Multi-layered Summarization of Spoken Document Archives by Information Extraction and Semantic Structuring

Lin-shan Lee, Sheng-yi Kong, Yi-cheng Pan, Yi-sheng Fu, Yu-tsun Huang

Speech Lab, College of EECS
National Taiwan University, Taipei, Taiwan, Republic of China
lslee@gate.sinica.edu.tw

Abstract

The spoken documents are very difficult to be shown on the screen, and very difficult to retrieve and browse. It is therefore important to develop technologies to summarize the entire archives of the huge quantities of spoken documents in the network content to help the user in browsing and retrieval. In this paper we propose a complete set of multi-layered technologies to handle at least some of the above issues: (1) Automatic Generation of Titles and Summaries for each of the spoken documents, such that the spoken documents become much more easier to browse, (2) Global Semantic Structuring of the entire spoken document archive, offering to the user a global picture of the semantic structure of the archive, and (3) Query-based Local Semantic Structuring for the subset of the spoken documents retrieved by the user's query, providing the user the detailed semantic structure of the relevant spoken documents given the query he entered. The Probabilistic Latent Semantic Analysis (PLSA) is found to be helpful, and an initial prototype system for the functions mentioned above has been successfully developed, in which the broadcast news archive in Mandarin Chinese is taken as the example archive.

Index Terms: summarization, spoken document archive, semantic structure.

1. Introduction

In the future network era, the digital content over the network will include all the information activities for human life. Apparently, the most attractive form of the network content will be in multimedia including speech information, and such speech information usually tells the subjects, topics and concepts of the multi-media content. As a result, the spoken documents associated with the network content will become the key for retrieval and browsing. However, unlike the written documents with well structured paragraphs and titles, the multi-media and spoken documents are both very difficult to retrieve or browse, since they are just audio/video signals, very difficult to be shown on the screen, and the user can not go through each of them from the beginning to the end during browsing. As a result, it will be very important to develop a set of technologies to summarize the entire archives of the spoken documents to help the user to browse and retrieve the multimedia/spoken documents [1]. Such summarization technologies for the entire archives at least need a few key elements: information extraction (to extract key information from the spoken documents), document archive structuring (to organize the archive of spoken documents into some form of hierarchical structures) and query-based (able to respond to the user's query to offer the information about a subset of the archive relevant to user's interest).

Note that while some of the technologies mentioned above have been studied or explored to a good extent, most of the work has been performed independently within individual scopes. Great efforts have been made to try to integrate several of these technologies for a specific application , and several well-known research projects have been successfully developed towards this goal. Examples include the Informedia System at Carnegie Mellon University [2], the Multimedia Document Retrieval Project at Cambridge University [3], the Rough'n'Ready System at BBN technologies [4], the Speech Content-based Audio Navigator (SCAN) System at AT&T Labs-Research [5], the Broadcast News Navigator at MITRE Corporation [6], the SpeechBot Audio/Video Search System at Hewlett-Packard (HP) Labs [7], the National Project of Spontaneous Speech Corpus and Processing Technologies of Japan [8], and the NewsBlaster project of Columbia University [9].

In this paper we propose a complete set of multi-layered technologies to handle at least some of the above issues: (1) Automatic Generation of Titles and Summaries for each of the spoken documents, (2) Global Semantic Structuring of the entire spoken document archive, and (3) Query-based Local Semantic Structuring for the subset of the spoken documents retrieved by the user's query. All of the above have to do with the analysis of the semantics carried by the spoken documents. Here we propose that the Probabilistic Latent Semantic Analysis (PLSA) recently developed for semantic analysis is very helpful. An initial prototype system for the functionalities mentioned above has been successfully developed, in which the broadcast news archive in Mandarin Chinese is taken as the example archive.

2. Proposed Approaches

2.1. Probabilistic Latent Semantic Analysis (PLSA)

The set of documents $\{d_i, i = 1, 2, \ldots, N\}$ have been conventionally analyzed by the terms $\{t_j, j = 1, 2, \ldots, L\}$ they may include, usually with statistical approaches. In recent years, efforts have also been made to establish a probabilistic framework for such purposes with improved model training algorithms, of which the Probabilistic Latent Semantic Analysis (PLSA)[10] is often considered as a representative. In PLSA, a set of latent topic variables is defined, $\{T_k, k = 1, 2, \ldots, K\}$, to characterize the "term-document" co-occurrence relationships.Both the document d_i and the term t_j are assumed to be independently conditioned on an associated latent topic T_k. The conditional probability of a document d_i generating a term t_j thus can be parameterized by

$$P(t_j|d_i) = \sum_{k=1}^{K} P(t_j|T_k)P(T_k|d_i). \qquad (1)$$

Notice that this probability is not obtained directly from the frequency of the term t_j occurring in d_i, but instead through $P(t_j|T_k)$, the probability of observing t_j in the latent topic T_k, as well as $P(T_k|d_i)$, the likelihood that d_i addresses the latent topic T_k. The PLSA model can be optimized with the EM algorithm by maximizing a carefully defined likelihood function [10].

