CHAPTER 13

SPOKEN DOCUMENT RETRIEVAL AND SUMMARIZATION

Berlin Chen,[†] Hsin-Min Wang[‡] and Lin-Shan Lee[§]
[†]*National Taiwan Normal University, Taipei*
[‡]*Academia Sinica, Taipei*
[§]*National Taiwan University, Taipei*
E-mail: berlin@csie.ntnu.edu.tw, whm@iis.sinica.edu.tw, lslee@gate.sinica.edu.tw

Huge, continually increasing quantities of multimedia content including speech information are filling up our computers, networks and lives. It is obvious that speech is one of the most important sources of information for multimedia content, as it is the speech of the content that tells us of the subjects, topics and concepts. As a result, the associated spoken documents of the multimedia content will be key for content retrieval and browsing. Substantial efforts along with very encouraging results for spoken document transcription, retrieval, and summarization have been reported. This chapter presents a concise yet comprehensive overview of information retrieval and automatic summarization technologies that have been developed in recent years for efficient spoken document retrieval and browsing applications. An example prototype system for voice retrieval of Chinese broadcast news collected in Taiwan will be introduced as well.

1. Introduction

Speech is the primary and most convenient means of communication between humans.[1] In the future of networks, digital content over the network will include all the information relating to our daily life activities, from real-time information to knowledge archives, from work environments to private services. Naturally, the most attractive form of content is multimedia, including speech which carries the information that tells us of the subjects, topics and concepts of the multimedia content. As a result, the spoken documents associated with the network content will be key in retrieval and browsing activities.[2]

At the same time, the rapid development of network and wireless technologies is making it possible for people to access network content not only from offices and homes, but from anywhere, at any time with the use of small, hand-held devices such as personal digital assistants (PDAs) and cell phones. Today, our access to the network is primarily text-based. Users need to enter instructions by keying in words or texts, and the network or search engine in turn offers text materials for the user to select. These users therefore interact with the network or search engine and

obtain the desired information via the text-based media. In the future, almost all text functions can be performed with speech. The users' instructions can be entered with speech just as well. Speech is a convenient user interface suitable for all the different kinds of devices and it is especially good for smaller, hand-held devices. The network content may be indexed, retrieved and browsed not only by text, but also by their associated spoken documents as well. Users may also interact with the network or the search engines by means of either text-based media or spoken, multi-modal dialogues. Text-to-speech synthesis can then be used to transform textual information in the content into speech when needed.

This chapter presents a concise yet comprehensive overview of the information retrieval and automatic summarization technologies that have been developed in recent years for efficient spoken document retrieval and browsing applications. An example prototype system for voice retrieval of Chinese broadcast news collected in Taiwan will be introduced as well.

2. Information Retrieval

We will start with a brief review of information retrieval (IR). In the past two decades, most of the research in IR focused on text document retrieval, and the Text REtrieval Conference[3] (TREC) evaluations in the nineties are good examples. In conventional text document retrieval, a collection of documents $D = \{d_i, i = 1, 2, \ldots, N\}$ are to be retrieved by a user's query Q. This retrieval is based on a set of indexing terms specifying the semantics of the documents and the query, which are very often a set of keywords, or even all the words used in all the documents. The document retrieval problem can thus be viewed as a clustering problem, i.e., selecting the documents out of the collection which are in the class relevant to the query Q. The documents are usually ranked by a retrieval model (or ranking algorithm) based on the relevance scores between each of the documents d_i and the query Q evaluated with the indexing terms. In this way, those documents on the top of the list are most likely to be relevant. The retrieval models are usually characterized by two different matching strategies, namely, literal term matching and concept matching. These two strategies are briefly reviewed below.

2.1. *Literal Term Matching*

The vector space model (VSM) is the most popular model for literal term matching.[4] In VSM, every document d_i is represented as a vector $\vec{d_i}$. Each component $w_{i,t}$ in this vector is a value associated with the statistics of a specific indexing term (or word) t, both within the document d_i and across all the documents in the collection D,

$$w_{i,t} = f_{i,t} \cdot \ln\left(N/N_t\right), \tag{1}$$

where $f_{i,t}$ is the normalized term frequency (TF) for the term (or word) t in d_i, used to measure the intra-document weight for the term (or word) t; while $\ln\left(N/N_t\right)$ is

the inverse document frequency (IDF), where N_t is the total number of documents in the collection which include the term t, and N is the total number of documents in the collection D. IDF is to measure the inter-document discrimination ability for the term t, reflecting the fact that indexing terms appearing in more different documents are less useful in identifying the relevant documents. The query Q is also represented by a vector $\vec{Q}$ constructed in exactly the same way, i.e., with components $w_{q,t}$ in exactly the same form as in Equation 1. The cosine measure is then used to estimate the query-document relevance scores:

$$R\left(Q, d_i\right) = \left(\vec{Q} \cdot \vec{d_i}\right) / \left(\parallel \vec{Q} \parallel \cdot \parallel \vec{d_i} \parallel\right), \tag{2}$$

which apparently matches Q and d_i based on the terms literally. This model has been widely used because of its simplicity and satisfactory performance.

Literal term matching can also be performed with probabilities, the n-gram-based[5] and hidden Markov model (HMM)-based[6] approaches being good examples of this. In these models, each document d_i is interpreted as a generative model composed of a mixture of n-gram probability distributions for observing a query Q, while the query Q is considered as observations, expressed as a sequence of indexing terms (or words) $Q = t_1 t_2 \ldots t_j \ldots t_J$, where t_j is the j-th indexing term in Q and J is the length of the query, as illustrated in Figure 1. The n-gram distributions for the terms t_j, for example $P\left(t_j|d_i\right)$ and $P\left(t_j|t_{j-1}, d_i\right)$ for uni- and bigrams, are estimated from the document d_i and then linearly interpolated with the background uni- and bigram models estimated from a large outside (i.e., not part of the set used for training) text corpus C, $P\left(t_j|C\right)$ and $P\left(t_j|t_{j-1}, C\right)$. The relevance score for a document d_i and the query Q can then be expressed as, with uni- and bigram models,

$$P\left(Q|d_i\right) = \left[m_1 \cdot P\left(t_1|d_i\right) + m_2 \cdot P\left(t_1|C\right)\right] \cdot \prod_{j=2}^{J} \left[m_1 \cdot P\left(t_j|d_i\right) + m_2 \cdot P\left(t_j|C\right) + m_3 \cdot P\left(t_j|t_{j-1}, d_i\right) + m_4 \cdot P\left(t_j|t_{j-1}, C\right)\right], \tag{3}$$

which again matches Q and d_i based on the terms literally. The uni- and bigram probabilities, as well as the weighting parameters, $m_1, \ldots, m_4$, can be further optimized, for example, by the expectation-maximization (EM) or minimum classification error (MCE) training algorithms, given a training set of query examples with the corresponding query-document relevance information.[7]

