A Risk-Aware Modeling Framework for Speech Summarization

Berlin Chen, *Member, IEEE*, and Shih-Hsiang Lin, *Student Member, IEEE*

***Abstract*—Extractive speech summarization attempts to select a representative set of sentences from a spoken document so as to succinctly describe the main theme of the original document. In this paper, we adapt the notion of risk minimization for extractive speech summarization by formulating the selection of summary sentences as a decision-making problem. To this end, we develop several selection strategies and modeling paradigms that can leverage supervised and unsupervised summarization models to inherit their individual merits as well as to overcome their inherent limitations. On top of that, various component models are introduced, providing a principled way to render the redundancy and coherence relationships among sentences and between sentences and the whole document, respectively. A series of experiments on speech summarization seem to demonstrate that the methods deduced from our summarization framework are very competitive with existing summarization methods.**

***Index Terms*—Decision-making, language modeling, loss functions, risk minimization, speech summarization.**

I. Introduction

HUGE volumes of multimedia data are continuously filling up our computers, networks, and daily lives. Automatic summarization that facilitates users to quickly digest the important information conveyed by either a single or a cluster of documents plays an ever-increasing role in managing the multimedia content [1]. Due to the maturity of text summarization [2], this realm of research has been extended to speech summarization over the years [3]–[8]. Speech summarization is inevitably faced with the problem of incorrect information caused by recognition errors when using automatic speech recognition (ASR) techniques to transcribe the spoken documents into text forms. However, speech summarization also presents opportunities that do not exist for text summarization; for example, information cues about prosody/acoustics and emotion/speakers can help the determination of the importance and structure of spoken documents [4], [9].

Manuscript received November 04, 2010; revised February 23, 2011 and May 27, 2011; accepted May 30, 2011. Date of publication June 16, 2011; date of current version November 09, 2011. This work was sponsored in part by "Aim for the Top University Plan" of National Taiwan Normal University and Ministry of Education, Taiwan, and the National Science Council, Taiwan, under Grants NSC 99-2221-E-003-017-MY3, NSC 98-2221-E-003-011-MY3, NSC 100-2515-S-003-003, and NSC 99-2631-S-003-002. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gokhan Tur.

B. Chen is with the Department of Computer Science and Information Engineering, National Taiwan Normal University, Taipei 116, Taiwan (e-mail: berlin@csie.ntnu.edu.tw).

S.-H. Lin is with the Voice Division Research Center, Delta Electronics, Taipei 11491,Taiwan (e-mail: shlin@csie.ntnu.edu.tw).

Digital Object Identifier 10.1109/TASL.2011.2159596

Broadly speaking, a summary can be either abstractive or extractive [2]. In abstractive summarization, a fluent and concise abstract that reflects the key concepts of a document is generated, whereas in extractive summarization, the summary is usually formed by selecting salient sentences from the original document. The former requires highly sophisticated natural language processing (NLP) techniques, including semantic representation and inference, as well as natural language generation, while this would make abstractive approaches difficult to replicate or extend from constrained domains to more general domains. In addition to being abstractive or extractive, a summary may also be generated by considering several other aspects like being generic or query-oriented summarization, single-document or multi-document summarization, and so forth. Interested readers are encouraged to refer to [2] for an excellent and entertaining overview of document summarization. In this paper, we focus exclusively on generic, extractive speech summarization since it usually constitutes the essential building block for many other speech summarization tasks.

A spoken sentence to be selected as part of a summary may be considered from the following three factors (although one can still tackle the extractive summarization problem from a different point of view): 1) *salience*—the importance of the sentence itself, which is usually evident by its structure, location, prosodic or word-usage information, and many more; 2) *relevance*—the more relevant a sentence to the whole document or the other sentences in the document, the more likely it should be included in the summary; and 3) *redundancy*—the information carried by the sentence and that of the already selected summary sentences should cover different topics or concepts of the document. Quite a few studies with either supervised or unsupervised machine-learning methods have been designed to address the above three factors to a certain extent. For the salience factor, a typical example is to estimate the salience of each spoken sentence with supervised machine-learning techniques. It can be thought of as a two-class (i.e., summary and non-summary) sentence-classification problem [10]: a sentence with a set of indicative features is fed to the classifier (or summarizer) and a classification result is then output from it in view of these features. Summary sentences are subsequently ranked and compiled according to those classification results. Although such supervised summarizers are effective, most of them usually explicitly assume that sentences are independent of each other and each sentence is classified individually without allowing for the relationship among the sentences (the so-called "*bag-of-sentences*" assumption) [11]. The other potential shortcoming is that a set of handcrafted document-reference summary exemplars are required for training the summarizers; however, such

summarizers tend to limit their generalization capability and might not be readily applicable for new tasks or domains.

There is another school of thought that attempts to conduct document summarization using unsupervised machine-learning approaches, getting around the need for manual annotation of training data. The common basic idea behind these summarizers is typically based on the conception of the relevance (or similarity) of a sentence to other sentences [12]. Put simply, sentences bearing more similarity to the document itself (or the other sentences in the document) are deemed more relevant to the main theme of the document; such sentences thus will be selected as part of the summary. Moreover, unsupervised summarizers are usually constructed only on the basis of the lexical information without considering other sources of information, whereas imperfect speech recognition often leads to degraded performance when using the lexical information solely. On the other hand, for the last factor, *redundancy*, maximum marginal relevance (MMR) [13] is usually considered to be a good remedy. MMR performs sentence selection iteratively by striking the balance between topic relevance and coverage.

We have recently introduced a new perspective on the problem of speech summarization, saying that it can be approached with a modeling framework built on the notion of risk minimization [14], [15], which shows good promise to inherit the merits of most existing summarization methods, as well as to provide a general and flexible way to allow for the aforementioned three factors. Our work in this paper continues this general framework of research in several significant aspects: 1) we investigate leveraging several selection strategies and modeling paradigms to construct the component models involved in such a framework; 2) we explore various ways to devise the loss functions that can effectively render the dependence relationship among the sentences of a spoken document to be summarized; 3) more extra information cues are incorporated into the summarization framework that can further enhance the summarization performance; 4) we also provide extensive analysis and a series of experiments, showing that the methods deduced from such a risk-aware modeling framework are indeed very competitive with the existing speech summarization methods; and 5) we again confirm the added benefit of using non-lexical features for speech summarization.

The remainder of this paper is structured as follows. We begin by giving a brief review of the related work on extractive summarization in Section II, with a focus on supervised and unsupervised machine-learning methods. In Section III, we describe how to cast extractive speech summarization as a risk minimization problem, followed by a detailed elucidation of the proposed methods in Section IV. After that, the experimental setup and several sets of experiments and associated discussions are presented in Sections V and VI, respectively. Finally, Section VII concludes our presentation and suggests avenues for future work.

II. Related Work

Speech summarization can be conducted using either supervised or unsupervised machine-learning methods. In the following, we briefly review a few celebrated machine-learning methods that have been applied to speech summarization with varying degrees of success, as well as some other considerations pertaining to spoken documents.

