Gestural Cohesion for Topic Segmentation

Jacob Eisenstein and **Regina Barzilay** and **Randall Davis**
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
77 Massachusetts Ave., Cambridge MA 02139
{jacobe, regina, davis}@csail.mit.edu

Abstract

This paper explores the relationship between discourse structure and coverbal gesture. Using the idea of *gestural cohesion*, we show that coherent topic segments are characterized by homogeneous gestural forms, and that changes in the distribution of gestural features predict segment boundaries. Gestural features are extracted automatically from video, and are combined with lexical features in a hierarchical Bayesian model. Unsupervised inference is performed through Metropolis-Hastings sampling. The resulting multimodal system outperforms a verbal-only model, both with manual and automatically-recognized speech transcripts.

1 Introduction

When humans communicate face-to-face, discourse cues are expressed simultaneously through multiple channels. Previous research has extensively studied how discourse cues correlate with lexico-syntactic and prosodic features (Hearst, 1994; Hirschberg and Nakatani, 1998); this work informs multiple text and speech processing applications, such as automatic summarization and segmentation. Gesture is another communicative modality that frequently accompanies speech, yet its connection to discourse remains poorly understood.

This paper empirically demonstrates that gesture correlates with discourse structure. In particular, we show that automatically-extracted gesture features can be combined with lexical cues in a statistical model to predict discourse segmentation. Our method builds on the idea that coherent discourse segments are characterized by *gestural cohesion*; in other words, that such segments exhibit homogeneous gestural patterns. Lexical cohesion (Halliday and Hasan, 1976) forms the backbone of many verbal segmentation algorithms, on the theory that segmentation boundaries should be placed where the distribution of words changes (Hearst, 1994). With gestural cohesion, we explore whether the same idea holds for gesture features.

The motivation for this approach comes from a series of psycholinguistic studies suggesting that gesture supplements speech with meaningful and unique semantic content (Kendon, 1994; McNeill, 1992). We assume that repeated patterns in gesture are indicative of the semantic coherence that characterizes well-defined discourse segments. An advantage of this view is that gestures can be brought to bear on discourse analysis without undertaking the daunting task of recognizing and interpreting individual gestures. This is crucial because coverbal gesture – unlike formal sign language – rarely follows any predefined form or grammar, and may vary dramatically by speaker.

A key implementational challenge is automatically extracting gestural information from raw video and representing it in a way that can applied to discourse analysis. We employ a representation of *visual codewords*, which capture clusters of low-level motion patterns. For example, one codeword may correspond to strong left-right motion in the upper part of the frame. These codewords are then treated similarly to lexical items; our model identifies changes in their distribution, and predicts topic

boundaries appropriately. The overall framework is implemented in the form of a hierarchical Bayesian model, supporting flexible integration of multiple knowledge sources.

Experimental results support the hypothesis that gestural cohesion is indicative of discourse structure. Applying our algorithm to a dataset of face-to-face dialogues, we find that gesture features correlate with segment boundaries. Moreover, gesture appears to communicate unique information, improving segmentation performance over lexical features alone. The positive impact of gesture is most pronounced when automatically-recognized speech transcripts are used, but gestures improve performance even in combination with manual transcripts.

2 Related Work

Gesture and Discourse Much of the work on gesture in natural language processing has focused on multimodal dialogue systems in which the gestures and speech may be constrained, e.g. (Johnston, 1998). In contrast, we focus on improving discourse processing on unconstrained natural language between humans. This effort follows basic psychological and linguistic research on the communicative role of gesture (McNeill, 1992), including some research that made use of automatically acquired visual features (Quek et al., 2000). We extend these empirical studies with a statistical model of the relationship between gesture and discourse segmentation.

Hand-coded descriptions of body posture shifts and eye gaze behavior have been shown to correlate with topic and turn boundaries in task-oriented dialogue (Cassell et al., 2001). These findings are exploited to generate realistic conversational "grounding" behavior in an animated agent. The semantic content of gesture was leveraged – again, for gesture generation – in (Kopp et al., 2007), which presents an animated agent that is capable of augmenting navigation directions with gestures that describe the physical properties of landmarks along the route. Both systems generate plausible and human-like gestural behavior; we address the converse problem of *interpreting* such gestures.

