A System for Spoken Query Information Retrieval on Mobile Devices

Eric Chang, *Senior Member, IEEE*, Frank Seide, *Member, IEEE*, Helen M. Meng, Zhuoran Chen, Yu Shi, and Yuk-Chi Li

***Abstract*—With the proliferation of handheld devices, information access on mobile devices is a topic of growing relevance. This paper presents a system that allows the user to search for information on mobile devices using spoken natural-language queries. We explore several issues related to the creation of this system, which combines state-of-the-art speech-recognition and information-retrieval technologies.**

This is the first work that we are aware of which evaluates spoken query based information retrieval on a commonly available and well researched text database, the Chinese news corpus used in National Institute of Standards and Technology (NIST)'s TREC-5 and TREC-6 benchmarks. To compare spoken-query retrieval performance for different relevant scenarios and recognition accuracies, the benchmark queries—read verbatim by 20 speakers—were recorded simultaneously through three channels: headset microphone, PDA microphone, and cellular phone.

Our results show that for mobile devices with high-quality microphones, spoken-query retrieval based on existing technologies yields retrieval precisions that come close to that for perfect text input (mean average precision 0.459 and 0.489, respectively, on TREC-6).

***Index Terms*—Information retrieval, speech recognition, user interface.**

I. Introduction

THERE is increasing interest in incorporating speech technologies on mobile devices. Speech input provides the following benefits:

- potentially faster input speed compared to keypads and/or handwriting recognition;
- ability to be used on small form-factor devices: in many mobile devices, a microphone is already included as a part of the device and;
- natural and easy user interface.

There is also growth in the use of mobile devices for accessing information and multimedia content. This trend is especially noticeable in Asia. For example, NTT DoCoMo has over 30 million users accessing textual and multimedia information from the World Wide Web with the i-mode service. Similarly, China already has over 170 million cellular-phone users as of May 2002 and the use of short messaging service (SMS) has grown tremendously in the past year. There is a growing trend in increasing the capability of mobile devices such as cellular phones and personal digital assistants (PDAs) so that they are not merely communication and storage devices, but also information-access devices. With the upcoming 2.5G and 3G wireless networks, it will become possible for mobile phone users to download or stream music and other multimedia content through the wireless network to their mobile devices. For these applications, it is crucial for the users to be able to easily search for and obtain their desired contents.

In this paper, we describe a spoken-query information retrieval system that has been implemented for Mandarin Chinese. A prototype of the system currently exists which allows users to search for information over a mobile device, in our case a Compaq iPAQ PocketPC. The focus of this paper, however, is on offline evaluation using read queries. While the system can be adapted to any type of textual content, we evaluated our system on a database traditionally used in information retrieval research to evaluate the overall effectiveness of the system. As far as we know, this is the first published work which systematically evaluates the applicability of spoken-query information retrieval on a commonly available and well-researched information-retrieval benchmark. We are using a collection of Chinese news articles that had been used during the NIST TREC-5 and TREC-6 conferences [28].

We explore several issues related to the creation of this information search system. We incorporate special considerations related to the linguistic properties of the Chinese language. The basic unit of the Chinese text is the character and the basic unit of the Chinese phonological system is the syllable. The GB-2312 character set defines over 6700 characters. Approximately 3–4000 Chinese characters can fully cover the Chinese writing system and approximately 400 Chinese syllables can fully cover the Mandarin Chinese phonological system (where Mandarin is the official Chinese dialect). When the Chinese tones are considered, the number of syllables grows to about 1250. The language is monosyllabic in nature, where each character is pronounced as a syllable, with a many-to-many mapping. A Chinese word may contain one or several characters and there is no explicit word delimiters, hence a given Chinese character sequence may be segmented in multiple ways to form different word sequences that have different meanings. Consequently, much ambiguity exists in the problem of Chinese word segmentation. Our current task of spoken query information retrieval requires bridging the gap between a *spoken* query and a *textual* document collection. This may involve i) transforming

Manuscript received October 3, 2001; revised August 7, 2002. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Harry Printz.

E. Chang, F. Seide, and Y. Shi are with the Microsoft Research Asia, Beijing, China 100080 (e-mail: echang@microsoft.com).

H. M. Meng and Y.-C. Li are with the Chinese University of Hong Kong, Shatin, N.T., Hong Kong.

Z. Chen is with Peking University, Beijing, China.

Digital Object Identifier 10.1109/TSA.2002.804301

and representing the spoken query in terms of characters and performing retrieval by matching with documents in character space or ii) transforming and representing the textual documents in terms of syllables and performing retrieval by matching with queries in syllable space. In this paper, we organize our investigation as follows.

- We explore the use of different indexing (and retrieval) units for the spoken-query information-retrieval task. First of all, we evaluate the performance of the search engine when different indexing units are used, specifically character unigrams, overlapping character bigrams, syllable unigrams, and syllable bigrams. The overlap in M-gram formation circumvents the problem of word-tokenization ambiguity in Chinese. The use of bigrams should capture some sequential lexical constraints since the majority of Chinese words are bi-character (hence also bi-syllable) words. In these experiments, we assume "perfect speech recognition" by using the correct character and syllable sequences to form the respective M-grams. This part of our investigation establishes a reference retrieval-performance benchmark for spoken-query retrieval. Because Mandarin Chinese is a syllabic language with a finite set of approximately 400 syllables, it is particularly worthwhile to explore the feasibility of indexing textual contents by syllable M-grams. We have also attempted to boost this benchmark by applying a query expansion technique.
- This work also explores the effect of imperfect speech recognition on retrieval performance. Spoken queries recorded from three different sources (headset microphone, PDA microphone and cellular telephone) are used. Since the acoustic models of our speech recognizer are trained on headset microphone speech, the PDA microphone and cellular telephone speech present increasing channel distortions for recognition. Such distortions tend to cause recognition errors which in turn affect retrieval performance. Hence, we have incorporated a channel-compensation technique to counteract such effects. In order to transcribe the spoken query into characters or syllables, we have used both a large-vocabulary continuous speech recognizer (LVCSR) and a Chinese syllable recognizer. Character and syllable M-grams are subsequently derived from the transcriptions. It should be noted that LVCSR incorporates rich lexical constraints from the recognizer's vocabulary and language model, but is vulnerable to out-of-vocabulary (OOV) words such as named entities. Contrastively, a syllable recognizer is not affected by the OOV problem, but lacks lexical constraints. We have also applied speaker/channel adaptation during recognition in order to minimize recognition errors and their negative effects upon retrieval performance.
- We have studied the use of two speech recognition systems in the task for information retrieval. The first system is a high-accuracy, state-of-the-art large-vocabulary, continuous-speech recognition (LVCSR) system that is widely available [22]. The second system is a syllable-based recognizer that is trained using the publicly available MSR Mandarin Speech Toolkit database [3] and the publicly available HTK toolkit [27]. We are using the LVCSR system because we want to understand the benefit of a state-of-the-art system with well-trained acoustic and language models trained with an amount of data not typically seen in Mandarin speech-recognition research. On the other hand, we created the HTK based recognizer because all the major components are available and it will provide a better baseline for researchers who are interested in speech recognition research.

