# Laplacian Eigenmaps for Automatic Story Segmentation of Broadcast News

Lei Xie, *Member, IEEE*, Lilei Zheng, Zihan Liu, and Yanning Zhang, *Member, IEEE*

***Abstract*—We propose Laplacian Eigenmaps (LE)-based approaches to automatic story segmentation on speech recognition transcripts of broadcast news. We reinforce story boundaries by applying LE analysis to sentence connective strength matrix and reveal the intrinsic geometric structure of stories. Specifically, we construct a Euclidean space in which each sentence is mapped to a vector. As a result, the original inter-sentence connective strength is reflected by the Euclidean distances between the corresponding vectors and cohesive relations between sentences become geometrically evident. Taking advantage of LE, we present three story segmentation approaches: LE-TextTiling, spectral clustering and LE-DP. In LE-DP, we formalize story segmentation as a straightforward criterion minimization problem and give a fast dynamic programming solution to it. Extensive story segmentation experiments on three corpora demonstrate that the proposed LE-based approaches achieve superior performances and significantly outperform several state-of-the-art methods. For instance, LE-TextTiling obtains a relative F1-measure increase of 17.8% on CCTV Mandarin BN corpus as compared to conventional TextTiling; LE-DP achieves a high F1-measure of 0.7460, which significantly outperforms a recent CRF-prosody approach with an F1-measure of 0.6783 on TDT2 Mandarin BN corpus.**

***Index Terms*—Laplacian Eigenmaps (LE), spoken document retrieval, story segmentation, topic segmentation.**

## I. Introduction

STORY segmentation is to divide a text, audio, or video stream into a set of story units, each of which presents a central topic [1]. With the exponential proliferation of multimedia contents, there is an urgent need for automatic and effective story segmentation techniques. This is because story segmentation serves as an important precursor for a variety of media content management tasks, including topic detection and tracking, summarization, information extraction, content indexing, and retrieval [2]. For example, users of a broadcast news retrieval system usually expect short story clips relevant to a topic of interest rather than an entire news program.

Story segmentation task originates from early approaches on text segmentation [3]–[5]. With the overwhelming growth of multimedia repositories from broadcast media and the Internet, researchers have been focusing on segmenting audio/video contents, such as broadcast news [6], meetings [7], and lectures [8], etc. Lexical cohesion-based methods have drawn much interest for story segmentation on texts and speech recognition transcripts of spoken documents [7], [9]–[13]. Lexical cohesion [14] indicates that words in a story (or topic) hang together by semantic relations and different stories tend to employ different sets of words. Therefore, inter-sentence cohesive strength is measured across the text and story boundaries detection is performed on the sentence connective strength matrix.

This paper presents effective lexical cohesion approaches using Laplacian Eigenmaps (LE) for story segmentation on broadcast news transcripts. LE is a geometrically motivated algorithm recently proposed for data representation [15], [16]. We reinforce story boundaries by applying LE analysis to sentence connective strength matrix and reveal the intrinsic geometric structure of stories. We first construct a Euclidean space in which each sentence is mapped to a vector. As a result, the original inter-sentence connective strength is reflected by the Euclidean distances between the corresponding vectors and cohesive relations between sentences become geometrically evident. Based on the LE representation of sentence connective strength matrix, we then present three story segmentation approaches that can significantly improve performances as compared with several state-of-the-art methods.

The remainder of this paper is organized as follows. Section II briefly summarizes the related work on story segmentation and Section III presents the motivations of this study. Section IV describes our sentence connective strength measure and the sentence distance penalty factor. In Section V, we describe the LE method and its natural connection to story segmentation. Section VI presents the three LE-based story segmentation approaches, including LE-TextTiling, spectral clustering and LE-DP. The experimental evaluations and detailed analysis on the results are described in Section VII. Finally, we summarize this study and point out our future work in Section VIII.

## II. Related Work

Substantial work has been reported in the story segmentation literature. In general, approaches can be categorized to *detection*-based [6], [9]–[11], [13], [17], [18], and *model*-based [5], [8], [19], [20]. Detection-based methods directly locate story

Manuscript received November 20, 2010; revised March 16, 2011; accepted June 13, 2011. Date of publication June 27, 2011; date of current version November 11, 2011. This work was supported in part by the National Natural Science Foundation of China (60802085), 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). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Haizhou Li.

L. Xie, L. Zheng, and Y. Zhang are with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China (e-mail: lxie@nwpu.edu.cn; xielei21st@gmail.com; lzheng@nwpu-aslp.org; zhenglilei@mail.nwpu.edu.cn; ynzhang@nwpu.edu.cn).

Z. Liu is with the School of Creative Media, City University of Hong Kong, Hong Kong, China (e-mail: zhliu3@student.cityu.edu.hk; updogliu@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2011.2160853

boundaries through a set of intuitive cues or features; a detector or classifier is learned from the cues to make boundary identification. Model-based methods focus on topic modeling and segment a document into story units under some optimal criterion.

In order to detect story boundaries, various boundary cues have been explored in different modalities. Visual cues, e.g., anchor face and scene transition, are widely studied in the story segmentation task of TREC video retrieval evaluations (TRECVID) [18], [21]. For example, the presence of anchors is a salient indication of story boundaries in a broadcast news video [21]. Audio signal implies rich structural information for story segmentation. Broadcast news programs often use music clips, speaker changes, and significant pauses to signal news transitions [18], [22]. Speech prosodic cues have lately raised interest in event detection tasks [23], including story segmentation [6], [24]–[26]. Speakers naturally separate their speech to paragraphs, topics or subtopics through particular intonational, durational and energetic behaviors [27]. Shriberg *et al.* [24] and Tür *et al.* [26] proposed prosody-based story approaches that adopt a rich set of pause, pitch and intensity cues. Our previous work discovered that the pitch reset feature is affected by Chinese syllabic tones and tone normalized pitch reset is more effective for Chinese story segmentation [28], [29]. Recently, Wang *et al.* [22] have integrated multiple prosodic features using linear-chain conditional random fields (CRF). They have shown that CRF outperforms decision tree (DT), support vector machines (SVMs), and maximum entropy (ME) in prosody-based story segmentation.

Audio/visual cues depend on the editorial and production rules and these rules often vary from different media sources. In contrast, lexical cues are more generic since they probe story shifts by monitoring semantic variations across the text. Moreover, lexical cues can be extracted from both pure texts and multimedia sources such as speech recognition transcripts and video captions. Main lexical approaches involve lexical cohesion[7], [9]–[11], use of cue phrases [21], [30] and topic modeling [5], [19], [20]. Lexical cohesion is a textual quality that makes the sentences in a topic seems to hang together via inter-word semantic relations [14]. Text segments with similar vocabulary are more likely to be parts of a coherent topic. Repetition (i.e., co-occurrence) of words is the most common appearance of the lexical cohesion phenomenon. Based on this principle, much effort has been devoted to lexical cohesion approaches for text and story segmentation. Major approaches involve TextTiling [9], C99 [4], and lexical chaining [10].

TextTiling [9] is a typical lexical-cohesion-based text segmentation approach that has been recently introduced to segment spoken documents such as broadcast news (BN) [6], [12] and meetings [7]. The intuitive idea of TextTiling is that different topics usually employ different sets of words and shifts in vocabulary use are indications of topic changes. Hence, pairwise sentence similarities are measured across the text and a local similarity minimum implies a story boundary. Stokes *et al.* [10] embodied word cohesion by lexical chaining for story segmentation. In their SeLeCT system [10], related words are linked into chains and a high concentration point of chain starting and ending is an indication of a story boundary. Conventional lexical chaining method rigorously counts chain heads and tails at inter-sentence positions. Chan *et al.* [11] proposed to use the log-normal distribution to capture the statistical behavior of lexical chains for a more effective story segmentation approach. Some other lexical-cohesion-based segmentation methods aim at finding the optimal segmentation under some optimal criteria [8], [31], [32]. Recently, TextTiling and lexical chaining have been implemented on subword sequences (i.e., character/syllable $n$-grams) of Chinese broadcast news transcripts because subword units show robustness to speech recognition errors [12], [33]. Yang *et al.* [34] proposed a subword latent semantic analysis (LSA) based TextTiling approach that combined the concept matching merit of LSA and the robustness of subwords.

## III. Motivations of This Study

Despite tremendous efforts on lexical approaches, there still exist some problems that lead to inferior segmentation performance. First, the inter-sentence lexical cohesion is measured extrinsically, in most cases, by normalized inner product of word frequency vector representations of sentences. The intrinsic cohesive lexical behaviors, which assemble the central topic of a story, still remain concealed. Hearst *et al.* [9] pointed out that the vector space method is better at distinguishing similarities rather than differences. These approaches are reliable when the documents have sharp variations in lexical distribution, such as synthetic collections by concatenation of random texts [4]. However, in real-world broadcast news and spoken lectures, topic transitions are much smoother and the distributional variations are very subtle. As a result, sentence-similarity-based segmentation methods are sensitive to noises and show poor robustness to real-world documents. In fact, in broadcast news, sentence similarities in a story may be low and sentences from different stories may present high similarities. Therefore, we desire a method that explicitly reflects the intrinsic geometric structure of lexical cohesive relations and clearly discriminates different stories.

Second, besides lexical similarity, another obvious factor affecting the likelihood of two sentences is the distance between them in a text stream. Two sentences near each other are usually contained in one story unless a story boundary happens to locate between them. On the other hand, if the distance between two sentences is far longer than the regular length of a story, they improbably belong to a same story even if their lexical similarity is large enough. Previous methods typically consider this fact by only utilizing the similarity between adjacent parts of text or by directly limiting the segment length to a certain range. Some other approaches introduce empirical parameters in, such as the mean and standard deviation of story length of the particular corpus [32]. However, we notice that few approaches explicitly incorporate distance into the measure of inter-sentence connective strength.

In this paper, we show how to properly solve the above problems in Laplacian Eigenmaps (LE)-based approaches and significantly push forward the performance of lexical-cohesion-based story segmentation. Specifically, we measure the connective strength between a pair of sentences with both their lexical similarity and the distance between them in a text stream. We address the first problem by Laplacian Eigenmaps [15], a recently proposed, geometrically motivated method for data representation and dimensionality reduction. The LE method utilizes maps provided by the eigenvectors of the graph Laplacian matrix and eigenfunctions of the Laplace Beltrami operator on the manifold [16]. In contrast to the traditional dimensionality reduction methods such as principal components analysis (PCA) and multidimensional scaling (MS), the LE solution can reflect the intrinsic geometrical structure of the manifold which is not explicitly considered by the traditional methods [35]. Additionally, the locality preserving character makes the LE method relatively insensitive to outliers and noises and naturally connects the LE method to clustering. Specifically, we construct a Euclidean space in which each sentence is mapped to a vector. The original connective strength between sentences is thus reflected by the Euclidean distances between the corresponding vectors. As a result, the connective strength between sentences becomes geometrically evident for discriminating different stories. We also point out that the introduced sentence distance measure is also an essential condition for applying LE in story segmentation of real-world broadcast news programs. Taking advantage of the LE method, we present three story segmentation approaches, i.e., integrating the LE resultant with TextTiling (namely LE-TextTiling), carrying out a $k$-means clustering on LE resultant (i.e., spectral clustering [36]) and a new method that minimizes an intuitive LE-based optimal criterion using a fast dynamic programming solution (namely LE-DP). Extensive experiments show that the proposed approaches impressively outperform several state-of-the-art lexical methods and the LE-DP approach achieves the best performance.