2.2. *Concept Matching*

Both approaches mentioned above are based on matching terms (or words), which makes them face the problem of word usage diversity (or vocabulary mismatch) very often. This happens when the query and its relevant documents are using rather different sets of words. In contrast, the concept matching strategy tries to discover the latent topical information inherent in the query and the documents on which the

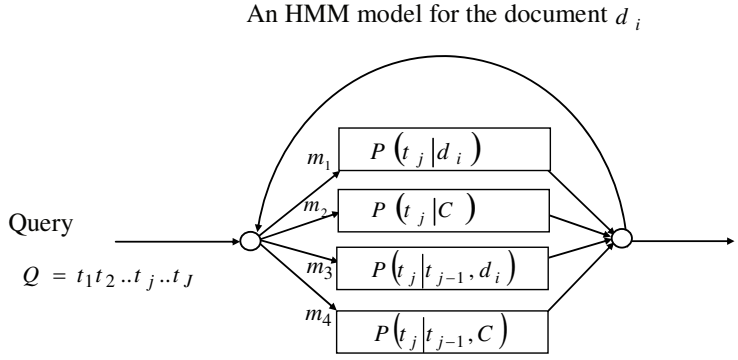

Fig. 1. An illustration of HMM-based retrieval model.

retrieval is to be done. The latent semantic indexing (LSI) model is a good example employing this strategy.[8,9] LSI starts with a "term-document" matrix W, describing the intra- and inter-document statistical relationships between all the terms and all the documents in the collection D, in which each term t is characterized by a row vector and each document d_i in D by a column vector of W. Singular value decomposition (SVD) is then performed on the matrix W in order to project all the term vectors and document vectors onto a single latent semantic space with significantly reduced dimensionality L:

$$W \approx \hat{W} = U\Sigma V^T \tag{4}$$

where $\hat{W}$ is the rank-L approximation to the "term-document" matrix W; U is the right singular matrix; Σ is the $L \times L$ diagonal matrix of the L singular values; V is the right singular matrix; and T denotes matrix transposition. In this way, the row/column vectors representing the terms/documents in the original matrix W can all be mapped to the vectors in the same latent semantic space with dimensionality L. As shown in Figure 2, in this latent semantic space, each dimension is defined by a singular vector and represents some kind of latent semantic concept. Each term t and each document d_i can now be properly represented in this space, with components in each dimension having to do with the weights of the term t and document d_i with respect to the dimension, or the associated latent semantic concept. While for the query Q or other documents that are not represented in the original analysis, they can be folded-in, i.e., similarly represented in this space, via some simple matrix operations. In this way, indexing terms describing related concepts will be close to each other in the latent semantic space even if they never co-occur in the same document, and the documents describing related concepts will be close to each other in the latent semantic space even if they do not contain the same set of words. So this is concept matching rather than literal term matching. The relevance score

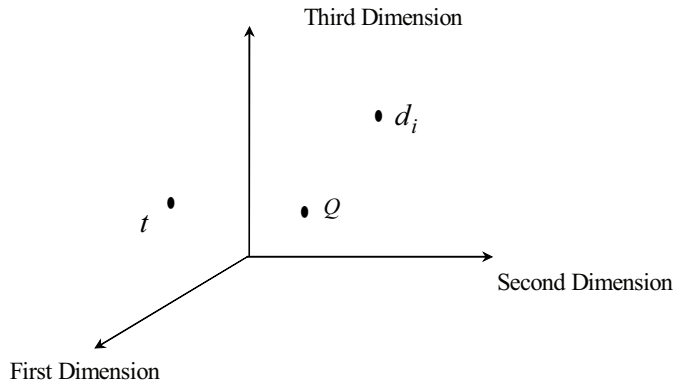

Fig. 2. Three-dimensional schematic representation of the latent semantic space and the LSI retrieval model.

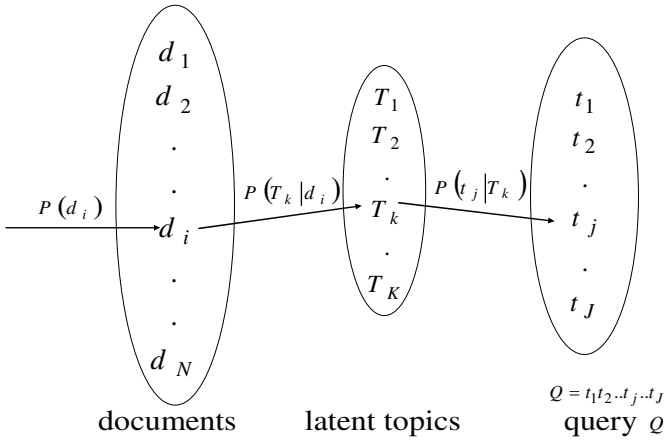

Fig. 3. Graphical representation of the PLSA-based retrieval model.

between the query Q and a document d_i is then estimated by computing the cosine measure between the corresponding vectors in this latent semantic space.

