A Hierarchical Bayesian Model for Topic Segmentation

Konrad P. Körding
Massachusetts Institute of Technology
Cambridge, MA 02139
kording@mit.edu

Thomas L. Griffiths
Brown University
Providence, RI 02912
tom_griffiths@brown.edu

Matthew Purver
Stanford University
Stanford, CA 94305
mpurver@stanford.edu

Joshua B. Tenenbaum
Massachusetts Institute of Technology
Cambridge, MA 02139
jbt@mit.edu

Abstract

Many streams of real-world data, such as conversations or body movements, consist of relatively coherent segments, each characterized by particular topics or controllers. Making sense of these data requires simultaneously segmenting the sequences and inferring the structure of the segments. We present a hierarchical Bayesian model that can be used to break a sequence of utterances or movements into segments with different distributions over topics or controllers. We apply this model to a database of meetings, showing that its unsupervised segmentation is competitive with other approaches, and a database of human hand movements, revealing some of the controllers for motions of the hand.

1 Introduction

Natural text or conversation streams, such as radio news or meeting discussions, often consist of a progression of semantically coherent segments. A particular topic or set of topics will be relevant for some number of utterances, until the point at which the discussion moves on to a new segment with a new distribution of topics. Given sufficient experience with a domain such as radio news or technical meetings, people can make sophisticated inferences about new data in that domain, inferring when segments end as well as the set of topics that characterize each segment.

The computational problems of identifying segments and inferring their structure are heavily intertwined. If we were told the location of segment boundaries, it would be straightforward to learn to analyze each segment using standard topic-based models for unsupervised categorization [1, 2, 3]. Likewise, if we were given a set of topics and told the distribution over topics at each point in the sequence, it would be straightforward to break the sequence into topically coherent segments. However, in many cases, we must simultaneously discover both the segment boundaries and the underlying topics.

The task of simultaneously segmenting a sequence and identifying the structure of the segments arises throughout perception, cognition, and action. In visual scene understanding, one may think of object categories as analogous to topics [4, 5]. Then consider watching a movie: within scenes, the "topics" (object categories) are constant, but change arbitrarily across scenes. Also, within a single image, the topics in one part of the image may be coherent but unrelated to those in another region of the same image (e.g., pedestrians on a street in the foreground, with boats on a lake in the background). In the human motor system, complex movements can be generated from a basis set of controllers [6]. A sequence of actions can be segmented into movements, each of which draws on a characteristic set of controllers ("motor topics"), with little coherence across movement segments. Building a system that learns to understand language, visual scenes, or action sequences requires solving the joint problems of inferring segment boundaries and discovering a set of topics or basis elements that can characterize the structure of each segment.

In this paper, we develop a formal framework for solving these joint inference problems. We define a hierarchical Bayesian model for sequential data which can be used to simultaneously divide a sequence into segments that have a common distribution over topics and infer the topics themselves. We build on previous work on probabilistic topic models that can infer a set of topics from a corpus in which words are divided into documents [1, 2, 3]. By treating each segment of a sequence as a "document", our approach extends these models to the case in which the boundaries between documents are unknown, providing a fully generative analogue of the model described in [7]. We apply this model to two kinds of data: a corpus of technical meetings, and a database of human hand movements.

2 Learning topics and segments

In specifying our model, we will use terminology appropriate for linguistic data. Assume we have a corpus of U utterances, ordered in sequence. The uth utterance consists of N_u words, chosen from a vocabulary of size W. The set of words associated with the uth utterance are denoted $\mathbf{w}_u$, and indexed as $w_{u,i}$. The entire corpus is represented by $\mathbf{w}$.

Following previous work on probabilistic topic models [1, 2, 3], we will model each utterance as being generated from a particular distribution over topics, where each topic is a probability distribution over words. The utterances are ordered sequentially, and we assume a Markov structure on the distribution over topics: with high probability, the distribution for utterance u is the same as for utterance $u-1$; otherwise, we sample a new distribution over topics. This pattern of dependency is produced by associating a binary switching variable with each utterance, indicating whether its topic is the same as that of the previous utterance. The joint states of all the switching variables define segments that should be semantically coherent, because their words are generated by the same topic vector. We will first describe this generative model in more detail, and then discuss inference in this model.