## IV. Connections Between Sentences

### A. Sentence Construction

A text document is composed of sentences (separated by delimiters, e.g., periods). The connections between sentences can be depicted by lexical similarities. However, sentence delimiters are not readily available in speech recognition outputs such as broadcast news LVCSR transcripts. In fact, sentence boundary detection in speech recognition transcripts is another challenging task [37]. A natural thought is to use significant pauses as delimiters, resulting in pause-separated *pseudo-sentences*. However, Hearst *et al.* [9] and Malioutov *et al.* [8] have pointed out that using actual sentences and pause-separated utterances can skew the cosine similarity scores. This is because the number of shared terms between two long sentences and between a long and a short sentence would probably yield incomparable scores. For text segmentation, blocks of actual sentences are usually used to ameliorate this problem [9]. Therefore, a preferable selection of *pseudo-sentence* in speech recognition transcripts is equal-sized units with a fixed number of consecutive terms (e.g., words) [8]. Lexical similarity is measured between consecutive pseudo-sentences and each start point of a pseudo-sentence is a candidate for story boundary detection. Fig. 1 shows an example of sentence construction. Term overlap between sentences is allowed in order to obtain adequate boundary candidates without severely restricting the length of sentence. The result of sentence construction is determined by two parameters: the length of sentence—$Len$ and the number of terms between start points of two adjacent sentences—$Step$. An example of sentence construction from 15 terms with $Len = 7$ and $Step = 4$ is shown in Fig. 1.

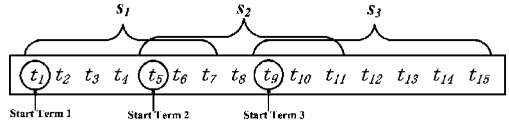


Fig. 1. Example of sentence construction with $Len = 7$ and $Step = 4$.

### B. Sentence Connective Strength

The lexical similarity between a pair of sentences $\mathbf{s}_i$, $\mathbf{s}_j$ is usually computed using the cosine similarity measure

$$\cos(\mathbf{s}_i, \mathbf{s}_j) = \frac{\sum_{k=1}^{M} s_{i,k} s_{j,k}}{\sqrt{\sum_{k=1}^{M} s_{i,k}^2 \times \sum_{k=1}^{M} s_{j,k}^2}} \tag{1}$$

where $\mathbf{s}_i$, $\mathbf{s}_j$ are term frequency vectors of the $i$th and the $j$th sentences, respectively. $s_{i,k}$ is the $k$th element of $\mathbf{s}_i$, representing the frequency of term $t_k$ appeared in the sentence $\mathbf{s}_i$. $M$ is the size of the whole vocabulary.

Considering that two sentences shall not belong to the same story if the distance between them is much longer than the ordinary length of one story, we incorporate the distance between sentences into the measure of sentence connective strength. The sentence connective strength measure is finally defined as

$$\mathrm{Co}(\mathbf{s}_i, \mathbf{s}_j) = \cos(\mathbf{s}_i, \mathbf{s}_j) \cdot \alpha^{|i-j|}. \tag{2}$$

The first part of (2) is the cosine similarity measure in (1) and the second part can be treated as a penalty factor of the distance $|i-j|$, where $\alpha$ is a constant parameter slight lower than 1.0. If the distance $|i-j|$ between sentences $\mathbf{s}_i$ and $\mathbf{s}_j$ is much larger than the ordinary length of a story, $\mathrm{Co}(\mathbf{s}_i, \mathbf{s}_j)$ will dramatically decrease by multiplying $\alpha^{|i-j|}$.

### C. Connection Matrix

When the connective strength measure is applied to all sentence pairs in a document, a connection matrix $\boldsymbol{C}$ is naturally constructed as

$$\boldsymbol{C} = \begin{bmatrix} \mathrm{Co}(\mathbf{s}_1, \mathbf{s}_1) & \mathrm{Co}(\mathbf{s}_1, \mathbf{s}_2) & \cdots & \mathrm{Co}(\mathbf{s}_1, \mathbf{s}_n) \\ \mathrm{Co}(\mathbf{s}_2, \mathbf{s}_1) & \mathrm{Co}(\mathbf{s}_2, \mathbf{s}_2) & \cdots & \mathrm{Co}(\mathbf{s}_2, \mathbf{s}_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Co}(\mathbf{s}_n, \mathbf{s}_1) & \mathrm{Co}(\mathbf{s}_n, \mathbf{s}_2) & \cdots & \mathrm{Co}(\mathbf{s}_n, \mathbf{s}_n) \end{bmatrix} \tag{3}$$

where $n$ is the number of sentences. It is easy to prove that $\boldsymbol{C}$ is symmetric and non-negative.

Fig. 2 shows two *dotplots* illustrating connection matrices with different sentence distance penalty factors ($\alpha = 1.00$ and $\alpha = 0.95$) for a half-an-hour-long broadcast news LVCSR transcript from China Central Television (CCTV). Note that high connection values are represented by dark pixels. We can see that each dotplot figure contains dark square regions along

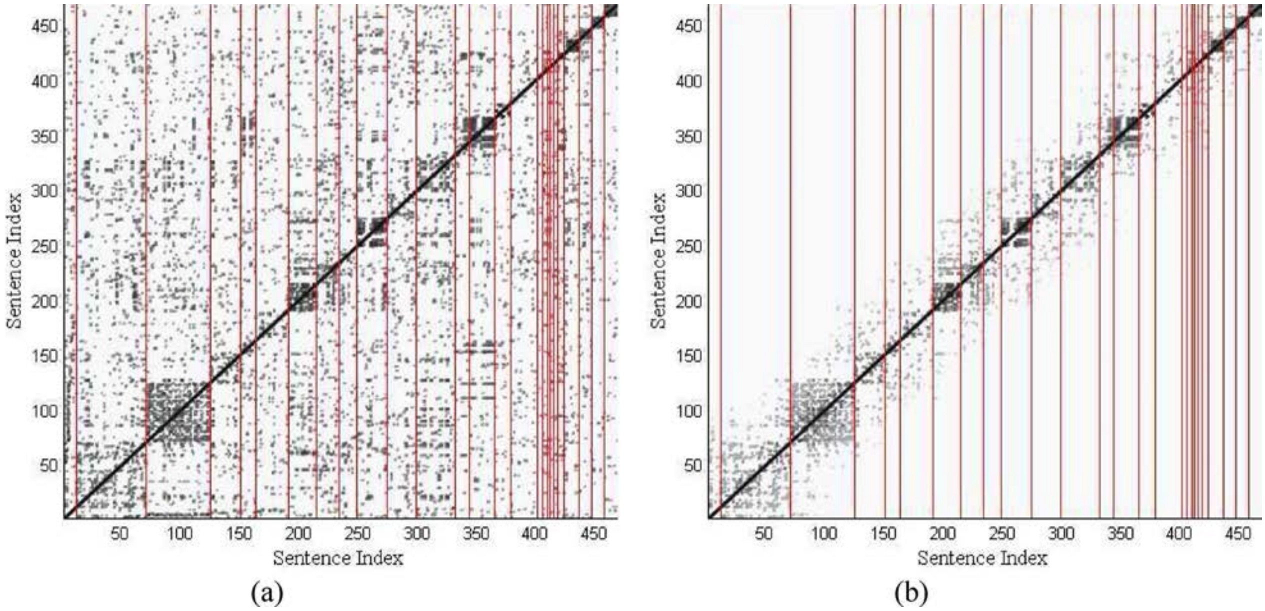


Fig. 2. Dotplots illustrating sentence connection matrices with sentence distance penalty $\alpha = 1.00$ (a) and $\alpha = 0.95$ (b) for a broadcast news transcript from CCTV. Vertical red lines denote real story boundary positions. (a) $\alpha = 1.00$. (b) $\alpha = 0.95$.

the diagonal with potential block structures delimited by story boundaries (red vertical lines). These regions indicate cohesive story segments with high sentence connective strength. On the other hand, sentences in different blocks tend to exhibit low similarities. However, we also observe that the dark blocks are not salient enough for direct story boundary identification. There are clear holes within each block and noisy dark dots scattered between different blocks. The former phenomenon is due to the fact that vocabulary use may vary throughout a story (e.g., via synonyms, generalizations, and specifications). The latter phenomenon is probably because of the high sentence similarities between different stories raised by common words and similar vocabulary usages. We can also find that noisy dark dots can be significantly reduced by setting an appropriate sentence distance penalty factor. In the next section, we employ a data representation method called Laplacian Eigenmaps to make the story blocks more salient and thus facilitate subsequent story boundary detection. Our goal is to unveil the intrinsic geometric story structures for boundary identification.

## V. Laplacian Eigenmaps: Unveil Story Structures

From the connection matrix, we can only see pairwise connections between sentences. More sophisticated relations hiding in sentence groups can hardly be seen directly. Hence, our objective is to convert sentence connection matrix $\boldsymbol{C}$ to a new matrix showing salient story boundaries. To this end, we use Laplacian Eigenmaps to map sentence term frequency vector $\mathbf{s}_i$ to a lower dimensional vector $\mathbf{y}_i$ so that the sentences belong to the same story stay as close together as possible and story boundaries can be clearly revealed.

LE [15] is a recently proposed data representation and dimensionality reduction procedure that can embed data into a Euclidean space in which the natural clusters in the data are implicitly emphasized. Specifically, the locality preserving character makes the LE method relatively robust to noises in data [15]. We believe this advantage benefits to story segmentation of speech recognition transcripts that have inevitable recognition errors.

### A. Laplacian Matrix

Given a set of data points (e.g., sentence term frequency vectors) $\mathbf{s}_1, \ldots, \mathbf{s}_n$, the pairwise data (e.g., sentence) connections $c_{ij}$ (e.g., $Co(\mathbf{s}_i, \mathbf{s}_j)$) and the connection matrix $\boldsymbol{C} := (c_{ij})_{(i,j=1,\ldots,n)}$, we define $\boldsymbol{D}$ to be the diagonal matrix with

$$d_i := \sum_{j=1}^{n} c_{ij}. \tag{4}$$

The unnormalized graph Laplacian matrix is thus defined as

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{C}. \tag{5}$$

Given a vector $\boldsymbol{v} = (v_1, \ldots, v_n)^T \in \mathbb{R}^n$, the following key identity can be easily verified:

$$\boldsymbol{v}^T \boldsymbol{L} \boldsymbol{v} = \frac{1}{2} \sum_{i,j=1}^{n} c_{ij} (v_i - v_j)^2. \tag{6}$$

It is easy to see that 1) $\boldsymbol{L}$ is symmetric and positive semi-definite; 2) the constant one vector $\mathbf{1} = (1, \ldots, 1)^T$ is an eigenvector of $\boldsymbol{L}$ with the eigenvalue 0; and 3) 0 is the smallest eigenvalue of $\boldsymbol{L}$.

