SUPPORT VECTOR MACHINES FOR DISCOURSE SEGMENTATION IN MEETINGS

Lyndon S. Kennedy

Deptartment of Electrical Engineering, Columbia University, New York, NY 10027
lyndon@ee.columbia.edu

ABSTRACT

We build a system for the automatic detection of topic boundaries in meeting recordings. We extract a number of prosodic and lexical features from a single tabletop microphone and the human-generated transcript. We use a support vector machine trained on these features to learn the properties of topic boundaries. In a detection experiment, we see results with up to 98% correct accept rate with a 5% false alarm rate, a significant gain in performance over simple modeling techniques, such as single-mixture Gaussian mixture models.

1. INTRODUCTION

There has been much recent interest in the automatic recognition and understanding of recordings of meetings. In the past couple of years, several multi-channel meeting recording corpora have sprung up from several different research sites, including the International Computer Science Institute, the National Institute of Standards and Technology, the Linguistic Data Consortium, and Carnegie Mellon University, and there have been numerous research efforts to attempt to automatically process and extract information from these recordings [1, 2, 3].

Meeting recordings are rich in information, but deficient in structure. By definition, there are a number of participants in meetings who engage in a discussion to come to some sort of conclusion or shared understanding on a number of topics by the end of the meeting. Knowledge of the topics discussed, the contributions of each participant, and the conclusions reached would be useful for the recognition and organization of meeting recordings. The automatic extraction of these pieces of knowledge, though quite simple for the human ear, is a huge challenge for the machine listening system. Meetings tend to be much less structured than other speech recognition tasks, such as broadcast news. The speech is natural and, therefore, contains many disfluencies and the recordings contain a lot of microphone noise.

In short, it is highly desirable to archive, index, and mine the recordings; however it is quite a challenge, in practice, to actually achieve these goals.

The automatic creation of meeting summaries is probably the highest-level goal for automated meeting recording understanding. This task would be helped greatly by methods for automatic extraction of topic-change points, which is what we set out to do in this paper.

We attempt to find topic change points by modeling the features of topic changes with pitch, word rate, overlapping speech, cue phrases, and other features and a support vector machine classifier.

In Section 2 we summarize the experimental data used in the experiment. Support vector machines and our feature set are described in Sections 3 and 4. In Section 5 we discuss the experiments that we have conducted and in Sections 6 and 7 we discuss the results and plans for future work.

2. EXPERIMENTAL DATA

We conduct our experiments on a subset of 25 meetings from the ICSI Meeting Recorder Corpus [1]. This set of data contains approximately 25 hours of multi-channel audio recordings of meetings with up to eight participants. Each of the participants is fitted with a high-quality, close-talking microphone. Additionally, there are 4 high-quality tabletop microphones and 2 lower-quality tabletop microphones. The meetings are hand-transcribed and include additional markings for microphone noise and human produced non-speech sounds (laughter, heavy breathing, etc.). In our experiments we use only one of the high-quality tabletop microphones and disregard the other available channels.

Ground truth topic boundaries are determined for the 25 meetings in our set by at least three annotators. Final annotations are found by majority agreement between the annotators. It is found that meetings contain 7.5 topic boundaries and 770 speaker turns on average. The topic annotation was conducted by Galley et al. [3]

3. SUPPORT VECTOR MACHINES

Support vector machines (SVMs) are relatively new tools which have received much attention in recent years due to their highly discriminative behavior and high performance on countless benchmark tests.

On a conceptual level, SVMs can be thought of as a method which tries to draw the optimal line (or hyperplane) of separation between two classes of training points by maximizing the distance from the decision boundary of the points nearest to the decision boundary.

SVMs are made practical by two additional features. The first is the introduction of slack variables, which decrease the capacity of the model, and allow for some misclassified points in the training set and increase the generalizability of the trained model. The second is the kernel trick, wherein the feature space is transformed into some nonlinear space, which enables SVMs to learn nonlinear decision boundaries. In this work we use radial basis functions as our kernel.

The mathematics behind SVMs are omitted for space. The interested reader should consult the tutorial on SVMs for more detail [4].

4. FEATURES

4.1. Pitch Features

Pitch features have been shown to be discriminating in nonlexical audio recognition tasks such as monologue topic segmentation [5] and hot-spot detection in meetings, but have not previously been used for topic segmentation in meetings.

