A Salience-Based Approach to Gesture-Speech Alignment

Jacob Eisenstein and C. Mario Christoudias
MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street
Cambridge, MA 02139
{jacobe+cmch}@csail.mit.edu

Abstract

One of the first steps towards understanding natural multimodal language is aligning gesture and speech, so that the appropriate gestures ground referential pronouns in the speech. This paper presents a novel technique for gesture-speech alignment, inspired by salience-based approaches to anaphoric pronoun resolution. We use a hybrid between data-driven and knowledge-based mtehods: the basic structure is derived from a set of rules about gesture salience, but the salience weights themselves are learned from a corpus. Our system achieves 95% recall and precision on a corpus of transcriptions of unconstrained multimodal monologues, significantly outperforming a competitive baseline.

1 Introduction

In face to face communication, speakers frequently use gesture to supplement speech (Chovil, 1992), using the additional modality to provide unique, non-redundant information (McNeill, 1992). In the context of pen/speech user interfaces, Oviatt finds that "multimodal ... language is briefer, syntactically simpler, and less disfluent than users' unimodal speech." (Oviatt, 1999)

One of the simplest and most direct ways in which gesture can supplement verbal communication is by grounding references, usually through deixis. For example, it is impossible to extract the semantic content of the verbal utterance "I'll take this one" without an accompanying pointing gesture indicating the thing that is desired. The problem of *gesture-speech alignment* involves choosing the appropriate gesture to ground each verbal utterance. This paper describes a novel technique for this problem. We evaluate our system on a corpus of multimodal monologues with no fixed grammar or vocabulary.

1.1 Example

```
[This]_1 thing goes over [here]_2 so
that it goes back ...
--------------------
1. Deictic: Hand rests on latch mechanism
2. Iconic:  Hand draws trajectory from
            right to left
```

In this example, there are three verbal references. The word "this" refers to the latch mechanism, which is indicated by the rest position of the hand. "Here" refers to the endpoint of the trajectory indicated by the iconic gesture. "It" is an anaphoric reference to a noun phrase defined earlier in the sentence; there is no accompanying gesture. The word "that" does not act as a reference, although it could in other cases. Not every pronoun keyword (e.g., this, here, it, that, etc.) will act as a reference in all cases. In addition, there will be many gestures that do not resolve any keyword.

2 Related Work

This research draws mainly from two streams of related work. Researchers in human-computer interaction have worked towards developing multimodal user interfaces, which allow spoken and gestural input. These systems often feature powerful algorithms for fusing modalities; however, they also restrict communication to short grammatically-constrained commands over a very limited vocabulary. Since our goal is to handle more complex linguistic phenomena, these systems were of little help in the design of our algorithm. Conversely, we found that the problem of anaphora resolution faces a very similar set of challenges as gesture-speech alignment. We were able to apply techniques from anaphora resolution to gesture-speech alignment.

2.1 Multimodal User Interfaces

Discussion of multimodal user interfaces begins with the seminal "Put-That-There" system (Bolt, 1980), which allowed users to issue natural language commands and use

deictic hand gestures to resolve references from speech. Commands were subject to a strict grammar and alignment was straightforward: keywords created holes in the semantic frame, and temporally-aligned gestures filled the holes.

More recent systems have extended this approach somewhat. Johnston and Bangalore describe a multimodal parsing algorithm that is built using a 3-tape, finite state transducer (FST) (Johnston and Bangalore, 2000). The speech and gestures of each multimodal utterance are provided as input to an FST whose output is a semantic representation conveying the combined meaning. A similar system, based on a graph-matching algorithm, is described in (Chai et al., 2004). These systems perform mutual disambiguation, where each modality helps to correct errors in the others. However, both approaches restrict users to a predefined grammar and lexicon, and rely heavily on having a complete, formal ontology of the domain.

In (Kettebekov et al., 2002), a co-occurrence model relates the salient prosodic features of the speech (pitch variation and pause) to characteristic features of gesticulation (velocity and acceleration). The goal was to improve performance of gesture recognition, rather than to address the problem of alignment directly. Their approach also differs from ours in that they operate at the level of speech signals, rather than recognized words. Potentially, the two approaches could compliment each other in a unified system.