### B. Optimal Mapping

Considering the problem of mapping the term frequency vectors $\mathbf{s}_i$ to lower dimensional vectors $\mathbf{y}_i$ so that the sentences in the same story stay as close together as possible, let

$$f : \mathbf{s}_i \mapsto \mathbf{y}_i \tag{7}$$

be such a mapping and $\mathcal{V}$ be the target space. A reasonable criterion for choosing an optimal mapping is to minimize the objective function

$$\sum_{i,j} \|\mathbf{y}_i - \mathbf{y}_j\|^2 c_{ij} \tag{8}$$

under appropriate constraints. If the connection $c_{ij}$ between sentences $\mathbf{s}_i, \mathbf{s}_j$ is strong, the distance between $\mathbf{y}_i, \mathbf{y}_j$ in the target space needs to be small enough to minimize the function above. $c_{ij}$ in the connection matrix $\boldsymbol{C}$ can be interpreted as a penalty factor for separating $\mathbf{y}_i$ and $\mathbf{y}_j$ apart in $\mathcal{V}$. A heavy penalty can be incurred if two sentences connected closely are mapped far away by $f$.

Assume the result of the mapping is an $n \times k$ matrix $\boldsymbol{Y}$, where the $i$th row of $\boldsymbol{Y}$ is the vector $\mathbf{y}_i$ that $\mathbf{s}_i$ is mapped to. According to (6), the objective function (8) can be rewritten as

$$\sum_{i,j} \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2 c_{ij} = \mathrm{tr}(\boldsymbol{Y}^T \boldsymbol{L} \boldsymbol{Y}). \tag{9}$$

To prevent $\boldsymbol{Y}$ from degenerating to a zero matrix or other matrices with its rank less than $k$, the following constraint is attached:

$$\boldsymbol{Y}^T \boldsymbol{D} \boldsymbol{Y} = \boldsymbol{I} \tag{10}$$

where $\boldsymbol{I}$ is the identity matrix.

Altogether, the problem of finding the optimal mapping can be written as follows:

$$\begin{aligned} &\underset{\boldsymbol{Y}}{\mathrm{argmin}} \quad \mathrm{tr}(\boldsymbol{Y}^T \boldsymbol{L} \boldsymbol{Y}) \\ &\text{subject to} \quad \boldsymbol{Y}^T \boldsymbol{D} \boldsymbol{Y} = \boldsymbol{I}. \end{aligned} \tag{11}$$

By the Rayleigh–Ritz theorem [38], we can see that the solution is provided by the eigenvectors corresponding to the smallest $k$ eigenvalues of the generalized eigenvalue problem

$$\boldsymbol{L}\boldsymbol{v} = \lambda \boldsymbol{D}\boldsymbol{v} \quad \text{or} \quad \boldsymbol{D}^{-1}\boldsymbol{L} = \lambda \boldsymbol{v}. \tag{12}$$

The same as in (6), 0 is the smallest eigenvalues of $\boldsymbol{D}^{-1}\boldsymbol{L}$ with the eigenvectors $\mathbf{1}$. The $n \times k$ matrix $\boldsymbol{Y}$, which is supposed to contain $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ as its rows, can be formed with the first $k$ eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ as its columns.

### C. Relations to Story Segmentation

Now we relate the LE method to the story segmentation task. First, we state that the sentences $\mathbf{s}_i$ and $\mathbf{s}_j$ are *directly connected* if and only if $c_{ij}$ is greater than zero. The relation of *indirectly connected* is recursively defined as follows:

- each pair of directly connected sentences are *indirectly connected* as well;
- if there exists a sentence $\mathbf{s}_h$ which is indirectly connected to both $\mathbf{s}_i$ and $\mathbf{s}_j$, then $\mathbf{s}_i$ and $\mathbf{s}_j$ are indirectly connected.

*1) Ideal Case:* An *ideal case* should meet the two conditions follows:

- any pair of sentences belonging to the same story are connected directly or indirectly;
- any pair of sentences belonging to different stories are unconnected.

In the ideal case, both the connection matrix $\boldsymbol{C}$ and the corresponding graph Laplacian matrix $\boldsymbol{L}$ have a block diagonal form as follows:

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_1 & & & \\ & \boldsymbol{C}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{C}_{N_s} \end{bmatrix}, \quad \boldsymbol{L} = \begin{bmatrix} \boldsymbol{L}_1 & & & \\ & \boldsymbol{L}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{L}_{N_s} \end{bmatrix}$$

where $N_s$ is the number of stories.

Note that each block $\boldsymbol{L}_i$ is a proper Laplacian matrix on its own, i.e., the Laplacian matrix corresponding to the $i$th story. Obviously, $\boldsymbol{L}_i$ has an eigenvalue 0 of multiplicity 1, and the corresponding eigenvector is the constant one vector on the $i$th story. As it is the same for all the block diagonal matrices, the eigenvalues of $\boldsymbol{L}$ are given by the union of the eigenvalues of $\boldsymbol{L}_i$ and the corresponding vectors of $\boldsymbol{L}$ are the eigenvectors of $\boldsymbol{L}_i$, which are filled with 0 at the positions of the other blocks. Thus, the matrix $\boldsymbol{L}$ has as many eigenvalues 0 as $N_s$, and an orthogonal basis for the eigenspace of eigenvalue 0 is $\{\mathbf{1}_{\boldsymbol{C}_1}, \mathbf{1}_{\boldsymbol{C}_2}, \ldots, \mathbf{1}_{\boldsymbol{C}_{N_s}}\}$, where $\mathbf{1}_{\boldsymbol{C}_i}$ is an 0–1 vector $(x_1, x_2, \ldots, x_n)$ with entries $x_t = 1$ if and only if entry $c_{tt}$ is contained in the $i$th block $\boldsymbol{C}_i$. Other alternative bases are linear combinations of this particular one; hence, they are also piecewise constant on the blocks.

Let $\boldsymbol{Y}$ be the matrix containing the vectors $\{\mathbf{1}_{\boldsymbol{C}_1}, \mathbf{1}_{\boldsymbol{C}_2}, \ldots, \mathbf{1}_{\boldsymbol{C}_{N_s}}\}$ as columns and $\mathbf{y}_i$ be the vector corresponding to the $i$th row of $\boldsymbol{Y}$. $\mathbf{y}_i$ is a unit vector in which the $N$th element is 1 and the others are 0 as $\mathbf{y}_i$ belongs to the $N$th story. We can see that in the ideal case, all sentences of a certain story are mapped to the identical vector. The relation between the sentences and stories is well revealed after mapping.

*2) Real Case:* As illustrated in Fig. 2, in real-world documents, we do not have a completely ideal situation where the connective strength between sentences in different stories is exactly 0, but if the connective strength between these sentences tend to be small, the Laplacian matrix can be treated as a perturbed version of the one in the ideal case. Formally,

$$\tilde{\boldsymbol{L}} := \boldsymbol{L} + \boldsymbol{H} \tag{13}$$

where $\boldsymbol{H}$ is the perturbation added to $\boldsymbol{L}$.

Let $\mathcal{U}$ be the eigenspace corresponding to the smallest $k$ eigenvalues of $\boldsymbol{L}$ and $\tilde{\mathcal{U}}$ be the analogous space for $\tilde{\boldsymbol{L}}$. By the Davis–Kahan theorem [39], we can see that if there exists an interval $[b, e]$ that contains the exactly $k$ smallest eigenvalues of $\boldsymbol{L}$ and $\tilde{\boldsymbol{L}}$, the distance between $\mathcal{U}$ and $\tilde{\mathcal{U}}$ is bounded by

$$dis(\mathcal{U}, \tilde{\mathcal{U}}) = \|\sin\Theta(\mathcal{U}, \tilde{\mathcal{U}})\| \leq \frac{\|\boldsymbol{H}\|}{\delta} \tag{14}$$

where $\delta$ is the difference between $e$ and the $(k+1)$st smallest eigenvalue of $\boldsymbol{L}$, and $\|\boldsymbol{H}\|$ is the Frobenius norm or the two-norm of $\boldsymbol{H}$. Notice that the distance between the two eigenspaces depends on the norm of perturbation $\|\boldsymbol{H}\|$ and the difference $\delta$. We denote the eigenvalues of $\boldsymbol{L}$ by $\lambda_1, \ldots, \lambda_n$ and the ones of $\tilde{\boldsymbol{L}}$ by $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n$. If we set the interval as $[0, \lambda_k]$, $\delta$ is

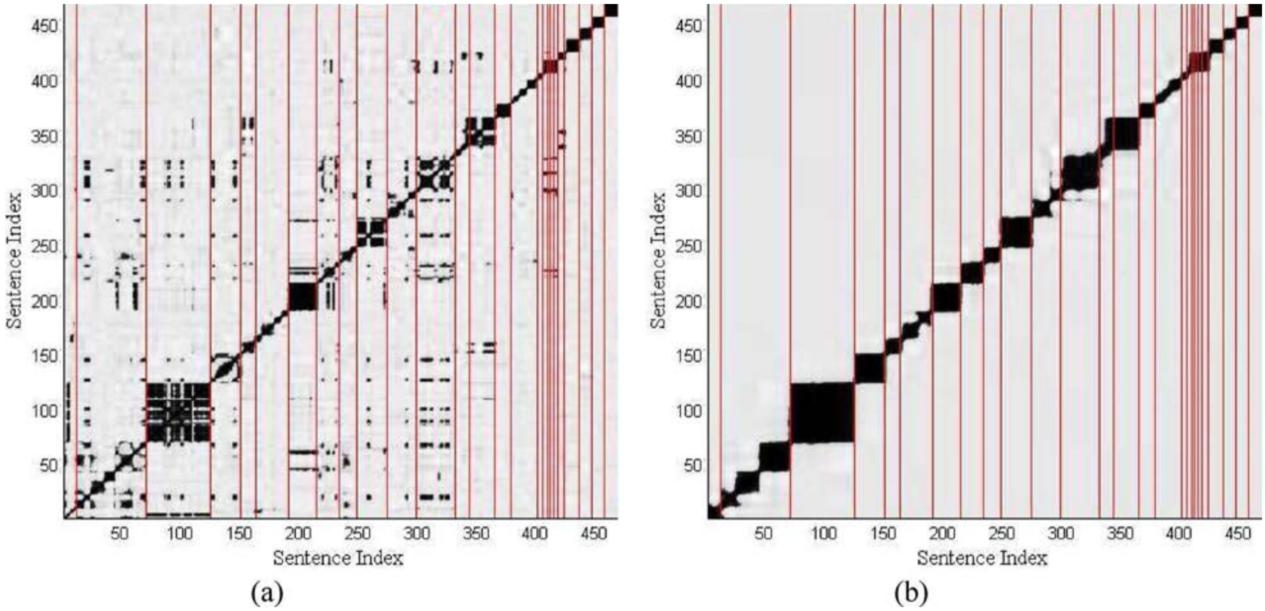


Fig. 3. Dotplots illustrating cosine similarities between sentences after LE mapping with sentence distance penalty $\alpha = 1.00$ (a) and $\alpha = 0.95$ (b) for the same broadcast news transcript in Fig. 2. Vertical red lines denote real story boundary positions. (a)

$|\lambda_{k+1} - \lambda_k|$. We can see from the Davis–Kahan theorem that the larger $\delta$ is, the smaller the distance between the ideal case and perturbed case is. On the other hand, if the perturbation $||\boldsymbol{H}||$ is too large, e.g., the noisy case shown in Fig. 2(a), the interval $[0, \lambda_k]$ may not be able to include both the first $k$ eigenvalues of $\boldsymbol{L}$ and $\tilde{\boldsymbol{L}}$. In this case, the effectiveness of Laplacian Eigenmaps becomes much weaker. Hence, this is another important reason why we incorporate the sentence distance penalty $\alpha$ into the sentence connective strength measure. Dotplots in Fig. 3 show the cosine similarities between sentences after LE mapping corresponding to the two dotplots in Fig. 2, respectively. We can see that it is much easier to differentiate the intra-story area and the inter-story area after using the sentence distance penalty $\alpha$ and the story boundaries are clearly revealed.

