Modeling Broadcast News Prosody Using Conditional Random Fields for Story Segmentation

Xiaoxuan Wang* ‡ Lei Xie* Bin Ma† Eng Siong Chng‡ and Haizhou Li†‡
* School of Computer Science, Northwestern Polytechnical University, China
E-mail: xwang@nwpu-aslp.org, lxie@nwpu.edu.cn
† Institute for Infocomm Research, Singapore
E-mail: mabin,hli@i2r.a-star.edu.sg
‡ School of Computer Engineering, Nanyang Technological University, Singapore
E-mail: aseschng@ntu.edu.sg

***Abstract*—This paper proposes to model broadcast news prosody using conditional random fields (CRF) for news story segmentation. Broadcast news has both editorial prosody and speech prosody that convey essential structural information for story segmentation. Hence we extract prosodic features, including pause duration, pitch, intensity, rapidity, speaker change and music, for a sequence of boundary candidates. A linear-chain CRF is used to label each candidate with boundary/non-boundary tags based on the prosodic features. Important inter-label relations and contextual feature interactions are effectively captured by CRF's sequential learning framework. Experiments show that the CRF approach outperforms decision tree (DT), support vector machines (SVM) and maximum entropy (ME) classifiers in prosody-based story segmentation.**

I. Introduction

Spoken documents, e.g., broadcast news, meetings and lectures, usually have multiple topics or sub-topics. For example, a one-hour broadcast news episode often includes a series of news stories, each addressing a central topic. The task of automatic story segmentation is to divide the spoken documents into topically homogeneous segments, known as *stories*. Segmentation is an important precursor that facilitates efficient information extraction, topic tracking, summarization, browsing, indexing and retrieval [1]. With the ever-increasing volumes of spoken content, automatic story segmentation techniques are highly in demand.

Automatic story segmentation approaches have focused on generative topic modeling [2] and story boundary detection [1], [3], [4], [5], [6]. The former category treats the word sequence (transcribed from speech) as observations of some pre-defined topics and topic labels are assigned to the speech transcripts by some optimal criterion. Story segmentation is simply obtained by marking boundaries between every two adjacent parts with different topic labels. For detection-based approaches, boundary candidates are first determined across the spoken document. Then story segmentation is viewed as a sequential classification/tagging problem, i.e., classifying each candidate into boundary or non-boundary based on a set of features/cues. Lexical cues have been extensively studied, such as word cohesiveness [3] and cue phrases [4]. For example, TextTiling [3] is classical lexical cohesiveness approach which based on an intuitive assumption that different topics usually employ different sets of words. Under graph representation, normalized cuts (N-cuts) [6] can optimize the sentence similarity within each story and dissimilarity across different stories. However, in lexical approaches to spoken document segmentation, speech-to-text is first performed by a large vocabulary continuous speech recognizer (LVCSR) and the inevitable speech recognition errors may affect the segmentation performance [6].

Spoken documents intrinsically have their *prosody* with an embedded rhythm on topic shifts [1], [5]. For example, broadcast news programs often follow editorial prosodic rules: (1) switch news topics by musical breaks or significant pauses; (2) two announcers report news stories in turn; (3) a studio announcer starts a topic and then passes it to a reporter for a detailed report. In TV newscasts, a similar prosodic cue is field-to-studio transition, i.e., each news story starts from a studio shot and then move to field shots [4]. Beside the editorial prosody, speakers naturally separate their discourses into different semantic units (e.g., sentences, paragraphs and topics) through *speech prosody* [7]. Prosodic contents of speech, e.g., intonational, durational and energy characteristics, are known to be relevant to discourse structures across languages [8]. Research has shown that listeners can perceive major discourse boundaries even if the speech itself is made unintelligible via spectral filtering [9]. This indicates that speech prosody conveys essential structural information of discourses. Therefore, prosodic cues have drawn much attention in structural event detection in speech, including disfluency detection, sentence and story segmentation [5], [10], [11], [1].

In this paper, we propose to model broadcast news prosody using conditional random fields (CRFs) for news story segmentation. A CRF is an undirected graphical model that defines a global log-linear distribution of the entire label sequence conditioned on the observation sequence [12]. The model has theoretical advantages in sequential classification: (1) it provides an intuitive method for integrating features from various sources; (2) it models the sequential/contextual information and labels a given candidate by considering its surrounding features and labels (i.e. global optimal labeling). Recently, CRFs have shown superior performances in various speech and langauge tasks such as POS tagging [12], shallow parsing [13], sentence boundary detection [10], pitch accent

Proceedings of the Second APSIPA Annual Summit and Conference, pages 253–256,
Biopolis, Singapore, 14-17 December 2010.

