IEICE TRANS. INF. & SYST., VOL.E95–D, NO.5 MAY 2012

PAPER Special Section on Recent Advances in Multimedia Signal Processing Techniques and Applications

Broadcast News Story Segmentation Using Conditional Random Fields and Multimodal Features

Xiaoxuan WANG[†a)], **Lei XIE**[†b)], **Mimi LU**[†c)], **Bin MA**[††d)], ***Nonmembers***, **Eng Siong CHNG**[†††e)], ***and*** **Haizhou LI**[††f)], ***Members***

SUMMARY In this paper, we propose integration of multimodal features using conditional random fields (CRFs) for the segmentation of broadcast news stories. We study story boundary cues from lexical, audio and video modalities, where lexical features consist of lexical similarity, chain strength and overall cohesiveness; acoustic features involve pause duration, pitch, speaker change and audio event type; and visual features contain shot boundaries, anchor faces and news title captions. These features are extracted in a sequence of boundary candidate positions in the broadcast news. A linear-chain CRF is used to detect each candidate as boundary/non-boundary tags based on the multimodal features. Important interlabel relations and contextual feature information are effectively captured by the sequential learning framework of CRFs. Story segmentation experiments show that the CRF approach outperforms other popular classifiers, including decision trees (DTs), Bayesian networks (BNs), naive Bayesian classifiers (NBs), multilayer perception (MLP), support vector machines (SVMs) and maximum entropy (ME) classifiers.

key words: *story segmentation, conditional random fields*

1. Introduction

With the development of multimedia and web technologies, ever-increasing multimedia collections are available, including those of broadcast news, meetings and lectures. Given the vast amount of multimedia data, automatic approaches for multimedia processing are urgently required, particularly for automatic indexing, summarization, retrieval, visualization and organization technologies. Among these technologies, automatic story (or topic) segmentation is an important precursor since other tasks usually assume the presence of individual topical documents. Story segmentation is a task that divides a stream of text, speech or video into topically homogeneous blocks known as *stories*. Specifically, for broadcast news (BN), a popular media repository, the objective is to segment continuous audio/video streams into distinct news stories, each addressing a central topic.

Manuscript received May 6, 2011.
Manuscript revised August 22, 2011.

[†]The authors are with the School of Computer Science, Northwestern Polytechnical University, China.
[††]The authors are with Institute for Infocomm Research, Singapore.
[†††]The author is with School of Computer Engineering, Nanyang Technological University, Singapore.
a) E-mail: xwang@nwpu-aslp.org
b) E-mail: lxie@nwpu.edu.cn
c) E-mail: mlu@nwpu-aslp.org
d) E-mail: mabin@i2r.a-star.edu.sg
e) E-mail: aseschng@ntu.edu.sg
f) E-mail: hli@i2r.a-star.edu.sg
DOI: 10.1587/transinf.E95.D.1206

Story boundary cues (features) from different modalities are of great importance for automatic story segmentation. Lexical cues reveal story boundaries via semantic variations across the text, which mainly include the exploration of word cohesiveness and the use of cue phrases [1]. For example, the TextTiling approach [2], [3] measures the lexical similarity between pairs of sentences in a text and local minima are detected as story boundaries. The lexical chaining method [4] chains related words such as word repetitions, and positions with high counts of chain starts and ends are considered as story boundaries. Recently, speech prosody has draw a considerable attention because it provides an acoustic knowledge source with an embedded rhythm on topic shifts [5], [6]. For example, broadcast news programs often follow editorial prosodic rules, such as the following. (1) News topics are separated by musical breaks or significant pauses; (2) two announcers report news stories in turn; (3) a studio anchor starts a topic and then passes it to a reporter for a detailed report. In addition to editorial prosody, speakers naturally separate their discourse into different semantic units (e.g., sentences, paragraphs and topics) through durational, intonational and intensity cues, known as *speech prosody* [5], [7]. Compared with lexical and acoustic cues, visual cues are more reliant on editorial rules and news production patterns. The transition of stories is usually followed by the change of video shots. For example, field-to-studio shot transition is a salient story boundary cue. This is because many broadcast news programs follow a clear pattern: each news story starts with a studio shot and then moves to field shots [8]. An anchor face is another visual feature indicating a topic transition [9]. Moreover, in a broadcast news video, a news story is often accompanied by a caption describing the content of the news.

Story segmentation approaches can be categorized into generative topic modeling [10]–[12] and story boundary detection [5], [8], [13]–[15]. The former category treats the word sequence (speech transcripts) as observations of some predefined topics, and topic labels are assigned to the speech transcripts under an optimal criterion. In the detection-based framework, boundary candidates are first determined across a spoken document. Story segmentation is then viewed as a sequential classification/tagging problem, i.e., each candidate is classified into a boundary or nonboundary based on a set of features. In this paper, we focus on the story boundary detection approach. Some recent studies have shown that integrating different features can significantly

improve the detection performance [8], [14], [16]. Decision trees (DTs) have been used for the integration of lexical and acoustic features [14], [17], owing to their effective ability to model feature interactions, to deal with missing features and to handle a large amount of training data. Tür *et al.* [14] adopted a hidden Markov model (HMM) to fuse features from different knowledge sources. Word usage and lexical cues were represented by a language model embedded in the HMM while prosodic cues, such as pause durations and pitch resets, were modeled by a DT based on automatically extracted acoustic features and alignments. The system developed in the *Informedia* project [13] was one of the earliest rule-based broadcast news video story segmentation systems, in which, some ad -hoc rules were designed to combine visual, acoustic and lexical features. Recently, support vector machine (SVM) [18] and maximum entropy (ME) models [9] have also been used for story segmentation.

Despite years of study, most of the previous research has focused on modeling features independently. However, time series data, such as speech and video, has a strong correlation among adjacent units. In particular, when it comes to the highest level of understanding such as story segmentation, global information is believed to be much more helpful. In this study, we employ a detection-based story segmentation approach and propose the integration of multimodal features using conditional random fields (CRFs) for news story segmentation. A CRF is an undirected graphical model that defines the global log-linear distribution of an entire label sequence conditioned on an observation sequence [19]. The model has theoretical advantages for sequential classification: (1) it provides an intuitive method for integrating features from various sources because there is no assumption of independence among features. This property of CRFs is used to help us to investigate the relations among features from intramodality to intermodality; (2) it models the sequential/contextual information and labels of a given candidate by considering its surrounding features and labels (i.e. global optimal labeling). In this way, a CRF models the conditional distribution of a label sequence given the feature sequence by globally combining both the feature-to-label and label-to-label correlations, which is thus a better framework for segmenting time series data. Recently, CRF modeling has exhibited superior performance in various speech and language tasks such as POS tagging [19], shallow parsing [20], sentence boundary detection [21], pitch accent prediction [22] and speech recognition [23].

