Assessing Prosodic and Text Features for Segmentation of Mandarin Broadcast News

Gina-Anne Levow
University of Chicago
levow@cs.uchicago.edu

Abstract

Automatic topic segmentation, separation of a discourse stream into its constituent stories or topics, is a necessary preprocessing step for applications such as information retrieval, anaphora resolution, and summarization. While significant progress has been made in this area for text sources and for English audio sources, little work has been done in automatic segmentation of other languages using both text and acoustic information. In this paper, we focus on exploiting both textual and prosodic features for topic segmentation of Mandarin Chinese. As a tone language, Mandarin presents special challenges for applicability of intonation-based techniques, since the pitch contour is also used to establish lexical identity. However, intonational cues such as reduction in pitch and intensity at topic boundaries and increase in duration and pause still provide significant contrasts in Mandarin Chinese. We first build a decision tree classifier that based only on prosodic information achieves boundary classification accuracy of 89-95.8% on a large standard test set. We then contrast these results with a simple text similarity-based classification scheme. Finally we build a merged classifier, finding the best effectiveness for systems integrating text and prosodic cues.

1 Introduction

Natural spoken discourse is composed of a sequence of utterances, not independently generated or randomly strung together, but rather organized according to basic structural principles. This structure in turn guides the interpretation of individual utterances and the discourse as a whole. Formal written discourse signals a hierarchical, tree-based discourse structure explicitly by the division of the text into chapters, sections, paragraphs, and sentences. This structure, in turn, identifies domains for interpretation; many systems for anaphora resolution rely on some notion of locality (Grosz and Sidner, 1986). Similarly, this structure represents topical organization, and thus would be useful in information retrieval to select documents where the primary sections are on-topic, and, for summarization, to select information covering the different aspects of the topic.

Unfortunately, spoken discourse does not include the orthographic conventions that signal structural organization in written discourse. Instead, one must infer the hierarchical structure of spoken discourse from other cues. Prior research (Nakatani et al., 1995; Swerts, 1997) has shown that human labelers can more sharply, consistently, and confidently identify discourse structure in a word-level transcription when an original audio recording is available than they can on the basis of the transcribed text alone. This finding indicates that substantial additional information about the structure of the discourse is encoded in the acoustic-prosodic features of the utterance. Given the often errorful transcriptions available for large speech corpora, we choose to focus here on fully exploiting the prosodic cues to discourse structure present in the original speech in addition to possibly noisy textual cues. We then compare the effectiveness of a pure prosodic classification to text-based and mixed text and prosodic based classification.

In the current set of experiments, we concentrate on sequential segmentation of news broadcasts into individual stories. This level of segmentation can be most reliably performed by human labelers and thus can be considered most robust, and segmented data sets are publicly available.

Furthermore, we consider the relative effectiveness prosodic-based, text-based, and mixed cue-based seg-

mentation for Mandarin Chinese, to assess the relative utility of the cues for a tone language. Not only is the use of prosodic cues to topic segmentation much less well-studied in general than is the use of text cues, but the use of prosodic cues has been largely limited to English and other European languages.

2 Related Work

Most prior research on automatic topic segmentation has been applied to clean text only and thus used textual features. Text-based segmentation approaches have utilized term-based similarity measures computed across candidate segments (Hearst, 1994) and also discourse markers to identify discourse structure (Marcu, 2000).

The Topic Detection and Tracking (TDT) evaluations focused on segmentation of both text and speech sources. This framework introduced new challenges in dealing with errorful automatic transcriptions as well as new opportunities to exploit cues in the original speech. The most successful approach (Beeferman et al., 1999) produced automatic segmentations that yielded retrieval results comparable to those with manual segmentations, using text and silence features. (Tur et al., 2001) applied both a prosody-only and a mixed text-prosody model to segmentation of TDT English broadcast news, with the best results combining text and prosodic features. (Hirschberg and Nakatani, 1998) also examined automatic topic segmentation based on prosodic cues, in the domain of English broadcast news.

