Leveraging Kullback–Leibler Divergence Measures and Information-Rich Cues for Speech Summarization

Shih-Hsiang Lin, *Member, IEEE*, Yao-Ming Yeh, and Berlin Chen, *Member, IEEE*

***Abstract*—Imperfect speech recognition often leads to degraded performance when exploiting conventional text-based methods for speech summarization. To alleviate this problem, this paper investigates various ways to robustly represent the recognition hypotheses of spoken documents beyond the top scoring ones. Moreover, a summarization framework, building on the Kullback–Leibler (KL) divergence measure and exploring both the relevance and topical information cues of spoken documents and sentences, is presented to work with such robust representations. Experiments on broadcast news speech summarization tasks appear to demonstrate the utility of the presented approaches.**

***Index Terms*—Kullback–Leibler (KL) -divergence, multiple recognition hypotheses, relevance information, speech summarization, topical information.**

I. Introduction

In THE era of Internet explosion, the information overload problem calls for considerable research effort to investigate efficient and effective technologies for managing the rapidly growing amount of textual information and multimedia content. Automatic summarization systems which enable users to quickly digest the important information conveyed by either a single or a cluster of documents are indispensible when dealing with this problem. The research of text summarization dates back to the late 1950s [1] and has continued to be an attractive subject of much research [2], [3].

A summary can be either abstractive or extractive. In abstractive summarization, a fluent and concise abstract that reflects the key concepts of a document (or set of documents) will be provided, whereas the summary is essentially formed by selecting salient sentences from the original document in extractive summarization. The former requires highly sophisticated natural language processing techniques, including semantic representation and inference, as well as natural language generation; this would make abstractive approaches difficult to replicate or extend from constrained domains to general domains. Thus, in recent years, researchers have tended to focus on extractive summarization. In addition to being extractive or abstractive, a summary may also be generated by considering factors from other aspects, e.g., being generic or query-oriented. A generic summary highlights the most salient information in a document, whereas a query-oriented summary presents the information in a document that is most relevant to a user's query. Interested readers can refer to [2] for a comprehensive overview of text summarization. Additionally, due to the maturity of text summarization, the research focus has been extended to speech summarization over the years. Speech summarization is anticipated to distill important information and remove redundant and incorrect information caused by recognition errors from spoken documents, enabling users to efficiently review spoken documents and understand the associated topics quickly [4]. It would also be useful for improving the efficiency of a number of potential applications like retrieval and mining of large volumes of spoken documents. Most of the existing studies usually assume that spoken documents equipped with the top one recognition transcripts in text form are available, based on which the well-established text summarization techniques can be applied [3].

Aside from traditional ad-hoc extractive text summarization methods, such as those based on document structure and style information [5], linguistic information [6], proximity [7] or significance measures [8] to identify salient sentences or paragraphs, machine-learning approaches with either supervised or unsupervised learning strategies have gained much attention and been applied with empirical success to many summarization tasks [9]. For supervised machine-learning approaches, the summarization task is normally cast as a two-class (summary/non-summary) sentence-classification problem: a sentence with a set of indicative features is input to the classifier (or summarizer) and a decision is then returned from it on the basis of these features. Representative supervised machine-learning summarizers include, but not limited to, Bayesian classifier [10], support vector machine (SVM) [11], and conditional random fields (CRFs) [12]. The major shortcoming of these summarizers is that a set of handcrafted document-reference summary exemplars are required for training the summarizers; nonetheless, manual annotation is often tedious and expensive in terms of time and personnel. Moreover, such summarizers trained on a specific domain might not be readily applicable to other domains due to, for example, different document genres or word usages. The other potential problem is the bag-of-sentences assumption implicitly made by most of these summarizers. Put differently, sentences are classified independently of each other, with little consideration

Manuscript received January 26, 2010; revised May 12, 2010; accepted July 21, 2010. Date of publication August 16, 2010; date of current version March 30, 2011. This work was supported in part by the National Science Council, Taiwan, under Grants NSC98-2221-E-003-011-MY3, NSC 99-2515-S-003-004, and NSC98-2631-S-003-002 and by National Taiwan Normal University under Grant 99T3060-1. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gokhan Tur.

The authors are with the Department of Computer Science and Information Engineering, National Taiwan Normal University, Taipei 116, Taiwan (e-mail: shlin@csie.ntnu.edu.tw; ymyeh@csie.ntnu.edu.tw; berlin@csie.ntnu.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.2010.2066268

of the dependence relationship among the sentences or the global structure of the document.

Another stream of thought attempts to conduct document summarization based on some heuristic rules or statistical evidences between each sentence and the document, getting around the demand for manually labeled training data. We may name them unsupervised summarizers. For example, the graph-based methods, including TextRank [13] and LexRank [14], to name a few, conceptualize the document to be summarized as a network of sentences, where each node represents a sentence and the associated weight of each link represents the lexical similarity relationship between a pair of nodes. Document summarization thus relies on the global structural information embedded in such conceptualized network, rather than merely considering the local features of each node (sentence). Put simply, sentences more similar to others are deemed more salient to the main theme of the document. Some other studies investigate the use of probabilistic models to capture the relationship between sentences and the document content [15]–[17]. Moreover, we have recently proposed a probabilistic generative ranking approach for speech summarization, which can perform the summarization task in a purely unsupervised manner [18]. Each sentence of a document to be summarized is treated as a probabilistic generative model or a language model for generating the document, and important sentences are selected according to their associated document-likelihoods. Even though the performance of unsupervised summarizers is usually worse than that of supervised summarizers, their domain-independent and easy-to-implement properties still make them attractive [9].

Most of the above methods can be equally applied to both text and speech summarization; the latter, in particular, presents unique difficulties, such as speech recognition errors, problems with spontaneous speech, and the lack of correct sentence or paragraph boundaries [3]. It has been shown that speech recognition errors are the dominating factor for the performance degradation of speech summarization when using recognition transcripts instead of manual transcripts, whereas erroneous sentence boundaries cause relatively minor problems [9], [19]–[21]. A straightforward remedy, apart from the many approaches to improving recognition accuracy, might be to develop more robust representations for spoken documents. For example, multiple recognition hypotheses, beyond the top scoring ones, are expected to provide alternative representations for the confusing portions of the spoken documents [22]–[24]

This paper extends our previous approach to speech summarization [18] in several significant ways: 1) we first evaluate the value of using multiple recognition hypotheses for representing spoken documents in the extractive speech summarization task; 2) a different summarization framework built on top of the Kullback–Leibler (KL) divergence measure [25], [26] is presented as well to properly accommodate other extra information cues for better summarization quality, where the sentence importance is calculated based on the "model distance" between a document and a sentence instead of the "document-likelihood"; and 3) two alternative sentence ranking strategies, namely, the sentence-wise strategy and the list-wise strategy, deduced from such a framework are extensively investigated.

The rest of this paper is structured as follows. Section II elucidates the summarization framework that leverages the KL-divergence measure. Section III discusses various ways to robustly represent the recognition hypotheses of spoken documents, including the confusion network (CN) [27] and the position specific posterior lattice (PSPL) [23]. Section VI describes some possible extensions, including the use of relevance and topical information cues to enhance the performance of the proposed summarization framework. Then, the experimental setup and a series of experiments and associated discussions are presented in Sections V and VI, respectively. Finally, in Section VII, some conclusions are drawn which provide avenues for future work.