The remainder of this paper is organized as follows. In the next section, we give an overview of our story segmentation system. In Sect. 3, we describe the proposed CRF approach for story segmentation. Section 4 reports the extraction of multi-modal features. We present our experimental results and analysis in Sect. 5 and summarize the paper in Sect. 6.

2. System Overview

The detection-based story segmentation system consists of three steps: candidate identification, feature extraction and boundary/nonboundary classification, as shown in Fig. 1. We model the story segmentation task as a sequential boundary/nonboundary classification/tagging problem. We first identify a set of candidates (i.e. potential story boundaries), denoted as $\mathcal{B}$, in the broadcast news stream. The principle of this step is to reduce the boundary search complexity and to ensure a low miss rate (high recall rate) of story boundaries at the same time. In this study, we consider all the silence and music positions (labeled by an audio classifier) as story boundary candidates. These positions can cover almost all the story boundaries because news broadcasts use silence breaks and music intervals to maintain the editorial tempo. A set of multimodal features, denoted as $\mathcal{F}$, which include acoustic, lexical and visual features, are then collected at these boundary candidates. We aim to classify the set of candidates, $\mathcal{B}$, into two classes (boundary and nonboundary) with the highest probability given the feature set $\mathcal{F}$:

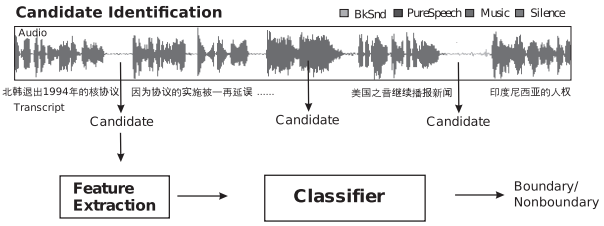


Fig. 1 Block diagram of the story segmentation approach.

$$\arg\max_{\mathcal{B}} P(\mathcal{B}|\mathcal{F}). \tag{1}$$

A CRF classifier, which is trained using multimodal features, is designed to carry out the classification. For performance comparison, several state-of-the-art classifiers, including three generative classifiers based on a DT, a BN and a NB, and three discriminative classifiers based on multilayer perceptions (MLP), support vector machines (SVMs) and maximum entropy (ME) are evaluated. We also investigate the effectiveness of features and how different features complement each other to improve the story segmentation performance.

3. Modeling Story Boundaries Using Conditional Random Fields

A CRF is a discriminative probabilistic model that has been used for labeling or segmenting sequential data [19]. It is a Markov random field in nature, where each random variable is conditioned on an observation sequence. In Fig. 2, a simple linear-chain CRF is illustrates which frequently used in sequential data labeling, which defines the conditional probability distribution $p(\mathcal{B}|\mathcal{F})$ of a label sequence $\mathcal{B} = (\mathcal{B}_1, \mathcal{B}_2, \cdots, \mathcal{B}_n)$ given an input observation sequence $\mathcal{F} = (\mathcal{F}_1, \mathcal{F}_2, \cdots, \mathcal{F}_n)$. Specifically for the story segmentation task, $\mathcal{B}$ represents a label sequence with story-boundary or non-story-boundary labels, and $\mathcal{F}$ is the feature observation sequence. We extract acoustic and visual features

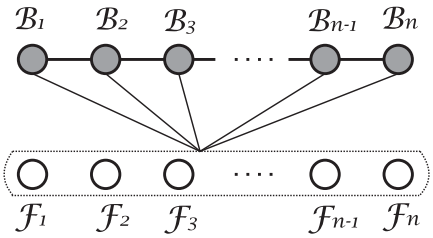


Fig. 2 Linear-chain CRF.

from the audio and video of broadcast news respectively, and search for lexical features based on speech recognition transcripts.

There are several benefits of using a CRF to model features: (1) A CRF is capable of accommodating statistically correlated features. Features from the same modality usually have semantic dependencies. For two lexical features, a lower lexical similarity usually accompany with a greater chaining strength at story boundary positions. A similar phenomenon can be observed for acoustic features where a long pause usually occurs with a change in speaker. However, different modality features always compliment each other, which is thus believed to lead to more robust segmentation by integrating different sources of information. (2) Modeling contextual feature information is beneficial for story segmentation. As one of the most conventional lexical similarity features, we often adopt the depth score [2], which reflects the contextual variantion tendency of lexical similarity, instead of using the lexical similarity directly indicting, that the contextual information is essential for this high-level structure task.

Starting with a training set with the reference labels and the extracted multi-modal features, we train a linear-chain CRF classifier that can label an input broadcast news stream with boundary and nonboundary tags at each candidate position. The decoding problem, i.e., finding the most likely label sequence $\hat{\mathcal{B}}$ for a given observation sequence, can be calculated as

$$\hat{\mathcal{B}} = \arg\max_{\mathcal{B}} p(\mathcal{B}|\mathcal{F}), \tag{2}$$

where the posterior probability takes the exponential form

$$p(\mathcal{B}|\mathcal{F}) = \frac{\exp \sum_k \lambda_k \cdot F_k(\mathcal{B}, \mathcal{F})}{Z_\lambda(\mathcal{F})}. \tag{3}$$

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

$$F_k(\mathcal{B}, \mathcal{F}) = \sum_i f_k(\mathcal{B}, \mathcal{F}, i). \tag{4}$$

where i ranges over all the input positions, and $f_k(\mathcal{B}, \mathcal{F}, i)$ is either a state function $s_k(\mathcal{B}, \mathcal{F}, i)$ of the entire observation sequence and the label transition at position i in the label sequence, or a transition function $t_k(\mathcal{B}, \mathcal{F}, i)$ of the label at position i and the observation sequence [24]. Z_λ is the normalization term, given by

$$Z_\lambda(\mathcal{F}) = \sum_{\mathcal{B}} \exp \sum_k \lambda_k \cdot F_k(\mathcal{B}, \mathcal{F}). \tag{5}$$

The CRF model is trained by globally maximizing the conditional distribution $p(\mathcal{B}|\mathcal{F})$ on a given training set. It can perform trade-off decisions at different sequence positions to achieve a globally optimal labeling. The most likely label sequence is found using the Viterbi algorithm.