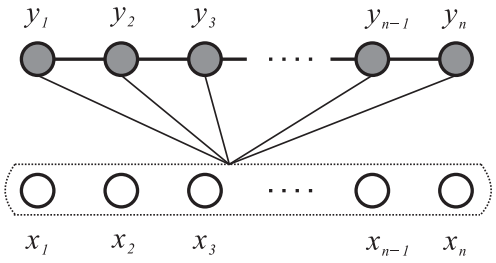

Fig. 1. A linear-chain conditional random field

prediction [14] and speech recognition [15]. In this paper, we show that CRF-based prosodic modeling of broadcast news can achieve better story segmentation performance as compared to several state-of-the-art classifiers.

II. Conditional Random Fields

A conditional random field (CRF) is a discriminative probabilistic model most often used for labeling or segmenting sequential data [12]. It is a random field in nature, where each random variable is conditioned on an observation sequence. Fig. 1 illustrates a simple linear-chain CRF frequently used in sequential data labeling, which defines a conditional probability distribution $p(\mathbf{y}|\mathbf{x})$ of label sequence $\mathbf{y} = (y_1, y_2, \cdots, y_n)$ given input observation sequence $\mathbf{x} = (x_1, x_2, \cdots, x_n)$. Specifically for the story segmentation task, $\mathbf{y}$ represents a label sequence with story-boundary or non-story-boundary labels and $\mathbf{x}$ is the feature observation sequence. The decoding problem, i.e., finding the most likely label sequence $\hat{\mathbf{y}}$ for the given observation sequence, can be calculated by

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}), \tag{1}$$

where the posterior probability takes the exponential form:

$$p(\mathbf{y}|\mathbf{x}) = \frac{\exp \sum_k \lambda_k \cdot F_k(\mathbf{y}, \mathbf{x})}{Z_\lambda(\mathbf{x})}. \tag{2}$$

$F_k(\mathbf{y}, \mathbf{x})$ are called feature functions defined over the observation and label sequences. The index k indicates different features, each of which has an associated weight λ_k. For input sequence $\mathbf{x}$, label sequence $\mathbf{y}$:

$$F_k(\mathbf{y}, \mathbf{x}) = \sum_i f_k(\mathbf{y}, \mathbf{x}, i) \tag{3}$$

where i ranges over input positions. Z_λ is the normalization term:

$$Z_\lambda(\mathbf{x}) = \sum_{\mathbf{y}} \exp \sum_k \lambda_k \cdot F_k(\mathbf{y}, \mathbf{x}). \tag{4}$$

The CRF model is trained by maximizing the conditional distribution $p(\mathbf{y}|\mathbf{x})$ on a given training set. The most likely label sequence is found using the Viterbi algorithm.

When $f_k(\mathbf{y}, \mathbf{x}, i) = f_k(y_{i-1}, y_i, \mathbf{x}, i)$, a first-order linear-chain CRF is formed, which only includes two sequential labels in the feature set. For an *N*-order linear-chain CRF, the feature function is defined as $f_k(y_{i-N}, \cdots, y_i, \mathbf{x}, i)$. Training is only practical for lower orders of N since the computational cost increases exponentially with N. Specifically, if we substitute $\mathbf{x}$ and $\mathbf{y}$ in Eq. (1)-(4) with x_i and y_i, the CRF model is downgraded to a maximum entropy (ME) model [16]. An ME classifier individually classifies each data sample without using contextual information, whereas a CRF models sequential information and performs a global optimal labeling.

III. Modeling Broadcast News Prosody using CRF

A. CRF for Story Segmentation

We transfer news story segmentation to a sequential classification task using linear-chain CRF modeling (Fig. 1) and prosodic features. We first determine story boundary candidate positions across the broadcast news audio stream. These candidates constitute the label sequence $\mathbf{y}$ to be decoded in the linear-chain CRF. The principle of candidate selection is to reduce the boundary search complexity and to maintain a lower miss rate of boundaries at the same time. For broadcast news audio, we consider all the silence and music positions as the story boundary candidates. These positions can recall almost all the story boundaries because broadcast news use silence breaks and music intervals to keep an editorial tempo. A set of prosodic features is then collected at the boundary candidate positions, which constitutes the observation sequence $\mathbf{x}$. Based on a training set with the reference labels and the extracted prosodic features, we train a linear-chain CRF classifier that can thus label an input broadcast news stream with boundary and non-boundary tags at each candidate position.

