PROSODY-BASED SENTENCE BOUNDARY DETECTION IN CHINESE BROADCAST NEWS

Lei Xie, Chenglin Xu and Xiaoxuan Wang

Shaanxi Provincial Key Laboratory of Speech and Image Information Processing,
School of Computer Science, Northwestern Polytechnical University, Xi'an

ABSTRACT

In this paper, we explore the use of prosodic features in sentence boundary detection in Chinese broadcast news. The prosodic features include speaker turn, music, pause duration, pitch, energy and speaking rate. Specifically, considering the Chinese tonal effects in pitch trajectory, we propose to use tone-normalized pitch features. Experiments using decision trees demonstrate that the tone-normalized pitch features show superior performance in sentence boundary detection in Chinese broadcast news. Furthermore, feature combination is able to achieve apparent performance improvement by intuitive feature interactive rules formed in the decision tree. Pause duration and a tone-normalized pitch feature contribute the most part of the feature usage in the best-performing decision tree.

Index Terms— sentence boundary detection, sentence segmentation, speech prosody, rich transcription

1. INTRODUCTION

Sentence boundary detection [1], or sentence segmentation [2], is an important precursor to downstream speech and language processing tasks, e.g., speech summarization [3], story segmentation [4] and machine translation [5]. Moreover, the output of current speech recognizers is only a word stream without important structural information such as punctuation and paragraphs. As we know, punctuation, in particular sentence boundaries, is crucial to legibility [1]. Therefore, detecting sentence boundaries in spoken documents has recently drawn much interest from researchers. Lexical and prosodic cues [6] have been studied in sentence boundary detetection in different genres, e.g., broadcast news [4], conversations [7] and meetings [8].

Recent efforts have shown that the prosodic aspects of speech, especially pause and pitch information, are effective indicators for detecting structural events, including topic boundary [9, 4], speech disfluency [1] and sentence boundary [4]. For example, prosody-based sentence segmentation studies specifically point out that sentence boundaries are often signaled by some combination of a long pause, a pre-boundary low tone and a pitch reset [4]. However, such prosody-based sentence boundary detection approaches may be different for a tonal language such as Mandarin Chinese. The use of the same prosodic features in Chinese sentence segmentation deserves further investigation. This is because the tonal aspect of Chinese may complicate the expressions of pitch features. As we know, Chinese syllable tones are expressed acoustically in pitch trajectories. In other words, different tones show different pitch ranges and trajectory evolving patterns. However, previous prosody-based sentence segmentation studies on Chinese spoken documents have paid little attention to the effects caused by tonality [10]. Our previous work on story segmentation, another important structural event detection task, has shown that Chinese tonal syllables complicate the expressions of pitch declination and reset and affect the effectiveness of pitch features in story boundary detection. Instead, our study suggests that *tone-normalized pitch features* are much more effective [9, 11].

In this paper, we aim to investigate the effectiveness of various prosodic features in sentence boundary detection in Mandarin broadcast news. The prosodic feature set is composed of speaker turn, music, pause duration, pitch, speaking rate and energy. Specifically, we propose tone-based pitch feature normalization strategies in order to alleviate the tonal effects and to improve sentence boundary detection performance. We carry out experiments to test the prosodic features and to unveil how different prosodic features can be integrated to improve performance of sentence boundary detection in Chinese broadcast news. Feature combination experiments show that pause duration and the proposed tone-normalized pitch feature are the most important prosodic features.