When $t_k(\mathcal{B}, \mathcal{F}, i) = t_k(\mathcal{B}_{i-1}, \mathcal{B}_i, \mathcal{F}, i)$, a first-order linear-chain CRF is formed, which includes only two sequential labels in the feature set. For an Nth-order linear-chain CRF, the feature function is defined as $t_k(B_{i-N}, \cdots, B_i, \mathcal{F}, i)$. The probability of a transition between labels depend not only on the current observation, but also on past, future observations and previous labels. Although there are only two classes in our label set, we consider that previous labels affect current decision making. In our task, it is impossible for two adjacent candidates to both be boundaries. In contrast, if several former labels are all non-boundaries, the current candidate has a higher probability of being a boundary. Training is only practical for lower values of N since the computational cost increases exponentially with N. Specifically, if we substitute $\mathcal{F}$ and $\mathcal{B}$ in Eqs. (2) – (5) with F_i and B_i, the CRF model is downgraded to an ME model. The ME classifier individually classifies each data sample without using any contextual information, whereas a CRF models sequential information and performs global optimal labeling.

4. Multimodal Feature Extraction

We extract story boundary features from lexical, audio and video modalities. Lexical features consist of lexical similarity, chain strength and overall cohesiveness; acoustic features involve pause duration, pitch, changes in the speaker and audio event type; visual features contain shot boundaries, anchor faces and news title captions.

4.1 Lexical Features

All lexical features are extracted from Chinese character (rather than Chinese word) unigram sequences based on the Mandarin Large Vocabulary Continuous Speech Recognition (LVCSR) transcripts. The corpora we evaluated the TDT2 (Topic Detection and Tracking) Mandarin audio corpus from Linguistic Data Consortium (LDC) and the homegrown China Central Television (CCTV) video corpus. We also obtained the transcription of the TDT2 corpus from LDC. For the CCTV corpus, we construct our own broadcast news recognizer [25]. The word error rate (WER) and character error rate (CER) are 37% and 20% for TDT2, and 25% and 18% for CCTV, respectively.

Lexical Similarity: Lexical cohesion indicates the lexical relationship between words within a story, while different stories employ different sets of words [2]. As a result,

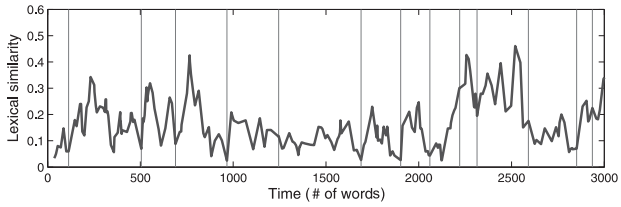


Fig. 3 Lexical similarity curve for a CCTV news episode. Vertical red lines denote the reference story boundaries.

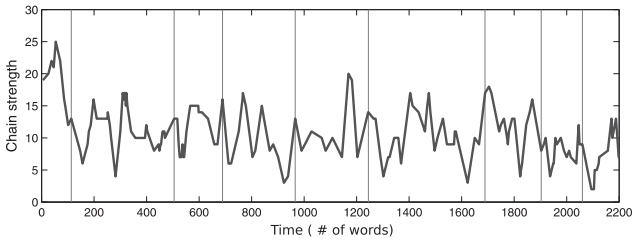


Fig. 4 Chain strength curve for a CCTV news episode. Vertical red lines denote reference story boundaries.

a story boundary may be detected from a shift in word usage or the lexical similarity between sentences. We extract lexical similarity scores as a story boundary feature from the broadcast news transcripts. To capture the variation tendency of lexical similarity, we also compute the difference among, which is denoted as SimDelta. The cosine similarity is calculated at each intersentence position g in the transcripts as follows:

$$\begin{aligned} lexscore(g) =& cos(\mathbf{v}_s, \mathbf{v}_{s+1}) \\ =& \frac{\sum_{i=1}^{I} v_{s,i} v_{s+1,i}}{\sqrt{\sum_{i=1}^{I} v_{s,i} v_{s,i} \sum_{i=1}^{I} v_{s+1,i} v_{s+1,i}}} \end{aligned} \quad (6)$$

where $\mathbf{v}_s$ and $\mathbf{v}_{s+1}$ are the term (i.e. word) frequency vectors for the sentences before and after g, respectively, and $v_{s,i}$ is the frequency of term w_i occurring in sentence s with a vocabulary size of I. Since sentence boundaries are not given in the speech recognition transcripts, we apply a block of fixed-length text as a sentence. Figure 3 shows a lexical similarity curve calculated from the speech recognition transcripts of a CCTV broadcast news episode. There is a good match between the story boundaries and the minima in the similarity curve.

Chain Strength: Lexical chaining is another embodiment of lexical cohesion. A lexical chain links up repeating terms where a chain starts at the first appearance of a term and ends at the last appearance of the term. Owing to lexical cohesion, chains tend to start at the begin of a story and terminate at the end of the story. Therefore, a high concentration of starting and/or ending chains is an indicator of a story boundary [26]. We measure the chaining strength at each inter-sentence position g as

$$chainstrength(g) = endchain(s) + startchain(s+1), \quad (7)$$

where $endchain(s)$ and $startchain(s+1)$ denote the number of chains ending at sentence s and the number of chains beginning at sentence $s+1$ of g, respectively. Similarly, fixed-length text blocks are used as 'sentences'. The variation tendency of chain strength is also adopted as a dimension of lexical features. We set up a maximal chain length and above which no chains are allowed. This is because some terms in a news story may reappear in another story. For example, some chains may span across the entire text if two items of news reporting the same topic are situated at the beginning and end of a news program. In Fig. 4, it shows a chain strength curve of a CCTV broadcast news episode. We can clearly observe that story boundary positions tend to have higher boundary strength scores.

Overall Cohesiveness: When a topic has sharp variations in the lexical distribution, the lexical similarity and chain strength, which focus on local cohesiveness, are reasonably effective. However, sometimes topic transitions among stories in broadcast news are smooth and the distributional variations are very subtle. Therefore, we adopt an overall cohesiveness that directly maximizes the total cohesiveness of all topic fragments extracted from the text. This boundary indicator can effectively detect smooth story changes.

The lexical cohesiveness of a fragment f is defined by

$$Cohscore(f) = A[length(f)] \sum_{i=1}^{I} [R(w_i) S(w_i)], \quad (8)$$

where w_i is the ith term of fragment f. $R(w)$ is the number of repetition of term w, indicating that each pair of identical words contained in fragment f contributes equally to the cohesiveness of f. Thus, the total contribution of word w_i is given by

$$R(w_i) = \sum_{k=1}^{Freq(w_i)-1} k = \frac{1}{2} Freq(w)[Freq(w_i) - 1], \quad (9)$$

where $Freq(w_i)$ is the term frequency of w_i in fragment f.