2.2. Automatic Generation of Titles and Summaries for Spoken Documents

The titles exactly complement the summaries for the user during browsing and retrieval. The user can easily select the desired document with a glance at the list of titles. He can then look through or listen to the summaries in text or speech form for the titles he selected.

Substantial efforts have been made in automatic generation of titles and summaries for spoken documents [11, 12, 13, 14, 15]. In this research, it was found that the topic entropy of the terms estimated from probabilities obtained in PLSA analysis is very useful in finding the key terms of the spoken documents in automatic generation of titles and summaries [16][17]. The topic entropy of a term t_j is evaluated from the topic distribution $\{P(T_k|t_j), k = 1, 2, \dots, K\}$ of the term obtained from PLSA and defined as:

$$EN(t_j) = -\sum_{k=1}^{K} P(T_k|t_j) \log P(T_k|t_j) \qquad (2)$$

Clearly, a lower entropy implies the term carries more topical information for a few specific latent topics, thus is more significant semantically.

In this research, it was found that the sentences selected based on the topic entropy can be used to construct better summaries for the spoken documents [16]. For title generation, a new delicate scored Viterbi approach was developed in this research based on the concept of the previously proposed statistical translation approach [12]. In this new approach, the key terms in the automatically generated summaries are carefully selected and sequenced by a Viterbi beam search using three sets of scores. This new delicate scored Viterbi approach was further integrated with the previously proposed adaptive K-nearest neighbor approach [18] to offer better results [17].

2.3. Global Semantic Structuring for the Spoken Document Archive

The purpose of global semantic structuring for spoken document archives is to offer an overall knowledge of the semantic content of the entire spoken document archive in some form of hierarchical structures with concise visual presentation to help the user to browse across the spoken documents efficiently. In this research, we developed successfully a new approach to analyze and structure the topics of spoken documents in an archive into a two-dimensional tree structure or a multi-layer map for efficient browsing and retrieval [19]. The basic approach used is based on the PLSA concept. In the constructed two-dimensional tree structure, the spoken documents are clustered by the latent topics they primarily address, and the clusters are organized as a two-dimensional map. The nodes on the map represent the clusters, each labeled by several key terms with the highest scores for the cluster. The nodes are organized on the map in such a way that the distances between nodes have to do with the relationships between the topics of the clusters, i.e., closely located nodes represent clusters with closely related topics. Every node can then be expanded into another two-dimension map in the next layer with nodes representing finer topics. In this way the entire spoken archive can be structured into a two-dimensional tree, or a multi-layered map, representing the global semantics of the archive [19]. This is very useful for browsing and retrieval purposes.

2.4. Query-based Local Semantic Structuring for Spoken Documents Relevant to User's Interests

The global semantic structure mentioned above is very useful, but not necessarily good enough for the user regarding his special information needs , very often represented by the query he entered to the information retrieval engine. The problem is that the query given by the user is usually very short and thus not specific enough, and as a result a large number of spoken documents are retrieved, including many noisy documents retrieved due to the uncertainty in the spoken document retrieval. However, as mentioned above, the spoken documents are very difficult to be shown on the screen and very difficult to browse. The large number of retrieved spoken document therefore becomes a difficult problem. It is thus very helpful to construct a local semantic structure for the retrieved spoken documents for the user to identify what he really needs to go through or to specify what he really wish to obtain. This semantic structure is localized to user's query, constructed from those retrieved documents only, thus needs to be much more delicate over a very small subset of the entire archive. This is why the global semantic structure proposed above in section 2.3 cannot be used here. Instead in this research we propose to construct a very fine topic hierarchy for the localized retrieved documents. Every node on the hierarchy represents a small cluster of the retrieved documents, and is labeled by a key term as the topic of the cluster. The user can then click on the nodes or topics to select the documents he wishes to browse, or to expand his query by adding the selected topics onto his previous query [20].

