Discriminating Capabilities of Syllable-Based Features and Approaches of Utilizing Them for Voice Retrieval of Speech Information in Mandarin Chinese

Berlin Chen, Hsin-min Wang, *Member, IEEE*, and Lin-shan Lee, *Fellow, IEEE*

***Abstract*—With the rapidly growing use of the audio and multimedia information over the Internet, the technology for retrieving speech information using voice queries is becoming more and more important. In this paper, considering the monosyllabic structure of the Chinese language, a whole class of syllable-based indexing features, including overlapping segments of syllables and syllable pairs separated by a few syllables, is extensively investigated based on a Mandarin broadcast news database. The strong discriminating capabilities of such syllable-based features were verified by comparing with the word- or character-based features. Good approaches for better utilizing such capabilities, including fusion with the word- and character-level information and improved approaches to obtain better syllable-based features and query expressions, were extensively investigated. Very encouraging experimental results were obtained.**

***Index Terms*—Confidence measure, retrieval of speech information, syllable-based features, term association matrix.**

I. Introduction

DUE TO THE prevalence of the Internet, huge quantities of information are being accumulated very rapidly and being made available to users. As a result, the primary obstacle for people to access the information is no longer the spatial or temporal distances, but instead the lack of efficient ways to retrieve the desired information. Information retrieval techniques which provide the users with convenient access to the desired information are therefore extremely attractive [1]. Most of the works on information retrieval have been focused on approaches using text input to retrieve text information. Substantial efforts and very encouraging results have been reported and practically useful systems have been successfully implemented along this direction [2]–[6]. Recently, with the advances in speech recognition technology [7]–[11], proper integration of information retrieval and speech recognition has been considered by many researchers. But most of such works tried to handle either the text information retrieval using speech queries [12], [13] or the speech information retrieval using text queries [14]–[24]. Only very limited works have considered the problem of speech information retrieval using speech queries [25], [26]. With the rapidly growing use of audio and multimedia information on the Internet, an exponentially increasing number of voice records such as broadcast radio, television programs, digital libraries and so on, are now being accumulated and made available. However, most of them are simply stored there and difficult for further reuse because of the lack of efficient retrieval technology. Development of the technology to retrieve speech information is thus becoming more and more important. In any case, retrieval of huge quantities of speech information using speech queries directly is apparently the most natural, convenient and attractive, although the technology involved will be the most difficult as well. For the Chinese language, because the language is not alphabetic and there exist a huge number of commonly used Chinese characters, the input of Chinese characters into computers is a very difficult and unsolved problem even today. As a result, voice retrieval of speech information will be much more important and attractive for Mandarin Chinese than for other languages.

Unlike text information, speech information cannot be retrieved at all by directly comparing the input speech queries with the voice records. Not only can the vocabularies, texts, and topic domains spoken in the voice records and the speech queries be completely different, but the differences in acoustic conditions such as speakers, speaking modes, and background noises add further complication. Therefore, both the speech queries and the voice records must be transcribed into some kind of content features using speech recognition techniques, based on which the relevance between the speech queries and the voice records can then be measured. As a result, accurate recognition of Mandarin speech with a high degree of variability in vocabularies, topic domains and acoustic conditions is certainly the first key issue in the problem to be discussed here. Such a high degree of variability apparently makes the desired accurate recognition very difficult, and a substantial percentage of recognition errors will inevitably happen. Such speech recognition errors definitely make the information retrieval techniques considered here significantly different from those used in the conventional text retrieval approaches, and a very high degree of robustness in these retrieval techniques is obviously needed.

The second issue for voice retrieval of Mandarin speech information is to choose appropriate content features to represent both the voice records as well as the speech queries, so that they can be used in evaluating the relevance measure in the retrieval processes [10], [15], [26]. There can be at least three areas of approaches: keyword-based, word-based, and subword-based

Manuscript received February 27, 2001; revised October 23, 2001. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. C.-C. Jay Kuo.