A. Supervised Summarizers

The supervised machine-learning approaches usually treat speech summarization as a two-class (summary and non-summary) sentence-classification problem: A spoken sentence S_i is characterized by a set of indicative features, such as lexical features [16], structural features [17], acoustic features [16], discourse features [18], relevance features [19], etc. Then, the corresponding feature vector X_i of S_i is taken as the input to the classifier. If the output (classification) score belongs to the positive class, S_i will be selected as part of the summary; otherwise, it will be excluded [10]. Specifically, the problem can be formulated as follows: Construct a sentence ranking model that assigns a classification score (or a posterior probability) of being in the summary class to each sentence; important sentences are subsequently ranked and selected according to these scores. To this end, several popular machine-learning methods could be utilized to serve the purpose, like Bayesian classifier (BC) [10], Gaussian mixture model (GMM) [20], hidden Markov model (HMM) [21], support vector machine (SVM) [22], maximum entropy (ME) [23], conditional random field (CRF) [11], [24], to name a few.

In general, these methods require a training set comprised of several documents and their corresponding handcrafted summaries to train the classifiers (summarizers). However, manual annotation is often expensive in terms of time and personnel. Moreover, such summarizers tend to limit their generalization capability and might not be readily applicable for new tasks or domains. Another major shortcoming of these summarizers is that most of them usually implicitly assume that sentences are independent of each other (or the so-called "*bag-of-sentences*" assumption) and classify each sentence individually without leveraging the dependence relationship among the sentences or the global structure of the document [11].

B. Unsupervised Summarizers

The unsupervised summarization approaches usually rely on some heuristic rules or statistical evidences (such as word occurrence statistics) between each sentence and the document, without recourse to manual annotation of training data. Most previous studies conducted along this line revolve around the conception of sentence centrality [12], [25]–[27]. That is, sentences more similar to others are deemed more relevant to the main theme of the document; such sentences thus will be selected as part of the summary. For example, the vector space model (VSM) approach represents each sentence of a document and the document itself as vectors in the index term space [12], and computes the relevance score between each sentence and the document (e.g., the cosine measure of the proximity between two vectors). Then, the sentences with the highest relevance scores are included in the summary. A natural extension is to represent each document and each sentence as vectors in a latent semantic space, instead of simply using the literal term information as that done by VSM. On the other hand, the graph-based methods, such as LexRank [25] and TextRank [27], 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 or topical similarity relationship between a pair of nodes. Document summarization thus exploits the global structural information conveyed by such conceptualized network, rather than merely considering the local features of each node or sentence.

However, due to the lack of document-summary reference pairs, the performance of the unsupervised summarizers might be worse than that of the supervised summarizers, but their domain-independent and easy-to-implement properties still make them attractive. Moreover, most of the unsupervised summarizers are usually constructed solely on the basis of the lexical information without leveraging other sources of information cues, and the imperfect speech recognition results often lead to severely degraded performance [19].

C. Spoken Documents

Most of the above-mentioned methods can be equally applied to both text and speech summarization [3], [4]; 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. It has been shown that speech recognition errors are the dominating factor for the performance degradation of speech summarization when using speech recognition transcripts instead of manual transcripts, whereas erroneous sentence boundaries cause relatively minor problems [5], [19]. As an illustration, it has been shown that when the speech recognition error rate is in the range between 20% and 40%, a severe performance drop of about 50% is encountered for summarizing broadcast news speech with speech recognition transcripts [19]. To relieve this problem, we may develop techniques that can robustly represent the spoken documents as a straightforward remedy, apart from the many approaches to improving speech recognition accuracy. For example, multiple recognition hypotheses, beyond the top scoring ones, are expected to provide alternative representations for the confusing portions of the spoken documents. A recent trend in speech summarization has been to pursue different ways for robustly representing the recognition hypotheses of spoken documents, such as the use of word lattices, confusion networks, and N-best lists [28], [29].

In addition, prosodic (acoustic) features, e.g., intonation, pitch, formant, energy, and pause duration, can provide important clues for speech summarization. Some recent work has revealed that exploring more non-lexical features such as the prosodic features is beneficial for speech summarization especially when the speech recognition accuracy is not perfect [16], [18], [30], although reliable and efficient ways to use such features still await further studies. The summaries of spoken documents can be presented in either text or speech form. 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 [3].

III. Risk-Aware Summarization Framework

Extractive summarization can be alternatively viewed as a decision-making process in which the summarizer attempts to select a representative subset of sentences from the original documents. Among the several analytical methods that can be employed for the decision-making process, the Bayes decision theory, which quantifies the tradeoff between various decisions and the potential cost that accompanies each decision [31], is perhaps the most suited one that can be used to guide the summarizer in choosing a course of action in the face of some uncertainties underlying the decision-making process. In formal terms, a decision-making problem may consist of four basic elements: 1) an observation O from a random variable $\mathbf{O}$; 2) a set of possible decisions (or actions) $a \in \mathbf{A}$; 3) the state of nature $\theta \in \mathbf{\Theta}$ which denotes the possible states existing in the problem; and 4) a loss function $L(a, \theta)$ which specifies the cost associated with a chosen decision a given that θ is the true state of nature. As an illustration, for a binary classification problem, the state of nature θ can be either "the positive class ($\theta = 1$)" or "the negative class ($\theta = 0$)," while the decision a means the class assignment for a particular observation. The expected risk associated with taking decision a is given by

$$R(a|O) = \int_{\theta} L(a, \theta) p(\theta|O) d\theta \tag{1}$$

where $p(\theta|O)$ is the posterior probability of the state of nature being θ given the observation O. The Bayes decision theory states that the optimum decision can be made by contemplating each action a, and then choosing the action a^* for which the expected risk is minimum:

$$a^* = \arg\min_{a} R(a|O). \tag{2}$$

The notion of minimizing the Bayes risk have recently attracted much attention and been applied with success to many natural language processing tasks, such as automatic speech recognition (ASR) [32], machine translation (MT) [33], and information retrieval (IR) [34]. However, as far as we are aware, this notion has never been extensively explored for either text or speech summarization.

Along this same vein, we formulate extractive speech summarization as a Bayes risk minimization problem in this paper. Without loss of generality, let us denote $\pi \in \mathbf{\Pi}$ as one of possible selection strategies which comprises a set of indicators used to address the importance of each sentence S_i in a document D to be summarized. For notational convenience, we refer to the kth action a_k as choosing the kth selection strategy π_k, and the observation O as the document D to be summarized. The expected risk of a certain selection strategy π_k is given by

$$R(\pi_k|D) = \int_{\pi} L(\pi_k, \pi) p(\pi|D) d\pi. \tag{3}$$

Consequently, the ultimate goal of extractive summarization could be stated as the search of the best selection strategy π^* from the space of all possible selection strategies that minimizes the expected risk defined as follows:

$$\begin{aligned} \pi^* &= \arg\min_{\pi_k} R(\pi_k|D) \\ &= \arg\min_{\pi_k} \int_{\pi} L(\pi_k, \pi) p(\pi|D) d\pi. \end{aligned} \tag{4}$$

As can be seen in (4), the realization of the Bayes decision theory for extractive speech summarization requires: 1) a practical definition of the selection strategy π; 2) an efficient and accurate way to estimate the probability of choosing a particular selection strategy π given D (i.e.,$p(\pi|D)$); and 3) an effective mechanism to measure the loss function between any two selection strategies (i.e., $L(\pi_k, \pi)$). In what follows, we will shed light on each of these three ingredients from various points of view.

A. Selection Strategy

A feasible selection strategy can be fairly arbitrary according to the underlying principle. For example, it could be a set of binary indicators denoting whether a sentence should be selected as part of summary or not. In addition, it may also be a ranked list used to address the importance degree of each individual sentence. Here, we present two different instantiations of it where the selection strategy can be either "*sentence-wise*" or "*list-wise*."

