Gesture in Automatic Discourse Processing

by

Jacob Eisenstein

Submitted to the Department of Electrical Engineering and Computer Science
on May 2, 2008, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Computers cannot fully understand spoken language without access to the wide range of modalities that accompany speech. This thesis addresses the particularly expressive modality of hand gesture, and focuses on building structured statistical models at the intersection of speech, vision, and meaning.

My approach is distinguished in two key respects. First, gestural patterns are leveraged to discover parallel structures in the meaning of the associated speech. This differs from prior work that attempted to interpret individual gestures directly, an approach that was prone to a lack of generality across speakers. Second, I present novel, structured statistical models for multimodal language processing, which enable learning about gesture in its linguistic context, rather than in the abstract.

These ideas find successful application in a variety of language processing tasks: resolving ambiguous noun phrases, segmenting speech into topics, and producing keyframe summaries of spoken language. In all three cases, the addition of gestural features – extracted automatically from video – yields significantly improved performance over a state-of-the-art text-only alternative. This marks the first demonstration that hand gesture improves automatic discourse processing.

Thesis Supervisor: Regina Barzilay
Title: Associate Professor

Thesis Supervisor: Randall Davis
Title: Professor

Words are but vague shadows of the volumes we mean.

Theodore Dreiser

1

Introduction

Speech is almost always accompanied by a range of other behaviors, including movements of the face, body, and hands (Rimé & Schiaratura, 1991). Of all these co-speech behaviors, hand gestures appear to be especially reflective of the speaker's underlying meaning (e.g., Figure 1-1). Many psychologists hypothesize that gesture is an integral part of spoken communication, helping to convey crucial semantic content (McNeill, 1992). If so, it is natural to ask whether automatic natural language processing systems can better understand speech by incorporating hand gestures. Can hand gestures fill in the gaps when words give only a "vague shadow" of the intended meaning?

Attempts to incorporate hand gestures in automatic language processing have been stymied by a number of practical and conceptual challenges. Gesture conveys meaning through a direct, visual medium – quite unlike speech, which is fundamen-

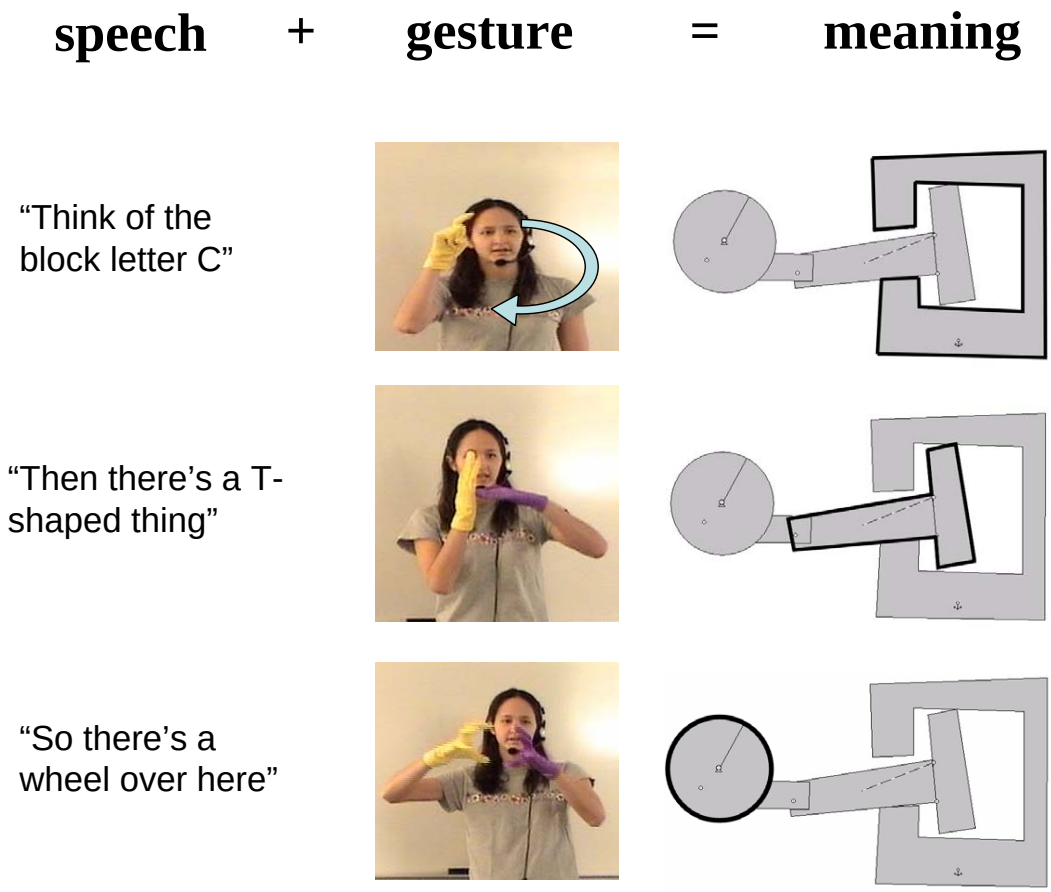


Figure 1-1: In this example, the speaker is describing the behavior of a piston, having seen an animation of the diagram on the right. Each line shows an excerpt of her speech and gesture, and highlights the relevant portion of the diagram. The speech expresses the spatial arrangement awkwardly, using metaphors to letters of the alphabet, while the gesture naturally provides a visual representation.

tally symbolic (Kendon, 2004). Thus, gesture does not easily lend itself to a compact, formal representation. The majority of co-speech gestures have neither predefined nor intrinsic meaning; rather, they are interpretable only in the surrounding linguistic context. Gesture is commonly believed to be highly idiosyncratic, so that its meaning may also vary widely by speaker. Finally, communicative gestures are only a fraction of the total set of hand motions that occur during speech. Automatic systems must be able to disattend other movements that may be semantically meaningless, such as when the speaker adjusts her glasses or hair.

To address these challenges, this dissertation offers two main ideas. First, I focus on identifying patterns *between* gestures, which are leveraged to discover parallel patterns in the discourse structure. Unlike previous research (e.g., Chen, Liu, Harper, & Shriberg, 2004; Cassell, Nakano, Bickmore, Sidner, & Rich, 2001), I do not attempt to assess the semantic or pragmatic contribution of individual gestures or movements. Speaker-specific idiosyncrasies and the modulating effects of linguistic context may make the form of individual gestures difficult to interpret directly. But even if individual gestures cannot be decoded, the relationships between gestures may be comprehensible.

As an example, Figure 1-2 shows automatically-extracted hand trajectories for two short gestures that occur roughly 30 seconds apart. It is hard to imagine inferring any meaning from these two trajectories when taken in isolation, but it is clear that the path of motion in the right hand is repeated quite closely. This repetition can serve as a clue for linguistic analysis. In this case, the second gesture (shown in the right panel) is accompanied by the ambiguous anaphoric pronoun "it," which refers back to the noun phrase "this thing," uttered during the performance of the first gesture (in the left panel). Recognizing the similarity of this pair of gestures can facilitate linguistic analysis even when the meaning of each individual gesture is unknown.

The second key idea is to build models of gesture without relying on gestural annotations. From a practical standpoint, annotating the form of gesture for a corpus of any reasonable size would be extremely time-consuming. Moreover, no annotation scheme formal enough for computational analysis has yet been shown to be sufficiently

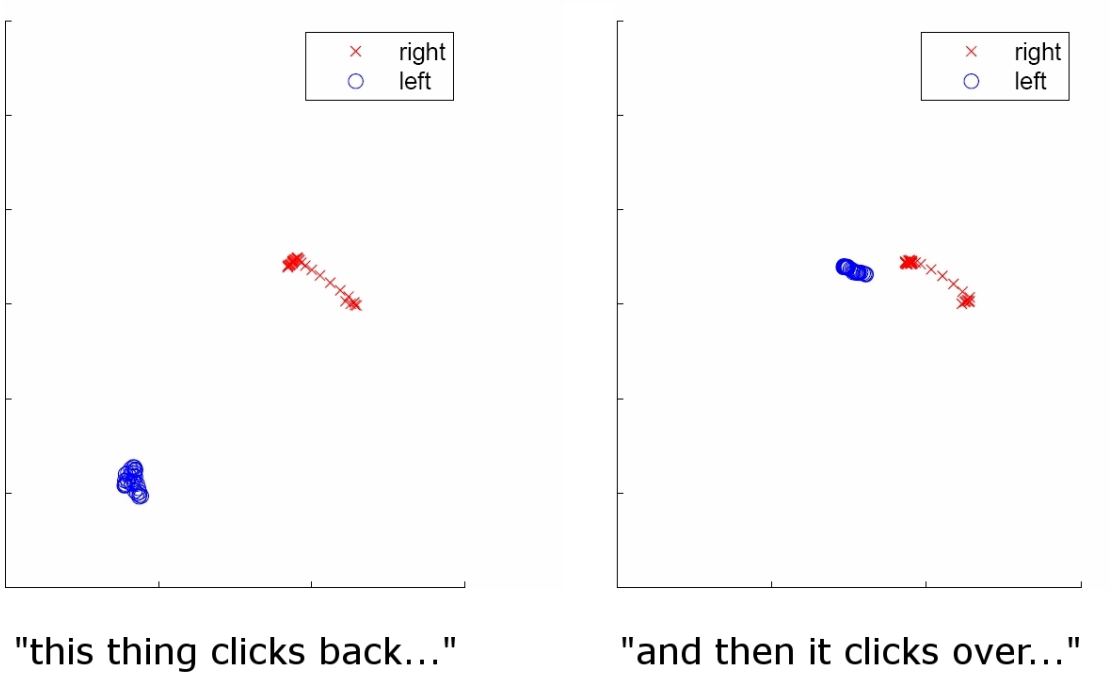


Figure 1-2: An example of two automatically extracted hand gestures. In each gesture, the left hand (blue) is held still while the right (red) moves up and to the left. The similarity of the motion of the right hand suggests a semantic relationship in the speech that accompanies each gesture: in this case, the noun phrase "this thing" and "it" refer to the same semantic entity.

flexible to describe all relevant aspects of a gesture's form. On a conceptual level, our ultimate goal is not to describe characteristics of gesture but to solve language processing problems that are critical to end-user applications. Thus, it seems best to learn directly from linguistic annotations whenever possible.

I avoid the need for gestural annotation by building custom statistical learning models that explicitly encode the relationship between gesture and speech. Such models are capable of learning about gesture by exploiting features of the language. This idea is applied in two ways: by leveraging linguistic annotations, and in a completely unsupervised framework. In both cases, gesture and verbal features are combined in a single joint model, maximizing the ability of each modality to disambiguate the other when necessary.

These two strategies – focusing on patterns between gestures, and learning models of gesture in the context of language – constitute the core technical innovations of this thesis. I demonstrate the applicability of these ideas to discourse processing on both local and global levels.

1.1 Gesture and Local Discourse Structure

In the previous section, the gestural trajectories shown in Figure 1-2 were used as an example of how gesture can help to disambiguate the relationship between the accompanying noun phrases. The problem of determining whether a pair of noun phrases refer to the same semantic entity is called *coreference resolution*, and may be considered a local-scale discourse phenomenon. Chapter 4 demonstrates that the similarity of the gestures accompanying a pair of noun phrases can help to predict whether they corefer.

To obtain maximum leverage from gestural similarity, is important to ensure that the hand movements being compared are indeed meaningful gestures. For any given hand motion, we may assess its *salience* – a measure of whether it is likely to be communicative. Gesture similarity and salience are learned using a novel architecture called *conditional modality fusion* – an application of hidden-variable conditional

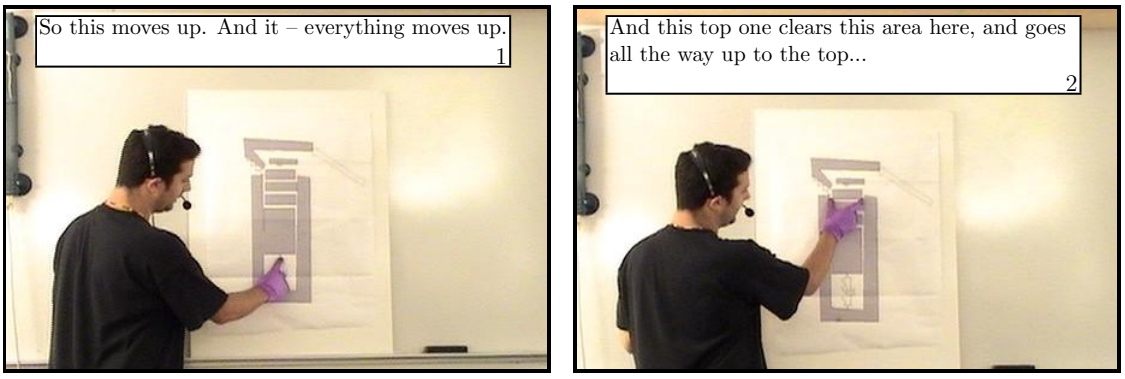


Figure 1-3: Two frames from a comic book summary generated by the system described in Chapter 4

random fields (Quattoni, Wang, Morency, Collins, & Darrell, 2007). This model operates without labels for gesture similarity or salience, learning directly from coreference annotations. The resulting system resolves noun phrases more accurately than a state-of-the-art baseline that uses only verbal features. Moreover, modeling gesture salience substantially increases the predictive power of gesture similarity information.

Gesture salience is then applied to another local discourse processing task: extracting keyframe summaries from video. The model of gesture salience learned on coreference is transferred directly to the keyframe summary task, again without any labeled data for summarization. The resulting system produces "comic books" in which the transcript is augmented with keyframes showing salient gestures (Figure 1-3). These comic books cohere well with human-generated summaries, outperforming state-of-the-art unsupervised keyframe extraction baselines.

1.2 Gesture and Global Discourse Structure

Gesture similarity is a property of pairs of gestures; in Chapter 5, this idea is extended to larger sets of gestures, under the name of *gestural cohesion*. This term draws a deliberate parallel to the well-known phenomenon of lexical cohesion, which measures the self-consistency of word use within a discourse segment (Halliday & Hasan, 1976). Lexical cohesion has been found to be an effective feature for high-level discourse

analysis, particularly for the task of topic segmentation: dividing a text into topically-distinct segments (Hearst, 1994). Chapter 5 investigates whether gestural cohesion can be used in the same way.

Lexical cohesion is an effective feature for discourse analysis because word choice is one way that semantics is conveyed in text and language. Thus, a change in the distribution of lexical items is predictive of a change in the intended meaning. Given the hypothesis that meaning is also conveyed via gesture, it seems reasonable that the distribution of gestural forms may predict segmentation in the same way. For each dialogue, a "lexicon" of gestural forms is acquired through unsupervised clustering.[1] The observed words and gestural forms are then combined in a novel, unsupervised model for segmentation. A Bayesian framework provides a principled way to combine the modalities: separate sets of language models are learned for gesture and speech, and the priors on these language models control the relative influence of each modality on the predicted segmentation. The resulting system produces more accurate segmentations than are obtained using only speech information.

Finally, the lexical representation constructed for topic segmentation is applied to answer a more fundamental question about gesture: to what extent do different speakers use the same gestural forms when describing a single topic? Even assuming that gestures convey semantic meaning, their form is shaped by the speaker's mental imagery, which may be highly idiosyncratic. Is it possible to show that for some topics, many speakers will use the same representational gestures? I build a lexicon of gestural forms across multiple speakers, and apply a hierarchical Bayesian model to quantify the extent to which the distribution over forms is determined by the speaker and topic. This yields the first quantitative evidence that the use of gestural forms to communicate meaning is consistent across speakers.

[1] Unlike the traditional sense of the term "lexicon," the gestural forms here do not necessarily have any pre-defined meaning.

1.3 Contributions

The main contribution of this thesis is a predictive analysis of the relationship between gesture and discourse structure. While previous research has identified correlations between gesture and discourse phenomena (e.g., Quek, McNeill, Bryll, Duncan, et al., 2002), this thesis presents systems that predict the discourse structure of unseen data, using gestural features that are automatically extracted from video. Moreover, adding gesture to state-of-the-art text-based systems yields significantly improved performance on well-known discourse processing problems. This demonstrates that gesture provides new and unique information for discourse processing.

A second contribution is the focus on relationships between gestures, which are used to detect parallel patterns in the discourse structure. This approach is the first to successfully uncover gesture's contribution to the underlying narrative semantics. It is a departure from earlier efforts in multimodal natural language processing, which tried to identify individual gestures and intonation patterns that act as pragmatic discourse cues (e.g., Chen et al., 2004; Shriberg, Stolcke, Hakkani-Tur, & Tur, 2000). This dissertation focuses on three specific gestural patterns: similarity, cohesion, and salience. Models for each pattern are learned from automatically extracted features without labeled data. These models are then demonstrated to be predictive of discourse structure.

Finally, the machine learning models themselves constitute an important contribution. The use of custom models that explicitly encode the role of each modality differs from previous research, which relied on generic machine learning methods (e.g., Chen et al., 2004). The models employed in this thesis are capable of learning about gesture in its linguistic context, rather than in the abstract. This permits learning about gesture directly from linguistic annotations. In addition, these models provide a principled approach to modality combination. Chapter 4 applies these ideas in a supervised, discriminative framework, using a novel hidden conditional random field architecture; Chapter 5 presents two novel unsupervised Bayesian models.

The remainder of the thesis is organized as follows. Chapter 2 assesses related

work on gesture, discourse processing, and other attempts to integrate non-verbal modalities into automatic language processing. Chapter 3 describes a novel gesture-speech dataset that makes this research possible. Chapter 4 applies the ideas of gestural similarity and salience to the local discourse problems of coreference resolution and keyframe extraction. The mechanism is *conditional modality fusion*, a novel discriminative technique for modality combination, which learns to identify salient gestures and filter away non-communicative hand motions. Chapter 5 develops the notion of gestural cohesion, with an application to topic segmentation. In addition, I use a novel hierarchical Bayesian model to demonstrate that the relationship between gestural forms and the discourse topics can generalize across speakers. Finally, the main ideas and contributions of the thesis are summarized in Chapter 6, where I also discuss limitations and directions for future work.

2

Related Work

This thesis builds on diverse streams of related work. First, any computational account of gesture and discourse should be informed by psychology and linguistics. These fields provide theoretical models of how gesture and speech combine to create meaning, as well as experimental results that shed light on how humans use these modalities to communicate. Section 2.1 summarizes relevant contributions from this area, and notes prior computational work that builds on psycholinguistic models of gesture.

The remaining portions of this chapter describe implemented systems that employ gesture or other non-verbal features. Section 2.2 describes multimodal interfaces and dialogue systems in which human users interact with computers using gesture and sometimes speech. Section 2.3 describes the application of prosody to natural language processing – a parallel line of research that faces similar challenges to those

dealt with in this dissertation.

2.1 Gesture, Speech, and Meaning

From the 1970s on, there has been increasing interest in the study of gesture from the psychological and linguistic communities. This has been fueled largely by the hope that gesture can provide clues to the organization of language and thought in the human mind (Kendon, 2004). In the course of trying to answer these high-level questions, psychologists and linguistists have developed valuable ideas and results that inform my research. In this section, I describe studies of gesture's communicative function, prior attempts to formalize these ideas in a computational framework, and briefly mention a few notable taxonomies and annotation systems for gesture.

2.1.1 The Role of Gesture in Communication

Gesture has long been understood to be closely linked to speech (Condon & Ogston, 1967; Kendon, 1972). The form and timing of gesture and speech mirror each other in ways that are obvious even from casual observation.[1] However, our understanding of the communicative role of gesture remains incomplete at best, particularly with respect to how gestures are understood. Indeed, psychologists continue to debate whether representational gestures affect the listener's comprehension at all. I briefly summarize arguments on both sides of this debate, and then review experimental results showing specific semantic and pragmatic functions played by gesture.

Do Listeners Understand Gestures?

While there can be little doubt that listeners understand certain, specialized gestures (e.g., navigational directions accompanied by pointing), some researchers have expressed skepticism about the communicative function of spontaneous, representational gestures. Part of the motivation for such skepticism is that gesture is employed

[1] The synchrony between speech, gesture, and other physical movements is surprisingly tight, incorporating even eye blinks (Loehr, 2007).

even in situations where it cannot possibly be viewed – for example, in telephone conversations, or when speaking to the blind (Rimé & Schiaratura, 1991). Such examples show that at least some gestures are not *intentionally* produced for the viewer's benefit. Are such gestures produced merely out of habit, or is there some other motivation? Some researchers argue that representational gestures are primarily for the benefit of the speaker, rather than listener. In particular, Krauss (2001) argues that by acting out an action or idea via gesture, the speaker may find it easier to produce the associated verbal form.

There is evidence for the view that gesture aids speech production. When told not to gesture, speakers becomes substantially more dysfluent, as increasing numbers of filled pauses (e.g. "um") are placed within grammatical clauses (Rauscher, Krauss, & Chen, 1996). The authors argue that this suggests the absence of gesture leads to difficulty with lexical retrieval. Speakers experience additional difficulties when discussing content with a spatial component, speaking more slowly and producing more dysfluencies overall.

Such findings are not limited to cases in which speakers were explicitly forbidden or prevented from gesturing. Goldin-Meadow, Nusbaum, Kelly, and Wagner (2001) observe that the absence of gesture increases the speaker's cognitive load regardless of whether speaker was instructed not to gesture or simply chose to produce speech without gesture. These results – as well the failure of some studies to show that listener comprehension benefits from gesture (Krauss, Morrel-Samuels, & Colasante, 1991) – lead Krauss (2001) to conclude that representational gestures primarily serve to aid the lexical retrieval of spatially-associated terms.

In other settings, gestures do appear to aid comprehension. Goldin-Meadow (2003) describes a series of studies showing that children can benefit from observing gestures produced by their teachers; furthermore, when students make errors, their gestures can reveal additional information about where they went wrong. Using videos of scripted scenarios, Kelly, Barr, Church, and Lynch (1999) find that comprehension improved significantly when both gesture and speech were present. Interesting, when asked to recall just the spoken communication, listeners often added information that

was actually communicated with gesture and not speech. This suggests that not only did the listeners draw information from gesture, they also encoded this information in a way that did not distinguish between spoken and gestural communication. Finally, electroencephalography (EEG) studies demonstrate that subjects register neurophysiological indicators of surprise when viewing videos in which the gesture and speech convey contradictory information (Kelly, Kravitz, & Hopkins, 2004). This supports the view that people attempt to interpret the semantics of the gestures that they observe.

While this debate is relevant to our main interest in gesture and automatic discourse processing, from a computational perspective we may remain agnostic about the extent to which listener comprehension benefits from viewing representational gestures. Human language processing is robust, and may succeed even when various cues are removed (Whitney, 1998).[2] Automatic natural language processing systems do not benefit from the same common-sense reasoning and background knowledge available to humans. Thus, even if representational gestures are generally redundant with speech – and thus, rarely necessary for human listeners – they may still be of great value to computer systems for the forseeable future.

Gesture and Local Semantics

We now explore a series of studies that identify specific types of semantic phenomena that are sometimes communicated with gesture. One such case is word-sense disambiguation: Holler and Beattie (2003) find that speakers are more likely to produce representational gestures in conjunction with homonyms (one example from the paper is "arms," meaning either the body part or weapons) than with other words. While the meaning of homonyms may be deducible from the larger discourse context, such disambiguation requires additional cognitive effort on the part of human listeners, and may pose substantial difficulties for automatic systems.

[2]However, evidence from eye tracking suggests that even in cases when people can deduce meaning without gesture, listeners use gesture to interpret the utterance more quickly than when only speech is available (Campana, Silverman, Tanenhaus, Bennetto, & Packard, 2005).

One type of information often conveyed by gesture that is apparently not deducible from the associated speech is the physical size of semantic entities. Beattie and Shovelton (2006) ask subjects to describe a set of cartoon narratives, and find that size is often communicated via gesture, and very rarely communicated redundantly in both speech and gesture. This suggests that size would *not* be deducible from the surrounding speech without the presence of gestures. Moreover, they find that speakers are more likely to communicate size via gesture when the size is particularly important to the overall story, as judged by a separate set of raters who were familiar with the underlying narrative but did not view the speakers' explanations.

Similarly, gesture may be used to differentiate spatial relations between objects. Lausberg and Kita (2003) find that when describing the spatial configuration of two objects, the horizontal position of each object is expressed via the hand that is used to represent it. Similarly, Melinger and Levelt (2004) find that some speakers use gesture to disambiguate spatial relationships when describing an abstract arrangement of colored nodes – however, they find that roughly half of the speakers in their study never used gestures. Those speakers who did gesture also produced significantly more ambiguous speech, suggesting that they were intentionally using gesture as a disambiguating modality. The idea that verbal ambiguity could predict the likelihood of salient gestures influenced the use of verbal features to help assess gesture salience (Section 4.4.2).

Both Kendon (2004) and McNeill (1992) note examples in which speakers assign spatial regions to individual characters or entities, and then refer back to these regions when the entities are mentioned again. A more systematic study of this phenomenon is presented by So, Coppola, Licciardello, and Goldin-Meadow (2005), who show that speakers tend to use gestures to communicate coreference when describing cartoon narratives. The observation that spatial location may disambiguate noun phrases helped motivate my work on noun phrase coreference (Chapter 4), which uses spatial location as one of the features that predicts gestural similarity.

Aside from the word sense disambiguation study (Holler & Beattie, 2003), in all of the cases discussed thus far, gesture is used *relationally*. For example, when describing

spatial arrangements, gestures are (presumably) not taken to indicate the absolute position of the object, but rather the position of the objects relative to each other. In this thesis, I have focused on how gestures reveal a relationship of semantic identity. However, these studies suggest that gesture also provides an imagistic representation of the specific ways in which entities are dissimilar – for example, in their size and placement. The exploitation of such contrastive gestures is an interesting avenue for future research.

Gesture and High-Level Discourse Structure

Chapter 5 deals with gesture and topic-level discourse structure. There, I identify two main types of cues that predict topic segmentation: inter-segmental boundary cues, and intra-segmental cohesion. Boundary cues are essentially pragmatic: they convey information about the locations of segment boundaries, and not about the content within the segment. This dissertation concerns gestures that communicate the speaker's underlying meaning, and so has focused on intra-segmental gestural cohesion – the repetition of gestural forms throughout a topically-coherent segment. However, it is important to note that psycholinguistic research suggests that gesture communicates discourse structure in both ways, suggesting future research on computational models that unify both types of cues.

An explicit, quantitative study of nonverbal cues for discourse segment boundaries was undertaken by Cassell et al. (2001). They begin by noting that "changes in the more slowly changing body parts occur at the boundaries of the larger units in the flow of speech," an observation presented earlier by Condon and Osgton (1971). Put another way, not only are gesture and speech tightly synchronized, but this synchronization occurs on multiple levels, such that small linguistic units (e.g., phrases) are synchronized with fast moving body parts (e.g., hand and fingers) and large discourse units (e.g., topic segments) are synchronized with slower-moving parts (e.g., the torso). Cassell et al. therefore hypothesize that whole-body posture shifts should be indicative of discourse segment boundaries. Using manual annotations of speaker posture in twelve videos, they find that posture shifts occur much more frequently at

segment boundaries, although some shifts still occur within segments. They exploit this correlation to generate realistic gestures in an animated avatar; to my knowledge no one has attempted to build a system that detects segment boundaries using posture shifts.

Gestural cues have also been observed for more fine-grained discourse phenomena. Kendon (1995) describes four conventionalized gestures that are used in Southern Italy, each of which conveys information about the role of the associated speech in the surrounding discourse. These include four specific gestures: "the purse hand," indicating that the associated utterance is a question; "praying hands," often indicating that the speaker asks the listener to accept the utterance as given; "the finger bunch," which often indicates that a statement is a high-level topic, rather than a comment; and "the ring," which tells the listener that a precise piece of information is being communicated. These gestures provide information on the level similar to the "dialogue acts" annotated in DAMSL (Core & Allen, 1997): descriptions of the meta-linguistic role played by sentence-level utterances. In this sense, such conventionalized gestures may provide a sort of pragmatic visual punctuation of discourse. However, Kendon notes that the Southern Italian linguistic community is well-known to have a large number of conventionalized gestural forms; analagous discourse-marking gestures may not exist in other linguistic communities.

The most relevant psycholinguistic work with respect to gestural cohesion derives from David McNeill's concept of *catchments*: recurring gestural themes that indicate patterns of meaning (1992, 2005). McNeill shows, though example, how unique iconic and metaphoric gestures accompany specific ideas, and how these gestures recur whenever the associated idea is discussed. Catchments may also display subtle – but semantically crucial – variations. In one of McNeill's examples (2005, pages 108-112), speakers describe a "Sylvester and Tweety" cartoon, in which Sylvester (a cat) crawls up a drain pipe two times: once on the outside, and once on the inside. In this example, speakers often uses similar gestures, but modulate the second instance to indicate that the action is now on the inside of the drain – moreover, this information is communicated in gesture even when one of the speakers forgets to mention it in

the speech.

McNeill argues that gesture also conveys meaning through *layers* of catchments, simultaneously providing information at the levels of narrative (the story), metanarrative (statements about the story), and paranarrative (statements about the speaker herself). In one such example, overlapping catchments occur at each layer, yielding extremely complex gestural forms (2005, pages 173-177).

While McNeill and his collaborators have demonstrated many compelling examples of catchments, there is little systematic data on how often catchments occur. There are other questions as well: which features of the gestural modality do catchments tend to employ? What governs the mapping between semantic concepts and gestural features? What sorts of ideas tend to be expressed via catchments, rather than (or in addition to) spoken language?

The answers to these questions bear obvious significance for any computational account of gestural catchments. A complete computational treatment of catchments would require automatically detecting the full set of gestural features that may be used to express catchments, and identifying the specific features that are relevant in any given gesture. Implementing the layered model – in which catchments occur at the narrative, metanarrative, and paranarrative levels – requires going even further, assigning gestural features to each layer of communication. No complete computational implementation of gesture catchments has yet been attempted, though some preliminary efforts will be discussed in Section 2.1.2.

Despite these challenges, this dissertation can be viewed as a partial implementation of the idea of catchments – indeed, the first such implementation that successfully *predicts* discourse phenomena on unseen examples. In Chapter 1, I emphasized a strategy of detecting patterns between gestures, rather than decoding individual gestures – this is directly inspired by the idea of catchments. By detecting similar pairs of gestures, this dissertation shows that at least some catchments can be recognized automatically from raw visual features. Chapter 4 goes further, showing that such automatically detected catchments can be used to make accurate predictions about noun phrase coreference. Chapter 5 extends this idea from pairs of gestures

to larger sets, demonstrating a predictive relationship between catchments and topic boundaries.

2.1.2 Computational Analysis of Gestural Communication

Prior computational research on gesture in human-human speech has rarely emphasized predictive analysis of linguistic phenomena from signal-level gesture features, as this dissertation does. Nonetheless, this work provides valuable insights as to what types of gestural features can be detected from video, and how automatically-extracted and hand-annotated gesture features correlate with language.

Francis Quek and colleagues have published a series of papers that explicitly address the idea of catchments. Quek, McNeill, Bryll, Duncan, et al. (2002) demonstrate how automatic hand tracking can supplement a manual discourse analysis. The paper provides automatically extracted hand positions and velocities for a portion of a single dialogue, and couples this with a close analysis of the associated discourse structure. While this does not go as far as showing how discourse structure might be predicted from the gestural features, it points the way towards the sorts of visual features that would be necessary for such an approach to succeed.

