Combining Words and Speech Prosody for Automatic Topic Segmentation

Andreas Stolcke, Elizabeth Shriberg, Dilek Hakkani-Tür
Gökhan Tür, Ze'ev Rivlin, Kemal Sönmez

Speech Technology and Research Laboratory,
SRI International, Menlo Park, CA
http://www.speech.sri.com

ABSTRACT

We present a probabilistic model that uses both prosodic and lexical cues for the automatic segmentation of speech into topic units. The approach combines hidden Markov models, statistical language models, and prosody-based decision trees. Lexical information is obtained from a speech recognizer, and prosodic features are extracted automatically from speech waveforms. We evaluate our approach on the Broadcast News corpus, using standard evaluation metrics. Results show that the prosodic model alone outperforms the word-based segmentation method. Furthermore, we achieve an additional reduction in error by combining the prosodic and word-based knowledge sources.

1. Introduction

Topic segmentation deals with the problem of automatically dividing a stream of text or speech into topically homogeneous blocks [1]. That is, given a sequence of (written or spoken) words, the aim is to find the boundaries where topics change. Topic segmentation is an important task for various language understanding applications, such as information extraction and retrieval, and text summarization. In this paper, we present our work on fully automatic detection of topic boundaries from speech input.

Past automatic topic segmentation systems have depended mostly on lexical information [6, 4, 1, 16, among others]. One problem for applying the text-based approach to speech input is the lack of typographic cues (such as headers, paragraphs, sentence punctuation and capitalization). On the other hand, speech provides an additional, nonlexical knowledge source through its durational, intonational, and energy characteristics, i.e., its *prosody*.

Prosodic cues are known to be relevant to discourse structure in spontaneous speech [8, 7, 14, among others], and can therefore be expected to play a role in indicating topic transitions. Furthermore, prosodic cues by their nature are relatively unaffected by word identity, and should therefore improve the robustness of lexical topic segmentation methods based on automatic speech recognition.

Past segmentation studies involving prosodic information have generally relied on hand-coded cues (with the notable exception of [5]). We therefore believe the present work to be the first that combines fully automatic extraction of both lexical and prosodic information for topic segmentation. Furthermore, we have adopted the strict evaluation paradigm used by the government-administered TDT-2 (Topic Detection and Tracking Phase 2) [15] program, allowing fair comparisons of various approaches both within this study and in relation to other work. The general framework for combining lexical and prosodic cues for tagging speech with various kinds of "hidden" structural information is a further development of our earlier work on sentence segmentation and disfluency detection for spontaneous speech [10, 12, 13].

2. Approach

Topic segmentation in the paradigm used by us and others [15] proceeds in two phases. In the first phase, the input is divided into contiguous strings of words assumed to belong to one topic each. We refer to this step as "chopping". For example, in textual input, the natural units for chopping are sentences (as can be inferred from punctuation and capitalization). For continuous speech input, the choices are less obvious; we compare several possibilities in our experimental evaluation. Here, for simplicity, we will use "sentence" to refer to units of chopping, regardless of the criterion used. In the second phase, the sentences are further grouped into contiguous stretches belonging to one topic, i.e., the sentence boundaries are classified into "topic boundaries" and "nontopic boundaries".[1]

Topic segmentation is thus reduced to a boundary classification problem. We will use T to denote the string of binary boundary classifications. Furthermore, our two knowledge sources are the (chopped) word sequence W and the stream of prosodic features F. Our approach aims to find the classification T with highest probability given the information in W and F

$$\operatorname*{argmax}_T P(T|W, F)$$

using statistical modeling techniques. In the following sections, we describe each of the elements of the overall model in turn: first, a model of the dependency between prosody F and topic segmentation T; second, a model relating words W and T; and finally, an approach for combining the models.

