*Artificial Intelligence for Engineering Design, Analysis and Manufacturing* (2011), *25*, 237–254.

doi:10.1017/S0890060411000060

# Using speech to identify gesture pen strokes in collaborative, multimodal device descriptions

JAMES HEROLD[1] AND THOMAS F. STAHOVICH[2]

[1]Department of Computer Science and Engineering, University of California, Riverside, California, USA
[2]Department of Mechanical Engineering, University of California, Riverside, California, USA

(RECEIVED May 10, 2010; ACCEPTED February 14, 2011)

**Abstract**

One challenge in building collaborative design tools that use speech and sketch input is distinguishing gesture pen strokes from those representing device structure, that is, object strokes. In previous work, we developed a gesture/object classifier that uses features computed from the pen strokes and the speech aligned with them. Experiments indicated that the speech features were the most important for distinguishing gestures, thus indicating the critical importance of the speech–sketch alignment. Consequently, we have developed a new alignment technique that employs a two-step process: the speech is first explicitly segmented (primarily into clauses), then the segments are aligned with the pen strokes. Our speech segmentation step is unique in that it uses sketch features for locating segment boundaries in multimodal dialog. In addition, it uses a single classifier to directly combine word-based, prosodic (pause), and sketch-based features. In the second step, segments are initially aligned with strokes based on temporal correlation, and then classifiers are used to detect and correct two common alignment errors. Our two-step technique has proven to be substantially more accurate at alignment than the existing technique that lacked explicit segmentation. It is more important that, for nearly all cases, our new technique results in greater gesture classification accuracy than the existing technique, and performed nearly as well as the benchmark manual speech–sketch alignment.

**Keywords:** Design Descriptions; Gesture/Object Stroke Classification; Multimodal Dialog; Speech Segmentation; Speech–Sketch Alignment

## 1. INTRODUCTION

Designers often communicate design concepts to each other with informal sketches, speech, and gestures. Although the importance of such communication has long been recognized by designers (Ullman et al., 1990), traditional design tools do not support this in any substantive way. Our long-term goal is to remedy this by creating computational techniques to enable collaborative design tools that support natural multimodal communication.

In previous work (Bischel et al., 2009), we conducted a study to examine the nature of multimodal communication in collaborative design. Specifically, we examined how designers use natural free-form sketching and speaking to describe the structure and behavior of a mechanical device. We found that both the sketch and speech are essential to such descriptions, and that typically neither modality can be understood without the other. In addition, the wide variety of information contained in the sketches makes them particularly challenging to interpret. Although many of the pen strokes portray device structure, others are gestures, such as arrows used to indicate motion, or circles used to single out a component being discussed. Figure 1a, which depicts a pair of C-clamp vise-grip pliers, is a typical sketch from the study. Consider the challenge such a drawing poses for any sketch-understanding software. To understand this sketch, it is first necessary to distinguish the *gesture strokes* (Fig. 1b) from the *object strokes* representing device structure or handwritten text (Fig. 1c).

Separating strokes in this way is valuable beyond the obvious purpose of facilitating sketch recognition. Most, if not all, gesture strokes have only temporary value. For example, gestures resolving deictic references or indicating the motion of a part may be superfluous once the discussion has moved on to a new topic. However, over time, such gestures accumulate (e.g., Fig. 1a), obscuring the sketch and hindering discussion. Detecting these and removing them from view when they are no longer needed may enable more efficient communication.

As part of our work in Bischel et al. (2009), we developed a technique for distinguishing gesture strokes from object

Reprint requests to: James Herold, Department of Computer Science and Engineering, University of California, Riverside, 3329 Utah Street, Riverside, CA 92507, USA. E-mail: jhero001@ucr.edu

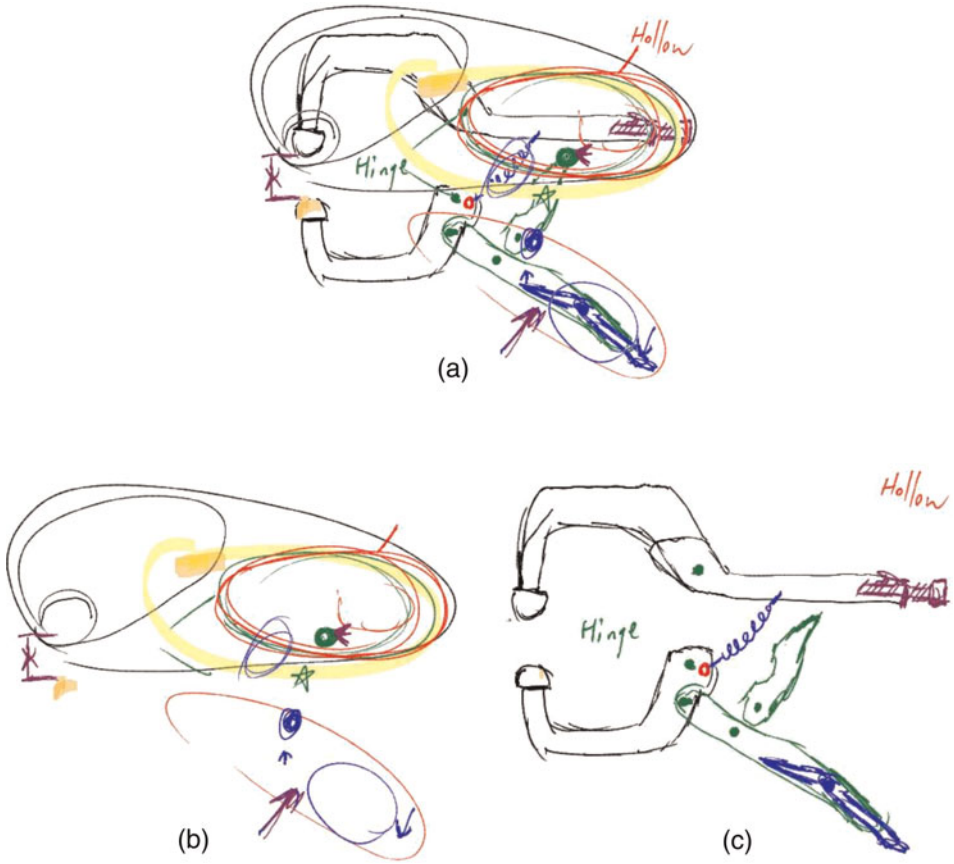


**Fig. 1.** (a) A sketch of C-clamp vise-grip pliers, (b) gesture pen strokes, and (c) pen strokes representing device structure and text. [A color version of this figure can be viewed online at journals.cambridge.org/aie]

strokes. The technique employs a statistical classifier that uses features of both the sketch and speech. The sketch features compute geometric properties of the strokes, and the spatial and temporal relationships between them. The speech features compute statistical properties of the speech aligned with each stroke. Experiments with the technique indicated that the speech modality is more important than the sketch modality for gesture/object classification: the single most effective feature for classification was a speech feature.

The importance of speech for gesture/object classification suggests that the accuracy of the speech–sketch alignment process is critical to gesture classification. The work in Bischel et al. (2009), used a "3-sec" alignment technique, in which the speech and sketch input were aligned based on temporal correlation. Each stroke was associated with the words that at least partially coincided with a temporal window extending three seconds on either side of the stroke. Our present work is focused on measuring the performance of this 3-sec alignment technique, and developing a new alignment technique to overcome some of its limitations, thus enabling more accurate gesture/object classification.

To evaluate the 3-sec alignment technique, we began by manually aligning the speech and sketch from the study in Bischel et al. (2009), as illustrated in Figure 2. We did this by first segmenting the speech primarily into clauses and then aligning these with the strokes to which they refer. Comparison of the 3-sec and manual alignment revealed that the former has substantial room for improvement. For example, for 41% of pen strokes, there was no intersection between the 3-sec alignment and the correct (manual) alignment.

Consequently, we sought to develop an improved alignment technique, which we modeled on our manual alignment process. The new technique employs an explicit speech segmentation process, followed by a segment–stroke alignment process. Because both processes employ statistical classifiers, we call our technique "classifier-based alignment" (CBA). Evaluation of the new technique demonstrated that it produces considerably more accurate alignment than the 3-sec technique. It is more important, however, that it results in substantially better gesture classification accuracy.

This work makes several contributions. First, we developed a technique for segmenting speech into meaningful clauses. The technique is well suited to the ungrammatical speech characteristic of multimodal dialog. The technique is effective, in part, because it uses information from the sketch input to help process the speech. Second, we developed a novel technique for aligning the segmented speech with the pen strokes to which it refers. These two efforts combine to produce an effective and accurate speech–sketch alignment technique for multimodal dialog. Third, we demonstrated that the

https://doi.org/10.1017/S0890060411000060 Published online by Cambridge University Press

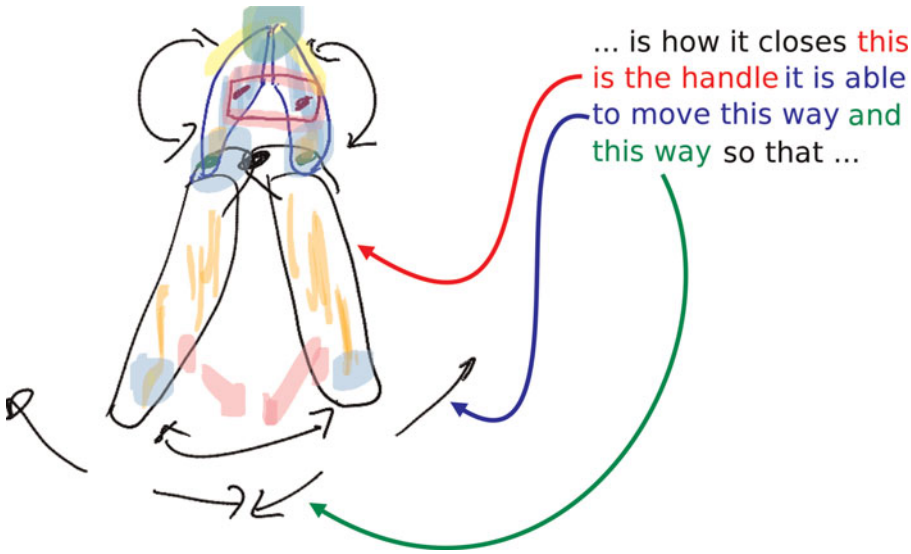


**Fig. 2.** An example of the alignment of strokes with the speech that refers to them. The bold arrows link the word groups with the associated pen strokes. [A color version of this figure can be viewed online at journals.cambridge.org/aie]

new alignment technique enables accurate classification of gesture and object strokes in a multimodal dialog.

The next section places this work in the context of related work. This is followed in Section 3 by a description of our study from Bischel et al. (2009) and the gesture classification technique we developed in that work. Section 4 describes our manual speech–sketch alignment process and presents an evaluation of the alignment accuracy of the 3-sec technique. Next, Section 5 describes our classifier-based speech–sketch alignment technique, including the speech segmentation technique it employs. Section 6 presents the gesture classification accuracy obtained using the two speech–sketch alignment techniques and compares this to the accuracy achieved via manual alignment. After a discussion of these results in Section 7, conclusions are presented in Section 8.

## 2. RELATED WORK

Multimodal systems date back at least to the work of Brown (1979), with subsequent early multimodal systems incorporating typed language and pointing with a mouse or light-pen (Wauchope, 1994; Woods et al., 1979). Bolt's "put-that-there" system (Bolt, 1980) was the first to incorporate early speech and three-dimensional pointing recognizers. Quickset (Cohen et al., 1997) explores a general architecture for multimodal fusion, but unlike our work, QuickSet is a command-based system, that is, the utterances are used as verbal replacements for mouse or menu commands. The iMap system handles free-hand gestures in a map–control user interface, using prosody cues to improve gesture recognition (Krahnstoever et al., 2002). The system in Johnston et al. (2002) provides a speech and pen interface to restaurant and subway information for New York City, but it is not a sketching system and has only text recognition and basic circling and pointing gestures for the graphical input modality.