The approach we used in this research for topic hierarchy construction is the Hierarchical Agglomerative Clustering and Partitioning algorithm (HAC+P) recently proposed for text documents [21]. This algorithm is performed on-line in real time on the retrieved spoken documents. It consists of two phases: an HAC-based clustering to construct a binary-tree hierarchy and a partitioning (P) algorithm to transform the binary-tree hierarchy to a balanced and comprehensive m-ary hierarchy, where m can be different integers at different splitting nodes. The first phase of HAC algorithm is kind of standard, based on the similarity between two clusters C_i and C_j and is performed bottom-up, while the second phase of partitioning is top-down. In this second phase, the binary-tree is partitioned into several sub-hierarchies first, and then this procedure is applied recursively to each sub-hierarchy. The point is that in each partitioning procedure the best level at which the binary-tree hierarchy should be cut in order to create the best set of sub-hierarchies has to be determined based on the balance of two parameters: the cluster set quality and the number preference score [20].

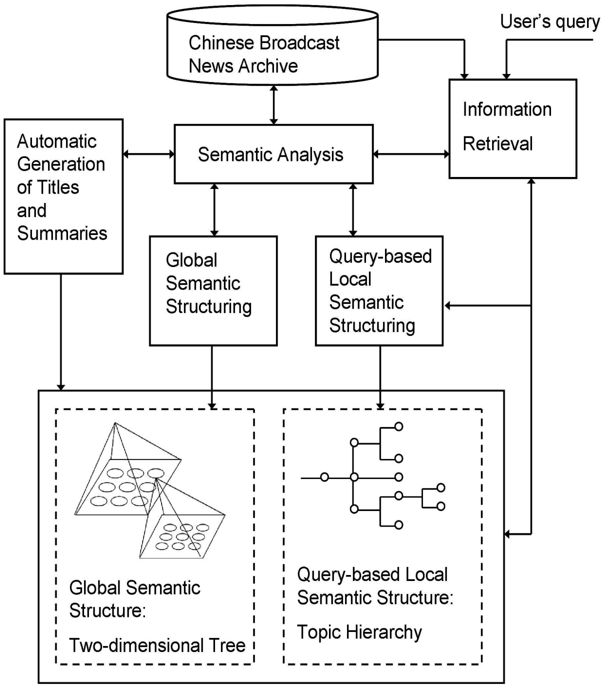
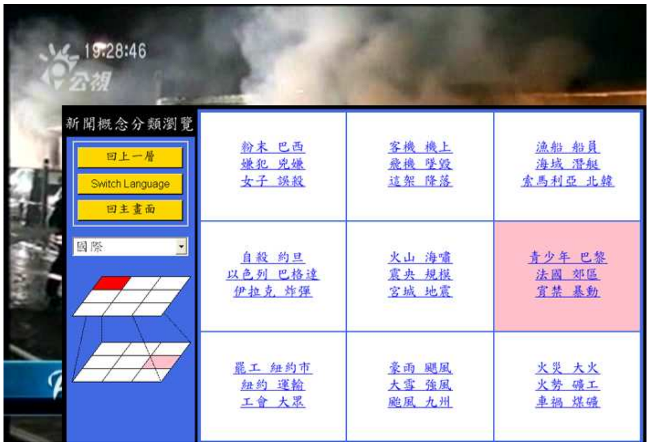

Chinese Broadcast News Archive
User's query
Information Retrieval
Automatic Generation of Titles and Summaries
Semantic Analysis
Global Semantic Structuring
Query-based Local Semantic Structuring
Global Semantic Structure:
Two-dimensional Tree
Query-based Local Semantic Structure:
Topic Hierarchy