We have studied the effect of using different speech-recognition results: Chinese characters versus syllables. Using a large-vocabulary continuous-speech recognition system to perform recognition provides the benefit of generating Chinese characters and generally higher accuracy. However, since the size of the lexicon for the large-vocabulary continuous-speech recognition engine is under 60 000 words, there will be many out-of-vocabulary words in normal usage. For example, while common city names may already be in the lexicon and thus are more easily recognized by the LVCSR engine, less common city names are not included in the lexicon and are difficult to be recognized correctly. Using syllables as recognition units can provide better ability to handle such out of vocabulary queries.

Another issue we explore in this paper is the adaptability of a speech-recognition system trained on high-quality speech recorded through headset microphones in handling new channel conditions such as the ones encountered when using personal digital assistants or cellular phones. While considerable amounts of high-quality speech data have been recorded historically in many languages for desktop dictation or command-and-control applications, the recording of similar amounts of data in new user scenarios such as over cellular phones or personal digital assistants is time consuming and expensive. In particular, as more devices are created in the future, it will be very expensive to collect large amounts of training data for all types of microphones and devices. In this paper, we study the use of speech recognition systems trained mainly on wide band, desktop data toward testing data collected under several scenarios. Also, we have developed a method to map the originally recorded high quality speech corpus to be closer to the target condition and its effectiveness is reported.

We examine the issue of channel effects by collecting a speech corpus which contains speech collected through three different channels (cellular phone, an iPAQ PocketPC PDA, and a headset microphone for baseline comparisons). By simultaneously recording the same speech through all three channels, we can more easily evaluate the channel's impact on recognition accuracy. It is found that while waveforms recorded on the iPAQ PocketPC are significantly different from those recorded on the desktop microphone, with some adaptation very good performance can be obtained on waveforms recorded with an iPAQ. However, the greater mismatch between data recorded on cellular phones and that on headsets brings up a more challenging speech recognition task.

The remainder of this paper is organized as follows: in Section II, we survey related work in speech query information retrieval, in Section III, the system architecture and the detailed descriptions of each individual component are provided. Section IV describes the experimental conditions and results. Sections V and VI present our conclusions and future extensions, respectively.

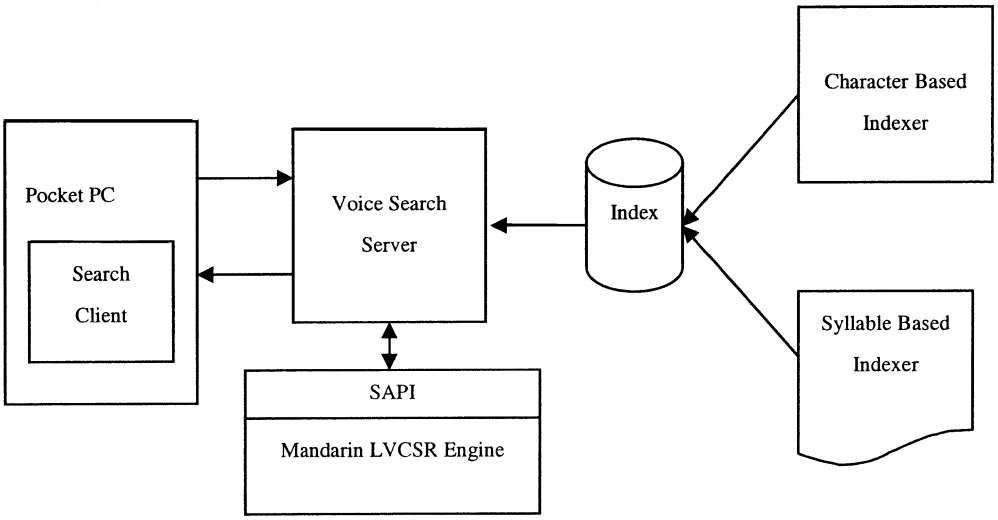


Fig. 1. System block diagram.

II. Previous Work

There has been relatively little work in spoken-query information retrieval compared with spoken-document retrieval. Kupiec *et al.* used a speaker-dependent speech-recognition system to recognize spoken keywords for information retrieval. A hypothesized phone sequence was generated from the speech-recognition system where each input keyword is spoken in isolation. The hypothesized phone sequence was also matched against possible keywords in the document database. Kupiec reported satisfactory results when the system was used to query articles in an encyclopedia [12].

More recently, there has been work in combining speech recognition with information retrieval using relatively small text or spoken document databases [1]. Chen *et al.* have previously compared the effect of using speech or text for both query and document source in information retrieval. Using spoken queries on text content was found to provide better retrieval performance compared to using text queries on spoken content [4]. However, this experiment was not conducted with widely available and well studied public databases.

In the related area of spoken-document retrieval, there has been extensive work reported over several years of TREC conferences [26]. Various approaches have been tried in indexing the spoken documents, ranging from using sub-word decomposition of English as representation units [10], [17] to using large-vocabulary recognition output of the spoken documents. More recently, there has also been work on using English news articles as queries for retrieving Mandarin spoken documents [15].

III. System Description

A. Overview

The prototype system consists of three components (Fig. 1). On the mobile device, there is a small client program which transmits spoken queries to the search engine. The data is transmitted to the engine in PCM format. Extensions to the client program that would allow features to be computed on the client, similar to distributed speech recognition approaches such as the Aurora project [18], is part of our future work.