2. THE APPROACH

We formulate the sentence boundary detection problem as a classification task, as shown in Fig. 1. Specifically, in an audio stream, we regard pause-separated inter-utterance boundary positions as sentence boundary candidates and label each candidate as sentence-boundary or non-sentence-boundary by a classifier. We expect that sentence boundaries can be discovered by a set of prosodic cues. Firstly, a set of features is extracted in the prosodic feature extraction region of each candidate in the audio stream. As shown in Fig. 1, the region of interest is a cross utterance window that covers the last word of the pre-boundary utterance, the pause interval and the first word of the post-boundary utterance. Secondly, a decision tree classifier is trained using the prosodic feature set

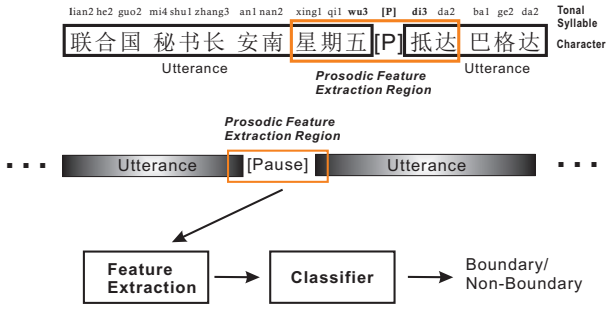

Fig. 1. Sentence boundary detection approach.

and its reference label (boundary/non-boundary). We choose decision tree as the classifier because it makes no assumption on the shape of the feature distributions, and most importantly, it provides visual interpretability of features. Finally, each candidate in the test audio stream is labeled as boundary or non-boundary using the decision tree classifier.

3. CORPUS

In this paper, we study sentence boundary detection on a subset of the TDT2 Mandarin broadcast news corpus[1], as shown in Table 1. In the TDT2 corpus, each audio file is provided with a speech recognition transcript and a reference transcript. The ASR transcript is composed of recognized words and pause labels with timing information, but it lacks sentence boundary labels (i.e. periods). The reference transcript is manually transcribed text with punctuation information. We manually add sentence boundary labels into the ASR transcripts according to the punctuation in the reference transcripts. We study the sentence boundary detection problem on the audio files and the modified ASR transcripts.

4. PROSODIC FEATURES

We extract a rich set of prosodic features at the candidate positions in the broadcast news audio stream. The feature set is composed of speaker turn, music break, pause duration, pitch, speaking rate and energy. In the following, we describe how to extract these features in detail.

4.1. Speaker Turn and Music

Broadcast news audio reports often consist of speaker changes among anchors, reporters and interviewees. Speaker turn positions are definitely sentence boundaries. We use a two-stage multi-feature integration approach to automatically detect speaker changes from broadcast news audio [12]. Speaker turn, namely SPK-T, is used as a binary feature with values 0 and 1.

In broadcasts news, music breaks are often used at intervals between different news stories. These story boundaries are certainly sentence boundaries. Therefore, we use an SVM binary tree approach [13] to detect music regions and whether music is show-up at candidate positions is used as a binary feature, namely MSC.

[1] http://www.ldc.upenn.edu/Projects/TDT2

Table 1. The corpus used in this study.

Source	TDT2 Mandarin (subset)
Nature	VOA broadcast news
Training Set	5 news episodes, 5 hours Boundary #: 1347 Non-Boundary #: 3760
Testing Set	2 news episodes, 2 hours Boundary #: 487 Non-Boundary #: 910

4.2. Pause Duration

Pause duration is one of the most important speech prosodic factors relevant to discourse structures. Speakers tend to use long pauses at major semantic boundaries. Previous studies on sentence segmentation suggest that pause duration is an effective sentence boundary indicator: the longer voiceless pause regions are, the more probably a sentence boundary is found. In our study, we obtain the pause duration information from the broadcast news ASR transcripts and the feature is named as PAU-DUR.

4.3. Pitch

Previous work has specifically pointed out that pitch (or F0) is a salient boundary indicator [4]. The pitch declination and reset phenomena are characterized by the tendency of a speaker to raise his/her pitch to the pitch topline at the beginning of a major speech unit, and lower it towards the pitch baseline at the end of the major speech unit. As a result, pitch undergoes a declination within the major speech unit and a reset between two major speech units. However, our previous work on story boundary detection [11], another important structural event detection task, has shown that Chinese tonal phenomena complicate the expressions of pitch declination and reset and affect the effectiveness of pitch features. Therefore, in this study, we propose tone-based pitch feature normalization strategies in order to alleviate the tonal effects and to improve sentence boundary detection performance.

