = 183, p < .01$). This suggests that the definite article is not used as frequently in combination with gestures that communicate meaning by hand location, and that computer systems may benefit by attending more to gesture during non-definite NPs.

Table 1. Average values for the FOCUS DISTANCE feature (Equation 1), computed for various linguistic phenomena

	coreferring	not coreferring	difference
all	48.4	74.8	26.4
pronouns	50.3	76.5	26.2
non-pronouns	46.5	73.5	27.0
definite NPs	50.8	70.0	19.2
non-definite NPs	48.0	75.7	27.7
"this"	44.5	71.8	27.3
non-"this"	50.2	76.1	25.9

The FOCUS DISTANCE feature is based on a Euclidean distance metric, but this need not be the case; it is possible that coreference is more sensitive to movement in either the y- or x- dimension. Figure 1 helps us to explore this phenomenon: it is a contour plot comparing relative hand position (indicated by the position on the graph) to the

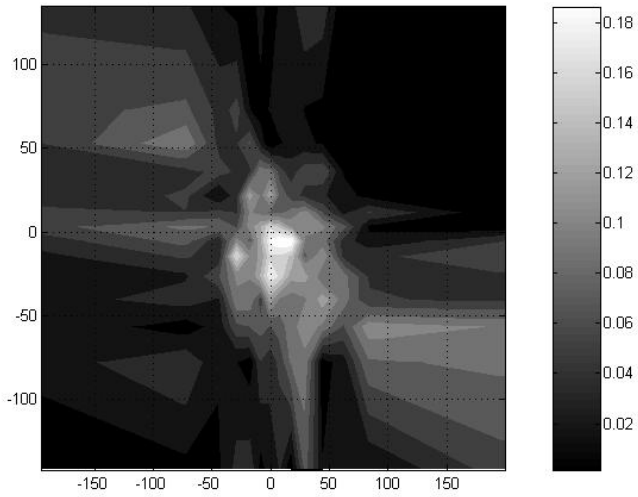

Fig. 1. A contour plot of coreference likelihood based on the relative position of the focus hands. The level of brightness indicates the likelihood of coreference, quantified in the color bar on the right.

likelihood of coreference (indicated by brightness). At (0,0), the hand positions during the candidate NPs are identical; at (0,50), the x-coordinates are identical but the y-coordinates of the focus hand at the time of the anaphoric noun phrase is 50 units higher than at the time of the antecedent NP.

The graph shows that the likelihood of coreference is greatest when the hand positions are identical (at the origin), and drops off nearly monotonically as one moves away from the center. However, there are some distortions that may be important. The drop-off in coreference likelihood appears to be less rapid in the y-dimension, suggesting the accuracy in the y-coordinate is less important than in the x-coordinate. This may be explained by the fact that the figures we used were taller than they were wide – the exact dimensions were 22 by 17 inches – or it may be a more general phenomenon. The decrease in coreference likelihood also appears to be less rapid as one moves diagonally from upper-left to lower-right. This may be an artifact of the fact that speakers more often stood to the left of the diagram, from the camera's perspective. From this position, it is easier to move the hand along this diagonal than from the upper-right to lower-left corners.

The choice of gesturing hand was also found to be related to coreference. As shown in Table 2, speakers were more likely to use the same hand in both gestures if the associated noun phrases were coreferent. They were less likely to use different hands, and they were more likely to gesture overall. These differences were found to be statistically significant ($p < .01, \chi^2 = 57.3, dof = 2$). Table 2 also shows how the relationship between hand choice and coreference varied according to the type of noun phrase. Hand choice was a significant feature for all types of NPs, except those beginning with the word "this" ($p = .12, \chi^2 = 4.28, dof = 2$).