1) Sentence-Wise Selection Strategy: For the sentence-wise selection strategy, we assume that summary sentences can be iteratively chosen (i.e., one at each iteration) from the original document until the aggregated summary reaches a predefined target summarization ratio. More concretely, the selection strategy is represented by a binary decision vector, of which each element corresponds to a specific sentence S_i in the document D and designates whether it should be selected as part of the summary or not. It turns out that the binary vector for each possible action will have just one element equal to 1 and all the others equal to zero (or the so-called "*one-of-n*" coding). For ease of notation, we denote the binary vector by S_i when the ith element has a value of 1. Therefore, (4) can be reduced to

$$\begin{aligned} S^* &= \arg\min_{S_i \in \widetilde{D}} R(S_i|\widetilde{D}) \\ &= \arg\min_{S_i \in \widetilde{D}} \sum_{S_j \in \widetilde{D}} L(S_i, S_j) P(S_j|\widetilde{D}) \end{aligned} \tag{5}$$

where $\widetilde{D}$ denotes the remaining sentences that have not been selected into the summary yet (i.e., the "*residual*" document); $P(S_j|\widetilde{D})$ reflects the importance degree of a sentence S_j given the residual document $\widetilde{D}$.

2) List-Wise Selection Strategy: The iterative (or greedy) selection procedure described above may sometimes result in a suboptimal selection. For example, the information carried by a verbose sentence would be succinctly depicted by one or more other concise (short) sentences which cover more topics of interest. To address this potential shortcoming, one may formulate the extractive summarization as a maximum convergence problem under a summary length constraint and solve the problem by exploiting some global inference algorithms, such as the integer linear programming (ILP) [35], [36] or graph-based submodular selection [37] methods. We, however, present here an alternative remedy to address the issue by exploring the so-called list-wise selection strategy under the risk minimization framework. Specifically, we contemplate every possible combination (or subset) of sentences in a spoken document as a candidate summary ψ_j and then the best summary can be constructed through the following equation:

$$Summary = \arg\min_{\psi_i \in \mathbf{\Psi}_D} \sum_{\psi_j \in \mathbf{\Psi}_D} L(\psi_i, \psi_j) P(\psi_j|D) \tag{6}$$

where $\mathbf{\Psi}_D$ denotes all possible combinations of sentences in a spoken document D (i.e., the set of all possible candidate summaries); $P(\psi_j|D)$ is the probability of ψ_j being the summary given the document D.

For practical implementation, it would be 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 in a document increases. To reduce the computational overhead, we can first use some prior knowledge, for example, the sentence-wise selection strategy, to select a set of possible summary sentences 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 constraint of the length of the target summary.

B. Evidence Modeling

The posterior probability $P(\pi|D)$ of a particular selection strategy π given the document D could be estimated from two different schools of thought: generative- and direct- modeling paradigms [38]. Each school has its own advantages and has shown promise in many NLP applications. Here, we illustrate how these two modeling paradigms can be adopted in the presented risk-aware summarization framework.

1) Generative Modeling: In generative modeling, the posterior probability $P(\pi|D)$ is evaluated through the data generation process. The basic idea behind this line of research assumes that the data is drawn from some parameterized probabilistic models and prior beliefs. By the application of Bayes' rule, the posterior probability $P(\pi|D)$ can be further decomposed as

$$P(\pi|D) = \frac{P(D|\pi)P(\pi)}{P(D)} \tag{7}$$

where $P(D|\pi)$ is the generative probability given a particular selection strategy π, $P(\pi)$ is the prior probability of π and $P(D)$ is the marginal probability of D, which, for example, can be approximated by the following expression when the possible selection strategies are countable:

$$P(D) = \sum_{\pi' \in \mathbf{\Pi}} p(D|\pi')P(\pi'). \tag{8}$$

2) Direct Modeling: Direct modeling, on the other hand, focuses on learning (or estimating) the probability of a direct mapping from an input variable (e.g., D) to an output variable (e.g., π). In other words, the posterior probability $P(\pi|D)$, viewed as a kind of discriminative model, is estimated directly without recourse to an intermediate step that explicitly represents the data generation process as done by generative modeling. It should be noted that direct modeling usually demands the training data equipped with labeled information for learning the associated

parameters. Interested readers may refer to [38] for a thorough discussion of generative modeling and direct modeling.

C. Loss Function

The loss function $L(\pi_i, \pi_j)$ introduced in the proposed risk-aware summarization framework is to measure the relationship between any pair of selection strategies. For example, in the sentence-wise selection strategy, when a given sentence is more dissimilar from most of the other sentences, it may incur higher loss as it is taken as the representative sentence (or summary sentence) to represent the main theme embedded in the other ones. Consequently, the loss function can be built on the grounds of the similarity measure. Here, we take the sentence-wise selection strategy as an example to illustrate how to measure the relationship between any pair of sentences through the design of meaningful loss functions $L(S_i, S_j)$, while the loss functions for the list-wise selection strategy can be constructed in the same spirit.

1) VSM Loss Function: We may fist represent each sentence S_i in vector form, where each dimension specifies the weighted statistic $z_{t,i}$, e.g., the product of the term frequency (TF) and inverse document frequency (IDF) scores associated with a word w_t, reflecting its importance to S_i. Then, the cosine similarity measure $Sim(S_i, S_j)$ is used to estimate the relevance between any given two selection strategies S_i and S_j (it is assumed that the relevance between two selection strategies is correlated with the similarity between them):

$$Sim(S_i, S_j) = \frac{\sum_{t=1}^{V} z_{t,i} \times z_{t,j}}{\sqrt{\sum_{t=1}^{V} z_{t,i}^2} \times \sqrt{\sum_{t=1}^{V} z_{t,j}^2}} \tag{9}$$

where V is the number of distinct words in the vocabulary. The loss function is thus defined by

$$L_{\mathrm{VSM}}(S_i, S_j) = 1 - Sim(S_i, S_j). \tag{10}$$

This means that $L_{\mathrm{VSM}}(S_i, S_j)$ is reversely proportional to the similarity measure $Sim(S_i, S_j)$ between sentences S_i and S_j.

2) KL-Divergence Loss Function: We may assume that if two sentences S_i and S_j are similar to each other, words w in each of them should be drawn from the same probability distribution. Therefore, we can use the KL-divergence measure, which assesses the relationship between any pair of probability distributions from a rigorous information-theoretic perspective, to quantify how close any two sentences S_i and S_j are [39]

$$L_{\mathrm{KL}}(S_i, S_j) = \sum_{w \in \mathbf{w}} P(w|S_j) \log \frac{P(w|S_j)}{P(w|S_i)} \tag{11}$$

where w denotes a specific word in the vocabulary set $\mathbf{w}$. It should be borne in mind that the closer the sentence generative model $P(w|S_i)$ (see Section IV for more details about the parameter estimation) to the sentence generative model $P(w|S_j)$, the more likely S_i is relevant to S_j. Therefore, (11) is a kind of the probability distance between the sentence generative models of S_i and S_j.

IV. Implementation

In this section, we elaborate a few practical implementation and technical details involved in the risk-aware summarization framework.