In recent years, new attempts have been made to establish probabilistic frameworks for the above latent topical approach. They include improved model training algorithms and the probabilistic latent semantic analysis (PLSA or aspect model),[10,11] which is often considered as a representative of this category. PLSA introduces a set of latent topic variables, $\{T_k, k = 1, 2, \ldots, K\}$, to characterize the "term-document" co-occurrence relationships, as shown in Figure 3. A query Q is again treated as a sequence of observed terms (or words), $Q = t_1 t_2 \ldots t_j \ldots t_J$, while the document d_i and a term t_j are both assumed to be independent conditioned on an associated latent topic T_k. The conditional probability of a document d_i

generating a term t_j thus can be parameterized by

$$P\left(t_j|d_i\right) = \sum_{k=1}^{K} P\left(t_j|T_k\right) \cdot P\left(T_k|d_i\right) \tag{5}$$

When the terms in the query Q are further assumed to be independent given the document, the relevance score between the query and document can then be expressed as:

$$P\left(Q|d_i\right) = \prod_{j=1}^{J} \left[\sum_{k=1}^{K} P\left(t_j|T_k\right) \cdot P\left(T_k|d_i\right) \right] \tag{6}$$

Notice that this relevance score is not obtained directly from the frequency of the respective query term t_j occurring in d_i, but instead through the frequency of t_j in the latent topic T_k as well as the likelihood that d_i generates the latent topic T_k. A query and a document thus may have a high relevance score even if they do not share any terms in common, which is therefore concept matching. The PLSA model can be trained in an unsupervised way by maximizing the total log-likelihood L_T of the document collection $\{d_i, i = 1, 2, \ldots, N\}$ in terms of the unigram $P\left(t_j|d_i\right)$ of all terms t_j observed in the document collection, using the EM algorithm:

$$L_T = \sum_{i=1}^{N} \sum_{j=1}^{N'} c\left(t_j, d_i\right) \cdot \log P\left(t_j|d_i\right) \tag{7}$$

where N is total number of documents in the collection, N' is the total number of different terms observed in the document collection, $c\left(t_j, d_i\right)$ is the frequency count for the term t_j in the document d_i, and $P\left(t_j|d_i\right)$ is the probability obtained above in Equation 5.

2.3. *Spoken Documents and Queries*

All the retrieval models mentioned above can in fact be equally applied to text or spoken documents with text or spoken queries. The additional, albeit important, difficulties for spoken documents and queries are the inevitable speech recognition errors: problems of spontaneous speech such as pronunciation variation as well as disfluencies, and the out-of-vocabulary (OOV) problem for words outside the vocabulary of the speech recognizer. A principal approach to the former, apart from the many approaches improving recognition accuracy, is to develop more robust indexing terms for audio signals. For example, multiple recognition hypotheses obtained from M-best lists, word graphs, or "sausages" can provide alternative representatives for the confusing portions of the spoken query or documents.[12] Improved scoring methods using different confidence measures, for example, posterior probabilities incorporating acoustic and language model likelihoods, or other measures considering relationships between the recognized word hypotheses,[13,14] as well as prosodic features including pitch, energy stress and duration measure,[15] can also

help to weight the term hypotheses properly. The use of subword units – for example, phonemes for English[13] and syllables for Chinese,[12,16] or segments of them – rather than words as indexing terms mentioned above has also been shown to be very helpful. Special considerations of using syllable-level indexing features for Chinese spoken document retrieval will be discussed in the next section. In addition, another set of approaches try to expand the representation of the query and documents not only using conventional IR techniques such as pseudo relevance feedback,[17] but based on the acoustic confusion statistics and/or semantic relationships among the word- or subword-level terms derived from some training corpus, and these have been shown to be very helpful as well.[14]

3. Considerations Of Using Syllable-Level Indexing Features For Chinese Spoken Document Retrieval

3.1. *Characteristics of the Chinese Language*

In the Chinese language, because every one of the large number of characters (at least 10,000 commonly used) is pronounced as a monosyllable and is itself a morpheme with its own meaning, new words are very easily generated everyday by combining a few characters or syllables. For example, the combination of the characters "電 (electricity)" and "腦 (brain)" gives us the rather new Chinese word "電腦 (computer)", and the combination of "股 (stock)", "市 (market)", "長 (long)", and "紅 (red)" gives the business domain the word "股市長紅 (the market remains bullish for long)". In many cases, the meanings of these new words are somewhat related to the meaning of the component characters. Examples of such new words also include many proper nouns such as personal names and organization names which are simply arbitrary combinations of a few characters, as well as many domain-specific terms, like in the above examples. Many of such words are very often the focus in IR functions, because they typically carry the core information, or characterize the subject topic. But in many cases these important words for retrieval purposes are simply not included in any lexicon. It is therefore believed that the OOV problem is a particularly important issue for Chinese IR, and this makes using syllable-level statistical characteristics attractive, logical and even necessary to deal with this problem.

Actually, the syllable-level information makes great sense for the retrieval of Chinese information due to the largely monosyllabic structure of the language. Although there are more than 10,000 commonly used Chinese characters, an elegant feature of the Chinese language is that all its characters are monosyllabic and the total number of phonologically allowed Mandarin syllables is only 1,345. So a syllable is usually shared by many homonym characters with completely different meanings. Each Chinese word is then composed of one to several characters (or syllables), thus the combination of these 1,345 syllables actually gives an almost unlimited number of Chinese words. In other words, each syllable may stand for many different characters with different meanings, while the combination of several specific syllables

very often gives only very few, if not unique, homonym polysyllabic words. As a result, comparing the input query and the documents to be retrieved based on the segments of several syllables may provide a very good measure of relevance between them.

In fact, there are other important reasons to use syllable-level information. We know that almost every Chinese character is a morpheme with its own meaning, and each of them have quite independent linguistic roles. As a result, the construction of Chinese words from its characters is indeed rather flexible. To illustrate this phenomenon, in many cases, different words describing the same or similar concepts can be constructed by slightly different combinations of characters. For example, both "中華文化 (Chinese culture)" and "中國文化 (Chinese culture)" have the same meaning, but the second characters in these two words are different. Another realization of this different-characters-same-meaning phenomenon is that a longer word can be arbitrarily abbreviated into shorter words, as in "國家科學委員會 (National Science Council)", which can be abbreviated into "國科會", with the same referent. The shorter word is made up of only the first, the third and the last characters of the first word. Furthermore, exotic words from foreign languages are very often translated into different Chinese words based on its pronunciation. To illustrate, "Kosovo" may be translated into "科索沃 /ke1-suo3-wo4/", "柯索佛 /ke1-suo3-fo2/", "克索夫 /ke1-suo3-fu1/," and so on, but these words usually have some syllables in common, or they can even have exactly the same syllables. Therefore, an intelligent IR system needs to be able to handle such word or terminological flexibilities, such that when the input queries include some words in one form, the desired spoken documents can be retrieved even if they include the corresponding words in different forms. The comparison between the spoken queries and the spoken documents directly at the syllable-level does allow for such flexibilities to some extent, since the "words" are not necessarily constructed during the retrieval processes, while the different forms of words describing the same or relevant concepts very often do have some syllables in common.

