A Probabilistic Generative Framework for Extractive Broadcast News Speech Summarization

Yi-Ting Chen, Berlin Chen, *Member, IEEE*, and Hsin-Min Wang, *Senior Member, IEEE*

***Abstract*—In this paper, we consider extractive summarization of broadcast news speech and propose a unified probabilistic generative framework that combines the sentence generative probability and the sentence prior probability for sentence ranking. Each sentence of a spoken document to be summarized is treated as a probabilistic generative model for predicting the document. Two matching strategies, namely literal term matching and concept matching, are thoroughly investigated. We explore the use of the language model (LM) and the relevance model (RM) for literal term matching, while the sentence topical mixture model (STMM) and the word topical mixture model (WTMM) are used for concept matching. In addition, the lexical and prosodic features, as well as the relevance information of spoken sentences, are properly incorporated for the estimation of the sentence prior probability. An elegant feature of our proposed framework is that both the sentence generative probability and the sentence prior probability can be estimated in an unsupervised manner, without the need for handcrafted document-summary pairs. The experiments were performed on Chinese broadcast news collected in Taiwan, and very encouraging results were obtained.**

***Index Terms*—Extractive spoken document summarization, probabilistic generative framework, language model (LM), relevance model (RM), topical mixture model.**

I. Introduction

HUGE quantities of audio–visual content continue to grow and fill our computers, networks, and daily lives. It is obvious that speech is one of the most important sources of information about this content. Therefore, how to access audio–visual content based on associated spoken documents has become an active focus of much research in recent years [1], [2]. Spoken documents are often automatically transcribed into words; however, incorrect speech recognition results (such as recognition errors and inaccurate sentence or paragraph boundaries) and redundant acoustic effects (generated by disfluencies, fillers, and repetitions) prevent documents from being accessed easily. Spoken document summarization, which tries to distill important information and remove redundant and incorrect information from spoken documents, can help users review documents efficiently and understand associated topics quickly.

Automatic summarization of text documents dates back to the early 1950s. Nowadays, the research is extended to cover a wider range of tasks, including multidocument, multilingual, and multimedia summarization [3]. Broadly speaking, summarization can be either extractive or abstractive. Extractive summarization selects indicative sentences, passages, or paragraphs from an original document according to a target summarization ratio and concatenates them to form a summary. Abstractive summarization, on the other hand, produces a concise abstract of a certain length that reflects the key concepts of the document [4], [5]. The latter is more difficult to achieve; thus, in recent years, research has focused on the former. Summarization can also be either generic or query-based. A generic summary highlights the most salient information in a document, whereas a query-based summary presents the information in a document that is most relevant to the user's query.

The wide variety of extractive summarization approaches that have been developed and applied to spoken document summarization can in general be classified into three categories: 1) approaches based on sentence structure or location information; 2) approaches based on proximity or significance measures; and 3) approaches based on sentence classification. In [6] and [7], the authors suggested that important sentences can be selected from the significant parts of a document, e.g., the introduction and conclusion. However, such approaches can only be applied to documents in some specific domains or documents that have some specific structures. In contrast, approaches based on proximity or significance measures [3] attempt to select salient sentences based on the statistical features of the sentences or the words in the sentences, such as the term frequency (TF), inverse document frequency (IDF), N-gram scores, and the topic or concept information. Associated methods based on these features have attracted much attention in recent years. For example, the vector space model (VSM) and the maximum marginal relevance (MMR) method [8] represent the whole document and each of its sentences in vector form consisting of statistical features, and then select important sentences based on the proximity measure between the vector representations of the document and its sentences; the latent semantic analysis (LSA) method [9] estimates the significance of a sentence by projecting the vector representation of the sentence into the latent semantic space of the document; and the sentence significance score method (SIG) [10], [11] estimates the significance

Manuscript received November 15, 2007; revised July 20, 2008. Current version published December 11, 2008. This work was supported in part by the National Science Council of Taiwan under Grants NSC95-2221-E-003-014-MY3 and NSC95-2422-H-001-031. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ruhi Sarikaya.

Y. T. Chen was with the Department of Computer Science and Information Engineering, National Taiwan Normal University, Taipei, Taiwan. She is now with the Institute of Information Science, Academia Sinica, Taipei, Taiwan (e-mail: g93470070@csie.ntnu.edu.tw).

B. Chen is with the Department of Computer Science and Information Engineering, National Taiwan Normal University, Taipei, Taiwan (e-mail: berlin@csie.ntnu.edu.tw).

H.-M. Wang is with the Institute of Information Science, Academia Sinica, Taipei, Taiwan (e-mail: whm@iis.sinica.edu.tw).

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.2008.2005031

of a sentence by linearly combining a set of statistical features of the sentence. In addition, a number of classification-based methods that use statistical features and/or sentence structure information have also been developed, including the Gaussian mixture model (GMM) [9], the hidden Markov model (HMM) [12], the Bayesian classifier [13], the support vector machine (SVM) [14], the conditional random fields (CRF) method [15], and the logistic regression model [16]. Under these methods, sentence selection is usually formulated as a binary classification problem; that is, a sentence can be included in a summary or omitted. These methods, however, need a training set comprised of documents and corresponding handcrafted summaries (or labeled data) for training the classifiers. In recent years, there has also been some research on exploring extra information clues (e.g., word-clusters, WordNet, or event relevance) [17]–[19] and novel ranking algorithms [20] for extractive text document summarization. Interested readers can refer to [3] for a comprehensive overview of the principal trends and the classical approaches for text summarization.

Although the above approaches can be applied to both text and spoken documents, the latter presents unique difficulties, such as recognition errors, problems with spontaneous speech, and the lack of correct sentence or paragraph boundaries. To avoid redundant or incorrect content while selecting important and correct information, multiple recognition hypotheses, confidence scores, language model scores, and other grammatical knowledge have been utilized [10], [11]. In addition, prosodic features (e.g., intonation, pitch, energy, and pause duration) can provide important clues for summarization; although reliable and efficient ways to use these prosodic features are still under active research [21], [22]. Summaries of spoken documents can be presented in either text or speech format. The former has the advantage of easier browsing and further processing, but it is subject to speech recognition errors, as well as the loss of the speaker's emotional/prosodic information, which can only be conveyed by speech signals.

In this paper, we consider generic, extractive summarization of Chinese broadcast news speech. A unified probabilistic generative framework that combines the sentence generative probability and the sentence prior probability for sentence ranking is proposed [23]–[27]. The sentence generative probability can be taken as a relevance measure between a document and a given sentence of the document, while the sentence prior probability is a measure of the importance of the sentence itself. A remarkable feature of our proposed framework is that both the sentence generative probability and the sentence prior probability can be estimated in an unsupervised manner, without the need for handcrafted document-summary pairs. Various kinds of modeling structures and summarization features are investigated as well. The performance of our proposed models is verified by comparison with a number of existing summarization models.

The remainder of this paper is organized as follows. In Section II, we elucidate our proposed probabilistic generative framework, which can leverage various kinds of sentence generative models and sentence prior probabilities for extractive spoken document summarization. The experiment setup and a series of spoken document summarization experiments are presented in Sections III and IV, respectively. We then present our conclusions in Section V.

II. Spoken Document Summarization

In this section, we begin by introducing the proposed probabilistic generative framework for extractive spoken document summarization and then discuss the structural characteristics of various sentence generative models and the features used for modeling the sentence prior probability.

A. Probabilistic Generative Framework

In the probabilistic generative framework, the importance of a sentence S in a document D to be summarized can be modeled by $P(S \mid D)$, i.e., the posterior probability of the sentence S given the document D. According to Bayes' rule, $P(S \mid D)$ can be expressed as [28]

$$P(S \mid D) = \frac{P(D \mid S)P(S)}{P(D)} \tag{1}$$

where $P(D \mid S)$ is the sentence generative probability, i.e., the likelihood of D being generated by S, $P(S)$ is the prior probability of S being important, and $P(D)$ is the prior probability of D. Note that $P(D)$, in (1), can be omitted because it is identical for all sentences and will not affect their ranking. The sentence generative probability $P(D \mid S)$ can be taken as a relevance measure between the document D and the sentence S, while the sentence prior probability $P(S)$ is, to some extent, a measure of the importance of the sentence itself. Therefore, all the sentences of the spoken document D can be ranked according to the product of the sentence generative probability $P(D \mid S)$ and the sentence prior probability $P(S)$. Then, the sentences with the highest probabilities are selected and sequenced to form a summary. Fig. 1 illustrates extractive spoken document summarization using the probabilistic generative framework.

