ACOUSTIC TEXTTILING FOR STORY SEGMENTATION OF SPOKEN DOCUMENTS

Lilei Zheng[1,2], *Cheung-Chi Leung*[2], *Lei Xie*[1], *Bin Ma*[2], *Haizhou Li*[2]

[1]Shaanxi Provincial Key Laboratory of Speech and Image Information Processing,
School of Computer Science, Northwestern Polytechnical University, China
[2] Institute for Infocomm Research, A*STAR, Singapore

ABSTRACT

We propose an acoustic TextTiling method based on segmental dynamic time warping for automatic story segmentation of spoken documents. Different from most of the existing methods using LVCSR transcripts, this method detects story boundaries directly from audio streams. In analogy to the cosine-based lexical similarity between two text blocks in a transcript, we define the acoustic similarity measure between two pseudo-sentences in an audio stream. Experiments on TDT2 Mandarin corpus show that acoustic TextTiling can achieve comparable performance to lexical TextTiling based on LVCSR transcripts. Moreover, we use MFCCs and Gaussian posteriorgrams as the acoustic representations in our experiments. Our experiments show that Gaussian posteriorgrams are more robust to perform segmentation for the stories each with multiple speakers.

Index Terms— story segmentation, topic segmentation, segmental dynamic time warping, TextTiling, spoken document processing

1. INTRODUCTION

Story segmentation aims to partition a text, audio or video stream into a sequence of topically coherent segments named as stories. It is a necessary pre-processing step for various tasks such as topic categorization and tracking, summarization, information extraction, indexing and retrieval [1, 2].

With progress in large vocabulary continuous speech recognition (LVCSR), lexical cohesion based methods have drawn much attention for story segmentation of spoken documents [1, 3]. TextTiling [4] is a typical and efficient lexical cohesion based approach that has been introduced to segment spoken documents such as broadcast news (BN) [1] and meeting recordings [3]. While TextTiling performs boundary identification using the similarities between adjacent sentences, some other methods such as dynamic programming (DP) [5] take into account some global criteria based on all the inter-sentence similarities in a spoken document. One disadvantage of the lexical cohesion based methods is that the performance heavily relies on an LVCSR system, which is built using considerable linguistic resources and requires tremendous effort to collect a large amount of training data. Consequently, researchers have been interested in processing spoken documents without using LVCSR. Specifically, speech prosodic cues have received attention for story segmentation task [1, 2]. However, prosodic cues depend on editorial and production rules which vary from different media sources. Finding acoustic pattern repetitions in speech [6, 7] has been another interesting research topic recently.

In this paper, we present an acoustic TextTiling method to detect story shifts without using LVCSR. While lexical TextTiling measures semantic variations between adjacent stories in a transcript, our method achieves this directly from the corresponding audio stream. The core technique for acoustic TextTiling is a segmental dynamic time warping (SDTW) algorithm proposed by Park and Glass for word discovery [6]. The SDTW algorithm finds alignment paths between two given utterances in vector representation. These paths probably correspond to similar acoustic patterns (and therefore common words and phrases) in the two utterances. Malioutov *et al.* [8] employed the SDTW algorithm for story segmentation on lectures each with a single speaker. However, there are usually multiple speakers present in real-world spoken documents such as meeting recordings or news programs. So we propose to use acoustic TextTiling for story segmentation of spoken documents with multiple speakers.

We use the SDTW algorithm to find repeated acoustic patterns in adjacent pseudo-sentences in a spoken document. Acoustic TextTiling is applied to measure similarities between the adjacent pseudo-sentences based on the repeated patterns. The local minima of the similarity values indicate the boundaries between stories. We use mel frequency cepstral coefficients (MFCCs) and Gaussian posteriorgrams as the signal representations. The latter has been demonstrated to be effective for discovering patterns between utterances from different speakers [9]. In our experiments, we compare the two representations for story segmentation of spoken documents with multiple speakers.