The authors are with the Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C. (e-mail: berlin@iis.sinica.edu.tw; lsl@iis.sinica.edu.tw; whm@iis.sinica.edu.tw).

Publisher Item Identifier 10.1109/TSA.2002.802541.

approaches. For the keyword-based approaches [10], one can define a set of keywords for the voice records in advance, and whenever some keywords are spotted from the speech queries, the voice records with those or relevant keywords can then be retrieved. This approach is efficient and cost-effective, especially for retrieval of static databases for which the primary search words do not change frequently. However, it is not always easy to define a set of adequate keywords for all the speech documents to be retrieved even if we know the contents of all of them in advance, which is almost impossible especially when the speech documents keep on growing very fast on the Internet every day. The out-of-vocabulary problem always exists no matter how large the keyword set is. Also, spotting of keywords becomes difficult if the keyword set becomes very large. Such considerations naturally lead to the word-based approaches, in which large vocabulary recognition techniques instead of keyword spotting techniques are used. Once both the voice records and the input speech queries are fully recognized into texts (words/characters), many well-developed text retrieval techniques can be directly applied [17]–[20]. However, even for such an approach, the out-of-vocabulary problem is still an issue [27]–[29], since a large vocabulary speech recognizer also needs a predefined lexicon, and some special words important for retrieval purposes may be simply outside of this predefined lexicon, which is true for the Chinese language as explained below. This leads to the concept of making a comparison directly on the level of subword units instead, or the subword-based approaches. Because it is much easier to obtain all necessary subword units to cover all possible pronunciations of a given language, the out-of-vocabulary problem may be somehow avoided if the relevance measure is evaluated directly on the level of subword units instead of on the word level [15], [24], [28], [29]. Because in such approaches the subword units are not necessarily decoded into words, the retrieval is therefore not limited by a finite lexicon. Of course, it is always possible to try to integrate more than one of the approaches mentioned above, the keyword-based, the word-based, and the subword-based. But how much incremental discriminating capabilities can be obtained by such integration certainly requires more investigation.

In this paper, considering the monosyllabic structure of the Chinese language, the syllable-based approach is selected as a special case of the subword-based approach, and a whole class of indexing features for retrieval of Mandarin speech information using syllable-level statistical characteristics is investigated. The discriminating capabilities of such syllable-based approaches were verified by comparing to the character- and word-based approaches. The information fusion of indexing features of syllable-, character-, and word-levels as well as various improved approaches for better retrieval were also investigated. In all these studies, broadcast news is taken as the example database for retrieval due to its availability. The rest of this paper is organized as follows. Considerations of using syllable-level statistical characteristics are introduced in Section II. The broadcast news database and the speech recognition process used in this study are described in Sections III and IV. Section V presents the syllable-level indexing features and information retrieval model. Sections VI–VIII then present some initial experimental results. Sections IX and X further discuss improved techniques to produce better syllable-level indexing features and query expressions. Section XI gives the final experimental results comparing the discriminating capabilities of indexing features at different levels and the fusion of them when many improved approaches are applied. The concluding remarks are made in Section XII.

II. Considerations of Using Syllable-Level Characteristics for Mandarin Chinese

In the Chinese language, because each of the large number of characters (at least 10 000 are commonly used) is pronounced as a monosyllable, and is 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 a new word “電腦 (computer),” and the combination of the characters “股” “市 (market),” “長 (long),” and “紅 (red)” gives a new word “股市長紅 (stock price remains high for long)” in business news. In many cases, the meaning of these words more or less have to do with 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 just as the examples mentioned previously. Many of such words are very often the right key in information retrieval functions, because they usually 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 out-of-vocabulary problem is especially important for Chinese information retrieval, and this is a very important reason why the syllable-level statistical characteristics makes great sense in the problem here. In other words, the syllables represent characters with meaning, and in the retrieval processes they do not have to be decoded into words which may not exist in the lexicon.