2.1 A hierarchical Bayesian model

We are interested in where changes occur in the set of topics discussed in a sequence of utterances. To this end, let c_u indicate whether a change in the distribution over topics occurs at the uth utterance and let $P(c_u = 1) = \pi$. The distribution over topics associated with the uth utterance will be denoted $\theta^{(u)}$, and is a multinomial distribution over T topics, with the probability of topic t being $\theta_t^{(u)}$. If $c_u = 0$, then $\theta^{(u)} = \theta^{(u-1)}$. Otherwise, $\theta^{(u)}$ is drawn from a symmetric Dirichlet distribution with parameter α. The distribution is thus

$$P(\theta^{(u)}|c_u, \theta^{(u-1)}) = \begin{cases} \delta(\theta^{(u)}, \theta^{(u-1)}) & c_u = 0 \\ \frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T} \prod_{t=1}^{T} (\theta_t^{(u)})^{\alpha-1} & c_u = 1 \end{cases}, \tag{1}$$

where $\delta(\cdot, \cdot)$ is the Dirac delta function, and $\Gamma(\cdot)$ is the generalized factorial function. This distribution is not well-defined when $u = 1$, so we set $c_1 = 1$ and draw $\theta^{(1)}$ from a symmetric Dirichlet(α) distribution accordingly.

As in [1, 2, 3], we assume that each topic T_j is a multinomial distribution $\phi^{(j)}$ over words,

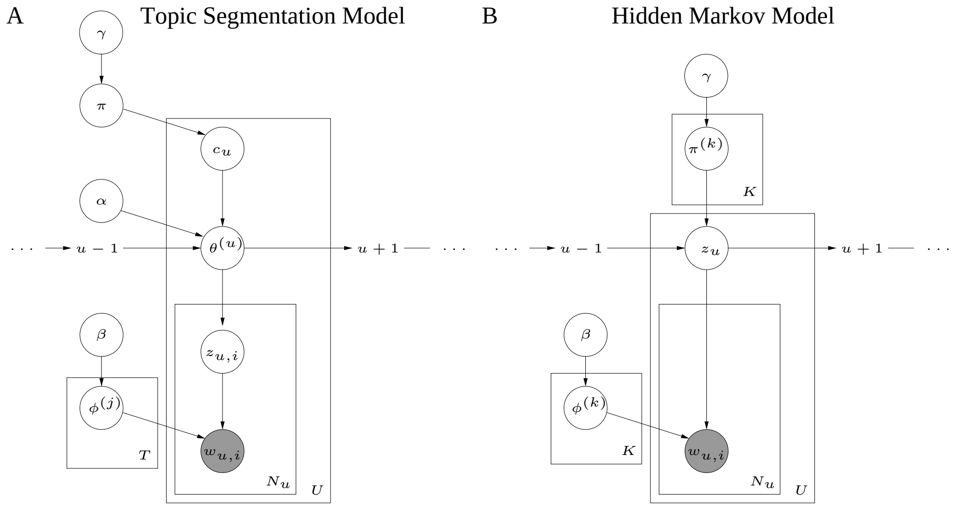


Figure 1: Graphical models indicating the dependencies among variables in A) the topic segmentation model and B) the hidden Markov model used as a comparison.

and the probability of the word w under that topic is $\phi_w^{(j)}$. The uth utterance is generated by sampling a topic assignment $z_{u,i}$ for each word i in that utterance with $P(z_{u,i} = t|\theta^{(u)}) = \theta_t^{(u)}$, and then sampling a word $w_{u,i}$ from $\phi^{(j)}$, with $P(w_{u,i} = w|z_{u,i} = j, \phi^{(j)}) = \phi_w^{(j)}$. If we assume that π is generated from a symmetric Beta(γ) distribution, each $\phi^{(j)}$ is generated from a symmetric Dirichlet(β) distribution, we obtain a joint distribution over all of these variables with the dependency structure shown in Figure 1A.