Work in discourse analysis (Nakatani et al., 1995; Swerts, 1997) in both English and Dutch has identified features such as changes in pitch range, intensity, and speaking rate associated with segment boundaries and with boundaries of different strengths.

3 Data Set

We utilize the Topic Detection and Tracking (TDT) 3 (Wayne, 2000) collection Mandarin Chinese broadcast news audio corpus as our data set. Story segmentation in Mandarin and English broadcast news and newswire text was one of the TDT tasks and also an enabling technology for other retrieval tasks. We use the segment boundaries provided with the corpus as our gold standard labeling. Our collection comprises 3014 stories drawn from approximately 113 hours over three months (October-December 1998) of news broadcasts from the Voice of America (VOA) in Mandarin Chinese. The transcriptions span approximately 740,000 words. The audio is stored in NIST Sphere format sampled at 16KHz with 16-bit linear encoding.

4 Prosodic Features

We employ four main classes of prosodic features: pitch, intensity, silence and duration. Pitch, as represented by f0 in Hertz, was computed by the "To pitch" function of the Praat system (Boersma, 2001). We then applied a 5-point median filter to smooth out local instabilities in the signal such as vocal fry or small regions of spurious doubling or halving. Analogously, we computed the intensity in decibels for each 10ms frame with the Praat "To intensity" function, followed by similar smoothing.

For consistency and to allow comparability, we computed all figures for word-based units, using the ASR transcriptions provided with the TDT Mandarin data. The words are used to establish time spans for computing pitch or intensity mean or maximum values, to enable durational normalization and pairwise comparison, and to identify silence duration.

It is well-established (Ross and Ostendorf, 1996) that for robust analysis pitch and intensity should be normalized by speaker, since, for example, average pitch is largely incomparable for male and female speakers. In the absence of speaker identification software, we approximate speaker normalization with story-based normalization, computed as $\frac{val-mean}{mean}$, assuming one speaker per topic[1]. For duration, we consider both absolute and normalized word duration, where average word duration is used as the mean in the calculation above.

Mandarin Chinese is a tone language in which lexical identity is determined by a pitch contour - or *tone* - associated with each syllable. This additional use of pitch raises the question of the cross-linguistic applicability of the prosodic cues, especially pitch cues, identified for non-tone languages. Specifically, do we find intonational cues in tone languages?

We have found highly significant differences based on paired t-test two-tailed, ($df = 1140, p < 0.0025$) for words in segment-final position, relative to the same word in non-final positions. (Levow, 2004). Specifically, word duration, normalized mean pitch, and normalized mean intensity all differ significantly for words in topic-final position relative to occurrences throughout the story. Word duration increases, while both pitch and intensity decrease. Importantly, reduction in pitch as a signal of topic finality is robust across the typological contrast of tone and non-tone languages, such as English (Nakatani et al., 1995) and Dutch (Swerts, 1997).

[1]This is an imperfect approximation as some stories include off-site interviews, but seems a reasonable choice in the absence of automatic speaker identification.

5 Classification

5.1 Prosodic Feature Set

The contrasts above indicate that duration, pitch, and intensity should be useful for automatic prosody-based identification of topic boundaries. To facilitate cross-speaker comparisons, we use normalized representations of average pitch, average intensity, and word duration. These features form a word-level context-independent feature set.

Since segment boundaries and their cues exist to contrastively signal the separation between topics, we augment these features with local context-dependent measures. Specifically, we add features that measure the change between the current word and the next word. [2] This contextualization adds four contextual features: change in normalized average pitch, change in normalized average intensity, change in normalized word duration, and duration of following silence.

5.2 Text Feature Set

In addition to the prosodic features, we also consider a set of features that exploit the textual similarity of regions to identify segment boundaries. We build on the standard information retrieval measures for assessing text similarity. Specifically we consider a $tf * idf$ weighted cosine similarity measure across 50 and 30 word windows. We also explore a length normalized word overlap within the same region size. We use the words from the ASR transcription as our terms and perform no stopword removal. We expect that these measures will be minimized at topic boundaries where changes in topic are accompanied by changes in topical terminology.