The analysis above shows that: 1) as long as the Frobenius norm of the perturbation $\boldsymbol{H}$ is not too large, the eigenspace $\tilde{\mathcal{U}}$ will not drift far away from $\mathcal{U}$. Although the sentences in a certain story will not be precisely mapped to an identical vector, we can expect the difference between these vectors to be relatively small; and 2) according to the restrictions in the Davis–Kahan theorem for Laplacian Eigenmaps, it is necessary to use an appropriate $\alpha$ to bring the real case closer to the ideal case so that the LE-based approaches can work in their right ways.

## VI. Story Segmentation Approaches

We adopt three story segmentation approaches, namely:

- LE-TextTiling: integrating the resultant of LE mapping, i.e., $\mathbf{y}_i$, into the classical TextTiling method [7], [9];
- Spectral Clustering [36]: carrying out a $k$-means clustering on $\boldsymbol{Y}$;
- LE-DP: a novel method that minimizes an optimal segmentation criterion of $\mathbf{y}_i$ using a fast dynamic programming solution.

### A. LE-TextTiling

In the classical TextTiling method [9], cosine similarity $\cos(\mathbf{s}_i, \mathbf{s}_{i+1})$ is calculated between each consecutive sentence pairs $(\mathbf{s}_i, \mathbf{s}_{i+1})$ across the text or the speech recognition transcript according to (1). Then $k$ inter-sentence positions with lowest cosine similarity values are considered as story boundaries. To take the advantage of LE, we introduce the resultant of mapping, $\mathbf{y}_i$, into the TextTiling algorithm. Specifically, we simply substitute $(\mathbf{s}_i, \mathbf{s}_{i+1})$ with $(\mathbf{y}_i, \mathbf{y}_{i+1})$ in the cosine similarity measure

$$\cos(\mathbf{y}_i, \mathbf{y}_{i+1}) = \frac{\sum_{j=1}^{k} y_{i,j} y_{i+1,j}}{\sqrt{\sum_{j=1}^{k} y_{i,j}^2 \times \sum_{j=1}^{k} y_{i+1,j}^2}} \tag{15}$$

and $k$ inter-sentence positions with lowest cosine similarity values are reported as story boundaries. We call this hybrid method LE-TextTiling.

### B. Spectral Clustering

TextTiling detects story boundaries with only local similarity measure, i.e., cosine of two consecutive sentences. In fact, we can treat story segmentation as a clustering problem which takes advantage of inter-sentence similarity information at the global level. Spectral clustering [36] is a typical application of LE in the field of modern clustering with superior performance over other clustering methods. From the graph cut point of view, spectral clustering can also be interpreted as an equivalent of the normalized cuts [40] algorithm. As introduced in [36], the algorithm usually employs a simple $k$-means algorithm on the resultant of LE mapping, $\boldsymbol{Y}$, and partition $n$ row vectors $\mathbf{y}_i$ into $k$ clusters. The basic $k$-means algorithm begins with an arbitrary set of cluster centers and usually lands into a local optimal solution. Therefore, we use an improved $k$-means algorithm called

$k$-means++ [41] to partition sentence vectors $\mathbf{y}_i$ into $k$ story clusters. This algorithm initials the centers with specific probabilities and consistently outperforms $k$-means in terms of accuracy and computation speed.

After clustering, each sentence $\mathbf{y}_i$ is labeled with a story cluster tag $(\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k)$ and the whole text is separated to $m$ segments each labeled with an identical cluster tag. In the ideal case, $m = k$ and boundaries between segments with different labels are reported as story boundaries. However, in fact, $m > k$ usually stands because sentences in the same story may be labeled with different cluster tags. Thus, we use the following heuristic steps to alternate the $m$ segments into $k$ segments.

1) If there are more than two segments having the same cluster tag in the text, keep the largest segment as the survivor and combine the other segments with their respective adjacent segments according to step 2. For example, suppose two segments are labeled as $\mathcal{C}_l$, i.e., $[\mathbf{y}_a, \ldots, \mathbf{y}_b] \in \mathcal{C}_l$ and $[\mathbf{y}_e, \ldots, \mathbf{y}_f] \in \mathcal{C}_l$. If $|a - b| > |e - f|$, then $[\mathbf{y}_a, \ldots, \mathbf{y}_b]$ is the survivor for $\mathcal{C}_l$.
2) For each segment to be combined, calculate the distances with its left and right adjacent segments and combine the segment with its closer adjacent segment. For example, suppose $[\mathbf{y}_e, \ldots, \mathbf{y}_f]$ is the segment to be combined and $[\mathbf{y}_c, \ldots, \mathbf{y}_d]$ and $[\mathbf{y}_g, \ldots, \mathbf{y}_h]$ are its left and right neighbors, respectively. If $\|\mathbf{y}_d - \mathbf{y}_e\|^2 < \|\mathbf{y}_f - \mathbf{y}_g\|^2$, then $[\mathbf{y}_e, \ldots, \mathbf{y}_f]$ is combined with $[\mathbf{y}_c, \ldots, \mathbf{y}_d]$, resulting in $[\mathbf{y}_c, \ldots, \mathbf{y}_d, \mathbf{y}_e, \ldots, \mathbf{y}_f]$.
3) Repeat step 1 and 2 until the number of segments equals to the number of clusters.

### C. LE-DP

As we pointed out in Section VI-B, spectral clustering has an apparent drawback for the story segmentation task: the segments after clustering have to be tuned by a heuristic patch. Therefore, we propose a dynamic programming (DP) solution that directly seeks the optimal story segmentation. Specifically, we formalize the process of segmentation as minimizing

$$\sum_{t=1}^{N_s} \left( \sum_{i,j \in Seg_t} \|\mathbf{y}_i - \mathbf{y}_j\|^2 \right) \tag{16}$$

where $\|\mathbf{y}_i - \mathbf{y}_j\|^2$ is the inter-sentence Euclidean distance in a story segment $Seg_t$, $N_s$ is the number of stories. Due to the linear constraint of the story segmentation task [8], we can obtain the global minimization of (16) using the following dynamic programming algorithm in polynomial time:

$$g(s,t) = \sum_{i=s}^{t} \sum_{j=s}^{t} \|\mathbf{y}_i - \mathbf{y}_j\|^2, \quad 1 \le s \le t \le n, \tag{17}$$

$$f(k,i) = \min_{k \le j \le i} \left\{ f(k-1, j-1) + g(j,i) \right\}, \quad 1 \le k \le N_s \tag{18}$$

$$b(k,i) = \operatorname*{argmin}_{k \le j \le i} \left\{ f(k-1, j-1) + g(j,i) \right\}, \quad 1 \le k \le N_s \tag{19}$$

$$\text{s.t. } f(1,i) = g(1,i), \; b(1,i) = 1, \quad 1 \le i \le n. \tag{20}$$

$g(s,t)$ is the score of a particular segment starting from the $s$th sentence to the $t$th sentence. $f(k,i)$ is the minimal cost of segmenting the first $i$ sentences into $k$ segments. $b(k,i)$ is used to recover the segment boundaries of the optimal segmentation. According to the principle of Occam's razor, if there are multiple solutions of (16), we choose the one with the fewest segments.

*1) Fast and Incremental Implementation:* The time complexity of natively calculating $g(s,t)$ is $O(|t-s|^2)$ depending on the implementation. As $1 \le s \le t \le N$, the average time complexity is $O(N^2)$, where $N$ is the number of sentences in the document. Therefore, calculating $g(s,t)$ for all possible pairs of $(s,t)$ separately has time complexity of $O(N^4)$. We now introduce an incremental method to reduce the time complexity to $O(N^2)$.

Equation (17) can be rewritten as

$$\begin{aligned} g(s,t) &= \sum_{i=s}^{t} \sum_{j=s}^{t} \|\mathbf{y}_i - \mathbf{y}_j\|^2 \\ &= \sum_{i=s}^{t-1} \sum_{j=s}^{t-1} \|\mathbf{y}_i - \mathbf{y}_j\|^2 + 2 \sum_{i=s}^{t} \|\mathbf{y}_t - \mathbf{y}_i\|^2 \\ &= g(s,t-1) + 2 \sum_{i=s}^{t} \|\mathbf{y}_t - \mathbf{y}_i\|^2. \end{aligned} \tag{21}$$

It is easy to obtain $\sum_{i=s}^{t} \|\mathbf{y}_t - \mathbf{y}_i\|^2$ in $O(N^2)$ time for all pairs of $(s,t)$ in advance. Therefore, using (21), we can obtain $g(s,t)$ by given $g(s, t-1)$ in $O(1)$. The complexity of calculating $g(s,t)$ for all pairs of $(s,t)$ is $O(N^2)$. The overall time complexity of the proposed DP solution (17)–(20) is $O(N^2)$.

## VII. Experiments and Analysis

### A. Corpus

To evaluate the proposed approaches, we carried out story segmentation experiments on three corpora: the TDT2 Mandarin broadcast news corpus, the TDT2 English broadcast news corpus (VOA data) and the CCTV Mandarin broadcast news corpus. In order to validate the generality of the proposed methods, experiments were performed on the three corpora with two different languages.

*1) TDT2 Mandarin BN Corpus:* The topic detection and tracking Phase 2 (TDT2) Mandarin corpus[1] contains about 53 hours Mandarin broadcast news audio from Voice of America (VOA). The 177 audio recordings contain 39 long programs and 138 short programs. The corpus provides manually annotated meta-data including story boundaries and speech recognition transcripts with word, character and base syllable error rates of 37%, 20% and 15%, respectively. We separated the corpus into two non-overlapping sets: a development set of 90 recordings with 1321 story boundaries for parameter tuning and a set of 87 recordings with 1262 story boundaries for performance testing.

*2) TDT2 VOA English BN Corpus:* The TDT2 VOA English corpus[2] includes 111 English broadcast news audio recordings

[1] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2001S93

[2] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99S84

from VOA (about 1 hour each audio). The word error rate of the associated speech recognition transcripts is 35%. We divided the news programs into a development set of 56 audio recordings with 2629 story boundaries and a test set of 55 audio recordings with 2627 story boundaries.