2.2 Anaphora Resolution

Anaphora resolution involves linking an anaphor to its corresponding antecedent in the same or previous sentence. In many cases, speech/gesture multimodal fusion works in a very similar way, with gestures grounding some of the same anaphoric pronouns (e.g., “this”, “that”, “here”).

One approach to anaphora resolution is to assign a salience value to each noun phrase that is a candidate for acting as a grounding referent, and then to choose the noun phrase with the greatest salience (Lappin and Leass, 1994). Mitkov showed that a salience-based approach can be applied across genres and without complex syntactic, semantic, and discourse analysis (Mitkov, 1998). Salience values are typically computed by applying linguistic knowledge; e.g., recent noun phrases are more salient, gender and number should agree, etc. This knowledge is applied to derive a salience value through the application of a set of predefined salience weights on each feature. Salience weights may be defined by hand, as in (Lappin and Leass, 1994), or learned from data (Mitkov et al., 2002).

Anaphora resolution and gesture-speech alignment are very similar problems. Both involve resolving ambiguous words which reference other parts of the utterance. In the case of anaphora resolution, pronomial references resolve to previously uttered noun phrases; in gesture-speech alignment, keywords are resolved by gestures, which usually precede the keyword. The salience-based approach works for anaphora resolution because the factors that contribute to noun-phrase salience are well understood. We define a parallel set of factors for evaluating the salience of gestures.

3 Our Approach

The most important goal of our system is the ability to handle natural, human-to-human language usage. This includes disfluencies and grammatically incorrect utterances, which become even more problematic when considering that the output of speech recognizers is far from perfect. Any approach that requires significant parsing or other grammatical analysis may be ill-suited to meet these goals.

Instead, we identify keywords that are likely to require gestural referents for resolution. Our goal is to produce an alignment – a set of *bindings* – that match at least some of the identified keywords with one or more gestures. There are several things that are known to contribute to the salience of candidate gesture-speech bindings:

- The relevant gesture is usually close in time to the keyword (Oviatt et al., 1997; Cohen et al., 2002)
- The gesture usually precedes the keyword (Oviatt et al., 1997).
- A one-to-one mapping is preferred. Multiple keywords rarely align with a single gesture, and multiple gestures almost never align with a single keyword (Eisenstein and Davis, 2003).
- Some types of gestures, such as deictic pointing gestures, are more likely to take part in keyword bindings. Other gestures (i.e., beats) do not carry this type of semantic content, and instead act to moderate turn taking or indicate emphasis. These gestures are unlikely to take part in keyword bindings (Cassell, 1998).
- Some keyword/gesture combinations may be particularly likely; for example, the keyword “this” and a deictic pointing gesture.

These rules mirror the salience weighting features employed by the anaphora resolution methods described in the previous section. We define a parameterizable penalty function that prefers alignments that adhere to as many of these rules as possible. Given a set of verbal utterances and gestures, we then try to find the set of bindings with the minimal penalty. This is essentially an optimization

approach, and we use the simplest possible optimization technique: greedy hill-climbing. Of course, given a set of penalties and the appropriate representation, any optimization technique could be applied. In the evaluation section, we discuss whether and how much our system would benefit from using a more sophisticated optimization technique. Later in this section, we formalize the problem and our proposed solution.

3.1 Leveraging Empirical Data

One of the advantages of the salience-based approach is that it enables the creation of a hybrid system that benefits both from our intuitions about multimodal communication and from a corpus of annotated data. The form of the salience metric, and the choice of features that factor into it, is governed by our knowledge about the way speech and gesture work. However, the penalty function also requires parameters that weigh the importance of each factor. These parameters can be crafted by hand if no corpus is available, but they can also be learned from data. By using knowledge about multimodal language to derive the form and features of the salience metric, and using a corpus to fine-tune the parameters of this metric, we can leverage the strengths of both knowledge-based and data-driven approaches.