B. Sentence Generative Model

1) LM-Based Sentence Generative Model: An LM can be applied in extractive spoken document summarization, where each sentence S of a document D to be summarized is treated as a probabilistic generative model comprised of N-gram distributions for predicting the document D; and the words (or terms) in D are taken as an input observation sequence. When only the unigrams are considered, the probability of the document D given the sentence S is expressed as [24]

$$P_{\mathrm{LM}}(D \mid S) = \prod_{w \in D} [\lambda \cdot P(w \mid S) + (1 - \lambda) \cdot P(w \mid C)]^{n(w,D)} \tag{2}$$

where λ is a weighting parameter and $n(w, D)$ is the occurrence count of the word w in D. The sentence model $P(w \mid S)$ and the collection model $P(w \mid C)$ are estimated, respectively, from the sentence S itself and a large external text collection C using the maximum-likelihood estimation (MLE) method [28]. The weighting parameter λ can be empirically tuned by using

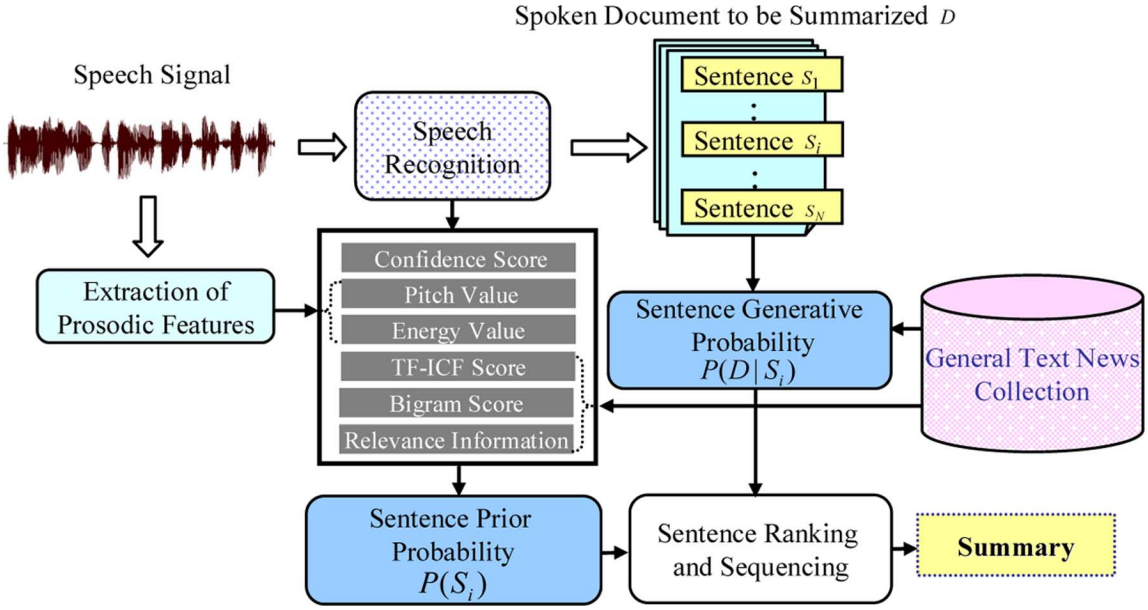

Fig. 1. Extractive spoken document summarization using the probabilistic generative framework.

a development data set, or optimized by applying the expectation-maximization (EM) training algorithm [29] to a training data set. Note that this relevance measure is computed according to the frequency that document words occur in the sentence, which is actually a form of literal term matching [1].

In the LM model defined in (2), the sentence model $P(w|S)$ is linearly interpolated with the collection model $P(w|C)$ such that there is some probability of generating every word in the vocabulary. However, the true sentence model $P(w|S)$ might not be accurately estimated by MLE, since the sentence only consists of a few words, and the occurrences of the words in the sentence are not in proportion to the probabilities of the words in the true model. Therefore, we employ the relevance model (RM) [30] to obtain a more accurate estimation of the sentence model. In the extractive spoken document summarization task, each sentence S of a document D to be summarized has its own associated relevant class R_s, which is defined as the subset of documents in the collection that are relevant to S. The relevance model of S is defined as the probability distribution $P(w|R_s)$, which gives the probability that we would observe a word w if we were to randomly select a document from the relevant class R_s and select a word from that document. After the relevance model of S has been constructed, it can be used to replace the original sentence model or it can be combined linearly with the original sentence model. Because we do not have prior knowledge about the subset of relevant documents for each spoken sentence S, we employ a local feedback-like procedure [24], [31] that takes S as a query and poses it to the information retrieval (IR) system to obtain a ranked list of documents. It is assumed that the top L documents returned by the IR system are relevant to S, and the relevance model $P(w|R_s)$ of S can be constructed by the following equation:

$$P(w|R_s) = \sum_{D_l \in \mathbf{D}_{\text{top}L}} P(D_l|S) \cdot P(w|D_l) \tag{3}$$

where $\mathbf{D}_{\text{top}L}$ is the set of top L retrieved documents, and the probability $P(D_l|S)$ can be approximated by the following equation using Bayes' rule:

$$P(D_l|S) \approx \frac{P(D_l) \cdot P(S|D_l)}{\sum_{D_u \in \mathbf{D}_{\text{top}L}} P(D_u) \cdot P(S|D_u)}. \tag{4}$$

A uniform prior probability $P(D_l)$ can be assumed for the top L retrieved documents, and the sentence likelihood $P(S|D_l)$ can be calculated using an equation similar to (2) if the IR system is implemented with the LM retrieval model [30], [32]. The relevance model $P(w|R_s)$ can then be combined linearly with the original sentence model $P(w|S)$ to form a more accurate sentence model

$$\hat{P}(w|S) = \alpha \cdot P(w|S) + (1-\alpha) \cdot P(w|R_s) \tag{5}$$

where α is a weighting parameter. The final sentence generative model (denoted as LM-RM) is thus expressed as

$$P_{\text{LM-RM}}(D|S) = \prod_{w \in D} [\lambda \cdot \hat{P}(w|S) + (1-\lambda) \cdot P(w|C)]^{n(w,D)}. \tag{6}$$

We can also use the retrieved relevant text document set to retrain the LM model directly. Since the relevant text documents retrieved for a given spoken sentence are statistically relevant to the spoken document that the spoken sentence belongs to, they might be used as the training data, instead of the spoken document, to obtain a more reliable parameter estimation of the LM model of the spoken sentence. For example, the weighting

parameter λ in (6) can be re-estimated with the retrieved relevant text document set $\mathbf{D}_{\text{top}L}$, using the following EM updating equation:

$$\hat{\lambda}=\frac{\sum_{D_l\in\mathbf{D}_{\text{top}L}}\sum_{w\in D_l} n(w,D_l)\cdot\frac{\lambda\cdot\hat{P}(w\,|\,S)}{\lambda\cdot\hat{P}(w\,|\,S)+(1-\lambda)\cdot P(w\,|\,C)}}{\sum_{D_u\in\mathbf{D}_{\text{top}L}}\sum_{w'\in D_u} n(w',D_u)}. \tag{7}$$

We denote this model as LM-RT.

2) STMM-Based Sentence Generative Model: Each sentence S of a spoken document D to be summarized can be also interpreted as a probabilistic sentence topical mixture model (STMM). In this model, a set of K latent topical distributions characterized by unigram language models are used to predict the words in the document, and each of the latent topics is associated with a sentence-specific weight [25]. That is, each sentence can belong to many topics. The probability of the document D given the sentence S is expressed as

$$P_{\text{STMM}}(D\,|\,S)=\prod_{w\in D}\left[\sum_{k=1}^{K}P(w\,|\,T_k)P(T_k\,|\,S)\right]^{n(w,D)} \tag{8}$$

where $P(w\,|\,T_k)$ and $P(T_k\,|\,S)$ denote, respectively, the probability of the word w occurring in a specific latent topic T_k and the posterior probability (or weight) of topic T_k conditioned on the sentence S. More precisely, the topical unigram distributions, $P(w\,|\,T_k), k=1,\ldots,K$, are the same for all sentences, but each sentence S has its own probability distributions over the latent topics, i.e., $P(T_k\,|\,S), k=1,\ldots,K$. Note that this relevance measure is not computed directly according to the frequency that the document words occur in the sentence. Instead, it is derived from the frequency of the document words in the latent topics as well as the likelihood that the sentence will generate the respective topics. Hence, STMM is actually a type of concept matching approach [1]. Structures similar to the presented topical mixture model have also been extensively investigated for IR tasks in recent years [33]–[35].