Table 2. Hand choice and coreference. All figures are percentages.

	same	*different*	*no gesture*
overall			
coreferring	59.9	19.9	20.2
not coreferring	52.8	22.2	25.1
all	53.2	22.0	24.8
anaphor is a pronoun			
coreferring	58.9	18.6	22.6
not coreferring	50.5	21.0	28.5
all	51.2	20.8	28.0
anaphor is definite NP			
coreferring	57.8	25.6	16.6
not coreferring	54.7	22.2	23.1
all	54.9	22.5	22.7
anaphor begins with "this"			
coreferring	65.6	21.8	12.6
not coreferring	61.6	23.0	15.4
all	61.8	22.9	15.3

The overall rate of detected gestures varied across linguistic phenomena, shown in column 4. The fewest gestures occurring during pronouns, and the most occurring during NPs beginning with the word "this." It is somewhat unsettling that the greatest proportion of gestures occurred during such NPs, and yet our comparative features were not better at predicting coreference – in fact, the WHICH HAND feature was worse for these NPs than overall.

5.1 Non-comparative Features

Non-comparative features cannot directly measure whether two noun phrases are likely to corefer, as they can be applied only to a single instant in time. However, they can be used for at least two purposes: to determine whether a given noun phrase is likely to have *any* coreferent NPs, and to assess whether the comparative features will be useful. In other words, non-comparative features may tell us whether gesture is worth attending to.

The first question – the relationship between non-comparative features and anaphoricity – is addressed in Table 3. For a given noun phrase, its "children" are subsequent NPs that corefer to it; its "parents" are preceding NPs to which it corefers. Columns two and three list the mean feature values, conditioned on whether the NP has children. Columns five and six are conditioned on whether the NP has parents.

As indicated by the table, the FOCUS SPEED and PURPOSE features predict the presence of "children" NPs when they take on low values. Both features attempt to quantify whether the hand is being held in position during the associated NP, or if it is in the process of moving to another location. The presence of a hold during a noun phrase seems to predict that future NPs will refer back to the present noun phrase; perhaps

Table 3. Non-comparative features, with 95% confidence intervals

feature	children	no children	p	parents	no parents	p
FOCUS SPEED	$.0553 \pm .0056$	$.0688 \pm .010$	.05	$.0634 \pm .0068$	$.0557 \pm .0079$	
JITTER	$1.37 \pm .066$	$1.36 \pm .079$		$1.39 \pm .067$	$1.34 \pm .079$	
PURPOSE	$.0561 \pm .0066$	$.0866 \pm .020$	.01	$.0656 \pm .0093$	$.0705 \pm .017$	
SYNCHRONIZATION	$.0273 \pm .043$	$.0515 \pm .049$		$.0281 \pm .046$	$.0504 \pm .045$	
SCALED SYNCH	$.095 \pm .096$	$.085 \pm .12$		$.0969 \pm .11$	$.0826 \pm .10$	

speakers are more likely to produce gestural holds when describing concepts that they know are important.

In contrast, none of the features were capable of predicting whether a noun phrase had parents – previously uttered NPs to which it refers. This is somewhat disappointing, as we had hoped to find gesture cues indicating whether an entity was being referred to for the first time. As always, it is possible that this information is carried in gesture, and our feature set is simply insufficient to capture it.

Non-comparative features as meta-features. Finally, we consider the possibility that non-comparative features can serve as meta-features, telling us when to consider the comparative features. If so, we would expect to observe a non-zero correlation between useful meta-features and the discriminability of the comparative features. That is, the meta feature is helpful if for some values, it can alert us that the comparative feature is likely to be highly discriminable.

For a given noun phrase NP_j, the discriminability of the FOCUS DISTANCE feature can be measured by subtracting the average FOCUS DISTANCE to all coreferring noun phrases NP_i from the average FOCUS DISTANCE to all non-coreferring noun phrases:

$$\begin{aligned} cr(NP_j) &= \text{Set of all noun phrases coreferent with } NP_j\text{, not including } NP_j \text{ itself.} \\ \overline{cr}(NP_j) &= \text{Complement of } cr(NP_j)\text{, not including } NP_j \text{ itself.} \\ \text{FD}(NP_j, NP_i) &= \text{FOCUS DISTANCE between } NP_j \text{ and } NP_i\text{, as defined in Equation 1} \end{aligned}$$

$$d(NP_j) = \frac{1}{|\overline{cr}(NP_j)|} \sum_{\overline{cr}(NP_j)} \text{FD}(NP_j, NP_i) - \frac{1}{|cr(NP_j)|} \sum_{cr(NP_j)} \text{FD}(NP_j, NP_i) \quad (9)$$

For each NP_j, we can measure the correlation of the non-comparative features at NP_j with the discriminability $d(NP_j)$, to see whether the non-comparative features are indeed predictive of the discriminability of the FOCUS DISTANCE feature. The results are shown in Table 4. Significance values are computed using the Fisher transform.