II. Proposed Summarization Framework

Extractive summarization produces a concise summary by selecting salient sentences or paragraphs from an original document according to a predefined target summarization ratio. Conceptually, it could be cast as an ad hoc information retrieval (IR) problem, where the document is treated as an information need and each sentence of the document is regarded as a candidate information unit to be retrieved according to its relevance (or importance) to the information need. Therefore, the ultimate goal of extractive summarization could be stated as the selection of the most representative sentences that can succinctly describe the main theme of the document. In the past several years, the language modeling (LM) approach have been introduced to a wide spectrum of IR tasks and demonstrated with good empirical success [28], [29]; this modeling paradigm has been successfully adopted for speech summarization recently [18].

In our previous work [18], each sentence S of a spoken document D to be summarized is treated as a probabilistic generative model for generating the document, and sentences are selected according to their generative probability $P(D|\theta_S)$, which can be approximated by

$$P(D|\theta_s) \approx \prod_{w \in D} P(w|\theta_s)^{c(w,D)} \tag{1}$$

where $c(w, D)$ is the occurrence count of a specific type of word (or term) w in D, reflecting that w will contribute more in the calculation of $P(D|\theta_S)$ if it occurs more frequently in D. The simplest way is to estimate the sentence model $P(w|\theta_S)$ on the basis of the frequency of words occurring in the sentence, with the maximum-likelihood estimation (MLE)

$$P(w|\theta_S) = \frac{c(w,S)}{|S|} \tag{2}$$

where $c(w, S)$ is the number of times that word w occurs in S and $|S|$ is the document length. However, the true sentence model might not always be accurately estimated by MLE, since the sentence consists of only a few sampled words and the portions of the words present are not the same as the probabilities of words in the true model. This phenomenon is especially prominent for the sentences of a spoken document when they are, respectively, represented solely by the best recognition transcript generated by automatic speech

recognition (ASR), i.e., the so-called 1-best recognition result, which is often error prone. To alleviate this problem, one may utilize other sophisticated techniques, such as language model smoothing or topic modeling, for better sentence model estimation [28]. Nevertheless, these approaches are restricted in the context of speech summarization, since they strive only to provide better estimation of each individual sentence model [cf. (1)] [18] without contemplating better ways to represent the whole spoken document.

In order to address the above drawback, we present a different summarization framework, building on the KL-divergence measure [25], [28], for important sentence selection or ranking, which assesses the relationship between the sentences of a document to be summarized and the document itself from a more rigorous information-theoretic perspective. For this idea to work, two different language models are involved in the KL-divergence measure: one for the whole document and the other for each sentence. We assume that words in the document are simple random draws from a language distribution describing some topics of interest and words in the sentences that belong to the summary should also be drawn from the same distribution. Therefore, we can use KL-divergence to quantify how close the document D and one of its sentences S are: the closer the sentence model $P(w|\theta_S)$ to the document model $P(w|\theta_D)$, the more likely the sentence would be part of the summary. The divergence of the sentence model with respect to the document model is defined by

$$\mathrm{KL}(\theta_D||\theta_S) = \sum_{w\in V} P(w|\theta_D)\log\frac{P(w|\theta_D)}{P(w|\theta_S)} \tag{3}$$

where w denotes a specific word in the vocabulary set V; and a sentence S has a smaller value (or probability distance) in terms of $\mathrm{KL}(\theta_D||\theta_S)$ is deemed to be more important. Then, the summary sentences of a given spoken document can be iteratively chosen (i.e., one at each iteration) from the spoken document in accordance with its corresponding divergence until the aggregated summary reaches a predefined target summarization ratio. Recently, Haghighi *et al.* [30] and Celikyilmaz and Hakkani-Tur [31] also employed a similar criterion for importance sentence selection in summarization of multiple text documents, and they also proposed the use of topical information for the purpose of language model smoothing. However, to our knowledge, this criterion has not yet been extensively explored in the context of speech summarization.

To go a step further, such an iterative (or greedy) selection procedure described above may sometimes result in suboptimal performance of importance sentence selection due to the following two main reasons: 1) a summary sentence is selected independently without considering the redundant information that might be contained in the already selected summary sentences and 2) the information carried by a verbose summary sentence sometimes would be succinctly depicted by one or more other concise (short) sentences which cover more topics of interest. To alleviate the first problem (i.e., the redundancy problem), the maximum marginal relevance (MMR) algorithm, which aims at excluding those sentences which are too similar to the already selected summary sentences, is always considered to be a representative approach. For the second problem (i.e., the global optimization problem), we may formulate the extractive summarization as a maximum convergence problem under a summary length constraint and to solve the problem by some global inference algorithms [32]. We, however, present here an alternative remedy to simultaneously deal with the above two problems on top of the KL-divergence measure. To do this, we consider every possible combination (or subset) of sentences in a spoken document as a candidate summary π and then compute its KL-divergence to the spoken document to be summarized. Therefore, the best summary π^* can be generated through the following equation:

$$\begin{aligned}\pi^* &= \operatorname*{argmin}_{\pi\in\Pi_D} \mathrm{KL}(\theta_D||\theta_\pi) \\ &= \operatorname*{argmin}_{\pi\in\Pi_D} \sum_{w\in V} P(w|\theta_D)\log\frac{P(w|\theta_D)}{P(w|\theta_\pi)}\end{aligned} \tag{4}$$

where Π_D denotes all possible combinations (i.e., the candidate summary set) of sentences in a spoken document D and θ_π denotes the model of a given candidate summary π (i.e., the summary model). It should be noted that the length of any possible candidate summary should satisfy the summary length constraint. For clarity of presentation, we hereafter term (3) the "sentence-wise" KL-divergence selection strategy and (4) the "list-wise" KL-divergence selection strategy, respectively.

We may compare the proposed KL-divergence methods [cf. (3) and (4)] with the LM method [cf. (1)] [18] from two aspects. On one hand, the ranking strategies of the KL-divergence methods are based on the probability distance between the document model and the sentence model (or the summary model), instead of the likelihood of the words in the document being generated by the sentence model as that done by the LM method. On the other hand, it is easy to show that the sentence-wise KL-divergence method [cf. (3)] can be degenerated to the LM method [18] once the document model $P(w|\theta_D)$ is estimated merely on the basis of the empirical frequency of words in the document [28].

One has also to bear in mind that, in analogy with the LM method, the true document or sentence model of the KL-divergence methods might not always be accurately estimated by MLE. However, the KL-divergence methods have the merit of being able to accommodate more elaborate model estimation techniques to improve summarization performance in a systematic way. For example, we can pair the KL-divergence methods with robust spoken document representations, such as that using multiple recognition hypotheses to offset the negative effect of inaccurate 1-best recognition results. Alternatively, we can also explore relevance or topical information cues [28] to get more accurate estimation of the document or sentence models employed in the KL-divergence methods. A detailed account on the above two possible refinements of the KL-divergence methods will be given in the following two sections, respectively.

III. Robust Representations of Spoken Documents

To tolerate errors resulting from imperfect ASR systems, there has a great deal of research effort directed towards utilizing word lattices or N-best lists to provide more alternative

recognition hypotheses in various speech transcription, translation, and retrieval tasks over the past several years [22], [24], [27], [33]. A word lattice is usually exploited to serve as an intermediate representation of the ASR output. It is a connected, directed acyclic graph where each arc includes a word hypothesis along with a posterior probability (combining acoustic and language model scores) as well as the time alignment information. It provides a rich set of alternative recognition hypotheses, and each path from the start node to the exit node stands for one hypothesis of spoken word sequences. However, since a word lattice often contains many confusing word hypotheses (including word arcs with very low posterior probabilities) and costs enormous storage space, various compact representations of the word lattice have been developed [23], [33]. In this paper, we investigate and compare the use of CN and PSPL for representing spoken documents and sentences for the purpose of speech summarization. Also worth mentioning is that we treat each sentence as an audio segment $\mathbf{o}$ for generating its own word lattice while the sentence boundaries are determined with the 1-best ASR transcript of the spoken document to be summarized [9].