A. Generative Modeling

1) Generative Probability: We explore the language modeling (LM) approach, which has been introduced in a wide range of IR tasks and demonstrated with good empirical success, to predict the generative probability $P(D|\pi)$, as shown in (7). In the LM approach, each selection strategy π (virtually, π may correspond to a sentence for the sentence-wise selection strategy, or a subset of possible summary sentences for the list-wise selection strategy) can be simply regarded as a probabilistic model for predicting the document. If we further assume that words are conditionally independent given π and their order is of no importance (i.e., the so-called "*bag-of-words*" assumption), then $P(D|\pi)$ can be decomposed as a product of unigram probabilities of words w generated by π

$$P(D|\pi) = \prod_{w \in D} P(w|\pi)^{c(w,D)} \tag{12}$$

where $c(w, D)$ is the number of times that index term (or word) w occurs in D, reflecting that w will contribute more in the calculation of $P(D|\pi)$ if it occurs more frequently in D. The simplest way is to estimate the probabilistic model $P(w|\pi)$ on the basis of the frequency of word w occurring in π, with the maximum-likelihood estimation (MLE)

$$P(w|\pi) = \frac{c(w, \pi)}{|\pi|} \tag{13}$$

where $c(w, \pi)$ is the number of times that word w occurs in π and $|\pi|$ is the number of words in π. In a sense, (12) belongs to a kind of literal term matching strategy and may suffer the problem of unreliable model estimation owing particularly to only a few sampled words present in π [39]. To mitigate this potential problem, a unigram probability (or background model) $P(w|BG)$ estimated from a general collection, which models the generic characteristics of words in the target language, is often used to smooth the generative model:

$$\hat{P}(w|\pi) = \lambda \cdot P(w|\pi) + (1 - \lambda) \cdot P(w|BG) \tag{14}$$

where λ is a weighting parameter. Interested readers may also refer to [39] an in-depth treatment of more elaborate ways to construct the generative model.

As an illustration, consider the sentence-wise selection strategy where the calculation of the probability $P(D|\pi)$ is equivalent to the calculation of sentence generative probability $P(D|S_j)$ if π corresponds to any arbitrary sentence S_j. The probability $P(D|S_j)$ can be interpreted as the likelihood of the (residual) document D being generated by S_j. If S_j generate D with a higher likelihood, it would be more likely to be a summary sentence. Phrased another way, $P(D|S_j)$ captures the degree of relevance of S_j to D. Following the same spirit, we can also use $P(D|\psi)$ to measure the similarity of a candidate summary (namely, a subset of possible summary sentences) ψ to D for the list-wise selection strategy.

2) Prior Probability: The prior probability $P(\pi)$, as shown in (7), can be regarded as the likelihood of a selection strategy π being important without seeing the whole document. It could be assumed uniformly distributed or estimated from a wide variety of factors, such as the positional information, the lexical information, the structural information or the inherent prosodic properties embedded in a sentence (or a subset of sentences) of the spoken document to be summarized.

Taking the sentence-wise selection strategy as an example, a straightforward way is to assume that the sentence prior probability $P(S_j)$ is set in proportion to the posterior probability of a sentence S_j being included in the summary class when observing a set of indicative features X_j of S_j derived from its structure, location, prosodic and word-usage information, or other sentence importance measures [10]. These features can be integrated in a systematic way into the proposed framework by taking the advantage of the learning capability of the supervised machine-learning methods. Specifically, the prior probability $P(S_j)$ can be approximated by

$$P(S_j) \approx \frac{p(X_j|\mathbf{S})P(\mathbf{S})}{P(X_j|\mathbf{S})P(\mathbf{S}) + P(X_j|\overline{\mathbf{S}})P(\overline{\mathbf{S}})} \tag{15}$$

where $P(X_j|\mathbf{S})$ and $P(X_j|\overline{\mathbf{S}})$ are the likelihoods that a sentence S_j with features X_j are generated by the summary class $\mathbf{S}$ and the non-summary class $\overline{\mathbf{S}}$, respectively; the prior probability $P(\mathbf{S})$ and $P(\overline{\mathbf{S}})$ are set to be equal in this research. To estimate $P(X_j|\mathbf{S})$ and $P(X_j|\overline{\mathbf{S}})$, several popular supervised classifiers (or summarizers), like BC, can be employed for this purpose.

On the other hand, the prior probability of each candidate summary $P(\psi)$ in the list-wise selection strategy can be estimated in a similar way as the sentence-wise selection strategy, or, alternatively, by considering the informativeness, clarity or redundancy of the constituent elements in the candidate summary (i.e., the subset of possible summary sentences). However, since in this research, the list-wise selection strategy is implemented by a two-stage selection procedure by first using the sentence-wise selection strategy to select a set of summary sentences to form the combinations of sentences as the possible candidate summaries, each candidate summary ψ, to a certain degree, is presumably representative enough. Thus, we might simply assume that the prior probability of each candidate summary $P(\psi)$ is uniformly distributed.

B. Direct Modeling

To accomplish direct modeling of the posterior probability $P(\pi|D)$, in this study, we employ the global conditional log-linear model (GCLM) [40] to fulfill this goal. In GCLM, the posterior probability of an output T given an observation (or input) R is represented by

$$P(T|R;\boldsymbol{\alpha}) = \frac{1}{Z(R,\boldsymbol{\alpha})} \exp\left(\Phi(T,R)\cdot\boldsymbol{\alpha}\right) \tag{16}$$

where $\Phi(T,R)$ is a feature vector used to characterize the relationship between T and R, $\boldsymbol{\alpha}$ is the corresponding parameter vector, $\Phi(T,R)\cdot\boldsymbol{\alpha}$ is the dot product of $\Phi(T,R)$ and $\boldsymbol{\alpha}$, and $Z(R,\boldsymbol{\alpha}) = \sum_{T'} \exp(\Phi(T',R)\cdot\boldsymbol{\alpha})$ is a normalization factor that depends on R and $\boldsymbol{\alpha}$. In the context of speech summarization, for a spoken document D and a given selection strategy π (i.e., π is a possible summary sentence for the sentence-wise selection strategy, or a subset of possible summary sentences for the list-wise selection strategy) associated with its feature vector X_π, the posterior probability $p(\pi|D)$ thus can be represented by

$$P(\pi|D;\boldsymbol{\alpha}) = \frac{1}{Z(D,\boldsymbol{\alpha})} \exp(X_\pi\cdot\boldsymbol{\alpha}) \tag{17}$$

where $Z(D,\boldsymbol{\alpha}) = \sum_{\pi'\in\boldsymbol{\Pi}} \exp(X_{\pi'}\cdot\boldsymbol{\alpha})$ and $X_{\pi'}$ is the feature vector for an arbitrary selection strategy π'. Given a set of training documents with their corresponding reference-summary information, the parameter vector $\boldsymbol{\alpha}$ can be estimated with the stochastic gradient-descent method [40].

C. Relation to Other Summarization Methods

In this subsection, we illustrate the relationship between our summarization framework (especially taking the pairing of the sentence-wise selection strategy and the generative modeling paradigm as an example) and a few existing summarization approaches. We start by considering a special case where a 0–1 loss function is used in (5), namely, the loss function will take value 0 if the two sentences are identical, and 1 otherwise. Then, (5) can be alternatively represented by

$$\begin{aligned} S^* &= \arg\min_{S_i\in\widetilde{D}} \sum_{S_j\in\widetilde{D}, S_j\neq S_i} \frac{P(\widetilde{D}|S_j)P(S_j)}{\sum_{S_m\in\widetilde{D}} P(\widetilde{D}|S_m)P(S_m)} \\ &= \arg\max_{S_i\in\widetilde{D}} \frac{P(\widetilde{D}|S_i)P(S_i)}{\sum_{S_m\in\widetilde{D}} P(\widetilde{D}|S_m)P(S_m)} \\ &= \arg\max_{S_i\in\widetilde{D}} P(\widetilde{D}|S_i)P(S_i) \end{aligned} \tag{18}$$

which actually provides a natural integration of a supervised summarizer (i.e., $P(S_j)$) and an unsupervised summarizer (i.e., $P(\widetilde{D}|S_j)$), as mentioned previously.