2.1. Prosodic Model

For modeling topic boundaries prosodically we used a wide range of features that were automatically extracted from the data. Let F_i be the features extracted from a window around the ith potential topic boundary (chopping boundary), and let T_i be the boundary type (boundary/no-boundary) at that position. We trained CART-style decision trees [2] to predict the ith boundary type, i.e., to estimate $P(T_i|F_i, W)$. The decision is only weakly conditioned on the word sequence W, insofar as some of the prosodic features depend on the phonetic alignment of the word models. We can thus expect the prosodic model estimates to be robust to recognition errors.

For training, we automatically aligned and extracted features from 70 hours of the Linguistic Data Consortium (LDC) 1997 Broadcast

[1] We do not consider the problem of detecting recurring, discontinuous instances of the same topic, a task known as "topic tracking" in the TDT paradigm.

News (BN) corpus. Topic boundary information determined by human labelers was extracted from the markup accompanying the word transcripts of this corpus.

We started with a large set of prosodic features capturing various durational and intonational aspects of speech prosody, as in [10]. We included features that, based on descriptive literature, we believed should reflect breaks in the temporal and intonational contour. We developed versions of such features that could be defined at each inter-word boundary, and which could be extracted by completely automatic means (no human labeling). Furthermore, the features were designed to the extent possible to be independent of word identities, for use with recognizer output.

The greedy nature of the decision tree learning algorithm implies that larger initial feature sets can give worse results than smaller subsets. Furthermore, it is desirable to remove redundant features for computational efficiency and to simplify interpretation of results. For this purpose we developed an iterative feature selection algorithm to find useful task-specific feature subsets. The algorithm combined elements of brute-force search with previously determined heuristics about good groupings of features. We used the entropy reduction of the overall tree after cross-validation, as a method for selecting a good set of features. Entropy reduction is the difference in test-set entropy between the prior class distribution and the posterior distribution estimated by the tree; it is a more fine-grained metric than classification accuracy, and is also more relevant to the model combination approach described later. The algorithm proceeds in two phases: in the first phase, the number of features is reduced, checking the effect of each feature on the performance by leaving out one feature at a time. The second phase then starts with the reduced number of features, and performs a beam search over all possible subsets of features. The decision tree paradigm also allows us to add, and automatically select, other (nonprosodic) features that might be relevant to the task.

We started with a set of 73 potential features. The iterative algorithm reduced this to a set of 5 features helpful for our task. Upon inspection, the following characteristics are modeled by the tree. We provide for each characteristic the relative frequency with which associated features are queried in the final decision tree; this gives an approximate indication of feature importance.

1. F0 differences across the boundary (44.0%). Several features compare the F0 following the boundary to F0 before the boundary. The F0s are measured over the duration of the words adjacent to the boundary, or over a fixed length window of 200 milliseconds. Values are either mean F0, or minimum/maximum F0, in the regions surrounding the boundary. The mean captures a range effect; the minimum and maximum values make the measure more sensitive to local variation, such as rising to accented syllables, and final pitch falls. Rather than using raw pitch tracks, all F0 features are based on an explicit model of pitch-halving/doubling, using straight-line stylizations for improved robustness [11].
2. Pause duration (36.3%). The duration of the nonspeech interval occurring at the boundary.[2]
3. Speaker change (15.5%). Whether or not a speaker change occurred at the boundary.
4. Gender (4.2%). We found stylistic differences between males and females in the use of F0 at topic boundaries. This is true even after proper normalization, e.g., equating the gender-specific non-topic boundary distributions. Additionally, we noted that non-topic pauses (i.e., chopping boundaries) are more likely to occur in male speech, a phenomenon that could have several causes and awaits further analysis.[3]

2.2. Language Model

For word-based modeling, we use standard language models and a hidden Markov model (HMM) based tagger. Similar to the Dragon HMM segmentation approach [16], we built an HMM, in which the states are topic clusters, and the observations are sentences (or chopped units). The resulting HMM forms a complete graph, allowing transition between any two topic clusters. The exact number of topic clusters is not important, as long as it is large enough to make two adjacent topics in the same cluster unlikely. The observation likelihoods for the HMM state represent the probability of generating a given sentence in a particular topic. The likelihoods are computed from unigram language models trained on the clusters, which are determined automatically using an unsupervised clustering algorithm, on the training data. All transitions within the same topic are given probability 1, while all transitions between topics are set to a global *topic switch penalty*, which is optimized on held-out training data. This parameter enables us to trade off between false alarms and misses. Once the HMM is trained, we use the Viterbi algorithm to search for the best state sequence and corresponding segmentation.