A. Confusion Network (CN)

A confusion network [27] is a multiple string alignment of the speech recognition results, which transforms all hypotheses in a word lattice into a sequence of equivalence clusters. The original purpose of CN is used to minimize the expected word errors by concatenating those words having the highest posterior probability in each equivalence cluster (or confusion set) to form the recognition output, where the posterior probability of each word hypothesis in a cluster can be thought of as the expected word count. In implementation, the transformation of a CN representation from a word lattice is fulfilled by a two-stage clustering procedure. The first stage is *intra-word clustering*, where clusters have word arcs with the identical orthography are grouped into a new equivalence cluster based on their temporal overlaps and word posterior probabilities. The second stage is then to perform *inter-word clustering* , where several heterogeneous clusters are iteratively grouped together according to their phonetic similarity. Interested reader is suggested to refer to [27] for a thorough and entertaining discussion of CN.

B. Position Specific Posterior Lattice (PSPL)

The basic idea of PSPL is to calculate the posterior probability of a word w occurring at a specific position l in a word lattice [23]. The position is defined as the path length from a start node of the lattice to a particular word. Since there might be more than one path contains the same word in the same position, one would need to sum over all possible paths in a lattice to compute the associated posterior probability (or the expected count) of a word w occurring at a given position l of the lattice. This computation can be accomplished by employing a modified forward–backward algorithm. For the forward search, the forward probability $\alpha(w)$ of a word w is partitioned into several more subtle probability masses $\alpha(w, l)$ according to the length of partial paths that start from the start node and end at w; while the procedure of the backward search remains unchanged. Eventually, the posterior probability of a given word w occurring at a given position l of the lattice can be easily calculated [23].

C. Pruning and Expected Count Computation

After the construction of CN or PSPL, a simple pruning procedure can be adopted to remove the unlikely word hypotheses (i.e., words with lower posterior probabilities) [23]. For each cluster (or position) l, the pruning procedure first finds the most likely word entry in it (i.e., the word with the highest posterior probability). Then, those word entries that have log posterior probabilities lower than that of the most likely one minus a predefined threshold τ are removed from l. As a final point, we can compute the expected frequency count of each word w in a given audio segment $\mathbf{o}$

$$E[c(w, \mathbf{o})] = \sum_{l} \sum_{w_l} P(w_l = w|\mathrm{LAT}(\mathbf{o})) \tag{5}$$

where w_l is an arbitrary word that occurs in cluster (or at position) l; $\mathrm{LAT}(\mathbf{o})$ denotes CN (or PSPL) of audio segment $\mathbf{o}$; $P(w_l = w|\mathrm{LAT}(\mathbf{o}))$ denotes the posterior probability of word w in cluster (or at position) l of audio segment $\mathbf{o}$.

IV. Incorporation of Relevance and Topical Information

As we have mentioned in Section II, the true document or sentence model (or summary model) might not always be accurately estimated when there are only a few words present in the (erroneous) recognition transcript of a spoken document or sentence. In order to shorten this potential defect, we explore two extra information cues, i.e., the relevance and topical information, to enhance the estimation of the document or sentence model employed in the KL-divergence methods.

A. Relevance Information

According to the principle of statistical analysis, reliable estimation of a probabilistic model can be obtained by using a large proportion of the data population being considered. Consequently, a simple and intuitive way to improve the accuracy of model estimation is to enlarge the size of the training data sample. In this paper, the notion of relevance class, originally proposed in the context of IR, is adopted here to facilitate accurate estimation of the document and sentence models used in the KL-divergence methods [34]. To illustrate, we take the sentence model of the sentence-wise KL-divergence method [cf. (3)] as an example. Each sentence S of the spoken document D to be summarized has its own associated relevance class R_S. This class is defined as the subset of documents in the collection that are relevant to the sentence S. The relevance model (RM) of the sentence S is therefore defined to be the probability distribution $P(w|\theta_{R_S})$, which gives the probability that we would observe a word w if we were to randomly select a document from the relevance class R_S and then pick up a random word from that document. Once the relevance model of the sentence S is constructed, it can be used to replace the original sentence model or to be combined with the original sentence model to produce a more accurate estimate. Because there is no prior knowledge

about the subset of relevant documents for each sentence S, a local relevance feedback-like procedure can be employed by taking S as a query and posing it to an IR system to obtain a ranked list of documents from a large document repository. The top L documents returned from the IR system are assumed to be the ones relevant to S, and the relevance model of S can be therefore constructed through the following equation:

$$P(w|\theta_{R_S}) = \sum\nolimits_{D_l \in \mathbf{D}_{\mathrm{Top}L}} P(w|\theta_{D_l}) P(D_l|S) \tag{6}$$

where $\mathbf{D}_{\mathrm{Top}L}$ is the set of the top L retrieved documents; and the probability $P(D_l|S)$ can be approximated by the following equation, through a simple mathematical manipulation

$$P(D_l|S) \approx \frac{P(D_l) \cdot P(S|\theta_{D_1})}{\sum_{d_u} \in \mathbf{D}_{\mathrm{TOP}L} P(D_u) \cdot P(S|\theta_{D_u})}. \tag{7}$$

A uniform prior probability $P(D_l)$ can be further assumed for the top L retrieved documents, and the sentence likelihood $P(S|\theta_{Dl})$ can be readily calculated if the IR system is implemented with a language modeling approach [28], [29]. After obtaining the relevance model, we can employ a two-stage smoothing strategy to form the final sentence model

$$\tilde{P}(w|\theta_S) = \alpha \cdot \{\beta \cdot P(w|\theta_S) + (1-\beta) P(w|\theta_{R_S})\} + (1-\alpha) \cdot P(w|\theta_C) \tag{8}$$

where $P(w|\theta_C)$ is the background model estimated from a general corpus, and the values of the interpolation weights α and β can be empirically set based on the development set, or further optimized by other estimation techniques [28], [29]. Along a similar vein, the relevance model $P(w|\theta_{R_D})$ for the spoken document D to be summarized can be constructed as well.

B. Topical Information

On the other hand, there probably would be word usage mismatch between a spoken document and one of its sentences even if they are topically related to each other. Therefore, instead of constructing the document and sentence (or summary) models of the KL-divergence methods based on literal term information (as previously described in Section II), we exploit probabilistic topic models [35] to represent a spoken document and its sentences through a latent topic space. For example, the associated document model of a spoken document to be summarized is interpreted as document topic model (DTM) consisting of a set of K shared latent topics $\{T_1, \ldots T_k, \ldots, T_K\}$ with document-specific topic weights $P(T_k|\theta_D)$, while each topic offers a unigram (multinomial) distribution $P(w|T_k)$ for observing an arbitrary word w of the vocabulary

$$P_{\mathrm{DTM}}(w|\theta_D) = \sum_{k=1}^{K} P(w|T_k) P(T_k|\theta_D). \tag{9}$$

The key idea we wish to illustrate here is that the probability $P_{\mathrm{DTM}}(w|\theta_D)$ of a word w given by a document D is not computed directly based on the frequency of w occurring in D, but instead based on the frequency of w in a latent topic T_k as well as the likelihood that D generates the respective topic T_k, which in fact exhibits some sort of concept matching [4]. Following the same spirit, the sentence (or summary) model can be derived as well.