Other applications of speaking and sketching include an early effort that used a diagram and written English text (Novak & Bulko, 1993), interesting in part because it used a blackboard to help establish the reference relationships between the graphical and text entities. BBN's Portable Voice Assistant (Oviatt, 2000) uses pen and voice input to enter and retrieve information on the Web. Their system integrates simultaneous speech and gesture inputs using a frame-based system. The Human-Centric Word Processor (Oviatt et al., 2000) enables radiologists to use pen-based selection gestures and command-based speech for postdictation correction of transcriptions. nuSketch COA Creator (Forbus et al., 2001) is designed as a general purpose multimodal architecture, allowing users to sketch and talk to add symbols to a military map using commands like "add severely restricted terrain." This system also uses command-based speech, and is focused on issues of reasoning about the content of the sketch rather than on recognition: the user assigns symbolic labels to the sketched objects.

Many systems have benefited from the series of empirical studies of multimodal communication in (Oviatt et al., 1997). Cassell (1998) was among the first to argue that natural, free-hand gestures can be relevant to human computer interaction, and presented a helpful framework for gestural interaction. Oviatt (1999) has demonstrated advantages of multimodal interfaces, noting that multimodal input simplifies the users' vocabulary and improves accuracy with accented speakers.

Our work is grounded in insights about how people use multimodal explanations to describe devices. Ullman et al. (1990) found that engineers commonly use five different categories of pen strokes in a sketch. His "support" and "draw" strokes are analogous to our categories of gesture and object strokes. Heiser and Tversky (2006) concluded that when there are numerous arrow gestures in a sketch, students can more easily understand the functionality of a device, illustrating the importance of gestures in a design sketch.

https://doi.org/10.1017/S0890060411000060 Published online by Cambridge University Press

Much of the previous work in understanding descriptions of mechanical devices has focused solely on sketching of structure (e.g., Bloomenthal & Zeleznik, 1998; Masry et al., 2005). By contrast, GIDeS++ (Silva & Cardoso, 2004) is a multimodal system specifically designed to understand descriptions of mechanical devices, but it uses pen strokes to replace mouse functionality rather than attempting to maintain a natural sketching environment. Likewise, ASSISTANCE (Oltmans, 2000) incorporates spoken behavioral descriptions to supplement the understanding of mechanical device sketches. However, it relies on limited vocabularies of speech patterns that must be explicitly identified in advance, where our system can adapt to new patterns via user-provided training data.

Hand and arm gestures have long been a topic of research. Kendon (1997) provides an overview of the study of gesture, dating back to work by Quintilianus (circa the first century) in which he details how an orator ought to use gesture in discourse. More relevant to our work, Kendon explores the organization of speech and gestures. He finds that speech is organized into "idea units" marked by prosodic features, such as pitch level and loudness, rather than by lexical properties. Similarly, gestures are organized into "gesture units." This suggests the need to segment our speech prior to aligning it with pen strokes. However, we segment speech based primarily on lexical considerations, and align each pen stroke with at most one speech segment.

Efron (1941) classifies hand and arm gestures across three dimensions: the trajectory of the gesture, whether the gesture involves the listener, and whether the gesture inherently contains semantic information. In our domain, gestures do not directly involve a listener, but they do contain semantic information that is frequently conveyed through shape.

We find parallels between the pen stroke gestures considered in our work and the hand and arm gestures studied by McNeill (1992). In McNeill's classification scheme, hand/arm gestures describing objects or actions are called *imagistic*, whereas those that do not evoke imagery are called *nonimagistic*. *Imagistic* gestures are further subdivided into *iconic* or *metaphoric* gestures. The former represent concrete concepts, such as a speaker illustrating how they threw a baseball by mimicking the action of throwing. The latter present abstract imagery, such as a person balling their fists and then quickly spreading their fingers to convey a metaphoric "explosion," illustrating frustration about the topic of discussion (McNeill, 1992). Nonimagistic gestures are also divided into two categories: *deictic* and *beats*. The former are pointing gestures, whereas the latter are typically involuntary movements of the hands made while speaking, and which carry no meaning.

Gestures are often understood in the context of accompanying speech. Oviatt et al. (1997) studied humans interacting with dynamic mapping software, quantifying the likelihood that speaking or sketching would occur first or that they would start simultaneously. This work was extended by Adler and Davis (2007) for design descriptions, who found consistent time delay patterns between when a pen stroke was drawn and when the related speech was spoken. The 3-sec speech–sketch alignment technique in Bischel et al. (2009) builds on this.

The findings of Oviatt et al. (1997) and Adler and Davis (2007) are at odds with findings of McNeill (1992), which suggest that speech always co-occurs with its referent gesture. This discrepancy is likely due to the differences in the domains considered. Oviatt and Adler consider speaking and drawing, whereas McNeill considers hand and arm gestures made during typical conversation. Although hand/arm gestures are often made with minimal effort or concentration, drawing can often require enough concentration so as to interrupt speaking. Likewise, drawing is inherently slower than hand/arm gesturing, which may contribute significantly to the differences in gesture/speech alignment between the two domains.

The work by Chai et al. (2004, 2005) sets interaction in the context of a dialogue, using context, semantics, and linguistic principles to resolve gestural references. Our task is different in that we must differentiate between gesture and object pen strokes and we consider ungrammatical, disfluent speech, whereas they assume the speech is unambiguous. Furthermore, they interpret interaction in the context of a predefined image, whereas we consider an incrementally created sketch whose meaning is not known in advance.

There have been several prior efforts focused on segmenting speech into phrases and sentences. For example, Nakai and Shimodaira (1994) describe a method that uses prosodic features to segment speech into accent phrases. A least-squares approach is used to find the optimum match between the speech and pitch pattern templates. A 97% segmentation accuracy is reported for a case in which the 30 best candidate segmentations are considered.

Most current techniques for identifying sentence boundaries in speech transcriptions are based on a hidden Markov model (Stolcke & Shriberg, 1996; Stolcke et al., 1998; Gotoh & Renals, 2000; Hwan Kim & Woodland, 2001). An n-gram language model is used to describe the joint distribution of words and sentence boundaries, which are modeled as events that occur between words. Many methods also use prosodic features for locating sentence boundaries. For example, Gotoh and Renals (2000) combine their n-gram language model with a prosodic model based on pause duration. Likewise, Hwan Kim and Woodland (2001) use a prosodic model based on 10 features. Stolcke and Shriberg (1996) included part of speech information in an n-gram language model, and found that this improves accuracy. In later work, Stolcke et al. (1998) augmented their n-gram language model with turn boundaries (change in speaker) and long pauses. All of these methods for locating sentence boundaries have been applied to telephone conversation and news broadcasts, whereas we consider a multimodal context with both speaking and sketching. In addition, although these methods classify interword events as boundaries and nonboundaries, we classify words according to their position in a speech segment. As described in Section 5.1, this allows us to take advantage of the frequent occurrence of single-word segments.

Sentence boundary detection methods vary in the way they combine the language and prosodic models. Stolcke et al.

https://doi.org/10.1017/S0890060411000060 Published online by Cambridge University Press

(1998) explore a variety of combination techniques including model interpolation, independent model combination, and joint modeling. In the latter case, a decision tree is used to combine posterior probabilities from the language model with prosodic features. Similarly, Liu et al. (2004) use a maximum entropy model to combine prosodic and word-level features. We do not use an explicit language model, but instead use a single classifier (Ada-boosted C4.5 decision tree) to directly combine word-based, prosodic (pause), and sketch-based features.

Our gesture classifier (Bischel et al., 2009) is related to the work of Patel et al. (2007) and Bishop, Svensen, and Hinton (2004) on separating text strokes from nontext strokes. These works differ from ours in considering only features from the sketch, where we examine the accompanying speech. In addition, in their work, text consists of a consistent set of letter and number glyphs, where the gestures in our domain are often unique, and frequently have the same shapes as object strokes.

Our gesture classifier also builds on work in shape recognition by using the kinds of features used by feature-based recognizers, such as in Patel et al. (2007) and Rubine (1991). Our system relies on some of the features these systems use, but it also extracts new features to address the special nature of identifying free-form gestures.

In examining properties of the accompanying speech, our gesture classifier does not try to understand it but simply identify it as that which accompanies either a gesture or object stroke. We do this with Bayesian filters (Graham, 2004) and Markovian filters (Yerazunis, 2004).

In summary, our work differs from much of the work in multimodal interfaces in that we consider free-form speech and sketching, rather than a predefined vocabulary. Similarly, although most multimodal systems use speech and sketch input as a substitute for mouse/menu commands, we consider the task of classifying sketch input as gesture and object strokes. Although many speech segmentation techniques exist, ours is novel in that it uses information from the sketch modality. In addition, it uses a single classifier to directly combine word-based, prosodic (pause), and sketch-based features. Finally, our speech–sketch alignment technique is novel in that it works from segmented speech and uses classifiers to detect and repair common alignment errors.

## 3. BACKGROUND: DISTINGUISHING GESTURE AND OBJECT STROKES

As described in Bischel et al. (2009), we conducted a study to characterize how designers use natural free-form sketching, speaking, and gesturing to communicate design descriptions to each other.[1] The study involved descriptions of four devices: C-clamp vise-grip pliers, bolt cutters, an air pump for inflating balls, and a door lock (Fig. 1, Fig. 2, and Fig. 3). The participants were 16 graduate and senior undergraduate mechanical engineering students at UC Riverside. Four were female. English was the primary language for 9 participants, but the speech of 10 participants was indistinguishable from that of native English speakers. Eleven participants received their engineering instruction in English. There were only 4 participants that both did not have English as their primary language and did not receive engineering instruction in English. Fourteen participants had previously taken a course in engineering drawing, and 7 had completed a team-based project-design course.

Each study session involved a pair of participants placed in separate rooms and allowed to communicate using tablet PCs, microphones, and headphones. The tablets provided a shared drawing environment with a pen, highlighter, and eraser, and the ability to select from several ink colors. The audio and drawing were recorded with timestamps.

During a session, one participant was asked to describe a device to his or her partner, who could ask clarifying questions. At the end of the description, both participants were asked survey questions about the structure and behavior of the device. To motivate effective dialog, the participants were informed that their compensation would be based on the accuracy of their answers. (All participants were given the maximum compensation.) The two participants repeated this process three times, switching roles, so that each participant described two devices. In all, a total of 48 device descriptions were collected.

Figures 1, 2, and 3 show typical examples of sketches collected in the study. As discussed above, these sketches contain two types of pen strokes: *object strokes* and *gesture strokes.* The former depict device structure or comprise text. The latter can be classified into two categories, adopted from the terminology developed by McNeill (1992). Strokes that demonstrate an action, such as an arrow illustrating the direction in which the handles of a pair of vice grips may move, are iconic gestures. Similarly, strokes that resolve deictic references from the speech modality are deictic gestures. These gestures may take many forms, such as tapping, circling, highlighting, and tracing. Object strokes could be considered iconic gestures, as they provide a representation of an object. However, we distinguish between object strokes and other iconic gesture strokes because our goal is to separate the representation of a device's structure from the description of its behavior.

### 3.1. Classifier design

As Figure 1 illustrates, there can be a comparable number of gesture and object strokes in a sketch, making it challenging to understand the final image. There is a clear need for techniques to separate the two types of strokes. This would at first appear to be a shape recognition problem solvable with standard shape recognizers such as those in Kara and Stahovich (2005) and Wobbrock et al. (2007). However, this problem is not amenable to such approaches for several reasons. First, gesture and object strokes can have arbitrary shapes, but shape recognizers require a predefined set of shapes. Second,

[1] This section presents an overview of work from (Bischel et al., 2009). For complete details, refer to Bischel et al. (2009).

https://doi.org/10.1017/S0890060411000060 Published online by Cambridge University Press

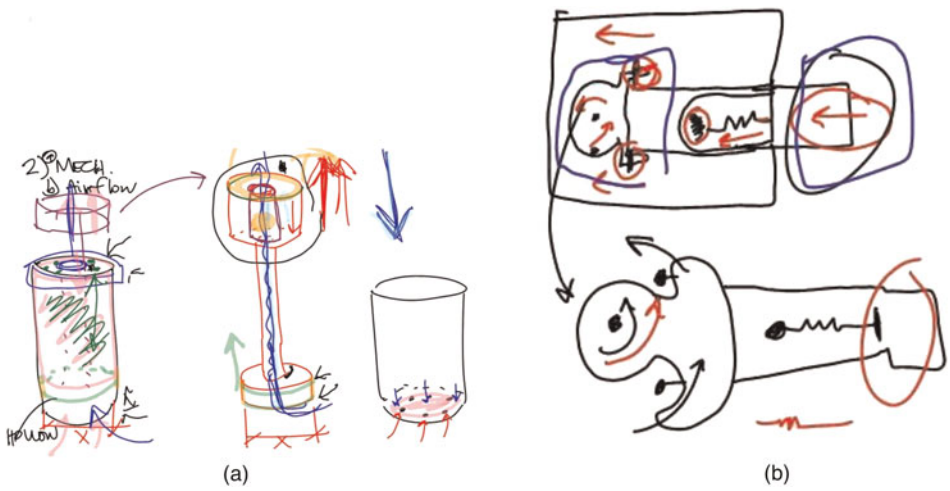


**Fig. 3.** Two of the devices from the study: (a) an air pump for inflating balls and (b) a door lock. [A color version of this figure can be viewed online at journals.cambridge.org/aie]

gesture and object strokes may be identical, and thus shape alone does not distinguish between the two classes of strokes. For example, a common selection gesture consists of tracing the shape of an object.