3.2. *Syllable-level Indexing Terms*

A whole class of syllable-level indexing terms were proposed by Chen *et al.*,[12] including overlapped syllable segments with length u $(A(u), u = 1, 2, 3, 4, 5)$ and syllable pairs separated by v syllables $(B(v), v = 1, 2, 3, 4)$. Considering a syllable sequence of 10 syllables $s_1 s_2 s_3 \ldots s_{10}$, examples of syllable segments are listed on the upper half of Table 1, while examples of syllable pairs on the lower half of the same table. For example, syllable segments of length $u = 3$ include such segments as $(s_1\ s_2\ s_3), (s_2\ s_3\ s_4)$, etc., while syllable pairs separated by $v = 1$ syllables include such pairs as $(s_1\ s_3), (s_2\ s_4)$, etc.

Considering the structural features of the Chinese language, combinations of these indexing terms are beneficial for the retrieval process. For example, as mentioned previously, each syllable represents some characters with their respective

Table 1. Various syllable-level indexing terms for an example syllable sequence $s_1\ s_2\ s_3 \ldots s_{10}$.

Syllable Segments	Examples
$A(u), u = 1$	$(s_1)(s_2)\ldots(s_{10})$
$A(u), u = 2$	$(s_1s_2)(s_2s_3)\ldots(s_9s_{10})$
$A(u), u = 3$	$(s_1s_2s_3)\ (s_2s_3s_4)\ldots(s_8s_9s_{10})$
$A(u), u = 4$	$(s_1s_2s_3s_4)\ (s_2s_3s_4s_5)\ldots(s_7s_8s_9s_{10})$
$A(u), u = 5$	$(s_1s_2s_3s_4s_5)\ (s_2s_3s_4s_5s_6)\ldots(s_6s_7s_8s_9s_{10})$
Syllable Pair Separated by v Syllables	Examples
$B(v), v = 1$	$(s_1s_3)\ (s_2s_4)\ldots(s_8s_{10})$
$B(v), v = 2$	$(s_1s_4)\ (s_2s_5)\ldots(s_7s_{10})$
$B(v), v = 3$	$(s_1s_5)\ (s_2s_6)\ldots(s_6s_{10})$
$B(v), v = 4$	$(s_1s_6)\ (s_2s_7)\ldots(s_5s_{10})$

meanings, and very often words with similar or relevant concepts have some syllables in common. Therefore syllable segments with length $u = 1$ makes sense in retrieval process. However, because each syllable is also shared by many homonymic characters, the syllable segments with length $u = 1$ may also cause ambiguity. Therefore it has to be combined with other indexing terms. On the other hand, more than 90% of most frequently used Chinese words are bi-syllabic,[12] so the syllable segments with length $u = 2$ definitely carry a plurality of linguistic information which are definitely useful as important indexing terms. Similarly, if longer syllable segments with $u = 3$ are matched between a document and the query, very often this brings about very important information for purposes of retrieval. On the other hand, because of the very flexible wording structure of Chinese, syllable pairs separated by v syllables are helpful in retrieval. For example, when the word "國家科學委員會 (National Science Council)" is abbreviated by including only the first, third and the last characters, syllable pairs separated by v syllables start to become useful. Furthermore, because substitution, insertion and deletion errors are inevitable and frequent during the recognition process, such indexing terms as syllable pairs separated by v syllables can also help to alleviate these problems.

3.3. *Information Fusion Using Word- and Syllable-Level Indexing Terms*

The characteristics of the Chinese language also lead to some special considerations for the spoken document retrieval task. That is, word-level indexing features possess more semantic information than syllable-level features; hence, word-based retrieval does enhance retrieval precision. Syllable-level indexing features behave more robustly in the areas of the Chinese word tokenization ambiguity issue, the abbreviation problem, the open vocabulary problem, and speech recognition errors, as mentioned above. Therefore, syllable-based retrieval enhances recall. Accordingly, there is good reason to fuse the information obtained from indexing the features of multiple levels. It has been shown that syllable-level indexing features are very

effective for Chinese spoken document retrieval, and retrieval performance can be improved further by integrating information from word-level indexing features.[12,16]

4. Spoken Document Summarization

Spoken document summarization, which aims at distilling important information and removing redundant and incorrect information from spoken documents, enables us to efficiently review spoken documents and understand their associated topics quickly. Although research into the automatic summarization of text documents dates back to the early 1950s, for nearly four decades, research work suffered from a lack of funding. However, the development of the World Wide Web led to a renaissance in the field and summarization was subsequently extended to cover a wider range of tasks, including multi-document, multilingual and multimedia summarization.[18] Document summarization in general can be either *extractive* or *abstractive*. Extractive summarization tries to select a number of indicative sentences, passages or paragraphs from the original document according to a target summarization ratio, and then sequence them together to form a summary. Abstractive summarization, on the other hand, tries to produce a concise abstract of desired length that can reflect the key concepts of the document. The latter appears to be more difficult, and recent approaches have been focusing more on the former. The approaches for extractive spoken document summarization have been in principle developed on the basis of either statistical models or probabilistic generative models. These two kinds of models will be briefly reviewed below in Sections 4.1 and 4.2, respectively; while special considerations of spoken document summarization will be briefly discussed in Section 4.3.