2.2 Inference

Assessing the posterior probability distribution over topic changes, **c**, given a corpus, **w**, can be simplified by integrating out the parameters θ, ϕ, and π. According to Bayes rule,

$$P(\mathbf{z}, \mathbf{c}|\mathbf{w}) = \frac{P(\mathbf{w}|\mathbf{z})P(\mathbf{z}|\mathbf{c})P(\mathbf{c})}{\sum_{\mathbf{z},\mathbf{c}} P(\mathbf{w}|\mathbf{z})P(\mathbf{z}|\mathbf{c})P(\mathbf{c})}. \tag{2}$$

Evaluating $P(\mathbf{c})$ requires integrating over π. Specifically, we have

$$P(\mathbf{c}) = \int_0^1 P(\mathbf{c}|\pi)P(\pi)\; d\pi = \frac{\Gamma(2\gamma)}{\Gamma(\gamma)^2} \frac{\Gamma(n_1 + \gamma)\Gamma(n_0 + \gamma)}{\Gamma(N + 2\gamma)}, \tag{3}$$

where n_1 is the number of utterances for which $c_u = 1$, and n_0 is the number of utterances for which $c_u = 0$. Computing $P(\mathbf{w}|\mathbf{z})$ proceeds along similar lines

$$P(\mathbf{w}|\mathbf{z}) = \int_{\Delta_W^T} P(\mathbf{w}|\mathbf{z}, \phi)P(\phi)\; d\phi = \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^T \prod_{t=1}^{T} \frac{\prod_{w=1}^{W} \Gamma(n_w^{(t)} + \beta)}{\Gamma(n_{\cdot}^{(t)} + W\beta)}, \tag{4}$$

where Δ_W^T is the T-dimensional cross-product of the multinomial simplex on W points, $n_w^{(t)}$ is the number of times word w is assigned to topic t in **z**, and $n_{\cdot}^{(t)}$ is the total number of words assigned to topic t in **z**. To evaluate $P(\mathbf{z}|\mathbf{c})$ we have

$$P(\mathbf{z}|\mathbf{c}) = \int_{\Delta_T^U} P(\mathbf{z}|\theta)P(\theta|\mathbf{c})\; d\theta. \tag{5}$$

The fact that the c_u variables effectively divide the sequence of utterances into segments that use the same distribution over topics simplifies solving the integral and we obtain:

$$P(\mathbf{z}|\mathbf{c}) = \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^{n_1} \prod_{u \in \mathcal{U}_1} \frac{\prod_{t=1}^{T} \Gamma(n_t^{(\mathcal{S}_u)} + \alpha)}{\Gamma(n_\cdot^{(\mathcal{S}_u)} + T\alpha)}, \tag{6}$$

where $\mathcal{U}_1 = \{u|c_u = 1\}$ and $\mathcal{U}_0 = \{u|c_u = 0\}$, and $\mathcal{S}_u$ denotes the set of utterances that share the same topic distribution (i.e. belong to the same segment) as u.

Equations 3, 4, and 6 allow us to evaluate the numerator of the expression in Equation 2. However, computing the denominator of this expression is intractable. Consequently, we sample from the posterior distribution using Markov chain Monte Carlo (MCMC) [8]. We will use Gibbs sampling, drawing the topic assignment for each word, $z_{u,i}$, conditioned on all other topic assignments, $\mathbf{z}_{-(u,i)}$, all topic change indicators, $\mathbf{c}$, and all words, $\mathbf{w}$, and then drawing the topic change indicator for each utterance, c_u, conditioned on all other topic change indicators, $\mathbf{c}_{-u}$, all topic assignments $\mathbf{z}$, and all words $\mathbf{w}$.

The conditional probabilities we need can be derived directly from Equations 3, 4, and 6. The conditional probability of $z_{u,i}$ indicates the probability that $w_{u,i}$ should be assigned to a particular topic, given other assignments, the current segmentation, and the words in the utterances. Cancelling constant terms, we obtain

$$P(z_{u,i}|\mathbf{z}_{-(u,i)}, \mathbf{c}, \mathbf{w}) = \frac{n_{w_{u,i}}^{(t)} + \beta}{n_\cdot^{(t)} + W\beta} \frac{n_{z_{u,i}}^{(\mathcal{S}_u)} + \alpha}{n_\cdot^{(\mathcal{S}_u)} + T\alpha}, \tag{7}$$

where all counts (i.e. the n terms) exclude $z_{u,i}$. The conditional probability of c_u indicates the probability that a new segment should start at u. In sampling c_u from this distribution, we are splitting or merging segments. Cancelling constant terms, we obtain