B. Broadcast News Prosody and Feature Extraction

We study story boundary cues of broadcast news through both *editorial prosody* and *speech prosody*. Editorial prosody refers to that broadcast news producers usually use prosodic cues, such as music, speaker change and significant pause, to switch news topics and sections [1], [5]. On the other hand, speech prosody, i.e., intonational, durational and energy aspects of speech, conveys structural, semantic, and functional information in all languages. Past research results suggest that speakers use prosody to impose structure on both spontaneous and read speech [1], [7], [10]. In this study, we investigate pause, pitch, intensity, rapidity, speaker change and music, resulting in a candidate prosodic feature set of 14 dimensions.

1) Pause Duration: Pause duration is one of the most important speech prosodic factors relevant to discourse structures. Speakers tend to use a long pause at large semantic boundaries. The average pause duration between different topics usually lasts longer than between sentences or lower semantic units. On the other hand, broadcast news producers usually insert a clear silence or a music clip between news stories. Previous work has shown that pause duration is quite effective for story segmentation of broadcast news [1], [4], [5], [10]. Fig. 2 shows the pause duration time trajectory of a VOA broadcast news episode. We can clearly see the pause melody, where pause duration is much salient at story boundary positions. Therefore, we use pause duration achieved from a home-grown audio classifier as a prosodic feature, namely PauD.

2) Pitch: Pitch declination and reset phenomena are characterized by the tendency of a speaker to raise his/her pitch to the topline at the beginning of a major speech unit, and lower it towards the pitch baseline at the end of the major speech unit. Therefore, pitch undergoes a declination within the major speech unit and a reset between two major speech units. Pitch declination and reset behaviors have been shown

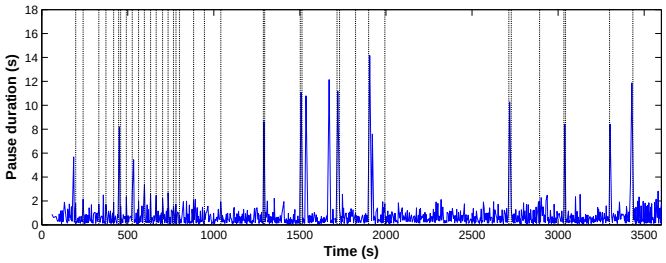

Fig. 2. Pause duration time trajectory for a VOA Mandarin broadcast news episode. Dotted vertical lines denote story boundaries

more pronounced at topic level than other smaller speech levels like utterances [1], [5], [10].

In this study, we extract pitch trajectory from broadcast news audio by the YIN pitch tracker [17]. The left and right nearest successive pitch contours of each boundary candidate (i.e. pause segment) are our regions of interest. A set of five pitch features are extracted for each boundary candidate, including pitch baseline before a candidate (PLbase), pitch topline after a candidate (PRtop), mean pitch before and after a candidate (PLmn and PRmn) and pitch reset (PReset, i.e. PRmn-PLmn). Since pitch is a speaker-dependent characteristic, we normalize pitch by speaker before pitch feature calculation [1]. For this purpose, a home-grown speaker change detector [18] is used to separate different speakers from audio.

3) Intensity: Previous work has shown that speech intensity has similar declination and reset behaviors with pitch [1], [7]. Speakers are likely to raise their speech volume at the beginning of major speech units, and lower their volume at the end of major speech units. Three energy features are extracted for each boundary candidate: mean speech energy for the word before and after a candidate (ELmn and ERmn), and energy reset (EReset, i.e. ERmn-ELmn). Similar to pitch, speaker-normalized energy is used.

4) Rapidity: Final lengthening and initial shortening effects are well-known durational cues indicating semantic boundaries [1], [7], [5]. They refer to the speaker's general behavior of slowing down the speaking rate at the end of a speech unit and speeding up the speaking rate at the beginning of another speech unit. In our study, we investigate the average syllable duration of the word before and after a candidate (SLmn and SRmn) and syllable duration reset (SReset, i.e. SRmn-SLmn).