If we further assume the prior probability $P(S_j)$ is uniformly distributed, the important (or summary) sentence selection problem has now been reduced to the problem of measuring the sentence generative probability $P(\widetilde{D}|S_j)$, or the relevance between the document and the sentence. By the same token, the important sentences of a document can be selected (or ranked) solely based on the prior probability $P(S_j)$ with the assumption of an equal sentence generative probability $P(\widetilde{D}|S_j)$.

V. Experiments Setup

A. Data

The summarization dataset employed in this study is a broadcast news corpus collected by the Academia Sinica and the Public Television Service Foundation of Taiwan between November 2001 and April 2003 [41], which 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 compiled between November 2001 and August 2002 was reserved for the summarization experiments. We chose 20 documents as the held-out test set while the remaining 185 documents as the training set. Twenty-five hours

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

	Training Set	Evaluation Set
Recording Period	Nov. 07, 2001 – Jan. 22, 2002	Jan. 24, 2002 – Aug. 20, 2002
Number of Documents	100	20.0
Average Duration per Document (in sec.)	129.4	141.3
Avg. Number of words per Document	326	290.3
Avg. Number of Sentences per Document	20.0	23.3
Avg. Character Error Rate	34.4%	36.2%

of gender-balanced speech from the remaining speech data were used to train the acoustic models for speech recognition. The data was first used to bootstrap the acoustic model training with the ML criterion. Then, the acoustic models were further optimized by the minimum phone error (MPE) discriminative training algorithm [42]. The average Chinese character error rate (CER) obtained for the 205 spoken documents was about 35%. Table I shows some basic statistics about the 205 spoken documents.

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 [43]. 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 employed to estimate the background model $P(w|BG)$ that is used to smooth the generative model $P(w|\pi)$ in (14).

B. Performance Evaluation

Three subjects were asked to create summaries of the 205 spoken documents for the summarization experiments as references (the gold standard) for evaluation. The reference summaries were generated by ranking the sentences in the manual transcript of a spoken document by importance without assigning a score to each sentence. For the assessment of summarization performance, we adopted the widely-used ROUGE measure [44]. It evaluates the quality of the summarization 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 summaries. Three variants of the ROUGE measure were used to quantify 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 [44].

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 research. Since increasing the summary length tends to increase the chance of getting higher scores in the recall rate of the various ROUGE measures and might not always select the right number of informative words in the automatic summary as compared to the reference summary, all the experimental results reported hereafter are obtained by calculating the F-scores of these ROUGE measures. Table II shows the levels of agreement (the Kappa statistic and ROUGE measures) 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 summary sentences for representing a given document.

TABLE II
AGREEMENT AMONG THE SUBJECTS FOR IMPORTANT SENTENCE RANKING FOR THE EVALUATION SET

Kappa	ROUGE-1	ROUGE-2	ROUGE-L
0.400	0.600	0.532	0.527

TABLE III
BASIC SENTENCE FEATURES USED BY SUPERVISED SUMMARIZERS

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

C. Features for Supervised Summarizers

Several features have been designed and widely-used in speech summarization, especially with the supervised machine-learning approaches [10], [19]. In this paper, we take BC as the representative supervised summarizer and use a set of 28 indicative features, as outlined in Table III, to characterize a spoken sentence, including the structural features, the lexical features, the acoustic features and the relevance feature. Interested readers may refer to [19] for detailed accounts on the characteristics of these features, and comparisons among them. Also noteworthy is that, for each kind of acoustic features, the minimum, maximum, mean, difference value and mean difference value of a spoken sentence are extracted. The difference value is defined as the difference between the minimum and maximum values of the spoken sentence, while the mean difference value is defined as the mean difference between a sentence and its previous sentence. All the 28 features are further normalized to zero mean and unit variance:

$$\hat{x}_m = \frac{x_m - \mu_m}{\sigma_m} \tag{19}$$

where μ_m and σ_m are, respectively, the mean and standard deviation of a feature x_m estimated from the training set (cf. Section V-A).

TABLE IV
RESULTS ACHIEVED BY THE BC AND LM SUMMARIZERS, RESPECTIVELY

	Text Documents (TD)			Spoken Documents (SD)		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
BC	0.445 (0.390 - 0.504)	0.346 (0.201 - 0.415)	0.404 (0.348 - 0.468)	0.369 [*0.314*] (0.316 - 0.426)	0.241 [*0.174*] (0.183 - 0.302)	0.321 [*0.276*] (0.268 - 0.378)
LM	0.387 (0.302 - 0.474)	0.264 (0.168 - 0.366)	0.334 (0.251 - 0.415)	0.319 [*0.266*] (0.274 - 0.367)	0.164 [*0.106*] (0.115 - 0.224)	0.253 [*0.204*] (0.215 - 0.301)

TABLE V
RESULTS ACHIEVED BY SEVERAL METHODS DERIVED FROM THE SENTENCE-WISE SELECTION STRATEGY CONDUCTED IN CONJUNCTION WITH THE GENERATIVE-MODELING PARADIGM

	Text Documents (TD)			Spoken Documents (SD)		
Loss	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
0-1	0.501	0.401	0.459	0.417	0.281	0.356
SIM	0.524	0.425	0.473	0.475	0.351	0.420
KL	0.531	0.429	0.484	0.467	0.336	0.409
MMR	0.529	0.426	0.479	0.475	0.351	0.420

VI. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. Baseline Experiments

At the outset, we evaluate the performance of a special case of the risk-aware summarization framework when conducting speech summarization with both the sentence-wise selection strategy and the generative-modeling paradigm, but using either the sentence generative model (denoted by the LM summarizer) or the sentence prior model (denoted by the BC summarizer), exclusively (cf. Section IV-A). Here the loss function is simply set to the 0–1 loss function. The corresponding results are detailed in Table IV, where the values shown in the parentheses are the associated 95% confidence intervals. It is also worth mentioning that TD denotes the summarization results obtained based on the manual transcripts of spoken documents while SD denotes the results using the speech recognition transcripts which may contain speech recognition errors and sentence boundary detection errors. In this research, sentence boundaries were determined by speech pauses [19]. For the TD case, the acoustic features were obtained by aligning the manual transcripts to their spoken documents counterpart by performing word-level forced alignment.

Furthermore, the ROUGE measures, in essence, are evaluated by counting the number of overlapping units between the automatic summary and the reference summary; the corresponding evaluation results, therefore, would be severely affected by speech recognition errors when applying the various ROUGE measures to quantify the performance of speech summarization. In order to get rid of the confounding effect of this issue, it is assumed that the selected summary sentences are 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. Consequently, we align the speech recognition transcripts of the summary sentences to their respective audio segments to obtain the correct (manual) transcripts for the summarization performance evaluation (i.e., for the SD case). On the other hand, the results obtained by directly evaluating the automatic summary based on its associated speech recognition transcripts also are shown in the brackets of Table IV for comparison.

We observe two phenomena from Table IV. One is that there are significant performance gaps between conducting summarization based on the manual transcripts and the erroneous speech recognition transcripts. The relative performance degradations are about 15%, 34%, and 23%, respectively, for ROUGE-1, ROUGE-2, and ROUGE-L measures. One explanation is that the erroneous speech recognition transcripts of spoken sentences would probably carry wrong information and thus deviate somewhat from representing the true theme of the spoken document. The other is that the supervised summarizer (i.e., BC) outperforms the unsupervised summarizer (i.e., LM). The better performance of BC can be further explained by two reasons. One is that BC is trained with the handcrafted document-summary sentence labels in the training set, while LM is instead conducted in a purely unsupervised manner. Another is that BC utilizes a rich set of lexical and non-lexical features to characterize a given spoken sentence, while LM is constructed solely on the basis of the lexical (unigram) information.