4.3.1. Basic Pitch Features

We obtain frame-level F0 values by the YIN pitch tracker [14]. The raw F0 values are then post-processed to account for tracking errors and are further normalized by z-score speaker normalization [9]. We calculate pitch range and reset features at both syllable and word levels in the prosodic feature extraction region (in Fig. 1). To this end, we force-aligned each audio file with its associated ASR transcript to obtain syllable boundaries.

We extract the pitch range features in both a narrower window and a wider window in the prosodic feature extraction region. Specifically, similar to [4], we measure the

minimum (baseline) and the average F0 values on the pitch trajectories of the last character and the last word in the utterance before the boundary; we measure the maximum (topline) and the average F0 values on the pitch trajectories of the first character and the first word in the utterance after the boundary. Cross-boundary pitch reset is also calculated by the difference between the post-boundary F0 maximum and the pre-boundary F0 minimum. Finally, the word-window pitch feature set (F0-WRD) includes F0-MIN-L-WRD, F0-MEAN-L-WRD, F0-MAX-R-WRD, F0-MEAN-R-WRD and F0-RESET-WRD; the character-window pitch feature set (F0-CHR) includes F0-MIN-L-CHR, F0-MEAN-L-CHR, F0-MAX-R-CHR, F0-MEAN-R-CHR and F0-RESET-CHR.

4.3.2. Tone-normalized Pitch Features

To alleviate the tonal effects, we also extract tone-normalized pitch features. We perform pitch normalization by:

$$\mathcal{F}_0 = (F_0 - \mu_{F_0}^{\tau})/\sigma_{F_0}^{\tau}, \quad (1)$$

where F_0 is a pitch value on a pitch trajectory in the prosodic feature extraction region and the pitch trajectory belongs to a character whose syllable tone is τ ($\tau = 1, 2, 3, 4, 5$ and 5 is the neutral tone). $\mu_{F_0}^{\tau}$ and $\sigma_{F_0}^{\tau}$ are the pitch mean and standard deviation calculated for all tonal syllables with tone τ in the corpus.

The tone-normalized pitch range and reset features are then calculated using tone-normalized pitch values according to the procedure described in Section 4.3.1. Finally, we obtain the character-window pitch feature set (F0-TN-CHR) and the word-window pitch feature set (F0-TN-WRD)[2].

4.4. Speaking Rate

Final lengthening and initial shortening are well-known durational cues indicating boundaries [15]. It refers to the speakers general behavior of slowing down his/her speaking rate at the end of a major speech unit and speeding up at the beginning of another major speech unit. Phone and rhyme durations [4] are often used as speaking rate features in sentence boundary detection in English spoken documents. In this work, we use syllable (character) duration instead because Chinese is a syllable-timed language. Specifically, in the prosodic feature extraction region shown in Fig. 1, we estimate the average syllable durations of the last word preceding an utterance boundary (SYL-DUR-L) and the first word following an utterance boundary (SYL-DUR-R). The average syllable duration is defined as the word duration divided by the number of syllables in the word. It is expected that the lengthening and shortening effect is more pronounced at sentence boundaries. The cross boundary syllable duration difference is also considered as a speaking rate feature, namely SYL-DUR-DIFF.

[2]To distinguish between features before and after tone normalization, the tone-normalized pitch features are named using "-TN-".

4.5. Energy

Usually, speakers tend to end long sentences with less energy, rather than taking an additional breath towards the end of a sentence. This often leads energy to a similar declination behavior with pitch. Therefore, in the feature extraction region shown in Fig. 1, we calculate the mean frame-level RMS values of the last word preceding an utterance boundary (namely ENG-L) and the first word following an utterance boundary (namely ENG-R). The cross boundary energy difference (ENG-DIFF) is also considered as another energy feature and three energy features are collected.