Actually, the syllable-level information makes great sense for retrieval of Chinese information due to the more general monosyllabic structure of the language. Although there exist more than 10 000 commonly used Chinese characters, a nice feature of the Chinese language is that all Chinese characters are monosyllabic and the total number of phonologically allowed Mandarin syllables is only 1345. So a syllable is usually shared by many homonym characters with completely different meanings. Each Chinese word is then composed of from one to several characters (or syllables), thus the combination of these 1345 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 exist other important reasons to use syllable-level information. Because almost every Chinese character is a morpheme with its own meaning, thus very often plays quite independent linguistic roles. As a result, the construction of Chinese words from characters is very often quite flexible. One example phenomenon is that in many cases different words describing the same or similar concepts can be constructed by slightly different characters, e.g., both "中華文化 (Chinese culture)" and "中國文化 (Chinese culture)" means the same, but the second characters used in these two words are different. rarily abbreviated into shorter words, e.g., "國家科學委員會 (National Science Council)" can be abbreviated into "國科會," which includes only the first, the third and the last characters. Furthermore, an exotic word in foreign languages can very often be translated into different Chinese words based on its pronunciation, e.g., "Kosovo" may be translated into "科索沃 /ke1-suo3-wo4/ ," "柯索佛/ke1-suo3-fo2/ ," "克索夫 /ke4-suo3-fu1/ ," "科索伏 /ke1-suo3-fu2/ ," "科索佛 /ke1-suo3-fo2/," and so on, but these words usually have some syllables in common, or even exactly the same syllables. Therefore, an intelligent retrieval system needs to be able to handle such wording flexibilities, such that when the input queries include some words in one form, the desired audio recordings can be retrieved even if they include the corresponding words in other different forms. The comparison between the speech queries and the audio recordings directly at the syllable-level does provide such flexibilities to some extent, since the "words" are not necessarily constructed during the retrieving processes, while the different forms of words describing the same or relevant concepts very often do have some syllables in common.

III. Broadcast News Database

In this paper, the radio news was recorded using a wizard FM radio connected to a PC, and digitized at a sampling rate of 16 kHz with 16-bit resolution. The data were collected from several radio stations, all located at Taipei, Taiwan, from December 1998 to July 1999. All the recordings were manually segmented into stories and transcribed. The database to be retrieved consists of 757 recordings (about 10.2 h of speech), and was collected from the Broadcasting Corporation of China (BCC). Each recording is a short news abstract (about 50 s of speech on the average) produced by one out of about 20 different anchors, including both male and female. Some recordings in the database contain background music. A different broadcast news speech database consisting of 453 stories (about 4 h of speech) collected from Police Radio Station (PRS), UFO Station (UFO), Voice of Han (VOH), and Voice of Taipei (VOT) produced by a different group of speakers was used for training the speaker-independent HMMs for automatic recognition of the broadcast news speech. Table I shows the average speaking rates of the radio news from respective sources. It can be found that, on the average, there were 5.8 characters/s in the testing radio news speech (BCC), while the average speaking rate for the training speech was 5.4 characters/s.

TABLE I
Speaking Rates of the Broadcast News Speech Used in This Study

Data Types	Training					Testing
Sources	PRS	UFO	VOH	VOT	Average	BCC
Average Speaking Rates (characters/sec)	4.9	6.2	4.8	5.7	5.4	5.8