The search engine passes the received speech samples to a LVCSR speech recognition engine. The recognized results are then used as the query sequence to generate a ranked list of relevant documents. To accelerate the retrieval process, syllable-based or character-based indices are pre-created using an indexer on the text corpus. Fig. 2 shows the user interface shown to the user. The recognized query is returned from the recognizer to the client and shown in the upper input window. The results from using the query sentence to search the database are shown in a ranked list below the input window. The user can then select any of the returned links to get the detailed content, or modify the query sentence by speech, soft keyboard, or handwriting.

To more formally study the effectiveness of the spoken query driven information retrieval approach, we focus on experimental results gathered on an offline system. The offline system allows us to flexibly change the indexing representation and the recognition engine. In the offline evaluations, the client component does not exist and the speech-recognition component is either an LVCSR or a syllable-unit based engine.

B. Information Retrieval Engine

1) Baseline System: Our information retrieval (IR) engine employs the well-known and commonly used vector-space model for information retrieval [19] due to its simplicity and good performance consistently verified by previous benchmarks. In this model, every word of the language is assumed to carry a certain concept and stands for a dimension in a high-dimensional space. A document, as well as a query, is represented in the space by a vector, where every element is associated with a particular word occurring in it. In this way the closeness of one document to one query, namely the similarity, can be calculated by the inner product of the two corresponding vectors. To execute a query, the engine orders the documents by their similarities to that query and returns the top N-ranked ones.

Fig. 2. Display shown to users after a spoken query has been processed. The query used in this illustration is: "Reform and growth in China's real estate industry."

In other words, the quantity of $sim(q,d)$, the similarity given the query q and the document d, is computed using the formula

$$sim(q,d) = \sum_k q_k \cdot d_k. \tag{1}$$

In this equation, q_k and d_k denote the weight of term k in the query and document respectively. Term weighting usually deviates among various systems and many variants have been proposed since the 1970's [21], [24]. The most popular family of weighing formulae is usually referred to as the TF.IDF approach, in which each weight is computed from three components: the term frequency (TF), the inverse document frequency (IDF) and some form of document length normalization.

A large variety of heuristic modifications of the basic TF.IDF formula has been proposed in the literature, including nonlinear compression of TF and IDF values (such as replacing by $\log(TF+1)$) and introduction of tunable parameters. Instead of trying to gain theoretic and practical understanding of the benefits and disadvantages of each individual formula variation, we took the approach of viewing the choice of the optimal weighting-formula version and its parameters as an optimization problem. We approached it in a blind data-driven manner using an evolutionary learning algorithm [6], [11].

We first combine various variations of the TF.IDF formula proposed in the literature as well as tuning parameters to generate a population of 200 weighting formula variations. Then those formulae that yield highest mAP scores for the TREC-5 queries (the development set) are selected. In the evolutionary learning iteration, they cross-over with each other so as to produce a new population. The procedure repeats five times and the best formula found undergoes further fine-tuning to get the final weighting formula. For example, for the indexing unit "char-UB" (see Section IV-A-II), we obtain the following optimal formula:

$$q_k = \log \frac{N}{DF'_k} \tag{2}$$

$$d_k = \frac{1+\log TF_{d,k}}{1+\log avgTF_d} \cdot \log \frac{N}{DF'_k} \cdot \frac{1}{p+L(d)} \tag{3}$$

$$DF'_k = \max\{DF_k, lowfreq\} \tag{4}$$

with $TF_{d,k}$ denoting the number of occurrences of term k in the document; N being the total number of documents in the corpus; DF_k denoting the number of documents that contain term k; $avgTF_d$ being the average term frequency in the document; $L(d)$ denoting the number of units in the document d; and two tunable parameters, p and $lowfreq$.

Some terms only have syntactic usefulness in the language and thus are considered as noncontent words. They are removed from the query beforehand, usually by looking up a manually crafted "stop list." In addition to the stop list, we also remove those terms that very frequently occur in the documents: A term k is included in the stop list if DF_k is more than $N/hifreq$, where $hifreq$ is another tunable parameter.

2) Query Expansion: To improve retrieval performance, modern information-retrieval systems usually use pseudo (blind) relevance-feedback techniques. First, the original query is used to retrieve a preliminary ranked list of documents. Then the R top weighted terms are extracted from the top S documents in the list, which are assumed to be relevant to the query. Finally, a new query is constructed by adding these terms and is used to generate the final result. This is actually a simplified version of the Rocchio algorithm [20]. The optimal values of R and S are also established by the evolutionary algorithm.

C. Choosing the Best Indexing Units for Chinese

The first difficulty in applying standard information-retrieval technology to Chinese articles is to find a suitable indexing unit. The Chinese language has around 3–4000 frequently used characters. Chinese text is written without segmentation—there is no natural spacing within a sentence. So there are no natural and unambiguous word units that can be used for indexing. This problem is especially challenging for infrequently occurring words such as proper names of people and locations.

Literature proposes to use sub-word units as the indexing units for speech-enabled information retrieval [17]. This poses

another problem: every spoken Chinese character is pronounced as a single syllable. Considering that the total number of Chinese syllables is approximately 400 (not considering tones), it is obvious that a syllable sequence has far less information than the corresponding character sequence.

Although successful sentence-segmentation algorithms have been proposed to handle the first issue, none of them can perform this task with 100% accuracy. Out-of-vocabulary words are the main problem. A promising approach is to use overlapping bigrams as the indexing units. Previous results show that this approach can achieve retrieval performance comparable to word-based approaches [28].

We have experimented with various types of possible indexing units in our information-retrieval system, in particular single characters (unigrams), overlapping character bigrams, single syllables (unigrams) and overlapping syllable bigrams. Bigrams with document frequency 1 are not used, assuming they just occur in the corpus by chance and do not actually carry meaningful concepts. In addition, we have experimented with the combination of unigrams and bigrams.

D. Speech Recognition System

We have performed experiments using both an LVCSR system and a syllable based speech recognition system on the collected corpus. One benefit of using an LVCSR system is its higher accuracy due to the extensive information provided by the lexicon and the powerful language model. Although large-vocabulary speech-recognition systems tend to have higher recognition accuracies, there will be cases when the query has out-of-vocabulary words and phrases such as names of people and locations. In these cases, a syllable-based speech-recognition system and a bi-syllable based indexing method may allow users to successfully find the desired content when the character-based method would fail. Also, the LCVSR engine available to us at this time, the Microsoft SAPI 5.0 recognition engine, was trained on high-quality headset–microphone data and does not match well to the data that we collected over the iPAQ and cellular phones. On the other hand, we have the opportunity to train more channel-dependent models with the syllable recognizer. Detailed descriptions of the two systems are presented below.