2. LEXICAL TEXTTILING FOR STORY SEGMENTATION

The classical TextTiling algorithm is composed of three steps: tokenization, lexical score determination and boundary identification. The tokenization step splits a text stream into individual lexical terms. In lexical score determination, the Text-

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

Tiling algorithm first divides the text into sentences. Lexical similarities are calculated at all sentence boundaries:

$$Cos(\mathbf{s}_i, \mathbf{s}_{i+1}) = \frac{\sum_{k=1}^{M} e_{i,k} e_{i+1,k}}{\sqrt{\sum_{k=1}^{M} e_{i,k}^2 \sum_{k=1}^{M} e_{i+1,k}^2}} \quad (1)$$

where $\mathbf{s}_i$ and $\mathbf{s}_{i+1}$ are term frequency vectors of the ith and the $(i+1)$th sentences respectively. $e_{i,k}$ is the kth element of $\mathbf{s}_i$, representing the frequency of term t_k appeared in the ith sentence. M is the vocabulary size.

In the next step, we adopt *relative scores* [2] instead of the lexical similarity scores for boundary identification:

$$\begin{aligned} Rel(\mathbf{s}_i, \mathbf{s}_{i+1}) = &(Cos(\mathbf{s}_{i-1}, \mathbf{s}_i) - Cos(\mathbf{s}_i, \mathbf{s}_{i+1})) \\ &+ (Cos(\mathbf{s}_{i+1}, \mathbf{s}_{i+2}) - Cos(\mathbf{s}_i, \mathbf{s}_{i+1})) \quad (2) \end{aligned}$$

If $Rel(\mathbf{s}_i, \mathbf{s}_{i+1})$ exceeds a pre-defined threshold γ, a story boundary is assigned between the ith and the $(i+1)$th sentences. The parameter γ is tuned on a development set.

3. ACOUSTIC TEXTTILING FOR STORY SEGMENTATION

The intuitive idea of lexical TextTiling is that different stories employ different sets of words and the shifts in vocabulary use indicate the boundaries between stories. Specifically, sentences in a story usually employ the same set of words such as person names and place names which are keys to topic discrimination. Analogously, we believe that the repetitions of similar acoustic patterns can be found in the utterances of a story, and this can reflect story shifts in a spoken document.

3.1. Sentence Construction

A text document is composed of sentences that separated by delimiters, e.g., periods. However, sentence delimiters are not readily available in speech. A natural thought is to use significant pauses as delimiters, resulting in pause-separated pseudo-sentences [5].

Firstly, pause-separated utterances are formed when we employ a voice activity detector (VAD) to detect significant pauses with duration longer than ψ in an audio recording. Secondly, pseudo-sentences are formed by concatenating a number of utterances. For instance, if a pseudo-sentence p contains utterances $(u_s, \cdots, u_e)$, the utterance u_{e+1} will be added in if: 1) the total duration of $(u_s, \cdots, u_e)$ is less than α; and 2) the pause region between u_e and u_{e+1} is less than β. Otherwise, p is formed by $(u_s, \cdots, u_e)$ and u_{e+1} is considered as the beginning of the next pseudo-sentence. The parameters are empirically set as ψ=0.32 second, α=10.2 seconds and β=0.96 second. It aims to prevent the pseudo-sentences from being too short or too long, because the similarity scores between two long sentences and between a long and a short sentence are probably incomparable [5].

3.2. Feature Vectors

For the vector representation of speech signals, we first use MFCCs that are widely used in speech processing applications, e.g., LVCSR. Each speech frame is represented by a standard 39-dimensional MFCCs (25ms windows, 10ms step). An additional process of whitening is carried out to make elements of the feature vectors uncorrelated and normalize the variance in each dimension [6].

Zhang *et al.* [9] has demonstrated that Gaussian posteriorgrams are more effective for comparing speech from different speaker in the task of pattern discovery. So we train a Gaussian mixture model (GMM) on all speech frames of whitened MFCCs to generate Gaussian posteriorgram vectors.