4 Formalization

We define a multimodal transcript M to consist of a set of spoken utterances S and gestures G. S contains a set of references R that must be ground by a gestural referent. We define a binding, $b \in B$, as a tuple relating a gesture, $g \in G$, to a corresponding speech reference, $r \in R$. Provided G and R, the set B enumerates all possible bindings between them. Formally, each gesture, reference, and binding are defined as

$$\begin{aligned} g &= \langle t_s^g, t_e^g, G_{type} \rangle \\ r &= \langle t_s^r, t_e^r, w \rangle \\ b &= \langle g, r \rangle \end{aligned} \tag{1}$$

where t_s, t_e describe the start and ending time of a gesture or reference, $w \in S$ is the word corresponding to r, and G_{type} is the type of gesture (e.g. *deictic* or *trajectory*).

An alternative, useful description of the set B is as the function $b(g)$ which returns for each gesture a set of corresponding references. This function is defined as

$$b(g) = \{r | \langle g, r \rangle \in B\} \tag{2}$$

4.1 Rules

In this section we provide the analytical form for the penalty functions of Section 3. We have designed these functions to penalize bindings that violate the preferences that model our intuitions about the relationship between speech and gesture. We begin by presenting the analytical form for the binding penalty function, ψ_b.

It is most often the case that verbal references closely follow the gestures that they refer to; the verbal reference rarely precedes the gesture. To reflect this knowledge, we parameterize ψ_b using a time gap penalty, α_{tg}, and a wrong order penalty, α_{wo} as follows,

$$\psi_b(b) = \alpha_{tg} w_{tg}(b) + \alpha_{wo} w_{wo}(b) \tag{3}$$

where,

$$w_{wo}(b) = \begin{cases} 0 & t_s^r \geq t_s^g \\ 1 & t_s^r < t_s^g \end{cases}$$

and $w_{tg} = |t_s^r - t_s^g|$

In addition to temporal agreement, specific words or parts-of-speech have varying affinities for different types of gestures. We incorporate these penalties into ψ_b by introducing a binding agreement penalty, $\alpha(b)$, as follows:

$$\psi_b(b) = \alpha_{tg} w_{wo}(b) + \alpha(b) \tag{4}$$

The remaining penalty functions model binding fertility. Specifically, we assign a penalty for each unassigned gesture and reference, $\psi_g(g)$ and $\psi_r(r)$ respectively, that reflect our desire for the algorithm to produce bindings. Certain gesture types (e.g., deictics) are much more likely to participate in bindings than others (e.g., beats). An unassigned gesture penalty is associated with each gesture type, given by $\psi_g(g)$. Similarly, we expect references to have a likelihood of being bound that is conditioned on their word or part-of-speech tag. However, we currently handle all keywords in the same way, with a constant penalty $\psi_r(r)$ for unassigned keywords.

4.2 Minimization Algorithm

Given G and R we wish to find a $B^* \subseteq B$ that minimizes the penalty function $\psi(B, G, R)$:

$$B^* = \arg\min_{\hat{B}} \psi(\hat{B}, G, R) \tag{5}$$

Using the penalty functions of Section 4.1 $\psi(\hat{B}, G, R)$ is defined as,

$$\psi(\hat{B}, G, R) = \sum_{b \in \hat{B}} \psi_b(b) + \psi_g(G_a) + \psi_r(R_a) \tag{6}$$

where

$$\begin{aligned} G_a &= \{g | b(g) = \emptyset\} \\ R_a &= \{r | b(r) = \emptyset\} \end{aligned} .$$

Although there are numerous optimization techniques that may be applied to minimize Equation 5, we have chosen to implement a naive gradient decent algorithm presented below as Algorithm 1. Observing the problem, note we could have initialized $B^* = B$; in other