There is a rich tradition of research in the realization of DTM. The probabilistic latent semantic analysis (PLSA) [36] and the latent Dirichlet allocation (LDA) [37] are often considered two basic representatives of this research line and have motivated many follow-up studies [28]. The difference between LDA and PLSA lies in the inference of model parameters: PLSA assumes the model parameters are fixed and unknown; while LDA places additional *a priori* constraints on the model parameters, i.e., thinking of them as random variables that follow some Dirichlet distributions. More concretely, LDA possesses fully generative semantics by treating the topic mixture weights of the document topic models as a whole as a K-parameter hidden random variable, rather than a large fixed set of individual parameters which are explicitly linked to the training set; LDA, thus, could overcome the over-fitting problem to some extent [37].

Meanwhile, rather than treating each entire spoken document or sentence as a document topic model, we can regard each word w_j of the language as a word topic model (WTM) [38], [39]. To get to this point, all words are assumed to share a same set of latent topic distributions but have different weights over these topics. The WTM model of each word w_j for predicting the occurrence of a particular word w is expressed by

$$P_{\mathrm{WTM}}(w|\theta_{w_j}) = \sum_{k=1}^{K} P(w|T_k) P(T_k|\theta_{w_j}) \tag{10}$$

where $P(w|T_k)$ and $P(T_k|\theta_{w_j})$, respectively, are the probability of a word w occurring in a specific latent topic T_k and the probability of the topic T_k conditioned on the WTM model of w_j. Then, for example, each document can be viewed as a composite WTM model, while the probability of a word w generated by a document D can be expressed by

$$P_{\mathrm{WTM}}(w|\theta_D) = \sum_{w_j \in D} P_{\mathrm{WTM}}(w|\theta_{w_j}) P(w_j|\theta_D). \tag{11}$$

The resulting composite WTM model for D, in a sense, can be thought of as a kind of language model for translating any word w_j occurring in D to an arbitrary word w of the language. The sentence (or summary) model can be constructed in a similar fashion. Due to limited space, we refer the reader to [35], [37], and [38] for a detailed account on the training of DTM and WTM. Furthermore, words represented in a latent topic space only offer coarse-grained concept clues about a document (or sentence) at the expense of losing the discriminative power among concept-related words in finer granularity. For the reason of better discrimination ability and probability smoothing, we use the same smoothing approach, introduced in Section IV-A for the relevance model, to form the final DTM or WTM models. For example, the unigram probability $P(w|\theta_D)$ of a word w occurring in the document D and the background model $P(w|\theta_C)$ are additionally used in association with the document or word topic model, i.e., $P_{\mathrm{DTM}}(w|\theta_D)$ or $P_{\mathrm{WTM}}(w|\theta_D)$.

TABLE I
STATISTICAL INFORMATION OF THE BROADCAST NEWS DOCUMENTS USED FOR THE SUMMARIZATION EXPERIMENTS

	Development Set	Evaluation Set
Recording Period	November 07, 2001 – January 22, 2002	January 23, 2002 – August 22, 2002
Number of Documents	100	105
Average Duration per Document (in sec.)	129.4	135.2
Avg. Number of words per Document	326	340
Avg. Number of Sentences per Document	20	20
Avg. Character Error Rate	34.4%	35.3%

V. EXPERIMENTAL SETUP

A. Speech and Text Corpora

The speech data set used in this research is the MATBN corpus [40], which is different from the set of broadcast news documents used in our previous studies [18]. It contains approximately 200 hours of Mandarin Chinese TV broadcast news collected by Academia Sinica and the Public Television Service Foundation of Taiwan between November 2001 and April 2003. The content has been segmented into separate stories and transcribed manually. Each story contains the speech of one studio anchor, as well as several field reporters and interviewees. A subset of 205 broadcast news documents (spoken documents that covered a wide range of topics) compiled between November 2001 and August 2002 was reserved for the summarization experiments.

A large number of text news documents collected by the Central News Agency (CNA) between 1991 and 2002 (the Chinese Gigaword Corpus released by LDC) were used. The documents collected in 2000 and 2001 were used to train N-gram language models for speech recognition with the SRI Language Modeling Toolkit [41]. In addition, a subset of about 14 000 text news documents, compiled during the same period as the broadcast news documents to be summarized, was used to estimate the relevance model in (6) and the background model in (8), as well as the word topic model in (10).

B. Spoken Documents for the Summarization Experiments

Three subjects were asked to create summaries of the 205 spoken documents for the summarization experiments as references (the gold standard) for evaluation. The summaries were generated by selecting 50% of the most important sentences in the reference transcript of a spoken document, and ranking them by importance without assigning a score to each sentence. The average Chinese character error rate (CER) obtained for the 205 spoken documents was about 35% [42], [43]. Table I shows some basic statistics about the 205 spoken documents.

C. Performance Evaluation

For the performance evaluation of summarization results, we adopted the widely used ROUGE measure [43] because of its higher correlation with human judgments [44], [45]. The ROUGE measure evaluates the quality of an automatic summary by counting the number of overlapping units, such as N-grams, longest common subsequences, or skip-bigram, between the automatic summary and a set of reference (or manual) summaries. Three variants of the ROGUE measure were used to assess the utility of the proposed methods. They are, respectively, the ROUGE-1 (unigram) measure, the ROUGE-2 (bigram) measure, and the ROUGE-L (longest common subsequence) measure. Generally speaking, the ROUGE-1 measure is to evaluate the informativeness of automatic summaries while the ROUGE-2 measure is to estimate the fluency of automatic summaries. On the contrary, ROUGE-L does not reward for fixed-length N-grams but instead for a combination of the maximal substrings of words, which works well in general for evaluating both content and grammaticality.

TABLE II
LEVELS OF AGREEMENT BETWEEN THE THREE SUBJECTS FOR IMPORTANT SENTENCE RANKING FOR THE EVALUATION SET

	Evaluation Measures		
	ROGUE-1	ROUGE-2	ROUGE-L
Agreement	0.675	0.645	0.631

The summarization ratio, defined as the ratio of the number of words in the automatic (or manual) summary to that in the reference transcript of a spoken document, was set to 10% in this study. Since increasing the summary length tends to increase the chances of getting higher scores in the recall rate of the various ROUGE measures and might not always select the right number of words in the automatic summary as compared to the reference summary, all the experimental results reported hereafter were obtained by calculating the F-scores of the ROUGE measures [44]. Table II shows the levels of agreement between the three subjects for important sentence ranking. Each of these values was obtained by using the summary created by one of the three subjects as the reference summary, in turn for each subject, while those of the other two subjects as the test summaries, and then taking their average. These observations seem to reflect the fact that people may not always agree with each other in selecting the important sentences for representing a given document.