B. Experiments on the Proposed Summarization Framework

We then turn our attention to investigate the utility of several methods deduced from the risk-aware summarization framework. We first evaluate the performance of the sentence-wise selection strategy conducted in conjunction with the generative-modeling paradigm [cf. (5) and (7)]. As can be seen from the first row of Table V, a simple combination of BC and LM [cf. (18)] can give about 4% to 5% absolute improvements as compared to the results of BC illustrated at the first row of Table IV. It shows the feasibility of combining the supervised with unsupervised summarizers. The result, to some extent, also confirms the complementary properties of lexical features and non-lexical features. Moreover, we consider the use of the loss functions defined in (10) (denoted by SIM) and (11) (denoted by KL), and the corresponding results are shown in the second and third rows of Table V, respectively. Consulting Table V, we notice two particularities. First, properly leveraging the loss function greatly boosts the summarization performance. Second, KL performs slightly worse than SIM for the SD case. A possible explanation is that the speech recognition errors will result in an inaccurate estimation of (13) since there are only a few words present in the erroneous speech transcription transcripts of spoken sentences.

TABLE VI
RESULTS ACHIEVED BY SEVERAL METHODS DERIVED FROM THE SENTENCE-WISE SELECTION STRATEGY CONDUCTED IN CONJUNCTION WITH THE GENERATIVE-MODELING PARADIGM AND WITH UNIFORM PRIOR PROBABILITY

	Text Documents (TD)			Spoken Documents (SD)		
Loss	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
0-1	0.387	0.264	0.334	0.319	0.164	0.253
SIM	0.405	0.281	0.348	0.365	0.209	0.305
KL	0.424	0.303	0.368	0.364	0.209	0.301
MMR	0.417	0.282	0.359	0.391	0.236	0.338

TABLE VII
RESULTS ACHIEVED BY SEVERAL METHODS DERIVED FROM THE SENTENCE-WISE SELECTION STRATEGY CONDUCTED IN CONJUNCTION WITH THE DIRECT-MODELING PARADIGM

	Text Documents (TD)			Spoken Documents (SD)		
Loss	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
0-1	0.512	0.415	0.463	0.423	0.294	0.360
SIM	0.513	0.417	0.470	0.430	0.300	0.371
KL	0.513	0.413	0.469	0.430	0.300	0.366
MMR	0.515	0.418	0.476	0.430	0.300	0.371

However, we believe that KL still has the merit of being able to accommodate more elaborate model estimation techniques to improve the performance in a systematic way (see Section VI-D for more details on this issue). Furthermore, one potential drawback of these two approaches is that they do not take the redundancy factor into account. To avoid such a problem of selecting a summary sentence having similar (redundant) information that is also contained in the already selected summary sentences, we may borrow the idea form the MMR method [13] and redefine the loss function shown in (10) as

$$L_{\mathrm{MMR}}(S_i, S_j) = 1 - \Big[\beta \cdot Sim(S_i, S_j) - (1-\beta) \cdot \max_{S' \in \mathbf{Summ}} Sim(S_i, S')\Big] \quad (20)$$

where $\mathbf{Summ}$ represents the set of sentences that have already been included into the summary and the novelty factor β is used to control the tradeoff between relevance and redundancy. That is, the loss function expressed in (20) is derived according to two criteria: 1) whether S_i is more similar (relevant) to S_j than the other sentences; and 2) whether S_i is less similar (relevant) to the set of summary sentences selected so far than the other sentences. As can be seen in the last row of Table V, it is evident that MMR delivers slightly higher summarization performance than SIM, which in turn verifies the merit of incorporating the MMR concept into the proposed framework for extractive summarization. If we further compare the results achieved by MMR with those of BC and LM as shown in Table IV, we can find significant improvements for both the TD and SD cases. To recap, for the TD case, this summarization method offers relative performance improvements of about 19%, 23%, and 19%, respectively, in the ROUGE-1, ROUGE-2, and ROUGE-L measures as compared to the BC baseline; while the relative improvements are 29%, 46%, and 31%, respectively, in the same measurements for the SD case. On the other hand, the performance gap between the TD and SD cases are reduced to a good extent by using the risk-aware summarization framework.

In the next set of experiments, we simply assume the sentence prior probability $P(S_j)$ is uniformly distributed; namely, we do not use any supervised information cue but use the unsupervised lexical information only:

$$S^* = \arg\min_{S_i \in \widetilde{D}} \sum_{S_j \in \widetilde{D}} L(S_i, S_j) P(\widetilde{D}|S_j). \quad (21)$$

The importance of a given sentence is thus considered from two angles: 1) the relationship between a sentence and the residual document (i.e., the sentence generative probability $P(\widetilde{D}|S_j)$), and 2) the relationship between the sentence and the other individual sentences (i.e., the value of the loss function $L(S_i, S_j)$). The corresponding results are illustrated in the Table VI. It should be noted that the coupling of the uniform prior probability and the 0–1 loss function is equivalent to the baseline LM approach. These results seem to reflect that the additional consideration of the "*sentence-sentence*" relationship is beneficial as compared to that considering only the "*document-sentence*" relevance information (cf. the second row of Table IV). It also gives competitive results as compared to the performance of BC for the SD case (cf. the first row of Table IV).

We then explore the performance of the sentence-wise selection strategy conducted in conjunction with the direct-modeling paradigm [cf. (17)] and the corresponding results are shown in Table VII. It is worth mentioning that we take the scores obtained by LM as an additional feature to augment the basic feature set (cf., Table III) for a fair comparison with the generative-modeling paradigm whose results are shown in Table V. Two observations can be made from Table VII. First, when a 0–1 loss function is being used, the summary sentences are selected solely based on the posterior probability of each sentence [cf. (17)]. As compared to the results shown in the first row of Table V, we can find that the direct-modeling paradigm seems to perform slightly better than the generative-modeling paradigm for all cases. Second, as can be seen from Table V, the use of the other loss functions, rather than the 0–1 loss function, provides moderate but consistent improvements. However, such improvements are not as apparent as those gains achieved by the sentence-wise selection strategy conducted in conjunction with the generative-modeling paradigm (cf. Table V). We speculate one possible reason is that the posterior probability estimated

TABLE VIII
RESULTS ACHIEVED BY SEVERAL METHODS DERIVED FROM THE LIST-WISE SELECTION STRATEGY CONDUCTED IN CONJUNCTION WITH THE GENERATIVE-MODELING PARADIGM

	Text Documents (TD)			Spoken Documents (SD)		
Loss	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
0-1	0.544	0.448	0.505	0.466	0.336	0.409
SIM	0.544	0.448	0.505	0.482	0.359	0.435
KL	0.527	0.427	0.486	0.486	0.364	0.436

by the direct-modeling paradigm is already good enough, which seems to overwhelm the added merit of using the various loss functions. However, the exact reason is still worthy of further investigation.