For these reasons, the gesture/object classifier described in Bischel et al. (2009) does not explicitly consider the shape of the pen stroke. Instead, each pen stroke is represented by features that are computed from both the sketch and speech input. The sketch features describe properties of the pen strokes, and the spatial and temporal relationships between them. The speech features describe properties of the speech aligned with each stroke. These features serve as inputs to a neural network which classifies a stroke as a gesture or object stroke.

### 3.1.1. Sketch features

The complete set of sketch features used for gesture/object classification is listed in Table 1. The first six features concern individual strokes, where $D_{SL}$ is the length of the pen stroke, $D_{SED}$ is the distance between its first and last points, $D_{AC}$ is the sum of the absolute value of the curvature along a stroke, and $D_{DC}$ is similar to curvature but is biased toward diagonal drawing directions. The ink density ($D_{ID}$) is a measure of the compactness of the stroke. The highlighter feature ($D_{HL}$) has a value of 1 if the stroke was made with a highlighter rather than an ordinary pen and is 0 otherwise.

The remaining 10 features describe the temporal and spatial relationships between strokes, in which $D_{DPS}$ and $D_{DNS}$ are the distance to the previous and next strokes, $D_{TPS}$ and $D_{TNS}$ are the time between the stroke and the previous and next strokes, $D_{TCS}$ is the time between the stroke and the closest previously drawn stroke, and $D_{ET}$ is the total elapsed time. The underlying color similarity ($D_{UCS}$) measures the extent to which earlier nearby strokes have the same color as the stroke. Underlying ink density ($D_{UID}$) is the density of the ink from other earlier pen strokes in the neighborhood (expanded bounding box) of the stroke. The two Hausdorff features (Kara & Stahovich, 2005) measure the extent to which a stroke traces underlying strokes. For each point on the stroke, the closest distance to a point on another earlier stroke is computed. Here, $D_{MHD}$ is the maximum of these closest distances, and $D_{AHD}$ is the average.

### 3.1.2. Speech features

To compute the speech features, it was first necessary to align the speech and sketch input, that is, determine which words are associated with each pen stroke. The 3-sec alignment technique presented in Bischel et al. (2009) was grounded in

**Table 1.** *Features for gesture versus object classification*

| Name | Description | Units |
|---|---|---|
| $D_{SL}$ | Stroke length | Pixel |
| $D_{SED}$ | Start to end distance | Pixel |
| $D_{AC}$ | Total absolute curvature | Radian |
| $D_{DC}$ | Diagonally biased curvature | Radian |
| $D_{ID}$ | Ink density | % |
| $D_{HL}$ | Highlighter | Boolean |
| $D_{DPS}$ | Distance to previous stroke | Pixel |
| $D_{DNS}$ | Distance to next stroke | Pixel |
| $D_{TPS}$ | Time to previous stroke | ms |
| $D_{TNS}$ | Tme to next stroke | ms |
| $D_{TCS}$ | Time to closest prior stroke | ms |
| $D_{ET}$ | Total elapsed time | ms |
| $D_{UCS}$ | Underlying color similarity | % |
| $D_{UID}$ | Underlying ink density | % |
| $D_{MHD}$ | Max. Hausdorff distance to underlying ink | Pixel |
| $D_{AHD}$ | Ave. Hausdorff distance to underlying ink | Pixel |
| $W_{TPS}$ | Time to previous speaker | ms |
| $W_{WC}$ | No. of words in temporal window | Word |
| $W_{BF}$ | Bayesian filter | Probability |
| $W_{TBF}$ | Thesaurus Bayesian filter | Probability |
| $W_{MF}$ | Markovian filter | Probability |

*Note:* $D_x$, sketch (drawing) feature; $W_x$, speech (word) feature.

https://doi.org/10.1017/S0890060411000060 Published online by Cambridge University Press

observations by Adler and Davis (2007) and Oviatt et al. (1997) suggesting that there is a strong temporal correlation between speaking and drawing. This technique employs a temporal window extending 3 s before and after the stroke. It is assumed that any words falling at least partially within this window are associated with the stroke. It is possible that a word may be associated with more than one stroke, or that a stroke may have no words associated with it.

The speech features associated with a pen stroke (Table 1) are computed from the speech aligned with it. To avoid inaccuracies inherent in current state of the art speech to text tools, the speech was manually transcribed and then Sphinx (Huang et al., 1993) was used to align the text with the recorded audio to find time stamps for the words. The words were also labeled with the identity of the speaker. Using manual transcriptions provides an upper bound on the contribution of the speech content to gesture classification. However, the speech may contain other valuable information, such as prosody, which was not considered.

The simplest speech features are the time to the previous speaker ($W_{TPS}$) and the number of words aligned with the stroke ($W_{WC}$). The other speech features concern the words themselves. Understanding grammatically correct speech is difficult enough; the speech considered here is ungrammatical, filled with pauses, repetitions, and disfluencies like "um" and "ah." Trying to perform semantic analysis on such ungrammatical text is intractable at present. As an alternative, statistical models are used to predict whether a set of words corresponds to a gesture or object stroke.

The first statistical speech feature ($W_{BF}$) is based on a Bayesian filter, a form of naive Bayesian classifier that has had some success in spam recognition (Graham, 2004). To construct the Bayesian filter, it is necessary to learn the conditional probability that a stroke is a gesture, given a specific word, $w_i$. Pr(*Gesture* | $w_i$ ) can be estimated from training data using Bayes' theorem:

$$p_i = \frac{\Pr(w_i|Gesture) \cdot \Pr(Gesture)}{\Pr(w_i)}$$

where Pr($w_i$|*Gesture*) is the conditional probability that word $w_i$ will be observed, given that a gesture stroke is observed; Pr(*Gesture*) is the prior probability of observing a gesture; and Pr($w_i$) is the prior probability of observing word $w_i$.

Participants in the study used a varied vocabulary to describe the same objects and gestures. If the Bayesian filter encounters a word that was not in the training corpus, it is unable to produce a probability. The thesaurus Bayesian filter feature ($W_{TBF}$) provides a remedy for this situation. It is computed much like $W_{BF}$, except that a thesaurus is used to generalize the training data. One strength of these two features is that they learn which words are most likely to coincide with gesture or object strokes. However, these features do not consider word order. The Markovian filter feature ($M_{MF}$) is analogous to the Bayesian filter features, but considers word sequences rather than individual words.

### 3.2. Results: Gesture/object classification accuracy

Bischel et al. (2009) used a form of holdout validation to evaluate the accuracy of the gesture classifier. The holdout set comprised 39 randomly selected sketches for training and 10 for testing. A conventional beam search approach (Aha & Bankert, 1994; Dash & Liu, 1997; Gupta et al., 2002) was used to determine which sets of features are the most effective at classification. To provide additional insights about which features are the most important, this process was performed three times: once considering only sketch features, once considering only speech features, and once considering both.

The best single sketch feature classifier used $D_{TNS}$ and achieved 69.5% accuracy. The best sketch-only classifier achieved 76.2% accuracy using nine features: $D_{SL}$, $D_{DPS}$, $D_{UID}$, $D_{MHD}$, $D_{AHD}$, $D_{HL}$, $D_{ET}$, $D_{AC}$, and $D_{DC}$. The best single speech-feature classifier used $W_{BF}$ and achieved 77.7%. The best speech-only classifier achieved 78.2% accuracy using three features: $W_{BF}$, $W_{TBF}$, and $W_{WC}$. The best classifier considering all features achieved 81.9% accuracy using six features: $D_{TCS}$, $D_{MHD}$, $D_{HL}$, $D_{ET}$, $W_{BF}$, and $W_{TPS}$.

In Bischel et al. (2009), this process was actually performed for four holdout sets. For all sets, the results were similar: the single best feature in all cases was either $W_{BF}$ or $W_{TBF}$. Similarly, for multifeature classifiers employing speech features, the best feature sets always contained at least one of these two features.

## 4. EVALUATION OF 3-SEC TECHNIQUE

As the results in the previous section demonstrate, the speech modality plays an important role in identifying gestures. For example, the two Bayesian filter features were the most important single features for classifying pen strokes as gesture or object strokes. The importance of speech suggests the need to examine the validity of the speech–sketch alignment technique that serves as the foundation for the speech features.

As the name suggests, the 3-sec alignment approach uses only temporal correlation to align the speech and sketch input. It is possible for this technique to associate speech with a stroke that is not logically related to it. To evaluate the performance of the 3-sec technique, we manually aligned the speech and pen strokes based on semantic information. We then compared the resulting alignment with that produced by the 3-sec technique. We also used the manually aligned speech to compute the speech features for our gesture classifier to determine if more accurate alignment would improve classification accuracy. The latter results are presented in Section 6.

### 4.1. Manual alignment

We manually aligned the speech and sketch modalities using a two-step approach. We first segmented the speech into small, meaningful statements. We then aligned each statement with the pen strokes, if any, to which it referred. The

https://doi.org/10.1017/S0890060411000060 Published online by Cambridge University Press

segmentation step proved to be difficult because the speech was terribly ungrammatical and disfluent, as is commonly the case in multimodal descriptions (Adler & Davis, 2007). Because of the nature of the speech, we could not use a simple segmentation strategy, such as decomposing the speech into grammatically correct clauses. Oviatt et al. (1997) suggest that in multimodal interactions, spoken phrases often follow a subject–verb–object pattern. We used this as the starting point for developing our manual segmentation approach. Our approach is similar to the Simple Metadata Annotation Specification (Strassel, 2004), but considers information from the sketch input.

Our manual segmentation comprises single "clauses" consisting of a subject, verb, and object; multiple logically related, sequential clauses; partial clauses; and filled pauses such as "uh" and "um." Note that filled pauses are identified purely on lexical grounds and are not prosodic features of the speech. Whenever possible, we segmented the speech into subject–verb–object "clauses." However, if the speaker moved on to a new thought before completing a clause, we segmented the incomplete thought into a partial clause. Likewise, a change in speaker before the completion of a clause also resulted in a partial clause.

Filled pauses could either comprise an entire segment or be included in a larger clause, depending on the circumstances. If the filled pause was in the middle of a set of words that otherwise formed a clause, the pause was grouped with that set of words. For example, "this handle uh moves here" is considered a single clause. Likewise, if the filled pause occurred immediately before the start of a clause, it was grouped with it. For example, "uh this handle moves here" would be segmented as a single clause if there were little delay between "uh" and "this." In all other cases, filled pauses were considered to form their own segments.

There were two occasional exceptions to our segmentation strategy. Two or more clauses were joined if they referred to the same pen stroke. We did this so that each stroke would be aligned with at most one speech segment. In addition, if a clause had multiple objects referring to different strokes, the objects were split into separate clauses. These two cases are the primary differences between our segmentation approach and that in Strassel (2004).

Figure 4 shows an example of the manual segmentation results. The first segment consists of the filled pause "uh." This pause was not combined with the subsequent clause because the time gap was too large. "it's kind of for cutting stuff" is a typical clause with subject "it," verb "is," and object "for cutting stuff," "when they uh attach" is also considered to be a clause with subject "they" and intransitive verb "attach." In addition, the filled pause "uh" is included in the segment because it occurs inside an otherwise valid clause. "Uh huh" is a segment consisting of two filled pauses in close succession. The phrase "and so both sides move the" is a partial clause; the speaker changed thoughts before completing it. The word "these" is again a partial clause representing a new idea. Finally, the phrase "this moves" is a clause with a subject and verb, but no object.