During training, a set of contemporary (or in-domain) text news documents $\mathbf{D}$ with corresponding human-generated titles (a title can be viewed as an extremely short summary of a document) can be collected to train the latent topical distributions $P(w\,|\,T_k)$ of the STMM model. For each document D_j of the text news collection $\mathbf{D}$, we treat the human-generated title H_j of D_j as an STMM model for generating D_j as follows:

$$P_{\text{STMM}}(D_j\,|\,H_j)=\prod_{w\in D_j}\left[\sum_{k=1}^{K}P(w\,|\,T_k)P(T_k\,|\,H_j)\right]^{n(w,D_j)}. \tag{9}$$

First, the K-means algorithm is used to partition all the titles of the document collection into K topical clusters in an unsupervised manner, after which the initial topical unigram distribution $P(w\,|\,T_k)$ for a cluster topic T_k is estimated according to the underlying statistical characteristics of the document titles assigned to it. In addition, the probability that each title will generate the topics, i.e., $P(T_k\,|\,H_j)$, is measured according to its proximity to the centroid of each respective cluster. Then, using the EM algorithm, the probability distributions $P(w\,|\,T_k)$ and $P(T_k\,|\,H_j)$ can be optimized by maximizing the total log-likelihood L_T of all the documents D_j in the collection $\mathbf{D}$ generated by their individual titles

$$L_T=\sum_{D_j\in\mathbf{D}}\log P_{\text{STMM}}(D_j\,|\,H_j). \tag{10}$$

We postulate that latent topical factors $P(w\,|\,T_k)$ properly constructed based on "document-title" relationships might provide very helpful clues for the subsequent spoken document summarization task. When performing extractive summarization of a broadcast news document D, we can apply the latent topical factors $P(w\,|\,T_k)$ trained in this way in (8), but use the EM algorithm to estimate the posterior probabilities, $P(T_k\,|\,S), k=1,\ldots,K$, on the fly by maximizing the log-likelihood of the document D generated by the STMM model. A detailed account of the process can be found in [25] and [35].

In most practical applications, the contemporary or in-domain text news documents used by spoken document summarization systems are not usually accompanied by "document-title" pairs for model training. Therefore, we also investigate the use of unsupervised training for STMM by exploiting all the sentences of the spoken (broadcast news) documents in the development set to construct the latent topical space [25]. That is, each sentence of a spoken document in the development set, regardless of whether it belongs to the reference summary or not, is treated as an STMM model and included in the construction of the latent topical distributions $P(w\,|\,T_k)$. Meanwhile, the probability distributions of the STMM models over the latent topics $P(T_k\,|\,S)$ are estimated on the fly during the summarization process. We denote this model as STMM-U.

3) WTMM-Based Sentence Generative Model: We also explore an alternative concept matching strategy, called the word topical mixture model (WTMM) [26], [36], to represent the sentence generative probability. Each word w_j of the language is treated as a WTMM M_{w_j} for predicting the occurrence of another word w

$$P(w\,|\,M_{w_j})=\sum_{k=1}^{K}P(w\,|\,T_k)P(T_k\,|\,M_{w_j}) \tag{11}$$

where $P(w\,|\,T_k)$ and $P(T_k\,|\,M_{w_j})$ are, respectively, the probability of a word w occurring in a specific latent topic T_k and the probability of a topic T_k conditioned on M_{w_j}. During the summarization process, we can linearly combine the associated WTMM models of those words involved in a sentence S to form a composite WTMM model of S. Then, the likelihood of the document D being generated by S can be expressed as

$$P_{\text{WTMM}}(D\,|\,S)=\prod_{w\in D}\left[\sum_{w_j\in S}\alpha_j\sum_{k=1}^{K}P(w\,|\,T_k)P(T_k\,|\,M_{w_j})\right]^{n(w,D)} \tag{12}$$

TABLE I
FEATURES EXPLOITED FOR MODELING THE SENTENCE PRIORITY PROBABILITY

Lexical Features	average TF-ICF score of words in a spoken sentence (F1) average bi-gram scores of word pairs in a spoken sentence (F2)
Prosodic Features	average pitch value of words in a spoken sentence (F3) average energy value of words in a spoken sentence (F4) maximum energy of words in a spoken sentence (F5)
Confidence Feature	average posterior probability of words in a spoken sentence (F6)
Relevance Feature	average similarity among the retrieved text documents for a spoken sentence (F7)

where the weighting parameter α_j is set in proportion to the frequency that w_j occurs in S, subject to $\sum_{w_j \in S} \alpha_j = 1$. In this paper, we investigate an unsupervised approach for training WTMM models. Each WTMM M_{w_j} of word w_j is trained by concatenating the words that occur within a context window of size m around each occurrence of w_j in the contemporary text news document collection. We postulate that these contextual words are relevant to w_j, and can therefore be used as an observation for training M_{w_j}. Interested readers may refer to [26] and [36] for details of the derivation of WTMM training using the EM algorithm.

C. Sentence Prior Probability

In the probabilistic generative framework for extractive spoken document summarization, the sentence prior probability in (1) can be regarded as the likelihood of a sentence being important in the document. Because the way to estimate the prior probability of a sentence is still an open issue, it is usually assumed uniformly distributed [24]–[26]. However, the sentences in a spoken document should not be considered equally important. In fact, a sentence's importance may depend on a wide variety of factors, such as the structural (positional and lexical) information, recognition accuracy, and inherent prosodic properties. Therefore, in this paper, we attempt to model the sentence prior probability (or importance) based on lexical, prosodic, and confidence features extracted from a spoken sentence. These features are presented in Table I. The TF-ICF score is similar to the conventional TF-IDF measure widely used in IR systems, but the value of inverse collection frequency (ICF) is calculated by [11]

$$\mathrm{ICF} = \log \frac{F_A}{F_w} \tag{13}$$

where F_w is the occurrence count of a word w in a large contemporary text news corpus, and F_A is the number of words in the corpus. In addition, the prosodic features are extracted from the broadcast news speech by using the Snack toolkit [37] and the methods described in [38]. The measure or score of each feature in Table I is normalized such that it can be taken as the sentence prior probability that satisfies $\sum_{S_i \in D} P(S_i) = 1$. Some of these features are used to calculate the sentence significance scores in [10] and [11], and included in the feature sets of the classification-based models in [9], [12], and [15], for spoken document summarization.

Nevertheless, the sentence prior probability might not be accurately estimated by the above-mentioned features, since the automatic transcript of a spoken document to be summarized usually contains recognition errors, incorrect boundaries, and redundant information. Hence, we also try to model the sentence prior probability by calculating the average similarity of documents in the relevant text document set $\mathbf{D}_{\mathrm{top}L}$ [27]. The documents are retrieved by the local feedback-like procedure for each spoken sentence S described in Section II-B. Our assumption is that the relevant text documents retrieved for a summary sentence might have the same or similar topics because a summary sentence is usually indicative for some specific topic related to the document. In contrast, the relevant text documents retrieved for a nonsummary sentence might cover diverse topics. Therefore, the relevance information estimated based on the similarity of documents in the relevant text document set $\mathbf{D}_{\mathrm{top}L}$ might be a good indicator for determining the importance of a spoken sentence. Consequently, the sentence prior probability can be approximated by using the sentence's relevance information as follows:

$$P(S) = \frac{\mathrm{avgSim}(S)}{\sum_{S_j \in D} \mathrm{avgSim}(S_j)} \tag{14}$$

where $\mathrm{avgSim}(S)$ is the average similarity of documents in the relevant text document set $\mathbf{D}_{\mathrm{top}L}$ for a spoken sentence S computed by

$$\mathrm{avgSim}(S) = \frac{\sum_{D_l \in \mathbf{D}_{\mathrm{top}L}} \sum_{\substack{D_u \in \mathbf{D}_{\mathrm{top}L} \\ D_l \neq D_u}} \frac{\vec{D}_l \cdot \vec{D}_u}{\|\vec{D}_l\| \cdot \|\vec{D}_u\|}}{L \cdot (L-1)} \tag{15}$$

where $\vec{D}_l$ is the TF-IDF vector representation of the document D_l, and L is the number of documents in the retrieved relevant text document set $\mathbf{D}_{\mathrm{top}L}$.