Explicitly referencing McNeill, Quek advocates a "Catchment Feature Model," in which a variety of salient visual features could be extracted from gesture and then applied to multimodal linguistic analysis (Quek, 2003). These include: detecting static "hold" gestures (Bryll, Quek, & Esposito, 2001), symmetric and anti-symmetric motion (Xiong, Quek, & McNeill, 2002), oscillatory motion (Xiong & Quek, 2006), and the division of space into semantically meaningful regions (Quek, McNeill, Bryll, & Harper, 2002). Quek and his coauthors give examples in which such features appear to correlate with linguistic phenomena, including topic segmentation (Quek et al., 2000), speech repairs (Chen, Harper, & Quek, 2002), and filled pauses (Esposito, McCullough, & Quek, 2001). My thesis builds on these ideas by showing that similar gesture features can *predict* discourse phenomena on unseen data.

Another way in which my thesis extends this line of work is through the application of machine learning methods. In the cited papers from Quek et al., detectors for each

gestural feature are constructed by hand. In many cases, these detectors include a large number of hand-tuned parameters – for example, Bryll et al. (2001) list twelve manual parameters for the static hold detector. Because these parameters are tuned on small datasets, it is unclear how closely the observed correlation between gesture features and discourse structure depends on their precise settings. In particular, such an approach may not scale to datasets that include multiple speakers and topics. For this reason, I have emphasized applying a learning-based approach, and in particular, learning about gesture in the context of a specific language processing task.

A learning-based approach to gestural communication is applied by Chen et al., who investigate the connection between gesture and sentence boundaries. While sentence boundary detection is not a discourse-level linguistic task, this work is clearly relevant. In their first paper on this topic, Chen et al. (2004) show that automatically extracted hand gesture features are predictive of sentence boundaries, improving performance above transcript-based features. They also investigate the connection with a third modality – prosody – and find that adding gesture does *not* yield significant performance gains over a system that combined verbal and prosodic features. Thus, with respect to sentence boundaries, there appears to be substantial overlap between the prosodic and gestural modalities. Their second paper replaces automatically extracted gesture features with hand annotations, and does obtain a statistically significant improvement over a system using verbal and prosodic features (Chen, Harper, & Huang, 2006).

Chen's research, while highly relevant to this dissertation, differs in a few illustrative respects. As noted, sentence segmentation is not a discourse-level phenomenon (since discourse, by definition, concerns the semantic relationships across sentences). As a result, the task faced by Chen et al. is somewhat different: rather than inferring the speaker's meaning, they query gesture for clues about how the speech is syntactically organized. As a result, they do not attempt to identify patterns of gestures that form a catchment structure, but instead search for individual gestures that serve as a sort of visual punctuation. Since prosody is known to play a similar role, particularly with respect to sentence boundaries (Shriberg et al., 2000), it is unsurprising that

they discovered a high degree of overlap between prosodic and gestural features on this task.

To summarize, my dissertation is influenced by prior computational research on the communicative impact of hand gesture, but extends this work in two key ways. Quek et al. helped to inspire the principle that meaning can be found in patterns of gesture, rather than individual gestures. However, they did not attempt to show that such patterns could be learned from data, and they did not demonstrate a predictive relationship between gestural patterns and linguistic phenomena. Chen et al. did take a machine learning-based approach, in the context of sentence segmentation. However, rather than searching for meaning in patterns of gesture, they treated gesture as a sort of visual punctuation, and thus found high redundancy with prosodic features. This thesis combines the strengths of both approaches, and is thus the first to show that gestural patterns can predict discourse structure.

2.1.3 Classifying and Annotating Gesture

In this thesis, I have generally avoided taxonomies or annotations for gesture. This is not because such formalisms have no value, but because I believe that such an approach induces a substantial startup cost for gesture research: designing an annotation scheme that is concise enough to be tractable yet detailed enough to be useful, and then producing a sufficiently large dataset of annotations. Nonetheless, existing annotation systems can shed light on the nature of gesture, and on our approach and dataset. This section briefly discusses two frequently-referenced taxonomies from the literature, and a recently-published annotation scheme.

Movement Phases

Kendon (1980) provides a hierarchical taxonomy of gesture with respect to its kinematic and temporal properties, shown in Figure 2-1. At the top level is the *gesture unit*, a period of activity that begins with an excursion from rest position, and ends when the hands return to repose. The *gesture phrase* is what is traditionally con-

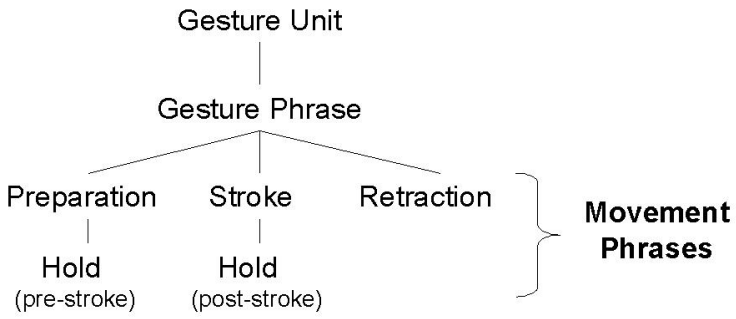


Figure 2-1: Kendon's taxonomy of gesture

sidered to be a single "gesture" – for example, pointing at something while talking about it, or tracing a path of motion. At the lowest level are *movement phases*: morphological entities that combine to create each gesture phrase. Every gesture phrase must have a *stroke*, which is considered by Kendon to be the content-carrying part of the gesture. In addition, there may also be a *prepare* phase, which initiates the gesture, and possibly a *retract* phase, bringing the hand back to rest position. A *hold* refers to a static positioning of the hand in gesture space, either before or after the stroke.

On the level of movement phases, Kendon's taxonomy provides a fine-scale temporal segmentation of gesture. An annotation according to this taxonomy could not be used to capture a gesture's semantics, as the taxonomy does not describe the gesture's form. However, the taxonomy does specify the stroke and hold phases as the most relevant portions of a gesture. This idea is relevant to the notion of gesture salience explored in Chapter 4, particularly the concept that the distance of the hands from rest position is predictive of the communicative role of the gesture. A fine-grained notion of salience, reflecting the distinctions proposed by Kendon, would be an interesting departure point for future research.

Types of Gesture Phrases

McNeill (1992) defines several types of gesture phrases: deictic, iconic, metaphoric, and beat. The following definitions are quoted and summarized from Cassell (1998):

- "**Deictics** spatialize, or locate in physical space..." Deictics can refer to actual physical entities and locations, or to spaces that have previously been marked as relating to some idea or concept.

- "**Iconic** gestures depict by the form of the gesture some features of the action or event being described." For example, a speaker might say "we were speeding all over town," while tracing an erratic path of motion with one hand.

- "**Metaphoric** gestures are also representational, but the concept they represent has no physical form; instead the form of the gesture comes from a common metaphor." For example, a speaker might say, "it happened over and over again," while repeatedly tracing a circle.

- "**Beat** gestures are small baton-like movements that do not change in form with the content of the accompanying speech. They serve a pragmatic function, occurring with comments on one's own linguistic contribution, speech repairs and reported speech." Speakers that emphasize important points with a downward motion of the hand are utilizing beat gestures.

This list nicely summarizes the various ways in which gestures can communicate information. However, as McNeill himself notes (2005, pages 41-42), the list should be thought of more as a set of dimensions of expressivity, rather than mutually exclusive bins. As discussed above, gestures are capable of communicating narrative, meta-narrative, and para-narrative information simultaneously. Thus, it is not difficult to find examples of gestures that occupy multiple places in this taxonomy: for example indicating a location in space (acting as a deictic) while simultaneously giving temporal emphasis (acting as a beat).

The data and methods in this dissertation emphasize deictic and iconic gestures. In the dataset described in Chapter 4, the presence of visual aids led speakers to produce a large number of deictic gestures (for a quantitative analysis, see Eisenstein & Davis, 2006), referring to specific areas on the diagram. No visual aids were permitted in the dataset from Chapter 5, and the resulting gestures were more often representational – iconic and metaphoric, by McNeill's taxonomy. Because the

dialogues focus on mechanical devices, iconics seem more likely than metaphorics, though no quantitative analysis of this dataset has yet been performed.

McNeill's taxonomy describes the communicative *function* of gesture rather than the communicative *content*. As with Kendon's taxonomy, even a perfect annotation will not tell us what the gestures actually mean. Moreover, it seems doubtful that the communicative function of a gesture can be evaluated without consideration of the surrounding linguistic context – a single form might appear as deictic or iconic depending on the accompanying speech. In a sense, McNeill's taxonomy describes more than gesture – it describes the role gesture plays in language, which cannot be understood without consideration of the speech itself. Indeed, viewers have difficulty reliably distinguishing deictic from iconic gestures when not permitted to consult the audio channel (Eisenstein & Davis, 2004).

FORM

Martell, Howard, Osborn, Britt, and Myers (2003) propose an annotation system named FORM, which describes the kinematic properties of gesture. FORM systematizes gesture in a variety of ways, such as dividing gesturing space into discrete bins, and categorizing all possible hand shapes. High quality FORM annotations – whether obtained through automatic or manual transcription – may ultimately facilitate gesture research by abstracting over some of the signal level noise in video. However, extracting communicative content from such a representation still poses some of the same problems faced when dealing with video directly: perceptually similar gestures may be appear to be quite different in the FORM representation, and only a few of the features in the representation will be relevant for any given gesture. These issues, coupled with the substantial implementational challenge of extracting a FORM representation from video, led me to avoid using FORM annotations in this dissertation.

2.2 Multimodal Interaction

Thus far, the area in which gestures have found the greatest application in implemented software systems is in multimodal user interfaces. Multimodal input permits users to control computer applications using speech and some sort of gestural modality – typically a pen, but in some cases, free-hand gestures. In general, both the language and gestures permitted by such systems are constrained by a limited vocabulary and fixed grammar. After reviewing some notable examples of multimodal input, I briefly present some relevant systems for multimodal output generation. In such research, information is conveyed to the user via avatars that are capable of both speech and gesture. Any such effort encodes theories about how modalities can combine to create meaning, raising interesting connections with this dissertation.

2.2.1 Multimodal Input

Multimodal input processing was pioneered by the "Put-That-There" system, which combined speech and pointing gestures (Bolt, 1980). Gestures were performed by manipulating a small cube augmented with tracked beacons, and the user was able to specify objects with a combination of words and pointing gestures, "e.g., move **that** to the right of the green square." The user's goal was to create and move colored geometric objects around a map.

Thus, from a very early stage, multimodal input research emphasized utilizing gesture to ground ambiguous pronouns with spatial reference. Subsequent systems extended this idea to more complex gestures. For example, QuickSet (Cohen et al., 1997) permitted deictic pen gestures indicating regions or sets of objects, and also recognized sketches of predefined symbols – though these cannot truly be said to be "gestures" in the same sense taken throughout this dissertation. A second innovation of QuickSet was the introduction of unification-based semantic parsing techniques to build a frame representation jointly from speech and gestural input (Johnston et al., 1997). However, this approach requires that users speak and gesture according to a fixed grammar, and is thus inapplicable to our interest in human-human dialogue.

More flexible probabilistic approaches were later considered (Chai, Hong, & Zhou, 2004; Bangalore & Johnston, 2004), though even these assume that the universe of possible referents is known in advance. While perfectly reasonable in the case of human-computer interaction, this assumption is generally implausible for discourse between people.

All of the systems described thus far in this section permit gesture only through pens or tracked pointers.[3] Sharma et al. argue that to move towards control by free hand gestures, we must design recognition algorithms that handle the gestures that occur in unconstrained human-human dialogue (Kettebekov & Sharma, 2000). As a basis for this work, they constructed a dataset of video recordings of weather forecasts; this was motivated in part by the professional recording quality, which facilitates hand tracking. According to Poddar, Sethi, Ozyildiz, and Sharma (1998), many hand motions in this dataset are well-described by a relatively small taxonomy: points, contours, and regions. Interestingly, even at this level of description, gesture processing is facilitated by taking the surrounding language into account: recognition was improved by both the transcipt (Sharma, Cai, Chakravarthy, Poddar, & Sethi, 2000) and the speaker's prosody (Kettebekov, Yeasin, & Sharma, 2005). This provides support for the view that the interpretation of co-speech gesture depends critically on the surrounding language.

2.2.2 Multimodal Output

While this dissertation focuses on processing natural multimodal language as input, a parallel track of research works to generate realistic-looking gesture and speech. As mentioned above, Cassell et al. (2001) describe a system that produces plausible posture shifts and gaze behaviors, based on the discourse structure. Nakano, Reinstein, Stocky, and Cassell (2003) present an empirical study of human-human interaction, showing a statistical relationship between hand-coded descriptions of head gestures and the discourse labels for the associated utterances (e.g., "acknowledgment," "an-

[3] An extension of QuickSet to free hand gestures is presented by Corradini, Wesson, and Cohen (2002).

swer," and "assertion"). They then demonstrate that these findings can be encoded in a model to generate realistic conversational "grounding" behavior in an animated agent.

These systems generate gestures that convey metanarrative content: information about the role that each utterance plays in the discourse. In contrast, Kopp, Tepper, Ferriman, and Cassell (2007) investigate how to produce gestures that convey narrative meaning directly. They present an animated agent that gives navigation directions, using hand gestures to describe the physical properties of landmarks along the route. In this system, the hand gestures are dynamically generated to reflect the characteristics of the semantic entity being described. As noted by Lascarides and Stone (2006), gestural form is generally underspecified by semantics, as there are multiple ways to express the same idea. One way to further constrain gesture generation is to produce gestures that observe the catchment structure proposed by McNeill (1992) and exploited in this dissertation. At this time, I am aware of no gesture generation system that attempts to implement this idea.

2.3 Prosody

Parallel to our interest in gesture, there is a large literature on supplementing natural language processing with *prosody* – a blanket term for the acoustic characteristics of speech, e.g. intonation and rhythm, that are not reflected in a textual transcription. Incorporating prosody into discourse processing poses similar challenges to those faced with gesture. Like gesture, prosody is a continuous-valued signal that does not easily lend itself to combination with the discrete representations usually employed for text. However, while this dissertation focuses on extracting narrative content from gesture, prosody has been used only at a metanarrative level, giving explicit cues of semantic and discourse structure.

2.3.1 Prosodic Indicators of Discourse Structure

Pierrehumbert and Hirschberg (1990) proposed that "intonational structures" help to group constituents for a compositional semantic analysis: a smooth intonational contour over a set of constituents suggests that they can be combined, while sharp discontinuities suggest a natural boundary point. This idea was applied to semantic processing using combinatory categorial grammars by Steedman (1990). In this line of research, prosodic contours act as intonational parentheses or commas: punctuation that serves to reveal the underlying combinatory semantic structure. Later research applied the idea of prosody-as-punctuation to statistical syntactic parsing (Gregory, Johnson, & Charniak, 2004; Kahn, Lease, Charniak, Johnson, & Ostendorf, 2005).

Prosodic features have also been applied to inter-sentential discourse analysis. For example, Grosz and Hirshberg (1992) showed that phrases beginning discourse segments are typically indicated by higher-than-normal pitch, and were preceded by unusually long pauses. Parentheticals – short digressions from the principal discourse segment topic – are indicated by a compressed pitch range. The relationship between these prosodic cues and discourse boundaries was more systematically investigated by Swerts (1997). In more recent research, similar prosodic features have been applied to topic and sentence segmentation, surpassing the performance of systems that use only textual features (Shriberg et al., 2000; Kim, Schwarm, & Osterdorf, 2004). The literature on prosody thus parallels Kendon's (1995) identification of gestural forms for specific discourse acts, and Chen's (2004, 2006) use of gesture as a sentence boundary cue. Rather than searching for a prosodic expression of the narrative semantic content, these researchers have identified pragmatic cues about the narrative structure.[4] While such an approach may be extensible to gesture, this would ignores gesture's capability to express narrative content directly through the gestural form.

[4]One exception is a recent effort to perform cohesion-based segmentation using only acoustic features (Malioutov, Park, Barzilay, & Glass, 2007). However, this work does not show that acoustic features yield an improvement over manual transcripts, so the acoustic features may be approximating standard lexical cohesion, rather than contributing additional information.

2.3.2 Modality Combination for Prosody and Speech

Another distinction from prior work on prosody relates to modality combination. Most of the papers on prosody train separate prosodic and textual classifiers, and then combine the posterior distributions (e.g., Y. Liu, 2004). Typically, this involves taking a weighted average, or multiplying the posteriors together, although one may also use the posterior probability from one classifier as a feature in another classifier. This approach, labelled *late fusion* in Section 4.4.3, was also used for gesture-speech combination by Chen et al. (2004, 2006). Alternatively, *early fusion* combines features from both modalities into a single feature vector – this technique was compared with various late fusion alternatives by Shriberg et al. (2000) and Kim et al. (2004).

Late fusion approaches often outperform early fusion because they explicitly acknowledge the differences between modalities; however, adding or multiplying posterior probabilities is ad hoc and theoretically unjustified. Such techniques may have been necessary because much of the research on prosody uses "off-the-shelf" machine learning components that were not originally intended to handle multiple modalities with very different characteristics. For example, Shriberg et al. (2000) model prosody with a decision tree because there is no obvious way to add prosodic features directly to an HMM-based language model.

In contrast, this dissertation employs custom models that explicitly encode each modality separately in the model structure. In the coreference task from Chapter 4, I employ a conditional model in which the potential function gates the gesture similarity features, while allowing the verbal features to be used in all cases. In the discourse segmentation task from Chapter 5, gesture and speech are modeled in a generative Bayesian framework; separate Dirichlet priors permit a different amount of smoothing in each modality. The application of custom, structured approaches to the model-combination problem is one of the major contributions of this dissertation.

3

Dataset

This chapter describes the set of video recordings used to perform the experiments in this dissertation.[1] The elicitation procedure for this dataset reflects a desire for balance between ecological validity and tractability for automatic processing. The resulting speech and gesture is spontaneous and unscripted but is shaped to be relevant to the main questions of this dissertation by the tasks and scenarios that were assigned to the participants. In particular, participants were asked to give *instructional presentations*, yielding speech and gesture that should be similar to classroom lectures and business presentations. The applicability of this dataset to other linguistic settings is an important topic for future research.

[1]This dataset is a subset of a larger effort performed collaboratively with Aaron Adler and Lisa Guttentag, under the supervision of Randall Davis (Adler, Eisenstein, Oltmans, Guttentag, & Davis, 2004). Portions of the dataset are available at `http://mug.csail.mit.edu/publications/2008/Eisenstein_JAIR/`

At the time of this writing, there exist some linguistic corpora that include visual data, but none are appropriate for this dissertation. The AMI corpus (Carletta et al., 2005) includes video and audio from meetings, but participants are usually seated and their hands are rarely visible in the video. The VACE corpus (Chen et al., 2005) also contains recordings of meetings, with tracking beacons attached to the speakers providing very accurate tracking. This corpus has not yet been released publicly.

Both AMI and VACE address seated meeting scenarios; however, gesture may be more frequent when speakers give standing presentations, as in classroom lectures or business presentations. There are many such video recordings available, but they have typically been filmed under circumstances that frustrate current techniques for automatic extraction of visual features, including camera movement, non-static background, poor lighting, and occlusion of the speaker. Rather than focus on solving these substantial challenges for computer vision, a new multimodal corpus was gathered in a manner that attempted to avoid such problems.

Participants Fifteen pairs of participants joined the study by responding to posters on the M.I.T. campus. The group included seventeen females and thirteen males, with ages ranging from 18 to 32; all participants were university students or staff. As determined by a pre-study questionnaire, all but six of the participants were native speakers of English. Of the remainder, four described themselves as "fluent," one as "almost fluent," and one spoke English "with effort." Participants registered in pairs, eliminating a known confound in which strangers often limit their gestures due to social inhibition in the early parts of a conversation.

Topics Each pair of participants conducted six short discussions. For each discussion, they were assigned a specific topic to ensure that the data would be meaningful and tractable. Five of the six topics related to the structure or function of a physical device: a piston, a pinball machine, a candy dispenser, a latchbox, and a small mechanical toy. The remaining topic was a short "Tom and Jerry" cartoon. These topics were chosen to encourage the use of concrete, representational gestures. How-

ever, the participants were given no explicit instructions about gesture. Diagrams for each topic are shown in Appendix C.

Procedure At the beginning of the experiment, one participant was randomly selected from each pair to be the "speaker," and the other to be the "listener." The speaker was given prior knowledge about the topic of discussion – usually in the form of an explanatory video – and was required to explain this information to the listener. The listener's role was to understand the speaker's explanations well enough to take a quiz later. The speaker stood behind a table, while the listener was seated.

The discussions were videotaped and were conducted without the presence of the experimenters. Discussions were limited to three minutes, and the majority of speakers used all of the time allotted. This suggests that more natural data could have been obtained by not limiting the explanation time. However, in pilot studies 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 elicitation was usually around 45 minutes.

Materials For the piston, pinball machine, candy dispenser, and latchbox, the speaker was privately shown a short video showing a simulation of the device; for the "Tom and Jerry" case, the speaker was shown the relevant cartoon; for the mechanical toy, the speaker was allowed to examine the physical object.

A variety of experimental conditions were considered, manipulating the presence of explanatory aids. In the "diagram" condition, the speaker was given a pre-printed diagram (all diagrams are shown in Appendix C, page 128). In the "no aid" condition, the speaker was not given any explanatory aids. Data from the "diagram" condition is used in Chapter 4, and the "no aid" condition is used in Chapter 5. There was also a "drawing" condition, in which the speaker was given a tracked whiteboard marker. Data from this condition is not used in this thesis.

The pinball machine was always presented first, as a "warmup" task, in the "diagram" presentation condition. The latchbox, candy dispenser, and piston were coun-

terbalanced against presentation conditions. No diagram was available for the cartoon and toy, so these were never assigned to the "diagram" condition. Except for the pinball machine, the order of presentation was counterbalanced against both the topic and presentation condition.

Recording and Annotations Speech was recorded using headset microphones. An integrated system controlled the synchronization of the microphones and video cameras. Speech was transcribed manually and with the Windows XP Microsoft Speech Recognizer. 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 (Huang, Alleva, Hwang, & Rosenfeld, 1993), yielding accurate timestamps for each transcribed word.

Video recording employed standard consumer camcorders. Both participants wore colored gloves to facilitate hand tracking. An automatic hand tracking system for this dataset is described in Section 4.2.1 (page 38). The extraction of spatio-temporal interest points is described in Section 5.2 (page 83).

Various linguistic annotations were applied to the dataset. Section 4.4.3 (page 56) describes noun phrase coreference annotation; Section 4.5.3 (page 71) describes the annotation of salient keyframes; Section 5.3.3 (page 93) describes the annotation of topic segmentation. Detailed statistics about the dataset can be found in Appendix B (page 125).

4

Gesture and Local Discourse Structure

This chapter describes the application of gesture to two local discourse processing problems: noun phrase coreference resolution, and the extraction of keyframe summaries. I show that models of gesture *similarity* and *salience* can be learned jointly, using labels only for noun phrase coreference. The resulting multimodal classifier significantly outperforms a verbal-only approach, marking the first successful use of gesture features on this problem. Modeling gesture salience is shown to further improve coreference performance; moreover, the learned model of gesture salience can be transferred to the keyframe extraction problem, where it surpasses competitive alternatives.[1]

[1]Some of the work in this chapter was published previously (Eisenstein, Barzilay, & Davis, 2008c; Eisenstein & Davis, 2007; Eisenstein, Barzilay, & Davis, 2007).

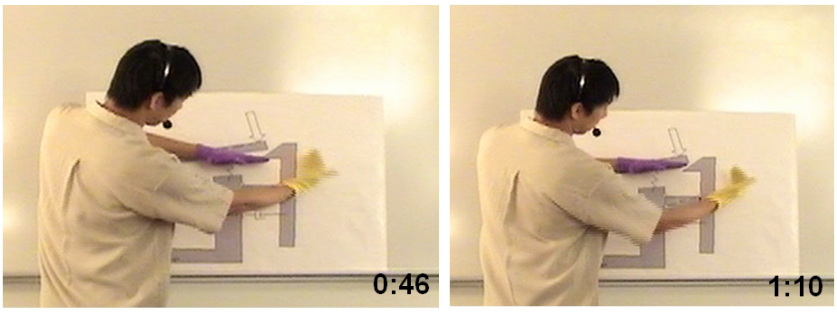


Figure 4-1: An excerpt of an explanatory narrative in which gesture helps to disambiguate coreference

4.1 Introduction

Coreference resolution is the task of partitioning the noun phrases in a document or dialogue into semantic equivalence classes; it has been studied for over thirty years in the AI community (Sidner, 1979; Kameyama, 1986; Brennan, Friedman, & Pollard, 1987; Lappin & Leass, 1994; Walker, 1998; Strube & Hahn, 1999; Soon, Ng, & Lim, 2001; Ng & Cardie, 2002). Resolving noun phrase coreference is an important step for understanding spoken language, with applications in automatic question answering (Morton, 1999) and summarization (Baldwin & Morton, 1998). This task is widely believed to require understanding the surrounding discourse structure (Sidner, 1979; Grosz & Sidner, 1986).

There are several ways to indicate that two noun phrases refer to the same semantic entity. Most trivially, it may be reflected in the orthography of the noun phrases. For example, consider the trio of noun phrases: "the big red ball," "the big round ball," and "the round, red ball"; the surface forms alone suggest that these noun phrases are likely to corefer. In other cases, coreference may be indicated by semantically similar but orthographically distinct words, e.g., "the man with the inappropriate clothing" and "Mr. Ugly-Pants" both indicate that the referent is a man who is poorly dressed, but more sophisticated linguistic processing is required to make such an inference.

With anaphoric pronouns, the surface form typically conveys little semantic information – in English, only the gender and number may be communicated. In text, the pronoun usually refers to the most recent, or most prominent, prior compatible noun phrase. However, the presence of hand gesture radically alters this situation. Gesture may be used to highlight the similarity between a pronoun and a previously spoken noun phrase, raising the prominence of noun phrases that otherwise would not be potential targets for coreference. Gestural similarity may be conveyed by assigning spatial locations to semantic entities (So et al., 2005), and then referring back to those locations. Alternatively, the similarity may conveyed through repeated motion patterns (McNeill, 1992).

Figure 4-1 shows an example in which gesture helps to explicate a coreference relation. Several sentences occur between the anaphoric pronoun "it" and the original noun phrase "this bar." However, the similarity of the gestures – in this case, both the location and organization of the hands – indicates that the noun phrases indeed refer to the same entity.

Thus, in the multimodal setting, gesture can be crucial to understand the speaker's meaning. Moreover, even when gesture is not the only cue to noun phrase coreference, it can reduce the burden on linguistic processing by acting as a supplemental modality. In either case, our goal is to identify *similar* gestures, which can be used as a clue to the semantic relationship of co-articulated noun phrases. In this way, gesture and speech combine to reveal meaning, without requiring the interpretation of individual gestures. Interpreting individual gestures requires reconstructing the visual metaphor governing the mapping between gestural features and semantic properties. This is especially difficult because the relationship between gesture and meaning is underspecified (Lascarides & Stone, 2006), permitting multiple gestural realizations of a single idea. By focusing on identifying similar gestures, the inherent difficulties of recognizing and interpreting individual gestures can be avoided.

While gestural similarity may be a useful clue for uncovering discourse structure, not at all hand movements are intended to be informative (Kendon, 1980). For example, "self-adaptors" are self-touching hand movements, such as adjusting one's glasses

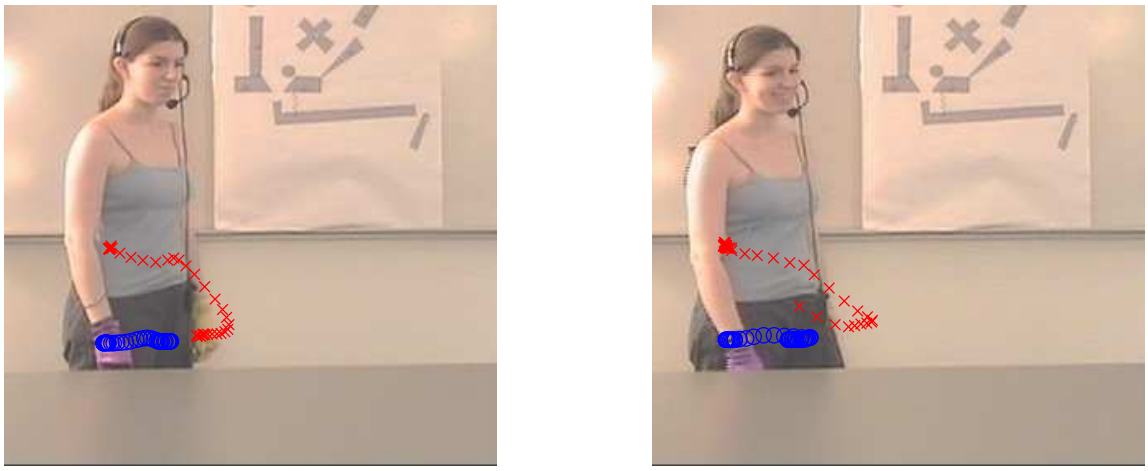


Figure 4-2: An example of a pair of "self-adaptor" hand motions, occurring at two different times in the video. The speaker's left hand, whose trajectory is indicated by the red "x," moves to scratch her right elbow. While these gestures are remarkably similar, the accompanying speech is unrelated.

or hair. Such movements are believed to have little direct communicative function, although they may function as proxies for stress (Beattie & Coughlan, 1999). Figure 4-2 shows two examples of a self-adaptor, which is repeated in a highly consistent form. It is probably inappropriate to draw inferences about the semantics of the co-articulated speech based on these hand movements. Thus, the idea of leveraging similarity of hand motion requires a qualification – we are interested in the similarity of *salient* gestures.

This chapter explores the connection between gestural similarity, salience, and meaning. Section 4.2 describes the extraction of a set of visual features used to characterize gestural similarity. Section 4.3 describes a novel gesture-speech combination technique called *conditional modality fusion*; it is distinguished from previous techniques in that it attempts to identify and isolate the contribution of salient gestures, ignoring spurious movements such as self-adaptors. Section 4.4 describes the application of these ideas to noun phrase coreference resolution, finding significant improvements over the speech-only case. Gesture salience is then exploited to automatically produce keyframe summaries in Section 4.5; these summaries are consistent with keyframes selected by human annotators. Finally, the ideas in this chapter are

summarized in Section 4.6.

4.2 From Pixels to Gestural Similarity

Gesture may be described on a myriad of dimensions, including: hand shape, location, trajectory, speed, and the spatial relation of the hands to each other. This section describes a set of features that characterize gestural similarity along some of these dimensions. The communicative power of the gestural medium derives from its inherent flexibility, and attempts to completely systematize the various ways in which gesture may convey meaning seem doomed to failure. The feature set presented here is thus limited, but has two desirable properties: the features can be extracted automatically from video, and are shown to be effective for language processing in the second part of this chapter. The development of additional gesture similarity features is always an important area of future research.