In this vein, hand-coded gesture features have been used to improve sentence segmentation, showing that sentence boundaries are unlikely to overlap gestures that are in progress (Chen et al., 2006). Features that capture the start and end of gestures are shown to improve sentence segmentation beyond lexical and prosodic features alone. This idea of gestural features as a sort of visual punctuation has parallels in the literature on prosody, which we discuss in the next subsection.

Finally, ambiguous noun phrases can be resolved by examining the similarity of co-articulated gestures (Eisenstein and Davis, 2007). While noun phrase coreference can be viewed as a discourse processing task, we address the higher-level discourse phenomenon of topic segmentation. In addition, Eisenstein and Davis focus primarily on pointing gestures directed at pre-printed visual aids. In our domain, speakers do not have access to visual aids, and thus pointing gestures are less frequent than "iconic" gestures, in which the form of motion is the principle communicative feature (McNeill, 1992).

Nonverbal Features for Topic Segmentation Research on nonverbal features for topic segmentation has primarily focused on prosody, under the assumption that a key prosodic function is to mark structure at the discourse level (Steedman, 1990; Grosz and Hirshberg, 1992; Swerts, 1997). The ultimate goal of this research is to find correlates of hierarchical discourse structure in phonetic features.

Today, research on prosody has converged on prosodic cues which correlate with discourse structure. Such markers include pause duration, fundamental frequency, and pitch range manipulations (Grosz and Hirshberg, 1992; Hirschberg and Nakatani, 1998). These studies informed the development of applications such as segmentation tools for meeting analysis, e.g. (Tur et al., 2001; Galley et al., 2003).

In comparison, the connection between gesture and discourse structure is a relatively unexplored area, at least with respect to computational approaches. One conclusion that emerges from our analysis is that gesture may signal discourse structure in a different way than prosody does: while specific prosodic markers characterize segment boundaries, gesture predicts segmentation through intrasegmental cohesion. The combination of these two

modalities is an exciting direction for future research.

3 Gesture Representation for Discourse Analysis

The units of our analysis are *codewords*, a compact representation of salient visual features in video. Codewords characterize frequently-occurring patterns of motion and appearance at a local scale: for example, one codeword might represent left-to-right motion in the upper part of the frame. We detect instances of codewords at specific locations and times throughout each video; the total set of codeword types forms a sort of visual vocabulary.

By listing the codewords that occur during a given period of time – such as a sentence – we obtain a succinct representation of the ongoing gestural activity. Distributions of codewords over time can be analyzed in similar terms to the distribution of lexical features. The codeword representation provides a straightforward way to assess gestural coherence: a change in the distribution of codewords indicates new visual kinematic elements entering the discourse.

The left panel of figure 1 shows the distribution of codewords in a single video; vertical lines indicate topic boundaries. Each column represents a sentence, and the blocks in a column indicate the codewords occurring during the sentence duration. While noisy, it is possible even from visual inspection to identify some connections between the segmentation and the distribution of codewords – for example, the second-to-last segment has a set of codewords starkly different from its neighbors. The right panel shows the lexical features in the same dialogue.

Computing Codewords from Video Codewords are extracted using techniques from the computer vision domain of *activity recognition* (Dollár et al., 2005; Efros et al., 2003). The goal of activity recognition is to classify video sequences into semantic categories: e.g., walking, running, jumping. Many recent approaches have focused on sparse low-level features called *spatio-temporal interest points*: high-contrast image regions – especially corners and edges – that undergo complex motion. The visual, spatial, and kinematic properties of these interest points are concatenated into feature vectors, which, while noisy, permit robust classification of the desired activities and behaviors.

As a simple example, a classifier may learn that a key difference between videos of walking and jumping is that walking is characterized by horizontal motion and jumping is characterized by vertical motion. Spurious vertical motion in a walking video is unlikely to confuse the classifier as long as the large majority of interest points move horizontally. Our hypothesis is that just as such low-level movement features can be applied in a supervised fashion to distinguish activities, they can be applied in an unsupervised fashion to group co-speech gestures into perceptually meaningful clusters.