Algorithm 1 Gradient Descent

Initialize $B^* = \emptyset$ and $B' = B$
repeat
  Let b_0 be the first element in B'
  $\delta_{max} = \psi(B^*, G, R) - \psi(\{B^*, b_0\}, G, R)$
  $b_{max} = b_0$
  for all $b \in B'$, $b \neq b_0$ **do**
    $\delta = \psi(B^*, G, R) - \psi(\{B^*, b\}, G, R)$
    if $\delta > \delta_{max}$ **then**
      $b_{max} = b$
      $\delta_{max} = \delta$
    end if
  end for
  if $\delta_{max} > 0$ **then**
    $B^* = \{B^*, b_{max}\}$
    $B' = B' - b_{max}$
  end if
  Convergence test: is $\delta_{max} < limit$?
until convergence

words, start off with all possible bindings, and gradually prune away the bad ones. But it seems likely that $|B^*| \leq min(|R|, |G|)$; thus, starting from the empty set will converge faster. The time complexity of this algorithm is given by $O(|B^*||B|)$. Since $|B| = |G||R|$, and assuming $|B^*| \propto |G| \propto |R|$, this simplifies to $O(|B^*|^3)$, cubic in the number of bindings returned.

4.3 Learning Parameters

We explored a number of different techniques for finding the parameters of the penalty function: setting them by hand, gradient descent, simulated annealing, and a genetic algorithm. A detailed comparison of the results with each approach is beyond the scope of this paper, but the genetic algorithm outperformed the other approaches in both accuracy and rate of convergence.

The genome representation consisted of a thirteen bit string for each penalty parameter; three bits were used for the exponent, and the remaining ten were used for the base. Parameters were allowed to vary from 10^{-4} to 10^3. Since there were eleven parameters, the overall string length was 143. A population size of 200 was used, and training proceeded for 50 generations. Single-point crossover was applied at a rate of 90%, and the mutation rate was set to 3% per bit. Tournament selection was used rather than straightforward fitness-based selection (Goldberg, 1989).

5 Evaluation

We evaluated our system by testing its performance on a set of 26 transcriptions of unconstrained human-to-human communication, from nine different speakers (Eisenstein and Davis, 2003). Of the four women and five men who participated, eight were right-handed, and one was a non-native English speaker. The participants ranged in age from 22 to 28. All had extensive computer experience, but none had any experience in the task domain, which required explaining the behavior of simple mechanical devices.

	Baseline	**Training**	**Test**
Recall	84.2%	94.6%	95.1%
σ	n/a	1.2%	5.1%
Precision	82.8%	94.5%	94.5%
σ	n/a	1.2%	5.0%

Table 1: Performance of our system versus a baseline

The participants were presented with three conditions, each of which involved describing the operation of a mechanical device based on a computer simulation. The conditions were shown in order of increasing complexity, as measured by the number of moving parts: a latchbox, a piston, and a pinball machine. Monologues ranged in duration from 15 to 90 seconds; the number of gestures used ranged from six to 58. In total, 574 gesture phrases were transcribed, of which 239 participated in gesture-speech bindings.

In explaining the devices, the participants were allowed – but not instructed – to refer to a predrawn diagram that corresponded to the simulation. Vocabulary, grammar, and gesture were not constrained in any way. The monologues were videotaped, transcribed, and annotated by hand. No gesture or speech recognition was performed. The decision to use transcriptions rather than speech and gesture recognizers will be discussed in detail below.

5.1 Empirical Results

We averaged results over ten experiments, in which 20% of the data was selected randomly and held out as a test set. Entire transcripts were held out, rather than parts of each transcript. This was necessary because the system considers the entire transcript holistically when choosing an alignment.

For a baseline, we evaluated the performance of choosing the temporally closest gesture to each keyword. While simplistic, this approach is used in several implemented multimodal user interfaces (Bolt, 1980; Koons et al., 1993). Kettebekov and Sharma even reported that 93.7% of gesture phrases were "temporally aligned" with the semantically associated keyword in their corpus (Kettebekov and Sharma, 2001). Our results with this baseline were somewhat lower, for reasons discussed below.

Table 1 shows the results of our system and the baseline on our corpus. Our system significantly outperforms

the baseline on both recall and precision on this corpus ($p < 0.05$, two-tailed). Precision and recall differ slightly because there are keywords that do not bind to any gesture. Our system does not assume a one-to-one mapping between keywords and gestures, and will refuse to bind some keywords if there is no gesture with a high enough salience. One benefit of our penalty-based approach is that it allows us to easily trade off between recall and precision. Reducing the penalties for unassigned gestures and keywords will cause the system to create fewer alignments, increasing precision and decreasing recall. This could be useful in a system where mistaken gesture/speech alignments are particularly undesirable. By increasing these same penalties, the opposite effect can also be achieved.