Once the sentence generative model and the sentence prior probability have been properly estimated, the sentences of the spoken document D to be summarized can be ranked by the product of the sentence generative probability and the sentence prior probability. The sentences with the highest probabilities are then selected and sequenced to form the final summary according to different summarization ratios.

III. EXPERIMENT SETUP

A. Speech and Text Corpora

The speech corpus was comprised of approximately 110 h of radio and TV broadcast news documents collected from several radio and TV stations in Taipei between 1998 and 2004 [39], [40]. From this corpus, a subset of 200 documents (1.6 h) collected in August 2001 was reserved for the summarization experiments [1] and divided into two equal parts. The first part was

TABLE II
SUMMARY FOR THE SPEECH CORPUS USED IN THIS PAPER

Total broadcast news speech		109.9 hours
Speech for the summarization experiments		1.6 hours
Speech for acoustic model training	With orthographic transcripts	4.0 hours
	Without orthographic transcripts	104.3 hours

TABLE III
DETAILED STATISTICS OF THE BROADCAST NEWS DOCUMENTS FOR THE SUMMARIZATION EXPERIMENTS

	Development Set	Evaluation Set
Recording period	August 1 – August 15, 2001	August 16 – August 31, 2001
Number of documents	100	100
Average duration per document (in sec.)	29.8	26.7
Average number of words per document	87	77
Average number of Chinese characters per document	171	153
Average number of sentences per document	10	9

taken as the development set, which formed the basis for tuning the parameters or settings. The second part was taken as the evaluation set; i.e., all the summarization experiments conducted on it followed the same training (or parameter) settings and model complexities, which were optimized based on the development set. Therefore, the experiment results can validate the effectiveness of the proposed approaches on comparable real-world data.

The remainder of the speech data was used to train the acoustic models for speech recognition, of which about 4.0 h of data with corresponding orthographic transcripts was used to bootstrap the acoustic model training. Meanwhile, 104.3 h of the remaining untranscribed speech data was reserved for unsupervised acoustic model training [41]. The acoustic models were further optimized by the minimum phone error (MPE) training algorithm [42].

A summary for the speech corpus used in this paper is shown in Table II, while the detailed statistics of the 200 broadcast news documents for the summarization experiments are given in Table III. It is worth mentioning that, for a spoken document to be summarized, we used the corresponding best scoring sequence of words (the one-best result) generated by the speech recognizer in the summarization experiments, while the number of sentences in the spoken document was simply determined based on the pause information provided by the speech recognizer (a pause with duration of more than 0.5 s was regarded as a sentence boundary). Though we believe that a more sophisticated sentence boundary detection algorithm using either prosodic or lexical information will be helpful for the summarization task, we do not have one at the moment.

We also used a large number of text news documents collected from the Central News Agency (CNA) between 1991 and 2002 (the Chinese Gigaword Corpus published by LDC) [43]. The text news documents collected in 2000 and 2001 were used to train N-gram language models for speech recognition with the SRI Language Modeling Toolkit [44]. The Chinese character error rate (CER) for the 200 broadcast news documents to be summarized was 14.17%.

A subset of approximately 14 000 text news documents collected in the same period as the broadcast news documents to be summarized (August 2001) was used to estimate the collection model $P(w \mid C)$ in (2), (6), and (7) for LM, LM-RM, and LM-RT and the latent topical distributions $P(w \mid T_k)$ in (8) and (12) for STMM and WTMM. It was also used to construct the relevant text document set for each spoken sentence (discussed in Sections II-B and C), and as the basis to estimate the model parameters for VSM, LSA, MMR, and SIG (see Section III-C).

B. Evaluation Metric

Three subjects were asked to create manual summaries of the 200 broadcast news documents as references for evaluation. The summaries were compiled by selecting 50% of the most important sentences in the reference transcript of a spoken (broadcast news) document and ranking them by importance without assigning a score to each sentence. The summarization results were tested by using several summarization ratios (10%, 20%, 30%, and 50%), defined as the ratio of the number of sentences in the automatic (or manual) summary to that in the reference transcript of a spoken document [1].

We used the ROUGE package (Version 1.5.5) [45] to evaluate the performance levels of the proposed models. The ROUGE measure evaluates the quality of the summarization by counting the number of overlapping units, such as N-grams and word sequences, between the automatic summary and a set of manual summaries. ROUGE–N is an N-gram recall measure, defined as follows:

$$\text{ROUGE–}N = \frac{\sum_{M \in \mathbf{M}_R} \sum_{\text{gram}_N \in M} \text{Count}_{\text{match}}(\text{gram}_N)}{\sum_{M \in \mathbf{M}_R} \sum_{\text{gram}_N \in M} \text{Count}(\text{gram}_N)} \quad (16)$$

where N denotes the length of the N-gram, M is an individual manual summary, $\mathbf{M}_R$ is a set of manual summaries, $\text{Count}_{\text{match}}(\text{gram}_N)$ is the maximum number of N-grams co-occurring in the automatic summary and the manual summary, and $\text{Count}(\text{gram}_N)$ is the number of N-grams in the manual summary. Since ROUGE-N is a recall measure, increasing the summary length (or the summarization ratio) tends to increase the chances of getting higher scores. In this paper, we mainly adopt the widely used ROUGE-2 measure [9], [21], which uses word bigrams as the matching units. The levels of agreement on the ROUGE-2 measure between the three subjects for important sentence ranking are about 0.53,

0.56, 0.61, and 0.68 for the summarization ratios of 10%, 20%, 30%, and 50%, respectively. In the last set of experiments, we will evaluate the best two summarization approaches using the ROUGE-1, ROUGE-2, ROUGE-3, and ROUGE-4 measures.

C. Conventional Summarization Models

We compare our proposed models with the following conventional summarization methods, which are commonly used for the spoken document summarization task: VSM, MMR, LSA, DIM, SIG, and SVM. Among them, VSM, MMR, LSA, DIM, and SIG are unsupervised models, while SVM is a supervised model. VSM is a typical literal term matching approach, and LSA is a typical concept matching approach [1].

VSM represents each sentence of a document and the whole document in vector form [1]. In this approach, each dimension specifies the weighted statistics, e.g., the product of the TF and IDF scores, associated with an index term (or word) in the sentence (or document). The sentences with the highest relevance scores (i.e., the cosine measure of two vectors) to the whole document are included in the summary. MMR is actually closely related to VSM [8] because it also represents each sentence of a document and the document itself in vector form and uses the cosine score for sentence selection. However, MMR performs sentence selection iteratively based on the criteria of topic relevance and coverage. The sentence is selected according to two criteria: 1) whether it is more similar to the whole document than the other sentences, and 2) whether it is less similar to the set of sentences $\mathbf{S}_l$ selected so far than the other sentences by the following formula:

$$\text{NextSen} = \max_{S_u}[\beta \cdot \text{sim}(S_u, D) - (1-\beta) \max_{S_j \in \mathbf{S}_l} \text{sim}(S_u, S_j)] \tag{17}$$

where β is a weighting parameter used to make a tradeoff between relevance and redundancy [8]. We set the parameter β at 0.6 in this study. Consequently, MMR not only selects relevant sentences for the summary, but also allows the summary to cover more topics (or concepts).

LSA, on the other hand, represents each sentence of a document as a vector in the latent semantic space of the document, which is constructed by performing singular value decomposition (SVD) on the "word-sentence" matrix of the document. The right singular vectors with larger singular values represent the dimensions of the more important latent semantic concepts in the document. Therefore, the sentences with the largest index values in each of the top L right singular vectors are considered as significant sentences and included in the summary [9]. DIM is an alternative LSA-based approach [7], [11] that computes the importance score of each sentence based on the norm of its vector representation in the lower m-dimensional latent semantic space; then, a fixed number of sentences with relatively large scores are selected to form the summary. The value of m is set at 1 because $m = 1$ yielded the best performance in the experiments on the development set. This result conforms with the results reported in [11].