We apply the Activity Recognition Toolbox (Dollár et al., 2005)[1] to detect spatio-temporal interest points in our dataset. At each interest point, we extract the brightness gradient of a small space-time volume of nearby pixels. PCA is applied to reduce this high dimensional vector to three principle components (Hastie et al., 2001). The spatial location of the interest point is added to the feature vector, resulting in a total of five dimensions. Finally, we apply clustering to the interest points in each video, arriving at a set of twenty codewords. These codewords are the final representation of visual features in our model.

Previous applications of gesture to NLP have often focused on tracking the speaker's hands, head, or torso (e.g., (Eisenstein and Davis, 2007)). Such approaches are powerful but are difficult to implement; worse, once tracking is lost, it is difficult to recover. For these reasons, low-level approaches based on interest points are increasingly popular for related computer vision problems.

4 Bayesian topic segmentation

Topic segmentation is performed in a Bayesian framework, using a model that is similar to previous hidden Markov model (HMM) techniques (e.g., (Tur et al., 2001)). Each sentence's segment assignment is encoded with a hidden variable, which is assumed to be generated by a Markov process. Observations – the words and gesture codewords – are generated by language models that are indexed ac-

[1] http://vision.ucsd.edu/∼pdollar/research/cuboids_doc/index.html

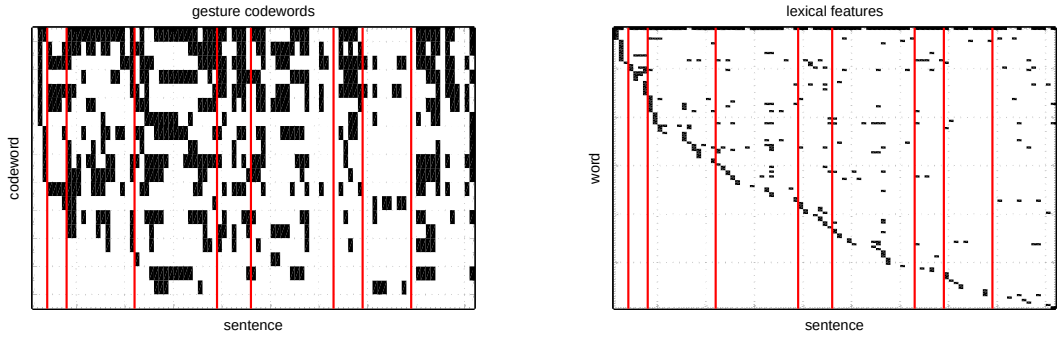

Figure 1: Distribution of gestural and lexical features by sentence. Segment breaks are indicated by red lines.

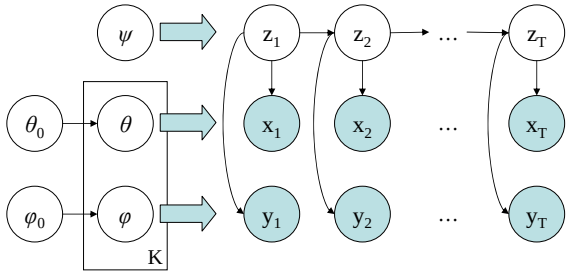

Figure 2: Plate diagram for Bayesian topic segmentation. Shaded nodes indicate observed variables. Thick shaded arrows indicate that the parameter to the left of the arrow impacts every element to the right, e.g., ψ is connected to every z_t.

cording to the segment. In such a framework, a high-likelihood segmentation will produce language models that are pure and distinct, thus maximizing the lexical coherence of each segment.

Figure 2 shows the plate diagram for this generative model, a Bayesian HMM. Each x_t represents the bag of words for the sentence t; y_t is the bag of gestures, and z_t is a positive integer indicating the segment assignment. The segment assignments are produced by a first-order Markov process, such that z_{t+1} is dependent on z_t and the parameter ψ. Words and gestures are generated by multinomial language models θ_{z_t} and ϕ_{z_t} respectively; z_t influences these observed variables by indexing a specific language model. Finally, each of the K language models are given symmetric Dirichlet priors with parameters θ_0 and ϕ_0 (Gelman et al., 2004).