Table 4 shows that low FOCUS SPEED and PURPOSE are indicative of a possibly useful gesture, which is expected, since they are indicative of a gestural hold. Additionally, the x-distance from the center of the body is also predictive of gesture discriminability. This reflects the fact that useful gestures typically refer to the diagram, and the speakers are usually mindful not to stand in front of it. The last two lines of the table show results for linear and interaction regression models, suggesting that a classifier built using these features could be strongly predictive of whether the current gesture is informative. This

would suggest a meta-learning system that could tell us when to pay attention to gesture and when to ignore it.

6 Gesture-Speech Synchronization

Given that gesture carries semantic content associated with speech, it is logical to ask how the speech and gesture are synchronized. Does the more informative part of gesture typically precede speech, follow it, or synchronize with it precisely? This is relevant from both an engineering and scientific perspective. To build systems that use gesture features, it is important to know the time window over which those features should be computed. From a psycholinguistic perspective, the synchronization of gesture and speech is thought to reveal important clues about how gesture and speech are produced and represented in the mind [7,8].

We compared the effectiveness of the FOCUS DISTANCE feature at assessing coreference identity, using discriminability as defined in Equation 9, evaluated at varying temporal offsets with respect to the start, midpoint, and end of the associated noun phrases. The results are shown in Figure 2. The optimal discriminability is found 108 milliseconds after the midpoint of the associated noun phrase. With respect to the speech onset, optimal discriminability is found 189 milliseconds after the speech onset. With respect to the end of the lexical affiliate, the optimal discriminability is found 81 milliseconds after the end point; while it is somewhat surprising to see the optimal discriminability after the end of the word, this may be the result of noise, as this graph is less sharply peaked than the other two.

Some of the existing psychology literature suggests that gesture strokes typically *precede* the associated speech [7,8]. Our data may appear to conflict with these results, but we argue that in fact they do not. The existing literature typically measures synchronization with respect to the *beginning* of the gesture and the beginning of speech. Note that the FOCUS DISTANCE feature only captures the semantics of deictic gestures, which McNeill defines as gestures that communicate meaning by hand position [2]. In particular, we believe that our results capture the beginning of the post-stroke "hold" phase of the gesture: exactly the moment at which the movement of the gesture ends, and the

Table 4. Non-comparative feature correlations with FOCUS DISTANCE discriminability

feature	r	**significance** (dof = 377)
FOCUS SPEED	-.169	$p < .01$
JITTER	.0261	
PURPOSE	-.1671	$p < .01$
SYNCHRONIZATION	-.0394	
SCALED SYNCHRONIZATION	-.0459	
Y-distance from bottom	.0317	
X-distance from body center	.215	$p < .01$
Regression, linear model	.288	$p < .01$
Regression, interaction model	.420	$p < .01$

hand rests at a semantically meaningful position in space. This analysis is supported by the finding that the FOCUS DISTANCE discriminability is negatively correlated with hand movement speed (see Table 4). At the onset of gesture motion, the hand is not yet at a semantically relevant location in space, and so the discriminability of the FOCUS DISTANCE feature cannot capture when motion begins. In future work, we hope to consider whether segmentation of hand motion into movement phases could be automated, facilitating this analysis.

7 Related Work

The psychology literature contains many close analyses of dialogue that attempt to identify the semantic contribution of gesture (e.g., [2,9]). This work has been instrumental in documenting the ways in which gesture and speech interact, identifying features of gesture (termed "catchments"), and showing how they relate to semantic phenomena. However, much of this analysis has been *post facto*, allowing psychologists to bring to bear human-level common-sense understanding to the interpretation of the gestures in dialogues. In contrast, we focus our analysis on coreference – rather than asking "what does this gesture *mean*?", we ask the simpler, yes/no question, "do these two gestures refer to the same thing?" Thus, we are able to systematize our treatment of semantics without having to create an ontology for the semantics of the domain. This, in combination with automatic extraction of gesture features through computer vision, clears the way for a predictive analysis without human intervention. We believe such an analysis provides a useful complement to the existing work that we have cited.