1) SAPI 5.0 Mandarin Speech Recognition Engine: The LVCSR engine applied is a component in the publicly available Microsoft SAPI 5.0 SDK [22]. Acoustic models were trained on data from over 1000 speakers with each speaker speaking over 200 utterances while wearing a headset microphone. The decoder uses a lexicon of over 50 000 Chinese multi-character words and utilizes a language model which was trained on over 1 billion characters of Chinese text. The engine is an extension of the Whisper system with additional features added to handle the tonal variations that occur in Mandarin Chinese [2].

The audio recording obtained from the user is passed to the Microsoft SAPI engine through the web server to perform large-vocabulary continuous Mandarin speech recognition. When keeping the speaker's acoustic information in a user profile, the SAPI engine can perform unsupervised adaptation to the input audio data based on the background adaptation feature of SAPI [22]. In addition to standard HMM model adaptation techniques, the engine performs aggressive adaptation of the silence and speech means used by the acoustic front end for cepstral mean normalization. We have passed the same audio file to the SAPI engine consecutively for five times based on the same user profile and observed significant error rate reduction through this adaptation process, which we attribute mostly to the front-end adaptation. The character output after five iterations are used as our recognized Chinese string for information retrieval.

2) HTK-Based Syllable Speech Recognition System: The syllable speech recognizers were trained using the same procedure as the benchmarking system described in the Microsoft Research (MSR) Mandarin Speech Toolbox [3]. For the headset and iPAQ recordings, we used wide-band models trained on the MSR Mandarin Speech Toolbox corpus directly, while for the cellular-phone recordings, a set of narrow-band models was created using the channel-mapping technique described in the following section.

During recognition, a tonal model and a lexicon comprising the $\sim$ 1250 tonal syllables of Mandarin is used. However, since tone recognition without lexical or language-model constraints is highly unreliable, tone information was stripped from the recognition output in all spoken-query retrieval experiments.

3) Channel Mapping: Since a Mandarin cellular-phone training database of sufficient size was not available, we applied the following channel-mapping strategy to compensate for the channel differences between training data and testing data: First, we estimated the average power-spectral densities P_r and P_e of the down-sampled wide-band training data x_r and a small held-out portion x_e of the cellular-phone test data, respectively,

$$P_x(m) = \frac{1}{L}\sum_{l=1}^{L} P_x^l(m) \tag{5}$$

where L is the number of speech frames and P_x^l the power spectral density of the l-th speech frame x^l given by

$$P_x^l(m) = \frac{1}{N}\left|\sum_{n=0}^{N-1} x^l(n)w(n)\exp\left(\frac{-j2\pi nm}{N}\right)\right|^2, \quad m = 0, 1, \ldots, N-1. \tag{6}$$

Here, N is the length of speech frames including padded zeros and w is a Hamming window. We only used nonsilence frames to calculate the average power-spectral density, using an energy threshold to classify frames as speech or silence. Note that signal x_r and x_e should be long enough to satisfy the assumption that estimated spectral densities represent the channel more appropriately than the actual speech content.

Then, we estimated the average power-spectral density P_h of the mapping filter h and calculated the autocorrelation sequence R_h of h

$$P_h(m) = \frac{P_{x_e}(m)}{P_{x_r}(m)} \tag{7}$$

$$R_h(k) = \frac{1}{N}\sum_{m=0}^{N-1} P_h(m)\exp\left(\frac{j2\pi km}{N}\right), \quad k = 0, 1, \ldots, N-1. \tag{8}$$

<top>

<num> Number: CH52

<C-title> 中国房地产业的改革与发展

<E-title> Reform and Growth in China's Real Estate Industry

<C-desc> Description:

中国，房地产业，投资，规模，交易，转让，炒卖，暴利，

<E-desc> Description:

China, real estate industry, investment, scale, trade, transferring possession, wild selling, huge profits

<C-narr> Narrative:

相关文件应提到中国房地产业所面临的问题以及政府采取何种措施来促进房地产业的健康发展．房地产业的问题包括炒买炒卖房地产，开发规模过大，投资者获取超额的利润，国有土地出让过多过滥，交易行为不规矩，交易价格混乱等．政府对房地产业所采取宏观管理则包括开征土地增值税，实施土地使用管制，颁布城市房地产业管理法等来促进房地产业健康发展之措施．有关中国住房制度改革之文件亦属相关文件．

<E-narr> Narrative:

A relevant document should discuss problems faced in the real estate industry and the various measures adopted by the Government to promote healthy growth for the industry. Problems in the real estate industry include wild buying and selling of real estate, developing at an excessive scale, investors obtaining excessive profits, excessive and indiscriminate selling of public lands, unrestricted trade practices, trade price speculation, etc. Growth policies being adopted by the Government for macro-management of the industry include measures to promote the healthy growth in the industry by collecting a value added tax on land, implementing controls for land use, promulgating the urban control of land use law, etc. Documents about the reform of the housing system are also relevant.

</top>

Fig. 3. Sample topic from the TREC-6 evaluation.

Given R_h, the coefficients and the gain of an auto-regressive (AR) linear filter, h, can be estimated by the Levinson-Durbin recursion. Finally, we mapped waveforms in the training data x_r from their original channel to the test data channel by filtering them with this filter. More details about the acoustic channel mapping method can be found in [23].

The whole MSR Mandarin Speech Toolbox corpus was filtered to match the cellular-phone channel and a narrow-band HMM model was trained on both the unfiltered down-sampled MSR Mandarin Speech Toolbox corpus and the newly mapped training set. The mapped training set contributed to the HMM model only after the decision tree was generated. In experiments on the development set, we found that this approach reduced the syllable error rate by over three percentage points compared to using only down-sampled data without adding mapped data (error reduction from 52.4% to 49.1%).

IV. Experimental Results

A. Baseline Retrieval Results Using Text Queries

1) Text Database: The document collection used in the NIST TREC-5/6 Chinese tasks contains 164 789 documents from the People's Daily and the Xinhua News Agency. There was no word-segmentation information supplied. For the TREC-5 evaluation, NIST constructed 28 topics for this task but only 19 of them were used to generate official results. In the TREC-6 evaluation, 26 new topics were provided on the same collection. Human assessors manually selected and verified the articles in the database that are related to each topic. This allows systematic evaluation of information-retrieval performance. A sample query from the TREC-6 evaluation corpora is shown in Fig. 3. In our experiments, TREC-5 was used as a development set for parameter tuning, while TREC-6 acted as the evaluation set.