$S(w_i)$ is used to measure the interfragment discriminability for term w_i, reflecting the fact that terms appearing in more fragments are less useful for discriminating a specific fragment:

$$S(w_i) = \frac{Freq(w_i)}{Total(w_i)}, \quad (10)$$

where $Total(w_i)$ is the number of times that term w_i occurs in the whole text.

As a normalization factor, $A(length)$ should be decreased slowly when $length(f)$ is reasonably small as it should not offset the cohesiveness gained by the increase in word repetition. If a fragment is much longer than the average length of the topic, $A(length)$ should provide a considerable negative effect as a penalty factor. We found that an exponential function with a base close to 1.0 serves our needs well. Formally, the length factor is defined as

$$A(length(f)) = \alpha^{-length(f)}, \quad (11)$$

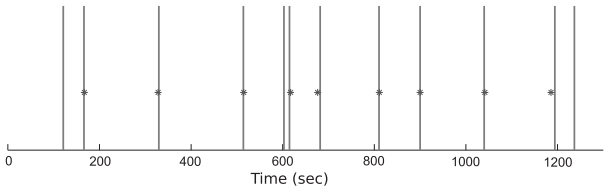


Fig. 5 Detected story boundaries (blue stars) obtained by the overall cohesiveness-based indicator compared with reference topic boundaries (vertical red lines) for a CCTV news episode.

where α is a constant parameter slightly larger than 1.0.

We define the *overall cohesiveness* of a text segment as the sum of the cohesiveness values of all fragments extracted from it, i.e.,

$$C(text) = \sum_{i=1}^{I} Cohescore(f_i). \quad (12)$$

To obtain the optimal text segments, we adopt the segmentation scheme for $C(text)$ by using a dynamic programming algorithm. Assume that the whole text consists of n words, represented as $w_1 w_2 \dots w_n$. Let $F(n)$ denote the objective function, i.e.

$$F(n) = max[C(w_1 w_2 \dots w_n)]. \quad (13)$$

The dynamic programming is conducted as follows:

$$F(i) = max_{0<=j<=i}[F(j) + Cohscore(w_{j+1} \dots w_i)], \quad (14)$$

with $F(0) = 0$. Figure 5 shows the segmentation results (blue stars) obtained by the overall cohesiveness-based indicator compared with reference story boundaries (vertical red lines) for a CCTV news episode. We align the boundary of each segment to its nearest pause (i.e., story boundary candidate) as the boundary indicator.

4.2 Audio Features

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 semantic boundaries. The pause duration between different stories usually lasts longer than that between sentences. Broadcast news producers usually insert a clear silence or a music clip between news stories. Previous works have shown that pause duration is effective for the story segmentation of broadcast news [5], [6]. Figure 6 shows the pause duration time trajectory of a VOA (Voice of America) broadcast news episode. We can clearly see the pattern of pauses, where the pause duration is much more salient at story boundaries. We used a home-grown audio classifier [27] to label a broadcast news audio stream into clips of six types: music, pure speech, speech with background sound, speech with music, background sound and silence. Here, for all the detected silences, pause duration is regarded as a prosodic feature, namely PauD.

Audio Event Type: According to the editorial rules

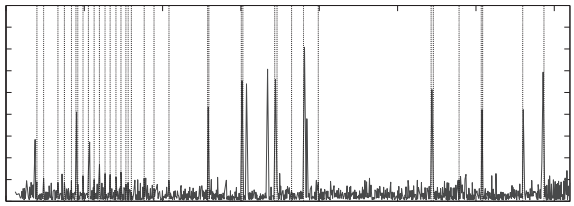


Fig. 6 Pause duration time trajectory for a VOA broadcast news episode. Dotted vertical lines denote story boundaries.

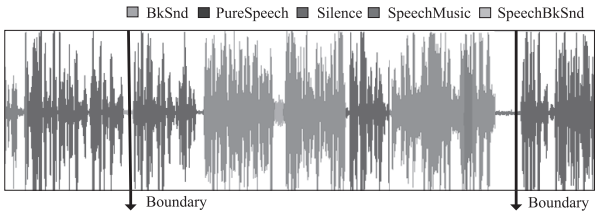


Fig. 7 Clip of annotated CCTV broadcast news audio. News stories usually starts from clean speech.

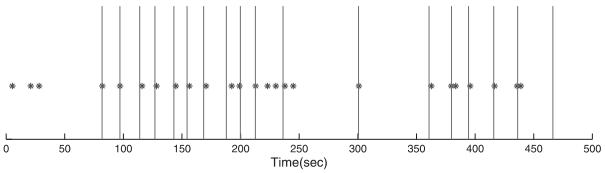


Fig. 8 Detected speaker changes (blue stars) compared with reference story boundaries (vertical red lines) for a brief news audio clip in the TDT2 corpus.

of broadcast news, studio-to-field transitions often coincide with internews boundaries; a news story usually starts from clean speech (e.g., anchor speech in the studio) and rarely starts from noisy speech (e.g., field speech), as shown in Fig. 7. Studio speech is generally clean while field speech is often contaminated with diverse background noises from news scenes such as streets, factories and buildings. Therefore, the changes in the audio event type may indicate potential topic boundaries. We use an SVM binary tree (SVM-BT) approach [27] to hierarchically classify audio clips into six classes: pure speech, speech with noise, speech with music, music, silence and noise. The SVM-BT architecture can realize coarse-to-fine multiclass classification with high accuracy and efficiency.

Speaker change: Broadcast news programs usually contain various speakers, such as anchors, reporters and interviewees. Many news sessions are hosted by two anchors who report news in turn. For example, a male anchor and a female anchor often alternate with each other to announce the news in a news session. Figure 8 shows an example where most of the detected speaker changes are at story boundaries. Some news programs follow a clear syntax: a news story is introduced by an anchor in the studio, which is then followed by a detailed report from a field reporter or an interview. Therefore, in broadcast news, changes in speakers may coincide with story transitions. We use a two-stage

multifeature integration approach to automatically detect changes in speaker from broadcast news audio [28]. Speaker change is used as a binary feature (Change/Not-change for each candidate).

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 [7]. Therefore, pitch undergoes a declination within a major speech unit and a reset between two major speech units. Pitch declination and reset behaviors are observed more often at topic level than at smaller speech levels such as utterances [5], [6], [21].

In this study, we extract the pitch trajectory from broadcast news audio using the YIN pitch tracker [29]. The nearest left and right successive pitch contours of each boundary candidate (i.e., pause segment) are determined as our regions of interest. A set of three pitch features are extracted from each boundary candidate: the 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 the pitch contour by the speaker before pitch feature calculation. The speaker boundaries are automatically determined by the detected changes of speaker [28].