4.1. *Statistical Models*

As one example, the vector space model (VSM), originally formulated for IR, can be used to respectively represent each sentence of the document, as well as the whole document, in a vector form. Within the VSM, each dimension specifies the weighted statistics associated with an indexing term (or word) in the sentence or document, and the sentences that have the highest relevance scores (e.g., in the cosine measure) to the whole document are selected to be included in the summary. When the intended summary aims to cover the more important concepts as well as the different ones within or among documents, after the first sentence with the highest relevance score is selected, indexing terms in that sentence can be removed from the document. The document vector is then reconstructed accordingly, based on which the next sentence can be selected, and so on.[19] The latent semantic analysis (LSA) model for IR is another example of a model that can be used to represent each sentence of a document as a vector in the latent semantic space for that document. This space is constructed by performing SVD on the "term-sentence" matrix for that document. The right singular vectors with larger singular values represent dimensions for more

important latent semantic concepts in that document. Therefore, the sentences that have the largest index values in each of the top L right singular vectors are included in the summary.[19] A third statistical approach is carried out as follows: indicative sentences can be chosen from the document based on the sentence significance score (denoted as the *SenSig* model below). Given a sentence $S = \{t_1, t_2, \ldots t_j, \ldots t_J\}$ with length J, the sentence significance score $Sig\,(S)$ can be expressed using the following formula:

$$Sig\,(S) = \frac{1}{J}\sum_{j=1}^{J}\left[\beta_1 \cdot I\,(t_j) + \beta_2 \cdot F\,(t_j)\right] \tag{8}$$

where $I\,(t_j)$ is evaluated based on some statistical measure of term t_j (such as a product of term frequency (TF) and inverse document frequency (IDF)); $F\,(t_j)$ can be a linguistic measure of t_j (e.g., named entities and different parts-of-speech (POSs) are given different weights, ignoring function words); and β_1 and β_2 are tunable weighting parameters.[20] These selected sentences in all the above cases can also be further condensed and shortened by removing the less important terms, if a higher compression ratio is desired.

4.2. *Probabilistic Generative Models*

Extractive document summarization also can be performed with probabilistic generative models.[21,22] For example, the HMM model originally formulated in IR can be applied to extractive spoken document summarization.[22] Each sentence S of a spoken document d_i is instead treated as a probabilistic generative model (or an HMM) consisting of n-gram distributions for predicting the document, and the terms (or words) in the document d_i are taken as an input observation sequence. The HMM model for a sentence can be expressed as the following using unigram modeling:

$$P_{HMM}\,(d_i|S) = \prod_{t_j \in d_i}\left[\lambda \cdot P\,(t_j|S) + (1-\lambda)\cdot P\,(t_j|C)\right]^{c(t_j,d_i)} \tag{9}$$

where λ is a weighting parameter, and $c\,(t_j, d_i)$ is the occurrence count of a term t_j in d_i. For each sentence HMM, the sentence model $P\,(t_j|S)$ and the collection model $P\,(t_j|C)$ can be simply estimated, respectively, from each sentence itself and a large text collection based on the maximum likelihood estimation (MLE). The weighting parameter λ can be further optimized by taking the document d_i as the training observation sequence and using the following EM training formula:

$$\hat{\lambda} = \frac{\sum\limits_{t_j \in d_i} n\,(t_j, d_i)\cdot \frac{\lambda \cdot P(t_j|S)}{\lambda \cdot P(t_j|S) + (1-\lambda)\cdot P(t_j|C)}}{\sum\limits_{t_j \in d_i} n\,(t_j, d_i)} \tag{10}$$

Once the HMM models for the sentences are estimated, they can thus be used to predict the occurrence probability of the terms in the spoken document, and the

sentences with the highest probabilities are then selected and sequenced to form the final summary according to different summarization ratios.

In the sentence HMM, as previously shown in Equation 9, the sentence model $P(t_j|S)$ is linearly interpolated with the collection model $P(t_j|C)$ to have some probability of generating every term in the vocabulary. However, the true sentence model $P(t_j|S)$ might not be accurately estimated by the MLE, since the sentence consists of only a few terms and the portions of terms present in it are not the same as the probabilities of those terms in the true model. Therefore, we can explore the use of the relevance model (RM),[23,24] also originally formulated for IR, to get a more accurate estimation of the sentence model. In the extractive spoken document summarization task studied here, each sentence S of the document d_i to be summarized has its own associated relevant 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 of the sentence S is defined to be the probability distribution $P(t_i|R_s)$, which gives the probability that we would observe a term t_j, if we were to randomly select a document from the relevant class R_s and then pick up a random term from that document.[23] 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 better estimated model. Because there is no prior knowledge about the subset of relevant documents for each sentence S, a local feedback-like procedure can be employed by taking S as a query and posing it to the IR system to obtain a ranked list of documents. The top L documents returned from the IR system are assumed to be the ones relevant to S, and the relevance model $P(t_j|R_s)$ of S can be therefore be constructed through the following equation:

$$P(t_j|R_s) = \sum_{d_i \in \{d\}_{\mathrm{Top}L}} P(d_l|S) \cdot P(t_j|d_l) \tag{11}$$

where $\{d\}_{\mathrm{Top}L}$ is the set of top L retrieved documents; and the probability $P(d_l|S)$ can be approximated by the following equation using Bayes' rule:

$$P(d_l|S) \approx \frac{P(d_l) \cdot P(S|d_l)}{\sum_{d_u \in \{d\}_{\mathrm{Top}L}} P(d_u) \cdot P(S|d_u)} \tag{12}$$

A uniform prior probability $P(d_l)$ can be further assumed for the top L retrieved documents, and the sentence likelihood $P(S|d_l)$ can be calculated using an equation similar to Equation 3 once the IR system is implemented with the HMM retrieval model. Consequently, the relevance model $P(t_j|R_s)$ is linearly combined with the original sentence model $P(t_j|S)$ to form a more accurate sentence model:

$$\hat{P}(t_j|S) = \alpha \cdot P(t_j|S) + (1-\alpha) P(t_j|R_s) \tag{13}$$

where α is a weighting parameter. The final sentence HMM is thus expressed as:

$$\hat{P}_{HMM}(d_i|S) = \prod_{t_j \in d_i} \left[\lambda \cdot \hat{P}(t_j|S) + (1-\lambda) \cdot P(t_j|C)\right]^{c(t_j,d_i)} \quad (14)$$

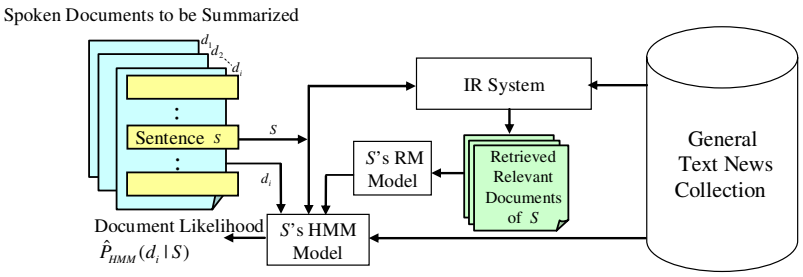

Fig. 4. A diagram of spoken document summarization jointly using the HMM and RM models.