Given two utterances, u_x and u_y, we can represent them as two time series of feature vectors, $(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ and $(\boldsymbol{y}_1, \cdots, \boldsymbol{y}_m)$, respectively. n and m denote the number of frames in u_x and u_y. We define a distance matrix $\boldsymbol{D}$ to measure the distances between the frame vectors of the two utterances. For MFCC vectors, the Euclidean distance measure is used:

$$d_{ij} = \|\boldsymbol{x}_i - \boldsymbol{y}_j\|, \quad (3)$$

and for Gaussian posteriorgram vectors, a negative log inner-product is used:

$$d_{ij} = -\log(\boldsymbol{x}_i \cdot \boldsymbol{y}_j), \quad (4)$$

where d_{ij} denotes the distance between the ith frame of u_x and the jth frame of u_y.

3.3. Segmental Dynamic Time Warping

The SDTW algorithm [6] divides the distance matrix $\boldsymbol{D}$ into a set of diagonal bands of width R and searches for an optimal alignment path within each band. A band overlap of 50% is used to take into account alignment paths across segmentation boundaries [9]. So the total number of bands N is $\lfloor \frac{n-1}{R} + \frac{m-1}{R} \rfloor$. For each band, an optimal alignment path that minimizes the total distance between the two time series of feature vectors is obtained using the traditional dynamic time warping (DTW) algorithm. Instead of an entire optimal alignment path, we are interested in portions of the optimal alignment path that probably correspond to similar acoustic patterns in the two utterances. Therefore, we divide the optimal alignment path into a number of fragments each with the average distance score smaller than a threshold θ and the length at least L as follows.

For an optimal alignment path with length N_p in a band, we use a *path refinement* algorithm proposed by Lin *et al.* [10] to locate its length-constrained minimum average (LCMA) fragment. The LCMA fragment is a fragment with the smallest average distance score and the length at least L in the given path. The average distance score of the LCMA fragment is:

$$f = \min_{1 \le s < t \le N_p} \frac{1}{t-s+1} \sum_{k=s}^{t} d_{i_k j_k}, \; t-s+1 \ge L \quad (5)$$

where $d_{i_k j_k}$ is the distance value of the kth entry in the path, i_k and j_k denote the row and column indices of this entry in matrix $\boldsymbol{D}$. If the average distance score f of the LCMA fragment is larger than a threshold θ, the LCMA fragment is discarded because it is less likely that this fragment corresponds to two similar acoustic patterns. Secondly, the *path refinement* algorithm is employed to iteratively find out all the fragments

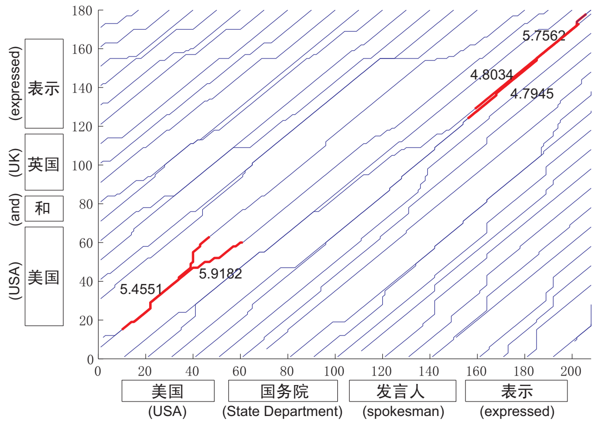

Fig. 1. Fragments discovered between two utterances by MFCC-based SDTW. R=10, L=50, θ=6.0.

in the optimal alignment path with the average distance score smaller than θ and the length at least L.

Figure 1 shows an example of the fragments discovered between two utterances by the SDTW algorithm. In this example, speech frames are represented by MFCCs. There are two word repetitions between the two utterances, the Chinese words "美国" (USA) and "表示" (expressed). We can observe that multiple fragments from different bands have been labeled by red bold lines in the cross regions where the same words occur in the two utterances. The fragment with the smallest average distance score 4.7945 corresponds to the repetition of word "表示" (expressed), and so do the fragments with the average distance scores 4.8034 and 5.7562. Since the band overlap of 50% is used, any cross region where similar acoustic patterns occur in two utterances would be covered by more than two bands. Thus multiple fragments from different bands are supposed to be found in the cross region. If only one fragment is found in the cross region, we treat the fragment as unreliable and discard it.