4.3 Video Features

Shot Boundary: It can be observed that, in broadcast news video, news story transitions are usually accompanied with a shot change. Therefore, it is reasonable to investigate whether there is a shot change at a story boundary candidate. We measure the block histogram difference between two adjacent video frames to decide whether a shot boundary exists. First, a frame k is divided into $M * N$ blocks and a gray-scale histogram $h(m, n, k)$ is calculated for each block (m, n). The histogram difference between frames k and $k + 1$ is calculated as

$$D(k, k+1) = \sum_{m=1}^{M} \sum_{n=1}^{N} |h(m, n, k) - h(m, n, k+1)|. \quad (15)$$

A shot boundary is detected if the calculated distance $D(k, k + 1)$ is larger than a preset threshold. Shot boundary is used as a binary feature (yes, no) for each story boundary candidate.

Anchor Face: According to the structural rules of broadcast news, many news stories begin with a studio anchor shot and then move to field shots. Previous research shows that the presence of an anchor face is an important visual cue for story boundary detection [18], [30]. We first use an AdaBoost detector to detect human faces in video frames, and then use a regression classifier to discriminate anchor faces from other detected non-anchor faces. On the basis of the characteristics of the anchor appearances, such as, face coordinates and size, the classifier labels video frames with anchor face counts (0,1,2). Figure 9 shows the anchor face counts for a CCTV news episode. We can clearly see that the anchor face count changes at some story boundaries. Therefore, we use the interframe anchor face count difference at candidate positions as a visual feature for story boundary detection.

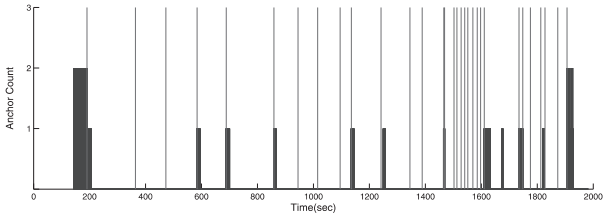


Fig. 9 Anchor face counts (blue bars) compared with reference story boundaries (vertical red lines) for a CCTV news episode.

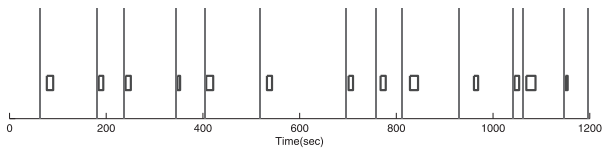


Fig. 10 Appearance of detected title captions (blue boxes) compared with reference story boundaries (vertical red lines) for a CCTV news episode.

News Title Caption: In broadcast news video, a news story is often accompanied by a caption indicating the title of the news. Hence, the appearance of a title caption is a clear story boundary indicator. We detect news title captions from broadcast news video on the basis of the color and structural information in the caption region. Since the title caption usually appears later than the news, we measure the time distance from a boundary candidate to the appearance of the next title caption as a feature. In Fig. 10, the blue boxes indicate the appearance of title captions and their durations. We can clearly see that almost every story boundary is associated with the subsequent appearance of a title caption.

5. Experiments

5.1 Experimental Setup

We carried out story segmentation experiments on two Mandarin broadcast news corpora, the LDC TDT2 Mandarin audio corpus and the homegrown CCTV video corpus, to evaluate the proposed approach. Table 1 shows details of the two corpora and the data organization in experiments. We extracted audio, video and lexical features for the CCTV video corpus, and audio and lexical features for the TDT2 audio corpus. We conducted the experiments using feature sets from a single modality (L, A, V) and integrated feature sets from multiple modalities (L+A, L+A+V). The full list of feature sets is shown in Table 2. Note that the position of the candidate (at the beginning of the broadcast news episode), namely Pos, was inserted into all the feature sets in the experiments. The Pos feature was used for time-

Table 1 Corpora for story segmentation experiments.

Corpus		TDT-2 Mandarin	CCTV
Source		VOA newscast	China Central TV
Media		audio, text	audio, video, text
No. of programs		177	71
Audio duration		53h	27h
LVCSR WER		37%	25%
Data assignment	Training	90 programs (1321 bnds)	40 (1209 bnds)
	Testing	87 programs (1262 bnds)	31 (892 bnds)

Table 2 Lexical, audio and video feature sets used in the experiments.

set	Features	Abbreviation	Value
Lexical	Lexical Similarity	LexSim	Continuous
	Similarity Variation	SimDelta	Continuous
	Chain Strength	ChStr	Continuous
	Chain Variation	ChainDelta	Continuous
	Global Cohesiveness	GlbCoh	Binary
Audio	Pause Duration	PseDur	Continuous
	Speaker Change	SpkChg	Binary
	Audio Event Type	AETyp	Discrete
	Pitch Left Mean	PLmn	Continuous
	Pitch Right Mean	PRmn	Continuous
	Pitch Reset	PReset	Continuous
Video	Shot Boundary	ShotBnd	Binary
	Anchor Count	AchrCnt	Triple
	Caption Distance	CapDist	Continuous

Table 3 Accuracy rates of feature extraction methods.

Features	ShotBnd	AchrCnt	Caption	AETyp	SpkChg
Accuracy	0.935	0.967	0.948	0.960	0.813

dependent heuristics. Table 3 gives the accuracies of detection in the cases of shot boundary detection, anchor counts, caption detection, audio event type detection and speaker change detection, which were tested on an extra set for validation. 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 evaluation standard, a detected story boundary is considered correct if it lies within a 15s tolerance window on each side of a manually annotated reference boundary.

Since the recorded broadcast news audio may include channel noises, we consider background sound positions together with silence and music positions as the story boundary candidates in the experiments. After feature extraction, some features, such as, GlbCoh, SpkChg and ShotBnd, must be aligned to an appropriate candidate because they do not always appear exactly at a candidate position. We aligned GlbCoh and ShotBnd to the nearest pause. Speaker change (SpkChg) points were matched with the nearest candidates on the left owing to the existence of a detection delay.

To maintain a reasonable dynamic range of feature values, we normalize all the continuous features to [0,1] using the formula

$$\mathcal{F}_v = \frac{F_v - F_{min}}{F_{max} - F_{min}}. \tag{16}$$

Table 4 Experimental results for CRF with different N and M on the CCTV corpus.