4.2.1 Hand Tracking from Video

The feature set used in this section is based on hand tracking, meaning that it is necessary to obtain estimates of the locations of the speaker's hands. This is done by estimating the pose of an articulated upper body model at each frame in the video, using color, motion, and image edges as cues. Search is performed using the particle filter – a sampling-based technique that enforces temporal smoothness across the model configuration. The system described in this section is implemented using OpenCV,[2] a library of image processing and computer vision algorithms. More details on the video recording can be found in Chapter 3.

Articulated Upper-Body Model

An instantiation of the articulated upper-body model is shown in the right panel of Figure 4-3. The model contains shoulder and elbow joints, a "twist" parameter allowing the entire body to rotate in the depth plane, and position on the x-axis.

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

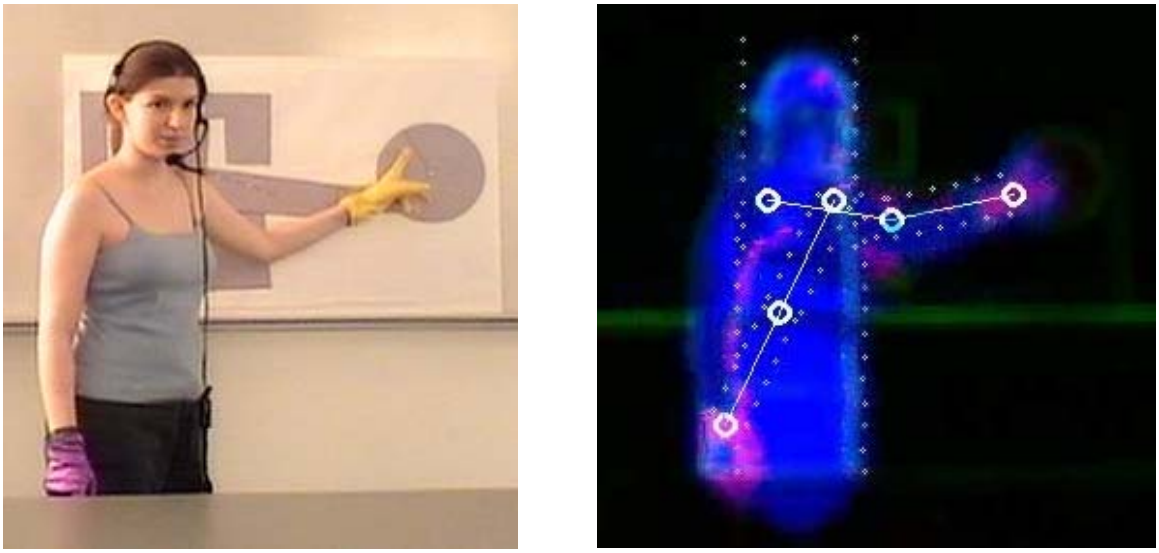


Figure 4-3: An example image from the dataset, with the estimated articulated model (right). Blue represents foreground pixels; red represents pixels whose color matches the hands; green represents image edges.

For each of these six parameters, the value, velocity, and acceleration are estimated. There are also six fixed parameters describing the dimensions of the articulated parts, which are tuned by hand for each speaker.

Visual Cues

The articulated model is fit to the video by leveraging color, motion, and edge cues. Speakers wore colored gloves, enabling the construction of a histogram of the expected color at the hand location; the likelihood of each model configuration was affected by how closely the observed color at the predicted hand location matched the known glove colors. Because the speaker was known to be the only source of motion in the video, pixels that differed from the background image are likely to be part of the speaker's body. Thus, the likelihood of each model configuration was also partially determined by how well it "covered" such estimated foreground pixels. A Canny filter (Forsyth & Ponce, 2003) was used to detect edges in the foreground portion of the image; the model configuration was also rated by how well its predicted edges lined up with these observed edges. Finally, a prior on model configurations enforced

physiological constraints: for example, reducing the probability of configurations with extreme joint angles.

Particle Filter

Using these cues and constraints, it is possible to search for model configurations that fit each individual frame in the video. However, more robust and rapid search can be performed by considering the video as a whole, leveraging temporal smoothness. This is done using a particle filter (Arulampalam, Maskell, Gordon, & Clapp, 2002), which maintains a set of weighted hypotheses about the model configuration. These weighted hypotheses are known as particles; the weights indicate an estimate of the probability that the hypothesized configuration is the true model state.

At each time step, particles may randomly "drift" to slightly different configurations, accounting for system dynamics. The particles are then reweighted, based on how well the hypothesized configuration matches the new observation. After reweighting, the set of particles is stochastically resampled; at each sampling step, the probability of a given particle being selected is proportional to its weight. Resampling tends to eliminate particles whose configuration does not match the observations, and creates multiple copies of those that do. The resulting online estimator is similar to the Kalman Filter, but better adapted to the non-Gaussian observation noise that typically affects vision applications (Arulampalam et al., 2002). The specific form of the particle filter employed here follows Deutscher, Blake, and Reid (2000).

Performance and Limitations

An informal review of the system output suggests that both hands were tracked accurately and smoothly more than 90% of the time, when not occluded. As shown in Figure 4-3, the system was able to correctly track the hands even when other parts of the articulated model were incorrect, such as the elbows; this is likely due to the strong cues provided by the colored gloves. It is difficult to assess the tracker performance more precisely without undertaking the time-consuming project of manually annotating the correct hand positions at each frame in the video. The main cause

Pairwise gesture features	
FOCUS-DISTANCE	the Euclidean distance in pixels between the average hand position during the two NPs
DTW-AGREEMENT	a measure of the agreement of the hand-trajectories during the two NPs, computed using dynamic time warping
SAME-CLUSTER	true if the hand positions during the two NPs fall in the same cluster
JS-DIV	the Jensen-Shannon divergence between the cluster assignment likelihoods

Table 4.1: The set of gesture similarity features

of error appears to be the lack of depth information. In particular, difficulties were encountered when speakers flexed their elbow joints in the depth dimension. Due to our single camera setup and the general difficulty of estimating depth cues (Forsyth & Ponce, 2003), such flexions in the depth dimension gave the appearance that the arm length itself was changing. Deutscher et al. (2000) show that this problem can be addressed with the use of multiple cameras.

4.2.2 Gesture Similarity Features

This section describes features that quantify various aspects of gestural similarity, listed in Table 4.1. Features are computed over the duration of each noun phrase. yielding a single feature vector per NP. While it is not universally true that the beginning and end points of relevant gestures line up exactly with the beginning and end of the associated words, several experiments have demonstrated the close synchrony of gesture and speech (McNeill, 1992; Loehr, 2007). The effectiveness of other gesture segmentation techniques is left to future work.

The most straightforward measure of gesture similarity is the Euclidean distance between the average hand position during each noun phrase – the associated feature is called FOCUS-DISTANCE. Euclidean distance captures cases in which the speaker is performing a gestural "hold" in roughly the same location (Kendon, 2004). However, Euclidean distance may not correlate directly with semantic similarity. For example, when gesturing at a detailed part of a diagram, very small changes in hand position may be semantically meaningful, while in other regions positional similarity may be

defined more loosely. The ideal feature would capture the semantic *object* of the speaker's reference (e.g., "the red block"), but this is not possible in general because a complete taxonomy of all possible objects of reference is usually unknown.

Instead, a hidden Markov model (HMM) is used to perform a spatio-temporal clustering on hand position and speed. This clustering is used to produce the SAME-CLUSTER and JS-DIV features, as explained below. The input to the model are the position and speed of the hands; these are assumed to be generated by Gaussians, indexed by the model states. The states of the HMM correspond to clusters, and cluster membership can be used as a discretized representation of positional similarity. Inference of state membership and learning of model parameters are performed using the traditional forward-backward and Baum-Welch algorithms (Rabiner, 1989).

While a standard hidden Markov model may be suitable, reducing the model's degrees-of-freedom can increase robustness and make better use of available training data. Reducing the number of degrees-of-freedom means that we are learning simpler models, which are often more general. This is done through *parameter tying*: requiring some subsets of model parameters to take the same values (Bishop, 2006). Three forms of parameter tying are employed:

1. Only one state is permitted to have an expected speed greater than zero. This state is called the "move" state; all other states are "hold" states, and their speed observations are assumed to be generated by zero-mean Gaussians. Only a single "move" state is used, because position seems most likely to be relevant for static gestures.

2. Transitions between distinct hold states are not permitted. This reflects the common-sense idea that it is not possible to transition between two distinct positions without moving.

3. The outgoing transition probabilities from all hold states are assumed to be identical. Intuitively, this means that the likelihood of remaining within a hold state does not depend on where that hold is located. While it is possible to

imagine scenarios in which this is not true, it is a reasonable simplification that dramatically reduces the number of parameters to be estimated.

Two similarity features are derived from the spatio-temporal clustering. The SAME-CLUSTER feature reports whether the two gestures occupy the same state for the majority of the durations of the two noun phrases. This is a Boolean feature that indicates whether two gestures are in roughly the same area, without need for an explicit discretization of space. However, two nearby gestures may not be classified as similar by this method if they are near the boundary between two states or if both gestures move between multiple states. For this reason, the similarity of the state assignment probabilities is quantified using the Jensen-Shannon divergence, a metric on probability distributions (Lin, 1991). JS-DIV is a real-valued feature that provides a more nuanced view of the gesture similarity based on the HMM clustering. Both SAME-CLUSTER and JS-DIV are computed independently for models comprising five, ten, and fifteen hold positions.

Thus far, our features are designed to capture the similarity between static gestures; that is, gestures in which the hand position is nearly constant. These features do not capture the similarity between gesture trajectories, which may also be used to communicate meaning. For example, a description of two identical motions might be expressed by very similar gesture trajectories. The DTW-DISTANCE feature quantifies trajectory similarity, using dynamic time warping (Huang, Acero, & Hon, 2001). This technique finds the optimal match between two time sequences, permitting a non-linear warp in the time dimension. Dynamic time warping has been used frequently in recognition of predefined gestures (Darrell & Pentland, 1993).

The continuous-valued are binned using WEKA's default supervised binning class, which is based on the method of Fayyad and Irani (1993).[3] This method recursively partitions the domain of an attribute value by adding cut points. Cut points are placed so as to minimize the class label impurity on each side of the cut. For example, in binning the FOCUS-DISTANCE feature the method first divides the attribute domain at the point that best separates positive and negative labeled examples. Additional

[3]The specific class is `weka.filters.supervised.attribute.Discretize`.

cut points are added until a termination criterion is reached, based on the minimum description length (Bishop, 2006).

Finally, note that feature set currently supports only single-hand gestures. The articulated upper body model makes it possible to estimate the distance of each hand from the body center. The more distant hand is used in all cases.

4.3 Gesture Salience

Section 4.1 introduced the hypothesis that gesture similarity is an important cue for analyzing discourse structure. However, not all hand movements are meaningful gestures. The psychology literature suggests that human viewers consistently identify a subset of hand motions as intentionally communicative, and disattend other, non-communicative movements (Kendon, 1978). A key claim of this thesis is that the same ability should be incorporated in multimodal discourse processing systems.

Hand movements that are relevant to the speaker's communicative intent will be referred to as *salient*. Our goal is to learn to estimate the salience of the hand movements that accompany critical parts of the speech – in the case of coreference, the focus is on noun phrases. As stated, the definition of salience implies that it is a property that could be annotated by human raters. In principle, such annotations could then be be used to train a system to predict salience on unseen data.

However, such labeling is time-consuming and expensive. Rather than addressing salience generally, it may be advantageous to use a more functional definition: salient gestures are hand movements that improve automatic language processing. As we will see, restating salience in this way permits it to be estimated without labeled data and situates the concept of gesture salience within the context of a specific language processing problem and feature set. For example, a given gesture may be irrelevant to coreference, as it may be communicating something other than noun phrase identity. Alternatively, a gesture may be salient for the relevant discourse processing task but may communicate in a way that cannot be captured by the available features. In

both cases, it would be better to treat the gesture as not salient for the purposes of natural language processing.

This section describes a novel approach to assessing gesture salience, called *conditional modality fusion*. This approach does not require an annotated training set and gives an estimate of salience that is customized both for the task and feature set. Conditional modality fusion learns gesture salience jointly with the target language processing task – in this case, noun phrase coreference. Gesture salience is modeled with a hidden variable; gesture features influence the coreference label prediction only when the hidden variable indicates that the gestures are salient. Thus, in maximizing the likelihood of the training data – which does not include labels for gesture salience – conditional modality fusion nonetheless learns to predict which gestures are likely to be helpful.

More formally, assume that the goal is to predict a binary label $y \in \{-1, 1\}$, representing a single binary coreference decision of whether whether two noun phrases refer to the same entity. The hidden variable $\mathbf{h}$ describes the salience of the gesture features. The observable features are written as $\mathbf{x}$, and the goal of training is to learn a set of weights $\mathbf{w}$. Conditional modality fusion learns to predict y and $\mathbf{h}$ jointly, given labeled training data only for y. Marginalizing over the hidden variable $\mathbf{h}$,

$$\begin{aligned} p(y|\mathbf{x};\mathbf{w}) &= \sum_{\mathbf{h}} p(y,\mathbf{h}|\mathbf{x};\mathbf{w}) \\ &= \frac{\sum_{\mathbf{h}} \exp\{\psi(y,\mathbf{h},\mathbf{x};\mathbf{w})\}}{\sum_{y',\mathbf{h}} \exp\{\psi(y',\mathbf{h},\mathbf{x};\mathbf{w})\}}. \end{aligned}$$

In the second line, the joint probability of y and $\mathbf{h}$ is modeled in terms of a ratio of exponentiated potential functions ψ. These functions representing the compatibility between the label y, the hidden variable $\mathbf{h}$, and the observations $\mathbf{x}$; this potential is parametrized by a vector of weights, $\mathbf{w}$. The numerator expresses the compatibility of the label y and observations $\mathbf{x}$, summed over all possible values of the hidden variable $\mathbf{h}$. The denominator sums over both $\mathbf{h}$ and all possible labels y', yielding the conditional probability $p(y|\mathbf{x};\mathbf{w})$.

This model can be trained by a gradient-based optimization to maximize the conditional log-likelihood of the observations. The unregularized log-likelihood and gradient are given by:

$$l(\mathbf{w}) = \sum_i \log p(y_i|\mathbf{x}_i;\mathbf{w}) \tag{4.1}$$

$$= \sum_i \log \frac{\sum_{\mathbf{h}} \exp\{\psi(y_i,\mathbf{h},\mathbf{x}_i;\mathbf{w})\}}{\sum_{y',\mathbf{h}} \exp\{\psi(y',\mathbf{h},\mathbf{x}_i;\mathbf{w})\}}, \tag{4.2}$$

$$\frac{\partial l_i}{\partial w_j} = \sum_{\mathbf{h}} p(\mathbf{h}|y_i,\mathbf{x}_i;\mathbf{w}) \frac{\partial}{\partial w_j}\psi(y_i,\mathbf{h},\mathbf{x}_i;\mathbf{w}) - \sum_{y',\mathbf{h}} p(\mathbf{h},y'|\mathbf{x}_i;\mathbf{w}) \frac{\partial}{\partial w_j}\psi(y',\mathbf{h},\mathbf{x}_i;\mathbf{w}).$$

The derivative of the log-likelihood is thus a difference of expectations. The first term is the expectation with the respect to only the hidden variable, using the training label y_i; the second term is the expectation with respect to both the hidden variable and the label. When these terms are equal the model cannot learn any more from this example and does not update the weights.

The use of hidden variables in a conditionally-trained model follows Quattoni et al. (2007). However, while this reference gives the general outline for hidden-variable conditional models, the form of the potential function depends on the role of the hidden variable. This is problem-specific, and a novel contribution of this thesis is the exploration of several potential functions, permitting different forms of modality fusion.

4.3.1 Models of Modality Fusion

Intuitions about the role of the hidden variable can be formalized in the form of the potential function ψ. This section considers several alternative forms for ψ, corresponding to different theories of gesture-speech integration. The models range from a simple concatenation of gesture-speech features to a structured fusion model that dynamically assesses the relevance of gesture features for every noun phrase.

All models are shaped by the goal of determining whether two noun phrases (NPs) are coreferent. Gesture salience is assessed at each NP, to determine whether the gestural features should influence our decision about whether the noun phrases corefer. We set $\mathbf{h} = \langle h_1, h_2 \rangle$, with $h_1 \in \{1, -1\}$ representing gesture salience during the first noun phrase (antecedent), and $h_2 \in \{1, -1\}$ representing gesture salience during the second noun phrase (anaphor).

Same-Same Model

In the trivial case, the hidden variable is ignored and features from both gesture and speech are always included. Since the weight vectors for both modalities are unaffected by the hidden variable, this model is referred to as the "same-same" model. Note that this is identical to a standard log-linear conditional model, concatenating all features into a single vector. This model is thus a type of "early fusion" (see Section 2.3), meaning that the verbal and non-verbal features are combined prior to training.

$$\psi_{ss}(y, \mathbf{h}, \mathbf{x}; \mathbf{w}) \equiv y(\mathbf{w}_v^\mathsf{T} \mathbf{x}_v + \mathbf{w}_{nv}^\mathsf{T} \mathbf{x}_{nv}) \tag{4.3}$$

$\mathbf{x}_v$ and $\mathbf{w}_v$ refer to the features and weights for the verbal modality; $\mathbf{x}_{nv}$ and $\mathbf{w}_{nv}$ refer to the non-verbal modality.

Same-Zero Model

Next, consider a model that treats the hidden variable as a gate governing whether the gesture features are included. This model is called the "same-zero" model, since the verbal features are weighted identically regardless of the hidden variable, and the gesture feature weights go to zero unless $h_1 = h_2 = 1$.

$$\psi_{sz}(y, \mathbf{h}, \mathbf{x}; \mathbf{w}) \equiv \begin{cases} y(\mathbf{w}_v^\mathsf{T} \mathbf{x}_v + \mathbf{w}_{nv}^\mathsf{T} \mathbf{x}_{nv}) + h_1 \mathbf{w}_h^\mathsf{T} \mathbf{x}_{h_1} + h_2 \mathbf{w}_h^\mathsf{T} \mathbf{x}_{h_2}, & h_1 = h_2 = 1 \\ y \mathbf{w}_v^\mathsf{T} \mathbf{x}_v + h_1 \mathbf{w}_h^\mathsf{T} \mathbf{x}_{h_1} + h_2 \mathbf{w}_h^\mathsf{T} \mathbf{x}_{h_2}, & \text{otherwise.} \end{cases} \tag{4.4}$$

The features $\mathbf{x}_h$ and weights $\mathbf{w}_h$ contribute to the estimation of the hidden variable $\mathbf{h}$. They may include some or all of the features from $\mathbf{x}_v$ and $\mathbf{x}_{nv}$, or different features. These features are assessed independently at each noun phrase, yielding $\mathbf{x}_{h_1}$ for the antecedent and $\mathbf{x}_{h_2}$ for the anaphor. A description of the hidden variable features that are used for coreference resolution is found in Section 4.4.2.

This model reflects the intuition that gesture similarity features (indicated by $\mathbf{x}_{nv}$) are relevant only when the gestures during *both* noun phrases are salient. Thus, these features contribute towards the overall potential only when $h_1 = h_2 = 1$.

To see how gesture salience can be learned from this potential function, it is helpful to consider cases. If the current model of gesture similarity contradicts the training label, then the dot product $y\mathbf{w}_{nv}^{\mathsf{T}}\mathbf{x}_{nv} < 0$. The log-likelihood is maximized when the numerator of equation 4.2 goes to infinity for all labeled examples; in other words, we want the potential function ψ to be large in all cases in the training set. Since $y\mathbf{w}_{nv}^{\mathsf{T}}\mathbf{x}_{nv} < 0$, the value of the potential function ψ will be higher when h_1 or $h_2 = -1$. Thus, when the model of gesture similarity contradicts the coreference label, this serves as a *de facto* negative training example for gestures salience. Similarly, cases in which the model of gesture similarity agrees with the coreference label serve as *de facto* positive examples of gesture salience.

Different-Zero Model

We may add flexibility to our model by permitting the weights on the verbal features to change with the hidden variable. This model is called the "different-zero" model, since a different set of verbal weights ($\mathbf{w}_{v,1}$ or $\mathbf{w}_{v,2}$) is used depending on the value of the hidden variable. Such a model is motivated by empirical research showing that speech is different when used in combination with meaningful non-verbal communication, compared to unimodal language (Kehler, 2000; Melinger & Levelt, 2004).

The formal definition of the potential function is:

$$\psi_{dz}(y, \mathbf{h}, \mathbf{x}; \mathbf{w}) \equiv \begin{cases} y(\mathbf{w}_{v,1}^{\mathsf{T}}\mathbf{x}_v + \mathbf{w}_{nv}^{\mathsf{T}}\mathbf{x}_{nv}) + h_1\mathbf{w}_h^{\mathsf{T}}\mathbf{x}_{h_1} + h_2\mathbf{w}_h^{\mathsf{T}}\mathbf{x}_{h_2}, & h_1 = h_2 = 1 \\ y\mathbf{w}_{v,2}^{\mathsf{T}}\mathbf{x}_v + h_1\mathbf{w}_h^{\mathsf{T}}\mathbf{x}_{h_1} + h_2\mathbf{w}_h^{\mathsf{T}}\mathbf{x}_{h_2}, & \text{otherwise.} \end{cases} \tag{4.5}$$

Other Models

Thus far, we have encountered three models of increasing complexity. The "different-different" model is one step more complex, including two pairs of weight vectors for both verbal and gestural features (see Equation 4.6). In this model, the distinction between verbal and non-verbal features ($\mathbf{x}_v$ and $\mathbf{x}_{nv}$) evaporates, and there is no reason that the hidden variable $\mathbf{h}$ should actually indicate the relevance of the non-verbal features. In addition, the high degree of freedom of this model may lead to overfitting.

$$\psi_{dd}(y, \mathbf{h}, \mathbf{x}; \mathbf{w}) \equiv \begin{cases} y(\mathbf{w}_{v,1}^{\mathsf{T}}\mathbf{x}_v + \mathbf{w}_{nv,1}^{\mathsf{T}}\mathbf{x}_{nv}) + h_1\mathbf{w}_h^{\mathsf{T}}\mathbf{x}_{h_1} + h_2\mathbf{w}_h^{\mathsf{T}}\mathbf{x}_{h_2}, & h_1 = h_2 = 1 \\ y(\mathbf{w}_{v,2}^{\mathsf{T}}\mathbf{x}_v + \mathbf{w}_{nv,2}^{\mathsf{T}}\mathbf{x}_{nv}) + h_1\mathbf{w}_h^{\mathsf{T}}\mathbf{x}_{h_1} + h_2\mathbf{w}_h^{\mathsf{T}}\mathbf{x}_{h_2}, & \text{otherwise.} \end{cases} \tag{4.6}$$

All of these models assume that the verbal features are always relevant, while the gesture features may sometimes be ignored. In other words, the salience of the verbal features has been taken for granted. One might consider alternative potential functions such as a "zero-same" model, in which the verbal features were sometimes ignored. As gesture unaccompanied by speech is extremely rare in the dataset, such models were not considered.

4.3.2 Implementation

The objective function (equation 4.1, page 46) is optimized using a Java implementation of L-BFGS, a quasi-Newton numerical optimization technique (D. C. Liu & Nocedal, 1989). L2-norm regularization is employed to prevent overfitting, with cross-validation to select the regularization constant.

Although standard logistic regression optimizes a convex objective, the inclusion of the hidden variable renders the objective non-convex. Thus, convergence to a global optimum is not guaranteed, and results may differ depending on the initialization. Nonetheless, non-convexity is encountered with many models in natural language processing and machine learning generally, such as Baum-Welch training of hidden Markov models (HMMs) (Rabiner, 1989) or hidden-state conditional random fields (Quattoni et al., 2007; Sutton & McCallum, 2006). Often, results can be shown to be reasonably robust to initialization; otherwise, multiple restarts can be used to obtain greater stability. The empirical evaluation presented in Section 4.4.4 shows that our results are not overly sensitive to initialization. In all other experiments, weights are initialized to zero, enabling the results to be reproduced deterministically.

4.4 Gestural Similarity and Coreference Resolution

This section describes the application of gesture similarity and conditional modality fusion to the problem of noun phrase coreference. The first part of the section presents the framework for this approach, which is based on existing text-based features and techniques. Next, I describe the setup and results of experiments showing that gesture similarity improves coreference resolution beyond traditional text-based approaches and that conditional modality fusion dramatically increases the power of gesture features by identifying salient gestures.

4.4.1 Verbal Features for Coreference

The selection of verbal features is guided by the extensive empirical literature on text-based coreference resolution (Soon et al., 2001; Ng & Cardie, 2002; Strube & Müller, 2003; Daumé III & Marcu, 2005). The proliferation and variety of features that have been explored is a consequence of the fact that coreference is a complex discourse phenomenon. Moreover, the way in which coreference is expressed depends

feature	type	description
pairwise verbal features		
NP-DIST	centering-based	the number of noun phrases between i and j in the document
SENT-DIST	centering-based	the number of sentences between i and j in the document
BOTH-SUBJ	centering-based	true if both i and j precede the first verb of their sentences
SAME-VERB	centering-based	true if the first verb in the sentences for i and j is identical
EXACT-MATCH	similarity	true if the two NPs are identical
OVERLAP	similarity	true if there are any shared words between i and j
STR-MATCH	similarity	true if the NPs are identical after removing articles
NONPRO-STR	similarity	true if the antecedent i and the anaphor j are not pronouns, and str-match is true
PRO-STR	similarity	true if i and j are pronouns, and str-match is true
J-SUBSTRING-I	similarity	true if j is a substring of i
I-SUBSTRING-J	similarity	true if i is a substring of j
EDIT-DISTANCE	similarity	a numerical measure of the string similarity between the two NPs
NUMBER-MATCH	compatibility	true if i and j have the same number
single-phrase verbal features		
PRONOUN	centering-based	true if the NP is a pronoun
HAS-MODIFIERS	centering-based	true if the NP has adjective modifiers
INDEF-NP	centering-based	true if the NP is an indefinite NP (e.g., *a fish*)
DEF-NP	centering-based	true if the NP is a definite NP (e.g., *the scooter*)
DEM-NP	centering-based	true if the NP begins with *this*, *that*, *these*, or *those*
COUNT	centering-based	number of times the NP appears in the document
lexical features	centering-based	lexical features are defined for the most common pronouns: *it*, *that*, *this*, and *they*

Table 4.2: The set of verbal features for multimodal coreference resolution. In this table, i refers to the antecedent noun phrase and j refers to the anaphor.

on the type of discourse in which it appears; relevant factors include the modality (e.g., speech vs. language), genre (e.g., meeting vs. lecture) and topic (e.g., politics vs. scientific subject). Although certain feature types are application-specific, three classes of features – centering-based, similarity, and compatibility features – are useful across most coreference applications. These classes form a basis for the verbal features selected here. Table 4.2 provides a brief description of the verbal feature set. Examples from the transcript in Appendix A.1 (page 120) provide a more detailed explanation of these features and motivate their use.

- **Centering-based features:** This set of features captures the relative prominence of a discourse entity, and its likelihood to act as a coreferent for a given phrase. These features are inspired by the linguistic analysis formalized in Centering Theory, which models the inter-sentential coherence of discourse (Grosz, Joshi, & Weinstein, 1995; Walker, Joshi, & Prince, 1998; Strube & Hahn, 1999; Kibble & Power, 2004). Centering theory posits that at any point of a coherent discourse, only one semantic entity is the *focus* of attention. Local discourse is then characterized in terms of focus transitions between adjacent sentences.

  Existing machine-learning based coreference systems generally do not attempt to fully implement centering-style analysis.[4] Instead, a number of centering-related features are included. For example, the BOTH-SUBJ feature helps to identify transitions in which the same entity remains in focus (these are known as *continue* transitions). According to centering theory, such transitions are common in locally-coherent discourse, and therefore coreference assignments that are consistent with this principle may be preferable. Transitions are also characterized in terms of their span (NP-DIST and SENT-DIST). Transitions that involve short gaps are preferred over transitions with long gaps.

  Another important set of Centering-related features is defined at the level of a single phrase. The syntactic role of a phrase in a sentence – captured in fea-

[4]Such an implementation is challenging in several respects. As noted by Poesio, Stevenson, Eugenio, and Hitzeman (2004), centering theory permits multiple possible computational instantiations, which may yield very different analyses. Additionally, implementing centering depends on obtaining detailed syntactic information, which is difficult for spoken language.

tures such as PRONOUN, HAS-MODIFIERS and INDEF-NP – indicates its discourse prominence and therefore its likelihood to be a coreference antecedent. For example, consider an utterance from lines 12 and 13 in Appendix A.1: "and this spring is active meaning that it's going up and down." Here, the anaphor "it" clearly refers to the antecedent "this spring." The fact that the antecedent is a demonstrative noun phrase (beginning with "this")[5] and that the anaphor is a pronoun also suggest coreference is likely. In addition to the syntactic status, frequency information is commonly used to approximate topical salience of an entity in a text (Barzilay & Lapata, 2005). This phenomenon is modeled by the COUNT feature.

- **Similarity features:** A simple yet informative set of coreference cues are based on string-level similarity between noun phrases. For example, the reference between "this spring" in line 12 of Appendix A.1 and the identical noun phrase in line 5 can be resolved by the exact match of the surface forms. Unsurprisingly, string match is often found to be the single most predictive feature because a discourse entity is commonly described using identical or similar noun phrases (Soon et al., 2001).

  Similarity information is captured in eight features that quantify the degree of string overlap. For example, the feature (EXACT-MATCH) indicates full overlap between noun phrases, while the feature (OVERLAP) captures whether two phrases share any common words. In the context of coreference resolution, noun phrase match is more informative than pronoun match, so in each syntactic category, distinct features for matching strings are applied (e.g., NONPRO-STR vs. PRO-STR), following (Ng & Cardie, 2002). Surface similarity may also be quantified in terms of EDIT-DISTANCE (Strube, Rapp, & Müller, 2002).

[5]Simple string matching techniques are used to assess phrase types: definite noun phrases are those beginning with the article "the"; indefinite noun phrases begin with "a" or "an"; demonstrative noun phrases begin with "this." Bare plurals are not marked as indefinites, and proper names do not appear in the dataset.

- **Compatibility features:** An important source of coreference information is compatibility between two noun phrases. For instance, the utterance "the ball" in line 11 cannot refer to the preceding noun phrase "these things," since they are incompatible in number. The NUMBER-MATCH feature captures this information, using a hand-coded heuristic to determine whether each noun phrase is singular or plural.

  Because the topic of discourse in the corpus relates to mechanical devices, almost all noun phrases are neuter-gendered. This eliminates the utility of features that measure gender compatibility. It is possible to use more complex semantic compatibility features – for example, derived from resources such as WordNet (Harabagiu, Bunescu, & Maiorano, 2001) or Wikipedia (Ponzetto & Strube, 2007) – but this is outside the scope of this thesis.

Some features traditionally used in coreference were avoided here. Features that depend on punctuation seem unlikely to be applicable in an automatic speech recognition setting, at least in the near future. In addition, while many systems in the MUC (Grishman & Sundheim, 1995) and ACE (Doddington et al., 2004) coreference corpora use "gazetteers" that list the names of nations and business entities, such features are not relevant to this corpus.