$$P(c_u|\mathbf{c}_{-u}, \mathbf{z}, \mathbf{w}) \propto \begin{cases} \frac{\prod_{t=1}^{T} \Gamma(n_t^{(\mathcal{S}_u^0)} + \alpha)}{\Gamma(n_\cdot^{(\mathcal{S}_u^0)} + T\alpha)} \frac{n_0 + \gamma}{N + 2\gamma} & c_u = 0 \\ \frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T} \frac{\prod_{t=1}^{T} \Gamma(n_t^{(\mathcal{S}_{u-1}^1)} + \alpha)}{\Gamma(n_\cdot^{(\mathcal{S}_{u-1}^1)} + T\alpha)} \frac{\prod_{t=1}^{T} \Gamma(n_t^{(\mathcal{S}_u^1)} + \alpha)}{\Gamma(n_\cdot^{(\mathcal{S}_u^1)} + T\alpha)} \frac{n_1 + \gamma}{N + 2\gamma} & c_u = 1 \end{cases} \tag{8}$$

where $\mathcal{S}_{u'}^1$ is $\mathcal{S}_{u'}$ for the segmentation when $c_u = 1$, $\mathcal{S}_{u'}^1$ is $\mathcal{S}_{u'}$ for the segmentation when $c_u = 0$, and all counts (e.g. n_1) exclude c_u. For this paper, we fixed α, β and γ at 0.01.

3 Results

3.1 Simulated data

To analyze the properties of this algorithm we first applied it to simulated data (Figure 2). The dataset was a sequence of 10,000 words out of a vocabulary of 25. Each segment in this dataset consisted of 100 successive words that shared the same topic distribution, with each subsequence of 10 words defined to be one utterance. The topic-word assignments were chosen such that when the words are aligned in a 5×5 grid the topics were binary bars. The topic distributions for the different segments were drawn from a Dirichlet distribution with $\beta = 0.1$. The resulting word sequence was supplied to the inference algorithm, which was run for 200,000 iterations, with samples collected after every 1,000 iterations to minimize autocorrelation. Figure 2 shows the inferred topic-word distributions and segment boundaries, which correspond well with those used to generate the data.

To compare with a similar but simpler model we applied a 10 state hidden Markov model (HMM) to the same data, using a similar Gibbs sampling algorithm. HMMs are often used for text segmentation (e.g., [9]). The key difference between the two models is shown in Figure 1. In the HMM, all variation in the content of utterances is modeled at a single

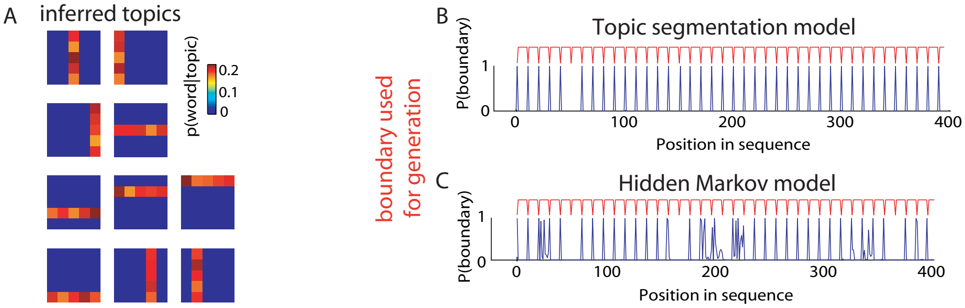


Figure 2: Simulated data. A) Inferred topics. B) Posterior probabilities of segment boundaries under the topic segmentation model and HMM, averaged over the last 100 samples.

level, with each segment having a distribution over words corresponding to a single state. The hierarchical structure of the topic segmentation model allows variation in content to be expressed at two levels, with each segment being produced from a linear combination of the distributions associated with each topic. Consequently, the topic segmentation model can often capture the content of a sequence of words by postulating a single segment with a novel distribution over topics, while the HMM has to frequently switch between states.