A set of 40 simple queries and the corresponding relevant news recordings were manually created to support the retrieval experiments. There were a total of 121 words (either monosyllabic or polysyllabic) in the 40 queries including 70% (85/121) of them treated as monosyllabic words. This is because many of the words in the queries were out-of-vocabulary words, such as personal names, exotic words, special terms for news events, and so on, and they simply were not included in the lexicon. This is a very normal situation in Chinese information retrieval. As long as a poly-character word is not included in the lexicon, in either speech recognition or word segmentation for texts it is simply taken as a string of a few mono-character words, because every Chinese character can always act as a mono-character word. As a result, in the query set each query contains three words, or four syllables (characters) on the average. On the other hand, each query had an average number of 23.3 relevant documents out of the 757 documents in the database, with the exact number ranging from 1 to 75. Two speakers (one male and one female) were asked to produce the 40 speech queries, respectively. To recognize the speech queries, another read speech database including 5.3 h of speech for phonetically balanced sentences and isolated words, produced by other 80 male and 40 female speakers, was used for training the speaker-independent HMMs for automatic recognition of the speech queries.

At the first glance, this spoken document retrieval task seems relatively easy since the retrieval target contains only the anchors' speech. However, this task was, in fact, rather difficult, because in almost all the cases, the query terms appeared at most only once, if not completely missing, in each of their relevant documents, and the queries were often very short (average 4.0 syllables). Detailed statistics of the queries and testing speech documents are shown in Table II.

IV. Mandarin Speech Recognition

A. Acoustic Processing and Language Modeling

Each frame of the speech data was represented by a 39-dimensional feature vector, consisting of 12 MFCCs and log energy, and their first and second differences. Utterance-based cepstral mean subtraction (CMS) was applied to all the training and testing materials. The acoustic units chosen for syllable recognition here were 112 right-context-dependent INITIALs and 38 context-independent FINALs, specially considering the phonetic structure of Mandarin syllables. Here, INITIAL is the initial consonant of the syllable and FINAL is the vowel (or diphthong) part but including optional medial or nasal ending. Each INITIAL was represented by an HMM with three states while each FINAL with four states. The Gaussian mixture number per state ranged from two to 16, depending on the quantity of training data. The silence model was a one-state HMM with 32 Gaussian mixtures trained with the nonspeech segments. In addition, the syllable-based and word-based N-gram language models were trained by a newswire text corpus consisting of 80 million Chinese characters collected from Central News Agency (CNA) in 1999, almost the same time frame as that when the broadcast news database used here

TABLE II
DESCRIPTION OF THE QUERIES AND TESTING SPEECH DOCUMENTS USED IN THIS STUDY

Number of Speech Documents to Be Retrieved	757 (10.2 Hrs)		
Number of Distinct Queries	40		
	Min.	Max.	Mean
Document Length (syllables/characters)	18	403	248.0
Query Length (syllables/Characters)	2	7	4.0
Number of Relevant Documents/Query	1	75	23.3

TABLE III
SYLLABLE, CHARACTER, AND WORD ACCURACIES FOR THE SPEECH QUERIES AND SPEECH DOCUMENTS, AFTER INTRODUCING THE WORD-LEVEL LANGUAGE MODEL

	Syllable	Character	Word
Speech Queries	89.20%	84.88%	72.62%
Speech Documents	73.37%	62.79%	43.37%

were collected. Word segmentation and phonetic labeling were performed for the training text corpus based on a 62k-word lexicon for training the N-gram language models.

B. *Speech Recognition*

A multipass search strategy was applied for speech recognition. In the first pass, Viterbi search [9] was performed based on the acoustic models and the syllable bigram language model, and the score at every time index was stored. In the second pass, a backward time-asynchronous A^* tree search [30], [31] generated the best syllable sequence based on the heuristic scores obtained from the first pass search and the syllable trigram language model. In the third pass, based on the state likelihood scores evaluated in the first pass search and the syllable boundaries of the best syllable sequence obtained in the second pass, the speech recognizer further performed Viterbi search on each utterance segment which may include a syllable and produced several most possible syllable candidates, and a syllable lattice was thus constructed. The syllable accuracies achieved at this stage for the speech documents and the speech queries were 64.9% and 81.5%, respectively. We further constructed the word graph from the syllable lattice based on the 62k-word lexicon mentioned above and performed dynamic programming on the word graph to find the best word sequence using the word unigram and bigram language models. The finally obtained word sequence was also automatically converted into its equivalent character- and syllable-level sequences to be used in the retrieval tasks. The final speech recognition results are presented in Table III. It can be found that the syllable accuracies for the speech documents and the speech queries were improved to 73.37% and 89.20%, respectively, while the character accuracies were 62.79% and 84.88%, and the word accuracies were 43.37% and 72.62%, respectively. Note that the word accuracies were relatively low and the character accuracies in the middle. One of the reasons is the relatively high out-of-vocabulary word percentages in the broadcast news database as discussed previously.