Figure 1: *The block diagram of the initial prototype system*

3. An Initial Prototype System

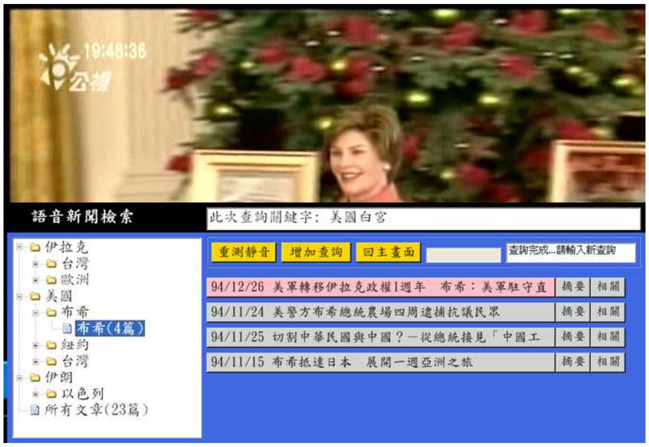
An initial prototype system has been successfully developed. The broadcast news are taken as the example spoken/multi-media documents. The broadcast news archive to be summarized includes two sets, all in Mandarin Chinese. The first has roughly 85 hours of about 5,000 news stories, recorded from radio/TV stations in Taipei from Feb. 2002 to May 2003. No video signals were kept with them. The character and syllable error rates of 14.29% and 8.91% respectively were achieved in the transcriptions. The second set has roughly 25 hours of about 800 news stories, all including the video signal parts, recorded from a TV station in Taipei from Oct. to Dec. 2005. The character and syllable error rates for this set is 20.92% and 13.90%.

For those news stories with video signals, the video signals were also summarized using video technologies, for example, video frames for human faces, with moving objects and scene changes are more important, and the length of the video summary is based on the length of the speech summary. For the global semantic structure, a total of six two-dimensional tree structures were obtained for six categories of news stories, e.g. world news, business news, sports news, etc. A 3x3 small map on the second layer of the tree for world news overlaid with the video signal is shown in Fig. 2. This is a map expanded from a cluster in the first layer covering all disasters happening worldwide. As can be found that on this map one small cluster is for airplane crash (墜機) and similar, one for earthquake (地震) and similar, one for hurricane (颶風) and similar, and so on. All news stories belonging to each node of the two-dimensional tree are listed under the node by their automatically generated titles. The user can easily browse through the titles or click to view either the summaries or the complete stories. With this structure it is much more easier for the user to browse the news stories either top-down or bottom-up. For the query-based local semantic structuring, the topic hierarchy constructed in real-time from the news stories retrieved by a query, “White House of United States (美國白宮),” is shown on the left lower corner of Fig 3, in which the three topics on the first layer are respectively Iraq (伊拉克), US (美國) and Iran (伊朗), and one of the node in the second layer below US is President George Bush (布希). When the user clicks the node of President George Bush, the relevant news stories are listed on the right lower corner by their automatically generated titles. The user can then click the “summary” button to view the summary, or click the titles to view the complete stories. Such information are overlaid with the news retrieved with the highest score.



粉末 巴西 嫌犯 兇嫌 女子 誤殺	客機 機上 飛機 墜毀 這架 降落	漁船 船員 海域 潛艇 索馬利亞 北韓
自殺 約旦 以色列 巴格達 伊拉克 炸彈	火山 海嘯 震央 規模 宮城 地震	青少年 巴黎 法國 郊區 宵禁 暴動
罷工 紐約市 紐約 運輸 工會 大眾	豪雨 颶風 大雪 強風 颱風 九州	火災 大火 火勢 礦工 車禍 煤礦

Figure 2: A 3x3 map on the second layer expanded from a cluster on the first layer of the global semantic structure for world news.



此次查詢關鍵字： 美國白宮

查詢完成...請輸入新查詢

- 伊拉克
  - 台灣
  - 歐洲
- 美國
  - 布希
    - 布希(4篇)
  - 紐約
  - 台灣
- 伊朗
  - 以色列
- 所有文章(23篇)

94/12/26 美軍轉移伊拉克政權1週年 布希：美軍駐守直

94/11/24 美警方布希總統農場四周逮捕抗議民眾

94/11/25 切割中華民國與中國？—從總統接見「中國工

94/11/15 布希抵達日本 展開一週亞洲之旅

Figure 3: The result of query-based local semantic structuring for a query of “White House of United States”

Table 1: *Evaluation results for title generation.*

Approaches	Precision	Recall	F1	Relevance	Readability
ST	0.0783	0.1513	0.1032	3.615	1.874
AKNN	0.1315	0.1074	0.1183	3.594	4.615
Proposed	0.2176	0.2119	0.2147	4.102	4.332

4. Performance Evaluation

The performance evaluation of some key technologies are briefly summarized here.

4.1. Performance Evaluation of Automatic Generation of Titles and Summaries

118 broadcast news stories recorded at Taipei were used as the test documents in the evaluation for title generation, compared to human-generated reference titles. The objective performance measures used were precision, recall, and F1 scores calculated from the number of identical terms in computer-generated and human-generated titles. In addition, five-level subjective human evaluation was also performed, where 5 is the best and 1 is the worst, with two different metrics, ”Relevance” calibrating how the titles are related to the documents, and ”Readability” indicating how the titles are readable. In subjective human evaluation, each subject was given the reference titles with reference scores for some reference documents. The results for the previously proposed statistical Translation (ST) approach [12], Adaptive K-nearest-neighbor (AKNN) approach [18] and the new approach proposed here are listed in table 1. It can be found that the proposed approach performs significantly better in all measures, except with slightly lower readability than AKNN.