To go a step further, we evaluate the utility of the list-wise selection strategy conducted in conjunction with the generative-modeling paradigm, and the corresponding results are shown in Table VIII. As can be seen, the list-wise selection strategy consistently outperforms the sentence-wise selection strategy (cf. Table V) even under the assumption of uniformly distributed priors $P(\psi)$. In the case of the 0–1 loss function, the selection of summary sentences relies solely on the list generative probability. It can be also seen that the 0–1 loss function performs on par with SIM (and even better than KL) in the TD case, and the performance gaps between the 0–1 loss function and SIM (and between the 0–1 loss function and KL as well) are reduced to a certain extent in the SD case as compared to that of the sentence-wise counterparts shown in Table V. One reason for this may be that speech summarization using the list generative probability alone is quite good enough for the list-wise selection strategy, leading to that the contributions made by further incorporating the various loss functions (viz. SIM and KL) are less pronounced.

The above results seem to demonstrate that the list-wise selection strategy overcomes the problem of suboptimal performance faced by most of the current commonly-used sentence-wise selection strategies for extractive summarization. To better understand why it outperforms the sentence-wise selection strategy, we further analyze the average number of sentences respectively selected by these two strategies subject to the same length constraint. We observed the list-wise selection strategy selects about 5.1 sentences on average while the sentence-wise selection strategy selects about 4.1 sentences into the summary under the same word length constraint (i.e., 10% summarization ratio). In other words, these statistics reveal that the list-wise selection strategy can, to some extent, avoid selecting verbose sentences into the summary.

C. Comparison With Conventional Summarization Methods

Furthermore, we compare our proposed summarization methods with a few existing summarization methods that have been well-practiced in various summarization tasks, including LEAD, VSM [12], LexRank [25], ILP [35], SVM, and CRF [11]; the corresponding results for the SD case are shown in Table IX. It should be noted that the LEAD-based method simply extracts the first few sentences from a document as the summary. For ILP method, we follow [35] to define the associated cost function and constraints and use the "lpsolve" [45] software as the ILP solver to find the optimal solution. From Table IX, we observe the following remarks. First, to our surprise, CRF does not provide superior results as compared to the other summarization methods. One possible explanation is that the structural evidence of the spoken documents in the test set is not strong enough for CRF to show its advantage of modeling the global structural information among sentences. Second, LexRank gives a very promising performance in spite that it only utilizes lexical information in an unsupervised manner. This somewhat reflects the importance of capturing the global relationship for the sentences in the spoken document to be summarized. Third, ILP performs better than VSM. The results confirm the utility of the global inference algorithm for extractive summarization. By and large, the evidence accumulated so far seems to suggest that our proposed methods can provide substantial improvements compared to these conventional summarization methods.

TABLE IX
RESULTS ACHIEVED BY SEVERAL CONVENTIONAL SUMMARIZATION METHODS

Method		Spoken Documents (SD) ROUGE-1	ROUGE-2	ROUGE-L
Unsupervised Summarizer	LEAD	0.312	0.168	0.251
	VSM	0.337	0.189	0.277
	LexRank	0.348	0.204	0.294
	ILP	0.356	0.203	0.312
Supervised Summarizer	SVM	0.362	0.215	0.290
	CRF	0.358	0.220	0.291

TABLE X
SUBJECTIVE EVALUATION FOR THE AUTOMATIC SUMMARIES GENERATED BY THE BC SUMMARIZATION METHOD AND THE PROPOSED METHOD

	BC	Proposed Method
Informativeness	3.59 (0.25)	4.09 (0.18)
Readability	3.43 (0.30)	3.57 (0.24)

In addition, we further evaluate the summarization results obtained by the BC summarization method and the proposed summarization method (i.e., conducting the list-wise selection strategy in conjunction with the generative-modeling paradigm and with the KL loss function) by using the Mean Opinion Score (MOS) test. Six graduate students were invited to evaluate the automatic summaries given that the associated reference transcripts were provided. They were asked to judge the informativeness and readability of automatic summaries by assigning scores ranging from 1.0 to 5.0, where 5.0 represents the best rating and 1.0 represents the worst rating. The average results of the subjective evaluation are shown in Table X, where the numbers shown in the parentheses are the associated standard deviations of the scores. We can see that the quality of the proposed method is better than BC in terms of informativeness of automatic summaries. On the other hand, the average readability scores show no significant difference between these two methods. One possible explanation is that summary sentences are simply presented in the summary according to their original

order in the document without any further post-processing (e.g., reparagraphing or rephrasing). Hence, it would sometimes hurt the readability of automatic summaries.

D. Incorporation of Extra Information Cues

In the final set of experiments, we further advance the proposed risk-aware summarization framework by incorporating some extra information cues, including the use of multiple recognition hypotheses and the use of more training data, for achieving robust estimation of the sentence generative model and the loss function. Here, we exploit the position specific posterior lattices (PSPL) [46] to represent the spoken sentences of a document to be summarized, since it achieved the best performance over the other mechanisms in our previous work [28]. On the other hand, as mentioned previously, because there are only a few words present in a spoken sentence, it would be difficult to make reliable estimation of the sentence generative model with the MLE criterion when adopting the generative-modeling paradigm. A simple and intuitive way is to adopt the conception of relevance class [39], originally proposed in the context of IR, to facilitate accurate estimation of the sentence generative models. We may assume that each sentence S_j of the spoken document D to be summarized has its own associated relevance class R_{S_j}. This class is defined as the subset of documents in the collection that are relevant to the sentence S_j. The relevance model (RM) of the sentence S_j is therefore defined to be the probability distribution $P(w|R_{S_j})$, 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_j} and then pick up a random word from that document. Once the relevance model of the sentence R_{S_j} is constructed, it can be used to replace the original sentence generative model or to be combined with the original sentence generative model to produce a more accurate estimate. Since there is no prior knowledge about the subset of relevant documents for each sentence S_j, a local relevance feedback-like procedure can be employed by taking S_j as a query and posing it to an IR system to obtain a ranked list of documents from a large document repository (here we take the text news document collection mentioned in Section V-A as the repository). The top M documents returned from the IR system are assumed to be the ones relevant to S_j, and can be therefore used to approximate the relevance model for S_j.

Taking the sentence-wise selection strategy as an illustration, Table XI shows the summarization results when respectively using the PSPL and RM cues and with some specific settings of the sentence prior model and the loss function. Comparing to the results (with the same corresponding settings) shown in Tables V and VI, it turns out that incorporating these two kinds of information cues provides additional performance gains for speech summarization. We may thus expect that exploring more extra information cues and sophisticated modeling paradigms, such as rhetorical information [47], topic modeling [48]–[50], to name a few, will enhance the component models of the proposed risk-aware summarization framework.

TABLE XI
RESULTS ACHIEVED BY SEVERAL METHODS DERIVED FROM THE SENTENCE-WISE SELECTION STRATEGY CONDUCTED IN CONJUNCTION WITH THE GENERATIVE-MODELING PARADIGM AND SOME EXTRA INFORMATION CUES

			Spoken Documents (SD)		
Extra Cue	Prior	Loss	ROUGE-1	ROUGE-2	ROUGE-L
PSPL	Uniform	0-1	0.352	0.198	0.285
	BC	0-1	0.444	0.313	0.387
RM	Uniform	KL	0.389	0.236	0.327
	BC	KL	0.486	0.365	0.426

VII. CONCLUSION AND FUTURE WORK

In this paper, we have presented a risk-aware modeling framework for extractive speech summarization, which has the capability to select summary sentences in a sentence-wise manner or in a list-wise manner, in conjunction with a generative-modeling paradigm or a direct-modeling paradigm. Furthermore, we have also demonstrated how to systematically integrate several existing summarization methods into the proposed framework. The empirical results show that our proposed methods substantially boost the summarization performance when compared to a number of popular summarization methods. It is worth emphasizing that the list-wise selection strategy conducted in conjunction with the generative-modeling paradigm achieved the best results for speech summarization using either the manual transcripts or the speech recognition transcripts of spoken documents. How to implement the list-wise selection strategy more efficiently would be worthy of future investigation. We list below some other possible future extensions: 1) exploring more extra information cues and sophisticated modeling paradigms for this framework; 2) investigating various discriminative training criteria for training the component models of this framework, 3) extending and applying the proposed framework to multi-document summarization tasks, and 4) incorporating the summarization results into audio indexing for better retrieval and browsing of spoken documents.