4.4.2 Salience Features

Salience features are observable properties of the speech and gesture that give clues about whether the speaker is gesturing in a way that is meaningful for the language language processing task at hand. In equations 4.4-4.6, these features are represented by $\mathbf{x}_{h_1}$ and $\mathbf{x}_{h_2}$. Unlike the similarity-based features described above, salience features must be computable at a single instant in time, as they encode properties of individual gestures and the associated noun phrases.

Previous research has investigated which types of verbal utterances are likely to be accompanied by gestural communication (Melinger & Levelt, 2004; Kehler, 2000).

Single-phrase gesture features	
DIST-TO-REST	distance of the hand from rest position
JITTER	sum of instantaneous motion across the NP
SPEED	total displacement over the NP, divided by duration
REST-CLUSTER	true if the hand is usually in the cluster associated with rest position
MOVEMENT-CLUSTER	true if the hand is usually in the cluster associated with movement

Table 4.3: The set of gesture features for multimodal coreference resolution

However, this thesis is the first attempt to formalize this relationship in the context of a machine learning approach that predicts gesture salience.

Verbal Salience Features

Meaningful gesture has been shown to be more frequent when the associated speech is ambiguous (Melinger & Levelt, 2004). Kehler (2000) finds that fully-specified noun phrases are less likely to receive multimodal support. These findings lead us to expect that salient gestures should be more likely to co-occur with pronouns, and less likely to co-occur with definite noun phrases, particularly if they include adjectival modifiers. Moreover, gestures are most likely to be helpful for coreference when the associated noun phrase is ambiguous. To capture these intuitions, all single-phrase verbal features (Table 4.2) are included as salience features.

Non-verbal Salience Features

The non-verbal salience features are described in Table 4.3. Research on gesture has shown that semantically meaningful hand motions usually take place away from "rest position," which is located at the speaker's lap or sides (McNeill, 1992). The DIST-TO-REST feature computes the average Euclidean distance of the hand from the rest position, over the duration of the NP. Here, rest position is computed from the articulated upper body model; it is defined as the center of the body on the x-axis, and at a predefined, speaker-specific location on the y-axis.

Hand speed may also be related to gesture salience. The SPEED feature captures the overall displacement (in pixels) divided by the length of the noun phrase.

Writing $\mathbf{x}$ for the hand position and $t \in \{1, 2, \ldots, T\}$ for the time index, we define SPEED $= \frac{1}{T}(\mathbf{x}_T - \mathbf{x}_1)^2$, which is the Euclidean distance between the start and end positions, divided by time. The JITTER feature captures the average *instantaneous* speed: JITTER $= \frac{1}{T}\sum_{t=2}^{T}(\mathbf{x}_t - \mathbf{x}_{t-1})^\mathsf{T}(\mathbf{x}_t - \mathbf{x}_{t-1})$. This feature captures periodic or jittery motion, which will not be quantified by the SPEED feature if the end position is near the original position. Also, high JITTER often indicates that the tracker has lost the hand position, which would be reason to ignore the gesture features.

As described in Section 4.2.2, an HMM was used to perform a spatio-temporal clustering on the hand positions and velocities. If the most frequently occupied state during the NP is the one closest to the rest position, then the REST-CLUSTER feature is set to TRUE. As noted earlier, rest position is very rarely used for communicative gestures.

In addition, the HMM employs parameter tying to ensure that all states but one are static holds, and this remaining state represents the transition movements between those holds. Only this state is permitted to have an expected non-zero speed. If the hand is most frequently in this transitional state during the NP, then this is expressed through the MOVEMENT-CLUSTER feature, which then is set to TRUE. While transitioning between other holds, the hand position itself is less likely to be meaningful.

4.4.3 Evaluation Setup

The goal of the evaluation is twofold: to determine whether gesture features improve coreference resolution and to compare conditional modality fusion with other approaches for gesture-speech combination. This section describes the dataset, evaluation metric, baselines for comparison, and parameter tuning.

Dataset

As described in Chapter 3, the dataset for all experiments in this thesis consists of a set of videos of short dialogues. This chapter focuses on a subset of videos in which

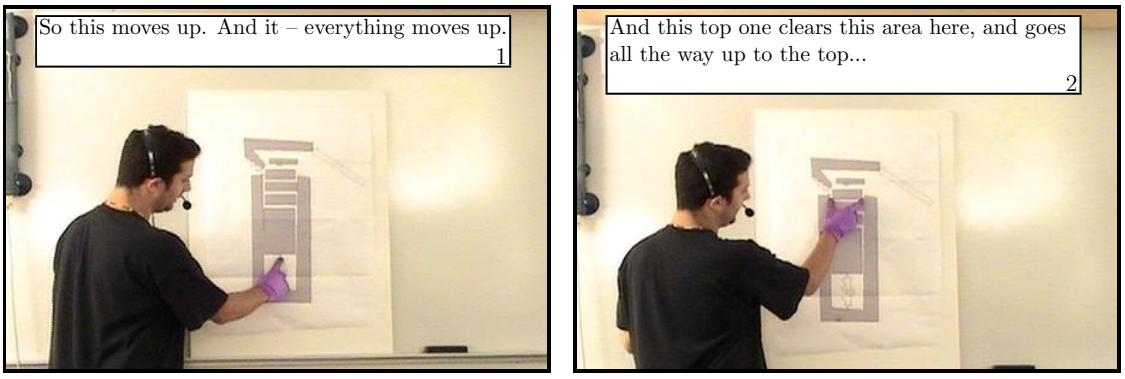


Figure 4-4: An excerpt of an explanatory narrative from the dataset

one of the participants was provided a pre-printed diagram showing a schematic of a mechanical device, which was the subject of discussion (see Figure 4-4).

The interpretation of gestures in this condition is often relatively straightforward; many, if not most of the gestures involve pointing at locations on the diagram. Visual aids such as printed or projected diagrams are common to important application areas, including business presentations, classroom lectures, and weather reports. Thus, this restriction does not seem overly limiting to the applicability of the work. Another subset of the corpus contains dialogues in which no such visual aids were permitted. Chapter 5 describes experiments that utilize this portion of the corpus.

For the experiments in this chapter, sixteen videos from nine different speakers are used. Corpus statistics are given in Table B.1. The dataset includes a total of 1137 noun phrases; this is roughly half the number found in the MUC6 training set, a text-only dataset that is also used for coreference resolution (Hirschman & Chinchor, 1998).

There are some important differences between this corpus and commonly-used textual corpora in coreference resolution, such as MUC (Grishman & Sundheim, 1995). Topically, this corpus focuses on descriptions of mechanical devices rather than news articles.[6] Consequently, less emphasis is placed on disambiguating entities such as people and organizations, and more on resolving references to physical objects.

[6]Four different mechanical devices were used as topics of discussion: a piston, candy dispenser, latch box, and pinball machine. Images of each are shown in Appendix C (page 128).

The corpora also differ in genre, with this corpus comprised of spontaneous speech, while the MUC corpus includes edited text. Such genre distinctions are known to play an important role in patterns of reference (Strube & Müller, 2003) and language use generally (Biber, 1988).

Speech Transcription A wide range of possibilities exist regarding the fidelity and richness of transcribed speech. Choices include transcription quality, existence of punctuation and capitalization, the presence of sentence boundaries and syntactic annotations. Here, a perfect transcription of words and sentence boundaries is assumed,[7] but no additional punctuation is given. This is similar to much of the research on the SWITCHBOARD corpus of telephone conversations, e.g., (Kahn et al., 2005; Li & Roth, 2001), although automatic speech recognition (ASR) transcripts have also been used (e.g., Shriberg et al., 2000). Using ASR may more accurately replicate the situation faced by an application developer trying to implement a deployable automatic language processing system. However, such an approach would also introduce a certain arbitrariness, as results would depend heavily on the amount of effort spent tuning the recognizer. In particular, if the recognizer is not well-tuned, this approach risks overstating the relative contribution of gesture features, because the verbal features would then be of little value.

Noun Phrase and Coreference Annotations Coreference resolution requires noun phrase boundaries as a preprocessing step, and we provide gold-standard noun phrase annotation. For the goal of isolating the contribution of gesture features for coreference, it seems undesirable to deliberately introduce noise in the noun phrase boundaries. Gold standard noun phrase annotations have been used in previous research on coreference resolution, (e.g., McCallum & Wellner, 2004; Haghighi & Klein, 2007).[8] In addition, automatic noun phrase chunking is now possible with high

[7]Sentence boundaries were annotated by the author according to the NIST Rich Transcription Evaluation (NIST, 2003).

[8]The cited references do not include noun phrases unless they participate in coreference relations. This substantially simplifies the coreference task by eliminating the need to disambiguate "singleton" items. In the work described in this chapter, singleton noun phrases are included.

accuracy. F-measures exceeding .94 have been reported on textual corpora (Kudo & Matsumoto, 2001; Sha & Pereira, 2003); on transcripts of the SWITCHBOARD corpus, state-of-the-art performance exceeds .91 (Li & Roth, 2001).

The annotation of noun phrases followed the MUC task definition for "markable" NPs (Hirschman & Chinchor, 1998). Personal pronouns were not annotated, as the discourse focused on descriptions of mechanical devices. Such pronouns could easily be filtered out automatically. Annotation attempted to transcribe all other noun phrases.

The gold standard coreference and markable annotation was performed by the author, using both the audio and video information. Appendix A.1 (page 120) shows the coreference annotations for one conversation in the dataset. Additional coreference annotations were performed by a second rater, permitting an assessment of interrater agreement. This rater is a native speaker of English, a graduate student in computer science, and is not an author on any paper published in connection with this research. She annotated two documents, comprising a total of 270 noun phrases.

Using the interrater agreement methodology described by Passonneau (1997), a score of .65 is obtained on Krippendorf's alpha. This level of agreement is typical for coreference on spoken language. On a corpus of spoken monologues ("Pear Stories"; Chafe, 1980), Passonneau (2004) reports coreference scores ranging from .49 to .74, depending on the story. Using the TRAINS (Heeman & Allen, 1995) corpus of travel planning dialogues, Poesio and Artstein (n.d.) investigate a number of variations of interrater scoring schemes, though the maximum reported alpha score is .67. On a corpus of multi-party spoken dialogues, Müller (2007) finds that agreement for pronoun resolution is low, ranging from .43 to .52.

Evaluation Metric

Coreference resolution is often performed in two phases: a binary classification phase, in which the likelihood of coreference for each pair of noun phrases is assessed; and a global partitioning phase, in which the clusters of mutually-coreferring NPs are formed (e.g., Cardie & Wagstaff, 1999; Soon et al., 2001). This dissertation does

not address the global partitioning phase; it considers only the question of whether each pair of noun phrases in the document corefer. Moving from pairwise noun phrase coreference to global partitioning requires a clustering step to ensure that the pairwise decisions are globally consistent. Because conditional modality fusion operates at the level of binary coreference decisions, interposing another processing step only obscures our ability to measure the contributions of this technique.

Moreover, a global evaluation depends on the choice of the clustering algorithm and the mechanism for selecting the number of clusters (or, alternatively, the cut-off value on merging clusters). This parametrization is particularly challenging for our corpus because of the absence of a large dedicated development set, which could be used to set the number of clusters. Consequently, the bulk of evaluation is performed on the binary classification phase. However, for the purpose of comparing with prior work on coreference, a global evaluation is also performed, measuring the overall results after clustering.

For the binary evaluation, the area under the receiver-operating characteristic (ROC) curve (AUC) is used as a performance metric (Bradley, 1997). AUC evaluates classifier performance without requiring the specification of a cutoff. This metric penalizes misorderings – cases in which the classifier ranks negative examples more highly than positive examples. ROC analysis is increasingly popular, and has been used in a variety of NLP tasks, including the detection of action items in emails (Bennett & Carbonell, 2007) and topic segmentation (Malioutov & Barzilay, 2006).

The global evaluation uses the constrained entity-alignment f-measure (CEAF) for evaluation (Luo, 2005). This metric avoids well-known problems with the earlier MUC evaluation metric (Vilain, Burger, Aberdeen, Connolly, & Hirschman, 1995). The clustering step is performed using two standard techniques from the literature, described in Section 4.4.5. Future work may explore techniques that perform multimodal coreference resolution in a single joint step (e.g., Daumé III & Marcu, 2005). In this case, a global metric would be more appropriate to measure the contributions of gesture and conditional modality fusion.

Baselines

Conditional modality fusion (CMF) is compared against traditional approaches to modality combination for NLP tasks (see Section 2.3.2):

- **Early fusion**: The early fusion baseline includes all features in a single vector, ignoring modality. This is equivalent to standard maximum-entropy classification. Early fusion is implemented with a conditionally-trained log-linear classifier. It uses the same code as the CMF model, but always includes all features.

- **Late fusion**: Late fusion trains separate classifiers for gesture and speech and then combines their posteriors. The modality-specific classifiers are conditionally-trained log-linear classifiers, and again use the same code as the CMF model. For simplicity, a parameter sweep identifies the interpolation weights that maximize performance on the test set. Thus, it is likely that these results somewhat overestimate the performance of these baseline models. There are two versions of late fusion: additive and multiplicative combination of the unimodal posteriors.

- **No fusion**: The "no fusion" baselines are unimodal classifiers for gesture and speech. As with the other baselines, the learning algorithm is still a conditionally-trained log-linear classifier. The implementation uses the same code as the CMF model, but weights on features outside the target modality are forced to zero.

An important question is how these results compare with existing state-of-the-art coreference systems. The "no fusion, verbal features only" baseline provides a reasonable representation of prior work on coreference, by applying a maximum-entropy classifier to a set of typical textual features. A direct comparison with existing implemented systems would be ideal, but all such available systems rely on textual features that are inapplicable to our dataset, such as punctuation, capitalization, and gazetteers of country names and corporations. All systems in the evaluation are summarized in Table 4.4.

CMF different-different (DD)	Uses two different sets of weights for both verbal and gestural features, depending on the hidden variable (equation 4.6).
CMF different-zero (DZ)	Uses different weights on the verbal features depending on the hidden variable; if the hidden variable indicates non-salience, gesture weights are set to zero (equation 4.5).
CMF same-zero (SZ)	Uses the same weights on verbal features regardless of gesture salience; if the hidden variable indicates non-salience, gesture weights are set to zero (equation 4.4).
Early fusion (E)	Standard log-linear classifier. Uses the same weights on verbal and gestural features, regardless of hidden variable (equation 4.3).
Late fusion, multiplicative (LM)	Trains separate log-linear classifiers for gesture and verbal features. Combines posteriors through multiplication.
Late fusion, additive (LA)	Trains separate log-linear classifiers for gesture and verbal features. Combines posteriors through interpolation.
No fusion, verbal only (VO)	Uses only verbal features for classification.
No fusion, gesture only (GO)	Uses only gesture features for classification.

Table 4.4: Summary of systems compared in the coreference evaluation

Parameter Tuning

As the small size of the corpus did not permit dedicated test and training sets, results are computed using leave-one-out cross-validation, with one fold for each of the sixteen documents in the corpus. Parameter tuning was performed using cross-validation within each training fold. This includes the selection of the regularization constant, which controls the trade-off between fitting the training data and learning a model that is simpler (and thus, potentially more general). In addition, binning of continuous features was performed within each cross-validation fold, using the method described in Section 4.2.2. Finally, as noted above, model weights are initialized to zero, enabling deterministic reproducibility of the experiments.

model	AUC
1. CMF different-zero	**.8226**
2. CMF different-different	.8174
3. CMF same-zero	.8084
4. Early fusion (same-same)	.8109
5. Late fusion, multiplicative	.8103
6. Late fusion, additive	.8068
7. No fusion (verbal features only)	.7945
8. No fusion (gesture features only)	.6732

Table 4.5: Coreference performance, in area under the ROC curve (AUC), for systems described in Table 4.4

4.4.4 Results

Table 4.5 gives results for both the baseline and experimental conditions. The gesture-only condition (line 8) is significantly better than chance, which is .5 AUC ($p < .01, t(15) = 15.5$). This shows that it is possible to predict noun phrase coreference purely from gestural information, supporting the hypothesis that gestural similarity correlates with this discourse phenomenon.

Moreover, the multimodal systems all outperform the verbal-only baseline. Even the worst-performing model combination technique – additive late fusion (line 6) – is significantly better than the verbal-only case ($p < .01, t(15) = 3.44$). This shows that gestural similarity provides information not captured by the verbal features.

Conditional modality fusion outperforms all other model-combination approaches by a statistically significant margin. Compared with early fusion, the different-zero model for conditional modality fusion offers an absolute improvement of 1.17% in area under the ROC curve (AUC) – compare lines 1 and 4 in the table. A paired t-test shows that this result is statistically significant ($p < .01, t(15) = 3.73$). CMF obtains higher performance on fourteen of the sixteen cross-validation folds. Both additive and multiplicative late fusion perform on par with early fusion. The p-values of the significance tests for of all pairwise comparisons are shown in Table 4.6.

Early fusion with gesture features is superior to unimodal verbal classification by an absolute improvement of 1.64% AUC ($p < .01, t(15) = 4.45$) – compare lines 4 and

	DD	SZ	E	LM	LA	VO	GO
CMF different-zero (DZ)	.01	.01	.01	.01	.01	.01	.01
CMF different-different (DD)		.05	ns	ns	.05	.01	.01
CMF same-zero (SZ)			ns	ns	ns	.05	.01
Early fusion (E)				ns	ns	.01	.01
Late fusion, multiplicative (LM)					ns	.01	.01
Late fusion, additive (LA)						.01	.01
Verbal features only (VO)							.01
Gesture features only (GO)							

Table 4.6: P-values of the pairwise comparison between models. "ns" indicates that the difference in model performance is not significant at $p < .05$. The parentheses in the left column explain the abbreviations in the top line.

7 in Table 4.5. The additional 1.17% AUC provided by conditional modality fusion amounts to a relative 73% increase in the power of the gesture features.

The results are robust to variations in the regularization constant, which controls the tradeoff between fitting the training data and learning simpler, more general models. As shown in Figure 4-5, the performance of all methods are relatively consistent across a wide range of values for the regularization constant, with conditional modality fusion consistently outperforming the baseline alternatives.

As noted in Section 4.3.2, conditional modality fusion optimizes a non-convex objective, meaning that local search is not guaranteed to find the global optimum. One danger is that the observed results may be sensitive to initialization. This was tested by re-running the CMF different-zero method with five randomized initializations. The resulting standard deviation of the AUC is $1.09 * 10^{-3}$, indicating that performance is fairly stable. In all other experiments the weights were initialized to zero, enabling the results to be reproduced deterministically.

4.4.5 Global Metric

Coreference is traditionally evaluated with a global error metric. However, the research described in this chapter is directed specifically at the binary classification of coreference between pairs of noun phrases. Thus, the focus of evaluation has been on

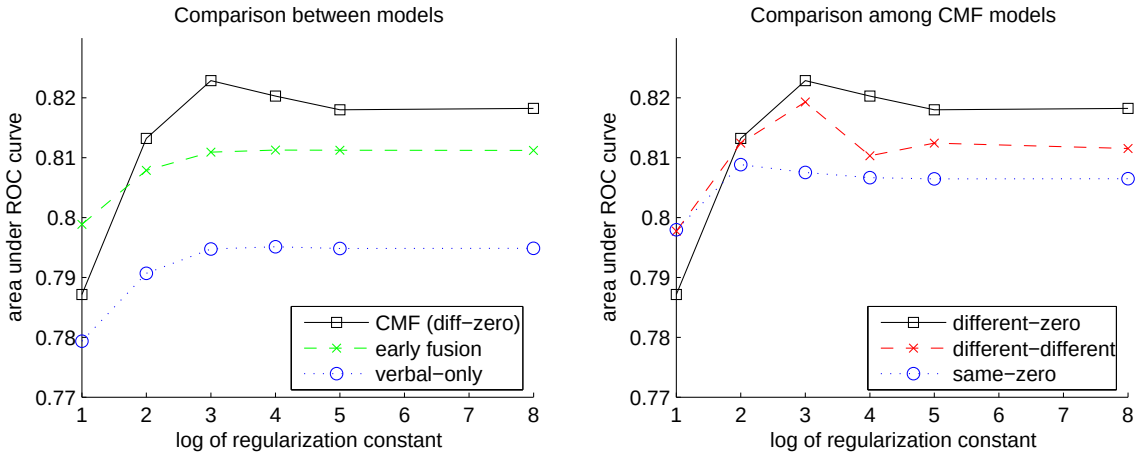


Figure 4-5: Results with regularization constant

model	**first-antecedent**	**best-antecedent**
CMF (different-zero)	55.67	56.02
CMF (different-different)	54.71	56.20
CMF (same-zero)	53.91	55.32
Early fusion (same-same)	54.18	55.50
Late fusion, multiplicative	53.74	54.44
Late fusion, additive	53.56	55.94
No fusion (verbal features only)	53.47	55.15
No fusion (gesture features only)	44.68	44.85

Table 4.7: CEAF global evaluation scores, using best clustering threshold

that specific portion of the larger coreference problem. Nonetheless, for the purpose of comparing with prior research on coreference, a more traditional global metric is also considered.

To perform a global evaluation, the noun phrases in the document were clustered using the pairwise coreference likelihoods as a similarity metric. Two clustering methods from the literature are considered. The **first-antecedent** technique resolves noun phrases to the first antecedent whose similarity is above a predefined threshold (Soon et al., 2001). The **best-antecedent** technique resolves each noun phrase to the most compatible prior noun phrase, unless none is above the threshold (Ng & Cardie, 2002).

Figure 4-6 shows the global scores plotted against the value of the clustering

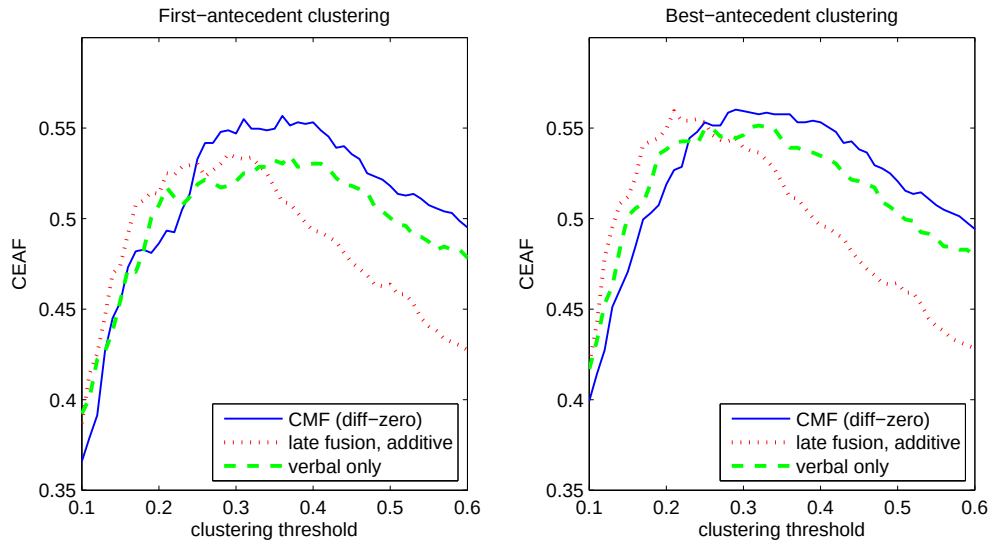
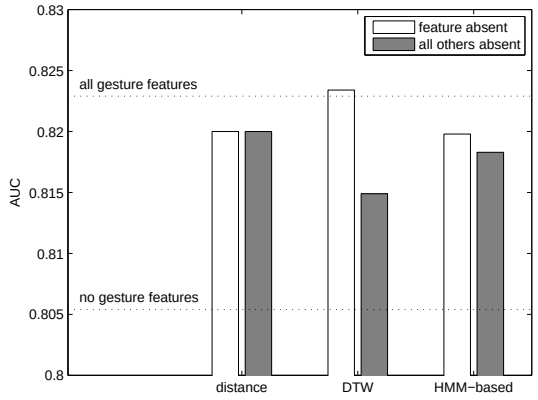


Figure 4-6: Global coreference performance, measured using CEAF scores, plotted against the threshold on clustering

threshold. For clarity, only the best performing system from each class is shown: for conditional modality fusion, this is the different-zero model; from the multimodal baselines, the additive late fusion model is plotted (the combination of additive late fusion and the best-antecedent clustering method is the best performing multimodal baseline); of the unimodal baselines, the verbal-features only system. Table 4.7 lists the performance of each method at its optimum clustering threshold. Ng (2007) reports a CEAF score of 62.3 on the ACE dataset, although the results are not directly comparable due to the differences in corpora.

As shown in these results, performance is sensitive to both the clustering method and the clustering threshold. Conditional modality fusion generally achieves the best results, and best-antecedent clustering generally outperforms the first-antecedent technique. Still, the advantage of conditional modality fusion is smaller here than with ROC analysis. ROC analysis demonstrates the advantage of conditional modality fusion more directly, while the global metric interposes a clustering step that obscures differences between the classification techniques. Nonetheless, the global metric may be a better overall measure of the quality of coreference for downstream applications such as search or summarization. In the text domain, some researchers have demon-



feature group	+	-
all gesture similarity features	.8226	.8054
FOCUS-DISTANCE	.8200	.8200
DTW-AGREEMENT	.8149	.8234
HMM-based	.8183	.8198

Figure 4-7: An analysis of the contributions of each set of gestural similarity features. The "plus" column on the left of the table shows results when only that feature set was present – this information is also shown in the white bars in the graph. The "minus" column shows results when only that feature was removed – this information is shown in the shaded bars. As before, the metric is area under the ROC curve (AUC).

strated global models of coreference that do not require separate classification and clustering phases (e.g., Daumé III & Marcu, 2005). Combining such models with conditional modality fusion is a topic for future work.

4.4.6 Feature Analysis

This evaluation permits a novel analysis comparing the linguistic contribution of the gesture similarity features in the presence of verbal features, enabling the investigation of which gesture features supply unique information over and above the verbal features. All statistical significance results are based on two-tailed, paired t-tests.

Figure 4-7 shows the contribution of three classes of gestural similarity features: FOCUS-DISTANCE, DTW-AGREEMENT, and the two HMM-based features (SAME-CLUSTER and JS-DIV). The top dotted line in the graph shows performance of the different-zero model with the complete feature set, and the bottom line shows performance of this model without any gestural similarity features.[9]

[9]Note that the baseline of "no gesture features" is higher than the "no fusion (verbal features only)" baseline from Table 4.5. Although the feature groups here are identical, the classifiers are different. The "no fusion (verbal features only)" baseline uses a standard log-linear classifier, while "no gesture features" uses conditional modality fusion, permitting two sets of weights for the verbal features, as shown in equation 4.5.

Each feature group conveys useful information, as performance with any one feature group is always better than performance without gestural similarity features ($p < .01, t(15) = 3.86$ for DTW-AGREEMENT, the weakest of the three feature groups). The performance using only the FOCUS-DISTANCE is significantly better than when only the DTW-AGREEMENT feature is used ($p < .05, t(15) = 2.44$); other differences are not significant. There appears to be some redundancy between the feature groups, as removing any individual feature group does not significantly impair performance if the other two feature groups remain.

While the gains obtained from gesture features are significant, it is important to ask whether this is the upper limit of information that can be obtained from gesture for this task. One way to address this question would be to ask human annotators to indicate what proportion of the noun phrases were disambiguated by gesture, though it is not obvious that such a determination could be made in every case. From inspection, I believe that improved visual processing and feature engineering could yield substantial performance gains. The tracking system has resolution on the order of the size of the entire hand (see Figure 4-3, page 39), and not on the level of individual fingers. However, the dataset contains many examples of gestures that refer to different entities but are distinguished only by the angle of the wrist – such references could not be distinguished by the current tracker. Another opportunity for improvement relates to temporal segmentation. Currently, gestural features are computed over the duration of the entire noun phrase. However, some pointing gestures are very brief, and averaging the features over a long window may obscure the portion of the gesture that conveys relevant information. Section 4.6 considers the possibility of learning more structured models of gesture salience, which might help to focus on the most critical portions of each gesture.

4.5 Keyframe Extraction

The previous sections show that estimating gesture salience through conditional modality fusion can improve performance on coreference resolution. However, it is not

clear whether these estimates of gesture salience cohere with human perception, or whether they improve performance for some reason that is artifactual to the specific implementation. This section explores the question of whether gestures judged to be salient by conditional modality fusion are also useful for human viewers. Specifically, gestures estimated to be salient are used to create keyframe summaries of video. The keyframes selected in this way are shown to match those selected by human raters; indeed, this technique outperforms comparable unsupervised text and image-based algorithms from prior work.

Section 4.5.1 explains the keyframe-based summarization task. The basic modeling approach is described in Section 4.5.2. The evaluation setup is presented in Section 4.5.3, and Section 4.5.4 gives the experimental results.

4.5.1 Motivation

The goal of keyframe summarization is to produce a "comic book," in which a textual transcript is augmented with panels, or *keyframes* – still images that clarify the accompanying text. Keyframe-based summaries allow viewers to quickly review key points of a video presentation, without requiring the time and hardware necessary to view the actual video (Boreczky, Girgensohn, Golovchinsky, & Uchihashi, 2000). A major assumption of this thesis is that textual transcriptions alone cannot capture all relevant information. A keyframe-based summary may supplement the transcript with the minimal visual information required for understanding. Figure 4-8 presents an excerpt from a summary produced by the system described in this section.

Existing techniques for keyframe extraction have usually focused on edited videos such as news broadcasts (e.g., Uchihashi, Foote, Girgensohn, & Boreczky, 1999; Boreczky et al., 2000; Zhu, Fan, Elmagarmid, & Wu, 2003). Such systems seek to detect large-scale changes in image features to identify different scenes, and then choose a representative example from each scene. This approach is poorly suited to unedited videos, such as recordings of classroom lectures or business presentations. In such videos, the key visual information is not the variation in scenes or camera angles, but the visual communication provided by the gestures of the speaker. Thus,

a better approach may be to capture keyframes that include salient gestures, using the model developed in Section 4.3.

4.5.2 Identifying Salient Keyframes

One possible method for identifying salient keyframes would be to formulate it as a standard supervised learning task, using a corpus in which salient gestures are annotated. However, such annotation can be avoided by bootstrapping from multimodal coreference resolution, using conditional modality fusion. By learning to predict the specific instances in which gesture helps coreference, we obtain a model of gesture salience. For example, we expect that a pointing gesture in the presence of an anaphoric expression would be found to be highly salient (as in Figure 4-1, page 35); a more ambiguous hand pose in the presence of a fully-specified noun phrase would not be salient. This approach does not identify *all* salient gestures, but does identify those that occur in the context of the selected language understanding task. In coreference resolution, only gestures that co-occur with noun phrases are considered. As noun phrases are ubiquitous in language, this should still cover a usefully broad collection of gestures.

Using the model for coreference resolution introduced in Section 4.3, we obtain a posterior distribution for the hidden variable, which controls whether the gesture features are included for coreference resolution. The basic hypothesis is that gestures that help coreference resolution are likely to be perceptually salient. The positive results on the coreference task support this claim, but do not demonstrate a direct link between the hidden variable and human perceptual judgments of gesture salience. By building keyframe summaries using gestures rated as salient by the model, it is possible to evaluate this hypothesis.