VI. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. Experiments on the KL-Divergence Methods and Different Representations of Spoken Documents

We first show the baseline performance of the language modeling (LM) approach to summarization [18], i.e., the LM method [cf. (1)], on the evaluation set by using the manual transcripts (denoted by "Manual") and the 1-best ASR transcripts (denoted by "1-best *"), respectively. This experiment, in fact, is equivalent to that using the sentence-wise KL-divergence method where the sentence models and the document model are simply estimated by MLE. The corresponding results are shown in Table III, where the values in the parentheses are the associated 95% confidence intervals. Looking at the first two rows of Table III, we see that there are significant performance gaps between summarization using the manual transcripts and the 1-best ASR transcripts. The relative performance degradations caused by using the 1-best ASR transcripts are roughly 25% for all ROUGE measures. The reasons for this phenomenon may lies in the following two factors. One is that the erroneous ASR transcripts of spoken sentences would carry wrong information and thus deviate somewhat from representing the true theme of the spoken document. On the other hand, the ROGUE measures are essentially based on counting the number of overlapping

TABLE III
BASELINE RESULTS ACHIEVED BY THE LM METHOD WITH RESPECT TO THE MANUAL TRANSCRIPTS AND THE1-BEST ASR TRANSCRIPTS

	Evaluation Measures		
	ROGUE-1	ROUGE-2	ROUGE-L
Manual	0.469 (0.430 - 0.507)	0.342 (0.295 - 0.386)	0.410 (0.370 - 0.448)
1-best*	0.327 (0.306 - 0.351)	0.171 (0.149 - 0.196)	0.273 (0.253 - 0.295)
1-best	0.358 (0.328 - 0.392)	0.242 (0.206 - 0.280)	0.332 (0.302 - 0.366)

TABLE IV
RESULTS ACHIEVED BY THE KL-DIVERGENCE METHODS WITH RESPECT TO VARIOUS REPRESENTATIONS OF SPOKEN DOCUMENTS (* DENOTES SIGNIFICANT IMPROVEMENTS OVER THE BASELINE LM METHOD USING 1-BEST ASR TRANSCRIPT ACCORDING TO THE t-TEST ($p<0.05$))

		Evaluation Measures		
		ROGUE-1	ROUGE-2	ROUGE-L
SentKL	1-best	0.358	0.242	0.332
	CN	0.367	0.249	0.339
	PSPL	0.377*	0.256	0.347
ListKL	1-best	0.376	0.247	0.344
	CN	0.378	0.251	0.352
	PSPL	0.394*	0.275*	0.364*

units between the automatic summary and the reference summary; the resulting evaluation results, therefore, would be severely affected by speech recognition errors when applying the various ROUGE measures to evaluate the performance of speech summarization. In order to get around the confounding effect of the latter factor, we assume that the resulting summary sentences can also be presented in speech form (besides text form) such that users can directly listen to the audio segments of the summary sentences to bypass the problem caused by speech recognition errors [8]. Then, we can align the ASR transcripts of the summary sentences to their respective audio segments to obtain the correct (manual) transcripts for evaluation. The corresponding results are shown in the last row ("1-best") of Table III. In this way, we can focus mainly on evaluating the correctness of audio segments extracted out of the spoken document while reducing (or ignoring) the confounding effect raised by speech recognition errors. Therefore, all the summarization results reported in the rest of this paper will follow this setup, unless otherwise stated.

We then turn our attention to investigate the utility of using CN and PSPL for representing spoken documents and sentences. The experimental results are shown in Table IV, where the row "1-best" shows the baseline results obtained by using the 1-best ASR transcripts, while Rows "CN" and "PSPL" are the results obtained by using the CN and PSPL representations, respectively. For these two representations, the pruning thresholds (as described in Section III-C) were tuned on the development set and then applied to the evaluation set. It also worth mentioning that "SentKL" is used to denote the sentence-wise KL-divergence selection strategy [cf. (3)] while "ListKL" the list-wise KL-divergence selection strategy [cf. (4)]. For practical implementation of ListKL, it is almost impossible to enumerate all possible combinations of summary sentences for forming the summary of a spoken document, due to the reason that the number of possible combinations would grow exponentially as the number of sentences increases. To reduce the computational overhead, we first use the SentKL to select the top 20% importance sentences having the lowest KL-divergence distances to the document as the candidates for being considered to be included in the summary, and then enumerate all possible combinations (or samplings) of these sentences under a specific summary length constraint of the length of the target summary (i.e., containing about 10% words of the original document). As can be seen from Table IV, using either CN or PSPL provides substantial performance boosts over the 1-best ASR transcripts for both SentKL and ListKL. Although not both of the two attempts (cf. the second and third rows of Table IV) give very significant improvements in ROUGE over the baseline LM method (cf. the third row of Table III or the first row of Table IV), the results still reflect the advantage of using multiple recognition hypotheses. It seems to justify our postulation that speech summarization could benefit from leveraging multiple recognition hypotheses, like CN or PSPL, for robust representations of spoken documents and sentences.

If we further compare between PSPL and CN, it shows that the former outperforms the latter. This might be explained by the fact that a given word arc of the original lattice will be exactly assigned into one particular cluster in CN, whereas a given word arc can belong to multiple clusters with different probabilities in PSPL. Phrased another way, CN performs some sort of hard-clustering of word arcs in the lattice while PSPL is a soft-clustering technique. As a result, when a strict threshold is applied in the pruning stage (as described in Section III-C), several content or informative words of CN might be pruned due to lower word posterior probabilities, but they would be more likely to be retained in PSPL since more than one cluster would contain instances of the same word arc.

Furthermore, ListKL consistently outperforms SentKL with respect to various representations of spoken documents. These results show the potential weakness of using the iterative (greedy) sentence selection strategy (e.g., SentKL) in speech summarization. In order to figure out why ListKL outperforms SentKL, we further analyze the average number of sentences respectively selected by ListKL and SentKL subject to the constraint (i.e., the resulting summary cannot contain words more that 10% of the original document). We observe that ListKL selected about 3.15 summary sentences on average while SentKL selected about 2.61 sentences into the summary under the same length constraint. These statistics reveal that ListKL can, to some extent, avoid selecting verbose sentences into the summary.

To take a step forward, we examine the impact of the average lattice size of a given spoken sentence contributed to the summarization performance, which is defined as the average number of words retained in the speech recognition output representation of a given spoken sentence after the pruning stage. Here, we take the pairing of PSPL and SentKL as an example. The associated results conducted on the development set (denoted by DEV) and the evaluation set (denoted by EVAL) are both graphically illustrated in Fig. 1. We see that the ROUGE-1 score goes up as the lattice size increases. It yields about 1% to 2% absolute performance gains as compared to summarization using merely the 1-best ASR transcripts. However, the performance becomes

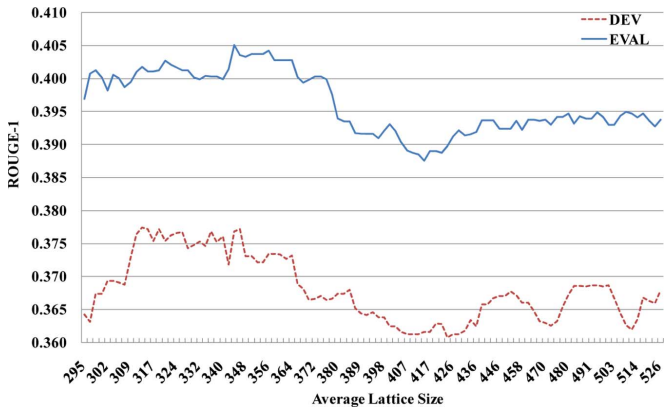

Fig. 1. Impacts of different lattice sizes on the summarization performance.

TABLE V
RESULTS ACHIEVED BY COMBING THE SENTENCE-WISE KL-DIVERGENCE METHOD WITH THE SENTENCE OR/AND DOCUMENT RELEVANCE INFORMATION (* DENOTES SIGNIFICANT IMPROVEMENTS OVER THE BASELINE LM METHOD ACCORDING TO THE t-TEST ($p<0.05$))

	Evaluation Measures		
	ROGUE-1	ROUGE-2	ROUGE-L
Manual	0.469	0.342	0.410
+RM-SEN	0.486	0.361	0.431
+RM-DOC	0.489	0.362	0.430
+Both	0.491	0.367	0.433
1-best	0.358	0.242	0.332
+RM-SEN	0.382*	0.268*	0.356*
+RM-DOC	0.374	0.259	0.346
+Both	0.384*	0.268*	0.358*

saturated, or shows noticeable drops, as PSPL contains too much confusion information.