*3) CCTV Mandarin BN Corpus:* We collected another Mandarin broadcast news corpus from CCTV. This corpus contains 71 news episodes with 27 hours of CCTV news audio and 2101 story boundaries. Each news episode includes a detailed news session (about 25 mins) and a brief news session (about 5 min). The audio was transcribed by the Julius LVCSR toolkit[3] with word, character, and base syllable error rates of 25%, 18%, 15.5%, respectively. We also downloaded the manual word transcripts associated with each news episode from the CCTV website as the complement references. We divided the corpus into a development set of 40 audio recordings with 1209 story boundaries and a test set of 31 audio recordings with 892 story boundaries.

### B. Experimental Setup

We experimented with four story segmentation methods: classical TextTiling [9], LE-TextTiling, spectral clustering and LE-DP on the three corpora. We also compared the proposed methods with several state-of-the-art methods on the TDT2 Mandarin corpus. For performance comparison, we followed the standard setup used by Hearst [9] and Malioutov [8], i.e., the number of stories ($N_s$) in each news program was set to be known in advance. For a more practical use, an automatic story number determination method can be considered as the pre-step of the LE-based approaches. Previous works showed that lexical measure on subword units was robust to speech recognition errors and out-of-vocabulary words in spoken document retrieval [42] and Chinese broadcast news segmentation [12]. Thus, we evaluated the story segmentation methods on both word and subword $n$-gram transcripts [12] for the three corpora. For the Chinese syllable $n$-gram experiments, we obtained the syllable sequences from the word transcripts using an in-house Mandarin word-to-syllable lexicon. For English speech recognition transcripts, we removed the uninformative words using a stopword list. The phonetic $n$-gram sequences were achieved from the word transcripts using the CMU dictionary in the English phoneme $n$-gram experiments. Additionally, for the English word level experiment, we reduced words to their linguistic roots by an English Porter stemmer.

For each method under evaluation, we first conducted empirical parameter tuning on the development set to obtain optimal parameter setting that achieved the best performance of story segmentation. Experiments were then carried out on the test set using the best-tuned parameters. The parameters are sentence length ($Len$), sliding step ($Step$) and sentence distance penalty $\alpha$. For the experiments on the two Mandarin corpora and the word level experiment on the TDT2 English corpus, the tuning range of $Len$ was from 10 to 70 with a step of 5. For the phonetic $n$-gram experiments on the TDT2 English corpus, the tuning range of $Len$ was from 100 to 200 with a step of 20. The sliding step ($Step$) was calculated by multiplying ($Len$) with a rate ranging from 0.1 to 0.9 with a step of 0.1. The tuning range of $\alpha$ was from 0.88 to 0.98 with a step of 0.01.

The evaluation criterion for story segmentation performance is *F1-measure*, i.e., the harmonic mean of *precision* and *recall*. They are defined as

$$\text{precision} = \frac{N_{\text{cor}}}{N_{ret}} \tag{22}$$

$$\text{recall} = \frac{N_{\text{cor}}}{N_{\text{ref}}} \tag{23}$$

and

$$F1\text{-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{24}$$

where $N_{\text{cor}}$ is the number of correctly detected story boundaries, $N_{\text{ref}}$ is the number of actual story boundaries (manual annotation), and $N_{ret}$ is the number of boundaries returned by the story segmentation approach. We followed the TDT2 evaluation standard[4] for all the experiments: a detected story boundary is considered correct if it lies within a 15-s tolerant window (about 30 words) on each side of a manually annotated reference boundary.

### C. Effectiveness of Mapping

Before reporting the experimental results, we first show the effectiveness of LE mapping on the sentence connection matrix. As stated in Section V-C, in the ideal case, the term frequency vectors are mapped to unit vectors with dimension of $N_s$ by LE. Comparing to the connection matrix $\boldsymbol{C}$, we define a new similarity matrix $\boldsymbol{T}$ to measure the relations between the unit vectors. It is easy to know that the similarity between vectors of a same story is 1 as they are exactly the same, and the similarity between vectors belong to different stories is 0 because the different unit vectors are pairwise orthogonal. The matrix $\boldsymbol{T}$ is represented in a diagonal form as the same as $\boldsymbol{C}$:

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{T}_1 & & & \\ & \boldsymbol{T}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{T}_{N_s} \end{bmatrix}$$

where the block $\boldsymbol{T}_i$ ($i = 1, \ldots, N_s$) is filled with 1.

In the real case, it is impossible to obtain such matrices $\boldsymbol{C}$ and $\boldsymbol{T}$. Actually, we have their real, perturbed versions $\tilde{\boldsymbol{C}}$ and $\tilde{\boldsymbol{T}}$ instead. Thus, a reasonable effectiveness indication of LE is to compare the similarity between $\tilde{\boldsymbol{C}}$ and $\boldsymbol{T}$ and the similarity between $\tilde{\boldsymbol{T}}$ and $\boldsymbol{T}$. We expect a higher similarity between $\tilde{\boldsymbol{T}}$ and $\boldsymbol{T}$, as the contribution of Laplacian Eigenmaps to reveal boundaries in the story segmentation task.

We define the similarity between matrices using the Frobenius inner product which induces the well-known Frobenius norm. The Frobenius inner product

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle = trace(\boldsymbol{A}^T \boldsymbol{B}) = trace(\boldsymbol{B}^T \boldsymbol{A}) \tag{25}$$

is a natural quantity to associate with the two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.

[3] http://julius.sourceforge.jp

[4] www.itl.nist.gov/iad/mig//tests/tdt/1998/doc/tdt2.eval.plan.98.v3.7.pdf

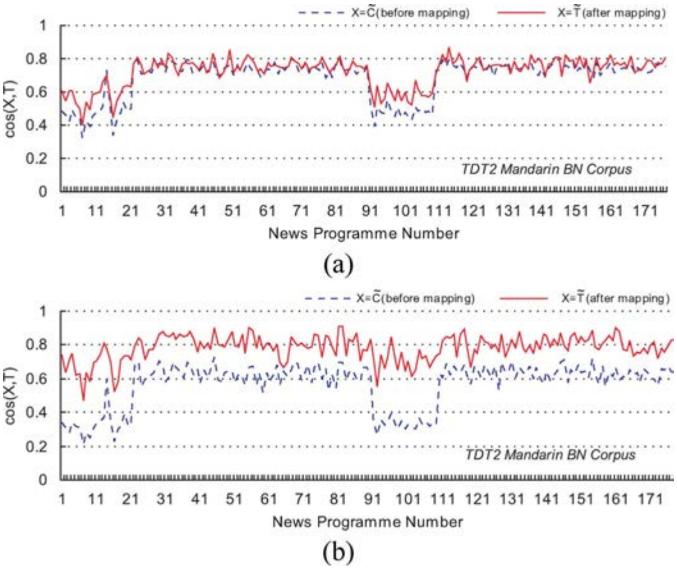


Fig. 4. Values of $\cos(\tilde{\boldsymbol{C}}, \boldsymbol{T})$ and $\cos(\tilde{\boldsymbol{T}}, \boldsymbol{T})$ for all programs in the TDT2 Mandarin BN corpus. Term frequency vectors $\mathbf{s}_i$ are calculated at (a) word level and (b) character-bigram level.

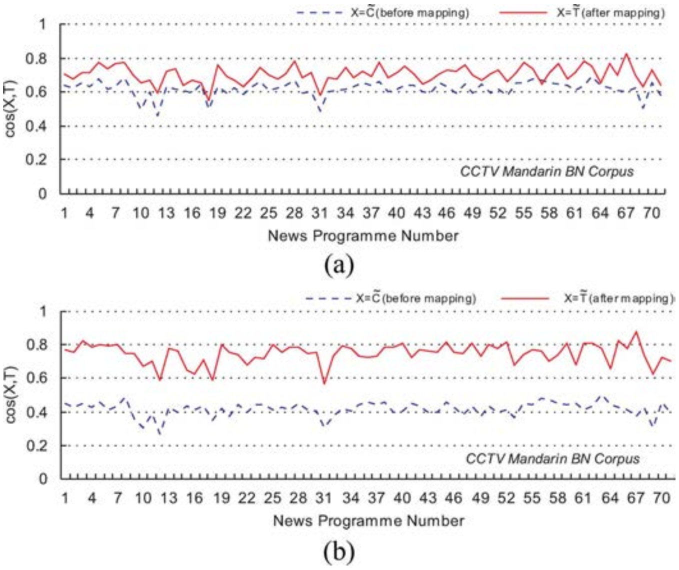


Fig. 5. Values of $\cos(\tilde{\boldsymbol{C}}, \boldsymbol{T})$ and $\cos(\tilde{\boldsymbol{T}}, \boldsymbol{T})$ for all programs in the CCTV Mandarin BN corpus. Term frequency vectors $\mathbf{s}_i$ are calculated at (a) word level and (b) character-bigram level.

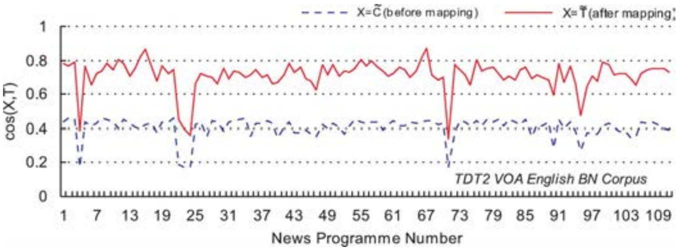


Fig. 6. Values of $\cos(\tilde{\boldsymbol{C}}, \boldsymbol{T})$ and $\cos(\tilde{\boldsymbol{T}}, \boldsymbol{T})$ for all programs in the TDT2 English BN corpus. Term frequency vectors $\mathbf{s}_i$ are calculated at word level.

We use cosine between $\boldsymbol{A}$ and $\boldsymbol{B}$ as the similarity measure, formulated as

$$\cos(\boldsymbol{A}, \boldsymbol{B}) = \frac{\langle \boldsymbol{A}, \boldsymbol{B} \rangle}{\langle \boldsymbol{A}, \boldsymbol{A} \rangle^{1/2} \langle \boldsymbol{B}, \boldsymbol{B} \rangle^{1/2}}. \tag{26}$$

In general, $0 \leq \cos(\boldsymbol{A}, \boldsymbol{B}) \leq 1$ and $\cos(\boldsymbol{A}, \boldsymbol{B}) = 1$ if and only if $\boldsymbol{A} = \boldsymbol{B}$.

The $\cos(\boldsymbol{X}, \boldsymbol{T})$ $(\boldsymbol{X} = \{\tilde{\boldsymbol{C}}, \tilde{\boldsymbol{T}}\})$ for the TDT2 Mandarin corpus, the CCTV Mandarin corpus and the TDT2 English corpus are plotted in Figs. 4–6, respectively. As expected, $\cos(\tilde{\boldsymbol{T}}, \boldsymbol{T})$ is greater than $\cos(\tilde{\boldsymbol{C}}, \boldsymbol{T})$ in most cases for the three different corpora, which clearly show the effectiveness of LE mapping. In each figure, the space between the two curves with different colors (red and blue) indicates the effectiveness of the mapping: the larger the space, the more effective the mapping. We also notice that the effectiveness of LE mapping is more salient at the character-bigram level as compared with the word level in Mandarin broadcast news. This observation is in accordance with the experimental results of story segmentation described in following sections.