Both systems perform worse on longer monologues. On the top quartile of monologues by length (measured in number of keywords), the recall of the baseline system falls to 75%, and the recall of our system falls to 90%. For the baseline system, we found a correlation of -0.55 (df = 23, $p < 0.01$) between F-measure and monologue length.

This may help to explain why Kettebekov and Sharma found such success with the baseline algorithm. The multimodal utterances in their corpus consisted of relatively short commands. The longer monologues in our corpus tended to be more grammatically complex and included more disfluency. Consequently, alignment was more difficult, and a relatively naïve strategy, such as the baseline algorithm, was less effective.

6 Discussion

To our knowledge, very few multimodal understanding systems have been evaluated using natural, unconstrained speech and gesture. One exception is (Quek et al., 2002), which describes a system that extracts discourse structure from gesture on a corpus of unconstrained human-to-human communication; however, no quantitative analysis is provided. Of the systems that are more relevant to the specific problem of gesture-speech alignment (Cohen et al., 1997; Johnston and Bangalore, 2000; Kettebekov and Sharma, 2001), evaluation is always conducted from an HCI perspective, in which participants act as users of a computer system and communicate in short, grammatically-constrained multimodal commands. As shown in Section 5.1, such commands are significantly easier to align than the natural multimodal communication found in our corpus.

6.1 The Corpus

A number of considerations went into gathering this corpus.[1] One of our goals was to minimize the use of discourse-related "beat" gestures, so as to better focus on the deictic and iconic gestures that are more closely related to the content of the speech; that is why we focused on monologues rather than dialogues. We also wanted the corpus to be relevant to the HCI community. That is why we provided a diagram to gesture at, which we believe serves a similar function to a computer display, providing reference points for deictic gestures. We used a predrawn diagram – rather than letting participants draw the diagram themselves – because interleaved speech, gesture, and sketching is a much more complicated problem, to be addressed only after bimodal speech-gesture communication is better understood.

For a number of reasons, we decided to focus on *transcriptions* of speech and gesture, rather than using speech and gesture recognition systems. Foremost is that we wanted the language in our corpus to be as natural as possible; in particular, we wanted to avoid restricting speakers to a finite list of gestures. Building a recognizer that could handle such unconstrained gesture would be a substantial undertaking and an important research contribution in its own right. However, we are sensitive to the concern that our system should scale to handle possibly erroneous recognition data. There are three relevant classes of errors that our system may need to handle: speech recognition, gesture recognition, and gesture segmentation.

- **Speech Recognition Errors**
  The speech recognizer could fail to recognize a keyword; in this case, a binding would simply not be created. If the speech recognizer misrecognized a non-keyword as a keyword, a spurious binding might be created. However, since our system does not require that all keywords have bindings, we feel that our approach is likely to degrade gracefully in the face of this type of error.

- **Gesture Recognition Errors**
  This type of error would imply a gestural misclassification, e.g., classifying a deictic pointing gesture as an iconic. Again, we feel that a salience-based system will degrade gracefully with this type of error, since there are no hard requirements on the type of gesture for forming a binding. In contrast, a system that required, say, a deictic gesture to accompany a certain type of command would be very sensitive to a gesture misclassification.

[1] We also considered using the recently released FORM2 corpus from the Linguistic Data Consortium. However, this corpus is presently more focused on the kinematics of hand and upper body movement, rather than on higher-level linguistic information relating to gestures and speech.

- **Gesture Segmentation Errors**
  Gesture segmentation errors are probably the most dangerous, since this could involve incorrectly grouping two separate gestures into a single gesture, or vice versa. It seems that this type of error would be problematic for any approach, and we have no reason to believe that our salience-based approach would fare differently from any other approach.