B. Experiments on Incorporating Relevance and Topical Information

In the next set of experiments, we explore the use of relevance information (cf. Section IV-A) for more accurate estimation of the sentence (denoted by RM-SEN) and document (denoted by RM-DOC) models used in the KL-divergence methods. The number of relevant documents retrieved was determined by the development set. Due to the reason that the computational cost for ListKL is expensive, we report here only the results obtained from SentKL with the relevance information. As shown in Table V, the summarization performance is consistently improved whenever either the sentence relevance information (RM-SEN) or the document relevance information (RM-DOC) is used for model estimation. Furthermore, the summarization performance of RM-SEN is superior to that of RM-DOC in the case of using the 1-best ASR transcripts. One possible explanation is that the spoken sentences are quite short as compared to the spoken document, and may contain erroneous transcripts; they, thus, require more statistical evidence contributed from the relevant documents for better sentence model estimation. On the other hand, the marriage of RM-SEN with RM-DOC (cf. the forth and the last rows in Table V) can provide additional gains, which leads to absolute improvements of about 2% to 3% as compared to the baseline summarization results obtained by merely using either the manual transcripts or the 1-best ASR transcripts.

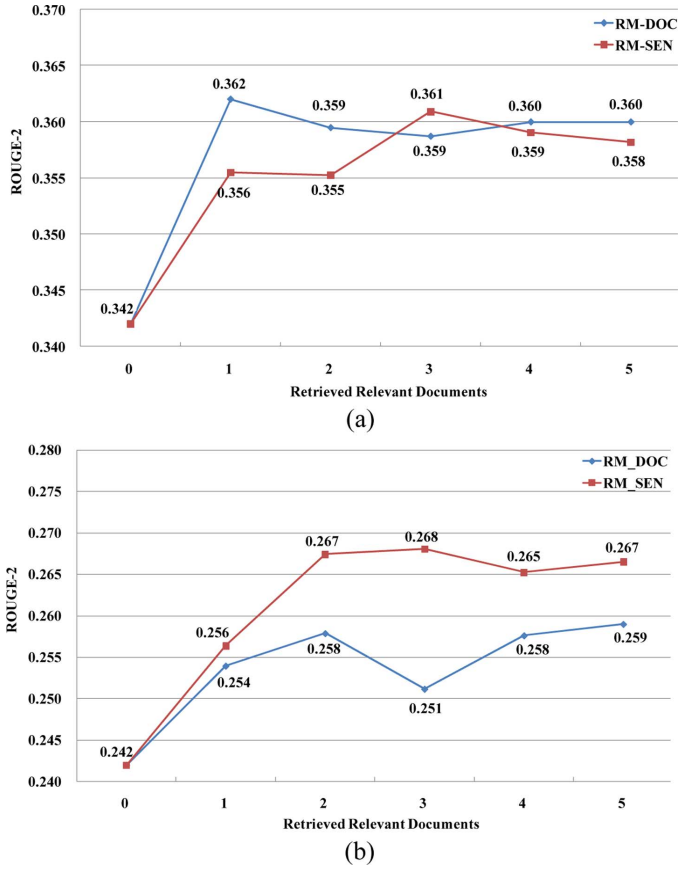

Fig. 2. Contributions of two relevance information cues made to the summarization performance. (a) Manual transcripts. (b) 1-best ASR Transcripts.

To get a better sense of the contributions of the relevance information made to the summarization performance, we further conduct a set of summarization experiments by varying the number of relevant documents being retrieved. Several observations are obtained from the results shown in Fig. 2. First, for the case of using the manual transcripts [cf. Fig. 2(a)], RM-DOC seems more useful than RM-SEN. A partial explanation for this may stem from the fact that RM-DOC plays the role for term expansion and re-weighting of the representation of a spoken document. It may therefore lower the divergence between the document and its "true" summary sentences, as they are modeled by the sentence-wise KL-divergence method. Second, the performance of RM-SEN tends to be improved as the number of retrieved relevant documents is increased. The results are in line with our expectation that incorporating the sentence model with the relevance model might lead to more accurate model estimation. However, the summarization performance is saturated or degraded slightly as the number of retrieved relevant documents is equal to or larger than 3. This phenomenon can be explained by the fact that the retrieved relevant documents might not always be truly relevant to the spoken document or the spoken sentences, and they may make the document model or the sentence models deviate from their original truth.

Further, we examine the utility for the additional use of the topical information. It is worth mentioning that both DTM and WTM are trained without supervision and have the same number of latent topics which was set to 16 in this study. Also note that LDA is taken as a study example for DTM since it

TABLE VI
RESULTS ACHIEVED BY COMBING THE KL-DIVERGENCE METHODS WITH THE TOPICAL INFORMATION CONVEYED BY DIFFERENT TOPIC MODELS (* DENOTES SIGNIFICANT IMPROVEMENTS OVER THE BASELINE LM METHOD ACCORDING TO THE t-TEST ($p<0.05$))

		Evaluation Units		
		ROGUE-1	ROUGE-2	ROUGE-L
SentKL	Manual	0.469	0.342	0.410
	+DTM	0.488	0.366	0.432
	+WTM	0.500*	0.382*	0.442*
	1-best	0.358	0.242	0.332
	+DTM	0.371	0.259	0.344
	+WTM	0.367	0.250	0.340
ListKL	Manual	0.472	0.351	0.423
	+WTM	0.501*	0.380*	0.450*
	1-best	0.376	0.247	0.344
	+WTM	0.397*	0.284*	0.372*

achieved better performance among various DTM models in the literature, and the associated parameters of LDA are estimated by the Gibbs sampling algorithm with symmetric Dirichlet prior distributions. The corresponding results achieved by SentKL incorporated with the topical information are shown in the upper part of Table VI, which reveal that the summarization performance receives substantial boosts from the additional incorporation of the topical information for the case of using the manual transcripts. On the contrary, it gives moderate (but consistent) improvements over the baseline model for summarization using the 1-best ASR transcripts. We attribute this phenomenon to the reason that imperfect ASR transcripts usually contain wrong information and would affect the correctness of the topic inference procedures. Comparing WTM to DTM, we see that WTM outperforms DTM when the manual transcripts are used. However, WTM seems to perform worse than DTM for the case of using the imperfect ASR transcripts. One possible speculation is that, unlike DTM, the model parameters of WTM [cf. (10)] are all estimated from an outside set of text news documents, which somewhat makes WTM unable to faithfully capture the topical relationship among words in the erroneous ASR transcripts. On the other side, we also illustrate the benefit of incorporating the topical information into the ListKL summarizer. Since the topic mixture weights of LDA for a new document have to be estimated online which would often be time-consuming, we only take the WTM as an example for illustration. The results are shown in the lower part of Table VI, which verifies the utility of constructing the ListKL summarizer with the topical information.

C. Comparison With Conventional Summarization Methods

In the final set of experiments, we compare our proposed summarization methods with a few existing summarization methods that have been widely used in speech summarization tasks, including two unsupervised summarizers, namely, vector space model (VSM) and LexRank, and one supervised summarizer, namely, SVM. VSM is a simple but effective literal term matching strategy which first represents each sentence of a spoken document and the spoken document itself in vector form, and then computes the relevance score between each sentence and the document (i.e., the cosine measure of the similarity between two vectors). The sentences with the highest relevance scores are included in the summary accordingly.