5.3 Classifier Training and Testing Configuration

We employed Quinlan's C4.5 (Quinlan, 1992) decision tree classifier to provide a readily interpretable classifier. Now, the vast majority of word positions in our collection are non-topic-final. So, in order to focus training and test on topic boundary identification, we downsample our corpus to produce training and test sets with a 50/50 split of topic-final and non-topic-final words. We trained on 2789 topic-final words [3] and 2789 non-topic-final words, not matched in any way, drawn randomly from the full corpus. We tested on a held-out set of 200 topic-final and non-topic-final words.

[2] We have posed the task of boundary detection as the task of finding segment-final words, so the technique incorporates a single-word lookahead. We could also repose the task as identification of topic-initial words and avoid the lookahead to have a more on-line process. This is an area for future research.

[3] We excluded a small proportion of words for which the pitch tracker returned no results.

5.4 Classifier Evaluation

5.4.1 Prosody-only Classification

The resulting classifier achieved 95.8% accuracy on the held-out test set, closely approximating pruned tree performance on the training set. This effectiveness is a substantial improvement over the sample baseline of 50%. Inspection of the classifier indicates the key role of silence as well as the use of both contextual and purely local features of both pitch and intensity. Durational features play a lesser role in the classifier.

5.4.2 Text and Silence-based Classification

In a comparable experiment, we employed only the text similarity and silence duration features to train and test the classifier. These features similarly achieved a 95.5% overall classification accuracy. Here the best classification accuracy was achieved by the text similarity measure that was based on the $tf * idf$ weighted 50 word window. The text similarity measures based on $tf * idf$ in the 30 word window and on length normalized overlap performed similarly. The combination of all three text-based features did not improve classification over the single best measure.

5.4.3 Combined Prosody and Text Classification

Finally we built a combined classifier integrating all prosodic and textual features. This classifier yielded an accuracy of 97%, the best overall effectiveness. The decision tree utilized all classes of prosodic features and performed comparably with only the $tf * idf$ features and with all text features. A portion of the tree is reproduced in Figure 1.

5.5 Feature Comparison

We also performed a set of contrastive experiments with different subsets of available features to assess the dependence on these features. [4] We grouped features into 5 sets: pitch, intensity, duration, silence, and text-similarity. For each of the prosody-only, text-only, and combined prosody and text-based classifiers, we successively removed the feature class at the root of the decision tree and retrained with the remaining features (Table 1).

We observe that although silence duration plays a very significant role in story boundary identification for all feature sets, the richer prosodic and mixed text-prosodic classifiers are much more robust to the absence of silence information. Further we observe that intensity and then pitch play the next most important roles in classification.

[4] For example, VOA Mandarin has been observed stylistically to make idiosyncratically large use of silence at story boundaries. (personal communication, James Allan).