A diagram of spoken document summarization jointly using the HMM and RM models is depicted in Figure 4.

4.3. *Spoken Documents*

The methods described above in Sections 4.1 and 4.2 are equally applicable to both text and spoken documents. However, spoken documents do involve extra difficulties like the handling of recognition errors, problems with spontaneous speech, and the lack of correct sentence or paragraph boundaries. In order to exclude the redundant and incorrect portions while selecting the important and correct information, multiple recognition hypotheses, confidence scores, language model scores and other grammatical knowledge have been utilized.[25] As an example, the above Equation 8 for the SenSig model may be extended as:

$$Sig(S) = \frac{1}{J}\sum_{j=1}^{J}[\beta_1 \cdot I(t_j) + \beta_2 \cdot F(t_j) + \beta_3 \cdot C(t_j) + \beta_4 \cdot G(t_j)] + \beta_5 \cdot H(S) \quad (15)$$

where $C(t_j)$ and $G(t_j)$ are obtained from the confidence score and n-gram score for the term t_j, $H(S)$ from the grammatical structure of the sentence S; and β_3, β_4 and β_5 are weighting parameters. In addition, prosodic features (e.g. intonation, pitch, energy, pause duration) can be used as important clues for summarization as well, although reliable and efficient approaches incorporating these features are still actively being studied.[25,26] The resulting summary of spoken documents can be generated in the form of either text or speech. Summaries in text have the advantage of easier browsing and further processing, but these are inevitably subject to speech

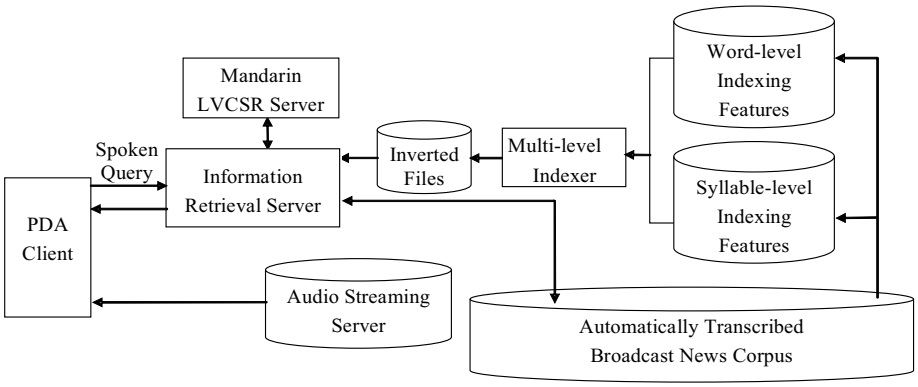

Fig. 5. The framework for voice retrieval of Chinese broadcast news.

recognition errors, as well as the loss of the speaker/emotional/prosodic information carried only by the speech signals. The speech form of summaries can preserve the latter information and is free from recognition errors, but it faces the difficult speech synthesis problem of smooth concatenation of speech segments.

5. A Prototype System of Chinese Spoken Document Retrieval and Summarization

5.1. *System Description*

A prototype system has been established in Taiwan that allows the user to search for Chinese broadcast news via the PDA using a spoken natural language query.[27] The framework of the system is shown in Figure 5. There is a small client program on the PDA, as illustrated in Figure 6, which transmits the speech waveform or acoustic feature data of the spoken query to the information retrieval server. The information retrieval server then passes the speech waveform or acoustic feature data to the large vocabulary continuous speech recognition (LVCSR) server.[28] The recognition result is then passed back to the information retrieval server to act as the query to generate a ranked list of relevant documents. When the retrieval results are sent back to the PDA, the user can first browse the summaries of the retrieved documents, which were generated beforehand by jointly using the HMM and RM models, and then click to read the automatic transcript of the relevant broadcast news documents or play the corresponding audio files from the audio streaming server. On the other hand, a huge collection of broadcast news documents are recognized offline by the broadcast news transcription system, and the resultant transcripts are then utilized by the multi-scale indexer to generate the word-level and syllable-level indexing terms.[12] Only the VSM model for literal term matching of the spoken query and the spoken documents was implemented here for simplicity, although our previous

experiments on Mandarin spoken document retrieval have demonstrated that the HMM retrieval model, and models with similar structure to the PLSA model, have superior retrieval performance over the VSM model.[7,11] The final retrieval indices, including the vocabularies and document occurrences of indexing terms of different types (word- and syllable-level indexing terms), are stored as inverted files[29] for efficient searching and comparison.

Fig. 6. A PDA-based broadcast news retrieval system that displays the retrieved broadcast news documents and their associated summaries for efficient browsing. The upper scrollable window lists the summaries of the retrieved documents, while the bottom one displays the automatic transcript of the selected document.