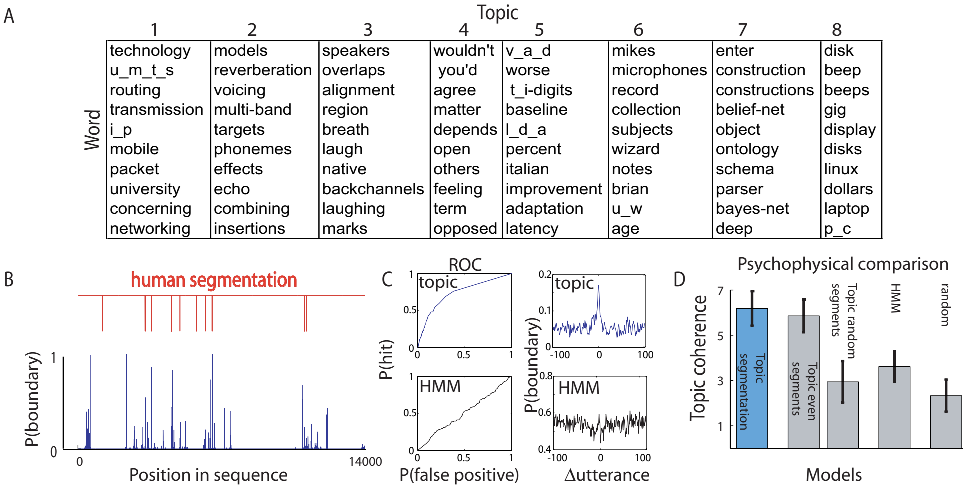
3.2 Segmenting meetings

We applied the algorithm to the ICSI meeting corpus [10], which consists of text transcripts of spoken multi-party meetings. We compared the results to two different human-annotated segmentations produced independently for certain portions of the corpus [11, 12]. These data were not supplied to the model: topic inference and segmentation was completely unsupervised, and the human judgments were only used to evaluate performance.

Data from all meetings were merged into a single dataset that contained 746,605 word tokens. We sampled for 200,000 iterations of MCMC, taking samples every 1,000 iterations. Figure 3A shows the most indicative words for the topics inferred at the last iteration.[1] Figure 3B shows an example of how the inferred topic segmentation probabilities at each utterance compare with the segment boundaries as judged by human raters. This relationship is further quantified in Figure 3C where ROC curves show that the segmentation probabilities can indeed be used to predict segmentation boundaries placed by humans. The boundaries that are placed by the topic segmentation model are often close to the boundaries placed by humans. Because the HMM cannot combine different topics it places a lot of segmentation boundaries, resulting in inferior performance. Using stemming and a bigram representation, however, might improve its performance [9].

To quantitatively compare models and address the question of how many topics we should be using we varied the number of topics. We assessed performance using the P_k error measure proposed by [13], which intuitively provides a measure of the probability that two points drawn from the meeting will be *incorrectly* separated by a hypothesized segment boundary – thus, lower P_k figures indicate better agreement with the human-annotated results. For the numbers of segments we are dealing with, a baseline of segmenting the discourse into equal-length segments gives a P_k of about 50%. We optimized a threshold on the posterior probability of a segment boundary for each model, finding P_k of 28.4%, 29.7%, 32.9% and 29.0% when using 2, 5, 10 or 20 topics respectively. Segmentation quality is thus hardly affected by the overall number of topics used. Using a similar procedure with a 10 state HMM we find a P_k value of 37.5%, although this performance is actually

[1] The ICSI corpus is drawn mainly from meetings of the ICSI speech group – thus topics discussed include speech recognition techniques, meeting recording, hardware setup etc.



Topic 1	2	3	4	5	6	7	8
technology	models	speakers	wouldn't	v_a_d	mikes	enter	disk
u_m_t_s	reverberation	overlaps	you'd	worse	microphones	construction	beep
routing	voicing	alignment	agree	t_i-digits	record	constructions	beeps
transmission	multi-band	region	matter	baseline	collection	belief-net	gig
i_p	targets	breath	depends	l_d_a	subjects	object	display
mobile	phonemes	laugh	open	percent	wizard	ontology	disks
packet	effects	native	others	italian	notes	schema	linux
university	echo	backchannels	feeling	improvement	brian	parser	dollars
concerning	combining	laughing	term	adaptation	u_w	bayes-net	laptop
networking	insertions	marks	opposed	latency	age	deep	p_c

Figure 3: Results from segmenting the meetings database. A) The words most indicative for each topic. B) Probability of a segment boundary, compared with human segmentation, for an arbitrary subset of the data. C) ROC curves for predicting segmentation boundaries defined by humans and the conditional probability of the model placing a boundary at an offset. D) Subjective topic coherence ratings.

achieved by exploiting an *anti*-correlation between the HMM segment boundaries and human judgments. Performance of our model exceeds that of another unsupervised system, based upon lexical coherence, which gives a P_k of 31.9% [11]. Combining the boundaries obtained by the topic segmentation algorithm with other features in a supervised way should allow reaching even better segmentation.