### *D. Experimental Results*

The experimental results of different story segmentation methods are compared in Tables I–III for the three corpora, respectively. In general, the proposed LE-DP method and spectral clustering achieve superior performances as compared with other methods. Especially, LE-DP obtains the best story segmentation performance at most word and subword levels. The highest F1-measures are 0.7260, 0.7460, and 0.6057 for the CCTV Mandarin corpus, the TDT2 Mandarin corpus and the TDT2 English corpus, respectively. These impressive results are all achieved by the proposed LE-DP method.

When we compare conventional TextTiling with LE-TextTiling, we observe that LE-TextTiling brings significant performance gains at all word/subword levels. For instance, F1-measure is increased from 0.4965 (TextTiling) to 0.5660 (LE-TextTiling) for the TDT2 English corpus. Even impressively, F1-measure is relatively increased by 11.5% (from 0.6282 to 0.7002) for the character-bigram level on the TDT2 Mandarin corpus; a significant F1-measure gain of 17.8% (from 0.5526 to 0.6509) is obtained for the character-unigram level on the CCTV Mandarin corpus. Fig. 7 shows the sentence cosine similarity curves for TextTiling (above) and LE-TextTiling (below) for a broadcast news program in the CCTV Mandarin corpus. We can clearly see that Laplacian Eigenmaps can effectively reinforce the story boundary positions. The original sentence cosine similarity curve $(\cos(\mathbf{s}_i, \mathbf{s}_{i+1}))$ fluctuates greatly over time and the story boundaries (vertical red lines) are not salient. In contrast, the sentence cosine similarity calculated on the LE vectors $(\cos(\mathbf{y}_i, \mathbf{y}_{i+1}))$ shows consistently high values within each story and clear valleys at inter-story positions, leading to more correctly detected story boundaries (red solid lines).

For performance comparison, we re-implemented several recent story segmentation approaches, including the graph cut approach [8] (Malioutov'06), modeling lexical chains' statistical behavior [11] (Chan'07), subword-LSA-based TextTiling [34] (Yang'08) and modeling broadcast news prosody using conditional random fields [22] (Wang'10). The best performances of these approaches on the TDT2 Mandarin corpus are shown in

TABLE I
STORY SEGMENTATION RESULTS (F1-MEASURE) OF PROPOSED METHODS ON CCTV MANDARIN BN CORPUS

| Approach | Word | | Unigram | | Bigram | | Trigram | | Quadgram | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Char. | Syl. | Char. | Syl. | Char. | Syl. | Char. | Syl. | Char. | Syl. |
| TextTiling | 0.5410 | 0.5237 | 0.5526 | 0.5526 | 0.5272 | 0.5283 | 0.5341 | 0.5376 | 0.5006 | 0.4948 |
| LE-TextTiling | 0.5757 | 0.5792 | 0.6509 | 0.6035 | 0.6324 | 0.6301 | 0.5572 | 0.5642 | 0.5399 | 0.5353 |
| Spectral Clustering | 0.5919 | 0.6058 | 0.6728 | 0.6335 | 0.6925 | 0.6821 | 0.6046 | 0.5942 | 0.5676 | 0.5792 |
| LE-DP | 0.6231 | 0.6324 | 0.6936 | 0.6613 | **0.7191** | **0.7260** | 0.6312 | 0.6393 | 0.6092 | 0.5884 |

TABLE II
STORY SEGMENTATION RESULTS (F1-MEASURE) OF PROPOSED METHODS ON TDT2 MANDARIN BN CORPUS

| Approach | Word | | Unigram | | Bigram | | Trigram | | Quadgram | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Char. | Syl. | Char. | Syl. | Char. | Syl. | Char. | Syl. | Char. | Syl. |
| TextTiling | 0.5483 | 0.5467 | 0.5902 | 0.5973 | 0.6282 | 0.6155 | 0.6052 | 0.6036 | 0.5752 | 0.5767 |
| LE-TextTiling | 0.6179 | 0.6195 | 0.6677 | 0.6511 | 0.7002 | 0.6954 | 0.6472 | 0.6495 | 0.6076 | 0.6266 |
| Spectral Clustering | 0.6432 | 0.6369 | 0.7041 | 0.7009 | 0.7358 | 0.7334 | 0.6741 | 0.6788 | 0.6392 | 0.6377 |
| LE-DP | 0.6329 | 0.6377 | 0.7215 | 0.7089 | **0.7460** | **0.7453** | 0.6930 | 0.6986 | 0.6693 | 0.6772 |

TABLE III
STORY SEGMENTATION RESULTS (F1-MEASURE) OF PROPOSED METHODS ON TDT2 ENGLISH BN CORPUS

| Approach | Word | Phoneme $n$-gram | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1-gram | 2-gram | 3-gram | 4-gram | 5-gram | 6-gram |
| TextTiling | 0.4965 | 0.4044 | 0.4575 | 0.4813 | 0.4986 | 0.4723 | 0.4813 |
| LE-TextTiling | 0.5660 | 0.4110 | 0.4649 | 0.5356 | 0.5672 | 0.5577 | 0.5360 |
| Spectral Clustering | 0.5988 | 0.3933 | 0.4776 | 0.5532 | 0.5775 | 0.5823 | 0.5486 |
| LE-DP | **0.6057** | 0.4053 | 0.4681 | 0.5528 | 0.5836 | 0.5845 | 0.5639 |

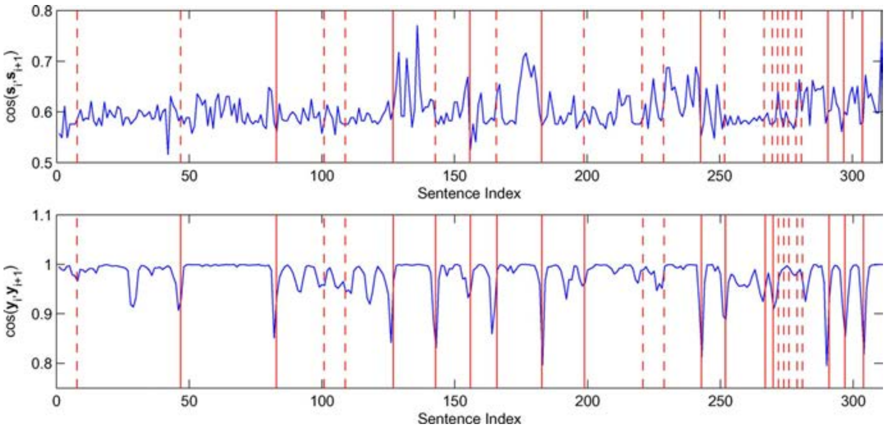


Fig. 7. Sentence cosine similarity curves for TextTiling (above) and LE-TextTiling (below) for a broadcast news program in CCTV Mandarin BN corpus. Red solid lines denote matches between red story boundaries and detected story boundaries and red dotted lines denote missed story boundaries.

Fig. 8. We can see that the three LE-based methods significantly outperform the others. We notice that after applying the Laplacian Eigenmaps technique, the classical TextTiling even surpasses those recent approaches. Impressively, the LE-DP approach achieves a relative 10% improvement over the recent CRF-Prosody approach [22].

### *E. Cross Language/Corpus Comparison*

As just discussed, the proposed LE-based methods show their effectiveness on all the three corpora and the LE-DP method consistently achieves the best performance. It indicates that our methods are applicable to story segmentation of news programs from different languages and media sources.

Comparing results for word and different subword levels on the two Mandarin corpora (Tables I and II), we observe that character/syllable unigrams and bigrams outperform trigrams, quadgrams and word in general. The superior performance of unigrams and bigrams is because: 1) the recognition error rates at character/syllable levels are much lower than the word level; 2) Chinese character/syllable units have the advantage of partial matching and this can partially recover the relations among mis-

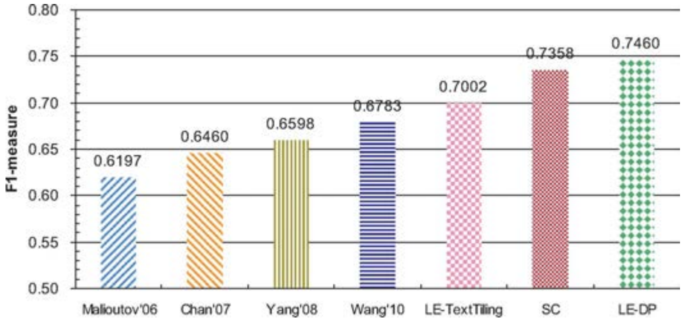


Fig. 8. Performance comparison between several recent story segmentation approaches on the TDT2 Mandarin BN corpus. SC denotes spectral clustering.

TABLE IV
DISTRIBUTION OF $n$-PHONEME WORDS IN THE ENGLISH CORPUS. STOPWORDS HAVE BEEN REMOVED

| Corpus | Number of phonemes in a word | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 11+ |
| TDT2 | 17413<br>(3.23%) | 23597<br>(4.38%) | 96970<br>**(18.00%)** | 100270<br>**(18.61%)** | 90020<br>**(16.71%)** | 72763<br>(13.51%) | 53455<br>(9.92%) | 36772<br>(6.83%) | 26868<br>(4.99%) | 9770<br>(1.81%) | 6353<br>(1.18%) | 4515<br>(0.84%) |

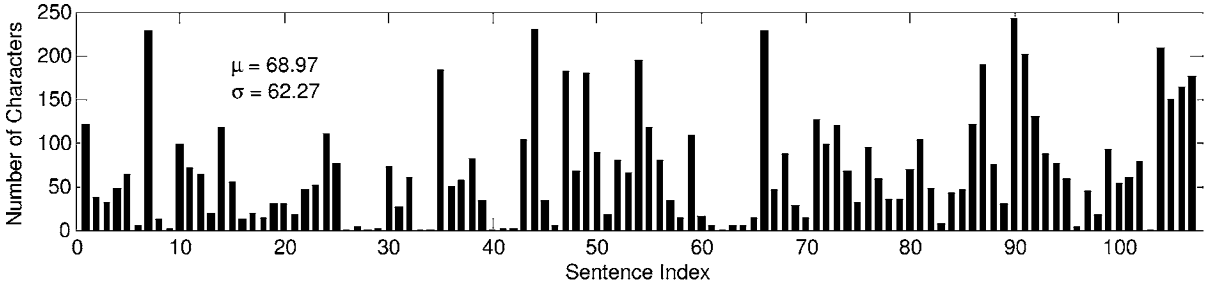


Fig. 9. Number of characters in the pause-separated pseudo-sentences for a broadcast news program in the CCTV Mandarin BN corpus. $\mu$ denotes mean and $\sigma$ denotes standard deviation. Pause threshold $\theta = 0.7$.

TABLE V
DISTRIBUTION OF $n$-CHARACTER WORDS IN THE TWO MANDARIN CORPORA

| Corpus | Number of characters in a word | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 4+ |
| TDT2 | 152675<br>**(39.09%)** | 204321<br>**(52.31%)** | 23061<br>(5.90%) | 7139<br>(1.83%) | 3407<br>(0.87%) |
| CCTV | 121962<br>**(37.47%)** | 180141<br>**(55.34%)** | 16692<br>(5.13%) | 5022<br>(1.54%) | 1698<br>(0.52%) |

recognized words; and 3) most frequently used words in Chinese are one-character- or two-character-long. From Table V, we can see that the proportions of one-character and two-character words are over 90% in the both Mandarin corpora.