5. EXPERIMENTAL STUDY

We carried out sentence boundary detection experiments on the use of both individual feature classes and combined features. We trained C4.5 decision tree classifiers using the Weka toolkit[3]. Since some features may have low discriminative abilities and some features may be highly correlated, we performed a feature selection procedure to find the optimal feature subset with highest F1-measure. Specifically, we adopted the backward elimination algorithm to seek the optimal subset by iteratively eliminating features whose absence did not decrease boundary detection performance. Classifier training and feature selection were performed on the training set and experimental results were reported on the testing set.

5.1. Experimental Results

Experimental results are summarized in Table 2. We can see that pause duration (PAU-DUR) achieves superior performance among different feature classes. The F1-measure achieved by this single feature is as high as 74.8%. The proposed word-level tone-normalized pitch feature set (F0-TN-WRD) achieves apparent performance improvement than the conventional pitch feature set (F0-WRD). It's interesting to notice that tone normalization is more effective for word (F0-TN-WRD VS F0-WRD) than character level (F0-TN-CHR VS F0-CHR). It is probably because the pitch trajectory for the same character might be different for different words, contexts and parts of sentences. Please note that the errors produced by ASR might also effect tone normalization of pitch. The F1-measure values for other feature sets are relatively low. This indicates that the use of these individual prosodic feature sets are not quite effective for sentence boundary detection. We notice that the music feature (MSC) can bring a 100% precision, but it can only recall quite a few sentence boundaries. We also observe that the feature selection procedure is able to improve the performance by removing features with low discriminative ability.

The combination of all 31 prosodic features[4] results in ap-

[3]Website: http://www.cs.waikato.ac.nz/ml/weka/. J48 is the implementation of C4.5 in Weka.

[4]The tone identities of the last syllable pre-boundary and the first syllable post-boundary are also used as two extra features, named as TONE-L and TONE-R, respectively.

Table 2. Experimental results for individual feature classes and feature combination.

Feature Set	Dim	All Features			Feature Selection			
		Recall	Prec.	F1	Recall	Prec.	F1	Selected Features
Speaker Turn	1	8.8	65.2	15.6	8.8	65.2	15.6	SPK-T
Music	1	1.8	100	3.6	1.8	100	3.6	MSC
Pause	1	68.6	82.3	74.8	68.6	82.3	**74.8**	PAU-DUR
F0-CHR	5	7.2	68.6	13	11.5	61.5	19.4	F0-RESET-CHR, F0-MEAN-L-CHR
F0-TN-CHR	5	3.9	73.1	7.4	4.3	70	8.1	F0-MAX-R-TN-CHR, F0-MEAN-L-TN-CHR
F0-WRD	5	7.8	71.7	14.1	30.2	66.5	41.5	F0-RESET-WRD, F0-MIN-L-WRD, F0-MEAN-L-WRD, F0-MEAN-R-WRD
F0-TN-WRD	5	31.6	67.5	43.1	33.5	66.3	44.5	F0-RESET-TN-WRD, F0-MIN-L-TN-WRD
Speaking Rate	3	0	0	0	0.6	75	1.2	SYL-DUR-DIFF, SYL-DUR-R
Energy	3	20.3	71.7	31.7	20.3	71.7	31.7	ENG-DIFF, ENG-R
Combined	31	61	84.9	71	72.1	84.2	**77.7**	PAU-DUR, SPK-T, SYL-DUR-L, F0-RESET-WRD F0-MIN-L-WRD, F0-MIN-L-TN-WRD, F0-MEAN-L-TN-WRD, TONE-L

Table 3. Feature usage in the best-performing decision tree.

Usage (%)	Feature
76.3	PAU-DUR
15.49	F0-MIN-L-TN-WRD
2.97	SYL-DUR-L
1.9	F0-RESET-WRD
1.38	TONE-L
1.2	F0-MEAN-L-TN-WRD
0.637	F0-MIN-L-WRD
0.123	SPK-T

parent performance degradation. The performance is dragged down by noneffective features. However, as expected, we obtain substantial performance improvement through feature selection. After feature selection, the best-performing tree identified a subset of eight features, which achieves the highest F1-measure of 77.7%. The relative F1-measure gain from the use of additional prosodic features (besides pause duration) is 3.9%. The best feature subset is composed of pause duration, speaker turn, pre-boundary syllable duration, 4 pitch features and the identity of the pre-boundary syllable tone.