We extract pitch estimates for each 10 ms segments of audio in the data set using the ESPS method in the Snack Sound Toolkit [6]. The pitch estimates were converted into an octave scale and then normalized according to the mean and variance of the active speaker's pitch. The active speaker was determined for each segment by aligning the transcription with the pitch estimate and choosing the currently speaking participant as the active speaker. In cases where multiple speakers were speaking, the participant who most recently began speaking is awarded the segment.

Continuous voiced segments (contiguous estimates with accuracy probabilities higher than a certain threshold) are then fit with a first-degree polynomial regression line to smooth out irregularities in the pitch estimates. The mean, variance, minimum, and maximum of the smoothed pitch estimates and the slopes of the smoothed estimates are calculated in 5 and 30 second windows before and after each candidate point.

4.2. Overlapping Speech

It has been observed that overlapping speech tends not to occur at the beginning of new topics [3]. An overlapping speech feature is calculated by examining the transcript and recording the percentage of overlapping speech segments in 30 second windows before and after each candidate point.

4.3. Rapidity of Speech

The rapidity of speech tends to increase at the start of new topics. A rapidity of speech feature is calculated by examining the transcript and counting the total number of words said per second by all participants in 30 second windows before and after each candidate point.

4.4. Unigram Cues

A number of key words have been identified as salient indicators of topic changes in this data set: "okay," "shall," "anyway," "we're," "alright," "let's," "should," "but," "so," "and," and "good." [3] The total number of occurrences of each of these terms is recorded for 30 second windows before and after each candidate point.

4.5. Duration

The percentage of voiced segments versus unvoiced segments in 30 second windows before and after each candidate point is also recorded.

5. EXPERIMENTS

We fuse our features and learn the lexical/acoustic characteristics of topic boundaries in meetings using support vector machines. The features are joined together as a 60-dimensional feature vector and used to train support vector machines. We perform 25-fold cross-validation, wherein we train a support vector machine [7] on 24 of the available meetings and test on the one held-out meeting. We do this 25 times, holding out each of the meetings once, and examine the average performance on held out meetings. In this process, we have a total of about 185 positive test points and about 20,000 negative test points.

When training the classifiers, we omit candidate points within 30 seconds of a ground-truth topic point to avoid confusing negative training points, which have features similar to positive training points due to the windowed nature of our features.

We evaluate the performance using correct accept and false alarm metrics, modified slightly to be more forgiving with small temporal errors. We define correct accept as the probability of detecting a boundary within some fixed window of a ground truth boundary point. We try several window sizes between 0.1 seconds and 30 seconds. We define false alarm as the probability of detecting a topic boundary at a candidate point, given that the candidate point is not, in reality, a topic boundary. We do not count false alarms that are within the fixed window of a topic boundary.

	30 sec	10 sec	0.1 sec
SVM	97.83%	94.57%	86.41%
Baseline	73.37%	66.85%	59.78%

Table 1. Correct acceptance rates of support vector machine and gaussian classifiers at differently-sized fuzzy windows at equal false acceptance rate of 5%.

We also build a baseline detection system using single gaussians to detect the topic boundaries. In the baseline system, we use the same features as in the SVM detector. We find single gaussian distributions for each class (topic and non-topic) by taking the sample mean and sample covariance over the feature sets. We conduct 25-fold cross-validation and evaluate the performance using variable windows.

Due to the large skew in the data (there are many more negative points than positive points) it is difficult to tune an SVM such that simply the sign of the distance away from the decision hyperplane is the best predictor of the class of a point. Instead, we evaluate our model by trying different thresholds on the distance from a test point to the hyperplane. Figure 2 shows the trade-off between correct detection and false alarm for different window sizes and different thresholds. In practice, the desired sensitivity could be chosen by cross-validation. Figure 1 shows the distance from the decision hyperplane for each testing point and the ground truth topic boundaries for a sample test meeting. Table 1 shows the correct acceptance rates for the SVM and the baseline Gaussian detectors at various window sizes and a fixed error rate of 5%.

6. DISCUSSION AND CONCLUSIONS

The feature sets described in Section 4 used with support vector machines as described in Section 5 seem to make for a highly discriminative topic boundary system. In Figure 1, we see that peaks (sometimes small, sometimes large) in the distance from the SVM decision boundary occur at nearly every ground-truth meeting boundary, showing that these points are the most unlike non-topic boundaries in the meeting at that ground truth topic boundaries are highly correlated with the distance from the SVM decision boundary.