As this example illustrates, manually segmenting the speech required considerable judgment. The task was performed by two researchers. Each segmented one-half of the speech and then verified the segmentation accuracy of the other half. Once the segmentation was completed, the two researchers then manually aligned the segments and pen strokes. Each stroke was aligned with at most one speech segment. However, a speech segment could be aligned with multiple strokes. As with the segmentation, the researchers divided the task and verified each other's work.

An alternative approach for annotating our data would have been for each researcher to annotate the entire corpus individually and then arbitrate the single, final annotation. This approach can lead to a more consistent annotation of the corpus than the cross-validation approach we used (Artstein & Poesio, 2005). We opted for our approach in the interest of expediency, and note that any inconsistencies between the two halves of the annotation will only hamper the performance of our statistical classifier.

Table 2 tabulates the results of the manual segmentation and alignment process. The data from the study contained 34,354 words forming 7454 speech segments. We found that 78.8% of the 6470 pen strokes were aligned with speech

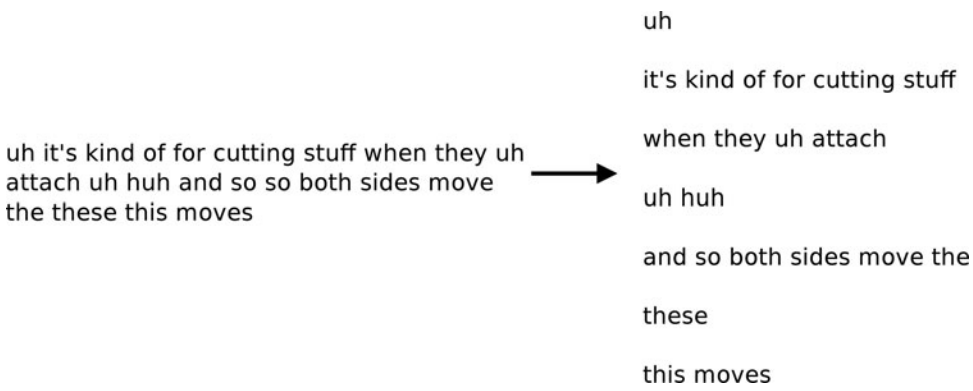


**Fig. 4.** Examples of manual segmentation: (left) raw speech and (right) segmented speech.

https://doi.org/10.1017/S0890060411000060 Published online by Cambridge University Press

**Table 2.** *Properties of manually aligned speech from user study*

| Attribute | Count |
|---|---|
| Words | 34,354 |
| Segments | 7,454 |
| Strokes | 6,470 |
| Words/segment, ave. | 4.6 |
| Words/associated segment, ave. | 8.3 |
| Segments associated with strokes | 22.5% |
| Strokes associated with segments | 78.8% |

segments, but only 22.5% of the segments were aligned with strokes. On average, there were 4.6 words per speech segment, but for segments aligned with strokes there was a much higher average of 8.3 words per segment.

### 4.2. Alignment accuracy of 3-sec technique

We evaluated the accuracy of the 3-sec alignment technique by direct comparison with the manual alignment, which constitutes the correct result. Specifically, we compared the set of words associated with each pen stroke in the two cases. Note that the 3-sec technique does not have an explicit segmentation step. Rather, any words that fall at least partially within the 3-sec temporal window of a stroke are associated with it. Thus, there is no notion of segmentation accuracy, and it is possible to evaluate accuracy only for those words that are associated with a pen stroke.

To illustrate the analysis, consider the speech and the accompanying gesture in Figure 5. The 3-sec approach has associated with this stroke the words "faces the other way this is uh like a." The correct association determined by the manual alignment process is "this is uh like a handle." In this case the 3-sec association begins too early and does not extend long enough. This situation occurred on average for 6% of the strokes (the average is computed over the 48 sketches).

There are a total of 14 possible relative arrangements of the 3-sec association and the correct (manual) association as shown in Figure 6. Each cell in the figure represents one of the possible arrangements. For example, cell 4 represents the arrangement from Figure 5. For clarity, the stroke itself is not represented in the various cells in Figure 6.

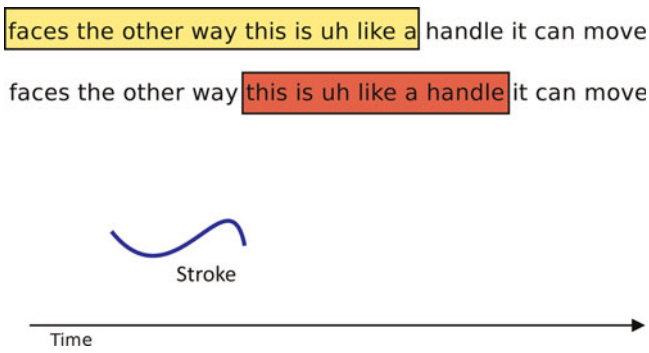


**Fig. 5.** A gesture pen stroke and the speech associated with it by the 3-sec technique (top bar) and the manual alignment process (bottom bar). [A color version of this figure can be viewed online at journals.cambridge.org/aie]

Cases 1 through 11 in Figure 6 are all cases in which the 3-sec approach associates words with strokes that should have associated words. Case 1 is when the 3-sec association exactly matches the correct result. This occurred on average for 2% of the strokes. Cases 2 through 9 are overlapping associations that do not perfectly match. These cases represent on average 56% of the strokes. Cases 10 and 11 are cases in which there is no overlap between the 3-sec association and the correct one. These cases represent on average 10% of the strokes. Case 12 describes strokes that should have no associated speech, but the 3-sec approach has made an association. This occurred on average for 26% of the strokes. Case 13 is the converse case in which there should be associated speech, but the 3-sec approach has associated none. This occurred on average for 5% of the strokes. Finally, case 14 describes situations in which the 3-sec approach has correctly associated no speech with a stroke. This case did not occur. On average, the 3-sec approach achieved the correct answer only 2% of the time (case 1), and 41% of the time the 3-sec association was completely disjoint from the correct result (cases 10–13).

## 5. IMPROVED SPEECH–SKETCH ALIGNMENT TECHNIQUE: CBA

As Figure 6 illustrates, the 3-sec technique does not accurately align the speech and sketch modalities. Consequently, we sought to develop an improved automatic alignment technique. We modeled the new technique on our manual process: the technique employs an explicit speech segmentation process, followed by a segment–stroke alignment process. Because both processes employ statistical classifiers, we call our alignment technique CBA.

### 5.1. Speech segmentation

Our approach to automatic segmentation uses a statistical classifier to classify words according to their position in a segment. We consider four classes of words: *start*, *middle*, *end*, and *only* words. As the names suggest, *start* and *end* words represent the start and end of a clause, respectively. All words in a clause other than these are defined as *middle* words. *Only* words are segments consisting of a single word, which is typically a filled pause. Figure 7 shows an example of the word classification for a passage of speech.

The word classifications are used to directly construct the speech segmentation. First, all valid segments are formed. Specifically, each *only* word is labeled as a segment. Likewise, each sequence of words that begins with a *start* word, ends with an *end* word, and has only *middle* words (if any) in between, is labeled as a segment.

Once all valid segments have been formed, a repair process is used to segment any remaining speech. The first and last

https://doi.org/10.1017/S0890060411000060 Published online by Cambridge University Press

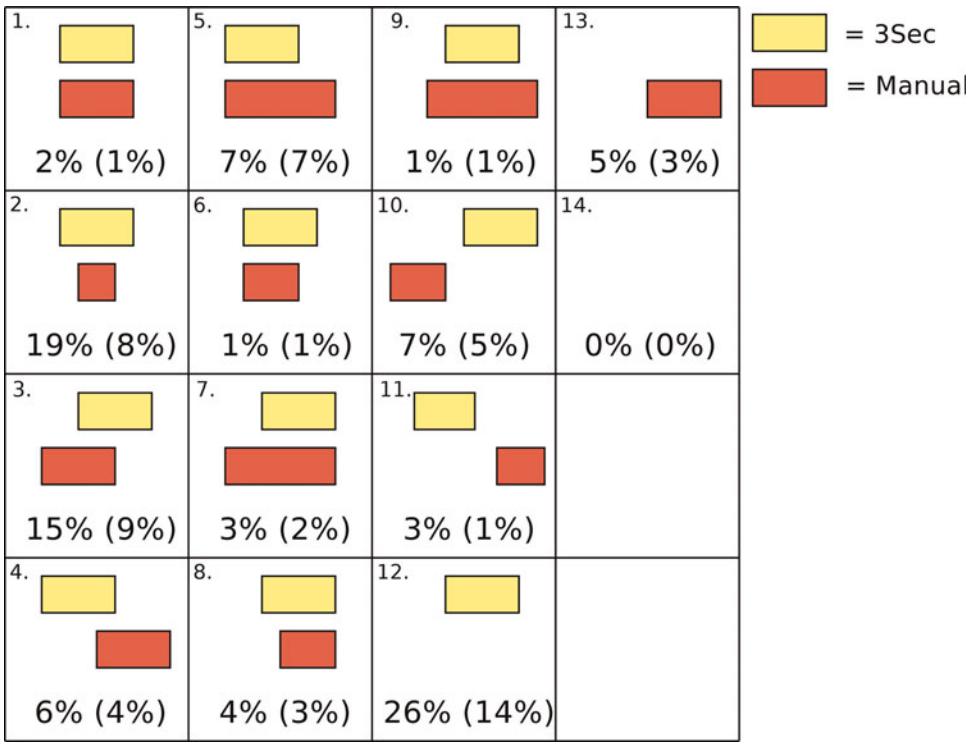


**Fig. 6.** The accuracy of the 3-sec technique. In each cell, the top bar represents the speech associated with a stroke by the 3-sec technique, and the bottom bar represents the manual (correct) association. Each cell represents a distinct relative arrangement of the associations and the frequency with which it occurs. The results are averaged over the 48 sketches. Standard deviations are included in parentheses. [A color version of this figure can be viewed online at journals.cambridge.org/aie]

words of the speech are always considered *start* and *end* words, respectively. Any unsegmented word immediately after a segment is treated as a *start* word, whereas any unsegmented word immediately before a segment is treated as an *end* word. A single word directly between two valid segments is considered an *only* word. After updating the word classifications in this fashion, any new valid segments are formed. The repair process is then repeated until all words have been segmented.

Consider a passage of speech that has been classified as: *start*, *middle*, *end*, *middle*, *middle*, *end*. In the initial segmentation pass, the first three words will be formed into a valid segment. Then, during the repair pass, the fourth word will be treated as a *start* word so that the last three words form a segment.

Our segmentation approach is based on four word classes. Many speech segmentation approaches such as (Stolcke & Shriberg, 1996; Stolcke et al., 1998; Gotoh & Renals, 2000; Hwan Kim & Woodland, 2001) classify interword

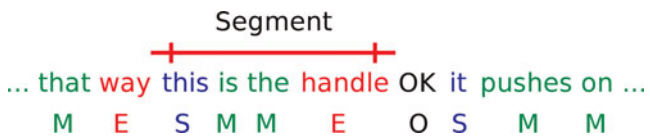


**Fig. 7.** The classification of the words in a spoken passage; S, start; M, middle; E, end; and O, only. [A color version of this figure can be viewed online at journals.cambridge.org/aie]

boundaries as segment events or nonsegment events. These approaches were developed for unimodal dialog such as the SWITCHBOARD corpus (Godfrey et al., 1992). We consider multimodal dialog in which the speech is highly disfluent and filled pauses are common. We designed our four-class approach to take advantage of the discriminatory power of single-word segments. This approach is also consistent with work in Hoffmann et al. (2001) and Luo (2008), suggesting that for some classification problems, decomposing a class into subclasses can result in higher accuracy.