To evaluate the quality of the inferred topics we did a psychophysical experiment in which seven human observers rated (on a scale of 1 to 9) the semantic coherence of 50 lists of 10 words (Figure 3D). Of these lists, 40 contained the most indicative words for each of the 10 topics from different models: the topic segmentation model, a topic model that had the same number of segments but with fixed evenly spread segmentation boundaries, a topic model with random segmentation boundaries, and the HMM. The other 10 lists contained random samples of 10 words from the words of the other 40 lists. Figure 3 shows that the topic segmentation model produced the most coherent topics, but using an even distribution of boundaries performs similarly. Topic quality is thus not very susceptible to the precise segmentation of the text. However, the topic segmentation model is able to identify meaningful segment boundaries at the same time as inferring topics.

3.3 Movement data

Prominent theories of motor control propose that the human movement system uses a number of controllers, and at given points in time switches between them or adaptively combines different controllers [6]. Such controllers could either be high-level state-dependent controllers or simple combinations between different muscles, where several muscles are controlled in conjunction [14]. Observing the movements of people should make it possible to infer the underlying controllers or "motor topics" from the movements they make.

Traditionally algorithms have been used that do not allow the segmentation of natural movements. For this reason inferring the properties of the controllers from natural movements has not been possible. Instead, laboratory experiments are used, where the onset and the target is clearly defined [15]. In some cases more natural movements have been used,

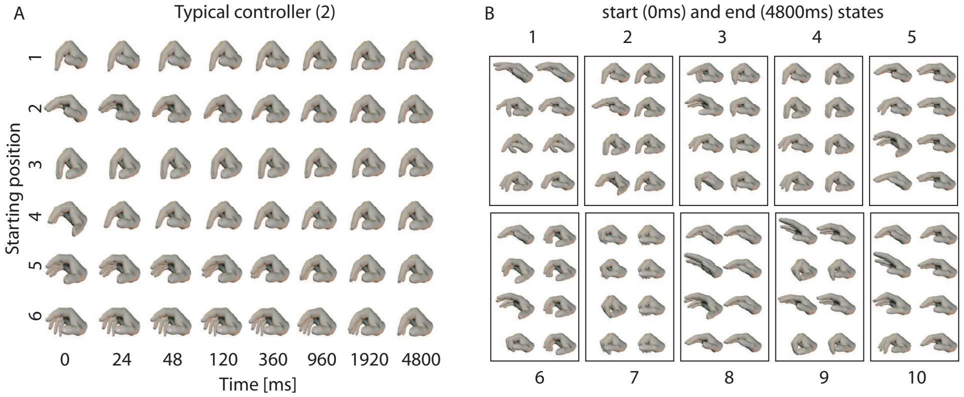


Figure 4: Inferring motor topics: A) for one typical controller the average evolution from different starting states is shown. B) for each of the controllers random combinations between starting state and end state (after about 5 sec. simulated time) are shown.

but without the segmentation only simple instantaneous statistics like principal components [16] or nonnegative coding dimensions [17] could be analyzed.

We examined whether nonlinear controllers characterized by posture-posture transitions can be revealed by a topic segmentation approach. One healthy male subject aged 31 participated in this study. For 200 minutes his right hand movements were measured using a CyberGlove (Virtual Technologies, Palo Alto, CA). The sensors were associated with 19 degrees of freedom of the hand and their readings was recorded by a lightweight backpack. Sensors were sampled at 42.5 Hz. Vector quantization was used to describe this dataset by a codebook of 200 vectors (83% of variance explained). The transitions between such states (4,260 bigrams) are used as input to the topic segmentation algorithm. With that much data (510,000 samples) we can characterize movements this subject makes during a typical few hours of his life (including a brief visit to a British pub).