6.2 Success Cases

Our system outperformed the baseline by more than 10%. There were several types of phenomena that the baseline failed to handle. In this corpus, each gesture precedes the semantically-associated keyword 85% of the time. Guided by this fact, we first created a baseline system that selected the nearest preceding gesture for each keyword; clearly, the maximum performance for such a baseline is 85%. Slightly better results were achieved by simply choosing the nearest gesture regardless of whether it precedes the keyword; this is the baseline shown in Table 1. However, this baseline incorrectly bound several cataphoric gestures. The best strategy is to accept just a few cataphoric gestures in unusual circumstances, but a naïve baseline approach is unable to do this.

Most of the other baseline errors came about when the mapping from gesture to speech was not one-to-one. For example, in the utterance "this piece here," the two keywords actually refer to a single deictic gesture. In the salience-based approach, the two keywords were correctly bound to a single gesture, but the baseline insisted on finding two gestures. The baseline similarly mishandled situations where a keyword was used without referring to any gesture.

6.3 Failure Cases

Although the recall and precision of our system neared 95%, investigating the causes of error could suggest potential improvements. We were particularly interested in errors on the training set, where overfitting could not be blamed. This section describes two sources of error, and suggests some potential improvements.

6.3.1 Disfluencies

We adopted a keyword-based approach so that our system would be more robust to disfluency than alternative approaches that depended on parsing. While we were able to handle many instances of disfluent speech, we found that disfluencies occasionally disturbed the usual relationship between gesture and speech. For example, consider the following utterance:

It has this... this spinning thing...

Our system attempted to bind gestures to each occurrence of "this", and ended up binding each reference to a different gesture. Moreover, *both* references were bound incorrectly. The relevant gesture in this case occurs after both references. This is an uncommon phenomenon, and as such, was penalized highly. However, anecdotally it appears that the presence of a disfluency makes this phenomenon more likely. A disfluency is frequently accompanied by an abortive gesture, followed by the full gesture occurring somewhat later than the spoken reference. It is possible that a system that could detect disfluency in the speech transcript could account for this phenomenon.

6.3.2 Greedy Search

Our system applies a greedy hill-climbing optimization to minimize the penalty. While this greedy optimization performs surprisingly well, we were able to identify a few cases of errors that were caused by the greedy nature of our optimization, e.g.

...once it hits this, this thing is blocked.

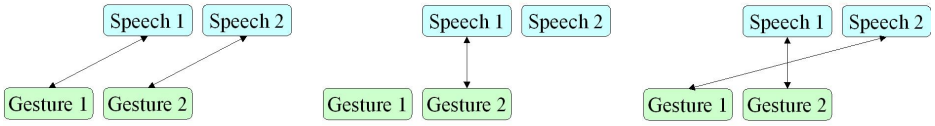
In this example, the two references are right next to each other. The relevant gestures are also very near each other. The ideal bindings are shown in Figure 1a. The earlier "this" is considered first, but from the system's perspective, the best possible binding is the second gesture, since it overlaps almost completely with the spoken utterance (Figure 1b). However, once the second gesture is bound to the first reference, it is removed from the list of unassigned gestures. Thus, if the second gesture were also bound to the second utterance, the penalty would still be relatively high. Even though the earlier gesture is farther away from the second reference, it is still on the list of unassigned gestures, and the system can reduce the overall penalty considerably by binding it. The system ends up crisscrossing, and binding the earlier gesture to the later reference, and vice versa (Figure 1c).

7 Future Work

The errors discussed in the previous section suggest some potential improvements to our system. In this section, we describe four possible avenues of future work: dynamic programming, deeper syntactic analysis, other anaphora resolution techniques, and user adaptation.

7.1 Dynamic Programming

Algorithm 1 provides only an approximate solution to Equation 5. As demonstrated in Section 6.3.2, the greedy choice is not always optimal. Using dynamic programming, an exhaustive search of the space of bindings can be performed within polynomial time.



(a) (b) (c)

Figure 1: The greedy binding problem. (a) The correct binding, (b) the greedy binding, (c) the result.