### 5.1.1. *Segmentation classifier and features*

Our segmentation classifier is an Ada-boosted C4.5 decision tree computed with WEKA (Hall et al., 2009). Each word is characterized by 25 features listed in Table 3. (The classifier considers the features of the word in question, as well as those of the word on either side.) The simplest feature is the word itself ($W_s$). Each word is also characterized by the parts of speech that it could possibly have in legal English usage, which is queried from the dictionary in the Stanford part of speech tagger (Toutanova & Manning, 2000). The rationale for these features is that different parts of speech may be more likely to occur in particular locations within a speech segment. For example, a verb is unlikely to be the first word in a segment. We define nine Boolean part of speech features indicating if the word could be a coordinating conjunction ($W_{CCN}$), determiner ($W_{DET}$), preposition ($W_{PRP}$), adjective

https://doi.org/10.1017/S0890060411000060 Published online by Cambridge University Press

**Table 3.** *Features for speech segmenter*

| Name | Description | Units |
|---|---|---|
| $W_{TNW}$ | Time to next word | ms |
| $W_{TPW}$ | Time to previous word | ms |
| $W_{TNR}$ | Time to next ratio | DQ |
| $W_{TPR}$ | Time to previous ratio | DQ |
| $W_S$ | The word | Text |
| $W_{CCN}$ | Coordinating conjunction | Boolean |
| $W_{DET}$ | Determiner | Boolean |
| $W_{PRP}$ | Preposition | Boolean |
| $W_{ADJ}$ | Adjective | Boolean |
| $W_{PP}$ | Personal pronoun | Boolean |
| $W_{ADV}$ | Adverb | Boolean |
| $W_{VRB}$ | Verb | Boolean |
| $W_{WHD}$ | Wh-determiner | Boolean |
| $W_{WHA}$ | Wh-adverb | Boolean |
| $W_{ACN}$ | Author change next | Boolean |
| $W_{ACP}$ | Author change previous | Boolean |
| $D_{TNS}$ | Time to next stroke | ms |
| $D_{TPS}$ | Time to previous stroke | ms |
| $D_{TNR}$ | Time to next ratio | DQ |
| $D_{TPR}$ | Time to previous ratio | DQ |
| $D_{ACN}$ | Author change next | Boolean |
| $D_{ACP}$ | Author change previous | Boolean |
| $D_{SL}$ | Stroke length | Pixel |
| $D_{SED}$ | Start to end distance | Pixel |
| $D_{DUR}$ | Stroke duration | ms |

*Note:* $D_x$, sketch (drawing) feature; $W_x$, speech (word) feature; DQ, dimensionless quantity.

($W_{ADJ}$), personal pronoun ($W_{PP}$), adverb ($W_{ADV}$), verb ($W_{VRB}$), wh-determiner ($W_{WHD}$), or wh-adverb ($W_{WHA}$). Wh-determiners are the words "what" and "which" used as determiners. Wh-adverbs are the words "how," "when," "whence," "where," and "why" used as adverbs. Note that we use the possible parts of speech, rather than the actual part of speech, because the latter is difficult to determine because the speech is highly ungrammatical and the sentence boundaries are as yet unknown.

Four of the features compute temporal relationships between the words. Here, $W_{TNW}$ and $W_{TPW}$ are the time to the next and previous words, respectively. To obtain a measure of the relative size of the time gap after a word, we compute the ratio of $W_{TNW}$ to the sum of the values of $W_{TNW}$ for the word and its two successors. We call this feature $W_{TNR}$; $W_{TPR}$ is an analogous feature that concerns the relative size of the gap before the word. A time gap that is large compared to the neighboring gaps (i.e., a large ratio) could indicate a segment boundary.

A change in speaker usually corresponds to a new segment. Thus, two features track changes in the "author" of the speech: $W_{ACN}$ is a Boolean feature that is true only when there is an author change immediately after the word; similarly, $W_{ACP}$ is true only when there is an author change immediately before the word.

A novel property of our segmentation technique is that we use information from the sketch modality. Specifically, we compute properties of the pen stroke drawn closest in time to the word. We refer to this as the "coincident stroke," although the word and stroke may not actually overlap in time. We characterize this stroke with three intrinsic properties: its arc length ($D_{SL}$), start to end distance ($D_{SED}$), and duration ($D_{DUR}$). The first two of these features are the same as those used with the gesture/object classifier described in Section 3.1.

Four features describe the temporal relationships between the coincident stroke and the other strokes. These features are analogous to those used to describe the temporal relationships between the words: $D_{TNS}$ and $D_{TPS}$ are the time to the next and previous strokes, respectively. Again, to obtain a measure of the relative size of the time gap after the coincident stroke, we compute the ratio of $D_{TNS}$ to the sum of the values of $D_{TNS}$ for the stroke and its two successors. We call this feature $D_{TNR}$; $D_{TPR}$ is an analogous feature that concerns the relative size of the time gap before the stroke.

The final two features track changes in the "author" of the pen strokes. Here, $D_{ACN}$ is a Boolean feature that is true only when there is an author change immediately after the coincident stroke, and $D_{ACP}$ is true only when there is an author change immediately before it.

### *5.1.2. Segmentation accuracy*

We performed leave one out cross-validation to evaluate our speech segmenter.[2] In each iteration of the cross-validation, the data from all but one sketch was used to train our classifier. We then used the trained classifier to predict the segment boundaries for the remaining sketch. The technique achieved an average accuracy of 92.7% at classifying words as *start*, *middle*, *end*, and *only* words.

To provide a more informative measure of accuracy, we directly compared our "classifier-based segmentation" with the manual segmentation. Specifically, we computed the fraction of the classifier-based segments that matched the manual segments within a tolerance ranging from zero to three words. The results are shown in Figure 8. On average, about 33.9% of the classifier-based segments exactly matched a manual segment, and about 75.9% matched within three words. In the latter case, the errors could be distributed on both ends of the segment as long as the total number of errors did not exceed three. For example, compared to the manual segment, the classifier-based segmentation could be missing one word at the beginning, and have two extra words at the end, or vice versa.

## 5.2. Stroke–speech alignment

Once the speech has been segmented, the next step is to align the segments with the pen strokes. We do this with a two-step process. First, segments are aligned with strokes based on simple temporal correlation. Second, we use a classifier to detect and repair two common alignment errors. The initial alignment borrows from the 3-sec approach. Each stroke is associated with the segment that has the greatest overlap with

[2] Cross-validation is a process of partitioning a data set into complementary training and testing sets. Multiple alternative partitions are considered, and the results from them are averaged.

https://doi.org/10.1017/S0890060411000060 Published online by Cambridge University Press

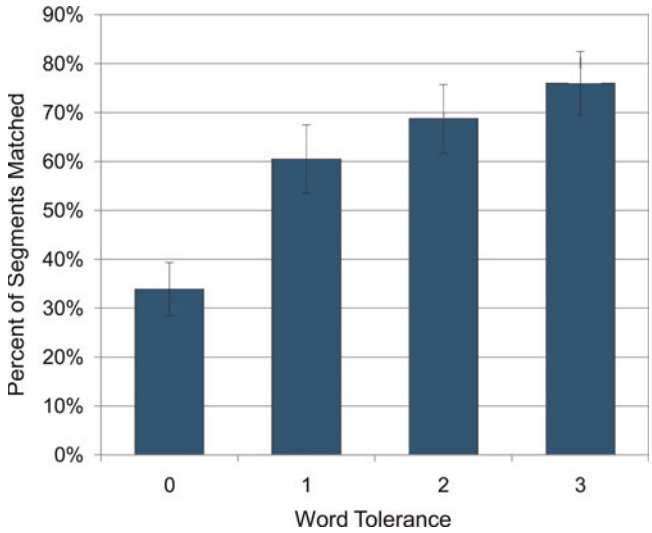


**Fig. 8.** Percentage of classifier-based segments matching (correct) manual segments within a tolerance. The results are averaged over the 48 sketches. [A color version of this figure can be viewed online at journals.cambridge.org/aie]

the stroke's 3-sec temporal window, that is, a window that extends 3 s before and after the stroke.

Using an analysis similar to that described in Figures 5 and 6, we computed the accuracy of the initial alignment to determine what improvements are necessary. The results are illustrated in Figure 9. The two most frequent problems are case 10 in which the initial association follows the correct association, and case 12 in which there is an association when there should be none. Case 10 occurs on average for 18% of pen strokes, whereas case 12 occurs for 26%.

Because of the prevalence of these two cases, we developed classifiers to detect them. The two classifiers are applied to each initial association. If a case 10 error is detected, the

**Table 4.** *Features used for segment–stroke alignment*

| Name | Description | Units |
|---|---|---|
| $D_{SL}$ | Arc length | Pixel |
| $D_{TNS}$ | Time to next stroke | ms |
| $D_{DUR}$ | Stroke duration | ms |
| $S_{DUR}$ | Segment duration | ms |
| $S_{NV}$ | Contains noun or verb | Boolean |
| $S_{WC}$ | Word count | Word |
| $S_{SC}$ | Stroke count | Stroke |
| $S_{SE}$ | Stroke start end duration | ms |
| $S_{SS}$ | Stroke start start duration | ms |
| $S_{TN}$ | Time to next segment | ms |
| $S_{TP}$ | Time to previous segment | ms |

*Note:* $D_x$, sketch (drawing) feature; $S_x$, feature. To train the case 10 classifier, all of the initial associations in the training set are labeled with a binary value indicating whether they are a case 10 error. An analogous approach is used to train the case 12 classifier.

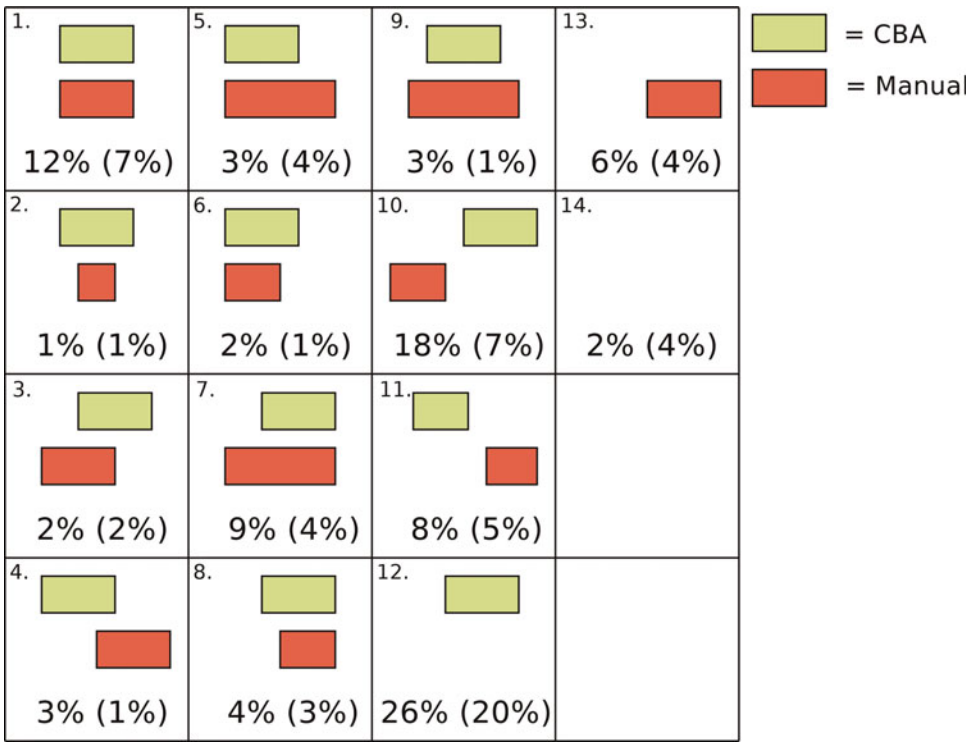


**Fig. 9.** The alignment accuracy after the first step of classifier-based alignment (CBA), that is, *before the final processing step*. In each cell, the top bar represents the speech associated with a stroke by the first step of CBA, and the bottom bar represents the manual (correct) association. Each cell represents a distinct relative arrangement of the associations and the frequency with which it occurs. The results are averaged over the 48 sketches. Standard deviations are included in parentheses. [A color version of this figure can be viewed online at journals.cambridge.org/aie]

https://doi.org/10.1017/S0890060411000060 Published online by Cambridge University Press