Ignoring the parallel English text, there are three sections in one topic, namely the *title*, the *description* and the *narrative*. Although all the sections are exploited in a typical run submitted for the TREC contest, we only choose the title section as our query sequence because it is written as a natural short phase, which, in our opinion, is more similar to phrases that would be used by humans to express their search requests.

2) Baseline Retrieval Results: To first evaluate the effectiveness of various indexing methods, we supplied the search engine with the correct character/syllable transcription of the queries. The results are listed in Table I. We compare retrieval performance for six different types of indexing units, including Chinese characters (char), Chinese syllables (syl), unigrams (U), overlapping bigrams (B), as well as unigrams plus bigrams

TABLE I
INFORMATION RETRIEVAL PERFORMANCE BASED ON DIFFERENT INDEXING UNITS USING CORRECT SCRIPT (MEASURED IN MEAN AVERAGE PRECISION, MAP)

indexing unit	TREC-5		TREC-6	
	query exp.	no query exp.	query exp.	no query exp.
char-U	0.281	0.205	0.324	0.289
char-B	0.340	0.176	0.471	0.323
char-UB	0.343	0.198	0.483	0.325
syl-U	0.093	0.036	0.076	0.052
syl-B	0.353	0.255	0.467	0.361
syl-UB	0.358	0.258	0.489	0.366

(UB). The weighting schemes (a set of term-weighting methods with corresponding parameters) were optimized individually for each type of indexing unit to ensure that their retrieval capabilities are impartially investigated. Numbers shown are the usual *mean average precision* (mAP) scores, a widely accepted criterion in the information retrieval community. Retrieval is performed both with and without query expansion.

We can draw the following conclusions.

1) Chinese-character unigrams do carry some information, far more than syllable unigrams.
2) Bigrams perform significantly better than unigrams as index units.
3) Syllable-level bigrams (syl-B) and the unigram-bigram combination (syl-UB) have similar retrieval performance as character bigrams (char-B) and the unigram-bigram combination (char-UB), even though characters carry more information than syllables.

The two indexing methods char-UB and syl-UB provide good retrieval performance that is comparable to that reported in [13], where a mean-average precision of 0.4755 is achieved on TREC-6 database under the same test conditions. We chose these for subsequent experiments. The parameter optimization was done as described in Section III-B1 using evolutionary algorithms. In this optimization, we did not include the popular Okapi (BM25) formula, which is commonly amongst the best-ranking information-retrieval criterions. While we are not aware of Okapi-based results on the TREC-6 data for title-only queries, [8] reports a mean-average precision score of 0.5129 on TREC-6 for long queries (title, description and narrative). For comparison, we applied our method to long queries. The resulting system, optimized using the TREC-5 queries (our development set), achieves a mean-average precision score of 0.5792 on TREC-6.

In this paper, we used nontonal syllables as representation units for indexing text documents even though Chinese syllables are tonal. There are several reasons for making this choice. First of all, the use of tonal syllables would significantly increase the size of syllable unigram and bigram indexing vectors. Secondly, the comparison between character bigram and syllable bigram shows that for text retrieval, the smaller vocabulary size of nontonal syllables already provides enough information for comparable retrieval results. Lastly, previous results in syllable recognition have shown that tonal syllable recognition error rate is significantly higher than nontonal syllable recognition (32.45% versus 22.66%, respectively) [3].

B. Speech Recognition Baseline Results for Spoken Queries

1) Speech Database: We have collected a database of spoken queries from 20 male speakers. Only male speakers are recorded for the testing set because the database used for training the syllable recognizer contains only male speakers. To study the effect of channel on speech recognition performance, the data were collected in a quiet office environment over three channels simultaneously: 1) an Andrea noise-canceling headset microphone, 2) a Compaq iPAQ PocketPC PDA using the original built-in microphone, and 3) a Motorola L2000 cellular phone running over a GSM network. The speakers were instructed to read the queries from the NIST TREC-5 and TREC-6 Chinese corpus, consecutively, in a natural fashion. The recordings were then manually verified and segmented. Each person spoke 108 sentences comprising the titles and descriptions of all 54 topics provided for the TREC-5 and TREC-6 benchmarks. Since we only perform evaluations on the title sections, we have been able to use the recordings of the 54 description sections per person for channel adaptation. We use ten speakers as our development set and the remaining ten speakers as our test set. The channel mapping methodology described in Section III-D3 was implemented using the data from the development set while all experiments reported are conducted on data from speakers in the test set.

2) Speech Recognition Baseline Results: We have trained two HMM models using HTK. For the desktop and iPAQ channels, a model was trained on the wide-band MSRCN Mandarin Speech Toolkit data described in [3]. For the cell-phone channel, we created a model on narrow-band data obtained by mapping the MSRCN wide-band data to cellular-phone channel conditions as described in Section III-D3 (the recordings used for channel mapping were not part of the test set). The test set consists of the first 54 sentences (title-part of the topics) from the ten evaluation set speakers. The average recognition error rates of the ten speakers for three channels are listed in Table II. The results for the ten development set speakers are also shown.

Supervised MLLR (maximum likelihood linear regression) and MAP (maximum a-posteriori) adaptation was performed on the original models to adapt them to the channel. Sentences 55 to 108 (description-part of the queries) from the ten speakers in the development set were used for performing this channel adaptation. Two iterations of adaptation were used with the HEAdapt tool in the HTK toolkit using the two-pass approach suggested by the HTK documentation: On the first pass, a global adaptation was performed for every output distribution

TABLE II
SYLLABLE ERROR RATES FOR THREE DIFFERENT CHANNELS USING THE SYLLABLE LOOP RECOGNIZER BASED ON HTK. THE TEST SET CONSISTS OF RECORDINGS FROM TREC-6 QUERIES

	wide-band				narrow-band			
	dev set		test set		dev set		test set	
	base	tonal	base	tonal	base	tonal	base	tonal
headset	29.9	52.3	37.7	58.8	-	-		
+adaptation	21.4	39.7	34.4	54.6				
iPAQ	46.1	64.9	41.1	61.6	-	-		
+adaptation	28.1	45.6	30.1	49.4	-	-		
cellular phone	-	-			49.1	72.0	46.1	69.9
+adaptation	-	-			37.6	54.8	40.2	59.5