5.2. *Evaluation of Chinese Spoken Document Retrieval*

In order to evaluate the performance level of the retrieval system, a set of 20 simple queries with length of one to several words, in both text and speech forms, was manually created. Four speakers (two males and two females) produced the 20 queries using an Acer n20 PDA with its original microphone in an environment with slight background noise. To recognize their spoken queries, another read speech corpus consisting of 8.5 hours of speech produced by an additional 39 male and 38 female speakers over the same type of PDA was used for training the speaker-independent acoustic models for recognition of the spoken queries. The character and syllable error rates for the spoken queries are 27.61% and 19.47%, respectively. The retrieval experiments were performed with respect to a collection of about 21,000 broadcast news stories. The final retrieval results are evaluated in terms of the mean average precision (mAP)[30] at different document cutoff values L, which computes the mean average precision when the top L documents have been presented to the user. The

formula can be expressed as:

$$m\text{AP}_L = \frac{1}{E}\sum_{e=1}^{E}\frac{1}{N'_e}\sum_{i=1}^{N'_e}\frac{i}{r_{e,i}} \tag{16}$$

where E is the number of queries, N'_e is the total number of documents that are relevant to query Q_e appearing among the top L documents, and $r_{e,i}$ is the position of the i-th document that is relevant to query Q_e appearing among the top L documents, counting down from the top of the ranked list. The retrieval results are shown in Table 2. Columns 3, 4, 5 respectively show the results using

Table 2. The retrieval results evaluated in terms of the mean average precision at different document cutoff values.

		Word	**Syllable**	**Word + Syllable**
Document Cutoff: 10	Text Query	0.9309	0.8885	0.9580
	Spoken Query	0.5533	0.6036	0.6617
Document Cutoff: 30	Text Query	0.8838	0.8165	0.9224
	Spoken Query	0.5270	0.5435	0.6465
Document Cutoff: 50	Text Query	0.8656	0.7834	0.9065
	Spoken Query	0.5242	0.5212	0.6386

word-level indexing features, syllable-level indexing features and both of them, which are evaluated at different document cutoff values and with either text or spoken queries. As can be seen, the word-level indexing features are better than the syllable-level features for the text queries, while using both levels results in significant improvements over using any of them alone. Moreover, the retrieval results for the spoken queries are much worse than those of the text queries, but the combination of word-level and syllable-level features helps to reduce the performance gap between the spoken and the text queries.

5.3. *Evaluation of Chinese Spoken Document Summarization*

A set of 200 broadcast news documents (1.6 hours) collected in August 2001 were used in the summarization experiments. The average Chinese character error rate (CER) for the automatic transcripts of these broadcast news documents was 14.17%. Three human subjects were instructed to do human summarization, and this was taken to be the references for evaluation, in two forms: the first, simply to rank the importance of the sentences in the corresponding reference transcript of the broadcast news document from the top to the middle, and the second, to write an abstract for the document manually by himself, with a length of about 25% of the original broadcast news document. Several summarization ratios were tested, which are the ratios of summary length to the total document length.[3] On the other hand, the ROUGE measure[31,32] was used to evaluate the performance levels of the proposed models and the other conventional models. It evaluates the summarization quality by counting overlapping units, such as the n-gram, word sequences and so forth,

Table 3. The results achieved by jointly using the HMM and RM models, and by using other summarization models under different summarization ratios.

Summarization Ratio	HMM+RM	VSM	LSA-1	LSA-2	SenSig	Random
10%	0.3078	0.2845	0.2755	0.2498	0.2760	0.1122
20%	0.3260	0.3110	0.2911	0.2917	0.3190	0.1263
30%	0.3661	0.3435	0.3081	0.3378	0.3491	0.1834
50%	0.4762	0.4565	0.4070	0.4666	0.4804	0.3096

between the automatic summary and a set of reference (or manual) summaries. ROUGE-N is an n-gram recall measure which is defined as follows:

$$ROUGE-N = \frac{\sum_{S \in \mathbf{S}_R} \sum_{gram_n \in S} Count_{match}\left(gram_n\right)}{\sum_{S \in \mathbf{S}_R} \sum_{gram_n \in S} Count\left(gram_n\right)} \tag{17}$$

where N stands for the length of the n-gram; S is an individual reference (or manual) summary; $\mathbf{S}_R$ is a set of reference summaries; $Count_{match}\left(gram_n\right)$ is the maximum number of n-grams co-occurring in the automatic summary and the reference summary; and $Count\left(gram_n\right)$ is the number of n-grams in the reference summary. In this study, the ROUGE-2 measure that used word bigrams as the matching units was adopted.

The summarization results obtained by using the HMM and RM models jointly are shown in the second column of Table 3, and the corresponding ROUGE-2 recall rates are about 0.31, 0.33, 0.37, and 0.48 for summarization ratios of 10%, 20%, 30% and 50%, respectively. Then, we try to compare these results with those obtained by using the conventional VSM,[19] LSA, and SenSig[20] models. Two variants of LSA, i.e., the one mentioned in Section 4.1[19] (LSA-1) and the one proposed by Hirohata *et al.*[33] (LSA-2), were both evaluated here. For a spoken document, LSA-2 simply evaluated the score of each sentence based on the norm of its vector representation in the lower L-dimensional latent semantic space, and a fixed number of sentences having relatively large scores were therefore selected to form the summary. The value of L was set to 5 in our experiments, which is just the same as that suggested by Hirohata *et al.*[33] The results for these models are shown in Columns 3 to 6 of Table 3, and the results obtained by random selection (Random) is also listed for comparison. As can be seen, HMM+RM is substantially better than VSM and LSA at lower summarization ratios, and is significantly superior to SenSig as well, which provide some evidence that the probabilistic generative model (HMM+RM) is indeed a good candidate for extractive spoken document summarization tasks.

6. Conclusion

The ever-increasing storage capability and processing power of computers have made vast amounts of multimedia content available to the public. Clearly, speech is one of the most important sources of information for multimedia content, as it gives important, if not key, information regarding the content. Therefore, multimedia

access based on associated spoken documents has been a focus of much active research. This chapter has presented a comprehensive overview of the information retrieval and automatic summarization technologies developed in recent years for efficient spoken document retrieval and browsing applications. An example prototype system for voice retrieval of Chinese broadcast news collected in Taiwan was also introduced.