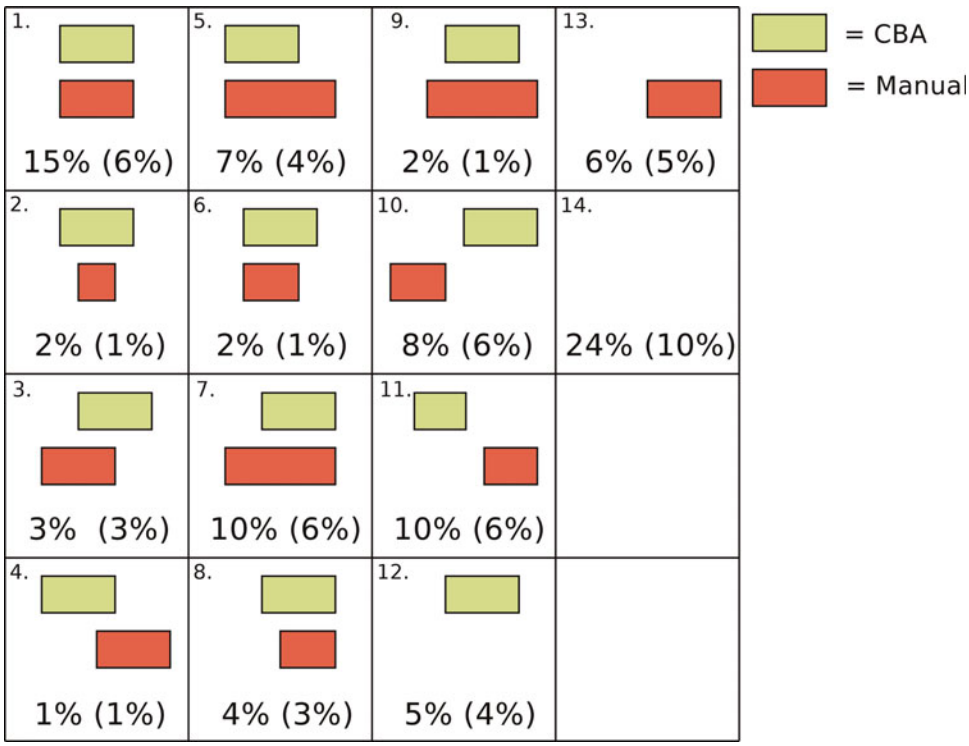


**Fig. 10.** The accuracy of the classifier-based alignment (CBA) technique. In each cell, the top bar represents the speech associated with a stroke by CBA, and the bottom bar represents the manual (correct) association. Each cell represents a distinct relative arrangement of the associations and the frequency with which it occurs. The results are averaged over the 48 sketches. Standard deviations are provided in parentheses. [A color version of this figure can be viewed online at journals.cambridge.org/aie]

association of the pen stroke is changed to the next earlier segment. If a case 12 error is detected, the association for the stroke is removed. In this fashion, the classifiers enable an efficient approach to improving the initial alignment.

The "case 10" and "case 12" classifiers are Ada-boosted C4.5 decision trees computed with WEKA (Hall et al., 2009). They consider features of both the speech segment and the initially associated pen stroke. There are a total of 11 features that are listed in Table 4.

The number of words in the segment is $S_{WC}$, and $S_{DUR}$ is its duration; $S_{SC}$ is the number of strokes associated with the segment, and the segment may also initially be associated with other strokes. A Boolean feature indicating if any of the words in the segment were tagged as a noun or verb by the Stanford part of speech tagger is $S_{NV}$ (Toutanova & Manning, 2000).[3] The intuition is that segments containing no nouns or verbs are generally uninformative and are unlikely to refer to a stroke. The initially associated pen stroke is characterized by its arc length ($D_{SL}$), duration ($D_{DUR}$), and the time to the next stroke ($D_{TNS}$).

Four other features describe temporal relationships: $S_{TN}$ is the time to the next segment, $S_{TP}$ is the time to the previous one, $S_{SE}$ is the time between the start of the segment and the end of the associated stroke, and $S_{SS}$ is the time from the start of the segment to the start of the associated stroke. Both of these features can have positive or negative values.

[3]At this point in the computation, the speech has been segmented into phrases, thus enabling the part of speech tagger to determine the actual part of speech of each word.

### 5.2.1. *Stroke–speech alignment accuracy*

To evaluate the performance of our two-step segment–stroke alignment technique, we again performed a leave one out cross-validation. In each iteration of the cross-validation, one sketch with speech was used for testing, whereas the others were used for training. We averaged the results across the 48 testing/training combinations.

Figure 10 compares the final alignment to the correct (manual) alignment. The case 10 and case 12 classifiers were clearly effective. The case 10 errors have been reduced from an average of 18% in Figure 9 to an average of only 8%. Likewise, the case 12 errors have been reduced from an average of 26% to an average of only 5%. Overall, after the second step of alignment, an average of 39% of the associations are perfect (cases 1 and 14). Furthermore, on average, only 29% of the associations are completely disjoint from the correct associations (cases 10–13).

To provide a more detailed evaluation of the alignment accuracy, we also computed the number of missing and extra words in each association. Extra words are those associated with the pen stroke that should not have been. Conversely, missing words are those that should have been associated

https://doi.org/10.1017/S0890060411000060 Published online by Cambridge University Press

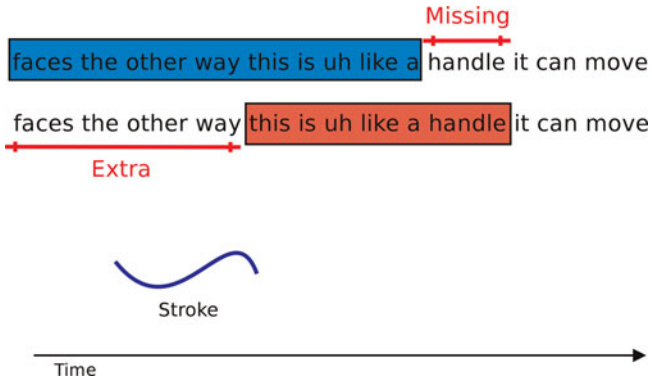


**Fig. 11.** Missing and extra words in speech aligned with a pen stroke. The top bar indicates the words actually associated with the pen stroke, and the bottom bar indicates the words that should have been associated. [A color version of this figure can be viewed online at journals.cambridge.org/aie]

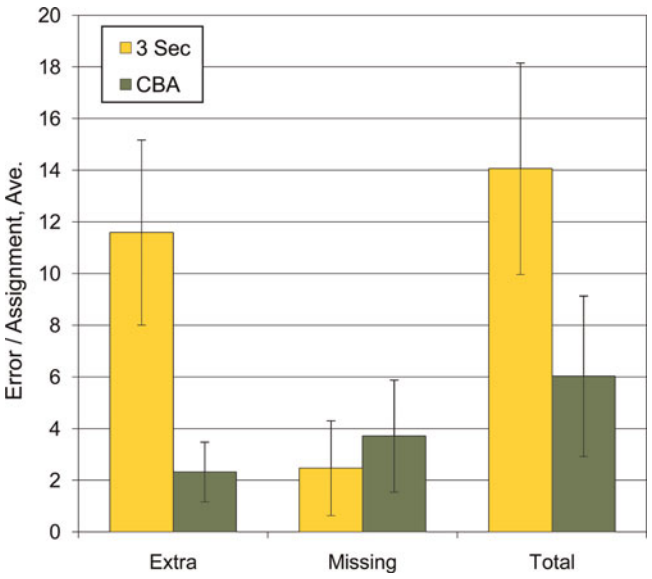


**Fig. 12.** The average number of words incorrectly aligned with each stroke for the 3-s and classifier-based alignment techniques. Averages are computed over the 48 sketches. [A color version of this figure can be viewed online at journals.cambridge.org/aie]

but were not. Consider the hypothetical example[4] in Figure 11. The stroke is associated with the words "faces the other way this is uh like a." The correct association (as determined by manual alignment) is the clause "this is uh like a handle." In this case, the words "faces the other way" are extra words, and "handle" is a missing word.

Figure 12 presents the missing/extra accuracy of both the 3-sec and CBA techniques. On average, the 3-sec approach has about 12 extra words and 2 missing words per stroke, while our classifier-based approach has only about 2 extra and 4 missing. Overall, the 3-sec approach has an average of 14 incorrect (missing plus extra) words per stroke, whereas our new approach has only 6. This is a 57% reduction in errors.

## 6. GESTURE/OBJECT CLASSIFICATION ACCURACY

Our purpose in creating an improved technique for speech–sketch alignment is to enable more accurate identification of gesture pen strokes like those in Figure 1b. Thus, to evaluate our CBA technique, we computed the gesture classification accuracy using our technique and compared this to the accuracy achieved with the 3-sec alignment technique. We also computed the accuracy using manual alignment to obtain an upper bound on the achievable gesture classification accuracy.

For this analysis, we used the gesture classifier just as described in Section 3.1, except that we used an Ada-boosted C4.5 decision tree computed with WEKA (Hall et al., 2009) rather than using a neural network. The sketch features were computed as before, whereas the speech features were computed using the speech–sketch alignment technique in question.

We computed accuracy via leave one out cross-validation, with one sketch used for testing and the others used for training. Our results are the average across the 48 testing/training combinations. We evaluated classification accuracy for four sets of features: the thesaurus Bayesian filter feature ($W_{TBF}$), the five most important features, the 10 most important features, and all features. We determined the top 5 and top 10 features using an information gain algorithm (Dash & Liu, 1997; Xing et al., 2001) as implemented by WEKA.[5] The top 5 features include the 2 Bayesian filter features ($W_{TBF}$, $W_{BF}$), the total elapsed time ($D_{ET}$), the time to the closest prior stroke ($D_{TCS}$), and the time to the next stroke ($D_{TNS}$). The top 10 features in addition include the time to the previous stroke ($D_{TPS}$), the distance to previous stroke ($D_{DPS}$), the distance to the next stroke ($D_{DNS}$), the maximum Hausdorff distance to the underlying ink ($D_{MHD}$), and the average Hausdorff distance to the underlying ink ($D_{AHD}$).

The gesture/object classification results are shown in Figure 13. [The accuracy in Fig. 13 differs from that in Bischel et al. (2009) because different classifiers were used, i.e., a neural network vs. Ada-boosted decision tree.] Typically, for a given set of features, the CBA resulted in better accuracy than the 3-sec alignment, and the manual alignment resulted in the best accuracy. Likewise, using more features typically resulted in better accuracy. There was one exception. The 3-sec approach achieved nearly its best accuracy when only the thesaurus Bayesian filter feature was used. For this single-feature case, the 3-sec approach actually achieved better accuracy than even the manual alignment. This is discussed in the next section.

[4] The speech is taken directly from the user study data. The hypothetical stroke was designed to monotonically increase along the horizontal axis, thus suggesting a drawing process evolving in time.

[5] As discussed in Section 3.2, the work in Bischel et al. (2009) employed a beam search approach to determine the best features. Although that approach may be more reliable, here we use information gain in the interest of expediency.

https://doi.org/10.1017/S0890060411000060 Published online by Cambridge University Press

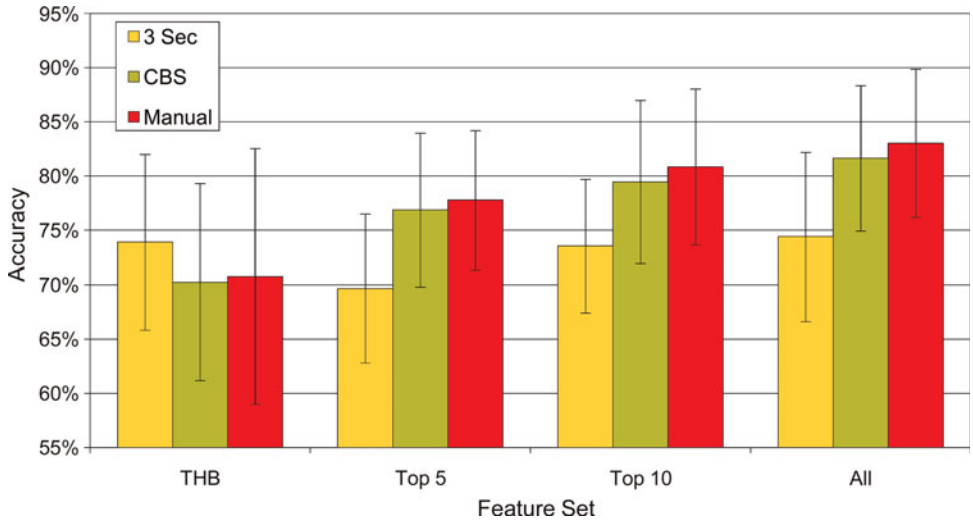


**Fig. 13.** The gesture/object classification accuracy versus speech–sketch alignment technique and number of features; THB, thesaurus Bayesian feature. The results are averaged over the 48 sketches. [A color version of this figure can be viewed online at journals.cambridge.org/aie]

## 7. DISCUSSION

Our speech segmenter achieved an average accuracy of 92.7% in classifying words as *start*, *middle*, *end*, and *only* words. Although the classification accuracy is high, a more important measure of accuracy is the segmentation accuracy: on average, 75.9% of the computed segments matched correct (manual) segments within a three-word tolerance.