As is common in speech recognition and other applications (Rabiner, 1989), we add a *left-right constraint*, such that $z_{t+1} \in \{z_t, z_t + 1\}$. This constraint limits the number of inter-state transitions to the desired number of segments, which we assume is specified in advance.[2] The parameter ψ specifies a distribution over segment durations. This is modeled using the negative-binomial distribution, an alternative to the Poisson distribution that permits greater variance. Given T sentences and K segments, the expected duration is T/K.

Our goal is to perform unsupervised inference on each document. Previous approaches to unsupervised inference in similar models include Gibbs sampling (Purver et al., 2006) and variational expectation-maximization (EM) (Beal, 2003). However, we find that the left-right constraint poses difficulties for both of these techniques. Gibbs sampling is known to converge slowly when the hidden variables are highly constrained (Gelman et al., 2004), as is the case here. Our experiments with variational EM showed that the left-right constraint makes it very sensitive to initialization.

One explanation for these difficulties is that both Gibbs sampling and variational EM attempt to search in the space of labellings. Given T sentences and K segments, there are K^T possible labellings, but only a small fraction are permissible segmentations, due to the left-right constraint. Rather than searching in this highly-constrained space, we prefer to search for segmentation points – given the left-right constraint, these search spaces are equivalent. Search is performed using the Metropolis-Hastings

[2]Evaluation is difficult when the target number of segments is unspecified, as there may be many equally appropriate segmentations at different levels of granularity. Prespecifying the desired segmentation granularity is common practice in topic segmentation, e.g. (Malioutov and Barzilay, 2006).

algorithm, a Markov Chain Monte Carlo (MCMC) technique.

4.1 Metropolis-Hastings for Segmentation

The Metropolis-Hastings algorithm samples from the configuration space of the model, and is guaranteed in the limit to draw samples from the posterior distribution of the hidden variables (Gelman et al., 2004). Metropolis-Hastings is a natural choice for inference in hierarchical Bayesian models because it permits search through arbitrary transformations of the model configuration.

Samples are generated from a proposal distribution q, and are stochastically accepted or rejected. The acceptance rate for a sample depends on two factors: the conditional probability of the sampled configuration given observations $\mathbf{x}$ and $\mathbf{y}$, and a correction if the proposal distribution is asymmetric. The probability of accepting a transformation from configuration S to $\tilde{S}$ is,

$$a(S, \tilde{S}) = \min\left[1, \frac{q(S|\tilde{S})p(\tilde{S}|\mathbf{x}, \mathbf{y})}{q(\tilde{S}|S)p(S|\mathbf{x}, \mathbf{y})}\right] \quad (1)$$

.

Here, the configuration S is the triple $\langle \mathbf{z}, \theta, \phi \rangle$.[3] Our proposal distribution is as follows: select any segmentation point with equal probability, and move the segmentation point left or right, with equal probability. The move distance is generated from an exponentially-decaying distribution, so that a move of one step is twice as likely as a move of two steps, and so on. Moves that eliminate segments or cross other segmentation points are rejected, and the configuration is left unchanged. This proposal distribution is symmetric, meaning that $q(S|\tilde{S}) = q(\tilde{S}|S)$, so those factors drop out of equation 1.

In the new configuration $\tilde{S}$, the segment indexes are updated appropriately, and the language models θ, ϕ are set to their expected posteriors, e.g., $E[\theta_k|\mathbf{x}_{z_t=k}, \theta_0]$. Due to multinomial-Dirichlet conjugacy, this expectation can be computed directly from the counts and the prior (Gelman et al., 2004). The conditional probability of a configuration is given by:

$$\begin{aligned} p(S|\mathbf{x}, \mathbf{y}) &\propto p(S)p(\mathbf{x}, \mathbf{y}|S) \\ &= p(\mathbf{z}|\psi)p(\theta|\theta_0)p(\phi|\phi_0)p(\mathbf{x}, \mathbf{y}|\mathbf{z}, \theta, \phi) \end{aligned} \quad (2)$$