The model assumes that 10 stochastic controllers (equivalent to topics), characterized by their posture-posture transition probabilities (bigram frequencies), are used, and that during each segment of movement a mixture of controllers may be active. Figure 4 shows the average behavior of the controllers. Most of the controllers (2-8 and 10) can be understood as moving the hand towards a given posture, which includes movements such as grasping. Some neural recordings [18] indicate that the nervous system uses such a coding scheme.

Controller 1 deserves special attention as the transitions encoded by it are almost exclusively transitions from each state to itself. It thus encodes keeping the hand stationary. Mixing controller 1 with any other controller allows slower movements. This controller is responsible for 39% of the transitions in the data. Controller 9 encodes two different ending positions. This might indicate that the number of controllers used was too small. Most of the controller transitions (83%) happen within segments giving evidence for the idea that indeed controllers are combined in an adaptive way as predicted by recent theories [6].

4 Conclusion

Probabilistic topic models can identify the topics expressed in a set of documents. Using these models requires that words be divided into documents. This is not the case in many natural settings where language models are relevant, such as modeling the content of conversations or meetings. Neuroscience is moving from describing neural data for simple stimuli and simple, well defined movements to analyzing progressively more natural data. In their everyday life animals progress from one movement target to the next and their mind wanders from one thought to the next. Algorithms that solve the problem of simultaneously identifying segments and their structure are necessary in order to make sense of data produced from natural behavior, whether those data concern speech or action. We have presented a hierarchical Bayesian model that can be used to solve this problem. This is a fully generative model for sequence data, and extends previous work on topic models to allow them to be applied in a range of novel settings. In addition to providing an effective method for identifying segments in meetings, this model can be used to identify some of the basic controllers of human motion.

Acknowledgements. This work was funded by the DARPA/SRI CALO project. JBT was supported by the P. E. Newton Career Development Chair. KPK was supported by a DFG Heisenberg Stipend. We want to thank J. Ingram and D. Wolpert for the motor data and M. Tresch for inspiring discussions.

References

[1] T. Hofmann. Probablistic latent semantic indexing. In *Proceedings of the Twenty-Second Annual International SIGIR Conference*, 1999.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Science*, 101:5228–5235, 2004.

[4] P. Moreels and P. Perona. Common-frame model for object recognition. In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

[5] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *MIT CSAIL AI Memo*, 2005.

[6] M. Haruno, D. Wolpert, and M. Kawato. Mosaic model for sensorimotor learning and control. *Neural Computation*, 13:2201–2220, 2001.

[7] D. Blei and P. Moreno. Topic segmentation with an aspect hidden Markov model. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001.

[8] W.R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, Suffolk, 1996.

[9] R. Barzilay and L. Lee. Catching the drift: Probabilistic content models, with applications to generation and summarization. *Proceedings of HLT-NAACL*, pages 113–120, 2004.

[10] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. The ICSI meeting corpus. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 364–367, 2003.

[11] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing. Discourse segmentation of multi-party conversation. In *Proc. 41st Annual Meeting of the Association for Computational Linguistics*, July 2003.

[12] A. Gruenstein, J. Niekrasz, and M. Purver. Meeting structure annotation: Data and tools, to appear.

[13] D. Beeferman, A. Berger, and J. Lafferty. Statistical models for text segmentation. *Machine Learning*, 34(1-3):177–210, 1999.

[14] E. Bizzi, A. D'avella, and M. Tresch. Modular organization of spinal motor systems. *Neuroscientist*, 8:437–442, 2002.

[15] M. Santello, M. Flanders, and J. Soechting. Patterns of hand motion during grasping and the influence of sensory guidance. *Journal of Neuroscience*, 22:1426–1435, 1998.

[16] E. Todorov and Z. Ghahramani. Analysis of the synergies underlying complex hand manipulation. In *Annual International Conference of the IEEE Engineering in Biology and Medicine Society*, 2004.

[17] M. Tresch, P. Saltiel, and E. Bizzi. The construction of movement by the spinal cord. *Nature Neuroscience*, 2:162–167, 1999.

[18] M. Graziano, C. Taylor, and T. Moore. Complex movements evoked by microstimulation of precentral cortex. *Neuron*, 34:841–851, 2002.