Liu et al. (2005) define a per boundary based metric for speech segmentation accuracy. This is defined as the sum of the false positive and false negative sentence boundaries normalized by the total number of interword boundaries. With their state of the art technique based on a conditional random field model, they achieve a boundary-based accuracy of 95.4% on conversational telephone speech. (They achieve higher accuracy on broadcast news thast is more grammatical than telephone conversation.) We achieve 86.8% accuracy on multimodal dialog that includes both speech and sketching. Our results compare favorably with theirs for several reasons. Sentence boundaries occur less frequently in their data than in ours: in their data set only 15.7% of interword boundaries are actually sentence boundaries, whereas in our data set 21.8% are. Thus, a naive classifier would perform better on their data than on ours. In addition, their accuracy is likely to benefit from a greater amount of training data: they trained on 480,000 words, whereas we trained on about 33,000. Finally, we consider different kinds of speech: theirs is unimodal whereas ours is multimodal.

Our speech segmentation approach is unique in that it demonstrates the usefulness of sketch features for locating segment boundaries in multimodal dialog. In addition, we use a single classifier to directly combine word-based, prosodic (pause between words),[6] and sketch-based features. Unlike many existing approaches (Stolcke & Shriberg, 1996; Stolcke et al., 1998; Gotoh & Renals, 2000; Hwan Kim & Woodland, 2001), we use a four-class (*start*, *middle*, *end*, and *only*) approach to locating segment boundaries. This approach was designed to take advantage of the discriminatory power of single-word segments. In future work, we plan to compare the performance of this approach to that of a more traditional approach in which interword boundaries are classified as either segment boundaries or nonboundaries. In addition, unlike traditional approaches, we do not explicitly consider word sequence: we have no n-gram language model. When processing a given word, our classifier does consider the previous and next words, but we do not use a Markovian approach. In future work, we plan to combine our technique with an explicit language model, but this will likely require a much larger data set. For example, to provide a benchmark for our results, we implemented the technique in Stolcke and Shriberg (1996) using a trigram language model. This approach performed poorly on our data: of the hypothesized segment boundaries, on average only 1.3% were true boundaries, whereas for our approach 79% were. It is likely that our corpus containing only about 34,000 words is too small to train the trigram model.

Our classification-based speech–sketch alignment technique performed significantly better than the 3-sec technique as indicated by multiple measures. On average, CBA aligned only 6 incorrect words (missing plus extra) per pen stroke, whereas the 3-sec approach had 14. Comparison of Figures 6 and 10 further illustrates the superiority of the CBA technique. For example, on average, CBA perfectly aligned the speech (cases 1 and 14) for 39% of pen strokes, whereas the 3-sec approach did this for only 2% of strokes. Similarly, for CBA an average of only 29% of the associations were completely disjoint from the correct associations (cases 10–13), whereas for the 3-sec approach 41% were. Likewise, for CBA an average of only 31% of the associations were par-

[6] The filled pauses discussed in Section 4.1 are not prosodic features but instead are a lexical concept. The time elapsed between words is the only prosodic feature we use.

https://doi.org/10.1017/S0890060411000060 Published online by Cambridge University Press

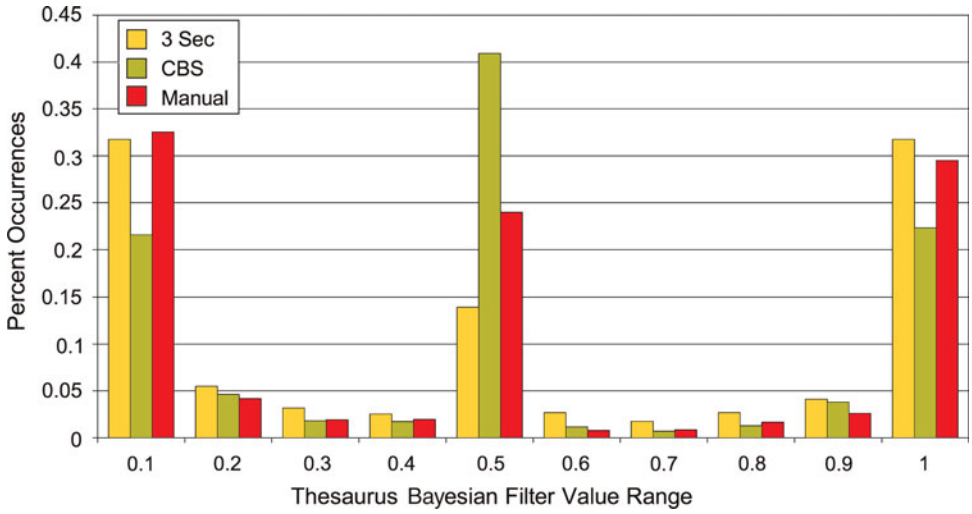


**Fig. 14.** Histograms of the values of the thesaurus Bayesian filter feature for different alignment techniques. [A color version of this figure can be viewed online at journals.cambridge.org/aie]

tially disjoint (cases 3–9), whereas for the 3-sec approach 56% were. Note that for some particular completely or partially disjoint cases, the 3-sec approach did have fewer errors than CBA. However, on the whole, CBA had far fewer completely and partially disjoint cases, and thus overall is significantly more accurate than the 3-sec approach.

Leaving aside the case of the single-feature classifier, the results in Figure 13 support our hypothesis that better speech–sketch alignment leads to better accuracy for classifying pen strokes as gestures or object strokes. Our classification-based alignment technique resulted in much greater accuracy than the 3-sec approach, and performed nearly as well as the manual alignment.

The single-feature case, however, is an interesting anomaly. To understand why the 3-sec alignment technique outperformed even the manual alignment when the classifier used only the thesaurus Bayesian filter feature, we examined the distribution of the values of this feature for the three alignment methods as shown in Figure 14. Comparatively speaking, the 3-sec alignment results in a bimodal distribution in which each stroke is either a gesture (feature value of 1) or not (feature value of 0). The other two methods, by contrast, have a greater percentage of cases with a probability of 0.5, which indicates that a stroke is equally likely to be a gesture or object stroke. Thus, with more accurate alignment, the thesaurus Bayesian filter is able to make predictions about fewer strokes.

Examining Figure 12 gives some additional insight into this anomaly. The 3-sec alignment technique tends to align many extra words with each pen stroke. These extra words may allow the thesaurus Bayesian filter to make predictions for strokes that do not actually have associated speech. For strokes that do have associated speech, we would expect that better alignment would result in better classification accuracy. To test this hypothesis, we evaluated gesture/object classification accuracy for only those strokes with associated speech as determined by the manual alignment. Here again, we computed accuracy via leave one out cross-validation, with one sketch used for testing and the others used for training. However, in this case only strokes with associated speech were included in the testing and training sets.

The results are shown in Figure 15. For strokes with associated speech, improved alignment does result in improved accuracy, even when only the thesaurus Bayesian filter is used. It appears that the overassociation of words by the 3-sec approach is useful when only speech is used for gesture/object classification. However, the benefit is quickly lost as additional features are used. Apparently, the noise introduced by over-alignment degrades the performance of the other features.

Currently, our system is designed to be applied once the device description has been completed. An important next step will be to adapt our system to work in real time so that strokes are classified as they are drawn. All of the features used for the various classifiers can be computed on the fly as they depend

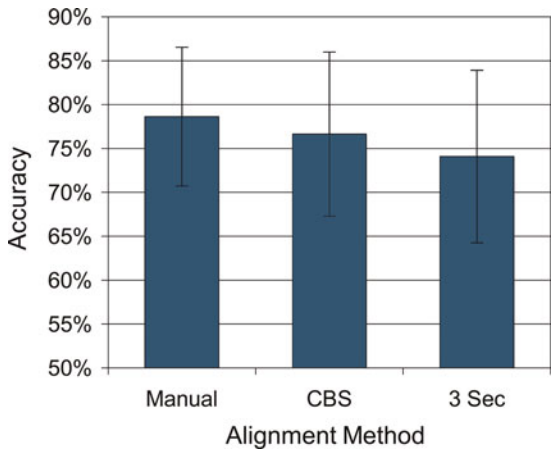


**Fig. 15.** The gesture/object classification accuracy for strokes known to have associated speech. Classification is based on only the thesaurus Bayesian filter feature. [A color version of this figure can be viewed online at journals.cambridge.org/aie]

https://doi.org/10.1017/S0890060411000060 Published online by Cambridge University Press

only on prior information. Thus, the primary challenge in creating a real-time system will be the problem of automatic speech recognition. The state of the art Sphinx-4 speech recognition system (Walker et al., 2004) achieves a word error rate of 7% with a vocabulary of 5000 words and a word error rate of 19% with a vocabulary of 60,000 words. The errors inherent in automatic speech recognition will clearly present challenges. However, we may be able to compensate for these errors by using additional prosodic features (we currently use only pause duration).

We evaluated our techniques using a nearly user-independent approach. The training data used when testing on a particular sketch was comprised of 44 sketches by other authors, and only three sketches from the primary author of the test sketch. It is likely that increasing the amount of user-specific training data will increase the accuracy of the system. Such training data has proven beneficial for other recognition tasks, such as hand-drawn symbol recognition (Field et al., 2010).

We have developed our techniques within the domain of collaborative engineering design, but they should generalize to many other domains. None of the features used by our classifiers are specific to mechanical devices or the task of designing; thus, we believe our techniques should be suitable for any domain in which the task involves drawing a sketch or diagram and explaining its elements. Examples of such domains include giving driving directions, explaining the solution to a problem in a physics lecture, and explaining a sports play.

## 8. CONCLUSION

We have presented a new technique for aligning speech and sketch input in multimodal dialog. It is designed for use in classifying pen strokes as gesture and object strokes. The technique, which we call CBA, employs a two-step process: the speech is first segmented into meaningful pieces (typically clauses), then the segments are aligned with pen strokes. Our speech segmenter uses a statistical classifier to classify words according to their position in a segment. We consider four classes of words: *start*, *middle*, *end*, and *only* words. The word classifications are then used to form speech segments. The segment-stroke alignment step initially uses temporal correlation to align segments with pen strokes. Classifiers are then used to detect and correct two common alignment errors.

Our classification-based speech–sketch alignment technique performed significantly better than the existing "3-sec" alignment technique, which is based solely on temporal correlation and has no explicit segmentation step. On average, our technique perfectly aligned the speech for 39% of pen strokes, whereas the 3-sec technique did this for only 2% of strokes. Furthermore, for our technique the aligned speech had no overlap with the correct alignment, on average, for only 29% of strokes. However, for the 3-sec technique there was no overlap for 41% of strokes. Finally, our technique had on average only six incorrectly aligned words (missing plus extra) per pen stroke, whereas the 3-sec approach had 14.

Our alignment technique is novel in that it uses information from the sketch modality for both the speech segmentation and alignment steps. Our results demonstrate that features from the sketch input are valuable for segmenting speech.

Our purpose in developing an effective speech–sketch alignment technique was to enable accurate identification of gesture pen strokes in multimodal dialog. Our gesture classifier uses features of the pen strokes and the speech aligned with them. Experiments with this classifier demonstrated that, when multiple speech and sketch features are used for classification, better alignment accuracy does lead to more accurate gesture classification. More precisely, when multiple features are used, our alignment technique resulted in much greater gesture classification accuracy than the 3-sec approach, and performed nearly as well as manual alignment. Inaccurate alignment was beneficial only when the gesture classifier used just a single statistical speech feature. In this case, the tendency of the 3-sec alignment technique to erroneously associate extra words with pen strokes allowed the gesture classifier to make predictions about pen strokes that in reality had no associated speech. Thus, in all but one unusual case, our new alignment technique enables substantially more accurate gesture classification than the prior technique.

## ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant 0729422. The authors are grateful to Matthew Ung for his assistance in segmenting and aligning the data.

## REFERENCES

Adler, A., & Davis, R. (2007). Speech and sketching: an empirical study of multimodal interaction. *SBIM '07: Proc. 4th Eurographics Workshop on Sketch-Based Interfaces and Modeling*, pp. 83–90. New York: ACM.

Aha, D.W., & Bankert, R.L. (1994). A comparative evaluation of sequential feature selection algorithms. *Proc. 5th Int. Workshop on Artificial Intelligence and Statistics*, pp. 1–7, Ft. Lauderdale, FL.

Artstein, R., & Poesio, M. (2005). Bias decreases in proportion to the number of annotators. *Proc. FG-MoL*, pp. 141–150.

Bischel, D., Stahovich, T., Peterson, E., Davis, R., & Adler, A. (2009). Combining speech and sketch to interpret unconstrained descriptions of mechanical devices. *IJCAI '09: Proc. 21st Int. Joint Conf. Artifical Intelligence*, pp. 1401–1406. San Francisco, CA: Morgan Kaufmann.

Bishop, C., Svensen, M., & Hinton, G. (2004). Distinguishing text from graphics in on-line handwritten ink. *Proc. Int. Workshop on Frontiers in Handwriting Recognition*, pp. 142–147.

Bloomenthal, K., & Zeleznik, R. (1998). SKETCH-N-MAKE: automated machining of CAD sketches. *Proc. ASME DETC '98*, pp. 1–11.

Bolt, R.A. (1980). "Put-that-there": voice and gesture at the graphics interface. *SIGGRAPH '80: Proc. 7th Annual Conf. Computer Graphics and Interactive Techniques*, pp. 262–270. New York: ACM.

Brown, D.C., Kwasny, S.C., Chandrasekaran, B., & Sondheimer, N.K. (1979). An experimental graphics system with natural language input. *Computers & Graphics 4(1)*, 13–22.

Cassell, J. (1998). A framework for gesture generation and interpretation. In *Computer Vision in Human–Machine Interaction* (Cipolla, R., & Pentland, A., Eds.), pp. 191–215. New York: Cambridge University Press.

Chai, J.Y., Hong, P., & Zhou, M.X. (2004). A probabilistic approach to reference resolution in multimodal user interfaces. *IUI '04: Proc. 9th Int. Conf. Intelligent User Interfaces*, pp. 70–77. New York: ACM Press.

Chai, J.Y., Prasov, Z., Blaim, J., & Jin, R. (2005). Linguistic theories in efficient multimodal reference resolution: an empirical investigation. *IUI '05: Proc. 10th Int. Conf. Intelligent User Interfaces*, pp. 43–50. New York: ACM.

Cohen, P.R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., & Clow, J. (1997). Quickset: multimodal interaction for distrib-

uted applications. *MULTIMEDIA '97: Proc. 5th ACM Int. Conf. Multimedia*, pp. 31–40. New York: ACM.
Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis 1*, 131–156.
Efron, D. (1941). *Gesture and Environment*. Morningside Heights, NY: King's Crown Press.
Field, M., Gordon, S., Peterson, E., Robinson, R., Stahovich, T., & Alvarado, C. (2010). Technical section: the effect of task on classification accuracy: using gesture recognition techniques in free-sketch recognition. *Computers & Graphics 34*, 499–512.
Forbus, K.D., Ferguson, R.W., & Usher, J.M. (2001). Towards a computational model of sketching. *Proc. 6th Int. Conf. Intelligent User Interfaces*, pp. 77–83. New York: ACM Press.
Godfrey, J.J., Holliman, E.C., & McDaniel, J. (1992). SWITCHBOARD: telephone speech corpus for research and development. *Proc. ICASSP*, Vol. 1, pp. 517–520.
Gotoh, Y., & Renals, S. (2000). Sentence boundary detection in broadcast speech transcripts. *Proc. ISCA Workshop: Automatic Speech Recognition: Challenges for the New Millennium ASR-2000*, pp. 228–235.
Graham, P. (2004). *Hackers and Painters, Big Ideas From the Computer Age*. New York: O'Reilly.
Gupta, P., Doermann, D., & DeMenthon, D. (2002). Beam search for feature selection in automatic svm defect classification. *Proc. Int. Conf. Pattern Recognition*, Vol. 2, p. 20212.
Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I.H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations Newsletter 11(1)*, 10–18.
Heiser, J., & Tversky, B. (2006). Arrows in comprehending and producing mechanical diagrams. *Cognitive Science 30(3)*, 581–592.
Hoffmann, A.G., Kwok, R.B.H., & Compton, P. (2001). Using subclasses to improve classification learning. *EMCL '01: Proc. 12th European Conf. Machine Learning*, pp. 203–213. London: Springer–Verlag.
Huang, X., Alleva, F., & Hon, H. (1993). The SPHINX-II speech recognition system: an overview. *Computer, Speech and Language 7*, 137–148.
Hwan Kim, J., & Woodland, P.C. (2001). The use of prosody in a combined system for punctuation generation and speech recognition. *Proc. EUROSPEECH*, pp. 2757–2760.
Johnston, M., Bangalore, S., Vasireddy, G., Stent, A., Ehlen, P., Walker, M., Whittaker, S., & Maloor, P. (2002). MATCH: an architecture for multimodal dialogue systems. *Proc. 40th Annual Meeting of the Association for Computational Linguistics*, pp. 276–383.
Kara, L.B., & Stahovich, T.F. (2005). An image-based, trainable symbol recognizer for hand-drawn sketches. *Computer Graphics 29(4)*, 501–517.
Kendon, A. (1997). Gesture. *Annual Review of Anthropology 26(1)*, 109–128.
Krahnstoever, N., Kettebekov, S., Yeasin, M., & Sharma, R. (2002). A real-time framework for natural multimodal interaction with large screen displays. *Proc. 4th Int. Conf. Multimodal Interfaces (ICMI 2002)*, Pittsburgh, PA.
Liu, Y., Stolcke, A., Shriberg, E., & Harper, M.P. (2004). Comparing and combining generative and posterior probability models: some advances in sentence boundary detection in speech. *Proc. Empirical Methods in Natural Language Processing*, Barcelona.
Liu, Y., Stolcke, A., Shriberg, E., & Harper, M. (2005). Using conditional random fields for sentence boundary detection in speech. *ACL '05: Proc. 43rd Annual Meeting on Association for Computational Linguistics*, pp. 451–458. Morristown, NJ: Association for Computational Linguistics.
Luo, Y. (2008). Can subclasses help a multiclass learning problem? *Proc. Intelligent Vehicles Symposium, 2008 IEEE*, pp. 214–219.
Masry, M., Kang, D., & Lipson, H. (2005). A freehand sketching interface for progressive construction of 3D objects. *Computers & Graphics 29(4)*, 563–575.
McNeill, D. (1992). *Hand and Mind: What Gestures Reveal About Thought*. Chicago: University of Chicago Press.
Nakai, M., & Shimodaira, H. (1994). Accent phrase segmentation by finding n-best sequences of pitch pattern templates. *Proc. 3rd Int. Conf. Spoken Language Processing (ICSLP 94)*, pp. 347–350.
Novak, G.S.J., & Bulko, W.C. (1993). Diagrams and text as computer input. *Journal of Visual Languages and Computing 4(2)*, 161–175.
Oltmans, M. (2000). *Understanding naturally conveyed explanations of device behavior.* MS Thesis. Massachusetts Institute of Technology.
Oviatt, S. (2000). Taming recognition errors with a multimodal interface. *Communications of the ACM 43(9)*, 45–51.
Oviatt, S., Cohen, P., Wu, L., Vergo, J., Duncan, L., Suhm, B., Bers, J., Holzman, T., Winograd, T., Landay, J., Larson, J., & Ferro, D. (2000). Designing the user interface for multimodal speech and pen-based gesture applications: state-of-the-art systems and future research directions. *Human–Computer Interaction 15(4)*, 263–322.
Oviatt, S., DeAngeli, A., & Kuhn, K. (1997). Integration and synchronization of input modes during multimodal human–computer interaction. *CHI '97: Proc. SIGCHI Conf. Human Factors in Computing Systems*, pp. 415–422. New York: ACM.
Oviatt, S.L. (1999). Mutual disambiguation of recognition errors in a multimodel architecture. *Proc. CHI 99 Conf. Human Factors in Computing Systems: The CHI is the Limit*, pp. 576–583. New York: ACM.
Patel, R., Plimmer, B., Grundy, J., & Ihaka, R. (2007). Ink features for diagram recognition. *Proc. SBIM '07*, pp. 131–138.
Rubine, D. (1991). Specifying gestures by example. *Computer Graphics 25*, 329–337.
Silva, N., & Cardoso, T. (2004). *GIDeS++—Using Constraints to Model Scenes*, Technical Report. Information Society Technologies.
Stolcke, A., & Shriberg, E. (1996). Automatic linguistic segmentation of conversational speech. *Proc. Int. Conf. Spoken Language Processing* (Bunnell, H.. & Idsardi, W., Eds.), Vol. 2, pp. 1005–1008. Philadelphia, PA.
Stolcke, A., Shriberg, E., Bates, R., Ostendorf, M., Hakkani, D., Plauche, M., Tur, G., & Lu, Y. (1998). Automatic detection of sentence boundaries and disfluencies based on recognized words. *Proc. Int. Conf. Spoken Language Processing* (Mannell, R., & Robert-Ribes, J., Eds.), *Vol. 5*, pp. 2247–2250. Sydney: Australian Speech Science and Technology Association.
Strassel, S. (2004). *Simple metadata annotation specification version 6.2.* Linguistic Data Consortium.
Toutanova, K., & Manning, C.D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. *Proc. 2000 Joint SIGDAT Conf. Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 63–70. Morristown, NJ: Association for Computational Linguistics.
Ullman, D.G., Wood, S., & Craig, D. (1990). The importance of drawing in the mechanical design process. *Computers & Graphics 14(2)*, 263–274.
Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., & Woelfel, J. (2004). *Sphinx-4: A Flexible Open Source Framework For Speech Recognition*, Technical Report TR-2004-139. Sun Microsystems.
Wauchope, K. (1994). *Eucalyptus: Integrating Natural Language Input With a Graphical User Interface*, Technical Report NRL/FR/5510-94-9711. US Naval Research Laboratory.
Wobbrock, J.O., Wilson, A.D., & Li, Y. (2007). Gestures without libraries, toolkits or training: a $1 recognizer for user interface prototypes. *UIST '07: Proc. 20th Annual ACM Symp. Interface Software and Technology*, pp. 159–168. New York: ACM.
Woods, W., Bates, L., Bobrow, R., Brachman, R., Cohen, P.R., & Klovstad, J. (1979). *Research in Natural Language Understanding*, Annual Report 4274. Bolt Beranek and Newman.
Xing, E.P., Jordan, M.I., & Karp, R.M. (2001). Feature selection for high-dimensional genomic microarray data. *Proc. 18th Int. Conf. Machine Learning*, pp. 601–608. San Mateo, CA: Morgan Kaufmann.
Yerazunis, W. (2004). The spam-filtering accuracy plateau at 99.9 percent accuracy and how to get past it. *Proc. MIT Spam Conf., 2004*.

**James Herold** earned his BS in computer science at California Polytechnic State University, Pomona, in 2004. Jim is currently a PhD student in computer science at the University of California, Riverside.

**Thomas F. Stahovich** received his BS in mechanical engineering from the University of California, Berkeley, in 1988. He received his MS and PhD from MIT in 1990 and 1995, respectively. He conducted his doctoral research at the MIT Artificial Intelligence Lab. After serving as an Assistant and Associate Professor of mechanical engineering at Carnegie Mellon University in Pittsburgh, PA, Dr. Stahovich joined the Mechanical Engineering Department at University of California, Riverside, in 2003, where he is currently a Professor and Chair. His research interests include pen-based computing, educational technology, design automation, and design rationale management.