Orders(N) \ Context(M)			0	1	2	3
Training	Lexical	1	0.7044	0.6981	0.7244	0.7354
		2	0.7243	0.7478	0.7451	**0.7581**
	Acoustic	1	0.7254	0.7334	0.7640	0.7604
		2	0.7438	0.7600	**0.7844**	0.7534
	Visual	1	0.5209	0.5769	0.5207	0.6458
		2	0.5207	**0.7178**	0.6888	0.7038
	L+A	1	0.7995	0.8226	0.8329	0.8263
		2	0.8140	0.8371	**0.8432**	0.8364
	L+A+V	1	0.8379	0.8528	0.8625	0.8729
		2	0.8374	0.8686	0.8766	**0.8832**
Testing	Lexical	1	0.6901	0.6830	0.7141	0.7119
		2	0.7006	0.7198	0.7310	**0.7361**
	Acoustic	1	0.7204	0.7334	0.7416	0.7420
		2	0.7404	**0.7518**	0.7365	0.7367
	Visual	1	0.5524	0.5299	0.5601	0.5607
		2	**0.7046**	0.6760	0.6992	0.6887
	L+A	1	0.7955	0.8086	0.8180	**0.8204**
		2	0.7965	0.8078	0.8095	0.8139
	L+A+V	1	0.8366	0.8516	**0.8576**	0.8559
		2	0.8397	0.8334	0.8565	0.8443

Table 5 Experimental results for CRF with different N and M on the TDT2 corpus.

Orders(N) \ Context(M)			0	1	2	3
Training	Lexical	1	0.6393	0.6350	0.6407	0.6296
		2	0.6702	**0.6769**	0.6738	0.6734
	Acoustic	1	0.7786	0.7743	0.7978	0.7989
		2	0.7826	0.7808	0.7691	**0.7897**
	L+A	1	0.7694	0.7883	0.8007	0.7988
		2	0.7768	0.7793	0.8004	**0.8229**
Testing	Lexical	1	0.6708	0.6652	0.6690	0.6629
		2	0.6964	0.7039	0.7034	**0.7175**
	Acoustic	1	0.7051	0.7056	0.7140	0.7241
		2	0.7094	0.6908	0.7123	**0.7269**
	L+A	1	0.7492	0.7665	0.7768	0.7802
		2	0.7717	0.7729	0.7847	**0.7981**

5.2 Story Segmentation with CRF

We trained a CRF boundary/nonboundary classifier using labeled candidates in a training set. We adopted the GRMM toolkit[†] to perform CRF training and testing after modifying it to support real-valued feature inputs. Different CRF orders N ($\mathcal{B} = \mathcal{B}_{i-N}, \cdots, \mathcal{B}_i$) and feature contexts M ($\mathcal{F} = \mathcal{F}_{i-M}, \cdots, \mathcal{F}_i, \cdots, \mathcal{F}_{i+M}$) were tested in order to achieve the best story segmentation performance. The order N is limited to 2 owing to the exponential computation cost for high orders and the data sparseness problem. The feature context M indicates the number of preceding and following features that are used in addition to the current $\mathcal{F}_i$.

Tables 4 and 5 show the story segmentation results using a CRF for the CCTV and TDT2 corpora, respectively. We also report the performance of training data used to compare evaluations. The results show that (1) with an increase

[†]http://mallet.cs.umass.edu/grmm/

Table 6 Experimental results (F1-measure) for different feature sets and classifiers on CCTV. L+A+V*: feature selection performed.

Classifier	Lexical	Acoustic	Visual	L+A	L+A+V	L+A+V*	Removed from feature selection
DT	0.6688	0.7224	0.6974	0.7744	0.8174	0.8187	AchrCnt
BN	0.6974	0.7361	0.6753	0.7907	0.8431	0.8432	PReset
NB	0.6696	0.5624	0.4087	0.7184	0.7453	0.7571	AETyp SimDelta ChStr
MLP	0.6934	0.7244	0.6934	0.7842	0.8054	0.8231	AchrCnt PRmn ChainDelta
SVM	0.6769	0.7125	0.4244	0.7845	0.8077	0.8077	—
ME	0.6767	0.6745	0.5881	0.7606	0.7973	0.8006	PLmn PRmn PReset ChainDelta
CRF	**0.7361**	**0.7518**	**0.7046**	**0.8204**	**0.8576**	**0.8607**	PReset

in the sequental/contextual information (M), the story segmentation performance is generally improved on both corpora; (2) multimodal feature integration significantly outperforms a single-modal feature set in terms of story segmentation. The best F1-measures for the lexical feature set (L), acoustic feature set (A) and visual feature set (V) on testing set are 0.7361, 0.7518, 0.7046 on the CCTV corpus, respectively. For the TDT2 corpus, the lexical feature set (L) and acoustic feature set (A) achieve F1-measure scores of 0.7175 and 0.7269, respectively. These results show that the features obtained from the three modalities can achieve comparable story segmentation performance. When features from different modalities are combined, the F1-measure is increased to 0.8204 (L+A, $N = 1, M = 3$) and 0.8576 (L+A+V, $N = 1$, $M = 2$) on the CCTV corpus and 0.7981 (L+A, $N = 2$, $M = 3$) on the TDT2 corpus. We found that results based on the CCTV corpus were always better than those on the TDT2 corpus. This is probably because of the different genres and style between CCTV and TDT2. For example, for CCTV broadcast news, at the end of programs, there are brief news stories that only contain one or two sentences which the anchors report in turn. For such brief stories, speaker change and pitch reset features are more effective in as indicators.

5.3 Comparison with Different Classifiers

For performance comparison, we also tested several popular classifiers, i.e., a C4.5 decision tree (DT), a naive Bayesian classifier (NB), RBF-kernel support vector machines (SVMs), multilayer perceptron (MLP), a Bayesian network (BN) and the maximum entropy classifier (ME). The Weka tookit[†] was used to train the DT, NB, SVM, MLP, BN and SVM classifiers, and the ME classifier was trained using the opennlp.maxent package[††].

Since some features may have low discriminative ability, we performed a feature selection procedure to find the optimal feature subset with the highest F1-measure. We adopted the backward elimination algorithm to search for the optimal subset by iteratively eliminating features whose absence did not decrease performance on different classifiers and corpora. Parameter tuning, classifier training and feature selection were performed on the training set and experimental results were reported on the testing set. All classifiers were equally tuned to obtain the best performance.

Experimental results for different classifiers on the CCTV and TDT2 corpora are listed in Tables 6 and 7, respectively. We clearly observe that the CRF classifier outperforms other classifiers for both individual feature sets and integrated feature sets on the two tested corpora. From the feature selection, we found that not all features contribute to story boundary detection for a particular classifier. Some features were removed owing to their low discriminative ability or because of the lower correlation with other more effective features. After feature selection, the highest F1-measure for the two corpora were 0.8607 (for CCTV) and 0.7981 (for TDT2). We also notice that feature selection approach selects different optimal subsets for different corpora and different classifiers.