For clarity, we consider only the verbal features, expanding $p(S|\mathbf{x})$ with only the parameter θ; the visual features $\mathbf{y}$ and ϕ are handled identically. We write n_i for the occupancy count of segment i, i.e., $n_i = \sum_t 1_{z_t=i}$, and $n_{i,j}$ as the count of times word j appears in segment i, i.e., $n_{i,j} = \sum_t 1_{x_t=j} 1_{z_t=i}$. NegBin$(n_i; \psi)$ indicates the probability of n_i under a negative binomial distribution with parameters ψ; similarly, Dir() represents a Dirichlet distribution, and Mult() represents a multinomial.

$$\begin{aligned} p(S|\mathbf{x}) &\propto \prod_i^K p(n_i|\psi)p(\theta_i|\theta_0)p(\mathbf{x}_{t:z_t=i}|\theta_i) \\ &= \prod_i^K \text{NegBin}(n_i; \psi)\text{Dir}(\theta_i; \theta_0) \prod_{\{t:z_t=i\}} \text{Mult}(x_t; \theta_i) \end{aligned}$$

We arrive at the following acceptance ratio:[4]

$$p(\tilde{S}|\mathbf{x})/p(S|\mathbf{x}) = \prod_i^K \frac{\text{NegBin}(\tilde{n}_i; \psi)}{\text{NegBin}(n_i; \psi)} \prod_j^W \frac{\tilde{\theta}^{\tilde{n}_{i,j}+\theta_0-1}}{\theta^{n_{i,j}+\theta_0-1}},$$

where W is the total number of words in the vocabulary, and $\theta_{i,j}$ is the parameter for word j in the multinomial θ_i.

By repeatedly drawing samples with this acceptance rate, we are guaranteed to converge to the posterior distribution over configuations. Annealing is applied to find the maximum *a posteriori* segmentation. The effect of annealing is to gradually decrease the probability of accepting moves that reduce the conditional likelihood. At a temperature of zero, only moves that increase the conditional likelihood are accepted, so the sampling converges to a final estimate. Annealing is frequently used to find MAP estimates in other applications of MCMC, particularly in natural language processing, e.g. (Goldwater and Griffiths, 2007).

[3] A fully Bayesian approach would integrate out the parameters θ and ϕ, such that the configuration S would only include the segmentation points. We leave this for future work.

[4] See http://XXXX for a full derivation.

5 Experimental Setup

Dataset Our dataset is composed of fifteen audio-video recordings of dialogues limited to three minutes in duration. The dataset includes nine different pairs of participants, and in each video one of five subjects is discussed: a "Tom and Jerry" cartoon, a "Star Wars" toy, and three mechanical devices, including a latchbox, a piston, a pez dispenser. One participant – "participant A" – was familiarized with the topic, and is tasked with explaining it to participant B, who is permitted to ask questions. Audio from both participants is used, but only video of participant A is used; we do not examine whether B's gestures are relevant to discourse segmentation. Neither participant has access to any supporting diagrams, and neither is permitted to draw.

Video is recorded using standard camcorders, with a resolution of 720 by 480 at 30 frames per second. The video is reduced to 360 by 240 grayscale images before visual analysis is applied. Audio is recorded using headset microphones. No manual postprocessing is applied to the video.

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

For automatic speech recognition, the default Microsoft speech recognizer was applied to each sentence, and the top-ranked recognition result was reported. As is sometimes the case in real-world applications, no speaker-specific training data is available, so the recognition quality is very poor – the word error rate is 77%.

Segment boundaries are specified as points that divide the dialogue into coherent topics. Segmentation points are required to coincide with sentence or turn boundaries. A second annotator, who is not an author on this paper, provided an additional set of segment annotations on six documents; on this subset of documents, the P_k between annotators was .306, and the WindowDiff was .325 (these metrics are explained in the next subsection). This is similar to the interrater agreement reported in (Malioutov and Barzilay, 2006).

Over the fifteen dialogues, a total of 7864 words were transcribed (524 per dialogue), spread over 1440 sentences or interrupted turns (96 per dialogue). There were a total of 102 segments (6.8 per dialogue), from a minimum of four to a maximum of ten. This rate of 15 sentences or turns per segment indicates relatively fine-grained segmentation; for examples, in the physics lecture corpus of (Malioutov and Barzilay, 2006), there are roughly 100 sentences per segment.