TABLE III
RETRIEVAL PERFORMANCE ON VARIOUS INDEXING UNITS ON BOTH CHINESE CHARACTERS AND ITS BASE-SYLLABLE'S REPRESENTATIONS GIVEN THE RECOGNITION CHARACTER ERROR RATE (CER)/SYLLABLE ERROR RATE (SER) FOR THE TREC-5 DATA (DEVELOPMENT SET) FOR BOTH THE LVCSR AND THE SYLLABLE-LOOP RECOGNIZER. THE EVALUATION-SET SPEAKERS WERE USED IN THIS EXPERIMENT. (WB: WIDE BAND, NB: NARROW BAND, AND GD: GENDER DEPENDENT)

recognizer	device	acoustic model	CER [%]	SER [%]	char-UB [mAP]	syl-UB [mAP]
reference	-	-	0.0	0.0	0.343	0.358
LVCSR (SAPI 5.0)	headset	WB/GD	33.1	29.2	0.235	0.230
		+adapt	22.0	18.8	0.265	0.257
	iPAQ	WB/GD	25.3	22.1	0.245	0.236
		+adapt	22.7	19.3	0.255	0.255
	cellular phone	NB/GD	70.2	66.7	0.066	0.061
		+adapt	50.2	46.9	0.138	0.138
syllable loop (HTK 2.2)	headset	WB/GD	-	29.9	-	0.135
		+adapt	-	21.4	-	0.143
	iPAQ	WB/GD	-	46.1	-	0.115
		+adapt	-	28.1	-	0.163
	cellular phone	NB/GD	-	49.1	-	0.093
		+adapt	-	37.6	-	0.107

of every model. Adaptation was done incrementally after every 54 utterances (every one speaker). Only MLLR was used in this pass. The second pass used this global transformation to transform the model set, producing better frame/state alignments which were then used to estimate a set of more specific transforms for specific groups of Gaussians. To identify the number of transforms that can be estimated using the current adaptation data, HEAdapt used a regression class tree to cluster together groups of output distributions that were to undergo the same transformation. The regression class tree was built by using the models and state occupation statistics generated by the last embedded re-estimation. One can see that adaptation to the channel provides substantial improvements in syllable recognition rates for all three types of channel conditions.

C. Results for Information Retrieval Using Spoken Queries

Table III shows the retrieval performance on various indexing units, including the combination of unigrams and bigrams (char-UB) and the corresponding syllable-based version of that combination (syl-UB). We have used both large-vocabulary Mandarin speech recognition (SAPI 5.0) as well as syllable-loop recognition (HTK 2.2) to perform speech recognition on the spoken queries. For syllable-level experiments using the character-level LVCSR output, we converted the character sequence to its base-syllable representation using the front-end processing module of Microsoft Research Asia's Mandarin text-to-speech engine that has been designed to handle the problem of homophones in Chinese.

The results show that with wide-band spoken queries, the results from the headset microphone and the mobile iPAQ are quite comparable. While the results are still worse compared to the result using perfect text (mAP of 0.255 versus 0.358 on TREC-5 and 0.459 vs. 0.489 on TREC-6), given that it is faster to input the queries using speech on mobile devices compared to using handwriting recognition or the soft keyboard, the overall efficiency of spoken-query information retrieval should be greater on a mobile device.

However, results with the cellular-phone channel data are not satisfactory. Without any adaptation to the channel, the best

TABLE IV
RETRIEVAL PERFORMANCE WITH VARIOUS INDEXING UNITS ON BOTH CHINESE CHARACTERS AND ITS BASE-SYLLABLE'S REPRESENTATIONS GIVEN THE RECOGNITION CHARACTER ERROR RATE (CER)/SYLLABLE ERROR RATES (SER) FOR THE TREC-6 DATA (EVALUATION SET) FOR BOTH THE LVCSR AND THE SYLLABLE-LOOP RECOGNIZER. THE-EVALUATION SET SPEAKERS WERE USED IN THIS EXPERIMENT. (WB: WIDE BAND, NB: NARROW BAND, AND GD: GENDER DEPENDENT)

recognizer	device	acoustic model	CER [%]	SER [%]	char-UB [mAP]	syl-UB [mAP]
reference	-	-	0.0	0.0	0.483	0.489
LVCSR (SAPI 5.0)	headset	WB/GD	21.0	16.9	0.398	0.425
		+adapt	12.9	9.2	0.444	0.463
	iPAQ	WB/GD	16.8	13.1	0.415	0.439
		+adapt	13.0	9.2	0.438	0.459
	cellular phone	NB/GD	74.6	71.0	0.082	0.091
		+adapt	50.6	45.9	0.217	0.228
syllable loop (HTK 2.2)	headset	WB/GD	-	37.7	-	0.241
		+adapt	-	34.4	-	0.248
	iPAQ	WB/GD	-	41.1	-	0.209
		+adapt	-	30.1	-	0.295
	cellular phone	NB/GD	-	46.1	-	0.134
		+adapt	-	40.2	-	0.168

result obtained for spoken queries recorded through cell phones is 0.134 (with the HTK syllable-loop recognizer). The result using the 16 KHz acoustic model in the SAPI 5.0 engine on 8 KHz data recorded from cellular phones is poor as expected. However, with five iterations of unsupervised self-adaptation for each spoken query, the error rate on the cellular-phone data dropped significantly (from 71% CER to 45.9% CER) and the resulting retrieval average precision increased to 0.228. Given that this is the result from self adaptation on just the testing query sentence, it is expected that an acoustic model trained from large amounts of speech collected over the telephone should achieve much lower error rates. For example, syllable error rate in the range of 22% has been reported for Mandarin syllable loop recognition on the Mandarin Across Taiwan (MAT) database [14].

The table also shows the effect of unsupervised self-adaptation, denoted by "+ adapt." Both character error rates (CER) and the syllable error rates (SER) are reduced by around 30% relatively after four rounds of iterative unsupervised adaptation, which we mostly attribute to the adaptation of the acoustic front-end's normalization parameters. Retrieval performance is improved by 10% on various indexing units.

We have also used HTK 2.2 to perform syllable recognition on our spoken queries and the results are also shown in Table III. Syllable error rate was lowest for headset-microphone data since they best match with the training data. The error rates on test data recorded on the cellular phone is worst because of the channel mismatch problem. The syllable error rates using the syllable-loop recognizer are substantially higher than those derived from the SAPI recognizer due to the following factors.