Table 7 Experimental results (F1-measure) for different feature sets and classifiers on TDT2. L+A*: feature selection performed.

Classifier	Lexical	Acoustic	L+A	L+A*	Removed from feature selection
DT	0.6829	0.7144	0.7381	0.7566	PRmn
BN	0.6744	0.6936	0.7652	0.7712	LexSim
NB	0.5934	0.6543	0.6960	0.7313	AETyp ChStr SimDelta
MLP	0.6652	0.7249	0.7654	0.7654	—
SVM	0.7076	0.7038	0.7572	0.7583	PLmn
ME	0.6687	0.6848	0.7441	0.7441	—
CRF	**0.7175**	**0.7269**	**0.7981**	**0.7981**	—

6. Conclusion

In this paper, we propose the integration of multimodal features using conditional random fields (CRFs) for the automatic segmentation of broadcast news stories. Features from different modalities, i.e., audio, visual and lexical modalities, are extracted for sequential boundary/nonboundary tagging of a story boundary candidate set. Sequential interlabel relations and contextual information are effectively captured by a linear-chain CRF. Experimental results for story segmentation have shown that (1) the CRF approach outperforms other competitive classifiers, i.e., DT, BN, NB, SVM and MLP; (2) multimodal feature integration shows significantly improved performance compared with features from single modalities.

Acknowledgement

This work was supported by the National Natural Science

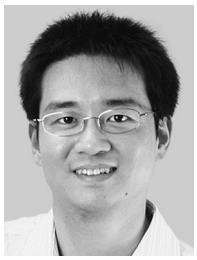
[†]http://www.cs.waikato.ac.nz/ml/weka/
[††]http://opennlp.sourceforge.net/

Foundation of China (60802085, 61175018), the China MOE Program for New Century Excellent Talents in University (2008), the Natural Science Basic Research Plan of Shaanxi Province (2011JM8009), and the Key Science and Technology Program of Shaanxi Province (2011KJXX29).

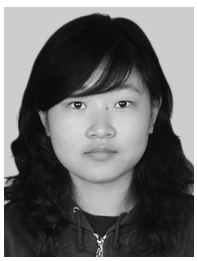
References

[1] M. Franz, J. McCarley, T. Ward, and W. Zhu, "Segmentation and Detection at IBM: Hybrid Statistical Models and Two-tiered Clustering," TDT-3 Workshop, 2000.
[2] M.A. Hearst, "TextTiling: Segmenting text into multi-paragraph subtopic passages," Computational Linguistics, vol.23, no.1, pp.33–64, 1997.
[3] X. Wang, L. Xie, B. Ma, E. Chng, and H. Li, "Phoneme lattice based TextTiling towards multilingual story segmentation," Proc. Interspeech, pp.1305–1308, 2010.
[4] S. Chan, L. Xie, and H. Meng, "Modeling the statistical behavior of lexical chains to capture word cohesiveness for automatic story segmentation," Proc. Interspeech, 2007.
[5] 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.
[6] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," Speech Commun., vol.32, no.1-2, pp.127–154, 2000.
[7] C.Y. Tseng, S.H. Pin, Y. Lee, H.M. Wang, and Y.C. Chen, "Fluent speech prosody: Framework and modeling," Speech Commun., vol.46, pp.284–309, 2005.
[8] W. Hsu, S. Chang, C. Huang, L. Kennedy, C. Lin, and G. Iyengar, "Discovery and fusion of salient multi-modal features towards news story segmentation," SPIE Electronic Imaging, 2004.
[9] H. Winston, H. Hsu, and S. Chang, "A statistical framework for fusing mid-level perceptual features in news story segmentation," International Conference on Multimedia and Expo, 2003.
[10] J. Yamron, I. Carp, L. Gillick, S. Lowe, and P. Van Mulbregt, "A hidden Markov model approach to text segmentation and event tracking," ICASSP, pp.333–336, 1998.
[11] J. Zhang, L. Xie, W. Feng, and Y. Zhang, "A subword normalized cut approach to automatic story segmentation of Chinese broadcast news," Information Retrieval Technology, pp.136–148, 2009.
[12] M. Lu, C. Leung, L. Xie, B. Ma, and H. Li, "Probabilistic latent semantic analysis for broadcast news story segmentation," Proc. Interspeech, pp.1301–1304, 2011.
[13] A. Hauptmann and M. Witbrock, "Story segmentation and detection of commercials in broadcast news video," IEEE International Forum on Research and Technology Advances in Digital Libraries, pp.168–179, 2002.
[14] G. Tür, D. Hakkani-Tür, A. Stolcke, and E. Shriberg, "Integrating prosodic and lexical cues for automatic Topic segmentation," Computational Linguistics, vol.27, no.1, pp.31–57, 2001.
[15] L. Xie, L. Zheng, Z. Liu, and Y. Zhang, "Laplacian eigenmaps for automatic story segmentation of broadcast news," IEEE Trans. Speech Audio Process., vol.20, no.1, pp.264–277, 2012.
[16] W. Qi, L. Gu, H. Jiang, X. Chen, and H. Zhang, "Integrating visual, audio and text analysis for news video," Proc. ICIP, pp.520–523, 2002.
[17] A. Rosenberg and J. Hirschberg, "Story segmentation of broadcast news in english, mandarin and arabic," Proc. HLT-NAACL, pp.125–128, 2006.
[18] W. Hsu, L. Kennedy, S. Chang, M. Franz, and J. Smith, "Columbia-IBM news video story segmentation in TRECvid 2004," CIVR, 2005.
[19] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," Proc. 18th ICML, pp.282–289, 2001.
[20] F. Sha and F. Pereira, "Shallow parsing with conditional random fields," Proc. HLT-NAACL, pp.213–220, 2003.
[21] 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 Trans. Audio Speech Language Process., vol.14, no.5, pp.1526–1540, 2006.
[22] G. Levow, "Automatic prosodic labeling with conditional random fields and rich acoustic features," Proc. IJCNLP, 2008.
[23] J. Morris and E. Fosler-Lussier, "Combining phonetic attributes using conditional random fields," Proc. Interspeech, pp.597–600, 2006.
[24] H.M. Wallach, "Conditional random fields: An introduction," Tech. Rep., 2004.
[25] L. Xie, Y. Yang, and Z.Q. Liu, "On the effectiveness of subwords for lexical cohesion based story segmentation of Chinese broadcast news," Information Sciences, vol.181, no.13, pp.287–2891, 2011.
[26] N. Stokes, J. Carthy, and A.F. Smeaton, "SeLeCT: A lexical cohesion based news story segmentation system," AI Communication, vol.17, no.1, pp.3–11, Jan. 2004.
[27] 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, vol.17, no.2, pp.101–112, 2011.
[28] L. Xie and G. Wang, "A two-stage multi-feature integration approach to unsupervised speaker change detection in real-time news broadcasting," ISCSLP, pp.1–4, 2008.
[29] A. de Cheveigné and H. Kawahara, "YIN, A fundamental frequency estimator for speech and music," J. Acoustical Society of America, vol.111, p.1917, 2002.
[30] C. Ma, B. Byun, I. Kim, and C. Lee, "A detection-based approach to broadcast news video story segmentation," ICASSP, pp.1957–1960, 2009.