TABLE VII
BASIC FEATURES USED BY SVM

Structural feature	1.Duration of the current sentence 2.Position of the current sentence 3.ength of the current sentence
Lexical Feature	1.Number of named entities 2.Number of stop words 3.Bigram language model scores 4.Normalized bigram scores
Acoustic Feature	1.The 1st formant 2.The 2nd formant 3.The pitch value 4.The peak normalized cross-correlation of pitch 5.The energy value
Relevance Feature	1.VSM feature

For SVM, we constructed a binary SVM summarizer with the radial basis function (RBF) as the kernel function and the baseline performance was obtained by using a set of 33 indicative features to characterize a spoken sentence, including the structural features, the lexical features, the acoustic features and the VSM relevance feature. For each kind of the acoustic features, the minimum, maximum, mean, difference value, and mean difference value of a spoken sentence where extracted [9]. The difference value is defined as the difference between the minimum and maximum values of the spoken sentence and mean difference values is defined as the mean difference between a sentence and its previous sentence. The features are outlined in Table VII, where each of them was further normalized to zero mean and unit variance.

The summarization results for these conventional methods are shown in Table VIII. We can see that SVM significantly outperforms VSM and LexRank. The results be attributed to two reasons. First, SVM is trained with the handcrafted document-summary labels of the documents in the development set while the other two methods are conducted in a purely unsupervised manner. Second, SVM utilizes a rich set of features to characterize a spoken sentence while the remaining two methods (VSM and LexRank) are constructed solely on the basis of the lexical information.

Comparing these results with those achieved by our proposed methods (cf. Tables IV–VI), several observations can be drawn. 1) When only the 1-best ASR transcripts are used, the SentKL performs slightly better than VSM and gives a competitive result as compared to LexRank. The summarization performance would become significantly better as we integrated extra information cues into the SentKL. 2) On the other hand, the pairing of ListKL and the topical information gives significant improvement over VSM and LexRank in the case of using merely 1-best ASR transcripts. 3) Our proposed unsupervised summarization methods cannot beat the supervised summarizer (i.e., SVM) for the same reasons mentioned earlier. However, their output (or score) can serve as a complementary feature to augment the feature set used for the supervised summarizer. We take, for example, the various scores obtained by SentKL as additional features to augment the basic feature set (defined in Table VII) for SVM, and the associated results are shown in Table IX. They demonstrate that incorporating the scores provided by SentKL, taken as an additional set of features for SVM, seems to improve the final summarization performance for the TD case. On the contrary, some augmented features hurt the original baseline performance for the SD case, probably due to the imperfect recognition transcripts. However, these results confirm again the

TABLE VIII
RESULTS ACHIEVED BY THREE CONVENTIONAL SUMMARIZATION METHODS COMPARED IN THIS PAPER (Δ DENOTES THE LIST-WISE KL-DIVERGENCE SELECTION STRATEGY WITH THE TOPICAL INFORMATION GIVES SIGNIFICANT IMPROVEMENTS OVER THESE MODELS ACCORDING TO THE t-TEST ($p<0.05$))

	Manual			1-best		
	ROGUE-1	ROUGE-2	ROUGE-L	ROGUE-1	ROUGE-2	ROUGE-L
VSM	0.435Δ	0.305Δ	0.373Δ	0.354Δ	0.241Δ	0.327Δ
LexRank	0.504	0.378	0.446	0.370Δ	0.249Δ	0.343Δ
SVM	0.637	0.562	0.603	0.462	0.363	0.434

TABLE IX
RESULTS ACHIEVED BY AUGMENTING VARIOUS RELEVANCE FEATURES, DERIVED BY THE SENTENCE-WISE KL-DIVERGENCE SELECTION STRATEGY, TO SVM (* DENOTES SIGNIFICANT IMPROVEMENTS OVER THE BASELINE SVM ACCORDING TO THE t-TEST ($p<0.05$))

	Manual			1-best		
	ROGUE-1	ROUGE-2	ROUGE-L	ROGUE-1	ROUGE-2	ROUGE-L
SVM	0.637	0.562	0.603	0.462	0.363	0.434
+KL_Manual	0.647	0.572	0.613	-	-	-
+KL_1-best	-	-	-	0.456	0.359	0.429
+KL_PSPL	-	-	-	0.459	0.361	0.432
+KL_DTM	0.644	0.569	0.611	0.459	0.362	0.433
+KL_WTM	0.635	0.557	0.599	0.462	0.368	0.435
+KL_RM	0.655*	0.582*	0.622*	0.465	0.367	0.439

utility of using multiple recognition hypotheses (cf. the third and forth rows of Table IX).

Further, the results obtained by augmenting the KL_RM feature give a quite comparable performance to those obtained by the human subjects when using the manual transcripts (as shown in Table II). They also give absolute improvements of about 1% to 3% as compared to the results obtained by the SVM summarizer with the basic feature set defined in Table VII.

D. Discussions and Summary

The results shown in Tables III–IX allow us to draw several conclusions. First, using multiple recognition hypotheses derived from CN or PSPL is effective especially when the 1-best ASR transcripts could not provide reliable lexical or topical information (cf. Table IV). We believe this initial attempt not only can benefit the KL-divergence summarization methods, but also can work well in conjunction with other summarization methods. Second, a notable characteristic of the KL-divergence summarization methods is that they can easily leverage various kinds of information sources in a systematic way. The experimental results have clearly supported this claim (cf. Tables IV–VI). Third, the relevant documents were retrieved by the unigram language modeling retrieval approach with merely the 1-best ASR transcripts for the experiments on using the relevance information (cf. Table V). The retrieved documents might contain irrelevant ones owing to the speech recognition errors. A more robust and accurate retrieval model would probably lead to a more substantial improvement of the summarization performance [28], [29]. Fourth, in this paper we employed only word or topic unigrams (multinomial distributions) for modeling the document and sentence models. One possible extension is using more sophisticated techniques like word bigrams or syntactic dependency information to enhance the model estimation. Fifth, the experimental results demonstrate the benefit of using the list-wise KL-divergence selection strategy (ListKL) (cf. Tables IV and VI). It overcomes, to certain extent, the problem of suboptimal performance faced by most of the current widely used sentence-wise selection strategies for extractive summarization. Sixth, Table IX highlights the importance of capturing the relevance of a spoken sentence to the whole spoken document. Seventh, the various ranking scores obtained from the proposed summarization methods, at this moment, are simply treated as an additional set of features for the supervised summarizer (i.e., SVM). More rigorous fusion of various summarizers (or features) still awaits further studies [43], [46]. Finally, our proposed method in essence is equally applicable and effective for both text and spoken document summarization tasks (cf. Tables V and IX).

VII. CONCLUSION

In this paper, we have investigated various ways to robustly represent spoken documents and sentences for speech summarization. We have also presented a KL-divergence-based summarization framework and conducted a series of experiments to verify its capabilities. The experimental results indeed confirm our expectation. Our future research directions include: 1) investigating more elaborate approaches to estimating the document and sentence models of this framework; 2) seeking other ways for representing the ASR output more robustly; and 3) incorporating the summarization results into audio indexing for better retrieval and browsing of spoken documents.

REFERENCES

[1] H. P. Luhn, "The automatic creation of literature abstracts," *IBM J. Res. Develop.*, vol. 2, pp. 157–165, 1958.
[2] I. Mani and M. T. Maybury, *Advances in Automatic Text Summarization*. Cambridge, MA: MIT Press, 1999.
[3] K. McKeown, J. Hirschberg, M. Galley, and S. Maskey, "From text to speech summarization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2005, pp. 997–1000.
[4] L. S. Lee and B. Chen, "Spoken document understanding and organization," *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 42–60, Sep. 2005.