As described in section 4.3, models of coreference resolution and gesture salience are learned jointly. After training, a set of weights $\mathbf{w}_h$ is obtained, allowing the estimation of gesture salience at each noun phrase. We sum over all possible values for y and h_2, obtaining $\sum_{y,h_2} \psi(y, \mathbf{h}, \mathbf{x}; \mathbf{w}) = h_1 \mathbf{w}_h^\mathsf{T} \mathbf{x}_{h_1}$. The potential for the case

when the gesture is salient is found by setting $h_1 = 1$, yielding $\mathbf{w}_h^\mathsf{T}\mathbf{x}_{h_1}$.[10] The key assumption is that this potential is a reasonable proxy for the informativeness of a keyframe that displays the noun phrase's accompanying gesture.

This potential function is used to generate an ordering on the noun phrases in the dialogue. Keyframes are selected from the midpoints of the top n noun phrases, where n is specified in advance by the annotator. Providing the system with the ground truth number of keyframes follows common practice from the textual summarization literature – summaries of different lengths are difficult to compare, as the summary duration is governed partially by the annotator's preference for brevity or completeness (Mani & Maybury, 1999). Each keyframe is given a caption that includes the relevant noun phrase and accompanying text, up to the noun phrase in the next keyframe. A portion of the output of the system is shown in Figure 4-8.

4.5.3 Evaluation Setup

The evaluation methodology for keyframe summaries is similar to the intrinsic evaluation developed for the Document Understanding Conference.[11] The quality of the automatically extracted keyframes is assessed by comparing them to human-annotated ground truth. This section describes the dataset, implementation, evaluation metric, and baseline systems.

Dataset

The dataset consists of a subset of the videos used in the coreference evaluation, described in Section 4.4.3. Of the sixteen videos used for the coreference evaluation, nine were manually annotated for keyframes. Of these, three are used in developing the system and the baselines, and the remaining six are used for final evaluation (these are indicated by asterisks in Table B.1 in Appendix B, page 125). There is no

[10] Note an identical value is obtained by considering the same noun phrase as the anaphor (x_{h_2}) and summing over all possible values of h_1.

[11] `http://duc.nist.gov`

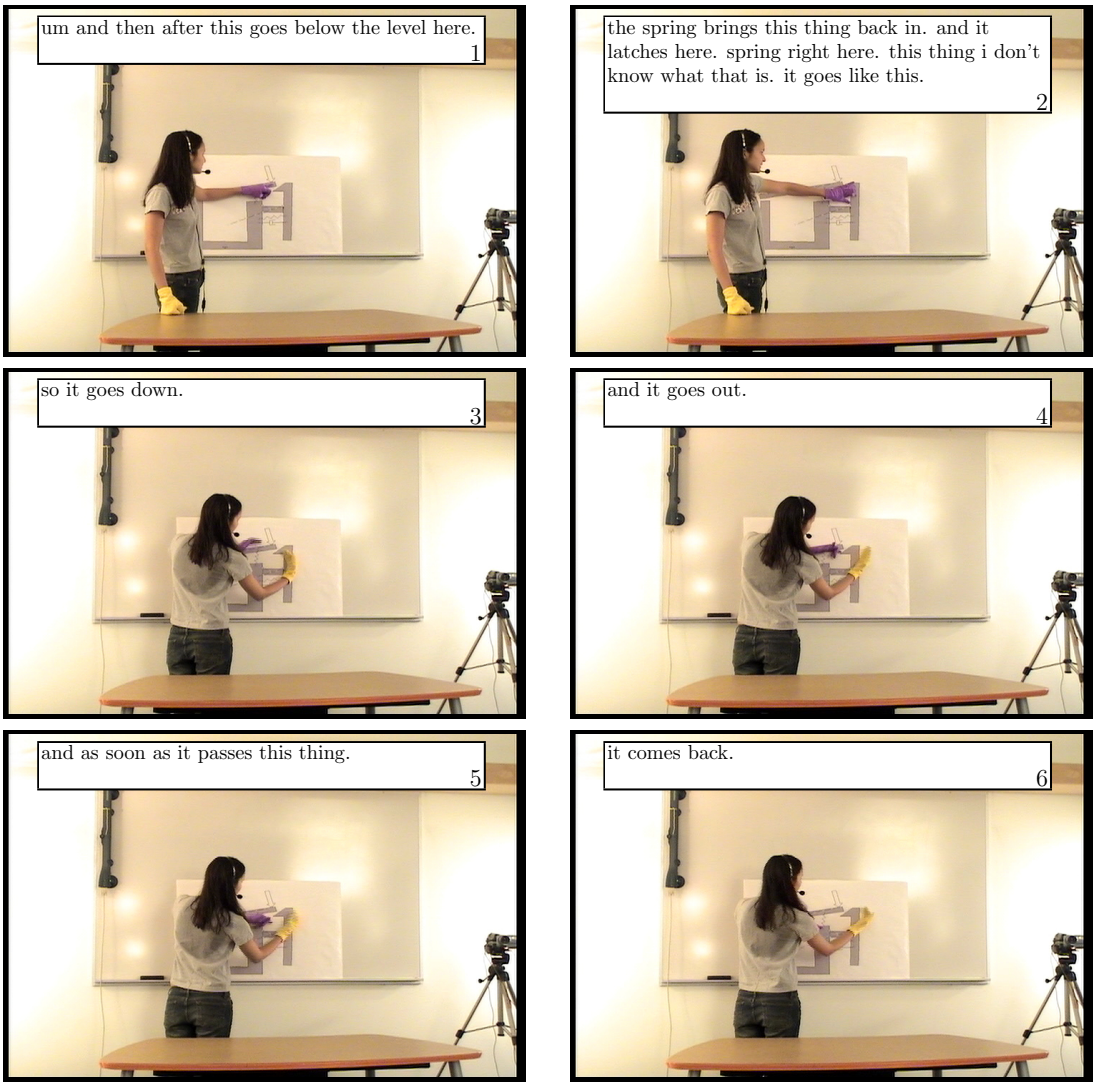


Figure 4-8: Example of the first six frames of an automatically-generated keyframe summary

R1 ↓, R2 →	in-keyframe	not-in-keyframe
in-keyframe	11992	5890
not-in-keyframe	9714	84305

Table 4.8: Agreement counts for the two raters, in numbers of frames

explicit training on the keyframe annotations, but the development set was used for evaluation as the system was under construction.

The specification of the ground truth annotation required that the keyframes capture all static visual information that the annotator deems crucial to understanding the content of the video. The number of selected frames was left to discretion; on average, 17.8 keyframes were selected per document, out of an average total of 4296 frames per document. Annotation was performed by the author and by another student who was not an author on any papers connected with this research. On a subset of two videos annotated by both raters, the raw interrater agreement was 86%, yielding a kappa of .52 (Carletta, 1996). Detailed statistics are given in Table 4.8.

One important difference between this multimodal corpus and standard sentence extraction datasets is that many frames may be nearly identical, due to the high frame rate of video. For this reason, the annotators marked regions rather than individual frames. Regions define equivalence classes, such that any frame from a given region conveys critical visual information, and the information conveyed by all keyframes in a given region is the same. Thus, if a single keyframe were selected from every ground truth region, the result would be the minimal set of keyframes necessary for a reader to fully understand the discourse. On average, 17.8 regions were selected from each video, spanning 568 frames, roughly 13% of the total number of frames per video.

Training Coreference Resolution

As described in Section 4.5.2, the current approach to keyframe extraction is based on a model for gesture salience that is learned from labeled data on coreference resolution. The training phase is performed as leave-one-out cross-validation: a separate set of weights is learned for each presentation, using the other fifteen presentations as a

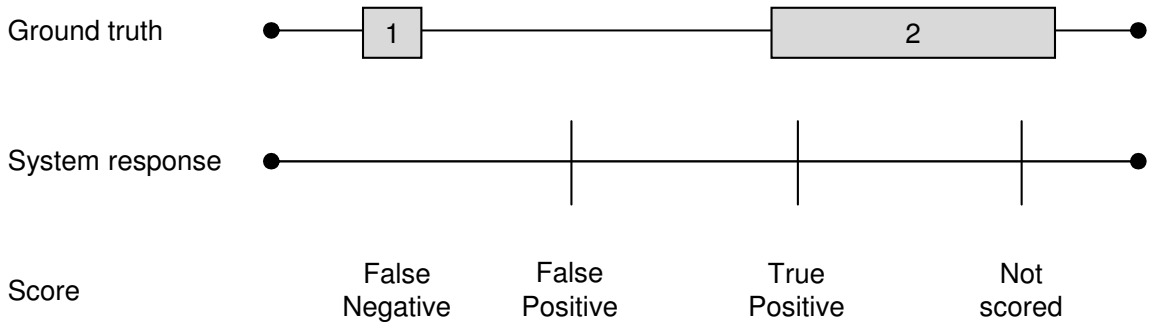


Figure 4-9: An example of the scoring setup

training set. The learned weights are used to obtain the values of the hidden variable indicating gesture salience, as described in Section 4.5.2.

Evaluation Metric

Figure 4-9 illustrates the scoring setup. The top row in the figure represents the ground truth; the middle row represents the system response, with vertical lines indicating selected keyframes; the bottom row shows how the response is scored.

For all systems the number of keyframes is fixed to be equal to the number of regions in the ground truth annotation. If the system response includes a keyframe that is not within any ground truth region, a false positive is recorded. If the system response fails to include a keyframe from a region in the ground truth, this is a false negative. A true positive is recorded for the first frame that is selected from a given ground truth region, but additional frames from the same region are not scored. The system is thus still penalized for each redundant keyframe, because it has "wasted" one of a finite number of keyframes it is allowed to select. Still, such false positives seem less grave than a true substitution error, in which a keyframe not containing relevant visual information is selected. Performance is quantified using the F-measure, which is the harmonic mean of recall and precision.

Baselines

The gesture salience keyframe extractor is compared against three baselines, presented in order of increasing competitiveness.

- **Random-keyframe:** The simplest baseline selects n keyframes at random from the video. This is similar to the "random sentence" baseline common in the textual summarization literature (Mani & Maybury, 1999). The number of keyframes selected in this baseline is equal to the number of regions in the ground truth. This baseline represents a lower bound on the performance that any reasonable system should achieve on this task. The reported scores are averaged across 500 independent runs.

- **NP-salience:** The NP-SALIENCE system is based on frequency-based approaches to identifying salient noun phrases (NPs) for the purpose of text summarization (Mani & Maybury, 1999). The salience heuristic chooses the most common representative tokens of the largest and most homogeneous coreference clusters.[12] The largest cluster is the one containing the most noun phrases; homogeneity is measured by the inverse of the number of unique surface forms. This provides a total ordering on NPs in the document; we select keyframes at the midpoint of the top n noun phrases, where n is the number of keyframe regions in the ground truth. One project for future work is to explore finding the best point *within* each noun phrase for keyframe selection.

- **Pose-clustering:** The final baseline is based purely on visual features. It employs clustering to find a representative subset of frames with minimum mutual redundancy. In a seminal paper on keyframe selection, Uchihashi et al. (1999) perform clustering on all frames in the video, using the similarity of color histograms as a distance metric. Representative images from each cluster are then used as keyframes. More recent video summarization techniques have improved the clustering algorithms (T. Liu & Kender, 2007) and the similarity metric (Zhu et al., 2003), but the basic approach of choosing exemplar keyframes from a clustering based on visual similarity is still widely used in state-of-the-art research on this topic (see Lew, Sebe, Djeraba, & Jain, 2006, for a survey).

[12] Here, coreference clusters are based on manual annotations.

In the current dataset, there is a single fixed camera and no change in the video except for the movements of the speaker. The color histograms are thus nearly constant throughout, precluding the use of color as a clustering feature. Instead, the tracked coordinates of the speaker's hands and upper body are used as the basic features; these are normalized, and a Euclidean distance metric is applied. In this setting, clusters correspond to typical body poses, and segments correspond to holds in these poses. Following Uchihashi et al. (1999), the video is divided into segments in which cluster membership is constant, and keyframes are taken at the midpoints of segments. The importance metric from this paper is used to rank segments; the top n are chosen, where n is the number of keyframes in the ground truth.

4.5.4 Results

The experimental results suggest that the estimates of gesture salience given by conditional modality fusion cohere with human perception. Table 4.9 compares the performance of the salience-based approach with the three baselines. Using paired t-tests, GESTURE-SALIENCE significantly outperforms all alternatives ($p < .05$ in all cases). The POSE-CLUSTERING and NP-SALIENCE systems are statistically equivalent; both are significantly better than the RANDOM-KEYFRAME baseline ($p < .05$).

This set of baselines is necessarily incomplete, as there are many ways in which keyframes extraction could be performed. For example, prosodic features could be used to identify moments of particular interest in the dialogue (Sundaram & Chang, 2003). In addition, a combination of baselines including visual and linguistic features may also perform better than any individual baseline. However, developing more complicated baselines is somewhat beside the point. The evaluation demonstrates that a simple yet effective technique for selecting meaningful keyframes can be obtained as a byproduct of conditional modality fusion.

A manual inspection of the system output revealed that in many cases our system selects a noun phrase that is accompanied by a relevant gesture, but the specific keyframe was slightly off. The current method always chooses the keyframe at the

Method	F-Measure	Recall	Precision
GESTURE-SALIENCE	**.404**	**.383**	**.427**
POSE-CLUSTERING	.290	.290	.290
NP-SALIENCE	.239	.234	.245
RANDOM-KEYFRAME	.120	.119	.121

Table 4.9: Comparison of performance on keyframe selection task

midpoint of the accompanying noun phrase; often, the relevant gesture is brief, and does not necessarily overlap with the middle of the noun phrase. Thus, one promising approach to improving results would be to "look inside" each noun phrase, using local gesture features to attempt to identify the specific frame in which the gesture is most salient.

Other errors arise because some key gestures are not related to noun phrases. For example, suppose the speaker says "it shoots the ball up," and accompanies only the word "up" with a gesture indicating the ball's trajectory. This gesture might be important to understanding the speaker's meaning, but since it does not overlap with a noun phrase, the gesture will not be identified by our system. Nonetheless, on balance the results show that focusing on noun phrases is a good start for linguistically-motivated keyframe extraction, and that this unsupervised approach is successful at identifying the noun phrases that require keyframes. As gesture is applied to other language tasks, it will be possible to model gesture salience at other phrase types, thus increasing the coverage for keyframe extraction.

4.6 Discussion

This chapter is motivated by the idea that gestural similarity sheds light on local discourse phenomena. We find that when semantically related noun phrases are accompanied by gestures, those gestures tend to be similar. Moreover, features that quantify gesture similarity improve noun phrase coreference when applied in concert with verbal features. This suggests that gesture provides a non-redundant source of information. This is not merely an engineering improvement; when asked to build a

visual summary of a dialog, human raters chose keyframes that were similar to those deemed helpful by the coreference system.

A second key finding is that a structured approach to multimodal integration is crucial to achieving the full linguistic benefits offered by gesture features. Rather than building separate verbal and gesture interpretation units – or simply concatenating their features – conditional modality fusion enables the construction of potential functions whose structure encodes the role of each modality. In particular, gesture supplements speech only intermittently, and therefore gesture salience is represented explicitly with a hidden variable. This approach yields a 73% relative improvement in the contribution of the gesture features towards coreference resolution. This improvement is attained by modeling gesture salience with a hidden variable and ignoring gestures that are not salient.

Conditional modality fusion induces an estimate of gesture salience within the context of a specific linguistic task. To test the generality of the salience model, the derived estimates were transferred to a different task: keyframe extraction. Without any labeled data on the keyframe task, this simple algorithm outperforms competitive unimodal alternatives. This suggests that the model of gesture salience learned from coreference coheres with human perception of gesture salience.

One interesting direction for future research to investigate richer models of gesture salience. The structure explored in this chapter is minimal – a binary variable to indicate the salience of a gesture for coreference resolution. I see this as a first step towards more complex structural representations for gesture salience that may yield greater gains in performance. For example, it is likely that gesture salience observes some temporal regularity, suggesting a Markov state model. Indeed, just such a state model is suggested by Kendon's taxonomy of "movement phases," as discussed in Section 2.1.3.

5

Gestural Cohesion and High-Level Discourse Structure

We now move from local discourse processing to document-level discourse structure. The concept of gestural similarity – measured over pairs of gestures – is extended to gestural cohesion, which describes the consistency of an entire set of gestures. Applying gestural cohesion to the task of topic segmentation yields a multimodal segmenter that performs substantially better than comparable text-only alternatives. A related approach is used to investigate the influence of discourse topic on gestural form. When multiple speakers describe a single topic, their gestures are significantly more similar than when each speaker describes a different topic. This shows that the

influence of discourse topic on gestural form can generalize across speakers.[1]

5.1 Introduction

This chapter will focus on *gestural cohesion*, which extends the idea of gestural similarity beyond pairs of gestures, allowing us to examine the self-consistency of the entire set of gestures accompanying a dialogue or discourse segment. The term "gestural cohesion" is chosen to evoke the parallel concept of lexical cohesion, which captures the self-consistency of word usage within a document or segment (Halliday & Hasan, 1976). Lexical cohesion is a powerful tool for high-level discourse analysis, and has been used repeatedly in tasks such as topic segmentation (e.g. Hearst, 1994; Tur, Hakkani-Tur, Stolcke, & Shriberg, 2001). The key observation in lexical cohesion-based approaches is that topically-coherent text is characterized by the repeated use of a consistent, limited subset of lexical items. When the distribution of lexical items changes abruptly – that is, when a large number of new words enter the discourse – this is taken to indicate that the discourse topic has changed.

Lexical cohesion is effective because meaning is partially expressed through word choice. A complete semantic analysis would also require syntactic and higher-level processing, but lexical cohesion provides a lightweight alternative that can be easily implemented for any text. If gestures communicate semantically relevant information, the discourse structure of a dialogue should also be mirrored in the cohesion structure of gestural features. In this chapter, gestural *codewords* are extracted from raw video, forming a visual lexicon.[2] Techniques from lexical analysis can then be applied directly to gestural features.

This chapter presents two sets of experiments centered on the idea of gestural cohesion. The first experiment focuses on discourse segmentation: the task of dividing a dialogue into sections with unique discourse topics. Both lexical and gestural

[1] Some of the work in this section was published previously (Eisenstein, Barzilay, & Davis, 2008b, 2008a).

[2] This is not a lexicon in the traditional sense, because there is no symbolic meaning assigned to each codeword.

cohesion are applied to this task, using a new, Bayesian approach that facilitates multimodal integration. The results demonstrate that gestural cohesion effectively supplements lexical analysis; the combined multimodal system outperforms lexical-only alternatives, for both manual and automatic transcripts.

Figure 5-1 gives some intuition about how lexical and gestural cohesion correlate with discourse segmentation. The upper part of the figure shows the distribution of lexical items in a single dialogue, with the manually annotated topic boundaries indicated by vertical red lines. Each column represents a sentence, and the blocks in a column indicate the words that the comprise the sentence. The lower part of the figure shows the distribution of automatically extracted gestural codewords for the same video – each column shows the codewords that occur during the duration of the sentence. While noisy, it is possible even from visual inspection to identify some connections between the segmentation and the distribution of codewords – for example, the second-to-last segment has a set of codewords starkly different from its neighbors.

Even if gesture mirrors discourse structure within a dialogue, the mapping between gesture and topic may not be consistent across speakers. Gestures express meaning through spatial metaphors; the extent to which they can be interpreted across speakers depends on the speaker-specificity of these spatial metaphors. The second technical portion of the chapter explores the existence of speaker-general gestural themes that characterize specific topics in this dataset, finding that a small but consistent percentage of gestural forms occur across speakers when they discuss a shared topic. This analysis is performed both in a Bayesian and classical framework.

The previous chapter used a portion of the dataset in which speakers had access to visual aids, emphasizing deictic gestures in which location was the principle communicative gestural feature. This chapter considers another portion of the dataset, in which no visual aids are permitted. Thus, a greater proportion of "iconic" or "illustrative" gestures are observed (McNeill, 1992, see also Section 2.1.3).[3] Consequently,

[3]Even without visual aids, speakers still perform some abstract deictic gestures, assigning space to ideas or concepts. (McNeill, 1992)

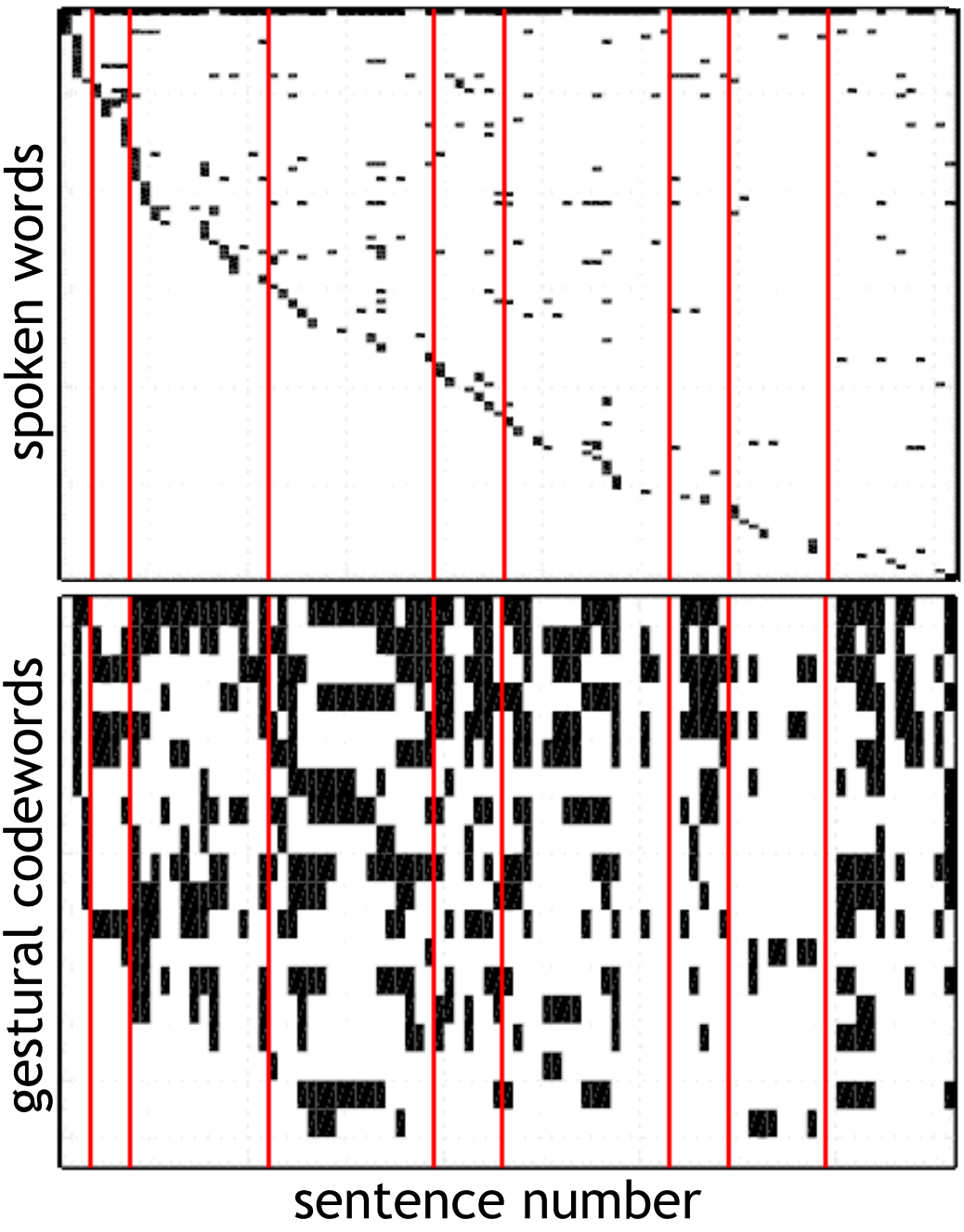


Figure 5-1: Distribution of gestural and lexical features by sentence. Each dark cell means that the word or keyword (indexed by row) is present in the sentence (indexed by the column). Manually annotated segment breaks are indicated by red lines.

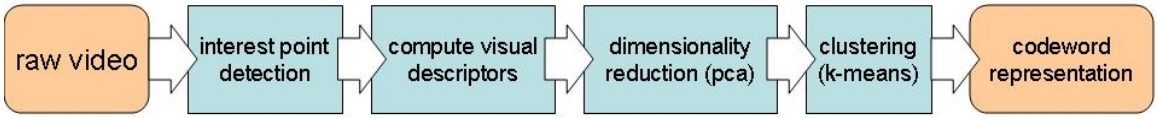


Figure 5-2: The visual processing pipeline for the extraction of gestural codewords from video

a new visual feature set is applied, emphasizing low-level descriptions of physical motion.

The chapter is structured as follows. Section 5.2 describes the visual features used in the experiments in this chapter. This section explains the extraction of low-level image descriptors, and shows how these image descriptors are converted into a lexicon of gestural forms. Gestural and lexical cohesion are applied to topic segmentation in Section 5.3, and a new Bayesian segmentation method is presented. Section 5.4 explores the question of whether gestural forms are shared across speakers for a given topic. A summary of these results is presented in Section 5.5.

5.2 A Codebook of Gestural Forms

This section describes the process of building a codebook representation, which permits the assessment of gestural cohesion. The core image-level features are based on *spatiotemporal interest points*, which provide a sparse representation of the motion in the video. At each interest point, visual, spatial, and kinematic characteristics are extracted and then concatenated into vectors. Principal component analysis (PCA) reduces the dimensionality to a feature vector of manageable size (Bishop, 2006). The feature vectors are then clustered, yielding a codebook of gestural forms. This video processing pipeline is shown in Figure 5-2; the remainder of the section describes the individual steps in greater detail.

5.2.1 Spatiotemporal Interest Points

Spatiotemporal interest points (Laptev, 2005) provide a sparse representation of video. The idea is to select a few local regions that contain high information content in both the spatial and temporal dimensions. The image features at these regions should be relatively robust to lighting and perspective changes, and they should capture the relevant movement in the video. Thus, the set of spatio-temporal interest points should provide a highly compressed representation of the key visual-kinematic features. Purely spatial interest points have been widely successful in a variety of image processing tasks (Lowe, 1999), and spatio-temporal interest points are beginning to show similar advantages for video processing (Laptev, 2005).

The use of spatiotemporal interest points is motivated by research in the computer vision domain of *activity recognition* (Efros, Berg, Mori, & Malik, 2003; Niebles, Wang, & Fei-Fei, 2006). The goal of activity recognition is to classify video sequences into semantic categories: e.g., walking, running, jumping. As a simple example, consider the task of distinguishing videos of walking from videos of jumping. In the walking videos, the motion at most of the interest points will be horizontal, while in the jumping videos it will be vertical. Spurious vertical motion in a walking video is unlikely to confuse the classifier, as long as the majority of interest points move horizontally. The hypothesis of this section is that just as such low-level movement features can be applied in a supervised fashion to distinguish activities, they can be applied in an unsupervised fashion to group co-speech gestures into perceptually meaningful clusters.

The Activity Recognition Toolbox (Dollár, Rabaud, Cottrell, & Belongie, 2005)[4] is used to detect spatiotemporal interest points for our dataset. This toolbox ranks interest points using a difference-of-Gaussians filter in the spatial dimension, and a set of Gabor filters in the temporal dimension. The total number of interest points extracted per video is set to equal the number of frames in the video. This bounds the complexity of the representation to be linear in the length of the video; however, the system may extract many interest points in some frames and none in other frames.

[4] `http://vision.ucsd.edu/~pdollar/research/cuboids_doc/index.html`

Figure 5-3: Circles indicate the interest points extracted at this frame in the video.

Figure 5-3 shows the interest points extracted from a representative video frame from the segmentation corpus. Note that the system has identified high contrast regions of the gesturing hand. From manual inspection, the large majority of interest points extracted in our dataset capture motion created by hand gestures. Thus, for this dataset it is reasonable to assume that an interest point-based representation expresses the visual properties of the speakers' hand gestures. In videos containing other sources of motion, preprocessing may be required to filter out interest points that are extraneous to gestural communication.

5.2.2 Visual Features

At each interest point, temporal and spatial brightness gradients are constructed across a small space-time volume of nearby pixels. Brightness gradients have been used in a variety ofcomputer vision applications (Forsyth & Ponce, 2003), and provide a fairly general way to describe the visual appearance of small image patches. However, even for a small space-time volume, the resulting dimensionality is still quite

large: a 10-by-10 pixel box across 5 video frames yields a 500-dimensional feature vector for each of the three gradients. For this reason, principal component analysis (Bishop, 2006) is used to reduce the dimensionality to a more manageable size. The spatial location of the interest point is added to the final feature vector.

This visual feature representation is at a lower level of abstraction than the articulated model used in Chapter 4; it is substantially lower-level than the descriptions of gestural form found in both the psychology and computer science literatures. For example, when manually annotating gesture, it is common to employ a taxonomy of hand shapes and trajectories, and to describe the location with respect to the body and head (McNeill, 1992; Martell, 2005). Working with automatic hand tracking, Quek et al. automatically compute perceptually-salient gesture features, such as holds (Bryll et al., 2001) and oscillatory repetitions (Xiong & Quek, 2006).

In contrast, the interest point representation takes the form of a vector of continuous values and is not easily interpretable in terms of how the gesture actually appears. However, this low-level approach offers several important advantages. Most critically, it requires no initialization and comparatively little tuning: it can be applied directly to any video with a fixed camera position and static background. Second, it is robust: while image noise may cause a few spurious interest points, the majority of interest points should still guide the system to an appropriate characterization of the gesture. In contrast, hand tracking can become irrevocably lost, requiring manual resets (Gavrila, 1999). Finally, the success of similar low-level interest point representations at the activity-recognition task provides reason for optimism that they may also be applicable to unsupervised gesture analysis.

5.2.3 Gesture Codewords for Discourse Analysis

The previous section describes the extraction of low-dimensional feature vectors that characterize the visual appearance at sparse spatiotemporal interest points. Using k-means clustering (Bishop, 2006), these feature vectors are grouped into *codewords*: a compact, lexicon-like representation of salient visual features in video. The number of clusters is a tunable parameter.

Codewords capture frequently-occurring patterns of motion and appearance at a local scale. For example, Figure 5-6 (page 102) shows three examples of a single codeword; in all cases, the salient visual characteristic is upward motion of a light object against a dark background.[5] Instances of codewords are detected at specific locations and times throughout each video. By grouping together visually similar interest points, the set of codeword types forms a sort of visual vocabulary for each video.

The codewords that occur during a given period of time, such as a spoken sentence, provide a succinct representation of the ongoing gestural activity. Distributions of codewords over time can be analyzed in similar terms to the distribution of lexical features. A change in the distribution of codewords indicates new visual kinematic elements entering the discourse. If the codeword representation succeeds at capturing salient perceptual features of gesture, then it will allow gestural cohesion to be assessed in much the same way as lexical cohesion.

5.3 Discourse Segmentation

This section describes how lexical and gestural cohesion can be combined to predict discourse segmentation. The goal is to divide each dialogue into topically coherent units. While a variety of algorithms have been applied to discourse segmentation, features based on lexical cohesion have formed the backbone of many such approaches (e.g. Hearst, 1994; Beeferman, Berger, & Lafferty, 1999). This section describes a new algorithm for discourse segmentation, permitting the flexible combination of lexical and gestural cohesion features in an integrated Bayesian framework. The resulting multimodal segmentation system outperforms unimodal, lexical-only approaches.