The F-measure results of the proposed summarization approach using N-gram co-occurrence statistics (ROUGE-1,2,3) and the longest common sub-sequence (ROUGE-L) evaluated with ROUGE package [22] with respect to human-generated summaries are shown in table 2 for summarization ratios of 10% and 30%. Here listed are two different ways to perform the automatic summarization: the well known and very successful significance score [14, 15], and the approach proposed in this research respectively. It can be found that the proposed approach is better in all scores.

4.2. Performance Evaluation of Global Semantic Structuring

The performance evaluation for the global semantic structuring was performed on the TDT-3 Chinese broadcast news corpus [19].

Table 2: *Evaluation results for automatic summary generation.*

	ROUGE-1		ROUGE-2		ROUGE-3		ROUGE-L	
Summarization Ratio	10%	30%	10%	30%	10%	30%	10%	30%
Significance Score	0.27	0.48	0.18	0.40	0.16	0.36	0.26	0.47
Proposed	0.36	0.54	0.30	0.47	0.29	0.44	0.36	0.53

A total of 47 different topics have been manually defined, and each news story was assigned to one of the topics, or as "out of topic". These 47 classes of news stories with given topics were used as the reference for the evaluation. We define the "Between-class to within-class" distance ratio as in equation (3),

$$R = \bar{d}_B / \bar{d}_W, \tag{3}$$

where $\bar{d}_B$ is the average of the distances between the locations of the two clusters on the map for all pairs of news stories manually assigned to different topics, and $\bar{d}_w$ is the similar average, but over all pairs of news stories manually assigned to identical topics. So the ratio R in equation (3) tells how far away the news stories with different manually defined topics are separated on the map. Apparently, the higher values of R the better. On the other hand, for each news story d_i, the probability $P(T_k|d_i)$ for each latent topic $T_k, k = 1, 2, \ldots, K$, was given. Thus the total entropy for topic distribution for the whole document archive with respect to the organized topic clusters can be defined as:

$$H = -\sum_{i=1}^{N} \sum_{k=1}^{K} P(T_k|d_i) \log(P(T_k|d_i)), \tag{4}$$

where N is the total number of news stories used in the evaluation. Apparently, lower total entropy means the news stories have probability distributions more focused on less topics. Table 3 lists the results of the performance measure for the proposed approach as compared to the well-known approach of Self-Organized Map (SOM) [23] for different choices of the "term" t_j used, i.e., W(words), S(2)(segments of two syllables), C(2)(segments of two characters), and combinations.

Table 3: *Evaluation results for the global semantic structuring.*

	Choice of Terms	Distance Ratio (R)		Total Entropy (H)
		Proposed	SOM	
(a)	W	2.34	1.11	5135.62
(b)	S(2)	3.38	1.04	4637.71
(c)	C(2)	3.65	1.03	3489.21
(d)	S(2)+C(2)	3.78	1.02	4096.68

4.3. Performance Evaluation of Query-based Local Semantic Structuring

The performance evaluation for the query-based local semantic structuring was performed using 20 queries to generate 20 topic hierarchies [20]. The average values of correctness (C) and coverage ratio (CR) were obtained with some manual efforts. The correctness (C) is the measure if all the key terms in the topic hierarchy is correctly located at the right node position. It can be evaluated by counting the number of key terms in the topic hierarchy which have to be moved manually to the right node position to produce a completely correct topic hierarchy. The coverage ratio (CR) is the percentage of the retrieved news stories which can be covered by the key terms in the topic hierarchy. On average a correctness (C) of 91% and a coverage ratio (CR) of 97% were obtained.

5. Conclusion

In this paper we proposed a whole set of technologies to offer multi-layered summarization for entire spoken archives for the purposes of efficient browsing and retrieval. This includes (1) Automatic Generation of Titles and Summaries for each of the spoken documents, (2) Global Semantic Structuring of the spoken document archive and (3) Query-based Local Semantic Structuring for the subset of spoken documents relevant to user's query. An initial prototype system was completed, and satisfactory performance was obtained.