In addition to the basic HMM segmenter developed by Dragon, we incorporated two additional states, for modeling the initial and final sentences of a topic segment. We reasoned that this approach can capture formulaic speech patterns used by broadcast speakers. Likelihoods for the start and end models are obtained as the unigram language model probabilities of the topic-initial and final sentences, respectively, in the training data. Note that a single start and end state are shared for all topics. Also, traversal of the initial and final states is optional in the HMM topology. We observed a 5% relative reduction in segmentation error over the baseline HMM topology using initial and final states. Because the topic-initial and final states are optional, our training of this model is probably suboptimal. Instead of labeling all topic-initial and final training sentences as data for the corresponding states, we should be training the model by using repeated forced alignments to find actual good examples of initial and final sentences (an approximate version of expectation-maximization [3]).

While constructing the topic language models, we used the pooled TDT Pilot and TDT-2 training data, which covers the transcriptions of Broadcast News from January 1992 through June 1994 and from January 1998 through February 1998, respectively (this corpus is distinct from the 1997 BN acoustic corpus used for prosodic model training and overall testing). We removed stories with fewer than

[2] The importance of pause duration is actually underestimated by this measure of feature use; as explained later, pause durations are already used during the chopping process, so that the decision tree is applied only to boundaries exceeding a certain duration. Separate experiments using boundaries below our chopping threshold show that the tree makes use of shorter pauses for segmentation decisions as well.

[3] For example, it could be that male speaker in BN are assigned longer topic segments on average, or that male speaker are more prone to pausing in general, or that males dominate the spontaneous speech portions where pausing is naturally more frequent.

300 and more than 3000 words, leaving 19,916 stories with an average length of 538 words without any stop words. Then we automatically constructed 100 topic language models, using the multipass *k*-means algorithm described in [16]. We did not smooth the individual topic language models, but instead interpolated them with the global unigram language model, which gave better results.

2.3. Model Combination

The word-based HMM was modified to use probabilities from the decision tree estimator as additional likelihood scores, with an empirically optimized weighting. To this end, we inserted a fictitious *boundary* observation between adjacent sentences, and introduced two more "boundary" states into the HMM topology. Between sentences, the model must pass one of the boundary states, denoting either the presence or absence of a topic boundary.

Likelihoods $P(F_i|T_i)$ for the boundary states are obtained from the prosodic model. The decision tree posterior probabilities must be converted to likelihoods, either by dividing them by priors or by training the decision trees on a balanced training set. We preferred the resampling method, so the following equations hold:

$$P(T_i|F_i) = \frac{P(F_i|T_i)P(T_i)}{P(F_i)} \propto P(F_i|T_i)P(T_i) \propto P(F_i|T_i) \quad .$$

Note $P(F_i)$ is a constant for different T_i, and $P(T_i) = 0.5$ by virtue of resampling.

3. Experiments and Results

Various models were evaluated on three hours (6 shows) from the 1997 BN corpus. To make best use of the available test data, we used a two-fold jack-knifing procedure to tune the model parameters (topic switch penalties, and model combination weights): parameters were tuned on each of two halves of the data, and then tested on the respective other half. Reported results represent the averages of these two trials. The error rates obtained in all experiments are according to the procedures set out in the DARPA Topic Detection and Tracking Project [15], with the NIST-TDT evaluation software. They represent a weighted detection error, using a particular choice of costs for false alarms and misses.