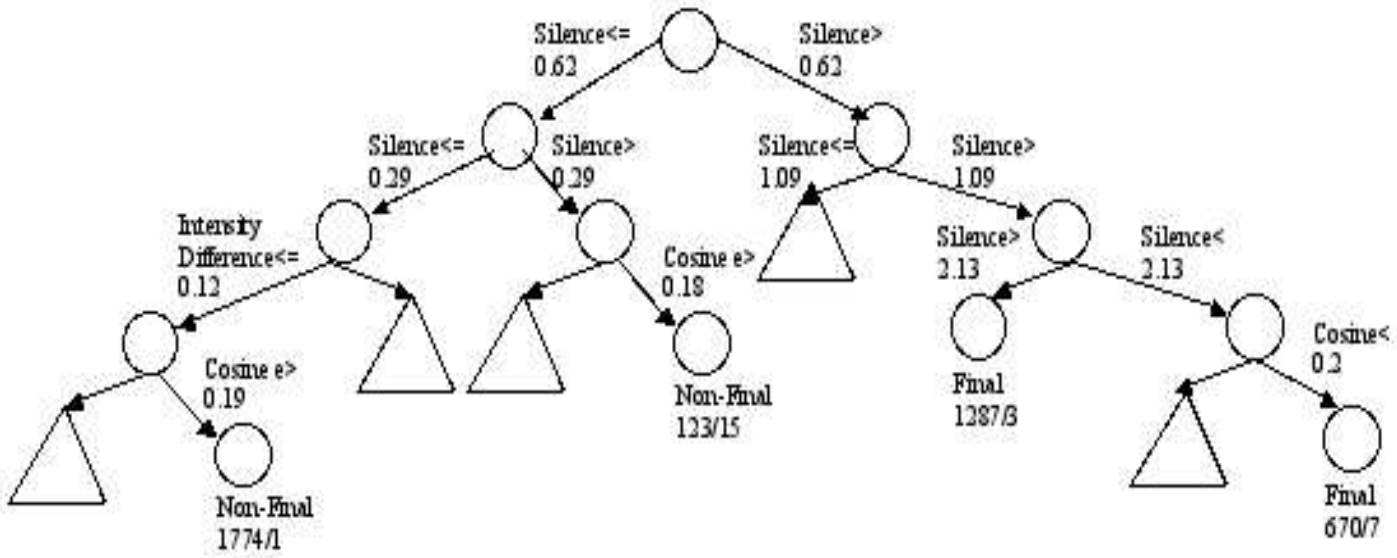

Figure 1: Decision tree classifier labeling words as segment-final or non-segment-final, using text and prosodic features

	Prosody-only		Text+Silence		Text+Prosody	
	Accuracy	Pct. Change	Accuracy	Pct. Change	Accuracy	Pct. Change
All	95.8%	0	95.5%	0	97%	0
Silence	89.4%	-6.7%	75.5%	-21%	91.5%	5.7%
Intensity	82.2%	-14.2%			86.4%	-11%
Pitch	64%	-33.2%			77%	-20.6%

Table 1: Reduction in classification accuracy with removal of features. Each row is labeled with the feature that is newly removed from the set of available features.

6 Conclusion and Future Work

We have demonstrated the utility of prosody-only, text-only, and mixed text-prosody features for automatic topic segmentation of Mandarin Chinese. We have demonstrated the applicability of intonational prosodic features, specifically pitch, intensity, pause and duration, to the identification of topic boundaries in a tone language. We observe similar effectiveness for all feature sets when all features are available, with slightly better classification accuracy for the hybrid text-prosody approach. These results indicate a synergistic combination of meaning and acoustic features. We further observe that the prosody-only and hybrid feature sets are much less sensitive to the absence of individual features, and, in particular, to silence features. These findings indicate that prosodic features are robust cues to topic boundaries, both with and without textual cues.

There is still substantial work to be done. We would like to integrate speaker identification for normalization and speaker change detection. We also plan to explore the integration of text and prosodic features for the identification of more fine-grained sub-topic structure, to provide more focused units for information retrieval, summarization, and anaphora resolution. We also plan to explore the interaction of prosodic and textual features with cues from other modalities, such as gaze and gesture, for robust segmentation of varied multi-modal data.

References

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

P. Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9–10):341–345.

B. Grosz and C. Sidner. 1986. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

M. Hearst. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*.

Julia Hirschberg and Christine Nakatani. 1998. Acoustic indicators of topic segmentation. In *Proceedings on ICSLP-98*.

Gina-Anne Levow. 2004. Prosody-based topic segmentation for mandarin broadcast news. In *Proceedings of HLT-NAACL 2004, Volume 2*.

D. Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press.

C. H. Nakatani, J. Hirschberg, and B. J. Grosz. 1995. Discourse structure in spoken language: Studies on speech corpora. In *Working Notes of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 106–112.

J.R. Quinlan. 1992. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

K. Ross and M. Ostendorf. 1996. Prediction of abstract labels for speech synthesis. *Computer Speech and Language*, 10:155–185.

Marc Swerts. 1997. Prosodic features at discourse boundaries of different strength. *Journal of the Acoustical Society of America*, 101(1):514–521.

G. Tur, D. Hakkani-Tur, A. Stolcke, and E. Shriberg. 2001. Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics*, 27(1):31–57.

C. Wayne. 2000. Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. In *Language Resources and Evaluation Conference (LREC) 2000*, pages 1487–1494.