V. RETRIEVAL APPROACHES USING SYLLABLE-LEVEL STATISTICAL CHARACTERISTICS

A. *Syllable-Level Indexing Terms*

Here, a whole class of syllable-level indexing terms were carefully defined, including overlapping syllable segments with length $N(S(N),\ N\ =\ 1,\ 2,\ 3,\ 4,\ 5)$ and syllable pairs separated by n syllables $(P(n),\ n\ =\ 1,\ 2,\ 3,\ 4)$. Considering a syllable sequence of ten syllables $S_1S_2S_3\cdots S_{10}$, examples of the former are listed on the upper half of Table IV, while examples of the latter on the lower half of Table IV. For example, overlapping syllable segments of length 3 $(S(N),\ N\ =\ 3)$ include such segments as $(S_1S_2S_3)$, $(S_2S_3S_4)$, $(S_3S_4S_5)$, etc., while syllable pairs separated by 1 syllable $(P(n),\ n = 1)$ include such pairs as (S_1S_3), (S_2S_4), (S_3S_5), etc. Considering the structural features of the Chinese language, combinations of these indexing terms make good sense for retrieval purposes. For example, as mentioned previously, each syllable represents some characters with meaning, and very often words with similar or relevant concepts have some syllables in common, even if some of such words are out-of-vocabulary. Therefore syllable segments with length 1 $(S(N),\ N\ =\ 1)$ (nonoverlapping monosyllables in this case) make sense in retrieval. However, because each syllable is also shared by many homonym characters each with a different meaning, syllable segments with length 1 $(S(N),\ N = 1)$ alone definitely cause serious ambiguity. Therefore, they have to be combined with other indexing terms. In fact, in the Chinese language, about 91% of the top 5000 most frequently used polysyllabic words are bi-syllabic [32], i.e., they are pronounced as a segment of two syllables. Therefore, the syllable segments with length 2 $(S(N),\ N = 2)$ definitely carry a plurality of linguistic information, and make great sense to be used as important indexing terms. Similarity, if longer syllable segments such as $S(N),\ N = 3$, are matched between a document and the query, very often very important information for retrieval may be captured in this way. On the other hand, because of the very flexible wording structure in the Chinese language as described previously, syllable pairs separated by n syllables are helpful in retrieval. Considering the example mentioned previously in Section II, the word "國家科學委員會 (National Science Council)" may be abbreviated as "國科會" including only the first, third, and the last characters. Syllable pairs separated by n syllables become apparently useful in such cases. Furthermore, because substitution, insertion and deletion errors always happen during the syllable recognition process, such indexing terms as syllable pairs separated by n syllables are also helpful in handling such syllable recognition errors. In summary, the monosyllables in Chinese represent characters carrying some meanings and concepts and may somehow take care of the out-of-vocabulary problem. The ambiguity caused by homonym characters sharing the same monosyllable may then be clarified by overlapping syllable segments with length N, $N > 1$, and syllable pairs separated by n syllables. The former may capture the information of polysyllabic words or phrases which are important for retrieval, and the latter may handle to some extent the problems arising from the flexible wording structure in the Chinese language such as the abbreviation problem as well as those problems due to speech recognition errors including substitutions, insertions, and deletions.