1) The SAPI 5.0 engine performs character recognition. Constraints from the lexicon and a powerful language model trained on over 1 billion characters of Chinese text are applied during recognition. Only afterwards, the character sequence is converted to syllables.
2) The amount of data used to train the syllable loop recognizer (100 speakers) is significantly less than the amount of data used to train the SAPI 5.0 Mandarin engine (over 1000 speakers).

Table IV shows the results of performing the same experiments on the TREC-6 set of queries. Although the general retrieval performance is substantially higher on the TREC-6 queries, the effects are widely consistent between TREC-5 and TREC-6. The recognition error rates for TREC-6 queries were significantly lower compared to TREC-5 queries for recognizers with similar settings. Accordingly, the gap between the mAP for spoken-query based systems was much closer to the mAP achieved with reference text input. For example, without self-adaptation on the incoming queries, the mAP of 0.439 was achieved for the iPAQ data set. With self-adaptation, the mAP increased to 0.459.

V. CONCLUSIONS

In this paper, we have described a system which allows the use of spoken queries to retrieve textual information from a database over mobile devices. We have demonstrated that retrieval performance on mobile devices with high-quality microphones such as the iPAQ PocketPC PDA is satisfactory compared with the performance one would obtain using a headset microphone or even with perfect text transcriptions. Such robust performance despite recognition errors demonstrates the effectiveness of using character and syllable bigram based approaches for indexing Mandarin text content. The performance of the system with data collected over cellular phones is far apart from those from the PDA. This result can be attributed to 1) the

reduction in information that occurs with the reduction in sampling rate from 16 000 Hertz to 8000 Hertz and 2) the lack of training data collected over the same channel. We introduced a channel mapping approach which maps the high bandwidth signal from the available training corpus to be more similar to waveforms collected over the cellular phones and obtained an improvement in recognition accuracy.

With the recent work in Aurora and distributed speech recognition (DSR) [18], it is anticipated that many mobile devices in the future will extract the features required for speech recognition on the device itself. This would allow for better extraction of suitable signals for speech recognition that will help to ameliorate the channel effects. Our result demonstrates that extracting features using the higher bandwidth signal should help substantially for the spoken query based information retrieval task. Another result in this paper demonstrates the effectiveness of adaptation to improve performance, both in terms of channel adaptation and unsupervised speaker adaptation. Future work will include creating systems which semi-autonomously adapt to data from different channels so that spoken queries from all channels can be recognized accurately.

VI. Future Work

For our future work, we plan to work toward a unified approach which will take into account the uncertainties in both speech recognition and information retrieval and provide them in a unified framework. Recent work in this area includes the work done by [16] and [17]. Also, a method to unify the large-vocabulary speech-recognition system and the syllable-based recognition and indexing system should provide better retrieval performance and provide more robustness toward the issue of the out-of-vocabulary problem.

Another area of future work is to create an appropriate display scheme for retrieval results. Since there are both relevant and irrelevant results for most queries, the organization of the returned results for efficient selection by the user is an important area worthy of studying. We are currently working on using text summarization techniques to efficiently display the gist of each returned article so that they are more easily scanned and reduce the size of area required to display each article [7]. Similarly, the clustering of search results based on their categories will speed up the scanning of the returned results [5]. Another issue worthy of studying is the feedback to be provided to the user in terms of what the system has recognized. While displaying the query recognized by the system may help the user to understand the search results, excessive misrecognitions in the recognized queries may confuse the user as well. Such issues will be explored further with future usability studies.

Acknowledgment

The system described has been built with contributions from many individuals. The authors wish to thank S. Ge, S. Di, D. Luan and C. Li for their contributions. We would also like to thank Dr. J. Gao and Dr. M. Zhou for providing the TREC Chinese database and many useful suggestions. They also gratefully acknowledge helpful comments and suggestions from the anonymous reviewers.

References

[1] B. Bai, B. Chen, H. M. Wang, L. F. Chien, and L. S. Lee, "Large-vocabulary Chinese text/speech information retrieval using Mandarin speech queries," in *Proc. Int. Symp. Chinese Spoken Language Processing*, 1998.

[2] E. Chang, J. Zhou, S. Di, C. Huang, and K. F. Lee, "Large vocabulary Mandarin speech recognition with different approaches in modeling tones," in *Proc. Int. Conf. Spoken Language Processing*, 2000.

[3] E. Chang, Y. Shi, J. Zhou, and C. Huang, "Speech lab in a box: A Mandarin speech toolbox to jumpstart speech related research toolbox," in *Proc. Eurospeech*, 2001.

[4] B. Chen, H. M. Wang, and L. S. Lee, "Retrieval of broadcast news speech in Mandarin Chinese collected in Taiwan using syllable-level statistical characteristics," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 2000.

[5] H. Chen and S. T. Dumais, "Bringing order to the web: Automatically categorizing search results," in *Proc. CHI'00, Human Factors in Computing Systems*, 2000.

[6] "Evolving the programs on a large scale," manuscript in preparation.

[7] S. Corston-Oliver, "Text compaction for display on very small screens," in *Proc. Workshop on Automatic Summarization, NAACL 2001*, Pittsburgh, PA, 2001.

[8] X. Huang and S. E. Robertson, "Okapi Chinese text retrieval experiments at TREC-6," in *Proc. 6th Text Retrieval Conf. (TREC-6)*, 1998, p. 137.

[9] X. D. Huang, A. Acero, C. Chelba, L. Deng, J. Droppo, D. Duchene, J. Goodman, H. Hon, D. Jacoby, L. Jiang, R. Loynd, M. Mahajan, P. Mau, S. Meredith, S. Mughal, S. Neto, M. Plumpe, K. Steury, G. Venolia, K. Wang, and Y. Wang, "Mipad, a multimodal interaction prototype," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 2001.

[10] K. S. Jones, G. J. F. Johns, J. T. Foote, and S. J. Young, "Experiments on spoken document retrieval," *Inform. Process. Manage.*, vol. 32, no. 4, pp. 399–417, 1996.

[11] J. Koza, *Genetic Programming*. Cambridge, MA: MIT Press, 1992.

[12] J. Kupiec, D. Kimber, and V. Balasubramanian, "Speech-based retrieval using semantic co-occurrence filtering," in *Proc. Human Language Technology Conf. (ARPA)*, 1994.