Previous approaches to discourse segmentation – including multimodal segmentation using prosody – are presented in Section 5.3.1. Section 5.3.2 describes the

[5]Note that in the topic segmentation experiments in Section 5.3, clustering is performed only within a single video, and not across speakers. Clustering across multiple videos and speakers is performed in Section 5.4.

Bayesian discourse segmentation model. Experiments for evaluating the contribution of gestural cohesion for discourse segmentation are presented in Section 5.3.3, with results given in Section 5.3.4.

5.3.1 Prior Work

Lexical Cohesion for Discourse Segmentation Hearst (1994) showed that lexical cohesion can be applied to discourse segmentation. Her approach, called TextTiling, computes an evolving metric of lexical cohesion and places segment boundaries at local minima of this metric. Later approaches use similar feature sets, but apply other segmentation algorithms, such as exponential models (Beeferman et al., 1999) and graph-theoretic techniques (Utiyama & Isahara, 2001; Malioutov & Barzilay, 2006).

Of particular relevance to this chapter are segmentation algorithms based on hidden Markov models (HMMs). One early example is the work of Yamron, Carp, Gillick, Lowe, and Mulbregt (1998), who construct an HMM by building topic-specific language models and then perform segmentation by finding the maximum likelihood path through the topics, for each document. Tur et al. (2001) apply a similar approach, but add special states to model features that occur at the beginning and end of segments. Both of these approaches train the topic models off-line, which is suboptimal, rather than learning them jointly with the segmentation.

Purver, Griffiths, Körding, and Tenenbaum (2006) overcome this problem, inferring the segmentation and topic models jointly via Gibbs sampling. However, like the earlier HMM-based approaches, they model topics across multiple documents. This means that new documents cannot be segmented unless they contain topics already observed in the training set. In addition, all known HMM-based approaches search in the space of segment *labels*, rather than in the space of segmentations. Each segmentation implies multiple possible labellings, because the label indices can be permuted; thus, the space of labellings is larger by a factor of $N!$, where N is the number of segments. The algorithm presented in Section 5.3.2 avoids both problems. It does not require a document-general model of topics, and so can operate on individual documents. In addition, it searches directly in the space of segmentations.

Nonverbal Features for Segmentation Research on nonverbal features for topic segmentation has focused primarily on prosody, under the assumption that a key prosodic function is to mark structure at the discourse level (Steedman, 1990; Grosz & Hirshberg, 1992; Swerts, 1997). The ultimate goal of such research is to find correlates of hierarchical discourse structure in phonetic features. Today, research on prosody has converged on a set of prosodic cues that correlate with discourse structure. Such markers include pause duration, fundamental frequency, and pitch range manipulations (Grosz & Hirshberg, 1992; Hirschberg & Nakatani, 1998). These studies informed the development of applications such as segmentation tools for meeting analysis, e.g. (Tur et al., 2001; Galley, McKeown, Fosler-Lussier, & Jing, 2003).

In general, attempts to apply prosody to discourse segmentation have focused on identifying prosodic markers of segment boundaries. Such markers are similar to cue phrases (Litman, 1996) – words or phrases that explicitly mark segment boundaries. In contrast, gestural cohesion seeks to identifying segmentation points that preserve intra-segmental consistency, paralleling lexical cohesion of Hearst (1994). This suggests that prosody and gesture convey discourse information in orthogonal ways. Thus, the combination of these two modalities may further improve performance, suggesting interesting possibilities for future work.

The connection between gesture and discourse structure is a relatively unexplored area, at least with respect to computational approaches. Quek et al. investigate the relationship between discourse segmentation and symmetric hand movements (Quek, Xiong, & McNeill, 2002) and the use of space (Quek, McNeill, Bryll, & Harper, 2002). While these papers demonstrate correlations between gestural features and segment boundaries, neither shows that gesture can be used to *predict* segment boundaries on unlabeled text. Another difference is that this prior work does not investigate whether gestural features supplement lexical cues with novel information. This line of research is discussed in more detail in Section 2.1.2.

5.3.2 Bayesian Topic Segmentation

Topic segmentation is performed in a Bayesian framework, with each sentence's segment index encoded in a hidden variable, written z_t. The hidden variables are assumed to be generated by a linear segmentation, such that $z_t \in \{z_{t-1}, z_{t-1}+1\}$. Observations – the words and gesture codewords – are generated by multinomial language models that are indexed according to the segment. In this framework, a high-likelihood segmentation will include language models that are tightly focused on a compact vocabulary. Such a segmentation maximizes the lexical cohesion of each segment. This model thus provides a principled, probabilistic framework for cohesion-based segmentation, and we will see that the Bayesian approach is particularly well-suited to the combination of multiple modalities.

Formally, our goal is to identify the best possible segmentation S, where S is a tuple: $S = \langle \mathbf{z}, \theta, \phi \rangle$. The segment indices for each sentence are written z_t; for segment i, θ_i and ϕ_i are multinomial language models over words and gesture codewords respectively. For each sentence, $\mathbf{x}_t$ and $\mathbf{y}_t$ indicate the words and gestures that appear. We seek to identify the segmentation $\hat{S} = \text{argmax}_S p(S, \mathbf{x}, \mathbf{y})$, conditioned on priors that will be defined below. The joint probability is written,

$$p(S, \mathbf{x}, \mathbf{y}) = p(\mathbf{x}, \mathbf{y}|S)p(S),$$

where

$$p(\mathbf{x}, \mathbf{y}|S) = \prod_i p(\{x_t : z_t = i\}|\theta_i)p(\{y_t : z_t = i\}|\phi_i), \tag{5.1}$$

$$p(S) = p(\mathbf{z}) \prod_i p(\theta_i)p(\phi_i). \tag{5.2}$$

The language models θ_i and ϕ_i are multinomial distributions, so the log-likelihood of the observations $\mathbf{x}_t$ is $\log p(\mathbf{x}_t|\theta_i) = \sum_j^W n(t,j) \log \theta_{i,j}$, where $n(t,j)$ is the count of word j in sentence t, and W is the size of the vocabulary. An analogous equation is used for the gesture codewords. Each language model is given a symmetric

Dirichlet prior α. As we will see shortly, the use of different priors for the verbal and gestural language models allows us to weight these modalities in a Bayesian framework. Finally, we model the probability of the segmentation $\mathbf{z}$ by considering the durations of each segment: $p(\mathbf{z}) = \prod_i p(\text{dur}(i)|\psi)$. A negative-binomial distribution with parameter ψ is applied to discourage extremely short or long segments.

Inference Crucially, both the likelihood (equation 5.1) and the prior (equation 5.2) factor into a product across the segments. This factorization enables the optimal segmentation to be found using a dynamic program, similar to those demonstrated by Utiyama and Isahara (2001) and Malioutov and Barzilay (2006). For each set of segmentation points $\mathbf{z}$, the associated language models are set to their posterior expectations, e.g., $\theta_i = E[\theta|\{x_t : z_t = i\}, \alpha]$.

The Dirichlet prior is conjugate to the multinomial, so this expectation can be computed in closed form:

$$\theta_{i,j} = \frac{n(i,j) + \alpha}{N(i) + W\alpha}, \tag{5.3}$$

where $n(i,j)$ is the count of word j in segment i and $N(i)$ is the total number of words in segment i (Bernardo & Smith, 2000). The symmetric Dirichlet prior α acts as a smoothing pseudo-count. In the multimodal context, the priors act to control the weight of each modality. If the prior for the verbal language model θ is high relative to the prior for the gestural language model ϕ then the verbal multinomial will be smoother, and will have a weaker impact on the final segmentation. The impact of the priors on the weights of each modality is explored in Section 5.3.4.

Estimation of priors The distribution over segment durations is negative-binomial, with parameter $\psi = \langle \bar{x}, k \rangle$, where $\bar{x}$ is the expected duration and k is a dispersion parameter.[6] In general, the maximum likelihood estimate of the dispersion parameter k cannot be found in closed form (Gelman et al., 2004).

Suppose we have a set of durations generated from the negative-binomial distribution, written $x_1 \ldots x_M$. Assuming a non-informative prior on k, $p(k|\mathbf{x}, \bar{x}) \propto p(\mathbf{x}|\bar{x}, k)$,

[6] This is different from the standard parametrization (Gelman, Carlin, Stern, & Rubin, 2004), but as the expectation can be found in closed form, such a reparametrization can easily be performed.

with the likelihood written as,

$$
\begin{aligned}
p(\mathbf{x}|\bar{x}, k) &= \prod_{i}^{M} \frac{\Gamma(k\bar{x} + x_i)}{x_i!\Gamma(k\bar{x})} \left(\frac{k}{k+1}\right)^{k\bar{x}} \left(\frac{1}{k+1}\right)^{x_i} \\
&= \sum_{i}^{M} \log\Gamma(k\bar{x} + x_i) - \log(x_i!) - log\Gamma(k\bar{x}) + k\bar{x}[\log k - \log(k+1)] - x_i \log(k+1).
\end{aligned}
\tag{5.4}
$$

We can maximize the log-likelihood in equation 5.4 using L-BFGS (D. C. Liu & Nocedal, 1989) a gradient-based search method. The gradient of the log-likelihood is,

$$
\begin{aligned}
dl/dk &= \sum_{i}^{M} \bar{x}\Psi(k\bar{x} + x_i) - \bar{x}\Psi(k\bar{x}) + (\bar{x}\log k + k\bar{x}\frac{1}{k}) - (\bar{x}\log(k+1) + k\bar{x}\frac{1}{k+1}) - x_i\frac{1}{k+1} \\
&= \sum_{i}^{M} \bar{x}[\Psi(k\bar{x} + x_i) - \Psi(k\bar{x}) + \log k - \log(k+1)] + \frac{\bar{x}}{k+1} - \frac{x_i}{k+1} \\
&= M\bar{x}[-\Psi(k\bar{x}) + \log k - \log(k+1)] + \bar{x}\sum_{i}^{M} \Psi(k\bar{x} + x_i).
\end{aligned}
$$

Using this gradient, it is possible to find the maximum likelihood parameters $\langle \bar{x}, k \rangle$ for any given set of segmentations. To jointly perform segmentation and parameter estimation, we iteratively segment and update the parameter estimates until convergence. This is equivalent to hard expectation-maximization (Bishop, 2006).

The other parameters are the symmetric Dirichlet priors on the language models. In the following experiments, these parameters are set using cross-validation. Sampling or gradient-based techniques may also be used to estimate these parameters, but this is left for future work.

Relation to other segmentation models Other cohesion-based techniques have typically focused on hand-crafted similarity metrics between sentences, such as cosine similarity (Galley et al., 2003; Malioutov & Barzilay, 2006). In contrast, the model described here is probabilistically motivated, maximizing the joint probability of the

segmentation with the observed words and gestures. Our objective criterion is similar in form to that of Utiyama and Isahara (2001); however, in contrast to this prior work, our criterion is justified by a Bayesian approach. Also, while the smoothing in our approach arises naturally from the symmetric Dirichlet prior, Utiyama and Isahara apply Laplace's rule and add pseudo-counts of one in all cases. Such an approach would be incapable of flexibly balancing the contributions of each modality.

5.3.3 Evaluation Setup

Dataset The dataset for the segmentation experiments is composed of fifteen audio-video recordings drawn from the corpus described in Chapter 3. As before, the videos are limited to three minutes in duration, and speakers mainly describe the behavior of mechanical devices – though the dataset for this section also includes four videos in which the speakers narrate the plot of a short "Tom and Jerry" cartoon. In this portion of the dataset, speakers were not permitted to use any visual aids; thus, there is no overlap between this dataset and the videos used in the previous chapter. Corpus statistics are found in Table B.2 (page 126).

Annotations and Data Processing All speech was transcribed by hand by the author, and time stamps were obtained using the SPHINX-II speech recognition system for forced alignment (Huang et al., 1993). Sentence boundaries are annotated according to the NIST (2003) specification, and additional sentence boundaries are automatically inserted at all turn boundaries. A stoplist of commonly-occurring terms unlikely to impact segmentation are automatically removed.

For automatic speech recognition (ASR), the default Microsoft speech recognizer was applied to each sentence and the top-ranked recognition result was reported. As is sometimes the case in real-world applications, no speaker-specific training data is available, so the recognition quality is very poor – the word error rate is 77%. An example of some output from the recognizer is shown in Figure 5-4.

Segmentation annotations were performed by the author, with the goal of selecting segment boundaries that divide the dialogue into coherent topics. Segmentation

	Reference Transcript	ASR Transcript
1	So this one's going to be kind of rough	So what can they can arise
2	Um so you got a it's another diagram	Ah
3	Like the ones we've been seeing	Like has been seeing
4	There's a kind of bucket thing	Has it that a bucket thing
5	Um there's a large block and three smaller blocks sitting on top of it	A and that there is a large block of three smaller blocks sitting on top of the
6	They're just free form	They just they form
7	And there's a a spring	And there's a day in the spring

Figure 5-4: A partial example of output from the Microsoft Speech Recognizer, transcribed from a video used in this experiment

boundaries are required to coincide with sentence or turn boundaries. A second annotator – a graduate student who is not an author on any paper connected with this research – provided an additional set of segment annotations on six documents. On this subset of documents, the P_k between annotators was .306, and the WindowDiff was .325 (these metrics are explained in the next subsection). This is similar to the interrater agreement reported by Malioutov and Barzilay (2006) – on a dataset of physics lectures, they found agreement ranging from .22 to .42 using the P_k metric.

Over the fifteen dialogues, a total of 7458 words were transcribed (497 per dialogue), spread over 1440 sentences or interrupted turns (96 per dialogue). There were a total of 102 segments (6.8 per dialogue), from a minimum of four to a maximum of ten. This rate of fourteen sentences or interrupted turns per segment indicates relatively fine-grained segmentation. In the physics lecture corpus used by Malioutov and Barzilay (2006), there are roughly 100 sentences per segment. On the ICSI corpus of meeting transcripts, Galley et al. (2003) report 7.5 segments per meeting, with 770 "potential boundaries," suggesting a similar rate of roughly 100 sentences or interrupted turns per segment.

The size of this multimodal dataset is orders of magnitude smaller than many other segmentation corpora. For example, the Broadcast News corpus used by Beeferman et al. (1999) and others contains two million words. The entire ICSI meeting corpus

contains roughly 600,000 words, although only one third of this dataset was annotated for segmentation (Galley et al., 2003). The physics lecture corpus of Malioutov and Barzilay (2006) contains 232,000 words. The task considered in this section is thus more difficult than much of the previous discourse segmentation work on two dimensions: there is less training data, and a finer-grained segmentation is required.

Metrics All experiments are evaluated in terms of the commonly-used P_k (Beeferman et al., 1999) and WindowDiff (WD) (Pevzner & Hearst, 2002) scores. These metrics are penalties, so lower values indicate better segmentations.

The P_k metric expresses the probability that any randomly chosen pair of sentences is incorrectly segmented, if they are k sentences apart (Beeferman et al., 1999). This is implemented by sliding a window of k sentences across the text and considering whether the sentences at the ends of the window are in the same segment. The P_k measure is the frequency with which that the reference and hypothesized segmentations disagree. Following tradition, k is set to half of the mean segment length.

The WindowDiff metric is a variation of P_k (Pevzner & Hearst, 2002). The P_k metric does not report an error if both the reference and hypothesized segmentations agree that the sentences at the endpoints of the sliding window are in different segments – even if the number of intervening segments between the endpoints is different. WD corrects this by again sliding a window of size k, but applying a penalty whenever the number of segments within the window differs for the reference and hypothesized segmentations.

Baselines Two naïve baselines are considered. Given that the annotator has divided the dialogue into K segments, the random baseline arbitrary chooses K random segmentation points. The results of this baseline are averaged over 1000 iterations. The equal-width baseline places boundaries such that all segments contain an equal number of sentences. Both the experimental systems and these naïve baselines were given the correct number of segments, and also were provided with manually anno-

Method	P_k	WD
1. Gesture only	.486	.502
2. ASR only	.462	.476
3. ASR + Gesture	**.388**	**.401**
4. Transcript only	.382	.397
5. Transcript + Gesture	**.332**	**.349**
6. random	.473	.526
7. equal-width	.508	.515

Table 5.1: For each method, the score of the best performing configuration is shown. P_k and WD are penalties, so lower values indicate better performance.

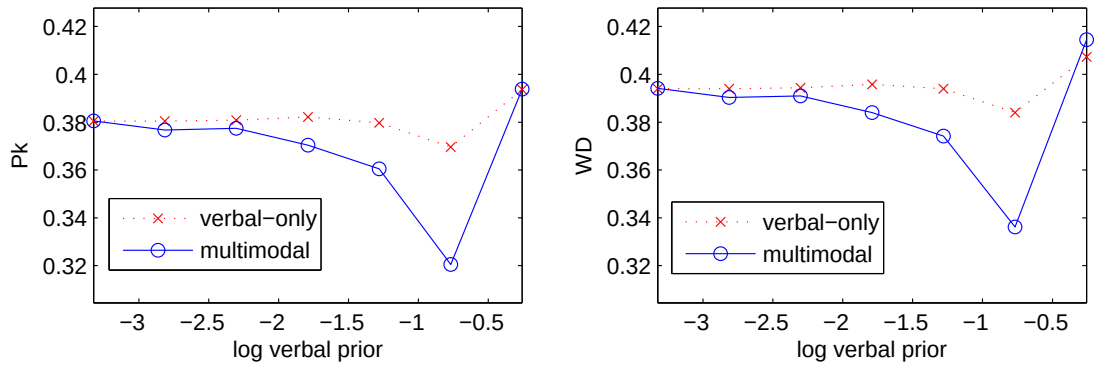


Figure 5-5: The multimodal and verbal-only performance using the reference transcript. The x-axis shows the logarithm of the verbal prior; the gestural prior is held fixed at the optimal value.

tated sentence boundaries – their task is to select the k sentence boundaries that most accurately segment the text.

5.3.4 Results

Table 5.1 shows the segmentation performance for a range of feature sets, as well as the two baselines. Given only gesture features, the segmentation results are poor (line 1), barely outperforming the baselines (lines 6 and 7). However, gesture proves highly effective as a supplementary modality. The combination of gesture with automatic speech recognition (ASR) transcripts (line 3) yields an absolute 7.4% improvement over ASR transcripts alone (line 4). Paired t-tests show that this result is statistically

significant ($t(14) = 2.71, p < .01$ for both P_k and WindowDiff).

As expected, segmentation quality is much higher when manual speech transcripts are available, compared to automatically recognized transcripts: without gesture features, both P_k and WindowDiff are roughly eight points better for manual transcripts than for ASR. However, even in this case, gesture features yield a substantial improvement, further reducing P_k and WD by roughly 5%. This result is statistically significant for both P_k ($t(14) = 2.00, p < .05$) and WindowDiff ($t(14) = 1.94, p < .05$). This suggests that gesture is not merely compensating for inadequate ASR transcripts, but adding new information not present in the lexical features.

Interactions of verbal and gesture features We now consider the relative contribution of the verbal and gestural features. In a discriminative setting, the contribution of each modality would be explicitly weighted. In a Bayesian generative model, the same effect is achieved through the Dirichlet priors, which act to smooth the verbal and gestural multinomials (see equation 5.3, page 91). For example, when the gesture prior is high and verbal prior is low, the gesture counts are smoothed and the verbal counts play a greater role in segmentation. When both priors are very high, the model will simply try to find equally-sized segments, satisfying the distribution over durations.

The effects of these parameters can be seen in Figure 5-5. The gesture model prior is held constant at its ideal value, and the segmentation performance is plotted against the logarithm of the verbal prior. Low values of the verbal prior cause it to dominate the segmentation; this can be seen at the left of both graphs, where the performance of the multimodal and verbal-only systems are nearly identical. High values of the verbal prior cause it to be over-smoothed, and performance thus approaches that of the gesture-only segmenter.

Comparison to other models Many models of cohesion-based topic segmentation have been proposed, though never for the multimodal case. While the focus of this thesis is not on topic segmentation algorithms, it is important to show that the technique applied here is competitive with the state of the art. Because the gestural

Method	P_k	WD
1. Gesture only	.423	.439
2. ASR only	.411	.565
3. ASR + Gesture	.399	.421
4. Transcript only	.390	.538
5. Transcript + Gesture	.399	.411

Table 5.2: Segmentation performance using TextSeg, with pre-specified number of segments

features are represented in a lexicon-like form, it is possible to apply previously implemented segmentation techniques without modification: in the multimodal conditions, the "sentence" is composed of both the spoken words and the gestural codewords that occur during the sentence duration.

Two alternative segmenters are considered. TextSeg (Utiyama & Isahara, 2001) uses a probabilistic approach that is somewhat similar to my Bayesian framework, as described at the end of Section 5.3.2. MinCutSeg (Malioutov & Barzilay, 2006) uses a somewhat different cohesion metric, but is specifically designed to segment speech transcripts. Another factor in the selection of these systems for comparison is that executables are publicly available online.[7]

Table 5.2 shows the performance of the TextSeg segmenter, using an evaluation setup in which the number of segments was specified by the annotator; an equivalent setup was used to generate the results for my technique shown in Table 5.1. The comparison is mixed: TextSeg is worse in all evaluations that include speech transcripts, but better in the **Gesture only** condition (line 1 in Tables 5.1 and 5.2). In the **ASR + Gesture condition** (line 3), TextSeg is only slightly worse, but it is several points worse in the **Transcript + Gesture** condition (line 5). In the unimodal verbal conditions, the TextSeg system scores very poorly on the WindowDiff metric – worse than the naïve baselines.

[7]For the Utiyama and Isahara segmenter, see `http://www2.nict.go.jp/x/x161/members/mutiyama/software.html#textseg`. For MinCutSeg, see `http://people.csail.mit.edu/igorm/acl06code.html`. Additional copies of each package, in the version used for this evaluation, can be found at `http://people.csail.mit.edu/jacobe/thesis/segmentation-baselines.html`

Method	P_k	WD
1. Gesture only	.547	2.057
2. ASR only	.398	.406
3. ASR + Gesture	.436	.489
4. Transcript only	.390	.397
5. Transcript + Gesture	.441	.561

Table 5.3: Segmentation performance using TextSeg, with automatically determined number of segments

From inspection, TextSeg generates many segments that include only a single sentence. It is possible that this behavior arises because TextSeg is not designed for conversational speech, which may include very short sentences when the conversants interrupt each other (see Appendix A.2, page 122). TextSeg is capable of determining the number of segments automatically – this may improve performance, as choosing a smaller-than-optimal number of segments may yield better performance than including many single-sentence segments. The results of this evaluation are shown in Table 5.3. While the transcript-only conditions (line 2) are much improved, the gesture-only and multimodal performance are substantially worse. From inspection, TextSeg does indeed generate fewer segments for the **ASR only** and **Transcript only** conditions, improving performance. However, on the multimodal and gesture-only conditions, it generates a much finer segmentation than desired, yielding very poor performance.

Overall, TextSeg segments the gestural and multimodal data fairly well when the number of segments is pre-specified and segments the lexical data well when the number of segments is determined automatically. It is possible that there may be some configuration of parameters that performs well on all data, but that is beyond the scope of this evaluation. TextSeg treats lexical and gestural features identically, making it incapable of modeling the different levels of noise in the two modalities. This suggests that the difficulties in applying TextSeg to multimodal data may be fundamental, and not correctable through alternative parameter settings.

MinCutSeg is another text segmentation algorithm, designed to handle the noise

Method	P_k	WD
1. Gesture only	.450	.488
2. ASR only	.433	.485
3. ASR + Gesture	.460	.502
4. Transcript only	.463	.504
5. Transcript + Gesture	.459	.500

Table 5.4: Segmentation performance using MinCutSeg, with pre-specified number of segments

inherent in automatic speech transcripts (Malioutov & Barzilay, 2006). MinCutSeg uses dynamic programming to search for the maximally cohesive segmentation, as in TextSeg and Section 5.3.2. However, MinCutSeg uses a smoothed cosine similarity metric to measure cohesion, rather than a probabilistic or Bayesian approach.

The results of applying MinCutSeg to the multimodal dataset are shown in Table 5.4 – in general they are poor, particularly in terms of the stricter WindowDiff metric. This may be because MinCutSeg is not designed for the fine-grained segmentation demanded by the multimodal dataset – Malioutov and Barzilay (2006) report an average segment length of 100 sentences on their physics lecture dataset, versus fourteen in the multimodal dataset. Malioutov and Barzilay (2006) also present results for a more finely segmented corpus of artificial intelligence lectures (the average segment includes 673 words, versus 1176 in the physics lectures; the number of sentences per segment is not reported). On this dataset, both MinCutSeg and TextSeg achieve a P_k of .37 and .38, and WindowDiff of .42. For TextSeg, these results are comparable to the transcript-only segmentation on the multimodal dataset (Table 5.3, lines 2 and 4); however, the performance of MinCutSeg is much worse on my dataset.

Summary The experiments described in this section show a novel relationship between gestural cohesion and discourse structure. Adding gestural cohesion substantially improves segmentation performance of text-only segmentation systems. This suggests that gestures provide unique information not present in the lexical features alone, even when perfect transcripts are available. These performance gains are made

possible by a novel Bayesian segmentation architecture, which outperforms alternative models for multimodal data.

5.4 Detecting Speaker-General Gestural Forms

Thus far, the research in this thesis has emphasized patterns of gestural forms within each dialogue. Crucially, this approach makes no assumption that multiple speakers will use similar gestures when describing the same semantic idea. Indeed, the proposed methods can succeed even if gestures are completely idiosyncratic, as long as each speaker exhibits some degree of self-consistency.

Nevertheless, the gestural codeword representation permits a novel cross-speaker analysis of the relationship between semantics and gestural form. By examining the distribution of codewords across speakers and dialogues with varying semantic topics[8], it is possible to quantify the extent to which speaker and topic shape the gestures that the speaker produces. This investigation is made possible by the visual processing pipeline described in Section 5.2, now applied to an entire dataset of videos rather than a single dialogue. An example of three interest points that are clustered together across speakers is shown in Figure 5-6.

The implications of such an investigation are both practical and theoretical. If each speaker employs a distinct, idiosyncratic set of gestural patterns for a given topic, then any attempt to process the semantics of gesture in a speaker-general fashion is likely to fail. On the theoretical side, this research is germane to the question of how human viewers extract content from co-speech gesture. Prior empirical research suggests that viewers are sensitive to the relationship between gesture and semantics (Kelly et al., 1999). But it is unknown whether viewers are interpreting gestures according to some speaker-general system, or if they dynamically build a speaker-specific model of gesture over the course of a conversation.

[8]The previous section described *segment-level* topics that describe a portion of a dialogue; we are now concerned with *document-level* topics that describe an entire dialogue.

Figure 5-6: The three rows show examples of interest points that were clustered together; all three include upward motion against a dark background. The center panel of each row shows the time when the interest point is detected; the left and right panels are 5 frames before and after, respectively.

The relationship between gestural form and meaning is an open question in psychology and cognitive science. Indeed, even the extent to which co-speech gesture affects listeners' comprehension is a subject of debate (see Section 2.1.1). Many researchers have focused on a micro-scale study of individual gestures and their relationship to discourse structure and semantics (e.g., Quek et al., 2000). An important complementary approach would be to investigate this phenomenon across a broad range of speakers and discourse topics, which is possible only with automated methods that can easily be applied to a large dataset.

This section describes a set of experiments using the codebook representation described in Section 5.2. Section 5.4.1 describes a hierarchical Bayesian model that learns a lexicon of gestural codewords, while jointly learning to associate codewords with specific speakers and topics. In Section 5.4.3, I describe a set of experiments that use this model to quantify the relative contribution of speaker and gesture. Section 5.4.4 presents results showing that discourse topic exerts a consistent influence on gestural form, even across speakers.

5.4.1 An Author-Topic Model of Gestural Forms

A hierarchical Bayesian model (Gelman et al., 2004) is employed to assess the relative contributions of speaker and topic to the form of the gesture. Hierarchical Bayesian models permit joint inference over sets of related random variables. In this case, the variables include the cluster memberships of each interest point, and an assignment of each interest point as generated by either the speaker or topic. This model thus allows us to estimate the proportion of interest points generated in a topic-specific, speaker-general fashion.

Inference begins with the following observed data: the speaker and topic for each dialog, and a low-dimensional description of each spatio-temporal interest point. This is the representation obtained after the dimensionality reduction step in Figure 5-2 (page 83). Each interest point is assumed to be generated from a mixture model. Interest points that are generated by the same mixture component should be visually similar. These components serve as cluster centers and are another instantiation of

gestural codewords. However, in this model, clustering is performed jointly in the hierarchical Bayesian model; this differs from the clustering in Section 5.2.3, which used the k-means algorithm as a preprocessing step. The sufficient statistics of the mixture components are shared across all dialogues, but the component weights vary by both topic and speaker. In this way, the model learns a global lexicon of visual forms, while jointly learning the distribution of visual forms with respect to speaker and topic.

The distribution over components for each speaker and topic is represented by a multinomial. A hidden auxiliary variable decides whether each codeword is drawn from the speaker-specific or topic-specific distributions. The parameter governing this hidden variable indicates the model's assessment of the relative importance of speaker and topic for gestural form.

The plate diagram for the model is shown in Figure 5-7. Each of the D dialogues is characterized by N_d visual features, which are written $\mathbf{x}_{d,i}$. Each visual feature vector $\mathbf{x}_{d,i}$ is generated from a multivariate Gaussian, $\mathbf{x}_{d,i} \sim \mathcal{N}(\mu_{z_{d,i}}, \sigma_{z_{d,i}})$, where $z_{d,i}$ indicates the codeword and σ is a diagonal covariance matrix. This induces a standard Bayesian mixture model over gesture features (Bishop, 2006). Each $z_{d,i}$ is drawn from either a speaker- or topic-specific multinomial, depending on the auxiliary variable $c_{d,i}$. If $c_{d,i} = 0$, then $z_{d,i} \sim \phi_{s_d}$, where s_d is the identity of the speaker for document d. If $c_{d,i} = 1$, then $z_{d,i} \sim \theta_{t_d}$, where t_d is the topic of document d. The distribution of c is governed by a binomial distribution with parameter λ.

Weakly informative conjugate priors are employed for all model parameters (Gelman et al., 2004). Specifically, the parameters μ and σ are drawn from a Normal-Inverse-Wishart distribution centered at the mean and variance of the observed data (Bishop, 2006). The multinomials ϕ and θ are drawn from symmetric Dirichlet priors, with parameter $\phi_0 = \theta_0 = .1$. The binomial parameter λ is drawn from a weakly informative beta prior with parameters $(.1, .1)$. As shown below, the use of conjugate priors ensures that standard closed-form posteriors can easily be found.