B. Information Retrieval Model

Vector space models widely used in many text information retrieval systems were adopted here [33]. A document was represented by a set of feature vectors, each consisting of information regarding one type of indexing terms. This is slightly different from a former approach, in which a single feature vector consisting of information regarding all different types of indexing terms together was used [28]. In this research, nine types of indexing terms [$S(N)$, $N = 1\sim5$, and $P(n)$, $n = 1\sim4$] were used to construct nine feature vectors for each document d

$$\vec{d_j} = (x_{j1}, x_{j2}, \ldots, x_{jt}, \ldots, x_{jM_j}), \qquad j = 1, 2, 3, \ldots, 9. \tag{1}$$

In this equation, $\vec{d_j}$ is the feature vector for the jth type of indexing terms, for example, $j = 2$ for $S(N)$, $N = 2$. The tth component of $\vec{d_j}$, x_{jt}, represents the score for a specific indexing term t, for example, a specific syllable segment (s_k, s_l) for the case of $j = 2$ for $S(N)$, $N = 2$. M_j is the total number of different specific indexing terms for the jth type. The value of x_{jt} is obtained by

$$x_{jt} = \left[1 + \ln\left(\sum_{i=1}^{n_t} c_t(i)\right)\right] \cdot \ln(N/N_t) \tag{2}$$

where $c_t(i)$, ranging from zero to 1, is the acoustic confidence measure [31] evaluated for the ith occurrence of the specific indexing term t within the document d, and n_t is the total frequency counts for the occurrences of the specific indexing term t in the document. Therefore, the value of $[1 + \ln(\sum_{i=1}^{n_t} c_t(i))]$ denotes the term frequency of the specific indexing term t but evaluated in terms of the acoustic confidence measure, and the logarithmic operation is to compress its distribution. The value of $\ln(N/N_t)$ is the inverse document frequency (IDF), where N_t is the total number of documents in the collection in which the specific indexing term t appears, and N is the total number of documents in the collection. The value of x_{jt} in (2) is set zero if the specific indexing term t did not appear in the document d.

Each utterance of the speech documents can be transcribed into a syllable sequence, or a syllable lattice. For the case of a syllable lattice, each utterance segment O which may be a syllable can have several syllable candidates. For a certain syllable candidate s of the utterance segment O, the acoustic confidence measure $c(s)$ is obtained with the following Sigmoid function:

$$c(s) = \frac{2}{1 + \exp(-\alpha \times [\log p(O|s) - \log p(O|s^*)])} \tag{3}$$

where $\log p(O|s)$ and $\log p(O|s^*)$ are the original acoustic recognition scores of the syllable s and its corresponding top one syllable candidate s^*, respectively, and the value of α is used to control the slope of the Sigmoid function. From (3), it is clear that $c(s) = 1$ if $s = s^*$, or s being a top 1 candidate. Also, $c(s)$ is always between zero and 1. With $c(s)$ in (3), the acoustic confidence measure of a specific indexing term t, c_t, is simply the average of the acoustic confidence measures $c(s)$ in (3) for all syllables involved in the specific indexing term t. The term $c_t(i)$ in (2) is then the acoustic confidence measure c_t for the ith occurrence of the specific indexing term t.