We observe different phenomena on the TDT2 English corpus when comparing results for word and different subword levels shown in Table III. In general, the word level outperforms all the subword levels and phoneme 4-gram and 5-gram achieve performances comparable to the word level. This observation is similar to Ng's result in phoneme based spoken document retrieval [42]. Table IV shows the distribution of words containing different number of phonemes in the TDT2 English corpus. Unlike the high concentration of one-character and two-character words in Chinese, the English words are distributed more evenly in the number of phonemes. For instance, in the TDT2 English corpus, the most frequently used words are 4-phoneme-long, which accounts for only 18.61% of the whole vocabulary. Therefore, a reasonable explanation to the less effectiveness of English subwords is as follows. No subword $n$-gram representation is appropriate for sentence similarity measure due to the flat distribution of words in the number of component phonemes. Although the use of subwords can recover some of the mis-recognized words, the robustness of subwords to speech recognition errors is canceled out by the matching failures raised by the sequential information of $n$-grams. Obviously, sentence similarity measure on the phoneme 4-gram representation (the subword level with best F1-measure) leads to matching failures for words with less than four component phonemes (that account for 25.61% of the vocabulary in the TDT2 corpus).

### F. Pseudo-Sentences: Fixed-Length-Based Versus Pause-Based

The experiential results show that fixed-length-based pseudo-sentence is an appropriate choice for lexical similarity measure. As discussed in Section IV-A, pause is a natural speech delimiter. Therefore, we also carried out story segmentation experiments on pause-separated pseudo-sentences for a sanity check.

A direct thought for pseudo-sentence construction is to use a pause threshold: a transcript is divided into sentence units at pause breaks with duration not less than a pre-defined threshold $\theta$. As explained in Section IV-A, pause-separated sentences lead to incomparable lexical similarity measures. Fig. 9 displays the numbers of characters in the pause-separated pseudo-sentences for a broadcast news LVCSR transcript in the CCTV Mandarin corpus. The threshold of pause $\theta$ is set to 0.7 second to keep a reasonable number of pseudo-sentences. We can clearly observe that the sentence lengths vary throughout the whole transcript and thus the similarity measure between two long sentences and between a long and a short sentence is incomparable.

To ameliorate the skewed sentence problem, we incorporate another parameter $\gamma$ with $\theta$ to prevent the length of a pseudo-sentence from being too short or too long. First, significant pauses $p_0, p_1, \ldots, p_m$ are found by a pre-defined threshold $\theta$, where $p_0$ is the beginning of the transcript. Second, $p_0$ is set to the start of the first pseudo-sentence and pause $p_k$ ($1 < k \leq m$) is considered as the end of the first pseudo-sentence if the accumulated speech duration between $p_0$ and $p_k$ is greater than or equal to the pre-defined length regulation factor $\gamma$. Then $p_k$ is supposed to be the beginning of the next pseudo-sentence. By iteration, pseudo-sentences are generated one by one accordingly. Fig. 10 shows the number of characters in the generated pseudo-sentences after using the length regulation factor $\gamma$ for the same broadcast news LVCSR transcript in Fig. 9. The value of $\theta$ is 0.2 second and the value of $\gamma$ is 11.0 seconds to keep the pseudo-sentences to a suitable number. It is easy to see that the skewed condition has been greatly alleviated.

Experimental results on the pause-separated pseudo-sentences for the test set of the CCTV Mandarin corpus is shown

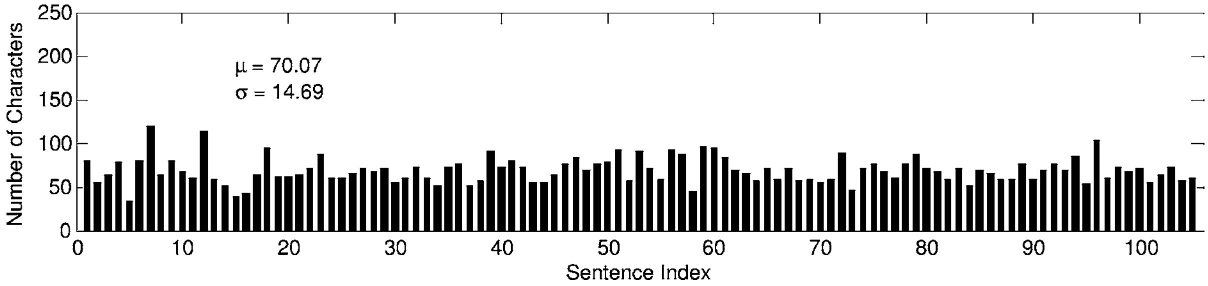


Fig. 10. Number of characters in the pause-separated pseudo-sentences for the same broadcast news program in Fig. 9 after sentence length regulation. $\mu$ denotes mean and $\sigma$ denotes standard deviation. Pause threshold $\theta = 0.2$ and length regulation factor $\gamma = 11.0$.

TABLE VI
STORY SEGMENTATION RESULTS (F1-MEASURE) ON THE PAUSE-SEPARATED PSEUDO-SENTENCES FOR THE CCTV MANDARIN CORPUS. CHAR.: CHARACTER, SYL.: SYLLABLE

| Approach | Word | | Unigram | | Bigram | | Trigram | | Quadgram | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Char. | Syl. | Char. | Syl. | Char. | Syl. | Char. | Syl. | Char. | Syl. |
| TextTiling | 0.5202 | 0.4925 | 0.5225 | 0.5272 | 0.5017 | 0.5110 | 0.5087 | 0.5087 | 0.4775 | 0.4751 |
| LE-TextTiling | 0.5491 | 0.5480 | 0.5942 | 0.5642 | 0.6092 | 0.5988 | 0.5445 | 0.5434 | 0.4960 | 0.4925 |
| Spectral Clustering | 0.5064 | 0.4983 | 0.5919 | 0.5653 | 0.5699 | 0.5723 | 0.4728 | 0.4728 | 0.4740 | 0.4821 |
| LE-DP | 0.5526 | 0.5561 | 0.6185 | 0.6023 | **0.6347** | **0.6370** | 0.5202 | 0.5306 | 0.5040 | 0.5064 |

TABLE VII
STORY SEGMENTATION RESULTS (F1-MEASURE) ON THE MANUAL TRANSCRIPTS OF THE CCTV MANDARIN CORPUS. CHAR.: CHARACTER, SYL.: SYLLABLE

| Approach | Word | | Unigram | | Bigram | | Trigram | | Quadgram | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Char. | Syl. | Char. | Syl. | Char. | Syl. | Char. | Syl. | Char. | Syl. |
| TextTiling | 0.5746 | 0.5653 | 0.6231 | 0.5931 | 0.5861 | 0.6000 | 0.5353 | 0.5364 | 0.5422 | 0.5422 |
| LE-TextTiling | 0.6058 | 0.6127 | 0.6555 | 0.6208 | 0.6775 | 0.6740 | 0.5988 | 0.5931 | 0.5480 | 0.5491 |
| Spectral Clustering | 0.6428 | 0.6486 | 0.6971 | 0.6879 | 0.7098 | 0.7283 | 0.6705 | 0.6578 | 0.6035 | 0.6139 |
| LE-DP | 0.6740 | 0.6671 | 0.7121 | 0.6879 | **0.7549** | **0.7491** | 0.6821 | 0.6844 | 0.6312 | 0.6335 |

in Table VI. The parameters $\theta$ and $\gamma$ were tuned on the development set to obtain the optimal settings. The tuning range of $\theta$ was from 0.1 to 0.4 with a step of 0.1 and the tuning range of $\gamma$ was from 8.0 to 12.0 by a step of 1.0. From Table VI, we observe that LE-DP performs the best at the character and syllable bigram levels. This observation is consistent with the fixed-length-based approaches shown in Table I. However, pause-based approaches demonstrate consistently inferior performances at all word and subword levels. For instance, at the syllable bigram level, the F1-measure scores of the pause-based LE-DP and the fixed-length-based LE-DP is 0.6370 and 0.7260, respectively. The apparent performance difference indicates that unequal sentence length adversely affects the lexical similarity measure between sentences and fixed-length sentence unit is more suitable to the story segmentation task.

### *G. Robustness of LE Methods to Speech Recognition Errors*

To analyze the effects from speech recognition errors, we experimented with TextTiling and the three LE-based approaches on the error-free manual transcripts in the CCTV Mandarin corpus. The experimental results are summarized in Table VII, which provide performance upper bounds of the word and different subword representations for story segmentation. We can see that the highest F1-measure on the manual transcript is 0.7549, achieved by the LE-DP method at the character bigram level.

Performance comparisons between the manual transcripts and the ASR transcripts are summarized in Table VIII. We can see that speech recognition errors adversely affect the story segmentation performances for all the methods and result in performance degradations as compared with error-free manual transcripts. This observation is consistent with our recent study on the effects from speech recognition errors [43]. At the same time, we observe that the LE-based approaches show only small F1-measure degradations as compared with TextTiling. For example, comparing at the best F1-measure level, the relative F1-measure degradation of LE-DP is only 3.83% (from 0.7549 to 0.7260) while that of conventional TextTiling is as high as 11.32% (from 0.6231 to 0.5526). This observation demonstrates that our proposed LE-based approaches show robustness to speech recognition errors.