SIG selects indicative sentences from a spoken document based on the lexical, grammar, and confidence scores [11]. For example, given a sentence $S = \{w_1, w_2, \ldots w_n, \ldots w_{N_S}\}$ of length N_S, the significance score of S can be expressed as

$$\text{Sig}(S) = \sum_{n=1}^{N_S} [\lambda_I I(w_n) + \lambda_L L(w_n) + \lambda_C C(w_n)] \tag{18}$$

where $I(w_n)$ is the product of the TF and ICF scores of a word w_n, $L(w_n)$ is the logarithmic bigram probability of w_n given its predecessor word w_{n-1} in S, which is estimated from a large contemporary text corpus; $C(w_n)$ is the confidence score of w_n, and λ_I, λ_L, and λ_C are weighting parameters for balancing these scores.

SVM is one of the representative supervised methods that are widely used in various text summarization tasks [14], [15]. The SVM summarizer is trained with the 100 document-summary pairs of the development set, using the three sets of features presented in Table I (excluding the confidence feature) and an additional set of prosodic features, such as the pitch variance, energy variance, pitch entropy, and energy entropy in the sentence. Note that the SVM summarizer trained with the manual summaries at a given summarization ratio is tested at the same summarization ratio. In this study, we implemented SVM with the SSVM Toolbox [46].

TABLE IV
RESULTS ACHIEVED BY DIFFERENT SENTENCE GENERATIVE MODELS, USING A UNIFORM SENTENCE PRIOR PROBABILITY

	LM	LM-RM	LM-RT	STMM	STMM-U	WTMM
10%	0.2932	0.3182	0.3316	0.3210	0.3016	0.3248
20%	0.3191	0.3264	0.3412	0.3333	0.3217	0.3324
30%	0.3705	0.3671	0.3739	0.3741	0.3618	0.3816
50%	0.4732	0.4774	0.4880	0.4605	0.4683	0.4581

IV. EXPERIMENT RESULTS

A. Experiment Results of the Sentence Generative Models

First, we evaluate the summarization performance of the proposed sentence generative models (LM, STMM, and WTMM) on the evaluation set. For the experiments in this section, the sentence prior probability $P(S)$ was assumed to be uniform, whereas a detailed account on the impact of using the non-uniform sentence prior probability will be given in Section IV-B. For the LM model, we use the relevant contemporary text document set $\mathbf{D}_{\text{top}L}$ retrieved for a spoken sentence by the local feedback-like procedure to construct its corresponding LM-RM and LM-RT models. L was set at 5 in the experiments. Moreover, we use the complete set of contemporary text news documents with corresponding human-generated titles to construct the STMM model, and use the development set to construct the STMM-U model (cf. Section II-B).

The summarization results of these models at different summarization ratios are shown in Table IV. It should be noted again that, since ROUGE-2 is a recall measure, increasing the summarization ratio tends to increase the chances of getting higher scores. From the table, we observe that the performance of STMM is generally better than that of STMM-U. This reveals that the document-title correspondence information in the

contemporary text news document set does provide good guidance on the construction of the latent topical distributions in the STMM model. We also observe that STMM compares quite well with WTMM; however, WTMM slightly outperforms STMM when the summarization ratio is 10%. One possible explanation is that WTMM directly models the relationship between words, and more training observations are available for model estimation in an offline manner; whereas STMM needs to update its weights over the latent topics (i.e., $P(T_k \mid S)$) on the fly during the summarization process, which might not be accurately estimated since a sentence of a broadcast news document usually only consists of a few words. Both STMM and WTMM clearly outperform LM at lower summarization ratios (10% and 20%). Interestingly, the opposite result is obtained when the LM model is combined with the relevance model estimated by the retrieved relevant text document set (i.e., LM-RM), or retrained by using the retrieved relevant text document set (i.e., LM-RT) directly. In most cases, the results of LM-RT are obviously better than those of STMM and WTMM, while the results of LM-RM are only slightly less accurate than those of STMM and WTMM. These findings show that the relevance information provided by the local feedback-like procedure can, to some extent, enhance the estimation of the parameters in the sentence generative models LM-RM and LM-RT. If we look into the training of LM-RM and LM-RT using the retrieved relevant text documents, it can be found that, for LM-RM, only the sentence model $P(w \mid S)$ is updated, while the weighting parameter λ is set at a fixed value for all sentences. In contrast, for LM-RT, both the sentence model $P(w \mid S)$ and the weighting parameter λ are updated. This might explain why LM-RT outperforms LM-RM. In brief, the best results achieved by using LM-RT with a uniform sentence prior probability across all spoken sentences are approximately 0.33, 0.34, 0.37, and 0.49 for summarization ratios of 10%, 20%, 30%, and 50%, respectively.

B. Nonuniform Sentence Prior Probability

As mentioned in Section II-C, the importance (or prior probability) of the sentences of a spoken document to be summarized may not be identical. Therefore, we try to model the sentence prior probability $P(S)$ by using different features, listed in Table I, which are extracted from the sentences. The measure or score of each feature must be normalized such that it can be taken as the sentence prior probability that satisfies $\sum_{S_i \in D} P(S_i) = 1$. The LM-RT and WTMM models are integrated with the sentence prior probabilities derived by different features because they achieved the best performance, as shown by the results in Table IV; they can also be regarded as representative methods for literal term matching and concept matching, respectively. The experiment results derived by LM-RT and WTMM, with the sentence prior probability modeled by using different features, are shown in Tables V and VI, respectively. Comparing these results with those in Table IV, we observe that the performance of both models at lower summarization ratios (10% and 20%) is significantly improved by incorporating the sentence prior probability, estimated according to F7, into the sentence ranking. F7 is the relevance feature, i.e., the average similarity among the top L retrieved text documents for a spoken sentence. L was set at 5 in the experiments. Table VII presents the average of the average similarity among the retrieved relevant text documents for the manual summary and nonsummary sentences of the evaluation set at different summarization ratios. It is observed that the retrieved relevant text documents for a summary sentence of a spoken document have a higher similarity than the retrieved relevant text documents for a nonsummary sentence, and the difference becomes smaller as the summarization ratio increases. These observations explain why incorporating the sentence prior probability derived by the relevance feature (F7) can boost the performance of both LM-RT and WTMM at lower summarization ratios. Moreover, as shown in Tables V and VI, in most cases, incorporating the prior probability estimated by either F1 (the average TF-IDF score of words in a spoken sentence) or F3 (the average pitch value of the words in a spoken sentence) can also improve the performance of both models considerably, though the improvements are not as significant as that yielded by incorporating the prior probability estimated by F7. The best results achieved by literal term matching (using LM-RT and F7) are approximately 0.36, 0.37, 0.39, and 0.49 for summarization ratios of 10%, 20%, 30%, and 50%, respectively, while the best results achieved by concept matching (using WTMM and F7) are approximately 0.38, 0.38, 0.38, and 0.46 for the same summarization ratios.

We also attempt to fuse several useful features (specifically, F1, F3, and F7) through a simple linear combination to obtain a better estimation of the sentence prior probability. The summarization results achieved by LM-RT and WTMM with different combinations of these features are shown in Tables VIII and

TABLE V
RESULTS ACHIEVED BY LM-RT, WITH THE SENTENCE PRIOR PROBABILITY MODELED BY USING DIFFERENT FEATURES

	F1	F2	F3	F4	F5	F6	F7
10%	0.3394	0.3408	0.3414	0.3347	0.3347	0.3203	0.3589
20%	0.3549	0.3448	0.3422	0.3443	0.3443	0.3430	0.3690
30%	0.3765	0.3696	0.3742	0.3622	0.3739	0.3765	0.3858
50%	0.4908	0.4823	0.4772	0.4884	0.4875	0.4884	0.4904

TABLE VI
RESULTS ACHIEVED BY WTMM, WITH THE SENTENCE PRIOR PROBABILITY MODELED BY USING DIFFERENT FEATURES

	F1	F2	F3	F4	F5	F6	F7
10%	0.3276	0.3288	0.3301	0.3064	0.3160	0.3276	0.3796
20%	0.3352	0.3359	0.3372	0.3160	0.3311	0.3352	0.3776
30%	0.3821	0.3778	0.3835	0.3543	0.3694	0.3821	0.3758
50%	0.4569	0.4619	0.4598	0.4632	0.4596	0.4619	0.4634