[13] K. L. Kwok, L. Grunfeld, and J. H. Xu, "TREC-6 English and Chinese retrieval experiments using PIRCS," in *Proc. 6th Text Retrieval Conf. (TREC-6)*, 1998.

[14] Y.-F. Liao, N. Wang, M. Huang, H. Huang, and F. Seide, "Improvements of the Philips 2000 Taiwan Mandarin benchmark system," in *Proc. Int. Conf. Spoken Language Processing*, Beijing, China, 2000.

[15] H. Meng, B. Chen, S. Khudanpur, G. Levow, W. K. Lo, D. Oard, P. Schone, K. Tang, H. M. Wang, and J. Wang, "Mandarin-English Information (MEI): Investigating translingual speech retrieval," in *Proc. Human Language Technology Conf.*, 2001.

[16] D. R. H. Miller, T. Leek, and R. M. Schwartz, "A hidden Markov model information retrieval system," in *Proc. SIGIR'99 ACM*, 1999.

[17] K. Ng, "Subword-Based Approaches for Spoken Document Retrieval," Ph.D. dissertation, Mass. Inst. Technol., 2000.

[18] D. Pearce and H.-G Hirsch, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. Int. Conf. Spoken Language Processing*, Beijing, China, 2000.

[19] G. Salton, A. Wong, and C. S. Yang, "A vector space model for information retrieval," *J. Amer. Soc. Inform. Sci.*, vol. 18, no. 11, pp. 613–620, Nov. 1975.

[20] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983.

[21] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inform. Process. Manage.*, vol. 24, no. 5, pp. 513–523, 1988.

[22] The Microsoft SAPI 5.0 SDK. [Online]. Available: http://www.microsoft.com/speech.

[23] Y. Shi, E. Chang, H. Peng, and M. Chu, "Power spectral density based channel equalization of large speech database for concatenative TTS system," in *Int. Conf. Spoken Language Processing*, 2002.

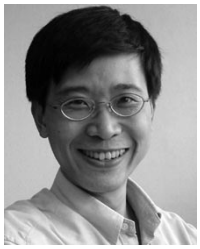
[24] A. Singhal, C. Buckley, and M. Mitra, "Pivoted document length normalization," in *Proc. 19th Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, 1996, pp. 21–29.

[25] X. Tong, C. X. Zhai, N. Milic-Frayling, and D. Evans, "Experiment on Chinese text indexing CLARIT TREC-5 Chinese track report," in *Proc. 5th Text Retrieval Conf. (TREC-5)*, 1997.

[26] *Proc. Text Retrieval Conf. (TREC)*, http://trec.nist.gov.

[27] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book, 2.2*.

[28] R. Wilkinson, "Chinese document retrieval at TREC-6," in *Proc. 6th Text Retrieval Conf. (TREC-6)*, 1998, pp. 25–30.



Eric Chang (M'96-SM'02) received the S.B., S.M., and Ph.D. degrees, all in electrical engineering, from the Massachusetts Institute of Technology, Cambridge, in 1990, 1990, and 1995, respectively.

He is currently Research Manager of the Speech Group at Microsoft Research Asia, which he joined in 1999. Prior to joining Microsoft Research, he was one of the founding members of the Research Group at Nuance Communications, where he worked from 1995 to 1999. He has also worked at MIT Lincoln Laboratory, Toshiba ULSI Research Center, and General Electric Corporate Research and Development Center. His research interests are speech technologies, machine learning and signal processing. He is currently serving on the editorial board of *Computer Speech and Language*.

Dr. Chang is a member of Tau Beta Pi, the IEEE Signal Processing Society, and the International Speech Communication Association.

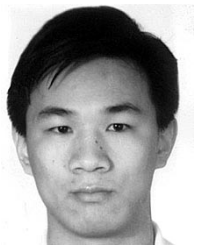
Frank Seide (M'94) was born in Hamburg, Germany. He received the Master degree ("Diplomingenieur") in electrical engineering from University of Technology of Hamburg-Harburg.

From 1993 to 1997, he worked on spoken-dialogue systems at the Speech Research Group of Philips Research in Aachen, Germany. He then transferred as one of the founding members of Philips Research East-Asia, Taipei, to lead a research project on Mandarin speech recognition. In June 2001, he joined the speech group at Microsoft Research Asia, Beijing. He has 19 scientific publications and ten patents. His current research interest includes speech-based information retrieval and novel approaches for acoustic modeling.

Helen M. Meng received the S.B., S.M., and Ph.D. degrees, all in electrical engineering from the Massachusetts Institute of Technology, Cambridge.

She has been a Research Scientist at the MIT Spoken Language Systems Group, where she worked on multilingual conversational systems. She is currently an Associate Professor in the Department of Systems Engineering and Engineering Management at The Chinese University of Hong Kong, which she joined in 1998. She established and directs the Human-Computer Communications Laboratory in the same department since 1999. Her research interest is in the area of human-computer interaction via multilingual spoken language systems, which integrate a suite of speech and language technologies, including speech recognition, natural language understanding, discourse and dialog modeling, language generation and speech synthesis. She is also working on translingual speech retrieval technologies.

She is a member of the IEEE Signal Processing Society and Sigma Xi.



Zhuoran Chen was born in China in 1977. He received the B.S. degree in computational mathematics from Zhongshan University, Guangzhou, China, in 1999. Currently he is pursuing the M.S. degree in computer science at Peking University, China.

His recent research interests include the theory and applications of computational complexity, evolutionary computation, bioinformatics, and information retrieval.

Yu Shi was born in Tianjin, China, in 1972. She received the B.S. degree in automation and the M.S. and Ph.D. degrees in pattern recognition and intelligent systems from Tsinghua University, Beijing, China, in 1996, 1998, and 2001, respectively.

Since 2001, she has been with Microsoft Research Asia, Beijing, as an Associate Researcher. Her current research interests are speech recognition and signal processing with applications in speech technologies. She has published 13 papers including two regular papers in the IEEE TRANSACTIONS ON SIGNAL PROCESSING. She also holds two patents on radar target recognition.

Yuk-Chi Li received the B.S. and M.Phil. degrees in 1999 and 2001, respectively, from the Department of Systems Engineering and Engineering Management, the Chinese University of Hong Kong.

His research interests are in Chinese spoken document retrieval, Chinese spoken query retrieval, and cross-lingual spoken document retrieval.