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

Two naïve baselines are evaluated. The random baseline arbitrarily selects k sentence breaks as segmentation points; results are averaged over 1000 iterations. The equal-width baseline places boundaries such that all segments contain an equal number of sentences. Thus, all systems – including these naïve baselines – were given the sentence boundaries and the correct number of segments; the task is to select the k sentence boundaries that most accurately segment the text.

Implementation All experiments use 10^5 sampling iterations. Implemented in Java, the total running time was less than two minutes. Longer sampling periods did not appreciably affect results. To obtain a final segmentation estimate, we used an annealing schedule that linearly reduced the temperature over the final $8 * 10^4$ iterations. We report the average results across ten different runs; the variance between runs was usually less than .01 on both evaluation metrics.

6 Results

Gesture Alone The first experiment examines the correlation of gesture features with segmentation boundaries, irrespective of verbal information. Each y_t is composed of the gesture codewords that occur during the sentence boundaries from the transcript, but lexical features are otherwise ignored. As shown on line 1 of Table 3, the gesture-only method achieve a P_k of .455, with a WindowDiff (WD) of

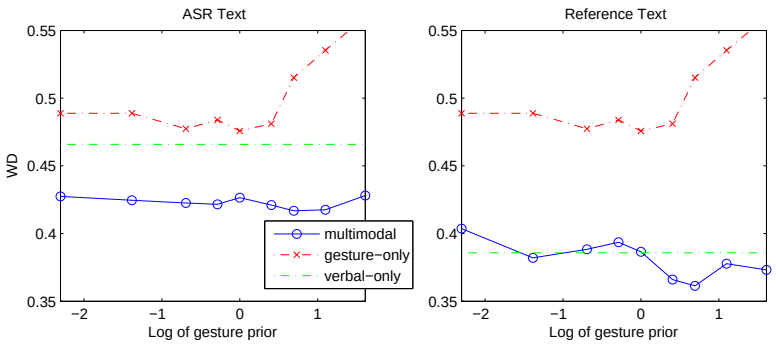

Figure 4: The multimodal and gesture performance are plotted with respect to the gesture prior; the verbal prior is held fixed at the optimal value. WD is a penalty, so lower scores indicate better segmentation.

Method	P_k	WD
1. gesture only	.455	.476
2. ASR only	.449	.466
3. ASR + gesture	.399	.415
4. transcript only	.348	.386
5. transcript + gesture	**.338**	**.361**
6. random	.473	.526
7. equal-width	.508	.515

Figure 3: For each method, the score of the best performing configuration is shown. P_k and WD are penalties, so lower values indicate better performance.

.476. This outperforms both naïve baselines (lines 6 and 7), supporting the hypothesis that gesture features predict discourse structure.

Gesture and Text Next, we examine whether the combination of gesture and verbal features outperforms verbal-only segmentation. First we consider ASR text, as returned by the Microsoft Speech Recognizer. Using ASR alone, the results are poor: line 2 of Table 3 shows a P_k of .449 and a WD of .466, only slightly better than the automatically recognized gesture features. However, when gesture features and ASR are combined, performance is substantially better than either modality in isolation, improving to .399 P_k and .415 WD (line 3 of the table). This represents an absolute gain of more than 4.5 percent on both metrics.

As expected, manual transcripts substantially increase performance: line 4 of the the table shows a $P_k = .348$ and WD = .386. Adding gesture, performance again improves: P_k drops to .338, and WD to .361. Even with perfect lexical features, automatically-recognized gestures add non-redundant information that improves discourse processing.

Interactions of Verbal and Gesture Features We now consider the relative contribution of the verbal and gesture features. In a discriminative setting, the contribution of each modality would be explicitly weighted. In a Bayesian generative model, the same effect is achieved through the use of the smoothing priors θ_0 (verbal) and ϕ_0 (gesture) – see equation 2 and Figure 2. For example, when the gesture prior is high and verbal prior is low, the gesture counts are smoothed, and the verbal counts play a greater role in segmentation. When both priors are very high, the model will simply try to find equally-sized segments (satisfying the distribution over durations).