TABLE VII
AVERAGE OF THE AVERAGE SIMILARITY AMONG THE RETRIEVED TEXT DOCUMENTS FOR THE REFERENCE SUMMARY AND NONSUMMARY SENTENCES OF THE EVALUATION SET AT DIFFERENT SUMMARIZATION RATIOS

	10%	20%	30%	50%
Summary sentences	0.1674	0.1362	0.1295	0.1148
Non-summary sentences	0.0891	0.0898	0.0859	0.0818

TABLE VIII
RESULTS ACHIEVED BY LM-RT, WITH THE SENTENCE PRIOR PROBABILITY MODELED BY COMBINING MULTIPLE FEATURES

	F1 and F3	F1 and F7	F3 and F7	F1, F3 and F7
10%	0.3639	0.3684	0.3684	0.3684
20%	0.3614	0.3696	0.3696	0.3696
30%	0.3809	0.3857	0.3807	0.3840
50%	0.4813	0.4909	0.4884	0.4884

TABLE IX
RESULTS ACHIEVED BY WTMM, WITH THE SENTENCE PRIOR PROBABILITY MODELED BY COMBINING MULTIPLE FEATURES

	F1 and F3	F1 and F7	F3 and F7	F1, F3 and F7
10%	0.3301	0.3836	0.3836	0.3836
20%	0.3372	0.3772	0.3772	0.3772
30%	0.3835	0.3748	0.3779	0.3728
50%	0.4578	0.4620	0.4633	0.4615

XI, respectively. Compared with the results in Tables V and VI, using multiple features instead of a single feature for sentence prior probability estimation improves the performance in almost all cases, except when F1 and F3 are fused. They seem not to be complementary to each other when a simple linear combination is used. Furthermore, the combinations that include F7 greatly enhance the performance of LM-RT and WTMM. However, at higher summarization ratios (e.g., 30% and 50%), the improvements made by the inclusion of F7 become less significant. Again, this is because the difference between the average similarities of the retrieved text documents for summary and nonsummary sentences is less significant at higher summarization ratios (cf. Table VII).

In the meantime, we are studying other available features. For example, the sentence prior probability can be estimated according to the position of a sentence in the spoken document (the front the sentence, the higher the prior probability it has) [7], [47]. However, the preliminary experiment results have shown that the use of such heuristic information does not always lead to consistent improvements across different spoken document summarization tasks [23], [48]. Moreover, we are also investigating better ways to fuse selected features, including using the whole sentence maximum entropy (WSME) model [49], [50], for more accurate estimation of the sentence prior probability [23]. Unfortunately, no apparent performance improvement over the simple linear combination has been evidenced thus far.

C. Comparison With Conventional Summarization Models

In the last set of experiments, we compare our proposed summarization models with a number of conventional summarization methods that are widely used in spoken document summarization tasks. The models are VSM, MMR, LSA, DIM, SIG, and SVM. The summarization results for these conventional methods are shown in Table X. We can see that the performances of the unsupervised summarization methods are comparable. It is interesting that MMR has the same performance

TABLE X
RESULTS ACHIEVED BY CONVENTIONAL SUMMARIZATION MODELS

	VSM	MMR	LSA	DIM	SIG	SVM
10%	0.3073	0.3073	0.3034	0.3187	0.3144	0.3425
20%	0.3188	0.3214	0.2926	0.3148	0.3259	0.3408
30%	0.3593	0.3678	0.3286	0.3383	0.3428	0.3719
50%	0.4485	0.4501	0.3906	0.4345	0.4666	0.4660

as VSM when the summarization ratio is 10%, and performs only slightly better than VSM at higher summarization ratios, despite that MMR is expected to outperform VSM because it is designed to allow the summary to cover more topics. This, in a sense, reflects that the issue of topic redundancy seems to have only a very limited impact on the accuracy of the automatic summarization studied here, probably due to the reason that each of the broadcast news documents to be summarized is short in its nature and centers on some specific topic or concept [39]. However, this issue still needs further investigation across different spoken document summarization tasks. On the other hand, SVM, the supervised summarization method, significantly outperforms all the conventional unsupervised summarization methods discussed here. The results achieved by SVM are approximately 0.34, 0.34, 0.37, and 0.47 for summarization ratios of 10%, 20%, 30%, and 50%, respectively.

Comparing these results with those achieved by our proposed methods, several observations can be drawn. 1) When a uniform sentence prior distribution is assumed, most of the sentence generative models are on par with the conventional unsupervised models, while LM-RT and WTMM (cf. Table IV) generally outperform the conventional unsupervised models. Note that both LM-RT and WTMM were also trained in an unsupervised manner, as described in Section II-B. 2) With a uniform sentence prior distribution, the performance of LM-RT or WTMM (cf. Table IV) is not as accurate as that of SVM at the summarization ratio of 10%, but it is better than SVM at higher summarization ratios. 3) When the sentence prior probability is properly modeled by a single useful feature (cf. Tables V and VI) or a combination of several features (cf. Tables VIII and IX), both LM-RT and WTMM outperform SVM by a substantial margin.

We further evaluate the performance of WTMM (using F7 for the sentence prior distribution) and SVM using the ROUGE-1, ROUGE-3, and ROUGE-4 measures. The results are shown in Tables XI (for WTMM) and XII (for SVM), where the values in the parentheses are the associated 95% confidence intervals. It is clear that WTMM is better than SVM in most cases. In addition, a five-level subjective human evaluation was performed on the summarization results for the summarization ratios of 20% and 30%, where five was the best and one was the worst. Six graduate students were invited to evaluate the automatic summaries given that the associated reference transcripts were provided. The average results of the human evaluation are shown in Table XIII, where the numbers in the parentheses are the corresponding standard derivations of the results. We can see that WTMM and SVM are comparable to each other in terms of human evaluation. Moreover, it is interesting to note that the

TABLE XI
RESULTS ACHIEVED BY WTMM, EVALUATED USING THE ROUGE-1, ROUGE-2, ROUGE-3, AND ROUGE-4 MEASURES

	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4
10%	0.4609 (0.39704 - 0.52230)	0.3796 (0.31379 - 0.44171)	0.3338 (0.26918 - 0.39664)	0.2947 (0.23481 - 0.35573)
20%	0.4606 (0.40336 - 0.51898)	0.3776 (0.31772 - 0.43619)	0.3291 (0.27096 - 0.38494)	0.2884 (0.23338 - 0.34093)
30%	0.4727 (0.42873 - 0.51695)	0.3758 (0.33058 - 0.42368)	0.3211 (0.27640 - 0.36709)	0.2733 (0.22850 - 0.31725)
50%	0.5808 (0.55394 - 0.60771)	0.4634 (0.43627 - 0.49149)	0.3861 (0.35923 - 0.41537)	0.3198 (0.29291 - 0.34798)

TABLE XII
RESULTS ACHIEVED BY SVM, EVALUATED USING THE ROUGE-1, ROUGE-2, ROUGE-3, AND ROUGE-4 MEASURES

	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4
10%	0.4198 (0.35573 - 0.48792)	0.3425 (0.27879 - 0.41320)	0.3056 (0.24328 - 0.36769)	0.2710 (0.21316 - 0.33236)
20%	0.4281 (0.36668 - 0.49019)	0.3408 (0.28045 - 0.40158)	0.2991 (0.24142 - 0.35503)	0.2610 (0.20646 - 0.31457)
30%	0.4726 (0.42522 - 0.51985)	0.3719 (0.32267 - 0.41847)	0.3194 (0.27302 - 0.36517)	0.2723 (0.22973 - 0.31586)
50%	0.5833 (0.55313 - 0.61026)	0.4660 (0.43611 - 0.49630)	0.3897 (0.36141 - 0.41873)	0.3234 (0.29513 - 0.35218)

TABLE XIII
RESULTS ACHIEVED BY WTMM AND SVM, EVALUATED BY SIX HUMAN SUBJECTS

	WTMM	SVM
20%	2.83 (0.11)	2.89 (0.12)
30%	3.57 (0.11)	3.55 (0.10)

human subjects have a tendency to give higher scores to the longer automatic summaries.