Two test conditions were used: forced alignments using the true words, and recognized words as obtained using a simplified version of the SRI Broadcast News recognizer [9], with a word error rate of 29%. We first present baseline results with word information alone, followed by results for the prosodic model and the combined model.

3.1. Chopping and Segmentation by Language Model

Unlike written text, the output of the automatic speech recognizer contains no sentence boundaries. Therefore, grouping words into (pseudo-)sentences (chopping) is a nontrivial problem while processing speech. Some pre-segmentation into roughly sentence-length units is necessary since otherwise the observations associated with HMM states are too inhomogeneous with regard to topic choice, causing very poor performance.

We investigated fixed-length blocks (based on number of words), turn boundaries (speaker change locations), pauses, and, for reference, actual sentence boundaries obtained from the transcripts, as chopping criteria. Table 1 gives the error rates for the four conditions, using the true word transcripts for testing. For the PAUSE condition, we empirically determined an optimal minimum pause duration threshold to use. Specifically, we considered pauses exceeding 0.66 second as potential topic boundaries in this (and all later) experiment. For the FIXED condition, a block length of 10 words was found to work best.

Chopping Criterion	Error Rate on Forced Alignments
FIXED	19.84%
TURN	22.78%
SENTENCE	20.56%
PAUSE	19.50%

Table 1: Error rates with various chopping criteria.

We conclude that a simple prosodic feature, pause duration, is an excellent criterion for the chopping step, working as well as or better than standard sentence boundaries.

As a side issue in our experiments, we wanted to verify that our test data (from the 1997 BN corpus) was comparable in difficulty to the official test corpus of the 1998 TDT-2 evaluations, for which we had only recognizer output (from a different system) available. Table 2 shows that the two test sets exhibit very similar results, justifying our use of the 1997 BN corpus for practical reasons.[4]

Test set	Error Rate on Forced Alignments	Error Rate on Recognized Words
TDT-2	NA	20.40%
BN'97	19.50%	20.86%

Table 2: Error rates using different corpora.

3.2. Segmentation using Prosody and Combined Models

Table 3 gives our results with forced alignments and recognized words for each of the individual models and the combined model. As shown, the error rate for the prosody model alone is lower than that for the language model, and combining both models gives further improvement. With the combined model, the error rate decreased by 22.97% relative to the language model, for the correct words, and by 19.27% for recognized words.

As discussed earlier, the results with the language model alone make use of prosody in the chopping step.

4. Summary and Discussion

Results so far indicate that prosodic information provides an excellent source of information for automatic topic segmentation, both by

[4] In particular, we chose the 1997 BN corpus because of the availability of detailed annotated transcripts for a variety of other tasks (such as sentence segmentation and named entities) that are the subject of current lexical-prosodic modeling research at SRI.

Model	Error Rate on Forced Alignments	Error Rate on Recognized Words
LM Only	19.50%	20.86%
Prosody Only	18.87%	19.85%
Combined	15.02%	16.84%

Table 3: Summary of error rates with individual and combined models, using pause duration as a chopping criterion.

itself and in conjunction with lexical information. Pause duration, a simple prosodic feature that is readily available as a by-product of speech recognition, proved extremely effective in the initial chopping phase, as well as being the most important feature used by prosodic decision trees. Additional prosodic features based on pitch were are also found to be relevant (and feasible) for automatic segmentation.

The HMM-based lexical topic segmentation approach [16] is easily extended to incorporate the decision tree posterior probabilities (as long as the tree is trained on a uniform prior distribution). The fact that the model combination gives a significant win indicates that the lexical and prosodic knowledge sources are sufficiently complementary for this simple combination approach (which assumes statistical independence).

The results obtained with recognized words (at a 29% word error rate) did not differ greatly from those obtained with correct word transcripts (7% error increase with LM, 5% error increase with prosody). Still, part of the appeal of prosodic segmentation is that it is inherently robust to recognition errors. This characteristic makes it even more attractive for use in domains with higher error rates due to poor acoustic conditions or more conversational speaking styles.