Our goal is to learn the relative importance of speaker versus topic, captured in the posterior distribution of the parameter λ, given observed data $\mathbf{x}, \mathbf{s}, \mathbf{d}$. Gibbs sam-

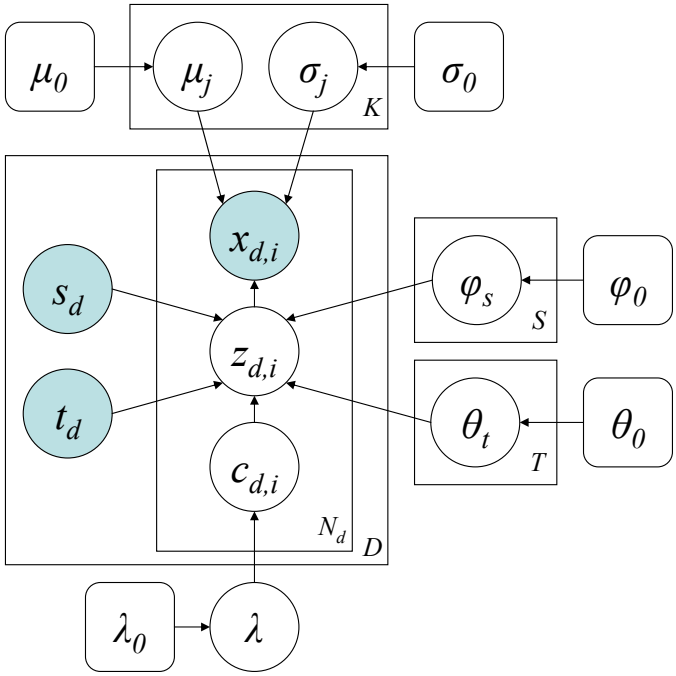


Figure 5-7: A plate diagram showing the dependencies in our model. Filled circles indicate observed variables, empty circles indicate hidden variables, and rounded rectangles indicate priors.

pling is a widely-used and easily-implemented technique for inference in hierarchical Bayesian models (Gelman et al., 2004); it involves repeatedly sampling over the posterior for each hidden variable with respect to the rest of the model configuration. After initializing the parameters randomly, Gibbs sampling is guaranteed in the limit to converge to the true distribution over the hidden variables, $p(\mathbf{z}, \mathbf{c}, \mu, \sigma, \lambda, \phi, \theta | \mathbf{x}, \mathbf{s}, \mathbf{t})$. The resulting sample set can be used to construct Bayesian confidence intervals for λ.

5.4.2 Sampling Distributions

Gibbs sampling requires posterior sampling distributions for all of the hidden variables. Rao-Blackwellization (Bishop, 2006) is used to reduce sampling variance by integrating out the parameters $\theta, \phi, \mu, \sigma$ and λ. This is possible through the use of conjugate priors. Thus it is necessary to sample only the hidden variables z and c.

The probability distribution of $z_{d,i}$ given all the variables in the model is written $p(z_{d,i}|\ldots)$; $z_{-(d,i)}$ denotes all z except $z_{d,i}$, and will be used later. $N(z_{d,i}, t_d, c_{d,i})$ denotes the count of times the codeword $z_{d,i}$ was drawn from the topic-specific distribution for topic t_d. This is computed as $\sum_{d' \leq D} \delta(t'_d, t_d) \sum_{i' \leq N_{d'}} \delta(z_{d',i'}, z_{d,i}) \delta(c_{d',i'}, c_{d,i})$, where the delta function takes the value one if the arguments are equal, and zero otherwise.

$$p(z_{d,i} = j|\ldots) \propto p(\mathbf{x}_{d,i}|\mu_j, \sigma_j) p(z_{d,i} = j|c_{d,i}, \phi_{s_d}, \theta_{t_d}) \tag{5.5}$$

$$p(z_{d,i} = j|c_{d,i}, \phi_{s_d}, \theta_{t_d}) = \begin{cases} \phi_{s_d}[j] & \text{if } c_{d,i} = 0 \\ \theta_{t_d}[j] & \text{if } c_{d,i} = 1, \end{cases}$$

where ϕ_{s_d} is the multinomial distribution indexed by the speaker s_d, and $\phi_{s_d}[j]$ is the entry for $z_{d,i} = j$ in that distribution. A student-T distribution is obtained by integrating out the parameters μ and σ from the first part of equation 5.5. This may be approximated by a moment-matched Gaussian (Gelman et al., 2004). Integrating out the parameters ϕ and θ,

$$\begin{aligned}
&p(z_{d,i} = j|c_{d,i}, z_{-(d,i)}, s_d, t_d, \phi_0, \theta_0) \propto \\
&\quad \int d\phi d\theta p(z_{d,i} = j|c_{d,i}, \phi_{s_d}, \theta_{t_d}) p(\phi_{s_d}|z_{-(d,i)}, \phi_0) p(\theta_{t_d}|z_{-(d,i)}, \theta_0) \\
&= \int d\phi d\theta (\phi_{s_d}[j]\delta(c_{d,i}, 0) + \theta_{t_d}[j]\delta(c_{d,i}, 1)) p(\phi_{s_d}|z_{-(d,i)}, \phi_0) p(\theta_{t_d}|z_{-(d,i)}, \theta_0) \\
&= \delta(c_{d,i}, 0) \int \phi_{s_d}[j] p(\phi_{s_d}|z_{-(d,i)}, \phi_0) d\phi + \delta(c_{d,i}, 1) \int \theta_{t_d}[j] p(\theta_{t_d}|z_{-(d,i)}, \theta_0) d\theta \qquad (5.6) \\
&= \delta(c_{d,i}, 0) \frac{N(j, s_d, c_{d,i} = 0) + \phi_0}{N(., s_d, c_{d,i} = 0) + K\phi_0} + \delta(c_{d,i}, 1) \frac{N(j, t_d, c_{d,i} = 1) + \theta_0}{N(., t_d, c_{d,i} = 1) + K\theta_0}. \qquad (5.7)
\end{aligned}$$

The derivation of line 5.7 from line 5.6 follows from standard Dirichlet-Multinomial conjugacy (Gelman et al., 2004), enabling the computation of the posterior probability of $z_{d,i}$ in a ratio of counts. Sampling c is more straightforward:

$$p(c_{d,i}|c_{-(d,i)}, z_{d,i}, \lambda_0, \ldots) \propto p(z_{d,i}|c_{d,i}, z_{-(d,i)}, t_d, s_d, \phi_0, \theta_0) \int p(c_{d,i}|\lambda) p(\lambda|c_{-(d,i)}\lambda_0) d\lambda.$$

The first part of the product is defined in equation 5.7. The integral can be handled analogously, as Beta-Binomial conjugacy is a special case of Dirichlet-Multinomial conjugacy,

$$\int p(c_{d,i}|\lambda) p(\lambda|c_{-(d,i)}\lambda_0) d\lambda = \frac{N(c_{d,i}) + \lambda_0}{N + 2\lambda_0}. \tag{5.8}$$

Both $z_{d,i}$ and $c_{d,i}$ are categorical variables, so it is possible to sample them jointly by considering all possible pairs of values. These parameters are tightly coupled, and sampling them together is thus likely to speed convergence (Gelman et al., 2004). The joint sampling distribution is given by

$$p(z_{d,i}, c_{d,i}| \ldots) = p(z_{d,i}|c_{d,i}, z_{-(d,i)}, s_d, t_d, \phi_0, \theta_0) p(c_{d,i}|c_{-(d,i)}, \lambda_0),$$

where the first part of the product is defined in equation 5.7 and the second part is defined in equation 5.8.

5.4.3 Evaluation Setup

Dataset

The dataset for this experiment is composed of 33 short videos from the corpus described in Chapter 3. As before, the topics consist of mechanical devices and a cartoon narrative, and dialogues were limited to three minutes in duration. As in the topic segmentation experiments in Section 5.3.3, speakers were not permitted to use visual aids. Many of the videos that were initially recorded could not be used in the

topic segmentation study because the audio was corrupted. Since audio plays no part in the experiments in this section, these videos can now be used. The parameters of this dataset are described in Table B.3 (page 127).

Implementation Details

The model from Section 5.4.1 includes four tunable parameters: the number of iterations of Gibbs sampling to run, the number of interest points to extract, the number of mixture components K, and the dimensionality of the gesture features after PCA. Gibbs sampling is performed along five parallel runs for 15000 iterations each. The first 5000 iterations are considered a "burn-in" period, and confidence intervals are estimated from the remaining 10000. The number of interest points extracted is set to 1/10 the number of frames in each video; on average, 390 interest points were extracted per video. The number of components was set to 100, and the dimensionality of the gesture features after PCA was set to 5. These parameter settings were made before the experiments were run and were not tuned with respect to the results, though the settings did reflect a desire for tractability in terms of speed and memory. In general, these settings impact the gesture clustering and do not directly affect the assignment of codewords to the speaker or topic; however, alternative settings may be considered in future work.

Experiments

The experiments analyze the influence of topic and speaker on gestural form from both Bayesian and frequentist perspectives.

Bayesian Analysis The first experiment estimates the number of gestural features that are generated in a topic-specific manner, using the model described in Section 5.4.1. This proportion should be represented by the parameter λ, the prior on the likelihood that each gestural feature is generated from the topic-specific model.

However, even if there were no topic-specific, speaker-general patterns, it is possible that the topic-specific model θ might somehow be used to overfit the data.

To isolate the extent to the which the topic-specific model is used for overfitting, the topic indicators were randomly shuffled in five baseline conditions. If the topic-specific model is used more frequently with the true topic indicators than in the randomized conditions, then this would suggest that the effect is due to a real correlation between topic and gestural form, and not simply to overfitting.

Frequentist Analysis Just as in text, lexical distributions are indicative of discourse topic (Hearst, 1994); thus, it may be helpful to examine the distribution of gestural codewords across topics. The Bayesian model builds a lexicon of gestures by clustering gesture features; this is done jointly with the assignment of gesture features to speakers and topics. Such a joint model is advantageous because it is possible to integrate over uncertainty in the clustering, rather than propagating the effects of a bad clustering decision to the other stages. However, it is illustrative to consider the maximum *a posteriori* (MAP) clustering induced by the model (Bishop, 2006) and investigate how the distribution of cluster tokens varies by topic.

To this end, the second experiment performs chi-squared analysis of the distribution of cluster membership, with respect to both topic and speaker. Chi-squared analysis allows us to test the null hypothesis that gestural forms are generated in a way that is independent of the discourse topic.

5.4.4 Results

The results of the Bayesian analysis are shown in Table 5.5 and Figure 5-8. With the correct topic labels, 12% of gestures are classified as topic-specific. When the topic labels are randomized, this average drops to less than 3%. Thus, the model uses the topic-specific codeword distributions mainly when the topic labels are actually informative, supporting the hypothesis of a connection between discourse topic and gestural form that transcends individual speakers.

Bayesian confidence intervals constructed from the samples show that these differences are robust. As indicated in Table 5.5, the confidence interval for the randomized conditions is much larger. This is expected, as each randomization of topic labels

condition	mean	upper	lower
true topic labels	.120	.136	.103
random topic labels	.0279	.0957	0

Table 5.5: Proportion of gestures assigned to the topic-specific model, with 95% confidence intervals

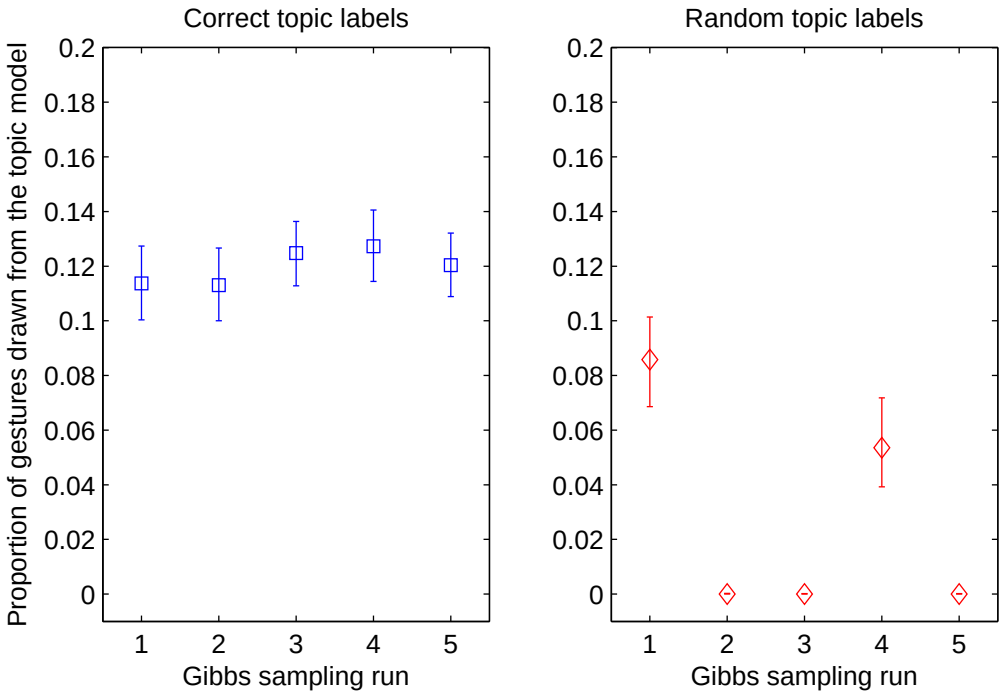


Figure 5-8: Proportion of gestures assigned to the topic model, per run

varies from the true labels to a different extent. Figure 5-8 illustrates this situation, showing the confidence intervals from each randomized run. In the topic-randomized condition, there is substantial between-run variance; in three of the runs, the topic exerts no influence whatsoever. In contrast, in the condition with correct topic labels, the influence of the topic-specific model is consistently in the range of 12%.

Next, the influence of topic and speaker on gestural form is analyzed using the classical chi-squared test. The maximum *a posteriori* (MAP) gesture feature clustering is obtained by selecting the iteration of Gibbs sampling with the highest likelihood.[9] The chi-squared test is used to determine whether the distribution of clusters differs

[9]In sampling-based inference, MAP estimates are often found by taking a mean or mode over multiple samples. In the case of estimating a clustering, this technique suffers from non-identifiability.

significantly according to topic and speaker (De Groot & Schervish, 2001).

Strong effects were found for both topic and speaker. For topics, $p < .01, \chi^2 = 1.12 * 10^4, \text{dof} = 439$.[10] For speakers, $p < .01, \chi^2 = 5.94 * 10^4, \text{dof} = 1319$. While the chi-squared values are not directly comparable – due to the different degrees of freedom – this experiment indicates an important effect from both the topic and speaker.

5.5 Discussion

This chapter demonstrates a novel relationship between gesture and high-level discourse structure. Within a dialogue, topically-coherent discourse segments are characterized by gestural cohesion. This internal consistency of gestural forms mirrors the well-known phenomenon of lexical cohesion. These results are obtained using a codebook representation, which clusters local video features into characteristic gestural forms. This same representation shows a connection between gesture and document-level topics, which generalizes across multiple speakers. In addition, this chapter demonstrates that even with a relatively lightweight visual analysis it is still possible to capture semantically-relevant aspects of gesture.

Cohesion is only one possibility for how gesture might predict topic segmentation. An alternative is gesture as "visual punctuation" – explicit discourse cues that predict segment boundaries. This is analogous to research on prosodic signatures of topic boundaries (e.g., Hirschberg & Nakatani, 1998). By design, the model presented in this chapter is incapable of exploiting such phenomena, as this thesis focuses on gestures that communicate narrative content. Thus, the performance gains obtained here cannot be explained by such punctuation-like phenomena; they are due to the consistent gestural themes that characterize coherent segments. However, the use of visual punctuation should be explored in the future, as the combination of

For example, two data points may appear in many different clusters, though usually together; even so, their modal cluster memberships may differ, causing them to be separated in the MAP estimate.

[10] Clusters with fewer than five examples were excluded, as the chi-squared text is not accurate for small bin values. Thus, the number of degrees of freedom is less than the expected $KT - 1$.

visual punctuations and cohesion may further improve segmentation performance. The interaction of the gesture and prosodic modalities suggests additional avenues of potentially fruitful research.

The results in Section 5.4 support the view that topic-specific gestural forms are shared across speakers. The frequency of such shared gestural forms is likely influenced by both the population of speakers and the topics of discussion. The speakers in the dataset are all American residents and fluent speakers of English. The extent to which gestural forms are shared across cultures is a key area for future research. Another important question is whether gestural forms are shared when the discourse topics are less concrete. Do multiple speakers use similar gestures when talking about, say, their circle of friends, or their ideas on politics?

While the dataset was designed to encourage speaker-general gestures, it is also true that any automatic vision-based technique for gesture analysis is likely to *over-state* speaker-specific factors. This is because it is difficult – if not impossible – to abstract away all features of the speaker's appearance. The visual features used here are brightness gradients and the location of movement. Brightness gradients are influenced by the speaker's skin tone and clothing; location of movement is influenced by anatomical factors such as the speaker's height. Thus, the likelihood of such visual features being clustered in a speaker-dependent manner is artificially inflated. With the development of robust vision techniques that describe gesture's visual form on a more abstract level, future work may show that topic exerts a greater influence than reported here.

Interest point features have not previously been used to describe co-speech gesture. This chapter demonstrates that they can be an effective technique to identify semantically-relevant gestural patterns. This result is encouraging, because such interest points are more robust and easier to compute than the tracking-based techniques used both in Chapter 4 and in other related work (e.g., Quek, McNeill, Bryll, Duncan, et al., 2002).

Still, the interest point representation can be extended in various ways. Individual interest points, as used here, are sufficient to describe a range of gestural forms, such

as handshapes and paths of motion. However, they do not account for higher-level phenomena, such as when both hands move in a synchronized or anti-synchronized fashion. Rather than assigning codewords to local visual features, it may be advantageous to consider sets of local features that frequently co-occur. Such an approach may result in a characterization of gesture that better coheres with human perception.

6

Conclusion

In face-to-face dialogue, meaning is communicated through a range of behaviors. Natural language processing research has long emphasized speech, but this thesis demonstrates that gesture plays an important role: hidden structures in discourse reveal themselves through patterns in gesture. Such gestural patterns are extracted from raw visual features, and the resulting multimodal discourse processing systems outperform state-of-the-art text-only alternatives.

The use of gestures to identify patterns in narrative content is a novel approach to multimodal language processing. This marks a substantial departure from prior computational research on prosody and gesture, which focused on building recognizers for specific pragmatic cues. However, the approach is well-supported by a tradition of psycholinguistic research on gestural *catchments* (McNeill, 1992). This thesis marks the first time that catchment theory has been used to predict discourse structure in

a computational framework.

This thesis has focused on two gestural patterns: cohesion across sets of gestures, and similarity of gestural pairs. Gestural similarity – computed from the output of an articulated upper-body tracker – was shown to predict noun phrase coreference in Chapter 4. Moreover, by identifying a subset of salient gestures among all other hand movements, a more refined model of gestural similarity was acquired, further improving the linguistic contribution. Chapter 5 introduced the idea of gestural cohesion, extending pairwise similarity to larger sets of gestures. Gestural cohesion – computed from spatiotemporal interest points – was demonstrated to improve unsupervised topic segmentation.

Throughout this thesis, models of gesture were learned without gesture-specific annotations. In Chapter 4, learning was driven by linguistic coreference annotations, which were exploited to learn both gesture similarity and gesture salience. Chapter 5 applied unsupervised learning in a joint model that incorporates both gesture and speech. While the details of the machine learning algorithms differ, both are structured models that explicitly encode the relationship between the gesture and speech modalities. This approach is motivated by the intuition that the interpretation of gesture depends crucially on the surrounding linguistic context.

6.1 Limitations

In general, the temporal segmentation of gestures in this dissertation has been completely driven by the speech. In Chapter 4, gesture features are computed over the duration of the noun phrase; in Chapter 5, interest points are computed over the duration of the sentence. In light of the well-known synchrony between gesture and speech (Condon & Ogston, 1967; Loehr, 2007), such an approach seems a reasonable way to approximate the temporal extent of gestures. However, a more precise and fine-grained model of gesture segmentation may be obtained by considering visual features of the gesture, and this may improve performance on linguistic processing.

While computer vision is not the focus of this dissertation, it is worth noting that the vision techniques used here are limited in ways that prevent their application to many of the videos currently available online. The articulated tracker described in Chapter 4 is heavily dependent on the colored gloves worn by the speakers. More robust hand tracking techniques would have to be applied to extend this application to more realistic videos. In addition, the vision techniques in both chapters rely on the fact that the speaker is the only source of motion in the video. The implementation of the articulated tracker and the use of spatiotemporal interest points would be substantially complicated by camera movement (e.g., panning and zooming) or the presence of moving entities other than the speaker.

Overall, these requirements made existing audio-video data unusable, forcing the acquisition of a new dataset. The resulting corpus is smaller than datasets used in other areas of natural language processing. Many researchers are interested in increasing the robustness and generality of computer vision. As new vision methods become more accessible to non-specialists, it is likely that such limitations can be overcome.

The topics of discussion in my dataset were deliberately restricted so as to encourage the use of direct, concrete gestures. To what extent do the gestural patterns observed and exploited in this thesis depend on having a topic of discourse that is directly physical? When the topic of discussion is more abstract, representational gestures employ physical metaphors (McNeill, 1992). It remains unknown how often such metaphoric gestures are used, and whether they can be exploited as effectively as the more straightforward iconic and deictic gestures that dominate this dataset.

Finally, the speakers in this dataset are permitted only to speak and gesture. In many real-world scenarios, speakers do a variety of other things: they pace around the room, draw on chalkboards, change slides, perform physics demonstrations, drink coffee, etc. The problem of identifying salient gestures becomes more complicated – and more crucial – when such additional activities are permitted. The application of this research to key target domains, such as classroom lectures and business meetings, depends on addressing this problem.

6.2 Future Work

Some future work has already been discussed; extensions that follow naturally from the technical content were presented at the end of the relevant chapters (Sections 4.6 and 5.5) so that the necessary details would remain fresh in the reader's mind. In this section, I consider future research that is somewhat further afield from the specific projects demonstrated in the thesis but is a continuation of the underlying ideas.

Contrastive Catchments From a theoretical perspective, much of this dissertation is motivated by David McNeill's concept of the gestural catchment. McNeill (1992) defines a catchment as a pair of gestures which, through their relationship to one another, convey some information about the discourse structure. I have focused exclusively on one type of catchment: the identity relationship. But catchments are also used to indicate *contrastive* relationships between discourse elements. McNeill describes many examples in which a speaker repeats a gesture but modulates some critical element, such as the speed or the handshape, indicating a key semantic distinction. Building an automatic system that could recognize such catchments requires a more fine-grained model of gesture, as well as a richer treatment of semantics. However, such an effort could substantially improve discourse processing, and may also help to answer fundamental linguistic questions about the frequency of such contrastive catchments and the situations in which they are used.

Richer Models of Similarity and Salience In a related vein, I have assumed that gesture similarity and salience are both separate and atomic. In fact they are likely neither: any pair of gestures will be similar in some ways and different in others, and whether the gestures are judged to be *holistically* similar depends on which features are salient. For example, if I produce a gesture with a complicated trajectory while standing in one part of the room, and then walk across the room and repeat it, do these two gestures mean the same thing? Only the absolute hand position has changed; if I am simply pacing across the room, this is probably irrelevant, but if I happen to be pointing at different regions on a large map, then the change in position may

be crucial. The similarity of two gestures can only be considered as a single atomic concept if we can simultaneously identify which *features* of the gesture are salient – that is, which features count towards a holistic assessment of similarity.

In principle, this model of gesture interpretation seems to require the ability to read the speaker's mind, so as to determine which aspects of the visual presentation are intended to be communicative. It is important to remember that human viewers appear to solve this problem effortlessly. While the relationship between gestural form and meaning is not generally conventionalized, it may be the case that the use of visual features to convey meaning does demonstrate some regularity – either through social convention or on some deeper cognitive level. One simple way in which it could be standardized is through a salience ordering of visual features – for example, movement patterns may always be more salient than handshapes. If this is the case, then viewers may not attend to handshape if the movement patterns of the gestures are different, but will attend to the handshape if the movement patterns are identical. Such a salience ordering is only one of many possibilities, and empirical research is required to understand how human viewers pick out the salient characteristics of gesture. Such research could directly inform automatic gesture processing systems, leading to dynamic models of gestural similarity that attend only to perceptually salient gestural features.

Learning Across Multiple Tasks One of the key ideas of this thesis is that models of gesture should be learned in the context of language processing, rather than in isolation. There are two main motivations for this view. First, learning about gesture in isolation would seem to require training from some general-purpose annotation of gestural form. Such annotations would be difficult to define and costly to produce. Second, as our ultimate goal and evaluation metric is performance on discourse-processing tasks, it is advantageous to train on these metrics directly, particularly since their annotation schemes are already well-defined.

But even while rejecting the idea of learning about gesture in isolation from language, we need not abandon the hope of learning models of gesture that generalize

across linguistic contexts. The projects described in this thesis leverage annotations from individual linguistic tasks, but more complex models may leverage multiple types of annotations. For example, given annotations for both noun phrase coreference and discourse segmentation, it may be possible to learn a model of gesture salience that applies to both tasks. If gesture salience is indeed a coherent concept outside the setting of noun phrase coreference, then combining linguistic annotations in this way should yield a more robust model, and increase the effective training set size.

In a sense, this approach amounts to learning task-general models of gesture, but in a bottom-up, data-driven way. In addition to the engineering advantages just mentioned, such research may be relevant from a purely linguistic standpoint. Such a model would, for example, permit an investigation of which language phenomena share coherent notions of gesture salience, and how gesture salience is expressed in visual and linguistic features. This thesis has shown that structured learning models can be used to incorporate linguistic ideas about gesture in a principled way. Future work may show that such models can also provide a new tool to study the linguistics of gesture.

A

Example Transcripts

A.1 Coreference Example

This section presents an example of the coreference annotations for a single conversation. Of the two participants, only the speech from the "presenter" is shown here – this is the participant who has seen a simulation of the relevant device, and is asked to explain it to the other participant. Annotated noun phrases are indicated using square brackets, and an index for the relevant coreference chain is indicated in parentheses. Note that first and second person pronouns are not annotated. The lines are numbered for reference in Section 4.4.

The topic of this converation was the pinball device, and the presenter was provided with the a printed version of the diagram shown in Figure C-3 (page 130).

```
1   ok so [(0) this object right here].
```

```
2   i’m going to attempt to explain what [(0) this] does to you.
3   [(1) this ball right here] is the only
4   [(1) this ball]
5   [(2) this spring]
6   [(3) this long arm]
7   and [(4) this] are [(5) the only objects that actually move].
8   no i take [(6) that] back.
9   [(7) this] rotates as well.
10   while [(8) these things] stay fixed.
11   what happens is [(1) the ball] comes down [(9) here].
12   and [(2) this spring] is active.
13   meaning that [(2) it’s] going up and down.
14   because [(4) this] will come up.
15   jostle [(3) that].
16   and then go around.
17   so [(3) it’ll] as [(4) this] raises [(3) it] up.
18   [(10) this hand] goes down.
19   and then [(10) it’ll] spring back up.
20   [(1) the ball] typically goes up [(11) here].
21   bounces off [(12) here].
22   gets caught in like [(13) a groove].
23   [(7) this] is continually moving around in [(14) a circle]
24   then [(15) this] happened three times
25   i watched [(16) a video] and [(15) it] happened [(17) three times]
26   [(1) the ball] never went through [(18) there] or [(19) over here]
27   [(1) it] always would get down back to [(20) here]
28   and then down through [(9) here]
29   sometimes [(21) this thing] would hit [(1) it] harder
30   and [(1) it] would go higher up
31   and sometimes [(1) it] would just kind of loop over
32   no no [(1) it] only came down through [(9) here]
33   i have no idea why there’s [(22) anchors] on [(23) here]
34   [(24) that] wasn’t really made clear to me
35   and yeah [(25) that’s] pretty much [(26) it]
36   [(1) it’s] essentially [(1) a bouncy ball]
37   but [(1) it] just pretty much drops like [(27) dead weight]
38   when [(1) it] hits [(28) something]
39   and that was [(26) it]
40   [(16) it] was probably like [(16) a forty five second video] at most
41   and [(29) it] happened [(17) three times] in [(16) that video]
42   so [(16) it] moves relatively quickly
43   not so much lodged as just like [(1) it] would like come down [(13) here]
44   and as [(7) this] is moving [(7) it] would
45   just kind of like dump [(1) it] into [(20) here]
46   [(7) it’s] more of something that’s in [(30) the way]
47   than actually [(31) a transfer]
48   because if [(7) this] wasn’t [(32) here]
49   [(1) it] would still fall down [(20) here] and then get in [(9) here]
50   that’s it
```

```
51   i'm not actually sure what [(0) this] does
52   [(0) it] looks like [(0) it] looks just like
53   [(0) this] on [(33) the computer screen]
54   so so um [(0) it] basically looks like
55   [(34) kind of a so so game of pinball]
56   [(35) that] was [(36) my understanding of it]
57   i'm not sure what else [(37) it's] supposed to do
58   ok we're done guys with [(37) this one]
```

A.2 Segmentation Example

This section presents an example of the topic segmentation annotations for a single conversation. Speaker `A` is tasked with explaining the behavior of a piston (Figure C-4, page 130) to speaker `B`, after watching a video demonstration. Speaker `A` is not permitted to use visual aids in this conversation.

When the participants interrupt each other, this is indicated by "`...`" after the speech. The start and end times are given in milliseconds in square brackets at the end of each line.