In Figure 2, we see that within a 30 second fuzzy window, we can detect over 96% of the topic boundaries with out making a false detection error, which demonstrates the highly discriminative nature of our system.

In Table 1, we see that the SVM-based system has significant and consistent gains over the baseline Gaussian-based system, thus demonstrating that the performance gains of the SVM-based system outweigh the overhead costs of implementing and training SVMs.

In conclusion, we see that pitch, duration, cue terms, overlapping speech, and word rate, used with support vector machines results in topic detection results with up to 98% correct accept rate with a 5% false alarm rate, a significant gain in performance over simple modeling techniques, such as single-mixture Gaussian mixture models.

7. FUTURE WORK

In future work, it may be helpful to investigate the contributions of each individual feature to the overall classification performance. It will be helpful to see the gains that each feature offers to aid in decisions about the trade-off between development and training time and performance gains. And in practice, sometimes a leaner feature set will perform better than a set full of bad features.

There are also still a number of features that may be useful for topic segmentation which would be helpful to examine in future works, including lexical cohesion, or lexical self-similarity and speaker-change patterns.

Other methods of feature fusion, such as Maximum Entropy Model, may also provide performance gains in topic detection.

8. REFERENCES

[1] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, , and A. Stolcke, “The meeting project at ICSI,” in *Proc. HLT*, 2001, pp. 246–252.

[2] S. Renals and D. Ellis, “Audio information access from meeting rooms,” in *Proc. Intern. Confer. on Acoustics, Speech and Signal Processing*, Hong Kong, 2003.

[3] M. Galley, K. McKeown, E. Fosler-Lussier, H. Jing, “Discourse Segmentation of Multi-Party Conversation,” in *Proc. 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, 2003.

[4] C. Burges, “A Tutorial on Support Vector Machines for Pattern Recognition” in *Data Mining and Knowledge Discovery*, 1998.

[5] B. Arons, “Pitch-based emphasis detection for segmenting speech recordings,” in *Proc. ICSLP*, Yokohama, 1994.

[6] Snack Sound Toolkit, http://www.speech.kth.se/snack/

[7] T. Joachims, “Making large-Scale SVM Learning Practical.” *Advances in Kernel Methods - Support Vector Learning*, B. Schlkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

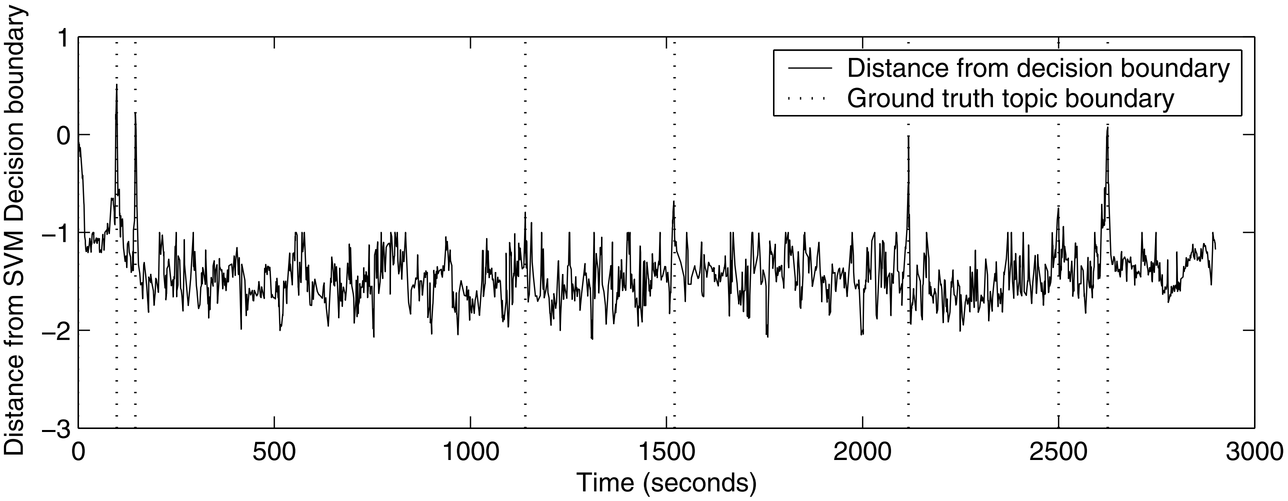

Fig. 1. Distance of test points from SVM decision boundary and topic-change points for an example test meeting.