Xiaoxuan Wang received the M.S. and B.S. degree in computer science from Northwestern Polytechnical University, Xi'an, China, in 2011 and 2008, respectively. She is currently a Research Assistant with School of Computing, National University of Singapore. From 2009 to 2010, she was with School of Computer Engineering, Nanyang Technological University, Singapore as a visiting student. She has published papers on major proceedings such as Interspeech, APSIPA. Her current research interest includes speech and language processing and spoken document retrieval



Lei Xie received the Ph.D. degree in computer science from Northwestern Polytechnical University, Xi'an, China, in 2004. He is currently a Professor with School of Computer Science, Northwestern Polytechnical University, Xi'an, China. From 2001 to 2002, he was with the Department of Electronics and Information Processing, Vrije Universiteit Brussel (VUB), Brussels, Belgium as a visiting scientist. From 2004 to 2006, he was a senior research associate in the Center for Media Technology (RCMT), School of Creative Media, City University of Hong Kong, Hong Kong. From 2006 to 2007, he was a postdoctoral fellow in the Human-Computer Communications Laboratory (HCCL), Department of Systems Engineering and Engineering Management, the Chinese University of Hong Kong. He has published more than 60 papers in major journals and proceedings, such as IEEE Transactions on Multimedia, IEEE Transactions on Audio, Speech and Language Processing, Information Sciences, Pattern Recognition, ACM/Springer Multimedia Systems, Interspeech, ICPR and ICASSP. His current research interests include speech and language processing, multimedia and human computer interaction.



Mimi Lu received the B.S. degree in electronic commerce from Northwestern Polytechnical University, Xi'an, China, in 2009, where she is currently working toward the M.S. degree in the School of Computer Science. From 2010 to 2011, she was with the Human Language Technology Department, Institute of Infocomm Research, A STAR, Singapore as an intern. Her current research interest includes speech and language processing and human computer interaction. She has published papers on major proceedings such as Interspeech, ISCSLP.

Bin Ma received the B.Sc. degree in Computer Science from Shandong University, China, in 1990, the M.Sc. degree in Pattern Recognition & Artificial Intelligence from the Institute of Automation, Chinese Academy of Sciences (IACAS), China, in 1993, and the Ph.D. degree in Computer Engineering from The University of Hong Kong, in 2000. He was a Research Assistant from 1993 to 1996 at the National Laboratory of Pattern Recognition in IACAS. In 2000, he joined Lernout & Hauspie Asia Pacific as a Researcher working on speech recognition. From 2001 to 2004, he worked for InfoTalk Corp., Ltd, as a Senior Researcher and a Senior Technical Manager for telephony speech recognition. He joined the Institute for Infocomm Research, Singapore in 2004. Now he works as a Senior Scientist and the Group Leader of Speech Processing Group. He is a Senior Member of IEEE, and is serving as a Subject Editor of speech communication. His current research interests include robust speech recognition, speaker & language recognition, spoken document retrieval, natural language processing and machine learning.

Eng Siong Chng is currently an Assistant Professor in the School of Computer Engineering, Nanyang Technological University, Singapore. He received his BEng (Hons) in Electrical and Electronics Engineering from the University of Edinburgh, U.K in 1991, and Ph.D. from the same University in 1996. Prior to joining NTU in 2003, he has worked in the following research centres and companies: 1) 1996, Institute of Physics and Chemical Research, Riken (http://www.bsp.brain.riken.jp/), as a post-doc working in the area of signal processing and classification, 2) 1996–1999, Institute of System Science (ISS, currently known as I2R http://www.i2r.a-star.edu.sg/) as a research staff to transfer the Apple-ISS speech and handwriting technologies to ISS, 3) 1999–2000, Lernout and Hauspie (now part of nuance http://www.nuance.com/) as a senior researcher in speech recognition, and 4) 2001–2002, Knowles Electronics (http://www.knowles.com/search/products/array_technologies.jsp) as a manager for the Intellisonic microphone array research. 5) 2003 onwards: Asst Professor, NTU's School of Computer Engineering. His research interests are in pattern recognition, signal, speech and video processing. He has published over 90 papers in international journals and conferences. He is currently leading the speech and language technology program (http://www3.ntu.edu.sg/home/aseschng/SpeechTechWeb/default.htm) in Emerging Research Lab at the School of Computer Engineering, NTU. Eng Siong is a senior member of IEEE since 2005.

Haizhou Li received the B.Sc., M.Sc., and Ph.D. degrees in electrical & electronic engineering from the South China University of Technology (SCUT), Guangzhou, in 1984, 1987, and 1990, respectively. Dr Li was a Research Assistant from 1988 to 1990 at the University of Hong Kong, Hong Kong, China. In 1990, he joined SCUT as an Associate Professor. From 1994 to 1995, he was a Visiting Professor at CRIN, Nancy, France. In 1995, he became the Manager of the ASR group at the Apple-ISS Research Centre in Singapore where he led the research of Apple's Chinese Dictation Kit for Macintosh. In 1999, he was appointed Research Director of Lernout & Hauspie Asia Pacific, where he oversaw the creation of the multimodal speech, pen and keyboard input solution for Chinese. From 2001 to 2003, he was the Vice President of InfoTalk Corp. Ltd. Since 2003, he has been with the Institute for Infocomm Research (I^2R), Singapore, where he is now the Principal Scientist and Head of Human Language Technology Department. His current research interests include automatic speech recognition, speaker and language recognition and natural language processing. Dr Li was named one of the two Nokia Visiting Professors 2009 by the Nokia Foundation in recognition of his contributions to Speaker and Language Recognition research. He was a recipient of the National Infocomm Award 2001 and the TEC Innovator's Award 2004 in Singapore. He is now an Associate Editor for *Springer International Journal of Social Robotics, IEEE Transactions on Audio, Speech and Language Processing*, and ACM Transactions on Speech and Language Processing. He is also an elected Board Member of International Speech Communication Association (2009–2013).