The effects of these parameters can be seen in Figure 4. The verbal prior is held constant at its ideal value, and the WD score is plotted against the logarithm of the gesture prior. The right panel shows the results on the reference transcripts, where lexical features are especially accurate. In this setting, low values of the gesture prior impair performance (yielding high WD scores), as the gesture features overwhelm the effect of the more accurate verbal features. This effect is not observed on the ASR transcripts, as the gesture and verbal features are equally predictive.

Figure 5 shows a contour plot of the multimodal system's WD score (indicated by color) against the two priors. These graphs show that performance is relatively robust to prior settings, although it decreases sharply if the verbal prior is set too high (thereby diminishing the impact of the ver-

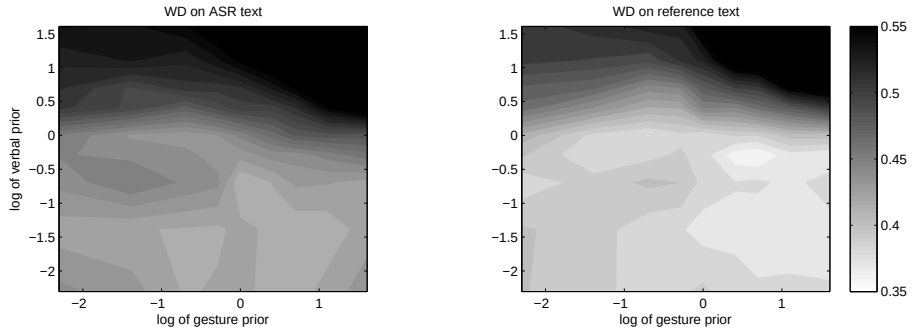

Figure 5: A contour plot of the WD penalty against both priors. Segmentation performance is represented by the color in the contour plot, with the gesture prior on the x-axis and the verbal prior on the y-axis.

bal features). Finally, we note that while the multimodal results are influenced by these priors, the same is true for the verbal-only systems. In future work we will consider automatic methods for setting these priors, such as additional Metropolis-Hastings moves (Goldwater and Griffiths, 2007).

Comparison to other models

While much of the research on topic segmentation focuses on written text, there are some comparable systems that also aim at unsupervised segmentation of spontaneous spoken language. For example, Malioutov and Barzilay (2006) segment a corpus of classroom lectures, using similar lexical cohesion-based features. With manual transcriptions, they report a .383 P_k and .417 WD on artificial intelligence (AI) lectures, and .298 P_k and .311 WD on physics lectures. This discrepancy suggests that segmentation scores are difficult to compare across domains; our results are in the range bracketed by these two extremes. The segmentation of physics lectures was at a very course level of granularity, while the segmentation of AI lectures was more similar to our annotations.

We applied the publicly-available executable for this algorithm to our data, but performance was poor, yielding a .434 P_k and .482 WD even when both verbal and gestural features were available. This may be because the technique is not designed for the relatively fine-grained segmentation demanded by our dataset (Malioutov, 2006).

7 Conclusions

This research shows a novel relationship between gestural cohesion and discourse structure. Automatically extracted gesture features are predictive of discourse segmentation when used in isolation; when lexical information is present, segmentation performance is further improved. This suggests that gestures provide unique information not present in the lexical features alone, even when perfect transcripts are available.

There are at least two possibilities for how gesture might impact topic segmentation: "visual punctuation," and cohesion. The visual punctuation view would attempt to identify specific gestural patterns that are characteristic of segment boundaries. This is analogous to research that identifies prosodic signatures of topic boundaries, such as (Hirschberg and Nakatani, 1998). By design, our model is incapable of exploiting such phenomena, as our goal is to investigate the notion of gestural cohesion. Thus, the performance gains demonstrated in this paper cannot be explained by such punctuation-like phenomena; we believe that they are due to the consistent gestural themes that characterize coherent topics. However, we are interested in pursuing the idea of visual punctuation in the future, so as to compare the power of visual punctuation and gestural cohesion to predict segment boundaries. The interaction of gesture and prosody suggests even further fruitful avenues of research.