We define $m[i,j]$ to be the penalty of the optimal subset $B^* \subset \{b_i, ..., b_j\} \in B$, $i \leq j$. $m[i,j]$ is implemented as a $k \times k$ lookup table, where $k = |B| = |G||R|$. Each entry of this table is recursively defined by preceding table entries. Specifically, $m[i,j]$ is computed by performing exhaustive search on its subsets of bindings. Using this lookup table, an optimal solution to Equation 5 is therefore found as $\psi(B^*, G, R) = m[1,k]$. Again assuming $|B^*| \propto |G| \propto |R|$, the size of the lookup table is given by $O(|B^*|^4)$. Thus, it is possible to find the globally optimal set of bindings, by moving from an $O(n^3)$ algorithm to $O(n^4)$. The precise definition of a recurrence relation for $m[i,j]$ and a proof of correctness will be described in a future publication.

7.2 Syntactic Analysis

One obvious possibility for improvement would be to include more sophisticated syntactic information beyond keyword spotting. However, we require that our system remain robust to disfluency and recognition errors. Part of speech tagging is a robust method of syntactic analysis which could allow us to refine the penalty function depending on the usage case. Consider that there at least three relevant uses of the keyword "this."

1. *This* movie is better than A.I.
2. *This* is the bicycle ridden by E.T.
3. The wheel moves like *this*.

When "this" is followed by a noun (case 1), a deictic gesture is likely, although not strictly necessary. But when "this" is followed by a verb (case 2), a deictic gesture is usually crucial for understanding the sentence. Thus, the penalty for not assigning this keyword should be very high. Finally, in the third case, when the keyword follows a preposition, a trajectory gesture is more likely, and the penalty for any such binding should be lowered.

7.3 Other Anaphora Resolution Techniques

We have based this research on salience values, which is just one of several possible alternative approaches to anaphora resolution. One such alternative is the use of *constraints*: rules that eliminate candidates from the list of possible antecedents (Rich and Luperfoy, 1988). An example of a constraint in anaphora resolution is a rule requiring the elimination of all candidates that disagree in gender or number with the referential pronoun. Constraints may be used in combination with a salience metric, to prune away unlikely choices before searching. The advantage is that enforcing constraints could be substantially less computationally expensive than searching through the space of all possible bindings for the one with the highest salience. One possible future project would be to develop a set of constraints for speech-gesture alignment, and investigate the effect of these constraints on both accuracy and speed.

Ge, Hale, and Charniak propose a data-driven approach to anaphora resolution (Ge et al., 1998). For a given pronoun, their system can compute a probability for each candidate antecedent. Their approach of seeking to maximize this probability is similar to the salience-maximizing approach that we have described. However, instead of using a parametric salience function, they learn a set of conditional probability distributions directly from the data. If this approach could be applied to gesture-speech alignment, it would be advantageous because the binding probabilities could be combined with the output of probabilistic recognizers to produce a pipeline architecture, similar to that proposed in (Wu et al., 1999). Such an architecture would provide multimodal disambiguation, where the errors of each component are corrected by other components.

7.4 Multimodal Adaptation

Speakers have remarkably entrenched multimodal communication patterns, with some users overlapping gesture and speech, and others using each modality sequentially (Oviatt et al., 1997). Moreover, these multimodal integration patterns do not seem to be malleable, suggesting that multimodal user interfaces should adapt to the user's tendencies. We have already shown how the weights of the salience metric can adapt for optimal performance against a corpus of user data; this approach could also be extended to adapt over time to an individual user.

8 Conclusions

This work represents one of the first efforts at aligning gesture and speech on a corpus of natural multimodal communication. Using greedy optimization and only a minimum of linguistic processing, we significantly outperform a competitive baseline, which has actually been implemented in existing multimodal user interfaces. Our approach is shown to be robust to spoken English, even with a high level of disfluency. By blending some of the benefits of empirical and knowledge-based approaches, our system can learn from a large corpus of data, but degrades gracefully when limited data is available.

Obviously, alignment is only one small component of a comprehensive system for recognizing and understanding multimodal communication. Putting aside the issue of gesture recognition, there is still the problem of deriving semantic information from aligned speech-gesture units. The solutions to this problem will likely have to be specially tailored to the application domain. While our evaluation indicates that our approach achieves what appears to be a high level of accuracy, the true test will be whether our system can actually support semantic information extraction from multimodal data. Only the construction of such a comprehensive end-to-end system will reveal whether the algorithm and features that we have chosen are sufficient, or whether a more sophisticated approach is required.