5.2. Feature Usage

To rank the prosodic features according to their importance, we calculated the usage for each feature appeared in the decision tree. Feature usage [4] is computed as the relative percentage frequency with which that feature is queried in the decision tree. We counted the number of times the decision tree asked a question of a given feature, over all test samples, divided by the total number of questions asked. Features used higher in the tree classify more samples and therefore have higher feature usage. The percentage feature usage for the best-performing decision tree is shown in Table 3. We observe that pause duration and a tone-normalized pitch feature (F0-MIN-L-TN-WRD) contribute the most part of the feature usage. Hence they are the most important prosodic features in the decision tree grown using combined features.

5.3. Prosodic Feature Interaction

In Fig. 2, we show the top part of the decision tree with the best performance. We can clearly observe the interactions between different prosodic feature classes, illustrating the inter-feature complementarity that reinforces the detection of sentence boundaries. Feature interactions can be summarized in terms of five major heuristic rules shown on the tree. These heuristic rules (labeled as #1 to #5 in Fig. 2) cover about 80% boundary/non-boundary decisions with high accuracy made on the testing set, as described in Table 4. Here, usage is defined as the ratio between the number of samples classified using the heuristic rule and the total number of samples in the testing set. Accuracy is defined as the ratio between the number of correctly classified samples and the total number of samples classified using the heuristic rule.

Heuristic rules #1 and #2 are consistent with general prosodic descriptions of sentence boundaries concluded in previous studies [4]. Sentence boundaries are usually signaled by longer pauses and pre-boundary lower F0 range. Interestingly, we discover that heuristic rules #3-#5 demonstrate how different features work together to make boundary/non-boundary decisions when individual boundary features are not salient enough. For example, heuristic rule #3 shows that pre-boundary F0 is the crucial sentence boundary decision factor when pause duration is neither too long nor too short.

6. CONCLUSIONS

We have explored the use of prosodic features in sentence boundary detection in Chinese broadcast news. Specifically, we extract a set of prosodic features in the boundary candidate positions and train a decision tree classifier to label each candidate as boundary or non-boundary. Through an experimental study, we has discovered that: (1) Tone-normalized pitch feature is more effective for sentence boundary detection in Chinese broadcast news; (2) Pause duration and pitch range (tone-normalized) features are the most important features; (3) Feature combination is able to achieve substantial performance improvement through feature interactions. In future

Table 4. Major heuristic rules discovered in the best-performing decision tree.

No.	Rule description	Boundary	Usage (%)	Accuracy (%)
1	Pause duration is short	N	85.13	94.26
2	Pause duration is long and pre-boundary F0 is low	Y	47.14	84.67
3	Pause duration is neither too short nor too long, but pre-boundary F0 average is low enough	Y	12.99	59.52
4	Pause duration is long, pre-boundary F0 baseline is not low enough, speaker change occurs and pitch reset is large	Y	4.45	84.51
5	Pause duration is neither too short nor too long, pre-boundary F0 mean is not low enough and pre-boundary lengthening is not pronounced	N	0.43	94.01

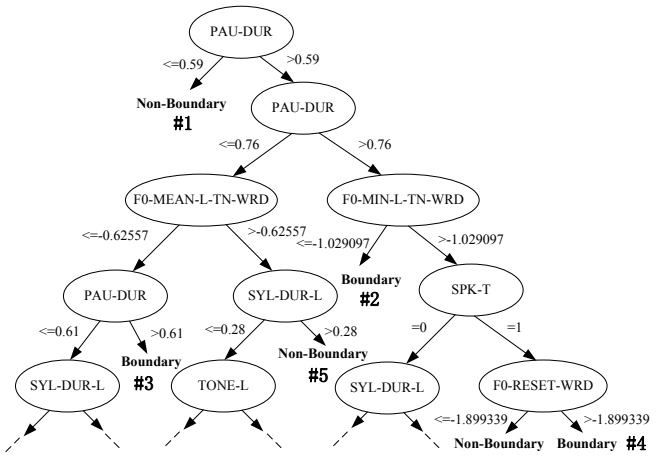

Fig. 2. Top levels of the best-performing decision tree.