REFERENCES

[1] M. Ostendorf, "Speech technology and information access," *IEEE Signal Process. Mag.*, vol. 25, no. 3, pp. 150–152, May 2008.
[2] I. Mani and M. T. Maybury, *Advances in Automatic Text Summarization*. Cambridge, MA: MIT Press, 1999.
[3] 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.
[4] 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.
[5] 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.
[6] G. Penn and X. Zhu, "A critical reassessment of evaluation baselines for speech summarization," in *Proc. Annu. Meeting Assoc. Comput. Linguist.*, 2008, pp. 470–478.
[7] S. Xie and Y. Liu, "Improving supervised learning for meeting summarization using sampling and regression," *Comput. Speech Lang.*, vol. 24, pp. 495–514, 2010.
[8] J. J. Zhang, R. H. Y. Chan, 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.
[9] J. Zhang and P. Fung, "Speech summarization without lexical features for Mandarin broadcast news," in *Proc. Human Lang. Technol. Conf. and North Amer. Chapt. Assoc. for Comput. Linguist. Annu. Meeting*, 2007.
[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] 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.
[12] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in *Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2001, pp. 19–25.
[13] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1998, pp. 335–336.
[14] S.-H. Lin and B. Chen, "A risk minimization framework for extractive speech summarization," in *Proc. Annu. Meeting Assoc. Comput. Linguist.*, 2010, pp. 79–87.
[15] S.-H. Lin, Y.-M. Yeh, and B. Chen, "Extractive speech summarization—From the view of decision theory," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 1684–1687.
[16] 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.
[17] S. Maskey and J. Hirschberg, "Automatic summarization of broadcast news using structural features," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2003, pp. 1147–1157.
[18] J. Zhang, H. Y. Chan, P. Fung, and L. Cuo, "A comparative study on speech summarization of broadcast news and lecture speech," in *Annu. Conf. Int. Speech Commun. Assoc.*, 2007, pp. 2781–2784.
[19] 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, no. 1, pp. 3:1–3:23, 2009.
[20] M. A. Fattah and F. Ren, "GA, MR, FFNN, PNN and GMM based models for automatic text summarization," *Comput. Speech Lang.*, vol. 23, no. 1, pp. 126–144, 2009.
[21] J. M. Conroy and D. P. O'learyc, "Text summarization via hidden Markov models," in *Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2001, pp. 406–407.
[22] A. Kolcz, V. Prabakarmurthi, and J. Kalita, "Summarization as feature selection for text categorization," in *Proc. Conf. Inf. Knowl. Manage.*, 2001, pp. 365–370.
[23] L. Ferrier, "A maximum entropy approach to text summarization," School of Artif. Intell., Univ. of Edinburgh, 2001.
[24] M. Galley, "A skip-chain conditional random field for ranking meeting utterances by importance," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2006, pp. 364–372.
[25] 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.
[26] D. R. Radev, H. Jing, M. Stys, and D. Ta, "Centroid-based summarization of multiple documents," *Inf. Process. Manage.*, vol. 40, pp. 919–938, 2004.
[27] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2005, pp. 404–411.
[28] 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.
[29] Y. Liu, S. Xie, and F. Liu, "Using N-best recognition output for extractive summarization and keyword extraction in meeting speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 5310–5313.
[30] S. Maskey and J. Hirschberg, "Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2005, pp. 621–624.
[31] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag, 1985.
[32] V. Goel and W. Byrne, "Minimum Bayes-risk automatic speech recognition," *Comput. Speech Lang.*, vol. 14, pp. 115–135, 2000.
[33] S. Kumar and W. Byrne, "Minimum bayes-risk decoding for statistical machine translation," in *Proc. Human Lang. Technol. Conf. North Amer. Chap. Assoc. Comput. Linguist. Annu. Meeting*, 2004.
[34] C. X. Zhai and J. Lafferty, "A risk minimization framework for information retrieval," *Inf. Process. Manage.*, vol. 42, no. 1, pp. 31–55, 2006.
[35] R. McDonald, "A study of global inference algorithms in multi-document summarization," in *Proc. Eur. Conf. Inf. Retrieval*, 2007.
[36] K. Riedhammer, D. Gillick, and B. Favre, "Long story short—Global unsupervised models for keyphrase based meeting summarization," *Speech Commun.*, vol. 52, no. 10, pp. 801–815, 2010.
[37] H. Lin, J. Bilmes, and S. Xie, "Graph-based submodular selection for extractive summarization," in *Proc. IEEE Workshop Autom. Speech Recogn. Understanding*, 2009, pp. 381–386.
[38] C. M. Bishop and J. Lasserre, "Generative or discriminative? Getting the best of both worlds," *Bayesian Statist.*, vol. 8, pp. 3–24, 2007.
[39] C. X. Zhai, *Statistical Language Models for Information Retrieval*. San Rafael, CA: Morgan & Claypool, 2008.
[40] B. Roark, M. Saraclar, and M. Collins, "Discriminative n-gram language modeling," *Comput. Speech Lang.*, vol. 21, no. 2, pp. 373–392, 2007.
[41] 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, no. 2, pp. 219–236, 2005.
[42] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, pp. 105–108.
[43] A. Stolcke, "SRILM—An extensible language modeling toolkit," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2005, pp. 901–904.
[44] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Workshop Text Summarizat. Branches Out*, 2004.
[45] M. Berkelaar, K. Eikland, and P. Notebaert, lp_solve 5.5, [Online]. Available: http://lpsolve.sourceforge.net/
[46] 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.
[47] J. J. Zhang, R. H. Y. Chan, 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.
[48] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum, "Topics in semantic representation," *Psychol. Rev.*, vol. 114, pp. 211–244, 2007.
[49] Y. Lu, Q. Mei, and C. X. Zhai, "Investigating task performance of probabilistic topic models—An empirical study of PLSA and LDA," *Inf. Retrieval*, 2010.
[50] B. Chen, "Word topic models for spoken document retrieval and transcription," *ACM Trans. Asian Lang. Inf. Process.*, vol. 8, no. 1, pp. 2:1–2:27, 2009.

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 and natural language processing, information retrieval, and artificial intelligence. He is the author/coauthor of over 100 academic publications.

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

Shih-Hsiang Lin (S'07) received the M.S. degree in information and computer education and the Ph.D. degree in computer science and information engineering from National Taiwan Normal University, Taipei, Taiwan, in 2007 and 2011, respectively.

He received the IBM Ph.D. fellowship award in 2010. He was also a research summer intern at IBM T. J. Watson Research Center in 2010. He is currently a Principal Engineer in the Voice Division Research Center, Delta Electronics, Taipei, Taiwan. His research interests are in the fields of large-vocabulary continuous speech recognition, natural language processing, and information retrieval.

Dr. Lin received the IBM Ph.D. Fellowship Award in 2010.