Another relevant area of research is in gesture generation, which has developed and exploited rich models of the relationships between gesture and semantics (e.g., [10]). However, there is an important difference between generation and recognition. Humans are capable of perhaps an infinite variety of gestural metaphors, but gesture generation need only model a limited subset of these metaphors to produce realistic gestures. To understand the gestures that occur in unconstrained human communication, a more complete understanding of gesture may be necessary.

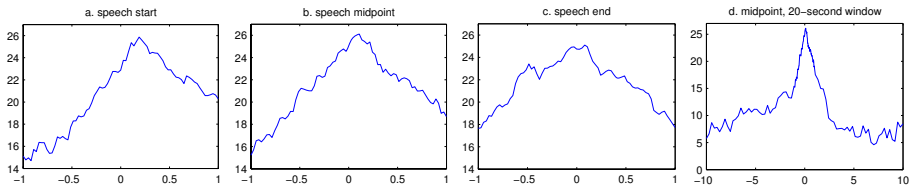

Fig. 2. Gesture-speech synchronization data. The y-axis is the discriminability of the FOCUS DISTANCE feature (Equation 9), and the x-axis is the offset in seconds. Part (d) shows a wider window than the other three plots. Parts (b) and (d) show the discriminability when the offset is computed from the speech midpoint. In parts (a) and (c), the offset is computed from the start and end points respectively.

8 Conclusion

This paper introduces a new computational methodology for addressing the relationship between gesture and semantics, avoiding time-consuming and difficult gesture annotation. We have found several interesting connections between gesture features and coreference, using a new corpus of dialogues involving diagrams of mechanical devices. In our data, the position of gestural holds appears to be the most important gesture feature, and other gesture features can help to determine when holds are occurring. In addition, an analysis of the timing of gesture/speech synchrony suggests that the most useful information for deictic gestures is located roughly 100 milliseconds after the midpoint of the lexical affiliate.

References

1. Grishman, R., Sundheim, B.: Design of the MUC-6 evaluation. In: Proceedings of the 6th Message Understanding Conference. (1995)
2. McNeill, D.: Hand and Mind. The University of Chicago Press (1992)
3. Adler, A., Eisenstein, J., Oltmans, M., Guttentag, L., Davis, R.: Building the design studio of the future. In: Making Pen-Based Interaction Intelligent and Natural, Menlo Park, California, AAAI Press (2004) 1–7
4. Huang, X., Alleva, F., Hwang, M.Y., Rosenfeld, R.: An overview of the Sphinx-II speech recognition system. In: Proceedings of ARPA Human Language Technology Workshop. (1993) 81–86
5. Deutscher, J., Blake, A., Reid, I.: Articulated body motion capture by annealed particle filtering. In: IEEE Conference on Computer Vision and Pattern Recognition. Volume 2. (2000) 126–133
6. Hirschman, L., Chinchor, N.: MUC-7 coreference task definition. In: Message Understanding Conference Proceedings. (1997)
7. Butterworth, B., Beattie, G.: Gesture and silence as indicators of planning in speech. In: Recent Advances in the Psychology of Language, Plenum Press (1978) 347–360
8. Morrel-Samuels, P., Krauss, R.M.: Word familiarity predicts temporal asynchrony of hand gestures and speech. Journal of Experimental Psychology: Learning, Memory and Cognition **18** (1992) 615–623
9. Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.F., Kirbas, C., McCullough, K.E., Ansari, R.: Multimodal human discourse: gesture and speech. ACM Transactions on Computer-Human Interaction (TOCHI) (2002) 171–193
10. Kopp, S., Tepper, P., Ferriman, K., Cassell, J.: Trading spaces: How humans and humanoids use speech and gesture to give directions. Spatial Cognition and Computation **In preparation** (2006)