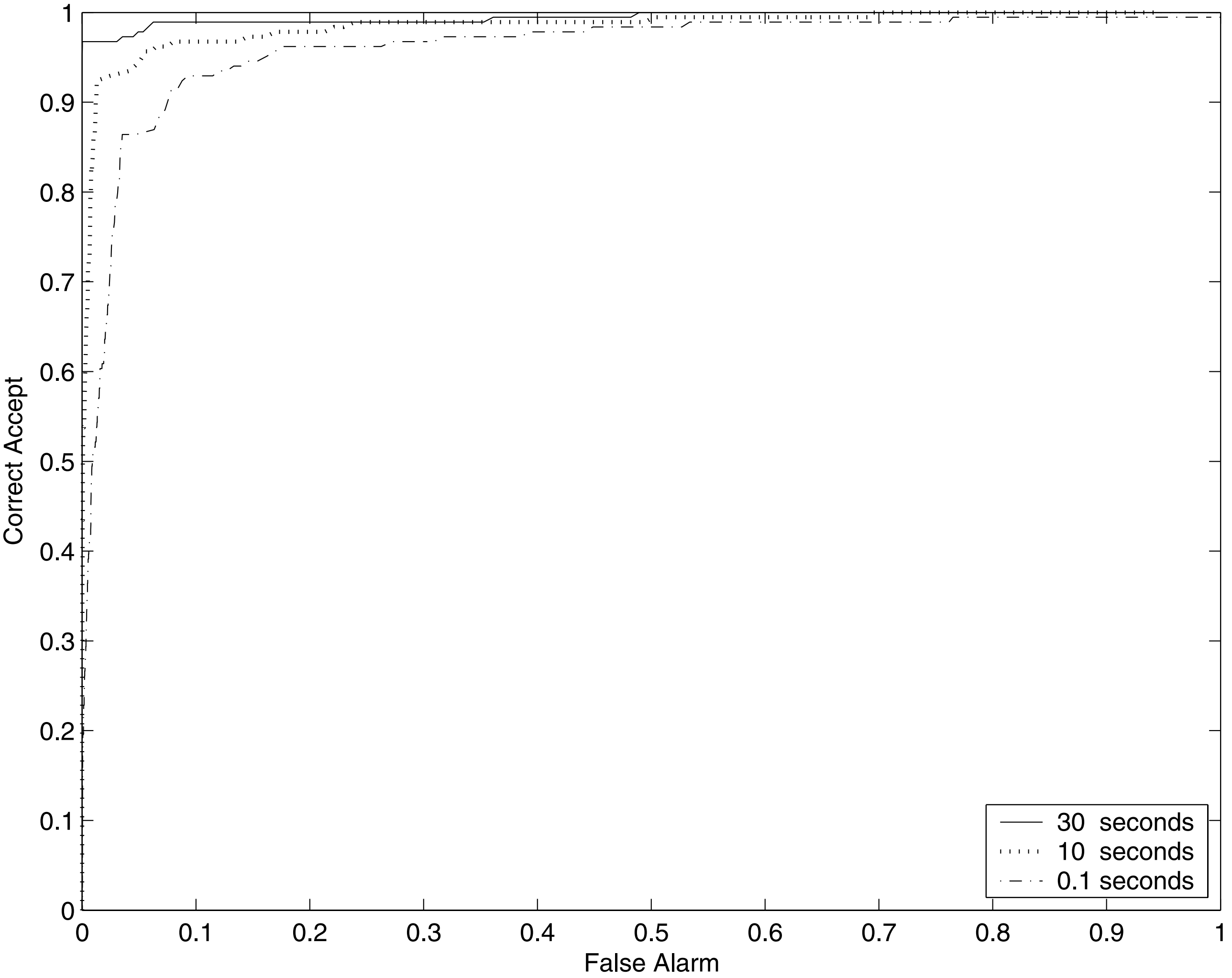

Fig. 2. Receiver operator characteristic curves for detection results from SVM classifier with different fuzzy window sizes and variable decision thresholds.