References

1. B. H. Juang and S. Furui, "Automatic Recognition and Understanding of Spoken Languagea First Step toward Natural Human-Machine Communication," in *Proc. IEEE 88(8)*, vol. 88(8), (2000), pp. 1142–1165.
2. L. S. Lee and B. Chen, "Spoken Document Understanding and Organization," *IEEE Signal Processing Magazine*, vol. 22(5), pp. 42–60, (2005).
3. Text retrieval conference (trec). [Online]. Available: http://trec.nist.gov/
4. G. Salton and M. E. Lesk, "Computer Evaluation of Indexing and Text Processing," *Journal of the ACM*, vol. 15(1), pp. 8–36, (1968).
5. J. M. Ponte and W. B. Croft, "A Language Modeling Approach to Information Retrieval," in *Proc. ACM SIGIR Conference on R&D in Information Retrieval*, (1998), pp. 275–281.
6. D. R. H. Miller, T. Leek, and R. Schwartz, "A Hidden Markov Model Information Retrieval System," in *Proc. ACM SIGIR Conference on R&D in Information Retrieval*, (1999), pp. 214–221.
7. B. Chen, H. M. Wang, and L. S. Lee, "A Discriminative HMM/N-Gram-Based Retrieval Approach for Mandarin Spoken Documents," *ACM Trans. on Asian Language Information Processing*, vol. 3, pp. 128–145, (2004).
8. G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. Harshman, L. A. Streeter, and K. E. Lochbaum, "Information Retrieval Using a Singular Value Decomposition Model of Latent Semantic Structure," in *Proc. ACM SIGIR Conference on R&D in Information Retrieval*, (1988), pp. 465–480.
9. J. R. Bellegarda, "Latent Semantic Mapping," *IEEE Signal Processing Magazine*, vol. 22(5), pp. 70–80, (2005).
10. T. Hofmann, "Probabilistic Latent Semantic Indexing," in *Proc. ACM SIGIR Conference on R&D in Information Retrieval*, (1999), pp. 50–57.
11. B. Chen, "Exploring the Use of Latent Topical Information for Statistical Chinese Spoken Document Retrieval," *Pattern Recognition Letters*, vol. 27(1), pp. 9–18, (2006).
12. 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. on Speech and Audio Processing*, vol. 10, pp. 303–314, (2002).
13. K. Ng and V. W. Zue, "Subword-Based Approaches for Spoken Document Retrieval," *Speech Communication*, vol. 32, pp. 157–186, (2000).
14. S. Srinivasan and D. Petkovic, "Phonetic Confusion Matrix Based Spoken Document Retrieval," in *Proc. ACM SIGIR Conference on R&D in Information Retrieval*, (2000), pp. 81–87.
15. B. Chen, H. M. Wang, and L. S. Lee, "Improved Spoken Document Retrieval by Exploring Extra Acoustic and Linguistic Cues," in *Proc. European Conference on Speech Communication and Technology*, (2001), pp. 299–302.
16. E. Chang, F. Seide, H. Meng, Z. Chen, Y. Shi, and Y. C. Li, "A System for Spoken

Query Information Retrieval on Mobile Devices," *IEEE Trans. on Speech and Audio Processing*, vol. l0, pp. 531–541, (2002).
17. A. Singhal and F. Pereira, "Document Expansion for Speech Retrieval," in *Proc. ACM SIGIR Conference on R&D in Information Retrieval*, (1999), pp. 34–41.
18. I. Mani and E. M. T. Maybury, *Advances in* A*utomatic* T*ext* S*ummarization*. (Cambridge. MA: MIT Press, 1999).
19. Y. Gong and X. Liu, "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis," in *Proc. ACM SIGIR Conference on R&D in Information Retrieval*, (2001), pp. 19–25.
20. J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell, "Summarizing Text Documents: Sentence Selection and Evaluation Metrics," in *Proc. ACM SIGIR Conference on R&D in Information Retrieval*, (1999), pp. 121–128.
21. B. Chen, Y. M. Yeh, Y. M. Huang, and Y. T. Chen, "Chinese Spoken Document Summarization Using Probabilistic Latent Topical Information," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal processing*, (2006), pp. 969–972.
22. Y. T. Chen, S. Yu, H. M. Wang, and B. Chen, "Extractive Chinese Spoken Document Summarization Using Probabilistic Ranking Models," in *Proc. International Symposium on Chinese Spoken Language Processing*, (2006).
23. W. B. Croft and J. L. (Eds.), *Language* M*odeling for* I*nformation* R*etrieva*. (Kluwer-Academic Publishers, 2003).
24. M. D. Smucker, D. Kulp, and J. Allan, CIIR T*echnical* R*eport:* D*irichlet* M*ixtures for* Q*uery* E*stimation in* I*nformation* R*etrieval*. (Center for Intelligent Information Retrieval, University of Massachusetts, 2005).
25. S. Furui, T. Kikuchi, Y. Shinnaka, and C. Hori, "Speech-to-Text and Speech-to-Speech Summarization of Spontaneous Speech," *IEEE Trans. on Speech and Audio Processing*, vol. 12, pp. 401–408, (2004).
26. S. Maskey and J. Hirschberg, "Comparing Lexical, Acoustic/Prosodic, Structural and Discourse Features for Speech Summarization," in *Proc. European Conference on Speech Communication and Technology*, (2005), pp. 621–624.
27. B. Chen, Y. T. Chen, C. H. Chang, and H. B. Chen, "Speech Retrieval of Mandarin Broadcast News via Mobile Devices," in *Proc. European Conference on Speech Communication and Technology*, (2005), pp. 109–112.
28. B. Chen, J. W. Kuo, and W. H. Tsai, "Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal processing*, (2004), pp. 777–780.
29. R. Baeza-Yates and B. Ribeiro-Neto, *Modern* I*nformation* R*etrieval*. (Addison-Wesley, 1999).
30. D. Harman, "Overview of the Fourth Text Retrieval Conference (TREC-4)," in *Proc. Fourth Text Retrieval Conference*, (1995), pp. 1–23.
31. C. Y. Lin. Rouge: Recall-oriented understudy for gisting evaluation (2003). [Online]. Available: http://www.isi.edu/ cyl/ROUGE/
32. ——, "Looking for a few Good Metrics: ROUGE and its Evaluation," *Working Notes of NTCIR-4*, vol. Supl. 2, pp. 1–8, (2004).
33. M. Hirohata, Y. Shinnaka, K. Iwano, and S. Furui, "Sentence Extraction-Based Presentation Summarization Techniques and Evaluation metrics," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal processing*, (2005), pp. 1065–1068.