Acknowledgements

We thank Robert Berwick, Michael Collins, Trevor Darrell, Randall Davis, Tracy Hammond, Sanshzar Kettebekov, Özlem Uzuner, and the anonymous reviewers for their helpful comments on this paper.

References

Richard A. Bolt. 1980. Put-That-There: Voice and gesture at the graphics interface. *Computer Graphics*, 14(3):262–270.

Justine Cassell. 1998. A framework for gesture generation and interpretation. In *Computer Vision in Human-Machine Interaction*, pages 191–215. Cambridge University Press.

Joyce Y. Chai, Pengyu Hong, , and Michelle X. Zhou. 2004. A probabilistic approach to reference resolution in multimodal user interfaces. In *Proceedings of 2004 International Conference on Intelligent User Intefaces (IUI'04)*, pages 70–77.

Nicole Chovil. 1992. Discourse-oriented facial displays in conversation. *Research on Language and Social Interaction*, 25:163–194.

Philip R. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen, and J. Clow. 1997. Quickset: Multimodal interaction for distributed applications. In *ACM Multimedia'97*, pages 31–40. ACM Press.

Philip R. Cohen, Rachel Coulston, and Kelly Krout. 2002. Multimodal interaction during multiparty dialogues: Initial results. In *IEEE Conference on Multimodal Interfaces*.

Jacob Eisenstein and Randall Davis. 2003. Natural gesture in descriptive monologues. In *UIST'03 Supplemental Proceedings*, pages 69–70.

Niyu Ge, John Hale, and Eugene Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 161–171.

David E. Goldberg. 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley.

Michael Johnston and Srinivas Bangalore. 2000. Finite-state multimodal parsing and understanding. In *Proceedings of COLING-2000*. ICCL.

Sanshzar Kettebekov and Rajeev Sharma. 2001. Toward natural gesture/speech control of a large display. In *Engineering for Human-Computer Interaction (EHCI'01). Lecture Notes in Computer Science*. Springer Verlag.

Sanshzar Kettebekov, Mohammed Yeasin, and Rajeev Sharma. 2002. Prosody based co-analysis for continuous recognition of coverbal gestures. In *International Conference on Multimodal Interfaces (ICMI'02)*, pages 161–166, Pittsburgh, USA.

David B. Koons, Carlton J. Sparrell, and Kristinn R. Thorisson. 1993. Integrating simultaneous input from speech, gaze, and hand gestures. In *Intelligent Multimedia Interfaces*, pages 257–276. AAAI Press.

Shalom Lappin and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

David McNeill. 1992. *Hand and Mind*. The University of Chicago Press.

Ruslan Mitkov, Richard Evans, and Constantin Orăsan. 2002. A new, fully automatic version of mitkov's knowledge-poor pronoun resolution method. In *Intelligent Text Processing and Computational Linguistics (CICLing'02)*, Mexico City, Mexico, February, 17 – 23.

Ruslan Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *COLING-ACL*, pages 869–875.

Sharon L. Oviatt, Antonella DeAngeli, and Karen Kuhn. 1997. Integration and synchronization of input modes during multimodal human-computer interaction. In *Human Factors in Computing Systems (CHI'97)*, pages 415–422. ACM Press.

Sharon L. Oviatt. 1999. Ten myths of multimodal interaction. *Communications of the ACM*, 42(11):74–81.

Francis Quek, David McNeill, Robert Bryll, Susan Duncan, Xin-Feng Ma, Cemil Kirbas, Karl E. McCullough, and Rashid Ansari. 2002. Multimodal human discourse: gesture and speech. *Transactions on Computer-Human Interaction*, 9(3):171–193.

Elaine Rich and Susann Luperfoy. 1988. An architecture for anaphora resolution. In *Proceedings of the Second Conference on Applied Natural Language Processing (ANLP-2)*, pages 18–24, Texas, USA.

Lizhong Wu, Sharon L. Oviatt, and Philip R. Cohen. 1999. Multimodal integration - a statistical view. *IEEE Transactions on Multimedia*, 1(4):334–341.