Although SVM can achieve quite comparable results in either ROUGE-N or the human evaluation, it, however, requires a set of handcrafted document-summary exemplars to learn its summarization capability, and tends to have poor performance in the absence of human supervision [48]. In contrast, most of the unsupervised summarization methods, including our proposed methods, usually consider the relevance (or proximity) of a sentence to the whole document, which might be more robust across different summarization tasks. Therefore, how to make use of unsupervised or semi-supervised learning to improve the performance of supervised summarizers when handcrafted labels are not available for training the supervised summarizers might be an important issue for spoken document summarization [48].

D. Discussions

The above experiment results seem to indicate that the proposed probabilistic generative framework and the associated summarization models are effective alternatives to the other summarization methods compared in this paper. For fair comparisons between these models, all the summarization experiments were carefully designed to avoid "testing on training"; i.e., all the training (or parameter) settings for our proposed summarization models and the conventional summarization models were trained (or tuned) by using the development set, and then applied to the evaluation set. Generally speaking, the training (or parameter) settings tuned on the development set performed rather well in the evaluation set.

A novel aspect of our proposed framework is that it can leverage various kinds of sentence generative models and sentence prior probabilities, and the estimation of the associated parameters can be conducted in a purely unsupervised manner, without the need for handcrafted document-summary pairs. Though STMM needs a set of contemporary text news documents with corresponding human-generated titles to train the latent topical distributions, we have developed an unsupervised training approach (i.e., STMM-U) to bypass this limitation. Moreover, the experiment results have confirmed our expectation that the relevance information of the spoken sentences, provided by the local feedback-like procedure, can greatly enhance the estimation of both the sentence generative model and the sentence prior probability for broadcast news speech summarization. The proposed summarization models in essence are equally applicable to both the text and spoken document summarization tasks, except that some features used for modeling the sentence prior probability are speech-specific. It is also worth noting that only simple word or topic unigrams (multinomial distributions) are employed for modeling the sentence generative probability in the proposed summarization models.

The additional, albeit important, difficulties for spoken document summarization are the inevitable speech recognition errors caused by problems of spontaneous speech, such as pronunciation variations as well as redundant acoustic effects, and the out-of-vocabulary (OOV) problem for words outside the vocabulary of the speech recognizer. Though the summarization methods, together with the associated experiments and evaluations, presented in this paper are not intended to focus on dealing with these problems, they still remain worthy of further investigation, especially when summarizing spontaneous spoken documents such as voice mails, lectures, and meeting recordings [2], [10], [22], [51]. A straightforward remedy, apart from the many approaches improving recognition accuracy, is to develop more robust representations for speech signals. For example, multiple recognition hypotheses, beyond the top scoring ones, obtained from N-best lists, word lattices, or confusion networks, can provide alternative (or soft) representations for the confusing portions of the spoken documents [52]. A scoring method using different confidence measures, e.g., posterior probabilities incorporating acoustic and language model likelihoods, measures considering relationships between adjacent word hypotheses, and prosodic features including pitch, energy stress, and duration measure, can also help to express the uncertainty of word occurrences and sentence boundaries [10], [50], [51]. Hence, sentence selection can be conducted on the basis of these representations. Moreover, the use of subword units (for example,

syllables or segments of them), as well as the combination of words and subword units, for representing the spoken documents has also been proven beneficial for spoken document summarization [24], [53]. One the other hand, the selected important sentences can be concatenated and further modified into a written article style by a sentence compaction scheme, which, for example, can employ a set of heuristic measures, including word concatenation scores and stochastic dependency grammar scores, and a dynamic programming technique to remove redundant acoustic effects, such as disfluencies, fillers, and repetitions [10].

The latent topical distributions of STMM and WTMM were trained offline before performing the summarization task. For a spoken document with unseen topics, the associated topical distributions of the sentences were simply approximated by the existing ones. It is worth mentioning that the approximation might lead to inaccurate estimation of the associated sentence generative models. Therefore, dynamic topic adaptation will be very important for better estimation of STMM and WTMM [34], [54]. It is also important to explore more features and characteristics inherent in the spoken documents, such as the speaking styles, emotional information, and rhetorical structures [55]. These features, together with the lexical, prosodic, confidence, and relevance features that we have investigated in this paper, should be fused under a more effective way for spoken document summarization.

V. Conclusion

We have proposed a probabilistic generative framework that combines the sentence generative probability and the sentence prior probability for extractive spoken document summarization. Each sentence of a spoken document to be summarized is treated as a probabilistic generative model for predicting the document. Various modeling approaches, including the language model (LM), the relevance model (RM), the sentence topical model (STMM), and the word topical mixture model (WTMM), have been extensively investigated for this purpose. In addition, several sets of lexical, prosodic, confidence, and relevance features have been properly incorporated for the estimation of the sentence prior probability. The results of experiments on Chinese broadcast news show that the proposed framework and associated models are good alternatives to the other summarization methods compared in this paper.

Acknowledgment

The authors would like to thank the reviewers for valuable comments that greatly improved the quality of this paper. The authors would also like to thank the Speech Processing Lab of National Taiwan University for providing the necessary speech and language data.

References

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

[2] K. Koumpis and S. Renals, "Content-based access to spoken audio," *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 61–69, Sep. 2005.

[3] *Advances in Automatic Text Summarization*, I. Mani and M. T. Maybury, Eds. Cambridge, MA: MIT Press.

[4] C. D. Paice, "Constructing literature abstracts by computer: Techniques and prospects," *Inf. Process. Manag.*, vol. 26, no. 1, pp. 171–86, 1990.

[5] M. Witbrock and V. Mittal, "Ultra summarization: A statistical approach to generating highly condensed non-extractive summaries," in *Proc. ACM SIGIR Conf. R&D in Inf. Retrieval*, 1999, pp. 315–316.

[6] P. B. Baxendale, "Machine-made index for technical literature-an experiment," *IBM J.*, Oct. 1958.

[7] M. Hirohata, Y. Shinnaka, K. Iwano, and S. Furui, "Sentence extraction-based presentation summarization techniques and evaluation metrics," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2005, pp. 1065–1068.

[8] G. Murray, S. Renals, and J. Carletta, "Extractive summarization of meeting recordings," in *Proc. Eur. Conf. Speech Commun. Technol.*, 2005, pp. 593–596.

[9] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in *Proc. ACM SIGIR Conf. R&D Inf. Retrieval*, 2001, pp. 19–25.

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

[11] M. Hirohata, Y. Shinnaka, K. Iwano, and S. Furui, "Sentence-extractive automatic speech summarization and evaluation techniques," *Speech Commun.*, vol. 48, no. 9, pp. 1151–1161, 2006.

[12] S. Maskey and J. Hirschberg, "Summarizing speech without text using hidden Markov models," in *Proc. HLT-NAACL*, 2006, pp. 89–92.

[13] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer," in *Proc. ACM SIGIR Conf. R&D Inf. Retrieval*, 1995, pp. 68–73.

[14] J. Zhang and P. Fung, "Speech summarization without lexical features for mandarin broadcast news," in *Proc. NAACL HLT, Companion Volume*, 2007, pp. 213–216.

[15] M. Galley, "Skip-chain conditional random field for ranking meeting utterances by importance," in *Proc. Empirical Methods in Natural Lang. Process.*, 2006, pp. 364–372.

[16] X. Zhu and G. Penn, "Evaluation of sentence selection for speech summarization," in *Proc. 2nd Int. Conf. Recent Adv. Natural Lang. Process. (RANLP-05), Workshop Crossing Barriers in Text Summarization Res.*, 2005, pp. 39–45.

[17] M. Amini, N. Usunier, and P. Gallinari, "Automatic text summarization based on word-clusters and ranking algorithms," in *Proc. Eur. Conf. Inf. Retrieval Res.*, 2005, pp. 142–156.

[18] K. Bellare, A. D. Sarma, A. D. Sarma, N. Loiwal, V. Mehta, G. Ramakrishnan, and P. Bhattacharya, "Generic text summarization using WordNet," in *Proc. Int. Conf. Lang. Resources Evaluation*, 2004, pp. 691–694.

[19] W. Li, M. Wu, Q. Lu, W. Xu, and C. Yuan, "Extractive summarization using inter- and intra- event relevance," in *Proc. Annu. Meeting Assoc. Comput. Linguist.*, 2006, pp. 369–376.

[20] D. Bollegala, N. Okazaki, and M. Ishizuka, "A bottom-up approach to sentence ordering for multi-document summarization," in *Proc. Annu. Meeting Assoc. Comput. Linguist.*, 2006, pp. 385–392.