work, we plan to use sequential labeling tools, e.g., conditional random fields (CRFs), to model the contextual prosodic information in Chinese broadcast news.

7. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (61175018), the Natural Science Basic Research Plan of Shaanxi Province (2011JM8009), the Key Science and Technology Program of Shaanxi Province (2011KJXX29), the Fok Ying Tung Education Foundation (131059) and a grant from Baidu.

8. REFERENCES

[1] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[2] M. Ostendorf, B. Favre, R. Grishman, D. Hakkani-Tür, M. Harper, D. Hillard, J. Hirschberg, J. Heng, J. G. Kahn, Y. Liu, S. Maskey, E. Matusov, H. Ney, A. Rosenberg, E. Shriberg, W. Wang, and C. Wooters, "Speech segmentation and spoken document processing," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 59–69, 2008.

[3] J. Mrozinski, E. W. D. Whittaker, P. Chatain, and S. Furui, "Automatic sentence segmentation of speech for automatic summarization," in *ICASSP*, 2006, vol. I, pp. 981–984.

[4] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, no. 1-2, pp. 127–154, 2000.

[5] J. Xu, R. Zens, and H. Ney, "Sentence segmentation using ibm word alignment model 1," in *Proceedings of 10th Annual Conference of the European Association for Machine Translation*, 2005, pp. 280–287.

[6] U. Guz, S. Cuendet, D. Hakkani-Tür, and G. Tur, "Co-training using prosodic and lexical information for sentence segmentation," in *Interspeech*, 2007, pp. 2597–2600.

[7] B. Roark, Y. Liu, M. Harper, R. Stewart, M. Lease, M. Snover, Z. Shafran, B. Dorr, J. Hale, A. Krasnyanskaya, and L. Young, "Reranking for sentence boundary detection in conversational speech," in *ICASSP*, 2006, vol. 1, pp. 545–548.

[8] J. Kolar, E. Shriberg, and Y. Liu, "Using prosody for automatic sentence segmentation of multi-party meetings," in *Proceedings of the 9th International conference on Text, Speech and Dialogue*, 2006, pp. 629–636.

[9] L. Xie, "Discovering salient prosodic cues and their interactions for automatic story segmentation in Mandarin broadcast news," *Multimedia Systems*, vol. 14, pp. 237–253, 2008.

[10] M. Zimmerman, Dilek Hakkani-Tr, James G. Fung, Nikki Mirghafori, L. Gottlieb, Elizabeth Shriberg, and Yang Liu, "The ICSI+ multilingual sentence segmentation system," in *Interspeech*, 2006, pp. 117–120.

[11] L. Xie, C. Liu, and H. Meng, "Combined use of speaker- and tone-normalized pitch reset with pause duration for automatic story segmentation in mandarin broadcast news," in *HLT-NAACL*, 2007, pp. 193–196.

[12] L. Xie and G. Wang, "A two-stage multi-feature integration approach to unsupervised speaker change detection in real-time news broadcasting," in *ISCSLP*, 2008, pp. 350–353.

[13] L. Xie, Z. Fu, W. Feng, and Y. Luo, "Pitch-density-based features and an svm binary tree approach for multi-class audio classification in broadcast news," *Multimedia Systems*, pp. 101–112, 2011.

[14] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *Journal of the Acoustic Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[15] C. Y. Tseng, S. H. Pin, Y. Lee, H. M. Wang, and Y. C. Chen, "Fluent speech prosody: Framework and modelling," *Speech Communication*, vol. 46, pp. 284–309, 2005.