5) Speaker Change and Music: Broadcast news programs usually involve various speakers, such as anchors, reporters, interviewees, etc. Many news sessions are hosted by two anchors and they report news in turn. For example, a male anchor and a female anchor usually alternate with each other to announce news in the VOA brief news session. Some news programs follow a clear syntax [4]: a news story is led in by an anchor in the studio, and then followed by a detail report from a field reporter or an interview. Therefore, in broadcast news, speaker changes may direct to story transitions. We use a home-grown speaker change detector [18] to automatically detect speaker changes from broadcast news audio. We align each detected speaker change to an appropriate candidate point (pause) since speaker change does not always show up exactly at a candidate point. The speaker change detector has an unavoidable delay in reporting speaker change points due to its window-based detection mechanism. Thus we align each detected speaker change point to its nearest candidate point. Speaker change (SpChg) is used as a binary prosodic feature (change/non-change for each candidate). Beside speaker change, the appearance of music (MSC) is also used as another binary feature for each candidate.

TABLE I
CORPUS FOR STORY SEGMENTATION EXPERIMENTS

Corpus		*TDT-2 Mandarin*
Source		VOA newscast, Feb. to June 1998
No. of programs		177
Audio duration		53h
WER		37%
Data assignment	Training	90 programs (1321 boundaries)
	Testing	87 programs (1262 boundaries)

IV. EXPERIMENTS

A. Corpus

We carried out story segmentation experiments on TDT2 VOA Mandarin corpus[1] to evaluate the proposed approach. Table I shows the details of the corpus and the data organization for experiments. We compared the detected story boundaries with the manually annotated boundaries in terms of *recall, precision* and their harmonic mean - *F1-measure*. According to the TDT standard, a detected story boundary is considered correct if it lies within a 15-second tolerant window on each side of a manually-annotated reference boundary.

B. Experimental Setup

We trained a CRF boundary/non-boundary classifier using the labeled candidates with features $F_k(\mathbf{x}, \mathbf{y})$ in the training set, where $\mathbf{x}$ represents the observation sequence and $\mathbf{y}$ refers to the corresponding reference label sequence. The Mallet [2] package was used to perform CRF training and testing. As Mallet package's simple interface only supported discrete feature inputs, we quantized the continuous features (all features except SpChg and MSC) into discrete formats by equal frequency binning. We tested different CRF order N ($\mathbf{y} = y_{i-N}, \cdots, y_i$) and feature context M ($\mathbf{x} = x_{i-M}, \cdots, x_i$) in order to achieve the best story segmentation performance.

We also compared the CRF approach with several state-of-the-art classifiers, i.e., decision tree (DT), support vector machines (SVM) and ME. We employed Quinlan's C4.5 decision tree, the tree building and testing were implemented by the Weka[3] toolkit. We used the *SVM*$^{\text{light}}$ toolkit[4] to build an RBF-kernel SVM classifier. The opennlp.maxent[5] package was used to perform the ME classifier training and testing.

Since features may be redundant or correlated and some features may have low discriminative ability, we further performed feature selections to achieve the best feature subsets for each classifiers. We adopted a greedy heuristic search algorithm, i.e backward elimination [19], to seek the optimal subset by iteratively eliminating the features whose absence

[1] http://projects.ldc.upenn.edu/TDT2/
[2] http://mallet.cs.umass.edu/
[3] http://www.cs.waikato.ac.nz/ml/weka/
[4] http://svmlight.joachims.org/
[5] http://opennlp.sourceforge.net/

TABLE III
FEATURE SELECTION RESULTS AND F1-MEASURE FOR DIFFERENT CLASSIFIERS

	F1-mea.	PauD	SpChg	MSC	PReset	PLbase	PRTop	PLmn	PRmn	EReset	ELmn	ERmn	SReset	SLmn	SRmn
DT	0.6445	✓	✓		✓				✓	✓			✓		
SVM	0.6453	✓	✓	✓			✓	✓	✓	✓	✓			✓	
ME	0.6506	✓	✓	✓	✓	✓		✓	✓			✓	✓		✓
CRF	**0.6783**	✓	✓	✓				✓	✓			✓			
TT	0.5632														
NCuts	0.6911														

TABLE II
EXPERIMENTAL RESULTS FOR CRFS WITH DIFFERENT N AND M

Context(M) \ Orders(N)	1	2	3	4
0	0.6543	0.6636	0.6783	0.6292
1	0.6639	0.6731	0.6328	0.636
2	0.6592	0.6402	0.6413	0.5822
3	0.6237	0.6296	0.5794	0.6481
4	0.6374	0.6412	0.5939	0.6000

do not decrease F1-measure. Classifier training and feature selection were carried out on the TDT2 training set; story segmentation results were reported on the TDT2 testing set.