3.4. Acoustic Similarity

To calculate the similarity between two utterances, the average distance scores of all the reserved fragments are used:

$$Utt\_Sim(u_x, u_y) = \sum_{i=1}^{N_f}(1 - \frac{f_i}{\theta}), \tag{6}$$

where N_f denotes the number of the fragments reserved, and f_i is the average distance score of the ith fragment.

Since each pseudo-sentence may contain several paused-separated utterances, the acoustic similarity between two pseudo-sentences p_i $(u_1, \cdots, u_K)$ and p_{i+1} $(v_1, \cdots, v_L)$ is defined as the sum of the similarities between the utterances in the two sentences:

$$Sent\_Sim(p_i, p_{i+1}) = \sum_{k=1}^{K}\sum_{l=1}^{L} Utt\_Sim(u_k, v_l), \tag{7}$$

where u_k is the kth utterance of the first sentence p_i, v_l is the lth utterance of the second sentence p_{i+1}, K and L denote the number of utterances in the two sentences, respectively. After that, we use *relative scores* as defined in Eq. (2) to identify the story boundaries.

4. EXPERIMENTS

4.1. Corpus and Experiment Setup

We experimented on the TDT2 Mandarin BN corpus[1] which contains about 53 hours of Mandarin BN audio from Voice of America (VOA). To appropriately compare the performance of using the MFCC and Gaussian posteriorgram representations for story segmentation, all the episodes with multiple speakers were selected. They include 48 short episodes with length around 10 minutes and 39 long episodes with length around 60 minutes. The corpus provides manually annotated meta-data including story boundaries and LVCSR transcripts with the word error rate of 37%. The 48 short episodes were divided into two non-overlapping sets: a development set of 24 episodes with 307 story boundaries for parameter tuning and a test set of 24 episodes with 308 story boundaries for performance evaluation. The 39 long episodes were also divided into a development set of 20 episodes with 607 story boundaries and a test set of 19 episodes with 654 story boundaries. According to TDT2 standards, a detected story boundary is considered correct if it lies within a 15-second tolerant window on each side of a manually-annotated reference boundary. The balanced F1-measure, i.e., the harmonic mean of precision and recall, was adopted as the evaluation criterion.

We experimented with lexical TextTiling on LVCSR transcripts and acoustic TextTiling on audio recordings represented using MFCCs and Gaussian posteriorgrams. We first conducted empirical parameter tuning on the development set to obtain optimal parameter setting that achieved the best performance of story segmentation, and carried out evaluation on the test set using the best-tuned parameters. The parameters were diagonal bands width R, fragment length constraint L and distance pruning threshold θ.

4.2. Results and Analysis

The segmentation performance of acoustic and lexical TextTiling approaches is summarized in Table 1. In this study, we take the word-based lexical TextTiling as the baseline [5]. We observe that acoustic TextTiling using the Gaussian posteriorgram representation achieves comparable F1-measures on both the short and long episodes (0.6596 and 0.3986 respectively) to word-based lexical TextTiling using LVCSR (0.6597 and 0.4197 respectively). Moreover, acoustic TextTiling using the MFCC representation achieves the best F1-measure of 0.7086 on the short episodes but the worst F1-measure of 0.3482 on the long episodes. These results are further confirmed by statistical significance tests.

Table 2 reports the performance in detecting word repetitions by using the MFCC and Gaussian posteriorgram representations in SDTW on (a) short episodes and (b) long episodes. The F1-measure used here is for the evaluation

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

Table 1. Segmentation performance of acoustic and lexical TextTiling (TT) approaches on (a) short episodes and (b) long episodes. GPs: Gaussian posteriorgrams

Approach	F1-measure	
	(a)	(b)
Acoustic TT on MFCCs	0.7086	0.3482
Acoustic TT on GPs	0.6596	0.3986
Word-based Lexical TT (*baseline*)	0.6597	0.4197

Table 2. Performance in detecting word repetitions by using MFCCs and Gaussian posteriorgrams in SDTW on (a) short episodes and (b) long episodes

(a)