### *H. Effectiveness of Sentence Distance Penalty $\alpha$*

Fig. 11 shows the F1-measure-versus-$\alpha$ curves of the proposed LE-DP method for the development and test sets of the three corpora. We can see that all the F1-measure curves rise rapidly to the peaks when $\alpha$ leaves 1.00 (i.e., when $alpha$ has no effect) and then fall continuously. All the F1-measure curves are almost flat with $\alpha$ in the range between 0.930 and 0.960. The F1-measure peak values on the development sets are reached when $\alpha = 0.958$, 0.942, 0.930 for the TDT2 Mandarin corpus, the TDT2 English corpus and the CCTV Mandarin corpus, respectively. Similarly, the F1-measure peak values on the test sets are reached when $\alpha = 0.932, 0.946, 0.956$for the corresponding corpora, respectively. From the observations, we conclude that: 1) a proper $\alpha$ between 0.930 and 0.960 can generally achieve good story segmentation performance for both development and

TABLE VIII
F1-MEASURE COMPARISON BETWEEN THE MANUAL TRANSCRIPTS (REF) AND ASR TRANSCRIPTS (ASR) ON THE CCTV MANDARIN CORPUS. RESULTS ON THE BEST F1-MEASURE LEVEL FOR EACH METHOD AND THE CHARACTER BIGRAM LEVEL ARE BOTH LISTED

| Approach | Best F1-measure level | | | Character bigram level | | |
|---|---|---|---|---|---|---|
| | REF | ASR | Relative degradation | REF | ASR | Relative degradation |
| TextTiling | 0.6231 (character unigram) | 0.5526 (character unigram) | 11.32% | 0.5861 | 0.5272 | 10.06% |
| LE-TextTiling | 0.6775 (character bigram) | 0.6509 (character unigram) | 3.92% | 0.6775 | 0.6324 | 6.66% |
| Spectral Clustering | 0.7283 (syllable bigram) | 0.6925 (character bigram) | 4.92% | 0.7098 | 0.6925 | 2.44% |
| LE-DP | 0.7549 (character bigram) | 0.7260 (syllable bigram) | 3.83% | 0.7549 | 0.7191 | 4.74% |

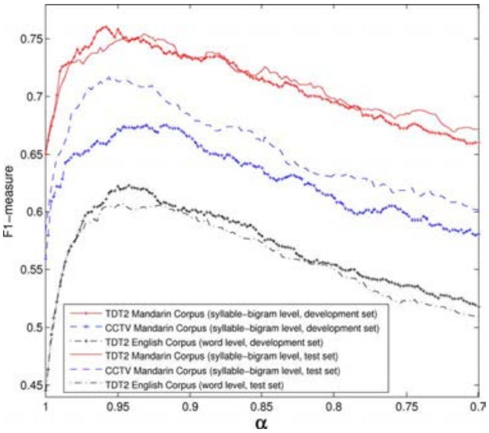


Fig. 11. F1-measure-versus-$\alpha$ curves for the LE-DP approach on the three corpora. A proper $\alpha$ between 0.930 and 0.960 can generally achieve good story segmentation performances on the development set and the test set, regardless of the corpus used.

test sets, regardless of the corpus used; and 2) considerable segmentation performance improvements can be achieved by incorporating distances between sentences into the measure of sentence connective strength. According to the Davis–Kahan theorem introduced in Section V-C, an appropriate sentence distance penalty factor can bring the real case closer to the ideal case, and hence the LE-based approaches work more effectively.

## VIII. CONCLUSION AND FUTURE WORK

In this paper, we have proposed to use Laplacian Eigenmaps to automatic story segmentation of broadcast news. Specifically, we apply LE analysis to sentence connective strength matrix in order to reveal the intrinsic cohesive lexical behaviors which assemble a story. We construct a Euclidean space in which each sentence is mapped to a vector. As a result, the cohesive relations between sentences becomes geometrically evident in the Euclidean space for discriminating different stories. Taking advantage of the LE method, we further present three segmentation approaches: LE-TextTiling, spectral clustering and LE-DP. Moreover, in order to make the LE-based approaches more effective to the segmentation task, we have introduced an explicit sentence distance penalty factor into the sentence connective strength measure. Experimental results on three corpora indicate that the proposed LE-based approaches achieve superior performances and significantly outperform several state-of-the-art story segmentation methods. Specifically, the LE-DP method consistently performs the best for the three corpora.

In the future, we plan to integrate a self-validated criterion [44] with the proposed LE-based approaches to automatically determine the number of stories in segmentation. Additionally, inverse document frequency (IDF), word/subword information integration, lattices, confusion networks, or other robust representations of recognition hypotheses will be considered for robust segmentation of versatile spoken documents with inferior speech recognition performance, e.g., spoken lectures and meeting recordings.

## ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their very helpful suggestions on the original version of the manuscript.

## REFERENCES

[1] *Topic Detection and Tracking: Event-Based Information Organization*, J. Allan, Ed.. Norwell, MA: Kluwer, 2002.

[2] L.-S. Lee and B. Chen, "Spoken document understanding and organization," *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 42–60, Sep. 2005.

[3] D. Beeferman, A. Berger, and J. Lafferty, "Statistical models for text segmentation," *Mach. Learn.*, vol. 34, no. 1-3, pp. 177–210, 1999.

[4] F. Y. Y. Choi, "Advances in domain independent linear text segmentation," in *Proc. NAACL*, 2000, pp. 26–33.

[5] J. Yamron, I. carp, L. Gillick, and P. Mulbregt, "A hidden Markov model approach to text segmentation and event tracking," in *Proc. ICASSP*, 1999, pp. 333–336.

[6] A. Rosenberg and J. Hirschberg, "Story segmentation of broadcast news in English, Mandarin and Arabic," in *Proc. HLT-NAACL*, 2006, pp. 125–128.

[7] S. Banerjee and I. A. Rudnicky, "A texttiling based approach to topic boundary detection in meetings," in *Proc. Interspeech*, 2006, pp. 57–60.

[8] I. Malioutov and R. Barzilay, "Minimum cut model for spoken lecture segmentation," in *Proc. ACL*, 2006, pp. 25–32.

[9] M. A. Hearst, "TextTiling: Segmenting text into multi-paragraph subtopic passages," *Comput. Linguist.*, vol. 23, no. 1, pp. 33–64, 1997.

[10] N. Stokes, J. Carthy, and A. Smeaton, "SeLeCT: A lexical cohesion based news story segmentation system," *J. AI Commun.*, vol. 17, no. 1, pp. 3–12, 2004.

[11] S. K. Chan, L. Xie, and H. M.-L. Meng, "Modeling the statistical behavior of lexical chains to capture word cohesiveness for automatic story segmentation," in *Proc. Interspeech*, 2007, pp. 2408–2411.

[12] L. Xie, J. Zeng, and W. Feng, "Multi-scale TextTiling for automatic story segmentation in Chinese broadcast news," in *Proc. Asia Inf. Retrieval Symp., LNCS*, 2008, vol. 4993, pp. 345–355.

[13] W.-K. Lo, W. Xiong, and H. Meng, "Automatic story segmentation using a Bayesian decision framework for statistical models of lexical chain features," in *Proc. ACL*, 2009.

[14] M. Halliday and R. Hasan, *Cohesion in English*. New York: Longman Group, 1976.

[15] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, pp. 1373–1396, 2002.

[16] M. Belkin, "Problem of learning on manifolds," Ph.D. dissertation, Univ. of Chicago, Chicago, IL, 2003.

[17] C. Ma, B. Byun, I. Kim, and C.-H. Lee, "A detection-based approach to broadcast news video story segmentation," in *Proc. ICASSP*, 2009, pp. 1957–1960.

[18] W. H. Hsu, L. S. Kennedy, S.-F. Chang, M. Franz, and J. Smith, "Columbia-IBM news video story segmentation in TRECVID 2004," in *Proc. CIVR*, 2005.

[19] C.-H. Wu and C.-H. Hsieh, "Story segmentation and topic classification of broadcast news via a topic-based segmental model and a genetic algorithm," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 8, pp. 1612–1623, Nov. 2009.

[20] J. M. Ponte and W. B. Croft, "Text segmentation by topic," in *Proc. ECDL*, 1997, pp. 113–125.

[21] 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*, 2004, vol. 5307, pp. 244–258.

[22] X. Wang, L. Xie, B. Ma, E. S. Chng, and H. Li, "Modeling broadcast news prosody using conditional random fields for story segmentation," in *Proc. APSIPA ASC*, 2010.

[23] Y. Liu, "Structural event detection for rich transcription of speech," Ph.D. dissertation, Purdue Univ., West Lafayette, IN, 2004.

[24] 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.

[25] G. A. Levow, "Prosody-based topic segmentation for Mandarin broadcast news," in *Proc. HLT-HAACL*, 2004, pp. 137–140.

[26] G. Tür and D. Hakkani-Tür, "Integrating prosodic and lexical cues for automatic topic segmentation," *Comput. Linguist.*, vol. 27, no. 1, pp. 31–57, 2001.

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

[28] 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 *Proc. HLT-NAACL*, 2007, pp. 193–169.

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

[30] S. Dharanipragada, M. Franz, J. Mccarley, S. Roukos, and T. Ward, "Story segmentation and topic detection in the broadcast news domain," in *Proc. DARPA Broadcast News Workshop*, 1999.

[31] O. Heinonen, "Optimal multi-paragraph text segmentation by dynamic programming," in *Proc. COLING-ACL*, 1998, pp. 1484–1486.

[32] P. Fragkou, V. Petridis, and A. Kehagias, "A dynamic programming algorithm for linear text segmentation," *J. Intell. Inf. Syst.*, vol. 23, no. 2, pp. 179–197, 2004.

[33] L. Xie and Y. Yang, "Subword lexical chaining for automatic story segmentation in Chinese broadcast news," in *Proc. PCM*, 2008, pp. 248–258.

[34] Y. Yang and L. Xie, "Subword latent semantic analysis for TextTiling-based automatic story segmentation of Chinese broadcast news," in *Proc. ISCSLP*, 2008, pp. 358–361.

[35] M. Belkin and P. Niyogi, "Towards a theoretical foundation for Laplacian-based manifold methods," *J. Comput. Syst. Sci.*, vol. 74, no. 8, pp. 1289–1308, 2008.

[36] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[37] 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, Lang. Process.*, vol. 14, no. 5, pp. 1526–1540, Sep. 2006.

[38] H. Lütkepohl, *Handbook of Matrices*. Chichester, U.K.: Wiley, 1997.

[39] C. Davis and W. M. Kahan, "The rotation of eigenvectors by a perturbation. III," *SIAM J. Numer. Anal.*, vol. 7, pp. 1–46, 1970.

[40] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[41] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in *Proc. SODA: ACM-SIAM Symp. Discrete Algorithms*, 2007.

[42] K. Ng, "Subword-based approaches for spoken document retrieval," Ph.D. dissertation, Dept. of Elect. Eng. and Comput. Sci., Mass. Inst. of Technol., Cambridge, 2000.

[43] L. Xie, Y. Yang, and Z.-Q. Liu, "On the effectiveness of subwords for lexical cohesion based story segmentation of Chinese broadcast news," *Inf. Sci.*, vol. 181, no. 13, pp. 2873–2891, Jul. 2011.

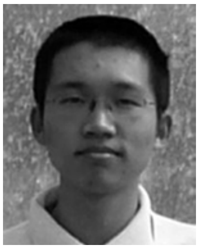
[44] W. Feng, J. Jia, and Z.-Q. Liu, "Self-validated labeling of Markov random fields for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1871–1887, Oct. 2010.

**Lei Xie** (M'07) 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. 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 the IEEE TRANSACTIONS ON MULTIMEDIA, 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.

**Lilei Zheng** received the B.S. degree in computer science and technology from Northwestern Polytechnical University, Xi'an, China, in 2009, where he is currently working toward the M.S. degree in the School of Computer Science.

From 2010 to 2011, he was with the Human Language Technology Department, Institute, Infocomm Research, A*STAR, Singapore, as an intern. His current research interest includes speech and language processing and human–computer interaction.



**Zihan Liu** received the B.S. degree in computer science and technology from Northwestern Polytechnical University, Xi'an, China, in 2010. He is currently pursuing the Ph.D. degree in the School of Creative Media, City University of Hong Kong.

His research interests include computer vision and machine learning.

**Yanning Zhang** (M'08) received the B.S. degree from the Dalian University of Technology, Dalian, China, in 1988 and the M.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 1993 and 1996, respectively.

She was a Visiting Researcher with the Chinese University of Hong Kong, Shatin, Hong Kong, and the University of Sydney, Sydney, Australia. She is currently a Professor with the School of Computer Science, Northwestern Polytechnical University. Her interests are concerned with image processing, pattern recognition, and computer vision.