TOPIC: intro and wheel

```
A: ok [18103 18483]
A: this one is actually pretty difficult to explain [20173 22693]
A: i wish they gave us something [23425 24385]
A: um ok just try to imagine what i'm going to point out to you here
   [25025 30965]
A: there's a circle right here attached to a straight brick and then
   another straight brick [30975 39705]
A: but at the end of this second straight brick is like a bar like
   this [39715 44475]
A: ok so we have a circle [44933 46063]
A: these are all connected by like a screw [46443 48403]
A: so there's a a wheel a bar coming off of it another bar and then
   like a t basically at the end of it [49675 59638]
B: uhum a long bar [60021 61361]
A: yeah [61390 61770]
B: go on [61761 62061]
```

TOPIC: the container

A: this is all inside of basically just picture a big square [62420 66740]
A: ok but there's an opening in the square right here for the bars to go through [67436 71376]
A: and then that big t thing is the width of the square [72065 76045]
A: want me to repeat some of that [79130 80520]
B: uhum the t thing [80641 81591]
B: but i [81601 81931]
B: is it [81971 82291]
B: you mean the vertical thing at the end of the two bricks [82401 84721]
A: yes exactly that is going vertical [84336 87096]
A: and everything else here is going ... [87166 88286]
B: but ... [88191 88351]
A: horizontal [88296 89166]
B: to the height of the whole square [88391 90001]

TOPIC: the function

A: what this machine does is it [90208 94998]
A: it's essentially a press [95068 96428]
A: that would be the easiest way to explain it [97330 98970]
A: um [99380 100040]
A: everything the wheel and the two bars are connected by one [102800 108340]
A: shit this is hard to explain [109090 110550]
A: in the wheel there's a screw [111615 115545]
A: that's attached to the [115961 116711]
A: wait just need to look at this [116781 117851]
A: there's a screw that's attached to the bar [118170 120520]
A: as the wheel turns [122273 123613]
A: the bar goes in [124143 125683]
A: but as the wheel turns out [126863 128403]
A: this arm will go out as well [129413 131073]
A: because they're connected [131083 132073]
A: so as the wheel like [132083 132943]
A: you know as that screw was back here [133053 134413]
A: the whole t thing is kind of back here [135180 136860]
A: but as it comes around [137160 138420]
A: everything goes out [138540 139540]
B: got it [139611 140091]
A: ok [141396 141806]
A: so the press works according to how the wheel's position is [142006 145476]
A: as the screw of the wheel comes facing as far as it can towards the press [146205 151145]
A: the t let's call it is pressing against the end of the ... [151753 154533]

TOPIC: comparison to a nut cracker

B: sort ... [154526 154836]
A: square [154543 155133]
B: of like one of those old fashion nut crackers [154846 157146]
A: i was actually thinking of a wine vinyard like [157096 159336]
A: just like that [159346 160066]
A: but a nut cracker is good ... [160563 161573]
B: you ... [161583 161723]
A: as well [161583 161943]
B: know [161733 161833]
B: where the like you twist it [161843 163473]
B: and the thing just presses the nut up against the end of the [163513 166963]
A: yeah ... [167073 167313]
B: ok ... [167433 167763]
A: except [167433 167733]
A: for there's no screw in this [167743 169193]
A: it's just a wheel [169203 170193]
A: and the nut crackers have a screw [171388 172838]
A: that you twist [172848 173638]
B: right but ... [173871 174671]
A: but the same philosophy [174866 176306]
A: and if something was in this square in the end of it [176536 178326]
A: it would get [178366 178816]

B

Dataset Statistics

number	speaker	topic	duration	words	sentences
1	03	pinball	3:02	455	95
2	03	candy	2:27	428	101
3*	04	latch	1:19	104	27
4	04	pinball	2:31	283	65
5*	06	pinball	3:05	325	69
6	07	pinball	2:23	192	47
7*	07	piston	0:47	48	8
8*	09	candy	3:02	404	100
9	09	pinball	3:01	421	109
10	10	pinball	3:01	362	89
11*	10	piston	3:02	313	69
12*	11	pinball	3:03	315	71
13	13	latchbox	2:20	347	72
14	13	pinball	3:11	221	51
15	15	pinball	2:30	378	87
16	15	candy	2:43	358	77
total			41:34	4954	1137

Table B.1: Corpus statistics for the dataset used in the experiments from Chapter 4. All videos were used in the coreference evaluation; asterisks indicate videos that were used in the keyframe evaluation.

number	speaker	topic	duration	words	sents	segments
1	03	cartoon	1:34	248	32	5
2	03	latchbox	2:41	393	54	6
3	04	piston	2:16	323	46	5
4	06	latchbox	3:04	461	91	9
5	07	latchbox	1:47	260	45	6
6	09	cartoon	3:01	691	186	10
7	09	toy	3:03	703	140	10
8	09	piston	3:10	673	153	9
9	10	cartoon	3:04	579	102	8
10	10	candy	2:47	488	100	6
11	10	toy	3:04	616	127	6
12	11	candy	3:05	530	93	6
13	13	cartoon	2:51	443	95	7
14	13	candy	3:02	604	104	5
15	15	piston	3:03	446	72	4
total			41:40	7458	1440	102

Table B.2: Corpus statistics for the experiments on discourse segmentation in Section 5.3

number	speaker	topic	duration	start	end
1	01	cartoon	0:38	0:00	0:38
2	01	pez	1:23	0:05	1:29
3	02	cartoon	1:06	0:01	1:07
4	02	pez	1:28	0:01	1:30
5	03	cartoon	1:09	0:05	1:15
6	03	latchbox	2:34	0:05	2:40
7	04	cartoon	2:49	0:04	2:53
8	04	piston	2:01	0:05	2:06
9	04	toy	2:58	0:05	3:03
10	05	cartoon	0:21	0:01	0:23
11	05	piston	1:27	0:02	1:30
12	05	toy	0:45	0:01	0:46
13	06	cartoon	2:53	0:03	2:56
14	06	latchbox	2:58	0:05	3:03
15	07	cartoon	0:52	0:04	0:56
16	07	latchbox	1:34	0:05	1:40
17	08	cartoon	1:26	0:04	1:30
18	09	cartoon	2:53	0:06	3:00
19	09	piston	3:07	0:03	3:11
20	09	toy	2:58	0:05	3:03
21	10	cartoon	2:56	0:06	3:03
22	10	pez	2:36	0:03	2:40
23	10	toy	2:56	0:06	3:03
24	11	cartoon	2:57	0:07	3:05
25	11	pez	2:37	0:27	3:05
26	12	cartoon	2:08	0:05	2:13
27	12	pez	2:51	0:13	3:05
28	12	toy	2:58	0:03	3:01
29	13	cartoon	2:19	0:07	2:26
30	13	pez	2:56	0:05	3:02
31	14	cartoon	1:26	0:10	1:36
32	14	latchbox	1:05	0:08	1:14
33	15	piston	2:40	0:21	3:01
total			70:02		

Table B.3: Corpus statistics for the experiments on speaker and topic-specific gestures in Section 5.4. To avoid including motion from the speaker entering or leaving the scene, only data between the start and end times are included, as indicated in the right two columns of the table.

C

Stimuli

This appendix includes still images of the six stimuli shown to the speakers in the dataset described in Chapter 3. For the first four stimuli, speakers were shown video simulations, using the Working Model software application (`http://www.design-simulation.com/WM2D/`). For the Star Wars toy, participants were able to see and manipulate the actual physical object. The final stimulus is a short "Tom and Jerry" cartoon.

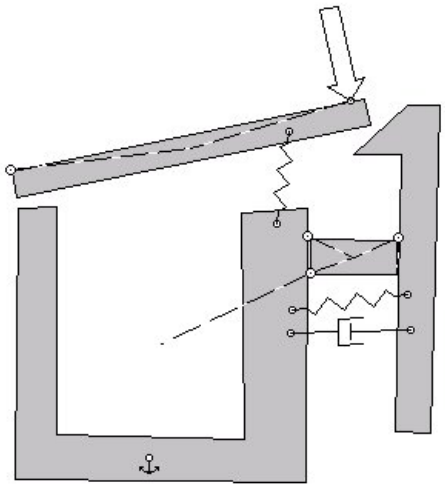


Figure C-1: Latching box

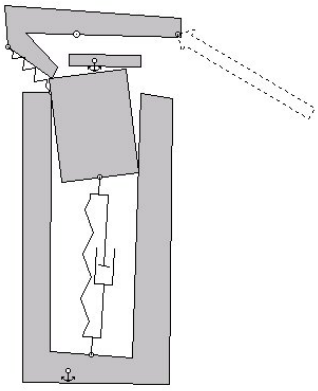


Figure C-2: Candy dispenser

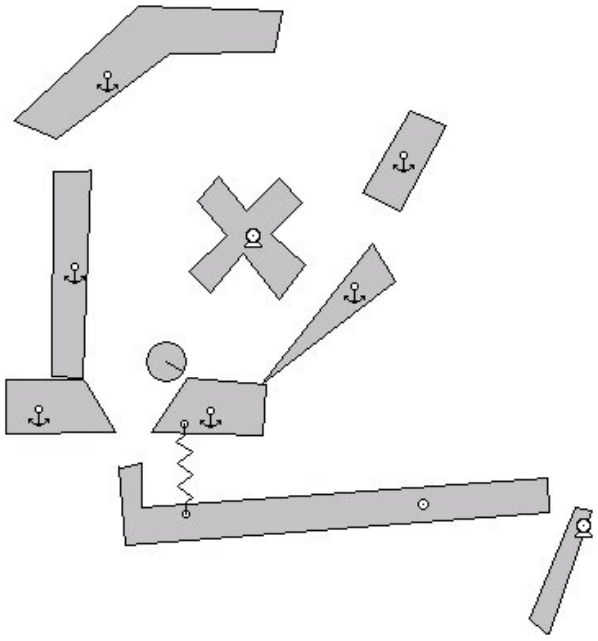


Figure C-3: Pinball machine

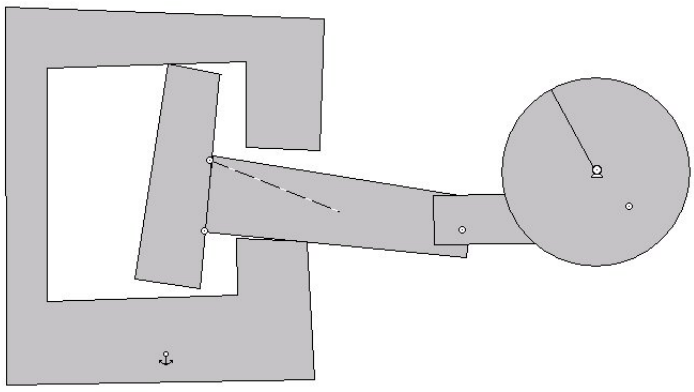


Figure C-4: Piston

Figure C-5: Star Wars toy

Figure C-6: Tom and Jerry cartoon

References

Adler, A., Eisenstein, J., Oltmans, M., Guttentag, L., & Davis, R. (2004). Building the design studio of the future. In *Proceedings of AAAI Workshop on Making Pen-Based Interaction Intelligent and Natural* (p. 1-7).

Arulampalam, S., Maskell, S., Gordon, N., & Clapp, T. (2002). A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, *50*(2), 174-188.

Baldwin, B., & Morton, T. (1998). Dynamic coreference-based summarization. In *Proceedings of EMNLP.*

Bangalore, S., & Johnston, M. (2004). Balancing data-driven and rule-based approaches in the context of a multimodal conversational system. In *Proceedings of HLT-NAACL'04* (p. 33-40).

Barzilay, R., & Lapata, M. (2005). Modeling local coherence: an entity-based approach. In *Proceedings of ACL* (p. 141-148).

Beattie, G., & Coughlan, J. (1999). An experimental investigation of the role of iconic gestures in lexical access using the tip-of-the-tongue phenomenon. *British Journal of Psychology*, *90*(1), 35-56.

Beattie, G., & Shovelton, H. (2006). When size really matters: How a single semantic feature is represented in the speech and gesture modalities. *Gesture*, *6*(1), 63-84.

Beeferman, D., Berger, A., & Lafferty, J. D. (1999). Statistical models for text segmentation. *Machine Learning*, *34*(1-3), 177-210.

Bennett, P., & Carbonell, J. (2007). Combining Probability-Based Rankers for Action-Item Detection. In *Proceedings of HLT-NAACL* (p. 324-331).

Bernardo, J. M., & Smith, A. F. M. (2000). *Bayesian theory.* Wiley.

Biber, D. (1988). *Variation across speech and language.* Cambridge University Press.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer.

Bolt, R. A. (1980). Put-that-there: Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on computer graphics and interactive techniques* (p. 262-270).

Boreczky, J., Girgensohn, A., Golovchinsky, G., & Uchihashi, S. (2000). An interactive comic book presentation for exploring video. In *Proceedings of CHI* (p. 185-192).

Bradley, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145-1159.

Brennan, S. E., Friedman, M. W., & Pollard, C. J. (1987). A centering approach to pronouns. In *Proceedings of ACL* (p. 155-162).

Bryll, R., Quek, F., & Esposito, A. (2001). Automatic hand hold detection in natural conversation. In *IEEE Workshop on Cues in Communication.*

Campana, E., Silverman, L., Tanenhaus, M., Bennetto, L., & Packard, S. (2005). Real-time integration of gesture and speech during reference resolution. In *Proceedings of the 27th Annual Meeting of the Cognitive Science Society.*

Cardie, C., & Wagstaff, K. (1999). Noun phrase coreference as clustering. In *Proceedings of the joint conference on empirical methods in natural language processing*

and very large corpora (p. 82-89).

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, *22*(2), 249-254.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., et al. (2005). The AMI meeting a corpus: a pre-announcement. In *Proceedings of the Workshop on Machine Learning for Multimodal Interaction (MLMI)* (p. 28-39).

Cassell, J. (1998). A framework for gesture generation and interpretation. In *Computer Vision in Human-Machine Interaction* (p. 191-215). Cambridge University Press.

Cassell, J., Nakano, Y. I., Bickmore, T. W., Sidner, C. L., & Rich, C. (2001). Non-verbal cues for discourse structure. In *Proceedings of ACL* (p. 106-115).

Chafe, W. (1980). *The pear stories: cognitive, cultural, and linguistic aspects of narrative production*. Ablex Pub. Corp., Norwood, NJ.

Chai, J. Y., Hong, P., & Zhou, M. X. (2004). A probabilistic approach to reference resolution in multimodal user interfaces. In *Proceedings of Intelligent User Interfaces* (p. 70-77).

Chen, L., Harper, M., & Huang, Z. (2006). Using maximum entropy (ME) model to incorporate gesture cues for sentence segmentation. In *Proceedings of ICMI* (p. 185-192).

Chen, L., Harper, M. P., & Quek, F. (2002). Gesture patterns during speech repairs. In *Proceedings of ICMI* (p. 155-160).

Chen, L., Liu, Y., Harper, M. P., & Shriberg, E. (2004). Multimodal model integration for sentence unit detection. In *Proceedings of icmi* (p. 121-128).

Chen, L., Rose, R. T., Parrill, F., Han, X., Tu, J., Huang, Z., et al. (2005). VACE multimodal meeting corpus. In *Proceedings of the Workshop on Machine Learning for Multimodal Interaction (MLMI)* (p. 40-51).

Cohen, P. R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., et al. (1997). Quickset: Multimodal interaction for distributed applications. In *Proceedings of ACM Multimedia* (p. 31-40).

Condon, W., & Ogston, W. (1967). A segmentation of behavior. *Journal of Psychiatric Research*, *5*(3), 221-235.

Condon, W., & Osgton, W. (1971). Speech and body motion synchrony of the speaker-hearer. *The Perception of Language*, 150-184.

Core, M., & Allen, J. (1997). Coding dialogs with the DAMSL annotation scheme. *AAAI Fall Symposium on Communicative Action in Humans and Machines*, 28-35.

Corradini, A., Wesson, R. M., & Cohen, P. R. (2002). A map-based system using speech and 3d gestures for pervasive computing. In *Proceedings of ICMI* (p. 191-196).

Darrell, T., & Pentland, A. (1993). Space-time gestures. In *Proceedings of CVPR* (p. 335-340).

Daumé III, H., & Marcu, D. (2005). A large-scale exploration of effective global features for a joint entity detection and tracking model. In *Proceedings of HLT-EMNLP* (p. 97-104).

De Groot, M. H., & Schervish, M. J. (2001). *Probability and statistics*. Addison Wesley.

Deutscher, J., Blake, A., & Reid, I. (2000). Articulated body motion capture by annealed particle filtering. In *Proceedings of CVPR* (Vol. 2, p. 126-133).

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., & Weischedel, R. (2004). The automatic content extraction (ace) program: Tasks, data, and evaluation. In *Proceedings of language resources and evaluation (LREC).*

Dollár, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *ICCV VS-PETS.*

Efros, A. A., Berg, A. C., Mori, G., & Malik, J. (2003). Recognizing action at a distance. In *Proceedings of ICCV* (p. 726-733).

Eisenstein, J., Barzilay, R., & Davis, R. (2007). Turning lectures into comic books with linguistically salient gestures. In *Proceedings of AAAI* (p. 877-882).

Eisenstein, J., Barzilay, R., & Davis, R. (2008a). Discourse topic and gestural form. In *Proceedings of AAAI, in press.*

Eisenstein, J., Barzilay, R., & Davis, R. (2008b). Gestural cohesion for discourse segmentation. In *Proceedings of ACL, in press.*

Eisenstein, J., Barzilay, R., & Davis, R. (2008c). Modeling gesture salience as a hidden variable for coreference resolution and keyframe extraction. *Journal of Artificial Intelligence Research*, *31*, 353-398.

Eisenstein, J., & Davis, R. (2004). Visual and linguistic information in gesture classification. In *Proceedings of ICMI* (p. 113-120).

Eisenstein, J., & Davis, R. (2006). Natural gesture in descriptive monologues. In *ACM SIGGRAPH 2006 Courses* (p. 26).

Eisenstein, J., & Davis, R. (2007). Conditional modality fusion for coreference resolution. In *Proceedings of ACL* (p. 352-359).

Esposito, A., McCullough, K. E., & Quek, F. (2001). Disfluencies in gesture: Gestural correlates to filled and unfilled speech pauses. In *Proceedings of IEEE Workshop on Cues in Communication.*

Fayyad, U. M., & Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of IJCAI* (Vol. 2, p. 1022-1027).

Forsyth, D. A., & Ponce, J. (2003). *Computer vision: A modern approach*. Prentice Hall.

Galley, M., McKeown, K. R., Fosler-Lussier, E., & Jing, H. (2003). Discourse segmentation of multi-party conversation. *Proceedings of ACL*, 562-569.

Gavrila, D. M. (1999). Visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, *73*(1), 82-98.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*. Chapman and Hall/CRC.

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Harvard University Press.

Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining math: gesturing lightens the load. *Psychological Science*, *12*(6), 516-22.

Gregory, M., Johnson, M., & Charniak, E. (2004). Sentence-internal prosody does not help parsing the way punctuation does. In *Proceedings of HLT-NAACL* (p. 81-88).

Grishman, R., & Sundheim, B. (1995). Design of the MUC-6 evaluation. In *Proceedings of the 6th message understanding conference.*

Grosz, B., & Hirshberg, J. (1992). Some intonational characteristics of discourse structure. In *Proceedings of ICSLP* (p. 429-432).

Grosz, B., Joshi, A., & Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, *21*(2), 203-225.

Grosz, B., & Sidner, C. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, *12*(3), 175-204.

Haghighi, A., & Klein, D. (2007). Unsupervised coreference resolution in a nonparametric bayesian model. In *Proceedings of ACL* (p. 848-855).

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in english.* Longman.

Harabagiu, S. M., Bunescu, R. C., & Maiorano, S. J. (2001). Text and knowledge mining for coreference resolution. In *Proceedings of NAACL* (p. 1-8).

Hearst, M. A. (1994). Multi-paragraph segmentation of expository text. In *Proceedings of ACL* (p. 9-16).

Heeman, P. A., & Allen, J. F. (1995). *The TRAINS 93 dialogues* (Tech. Rep. No. TN94-2). Department of Computer Science, University of Rochester.

Hirschberg, J., & Nakatani, C. (1998). Acoustic indicators of topic segmentation. In *Proceedings of ICSLP.*

Hirschman, L., & Chinchor, N. (1998). MUC-7 coreference task definition. In *Proceedings of the Message Understanding Conference.*

Holler, J., & Beattie, G. (2003). Pragmatic aspects of representational gestures: Do speakers use them to clarify verbal ambiguity for the listener? *Gesture*, *3*(2), 127-154.

Huang, X., Acero, A., & Hon, H.-W. (2001). *Spoken language processing.* Prentice Hall.

Huang, X., Alleva, F., Hwang, M.-Y., & Rosenfeld, R. (1993). An overview of the Sphinx-II speech recognition system. In *Proceedings of ARPA Human Language Technology Workshop* (p. 81-86).

Johnston, M., Cohen, P. R., McGee, D., Oviatt, S. L., Pittman, J. A., & Smith, I. (1997). Unification-based multimodal integration. In *Proceedings of ACL* (p. 281-288).

Kahn, J. G., Lease, M., Charniak, E., Johnson, M., & Ostendorf, M. (2005). Effective use of prosody in parsing conversational speech. In *Proceedings of HLT-EMNLP* (p. 233-240).

Kameyama, M. (1986). A property-sharing constraint in Centering. In *Proceedings of ACL* (p. 200-206).

Kehler, A. (2000). Cognitive status and form of reference in multimodal human-computer interaction. In *Proceedings of AAAI* (p. 685-690).

Kelly, S., Barr, D., Church, R., & Lynch, K. (1999). Offering a Hand to Pragmatic Understanding: The Role of Speech and Gesture in Comprehension and Memory. *Journal of Memory and Language*, *40*(4), 577-592.

Kelly, S., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language*, *89*(1), 253-260.

Kendon, A. (1972). Some relationships between body motion and speech. In *Studies in dyadic communication* (p. 177-210). Pergamon Press.

Kendon, A. (1978). Differential perception and attentional frame in face-to-face interaction: Two problems for investigation. *Semiotica*, *24*, 305-315.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of the utterance. In M. R. Key (Ed.), *The relation between verbal and non-verbal communication* (p. 207-227). Mouton.

Kendon, A. (1995). Gestures as illocutionary and discourse structure markers in Southern Italian conversation. *Journal of Pragmatics*, *23*(3), 247-279.

Kendon, A. (2004). *Gesture: Visible action as utterance.* Cambridge University Press.

Kettebekov, S., & Sharma, R. (2000). Understanding gestures in multimodal human computer interaction. *International Journal on Artificial Intelligence Tools*, *9*(2), 205–223.

Kettebekov, S., Yeasin, M., & Sharma, R. (2005). Prosody based audiovisual coanalysis for coverbal gesture recognition. *IEEE Transactions on Multimedia*, *7*(2), 234-242.

Kibble, R., & Power, R. (2004). Optimising referential coherence in text generation. *Computational Linguistics*, *30*(4), 401-416.

Kim, J., Schwarm, S. E., & Osterdorf, M. (2004). Detecting structural metadata with decision trees and transformation-based learning. In *Proceedings of HLT-NAACL* (p. 137-144).

Kopp, S., Tepper, P., Ferriman, K., & Cassell, J. (2007). Trading spaces: How humans and humanoids use speech and gesture to give directions. In T. Nishida (Ed.), *Conversational informatics: An engineering approach.* Wiley.

Krauss, R. (2001). Why do we gesture when we speak? *Current Directions in Pschological Science*, *7*(54-59).

Krauss, R., Morrel-Samuels, P., & Colasante, C. (1991). Do conversational gestures communicate? *Journal of Personality and Social Psychology*, *61*(5), 743-754.

Kudo, T., & Matsumoto, Y. (2001). Chunking with support vector machines. In *Proceedings of NAACL* (p. 1-8).

Lappin, S., & Leass, H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, *20*(4), 535-561.

Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision*, *64*(2-3), 107-123.

Lascarides, A., & Stone, M. (2006). Formal Semantics for Iconic Gesture. In *Proceedings of the 10th workshop on the semantics and pragmatics of dialogue (brandial)* (p. 64-71).

Lausberg, H., & Kita, S. (2003). The content of the message influences the hand choice in co-speech gestures and in gesturing without speaking. *Brain and Language*, *86*(1), 57-69.

Lew, M. S., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on*

Multimedia Computing, Communications and Applications, *2*(1), 1-19.

Li, X., & Roth, D. (2001). Exploring evidence for shallow parsing. In *Proceedings of CoNLL* (p. 1-7).

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, *37*, 145-151.

Litman, D. J. (1996). Cue phrase classification using machine learning. *Journal of Artificial Intelligence Research*, *5*, 53-94.

Liu, D. C., & Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, *45*, 503-528.

Liu, T., & Kender, J. R. (2007). Computational approaches to temporal sampling of video sequences. *ACM Transactions on Multimedia Computing, Communications and Applications*, *3*(2), 7.

Liu, Y. (2004). *Structural event detection for rich transcription of speech*. Unpublished doctoral dissertation, Purdue University.

Loehr, D. (2007). Aspects of rhythm in gesture and speech. *Gesture*, *7*(2), 179-214.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of ICCV* (Vol. 2, p. 1150-1157).

Luo, X. (2005). On coreference resolution performance metrics. In *Proceedings of HLT-EMNLP* (p. 25-32).

Malioutov, I., & Barzilay, R. (2006). Minimum cut model for spoken lecture segmentation. In *Proceedings of ACL* (p. 25-32).

Malioutov, I., Park, A., Barzilay, R., & Glass, J. (2007). Making sense of sound: Unsupervised topic segmentation over acoustic input. In *Proceedings of ACL* (p. 504-511).

Mani, I., & Maybury, M. T. (Eds.). (1999). *Advances in automatic text summarization*. MIT Press.

Martell, C. (2005). *FORM: An experiment in the annotation of the kinematics of gesture*. Unpublished doctoral dissertation, University of Pennsylvania.

Martell, C., Howard, P., Osborn, C., Britt, L., & Myers, K. (2003). *FORM2 kinematic gesture* (Tech. Rep. No. LDC2003V01). The Linguistic Data Consortium.

McCallum, A., & Wellner, B. (2004). Conditional models of identity uncertainty with application to noun coreference. In *Proceedings of NIPS* (p. 905-912).

McNeill, D. (1992). *Hand and mind*. The University of Chicago Press.

McNeill, D. (2005). *Gesture and thought*. The University of Chicago Press.

Melinger, A., & Levelt, W. J. M. (2004). Gesture and communicative intention of the speaker. *Gesture*, *4*(2), 119-141.

Morton, T. S. (1999). Using coreference in question answering. In *Proceedings of TREC.*

Müller, C. (2007). Resolving it, this, and that in unrestricted multi-party dialog. In *Proceedings of ACL* (p. 816-823).

Nakano, Y., Reinstein, G., Stocky, T., & Cassell, J. (2003). Towards a model of face-to-face grounding. In *Proceedings of ACL* (p. 553-561).

Ng, V. (2007). Shallow semantics for coreference resolution. In *Proceedings of IJCAI* (p. 1689-1694).

Ng, V., & Cardie, C. (2002). Improving machine learning approaches to coreference resolution. In *Proceedings of ACL* (p. 104-111).

Niebles, J. C., Wang, H., & Fei-Fei, L. (2006). Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. In *Proceedings of the British Machine Vision Conference.*

NIST. (2003). *The Rich Transcription Fall 2003 (RT-03F) Evaluation plan.*

Passonneau, R. J. (1997). *Applying reliability metrics to co-reference annotation* (Tech. Rep. No. CUCS-017-97). Columbia University.

Passonneau, R. J. (2004). Computing reliability for coreference annotation. In (Vol. 4).

Pevzner, L., & Hearst, M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, *28*(1), 19-36.

Pierrehumbert, J., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan, & M. Pollack (Eds.), *Intentions in communication* (p. 271-311). MIT Press.

Poddar, I., Sethi, Y., Ozyildiz, E., & Sharma, R. (1998). Toward natural gesture/speech HCI: A case study of weather narration. In *Proceedings of Perceptual User Interfaces* (p. 1-6).

Poesio, M., & Artstein, R. (n.d.). The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the workshop on frontiers in corpus annotations ii: Pie in the sky* (pp. 76–83).

Poesio, M., Stevenson, R., Eugenio, B. D., & Hitzeman, J. (2004). Centering: a parametric theory and its instantiations. *Computational Linguistics*, *30*(3), 309-363.

Ponzetto, S. P., & Strube, M. (2007). Knowledge Derived From Wikipedia For Computing Semantic Relatedness. *Journal of Artificial Intelligence Research*, *30*, 181-212.

Purver, M., Griffiths, T., Körding, K., & Tenenbaum, J. (2006). Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of ACL* (p. 17-24).

Quattoni, A., Wang, S., Morency, L.-P., Collins, M., & Darrell, T. (2007). Hidden conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(10), 1848-1852.

Quek, F. (2003). The catchment feature model for multimodal language analysis. In *Proceedings of ICCV.*

Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X., Kirbas, C., et al. (2002). Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction*, *9:3*, 171-193.

Quek, F., McNeill, D., Bryll, R., & Harper, M. (2002). Gestural spatialization in natural discourse segmentation. In *Proceedings of ICSLP* (p. 189-192).

Quek, F., McNeill, D., Bryll, R., Kirbas, C., Arslan, H., McCullough, K. E., et al. (2000). Gesture, speech, and gaze cues for discourse segmentation. In *Proceedings of CVPR* (Vol. 2, p. 247-254).

Quek, F., Xiong, Y., & McNeill, D. (2002). Gestural Trajectory Symmetries and Discourse Segmentation. In *Proceedings of ICSLP.*

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257-286.

Rauscher, F. H., Krauss, R. M., & Chen, Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science*, *7*(4), 226-231.

Rimé, B., & Schiaratura, L. (1991). Gesture and speech. In *Fundamentals of nonverbal behavior.* Press Syndicate of the University of Cambridge.

Sha, F., & Pereira, F. (2003). Shallow parsing with conditional random fields. In *Proceedings of NAACL* (p. 134-141).

Sharma, R., Cai, J., Chakravarthy, S., Poddar, I., & Sethi, Y. (2000). Exploiting speech/gesture co-occurence for improving continuous gesture recognition in weather narration. In *Proceedings of Face and Gesture Recognition.*

Shriberg, E., Stolcke, A., Hakkani-Tur, D., & Tur, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, *32*.

Sidner, C. L. (1979). *Towards a computational theory of definite anaphora comprehension in english discourse* (Tech. Rep. No. AITR-537). Massachusetts Institute of Technology.

So, W. C., Coppola, M., Licciardello, V., & Goldin-Meadow, S. (2005). The seeds of spatial grammar in the manual modality. *Cognitive Science*, *29*, 1029 - 1043.

Soon, W. M., Ng, H. T., & Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, *27*(4), 521-544.

Steedman, M. (1990). Structure and intonation in spoken language understanding. In *Proceedings of ACL* (p. 9-16).

Strube, M., & Hahn, U. (1999). Functional centering: grounding referential coherence in information structure. *Computational Linguistics*, *25*(3), 309-344.

Strube, M., & Müller, C. (2003). A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of ACL* (p. 168-175).

Strube, M., Rapp, S., & Müller, C. (2002). The influence of minimum edit distance on reference resolution. In *Proceedings of EMNLP* (p. 312-319).

Sundaram, H., & Chang, S.-F. (2003). Video analysis and summarization at structural and semantic levels. In W. C. S. D. Feng & H. Zhang (Eds.), *Multimedia information retrieval and management: Technological fundamentals and applications* (p. 75-94). Springer Verlag.

Sutton, C., & McCallum, A. (2006). An introduction to conditional random fields for relational learning. In L. Getoor & B. Taskar (Eds.), *Introduction to statistical relational learning* (p. 95-130). MIT Press.

Swerts, M. (1997). Prosodic features at discourse boundaries of different strength. *The Journal of the Acoustical Society of America*, *101*, 514.

Tur, G., Hakkani-Tur, D., Stolcke, A., & Shriberg, E. (2001). Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics*, *27*(1), 31-57.

Uchihashi, S., Foote, J., Girgensohn, A., & Boreczky, J. (1999). Video manga: generating semantically meaningful video summaries. In *Proceedings of ACM*

MULTIMEDIA (p. 383-392).

Utiyama, M., & Isahara, H. (2001). A statistical model for domain-independent text segmentation. In *Proceedings of ACL* (p. 491-498).

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., & Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of Message Understanding Conference (MUC)* (p. 45-52).

Walker, M. (1998). Centering, anaphora resolution, and discourse structure. In M. Walker, A. Joshi, & E. Prince (Eds.), *Centering theory in discourse* (p. 401-435). Oxford University Press.

Walker, M., Joshi, A., & Prince, E. (Eds.). (1998). *Centering theory in discourse*. Oxford University Press.

Whitney, P. (1998). *The psychology of language*. Houghton Mifflin.

Xiong, Y., & Quek, F. (2006). Hand Motion Gesture Frequency Properties and Multimodal Discourse Analysis. *International Journal of Computer Vision*, *69*(3), 353-371.

Xiong, Y., Quek, F., & McNeill, D. (2002). Hand gesture symmetric behavior detection and analysis in natural conversation. In *Proceedings of ICMI* (p. 179-184).

Yamron, J., Carp, I., Gillick, L., Lowe, S., & Mulbregt, P. van. (1998). A hidden markov model approach to text segmentation and event tracking. *Proceedings of IEEE Conference on Acoustics, Speech and Signal Processing*, *1*, 333-336.

Zhu, X., Fan, J., Elmagarmid, A., & Wu, X. (2003). Hierarchical video content description and summarization using unified semantic and visual similarity. *Multimedia Systems*, *9*(1), 31-53.