Approach	SDTW using MFCCs	SDTW using GPs
N_{tran}	3060	
N_{find}	2281	2634
N_{corr}	1479	1439
Precision	0.6484	0.5451
Recall	0.4833	0.4703
F1-measure	0.5538	0.5049

(b)

Approach	SDTW using MFCCs	SDTW using GPs
N_{tran}	13505	
N_{find}	18103	12303
N_{corr}	3643	3666
Precision	0.2012	0.2980
Recall	0.2698	0.2715
F1-measure	0.2305	0.2841

GPs: Gaussian posteriorgrams;
N_{tran}: number of word repetitions in LVCSR transcripts;
N_{find}: number of acoustic patterns discovered;
N_{corr}: number of correctly discovered acoustic patterns.

of pattern discovery. It is different from the one in Table 1. Since the optimal length constraint L for SDTW in audio streams is tuned to be 500 ms, the patterns containing only one Chinese character fail to be discovered. Therefore, for word repetitions in the LVCSR transcripts, we only take into account the words involving two or more Chinese characters.

We observe that the results of pattern discovery are consistent with the segmentation performance reported in Table 1. SDTW using MFCCs performs better than that using Gaussian posteriorgrams for the short episodes, but reversely for the long episodes. Using MFCCs as the acoustic representation is ineffective to discover patterns from the utterances with different speaker. In other words, it is effective to detect speaker changes in speech. In the short episodes, a story usually includes only one speaker, so speaker changes usually indicate story shifts. This is probably the reason why MFCC-based acoustic TextTiling performs the best in the segmentation of the short episodes. However, since multiple speakers may occur in a story of a long episode, speaker changes in a story may be incorrectly considered as story boundaries. This produces many false alarms which lead MFCC-based acoustic TextTiling to the worst F1-measure in the segmentation of the long episodes.

5. CONCLUSIONS

This paper proposes an acoustic TextTiling method based on SDTW to measure semantic variations directly from audio streams and identify the story boundaries. Our experimental results demonstrate that acoustic TextTiling perform comparably to lexical TextTiling for story segmentation of spoken documents with multiple speakers. In the future, we will extend our work to some segmentation methods in which global criteria [5] are used for story boundary identification. In this case, we will take into account all the inter-sentence acoustic similarities in a spoken document.

6. ACKNOWLEDGMENT

This work was supported by the National Natural Science 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).

7. REFERENCES

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

[2] 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, pp. 253–256.

[3] S. Banerjee and I. A. Rudnicky, “A TextTiling based approach to topic boundary detection in meetings,” in *Proc. Interspeech*, 2006, pp. 57–60.

[4] M.A. Hearst, “TextTiling: Segmenting text into multi-paragraph subtopic passages,” *Computational Linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[5] L. Xie, L. Zheng, Z. Liu, and Y. Zhang, “Laplacian eigenmaps for automatic story segmentation of broadcast news,” *IEEE Transactions on Audio, Speech & Language Processing*, vol. 20, no. 1, pp. 264–277, 2012.

[6] A. S. Park and J. R. Glass, “Unsupervised pattern discovery in speech,” *IEEE Transactions on Audio, Speech & Language Processing*, vol. 16, no. 1, pp. 186–197, 2008.

[7] M. Dredze, A. Jansen, G. Coppersmith, and K. W. Church, “NLP on spoken documents without ASR,” in *Proc. EMNLP*, 2010, pp. 460–470.

[8] I. Malioutov, A. S. Park, R. Barzilay, and James R. Glass, “Making sense of sound: Unsupervised topic segmentation over acoustic input,” in *Proc. ACL*, 2007, pp. 504–511.

[9] Y. Zhang and J. R. Glass, “Towards multi-speaker unsupervised speech pattern discovery,” in *Proc. ICASSP*, 2010, pp. 4366–4369.

[10] Y. L. Lin, T. Jiang, and K. M. Chao, “Efficient algorithms for locating the length-constrained heaviest segments with applications to biomolecular sequence analysis,” *JCSS: Journal of Computer and System Sciences*, vol. 65, pp. 570–586, 2002.