Finally, topic segmentation is only one form of discourse structure. We would like to explore the relationship between gesture and richer discourse structures, such as hierarchical segmentation (Grosz and Sidner, 1986) or Rhetorical Structure Theory (Mann and Thompson, 1988).

References

Matthew J. Beal. 2003. *Variational Algorithms for Approximate Bayesian Inference*. Ph.D. thesis, University of London.

Doug Beeferman, Adam Berger, and John D. Lafferty. 1999. Statistical models for text segmentation. *Machine Learning*, 34(1-3):177–210.

Justine Cassell, Yukiko I. Nakano, Timothy W. Bickmore, Candace L. Sidner, and Charles Rich. 2001. Non-verbal cues for discourse structure. In *Proceedings of the ACL*, pages 106–115.

Lei Chen, Mary Harper, and Zhongqiang Huang. 2006. Using maximum entropy (ME) model to incorporate gesture cues for sentence segmentation. In *Proceedings of ICMI*, pages 185–192. ACM Press.

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

Alexei A. Efros, Alexander C. Berg, Greg Mori, and Jitendra Malik. 2003. Recognizing action at a distance. *International Conference on Computer Vision*, pages 726–733.

Jacob Eisenstein and Randall Davis. 2007. Conditional modality fusion for coreference resolution. In *Proceedings of the ACL*, pages 352–359.

M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing. 2003. Discourse segmentation of multi-party conversation. *Proceedings of the ACL*, pages 562–569.

Andrew Gelman, John B. Carlin, Hal .S. Stern, and Donald .B. Rubin. 2004. *Bayesian data analysis*. Chapman and Hall/CRC.

Sharon Goldwater and Tom Griffiths. 2007. A fully bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the ACL*, pages 744–751, June.

Barbara Grosz and Julia Hirshberg. 1992. Some intonational characteristics of discourse structure. In *ICSLP*, pages 429–432.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London.

T. Hastie, R. Tibshirani, and J. Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Marti A. Hearst. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the ACL*, June.

J. Hirschberg and C.H. Nakatani. 1998. Acoustic Indicators Of Topic Segmentation. *ICSLP*.

Xuedong Huang, Fileno Alleva, Mei-Yuh Hwang, and Ronald Rosenfeld. 1993. An overview of the Sphinx-II speech recognition system. In *Proceedings of ARPA Human Language Technology Workshop*, pages 81–86.

Michael Johnston. 1998. Unification-based multimodal parsing. In *Proceedings of COLING-1998*, pages 624–630.

Adam Kendon. 1994. Do gestures communicate? a review. *Research on language and social interaction*, 27:175–200.

Stefan Kopp, Paul Tepper, Kim Ferriman, and Justine Cassell. 2007. Trading spaces: How humans and humanoids use speech and gesture to give directions. In Toyoaki Nishida, editor, *Conversational Informatics: An Engineering Approach*. Wiley.

Igor Malioutov and Regina Barzilay. 2006. Minimum cut model for spoken lecture segmentation. In *Proceedings of the ACL*, pages 25–32. Association for Computational Linguistics, July.

Igor Malioutov. 2006. Minimum cut model for spoken lecture segmentation. Master's thesis, Massachusetts Institute of Technology.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8:243–281.

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

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

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

M. Purver, T.L. Griffiths, K.P. Körding, and J.B. Tenenbaum. 2006. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the ACL*, pages 17–24.

Francis Quek, David McNeill, Robert Bryll, Cemil Kirbas, Hasan Arslan, Karl E. McCullough, Nobuhiro Furuyama, and Rashid Ansari. 2000. Gesture, speech, and gaze cues for discourse segmentation. In *Proceedings of CVPR*, volume 2, pages 247–254.

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

Mark Steedman. 1990. Structure and intonation in spoken language understanding. In *Proceedings of the ACL*, pages 9–16.

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

Gokhan Tur, Dilek Hakkani-Tur, Andreas Stolcke, and Elizabeth Shriberg. 2001. Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics*, 27(1):31–57.