[5] P. Baxendale, "Machine-made index for technical literature—An experiment," *IBM J. Res. Develop.*, pp. 354–361, 1958.
[6] R. Barzilay and M. Elhadad, "Using lexical chains for text summarization," in *Proc. Workshop Intell. Scalable Text Summarization*, 1997, pp. 10–17.
[7] J. Steinberger and K. Ježek, "Using latent semantic analysis in text summarization and summary evaluation," in *Int. Conf. Inf. Syst. Implement. Model.*, 2004, pp. 93–100.
[8] 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.
[9] S.-H. Lin, B. Chen, and H.-M. Wang, "A comparative study of probabilistic ranking models for Chinese spoken document summarization," *ACM Trans. Asian Lang. Inf. Process.*, vol. 8, pp. 3:1–3:23, 2009.
[10] J. Kupiec, J. Pedersen, and F. Chen, "A trainable document summarizer," in *Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1999, pp. 68–73.
[11] J. Zhang and P. Fung, "Extractive speech summarization using shallow rhetorical structure modeling," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1147–1157, Aug. 2010.
[12] D. Shen, J.-T. Sun, H. Li, Q. Yang, and Z. Chen, "Document summarization using conditional random fields," in *Proc. Int. Joint Conf. Artif. Intell.*, 2007, pp. 2862–2867.
[13] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," in *Proc. Conf. Empirical Methods in Natural Lang. Process.*, 2005, pp. 404–411.
[14] G. Erkan and D. R. Radev, "LexRank: Graph-based lexical centrality as salience in text summarization," *J. Artif. Intell. Res.*, vol. 22, pp. 457–479, 2004.
[15] R. Barzilay and L. Lee, "Catching the drift: Probabilistic content models, with applications to generation and summarization," in *Proc. Human Lang. Technol. Conf. North Amer. Chap. Assoc. Comput. Linguist. Annu. Meeting*, 2004, pp. 113–120.
[16] H. DauméIII and D. Marcu, "Bayesian query focused summarization," in *Proc. Annu. Meeting Assoc. Comput. Linguist.*, 2006, pp. 305–312.
[17] A. Nenkova, L. Vanderwende, and K. McKeown, "A compositional context sensitive multi-document summarizer: Exploring the factors that influence summarization," in *Proc. Annual Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2006, pp. 573–580.
[18] Y.-T. Chen, B. Chen, and H.-M. Wang, "A probabilistic generative framework for extractive broadcast news speech summarization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 1, pp. 95–106, Jan. 2009.
[19] H. Christensen, Y. Gotoh, and S. Renals, "A cascaded broadcast news highlighter," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 151–161, Jan. 2008.
[20] G. Penn and X. Zhu, "A critical reassessment of evaluation baselines for speech summarization," in *Annu. Meeting Assoc. Linguist.*, 2008, pp. 470–478.
[21] Y. Liu and S. Xie, "Impact of automatic sentence segmentation on meeting summarization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 5009–5012.
[22] 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.
[23] C. Chelba, J. Silva, and A. Acero, "Soft indexing of speech content for search in spoken documents," *Comput. Speech Lang.*, vol. 21, pp. 458–478, 2007.
[24] T. K. Chia, K. C. Sim, H. Z. Li, and H. T. Ng, "A lattice-based approach to query-by-example spoken document retrieval," in *Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2008, pp. 363–370.
[25] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.
[26] S. H. Lin and B. Chen, "Improved speech summarization with multiple-hypothesis representations and Kullback-Leibler divergence measures," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2009, pp. 1847–1850.
[27] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," *Comput. Speech Lang.*, vol. 14, pp. 373–400, 2000.
[28] C. X. Zhai, *Statistical Language Models for Information Retrieval*. San Rafael, CA: Morgan & Claypool, 2008.
[29] 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, pp. 128–145, 2004.
[30] A. Haghighi and L. Vanderwende, "Exploring content models for multi-document summarization," in *Proc. Human Lang. Technol. Conf. North Amer. Chapt. Assoc. Comput. Linguist. Annu. Meeting*, 2009, pp. 362–370.
[31] A. Celikyilmaz and D. Hakkani-Tur, "A hybrid hierarchical model for multi-document summarization," in *Proc. Annu. Meeting Assoc. Linguist.*, 2010, pp. 815–824.
[32] R. McDonald, "A study of global inference algorithms in multi-document summarization," in *Proc. Eur. Conf. Inf. Retrieval*, 2007.
[33] Z. H. Zhou, P. Yu, C. Chelba, and F. Seide, "Towards spoken-document retrieval for the internet: Lattice indexing for large-scale web-search architectures," in *Proc. Human Lang. Technol. Conf. North Amer. Chap. Assoc. Comput. Linguist.*, 2006, pp. 415–422.
[34] V. Lavrenko and W. Croft, "Relevance based language models," in *Proc. Annual Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2001, pp. 120–127.
[35] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum, "Topics in semantic representation," *Psychol. Rev.*, vol. 114, pp. 211–244, 2007.
[36] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, pp. 177–196, 2001.
[37] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.
[38] B. Chen, "Latent topic modeling of word co-occurrence information for spoken document retrieval," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 3961–3964.
[39] G. Y. Chen, H. S. Chiu, and B. Chen, "Latent topic modeling of word vicinity information for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 5394–5397.
[40] H.-M. Wang, B. Chen, J.-W. Kuo, and S.-S. Cheng, "MATBN: A Mandarin Chinese broadcast news corpus," *Int. J. Comput. Linguist. Chinese Lang. Process.*, vol. 10, pp. 219–236, 2005.
[41] A. Stolcke, "SRILM—An extensible language modeling toolkit," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2005, pp. 901–904.
[42] 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.
[43] S. H. Lin, Y. M. Chang, J. W. Liu, and B. Chen, "Leveraging evaluation metric-related training criteria for speech summarization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 5314–5317.
[44] C. Y. Lin, "ROUGE: Recall-oriented understudy for gisting evaluation," 2003. [Online]. Available: http://haydn.isi.edu/ROUGE/
[45] F. Liu and Y. Liu, "Exploring correlation between ROUGE and human evaluation on meeting summaries," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 187–196, Jan. 2010.
[46] S. H. Lin and B. Chen, "A risk minimization framework for extractive speech summarization," in *Proc. Annu. Meeting Assoc. Linguist.*, 2010, pp. 79–87.

Shih-Hsiang Lin (S'07) received the M.S. degree in information and computer education from National Taiwan Normal University, Taipei, Taiwan, in 2007. He is currently pursuing the Ph.D. degree in the Department of Computer Science and Information Engineering, National Taiwan Normal University.

His research interests are in the field of large-vocabulary continuous speech recognition, natural language processing, and information retrieval.

Mr. Lin is a student member of the ISCA and ACLCLP.

Yao-Ming Yeh received the B.S. degree in computer engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1981, the M.S. degree in computer science and information engineering from National Taiwan University, Taipei, Taiwan, and the the Ph.D. degree in the Department of Electrical and Computer Engineering, Pennsylvania State University, University Park, PA, in 1991.

In 1991, he joined the Graduate Institute of Information and Computer Education, National Taiwan Normal University, Taipei. He is currently a Professor in the Department of Computer Science and Information Engineering, National Taiwan Normal University. His research interests include parallel processing, fault-tolerant computing, web computing, and speech-based applications.

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, 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 the Graduate Institute of Computer Science and Information Engineering, National Taiwan Normal University, Taipei. He is currently a Professor in the Department of Computer Science and Information Engineering of the same university. His research interests generally lie in the areas of speech recognition, information retrieval, and machine learning.

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