6. References

[1] L.-S. Lee and B. Chen, "Spoken document understanding and organization," *IEEE Signal Processing Magazine*, vol. 22, no. 5, Sept. 2005.

[2] CMU Informedia Digital Video Library project [online]. Available : http://www.informedia.cs.cmu.edu/.

[3] Multimedia Document Retrieval project at Cambridge University [online]. Available : http://mi.eng.cam.ac.uk/research/Projects/Multimedia_Document_Retrieval/.

[4] D.R.H Miller, T. Leek, and R. Schwartz, "Speech and language technologies for audio indexing and retrieval," *Proc. IEEE*, vol. 88, no. 8, pp. 1338–1353, 2000.

[5] S. Whittaker, J. Hirschberg, J. Choi, D. Hindle, F. Pereira, and A. Singhal, "Scan: Designing and evaluating user interface to support retrieval from speech archives," in *Proc. ACM SIGIR Conf. R&D in Information Retrieval*, 1999, pp. 26–33.

[6] A. Merlino and M. Maybury, "An empirical study of the optimal presentation of multimedia summaries of broadcast news," in *Automated Text Summarization*, I. Mani and M. Maybury, Eds., pp. 391–401. Eds. Cambridge, MA:MIT Press, 1999.

[7] SpeechBot Audio/Video Search at Hewlett-Packard (HP) Labs [online]. Available : http://www.speechbot.com/.

[8] S. Furui, "Recent advances in spontaneous speech recognition and understanding," *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 1–6, 2003.

[9] Columbia Newsblaster project at Columbia University [online]. Available : http://www1.cs.columbia.edu/nlp/newsblaster/.

[10] T. Hofmann, "Probabilistic latent semantic analysis," *Uncertainty in Artificial Intelligence*, 1999.

[11] R. Jin and A. Hauptmann, "Automatic title generation for spoken broadcase news," in *Proc. of HLT*, 2001, pp. 1–3.

[12] M. Banko, V. Mittal, and M. Witbrock, "Headline generation based on statistical translation," in *Proc. of ACL*, 2000, pp. 318–325.

[13] B. Dorr, D. Zajic, and R. Schwartz, "Hedge trimmer: A parse-and-trim approach to headline generation," in *Proc. of HLT-NAACL*, 2003, vol. 5, pp. 1–8.

[14] S. Furui, T. Kikuchi, Y. Shinnaka, and C. Hori, "Speech-to-text and speech-to-speech summarization of spontaneous speech," *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 4, pp. 401–408, 2004.

[15] M. Hirohata, Y. Shinnaka, K. Iwano, and S. Furui, "Sentence extraction-based presentation summarization techniques and evaluation metrics," in *Proc. ICASSP*, 2005, pp. SP–P16.14.

[16] S.-Y. Kong and L.-S. Lee, "Improved spoken document summarization using probabilistic latent semantic analysis (plsa)," in *Proc. ICASSP*, 2006, To Appear.

[17] C.-C. Wang, "Improved automatic generation of titles for spoken documents using various scoring techniques," M.S. thesis, National Taiwan Univerisity, 2006.

[18] S.-C. Chen and L.-S. Lee, "Automatic title generation for chinese spoken documents using an adaptive k-nearest-neighbor approach"," in *Proc. European Conf. Speech Communication and Technology*, 2003, pp. 2813–2816.

[19] T.-H. Li, M.-H. Lee, B. Chen, and L.-S. Lee, "Hierarchical topic organization and visual presentation of spoken documents using probabilistic latent semantic analysis (plsa) for efficient retrieval/browsing applications," in *Proc. European Conf. Speech Communication and Technology*, 2005, pp. 625–628.

[20] Y.-C. Pan, C.-C. Wang, Hsieh Y.-C., T.-H. Lee, Y.-S. Lee, Y.-S. Fu, Y.-T. Huang, and L.-S. Lee, "A multi-modal dialogue system for information navigation and retrieval across spoken document archives with topic hierarchies," in *Proc. of ASRU*, 2005, pp. 375–380.

[21] S.-L. Chuang and L.-F. Chien, "A pratical web-based approach to generating topic hierarchy for text segments"," in *ACM SIGIR*, 2004, pp. 127–136.

[22] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. of Workshop on Text Summarization Branches Out*, 2004, pp. 74–81.

[23] T. Kohonen, S. Kaski, K. Lagus, J. Salojvi, J. Honkela, V. Paatero, and A. Saarela, "Self organization of a massive document collection," *IEEE Trans on Neural Networks*, vol. 11, no. 3, pp. 574–585, 2000.