C. Results and Analysis

Experimental results for CRF with different order (N) and feature context (M) are listed in Table II. We can observe that the best F1-measure is 0.6783, which is achieved when $N = 3$ and $M = 0$. This result indicates that modeling the sequential/contextual information can improve the story segmentation performance. However, further increase of M and N leads to performance degradation. This is probably due to training data sparseness for complex CRF modeling. Meanwhile, the computational cost grows exponentially with the increase of M and N.

Table III summarizes the feature selection results and performance comparison between different classifiers. We also listed the performances of two lexical-based approaches on the same corpus: the classic TextTiling approach (TT) [3] and a recent approach based on subword NCuts [6]. From the results, we can clearly see that the proposed CRF approach outperforms DT, SVM and ME; prosody-based story segmentation methods can obtain comparable and even better results over lexical-based methods. Feature selection results show that: (1) not all prosodic features are useful to the segmentation task; (2) different classifiers may select different features; (3) pause duration (PausD), speaker change (SpChg) and pitch after candidate (PRmn) are selected by all classifiers, which shows their significance to the story segmentation task.

V. CONCLUSIONS AND FUTURE WORKS

We proposed to model broadcast news prosody using CRFs, a sequential learning framework, for automatic story segmentation. We extracted prosodic features, including pause, pitch, intensity, rapidity, speaker change and music, which originate from both editorial prosody and speech prosody. Sequential inter-label relations and contextual feature interactions are effectively captured by a linear-chain CRF. Our results show that the CRF approach outperforms other competitive classifiers, i.e., DT, SVM and ME, in prosody-based story segmentation; the prosody approaches shows comparable performance with lexical-based approaches. Our future work involves directly using continuous observations (features) in the CRF approach, as well as integrating multi-modal cues (video, audio and lexis) using CRFs for a better story segmentation performance.

VI. ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China (60802085), the Research Fund for the Doctoral Program of Higher Education (20070699015), the Program for New Century Excellent Talents in University and the NPU Foundation for Fundamental Research (W018103).

REFERENCES

[1] L. Xie, "Discovering salient prosodic cues and their interactions for automatic story segmenation in Mandarin broadcast news," *Multimedia Systems*, vol. 14, pp. 237–253, 2008.
[2] 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, 1998, pp. 333–336.
[3] M. Hearst, "Textiling: Segmenting text info multiparagraph subtopic passages," *Computational Linguistics*, vol. 23, no. 1, pp. 33–64, 1997.
[4] W. Hsu, S. F. Chang, C. W. Huang, L. Kennedy, C. Y. Lin, and G. Iyengar, "Discovery and fusion of salient multi-modal features towards news story segmentation," in *Proc. SPIE*, vol. 5307, 2004.
[5] 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.
[6] J. Zhang, L. Xie, W. Feng, and Y. Zhang, "A Subword Normalized Cut Approach to Automatic Story Segmentation of Chinese Broadcast News," *Proc. AIRS*, pp. 136–148, 2009.
[7] 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.
[8] J. Vaissiére, "Language-independent prosodic features," in *Prosody*. Berlin: Springer, pp. 53–66.
[9] Swerts, M. and T. Geluykens, R., "Prosodic correlates of discourse units in spontaneous speech," in *Proc. ICSLP*, 2006, pp. 421 –424.
[10] 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.
[11] M. Zimmerman, D. Hakkani-Tür, J. Fung, N. Mirghafori, L. Gottlieb, E. Shriberg, and Y. Liu, "The ICSI+ multilingual sentence segmentation system," in *Proc. Interspeech*, 2006, pp. 117–120.
[12] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. on Machine Learning*, 2001, pp. 282–289.
[13] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," in *Proc. HLT-NAACL*, 2003, pp. 213–220.
[14] G. Levow, "Automatic prosodic labeling with conditional random fields and rich acoustic features," *Proc. IJCNLP*, 2008.
[15] J. Morris and E. Fosler-Lussier, "Combining phonetic attributes using conditional random fields," in *Proc. Interspeech*, 2006.
[16] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A Maximum Entropy Approach to Natural Language Processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.
[17] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, p. 1917, 2002.
[18] L. Xie and G. Wang, "A Two-Stage Multi-Feature Integration Approach to Unsupervised Speaker Change Detection in Real-Time News Broadcasting," in *Proc. ISCSLP*, 2008, pp. 350–353.
[19] I. Guyon, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.