[21] S. Maskey and J. Hirschberg, "Comparing lexical, acoustic/prosodic, structural and discourse Features for speech summarization," in *Proc. Eur. Conf. Speech Commun. Technol.*, 2005, pp. 621–624.

[22] K. Koumpis and S. Renals, "Automatic summarization of voicemail messages using lexical and prosodic features," *ACM Trans. Speech Lang. Process.*, vol. 2, no. 1, pp. 1–24, 2005.

[23] Y. T. Chen, H. S. Chiu, H. M. Wang, and B. Chen, "A unified probabilistic generative framework for extractive spoken document summarization," in *Proc. Eur. Conf. Speech Commun. Technol.*, 2007, pp. 2805–2808.

[24] Y. T. Chen, S. Yu, H. M. Wang, and B. Chen, "Extractive Chinese spoken document summarization using probabilistic ranking models," in *Proc. Int. Symp. Chinese Spoken Lang. Process.*, 2006, pp. 660–671.

[25] B. Chen, Y. M. Yeh, Y. M. Huang, and Y. T. Chen, "Chinese spoken document summarization using probabilistic latent topical information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2006, pp. 969–972.

[26] B. Chen and Y. T. Chen, "Word topical mixture models for extractive spoken document summarization," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2007, pp. 52–55.

[27] Y. T. Chen, S. H. Lin, H. M. Wang, and B. Chen, "Spoken document summarization using relevant information," in *Proc. IEEE Workshop Automatic Speech Recognition and Understanding*, 2007, pp. 189–194.

[28] J. Frederick, *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press, 1999.

[29] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, vol. 39, no. 1, pp. 1–38, 1977, Series B.
[30] , W. B. Croft and J. Lafferty, Eds., *Language Modeling for Information Retrieval*. Norwell, MA: Kluwer, 2003.
[31] J. Xu and W. B. Croft, "Query expansion using local and global document analysis," in *Proc. ACM SIGIR Conf. R&D in Inf. Retrieval*, 1996, pp. 4–11.
[32] B. Chen, H. M. Wang, and L. S. Lee, "A discriminative HMM/n-gram-based retrieval approach for Mandarin spoken documents," *ACM Trans. Asian Lang. Inf. Process.*, vol. 3, no. 2, pp. 128–145, 2004.
[33] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, pp. 177–196, 2001.
[34] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
[35] B. Chen, J. W. Kuo, Y. M. Huang, and H. M. Wang, "Statistical Chinese spoken document retrieval using latent topical information," in *Proc. Int. Conf. Spoken Lang. Process.*, 2004, pp. 1621–1625.
[36] H. S. Chiu and B. Chen, "Word topical mixture models for dynamic language model adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2007, pp. 169–172.
[37] "Snack Sound Toolkit." [Online]. Available: http://www.speech.kth.se/snack/
[38] C. L. Huang, C. H. Hsieh, and C. H. Wu, "Spoken document summarization using acoustic, prosodic and semantic information," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2005.
[39] B. Chen, H. M. Wang, and L. S. Lee, "Discriminating capabilities of syllable-based features and approaches of utilizing them for voice retrieval of speech information in mandarin Chinese," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 303–314, Jul. 2002.
[40] B. Chen, Y. T. Chen, C. H. Chang, and H. B. Chen, "Speech retrieval of mandarin broadcast news via mobile devices," in *Proc. Eur. Conf. Speech Commun. Technol.*, 2005, pp. 109–112.
[41] B. Chen, J. W. Kuo, and W. H. Tsai, "Lightly supervised and data-driven approaches to mandarin broadcast news transcription," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, pp. 777–780.
[42] S. H. Liu, F. H. Chu, S. H. Lin, H. S. Lee, and B. Chen, "Training data selection for improving discriminative training of acoustic models," in *Proc. IEEE Workshop Automatic Speech Recognition and Understanding*, 2007, pp. 284–289.
[43] "Central News Agency (CNA)." [Online]. Available: http://www.cna.com.tw/
[44] A. Stolcke, "SRI Language Modeling Toolkit." 2005 [Online]. Available: http://www.speech.sri.com/projects/srilm/, Version 1.4.4.
[45] C. Y. Lin, "ROUGE: Recall-oriented Understudy for Gisting Evaluation." 2003 [Online]. Available: http://haydn.isi.edu/ROUGE/
[46] "SSVM Toolbox." 2007 [Online]. Available: http://dmlab1.csie.ntust.edu.tw/
[47] R. Brandow, K. Mitze, and L. F. Rau, "Automatic condensation of electronic publications by sentence selection," *Inf. Process. Manag.*, vol. 31, no. 5, pp. 675–685, 1995.
[48] S. H. Lin, Y. T. Chen, H. M. Wang, and B. Chen, "A comparative study of probabilistic ranking models for spoken document summarization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 5025–5028.
[49] R. Rosenfeld, S. F. Chen, and X. Zhu, "Whole-sentence exponential language models: A vehicle for linguistic-statistical integration," *Comput. Speech Lang.*, vol. 15, no. 1, pp. 55–73, 2001.
[50] O. Chan and R. Togneri, "Prosodic features for a maximum entropy language model," in *Proc. Int. Conf. Spoken Lang. Process.*, 2006, pp. 1858–1861.
[51] T. Kawahara, M. Hasegawa, K. Shitaoka, T. Kitade, and H. Nanjo, "Automatic indexing of lecture presentations using unsupervised learning of presumed discourse markers," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 4, pp. 409–419, Jul. 2004.
[52] C. Chelba, T. J. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Process. Mag.*, vol. 25, no. 3, pp. 39–49, May 2008.
[53] S. Y. Kong and L. S. Lee, "Improved summarization of Chinese spoken documents by probabilistic latent semantic analysis (PLSA) with further analysis and integrated scoring," in *Proc. Int. Workshop Spoken Lang. Technol.*, 2006, pp. 26–29.
[54] J. T. Chien and M. S. Wu, "Adaptive Bayesian latent semantic analysis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 198–207, Jan. 2008.
[55] J. J. Zhang, H. Y. Chan, and P. Fung, "Improving lecture speech summarization using rhetorical information," in *Proc. IEEE Workshop Automatic Speech Recognition and Understanding*, 2007, pp. 195–200.

Yi-Ting Chen received the B.S. degree in computer science and information engineering from Tunghai University, Taichung, Taiwan, in 2004 and the M.S. degrees in computer science and information engineering from National Taiwan Normal University, Taipei, Taiwan, in 2007.

She was an Intern and then a Research Assistant with the Institute of Information Science, Academia Sinica, Taipei, from 2005 to 2007. Her research interests are in speech recognition, natural language processing, and information retrieval.

Berlin Chen (M'04) received the B.S. and M.S. degrees in computer science and information engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1994 and 1996, respectively, and the Ph.D. degree in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 2001.

He was with the Institute of Information Science, Academia Sinica, Taipei, from 1996 to 2001, and then with the Graduate Institute of Communication Engineering, National Taiwan University, from 2001 to 2002. In 2002, he joined National Taiwan Normal University, Taipei, where he is now an Associate Professor in the Department of Computer Science and Information Engineering. His current research activities center around robust and discriminative feature extraction, acoustic and language modeling, search algorithms for large-vocabulary continuous speech recognition (LVCSR), and speech retrieval, summarization, and mining.

Prof. Chen is a member of the ISCA and ACLCLP. He currently serves as a board member and chair of academic council of ACLCLP.

Hsin-Min Wang (S'92–M'95–SM'05) received the B.S. and Ph.D. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1989 and 1995, respectively.

In October 1995, he joined the Institute of Information Science, Academia Sinica, Taipei, Taiwan, as a Postdoctoral Fellow. He was promoted to Assistant Research Fellow and then Associate Research Fellow in 1996 and 2002, respectively. He was an Adjunct Associate Professor with National Taipei University of Technology and National Chengchi University. His major research interests include speech processing, natural language processing, spoken dialogue processing, multimedia information retrieval, and pattern recognition.

Dr. Wang was a recipient of the Chinese Institute of Engineers (CIE) Technical Paper Award in 1995. He is a life member of ACLCLP and IICM and a member of ISCA. He was a board member and chair of academic council of ACLCLP. He currently serves as Secretary-General of ACLCLP and as an editorial board member of the *International Journal of Computational Linguistics and Chinese Language Processing*.