Several aspects of our system are suboptimal. For example, we have not yet optimized the chopping stage relative to the combined model (only relative to the lexical-only segmenter). Also, the use of prosodic features other than just pause should further improve the overall performance. Ultimately, we want to eliminate the need to separate chopping and HMM classification stages, which is both theoretically unappealing and inconvenient in the optimization of the overall system.

5. Conclusion

We have presented our work on automatic topic segmentation from speech, using a combination of lexical and prosodic cues. Our results show that the prosodic model alone outperforms the word-based segmentation method, and an additional reduction in error can be achieved by combining the lexical and prosodic models.

Acknowledgments

We thank Becky Bates and Ananth Sankar for invaluable assistance in preparing the data for this study, as well as for many helpful discussions. This research was supported by DARPA and NSF under NSF grant IRI-9619921, and DARPA contract no. N66001-97-C-8544. The views herein are those of the authors and should not be interpreted as representing the policies of the funding agencies.

References

1. J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. In *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, pp. 194–218, Lansdowne, VA, 1998.
2. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Pacific Grove, CA, 1984.
3. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society, Series B*, 34:1–38, 1977.
4. M. A. Hearst. TexTiling: Segmenting text info multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.
5. J. Hirschberg and C. Nakatani. A prosodic analysis of discourse segments in direction-giving monologues. In *Proc. ACL*, pp. 286–293, Santa Cruz, CA, 1996.
6. H. Kozima. Text segmentation based on similarity between words. In *Proc. ACL*, pp. 286–288, Ohio State University, Columbus, Ohio, 1993.
7. D. J. Litman and R. J. Passonneau. Combining multiple knowledge sources for discourse segmentation. In *Proc. ACL*, pp. 108–115, MIT, Cambridge, MA, 1995.
8. S. Nakajima and J. F. Allen. A study on prosody and discourse structure in cooperative dialogues. *Phonetica*, 50:197–210, 1993.
9. A. Sankar, F. Weng, Z. Rivlin, A. Stolcke, and R. R. Gadde. The development of SRI's 1997 Broadcast News transcription system. In *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, pp. 91–96, Lansdowne, VA, 1998.
10. E. Shriberg, R. Bates, and A. Stolcke. A prosody-only decision-tree model for disfluency detection. In G. Kokkinakis, N. Fakotakis, and E. Dermatas, editors, *Proc. EUROSPEECH*, vol. 5, pp. 2383–2386, Rhodes, Greece, 1997.
11. K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub. Modeling dynamic prosodic variation for speaker verification. In R. H. Mannell and J. Robert-Ribes, editors, *Proc. ICSLP*, vol. 7, pp. 3189–3192, Sydney, 1998. Australian Speech Science and Technology Association.
12. A. Stolcke and E. Shriberg. Automatic linguistic segmentation of conversational speech. In H. T. Bunnell and W. Idsardi, editors, *Proc. ICSLP*, vol. 2, pp. 1005–1008, Philadelphia, 1996.
13. A. Stolcke, E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tür, and Y. Lu. Automatic detection of sentence boundaries and disfluencies based on recognized words. In R. H. Mannell and J. Robert-Ribes, editors, *Proc. ICSLP*, vol. 5, pp. 2247–2250, Sydney, 1998. Australian Speech Science and Technology Association.
14. M. Swerts and M. Ostendorf. Prosodic and lexical indications of discourse structure in human-machine interactions. *Speech Communication*, 22(1):25–41, 1997.
15. The Topic Detection and Tracking Phase 2 (TDT2) evaluation plan. available from `http://www.nist.gov/speech/tdt98/tdt98.htm`, 1998.
16. J. Yamron, I. Carp, L. Gillick, S. Lowe, and P. van Mulbregt. A hidden Markov model approach to text segmentation and event tracking. In *Proc. ICASSP*, vol. 1, pp. 333–336, Seattle, WA, 1998.