Prosody-based Topic Segmentation for 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, acoustic feature-based segmentation of other languages. In this paper, we focus on prosody-based 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. We demonstrate that 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 also build a decision tree classifier that, based only on word and local context prosodic information without reference to term similarity, cue phrase, or sentence-level information, achieves boundary classification accuracy of 89-95.8% on a large standard test set.

1 Introduction

Natural spoken discourse is composed 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, rather than on the lexical cues or term frequencies of the transcription.

In the current set of experiments, we concentrate on sequential segmentation of news broadcasts into individual stories. While a richer hierarchical segmentation is ultimately desirable, sequential story segmentation provides a natural starting point. This level of segmentation can also be most reliably performed by human labelers and thus can be considered most robust, and segmented data sets are publicly available.

Furthermore, we apply prosodic-based segmentation to Mandarin Chinese. 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. They also demonstrated that access to acoustic cues improves the ease and quality of human labeling.

3 Prosody and Mandarin

In this paper we focus on topic segmentation in Mandarin Chinese broadcast news. 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? The fact that emphasis is marked intonationally by expansion of pitch range even in the presence of Mandarin lexical tone (Shen, 1989) suggests encouragingly that prosodic, intonational cues to other aspects of information structure might also prove robust in tone languages.

4 Prosodic Features

We consider four main classes of prosodic features for our analysis and classification: 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 selected the highest ranked pitch candidate value in each voiced region. 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 compute 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 the pairwise comparisons reported below, 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.

5 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.

6 Prosodic Analysis

To evaluate the potential applicability of prosodic features to story segmentation in Mandarin Chinese, we performed some initial data analysis to determine if words in story-final position differed from the same words used throughout the story. This lexical match allows direct pairwise comparison. We anticipated that since words in Mandarin varied not only in phoneme sequence but also in tone sequence, a direct comparison might be particularly important to eliminate sources of variability. Features that differed significantly would form the basis of our classifier feature set.

[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.

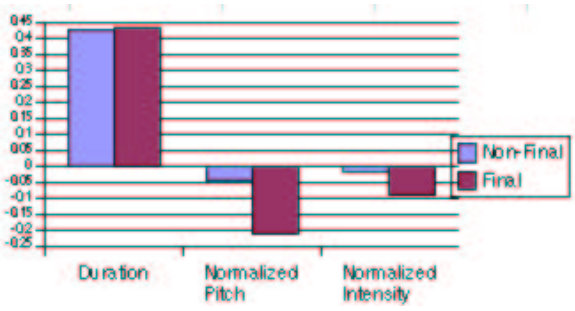

Figure 1: Differences in duration, normalized pitch, and normalized intensity between words in segment non-final and segment-final positions.

We found highly significant differences based on paired t-test two-tailed, ($df = 1140, p < 0.0025$), for each of the features we considered. Specifically, word duration, normalized mean pitch, and normalized mean intensity all differed significantly for words in topic-final position relative to occurrences throughout the story (Figure 1). Word duration increased, while both pitch and intensity decreased. A small side experiment using 15 hours of English broadcast news from the TDT collection shows similar trends, though the magnitude of the change in intensity is smaller than that observed for the Chinese.

These contrasts are consistent with, though in some cases stronger than, those identified for English (Nakatani et al., 1995) and Dutch (Swerts, 1997). The relatively large size of the corpus enhances the salience of these effects. We find, importantly, that reduction in pitch as a signal of topic finality is robust across the typological contrast of tone and non-tone languages. These findings demonstrate highly significant intonational effects even in tone languages and suggest that prosodic cues may be robust across wide ranges of languages.

7 Classification

7.1 Feature Set

The results 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. We also include absolute 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.

7.2 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.

7.3 Classifier Evaluation

The resulting classifier achieved 95.8% accuracy on the held-out test set, closely approximately pruned tree performance on the training set. This effectiveness is a substantial improvement over the sample baseline of 50%. [4] A portion of the decision tree is reproduced in Figure 2. Inspection of the tree 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. The classifier relies on the theoretically and empirically grounded features of pitch and intensity and silence, where it has been suggested that higher pitch and wider range are associated with topic initiation and lower pitch or narrower range is associated with topic finality.

We also performed a contrastive experiment where silence features were excluded, to assess the dependence on these features. [5] The resulting classifier achieved an accuracy of 89.4% on the heldout balanced test set, reinforcing the utility of pitch and intensity features for classification.

We performed a second set of contrastive experiments to explore the impact of different lexical tones on classification accuracy. We grouped words based on the lexical

[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.

[4] On a randomly sampled test set, there were no missed boundaries and a ≈5% false alarm rate was observed.

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

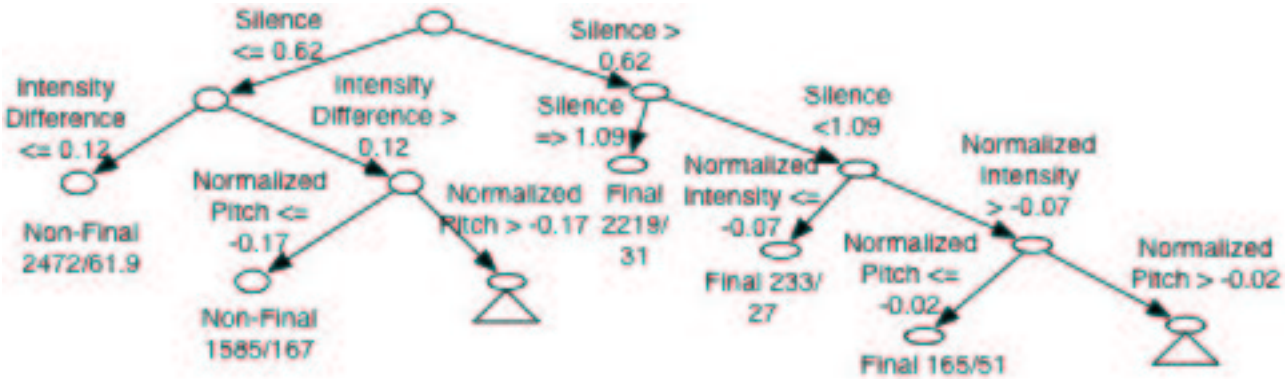

Figure 2: Decision tree classifier labeling words as segment-final or non-segment-final

tone of the initial syllable into high, rising, low, falling, and neutral groups. We found no tone-based differences in classification with all groups achieving 94-96% accuracy. Since the magnitude of the difference in pitch based on discourse position is comparable to that based on lexical tone identity, and the overlap between pitch values in non-final and final positions is relatively small, we obtain consistent results.

8 Conclusion and Future Work

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, Mandarin Chinese. We find highly significant decreases in pitch and intensity at topic final positions, and significant increases in word duration. Furthermore, these features in both local and contextualized form provide the basis for an effective decision tree classifier of boundary positions that does not use term similarity or cue phrase information, but only prosodic features. We also find that analogous to (Tur et al., 2001)'s work on an English story segmentation task, pause and pitch - both for the individual word and adjacency pair - play a crucial role; our findings for Chinese, however, identify a greater role played by intensity and durational contrasts.

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 prosodic and textual features and investigate the identification of more fine-grained sub-topic structure, to provide more focused units for information retrieval, summarization, and anaphora resolution.

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*.

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.

X.-N. Shen. 1989. *The Prosody of Mandarin Chinese*, volume 118 of *University of California Publications in Linguistics*. University of California Press.

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.