TABLE IV
VARIOUS SYLLABLE-LEVEL INDEXING TERMS FOR AN EXAMPLE SYLLABLE SEQUENCE $S_1 S_2 S_3 \cdots S_{10}$

Syllable Segments	Examples
S(N), N=1	(s_1) (s_2) ...(s_{10})
S(N), N=2	$(s_1 s_2)$ $(s_2 s_3)$...$(s_9 s_{10})$
S(N), N=3	$(s_1 s_2 s_3)$ $(s_2 s_3 s_4)$...$(s_8 s_9 s_{10})$
S(N), N=4	$(s_1 s_2 s_3 s_4)$ $(s_2 s_3 s_4 s_5)$...$(s_7 s_8 s_9 s_{10})$
S(N), N=5	$(s_1 s_2 s_3 s_4 s_5)$ $(s_2 s_3 s_4 s_5 s_6)$...$(s_6 s_7 s_8 s_9 s_{10})$
Syllable Pair Separated by *n* Syllables	Examples
P(n), n=1	$(s_{1,} s_3)$ $(s_2 s_4)$ ...$(s_8 s_{10})$
P(n), n=2	$(s_1 s_4)$ $(s_2 s_5)$ ...$(s_7 s_{10})$
P(n), n=3	$(s_1 s_5)$ $(s_2 s_6)$ ...$(s_6 s_{10})$
P(n), n=4	$(s_1 s_6)$ $(s_2 s_7)$ ...$(s_5 s_{10})$

A speech query is also represented by a total of nine feature vectors in exactly the same way as the documents. The Cosine measure was used to estimate the query-document relevance for the jth type of indexing terms

$$R_j\left(\vec{q_j}, \vec{d_j}\right) = \left(\vec{q_j} \bullet \vec{d_j}\right) \Big/ \left(\|\vec{q_j}\| \cdot \left\|\vec{d_j}\right\|\right) \tag{4}$$

where $\vec{q_j}$ is the feature vector for the speech query using the jth type of indexing terms. The overall relevance measure was then the weighted sum of the relevance measures of all types of indexing terms

$$R\left(\vec{q}, \vec{d}\right) = \sum_j w_j \cdot R_j\left(\vec{q_j}, \vec{d_j}\right) \tag{5}$$

where w_j is a weighting parameter obtained empirically.

VI. Initial Experimental Results Using Syllable-Level Feature Alone

The experimental results are discussed starting this section. In this research, the retrieval results are expressed in terms of *non-interpolated average precision* [34]. The retrieval experiments using perfect manual transcriptions of texts for both the queries and documents were also evaluated for reference, denoted as TQ (**T**ext **Q**ueries) and TD (**T**ext **D**ocuments), as compared to those with the erroneous transcriptions obtained from speech recognition denoted as SQ (**S**peech **Q**ueries) and SD (**S**peech **D**ocuments). This is why there are four sets of test results for the several tables presented below: TQ/TD, SQ/TD, TQ/SD, SQ/SD. In the case of TD or TQ, the acoustic confidence measure $c_t(i)$ in (2) was simply replaced by unity.

In the first set of experiments, only syllable-level features as mentioned above were used. Only the top one syllable candidates were included here. Extension to syllable lattices will be discussed latter. The syllable-level indexing terms as mentioned in Section V-B above were automatically constructed from the top 1 sequences of the syllable lattices for the queries and documents. The retrieval results are summarized in Tables V and VI.

The retrieval performance using different combinations of overlapping syllable segments alone, not including syllable pairs separated by n syllables, is shown in Table V. In this table, the first column is the results for using $S(N)$, $N=1$, only, the second column is the results when $S(N)$, $N=2$, was used in addition, the third when $S(N)$, $N=3$, further used in addition, and so on. In each case when an extra type of indexing term was added, the weighting parameter w_j in (5) was tuned empirically to give the best results. For all four categories of retrieval, TQ/TD, SQ/TD, TQ/SD, SQ/SD, it can be found from Table V that $S(N)$, $N=1$, gives reasonable performance of retrieval, but the retrieval performance was significantly improved when the syllable segments for indexing were extended from $N=1$ to $N=2$, and another limited incremental improvement was obtained when extended to $N=3$. But the performance was kind of saturated when $N=3$. Further increasing N up to four or five was not helpful any more. Apparently $S(N)$, $N=1{\sim}3$, carry plenty of linguistic information for retrieval for the Chinese language probably due to the reasons mentioned in Section V-A, including the fact that 91% of the top 5000 most frequently used polysyllabic Chinese words are bi-syllabic, and words or phrases up to three syllables do bring very useful information. These results are in general in parallel with those obtained previously using the phone-level (a phone is a smaller acoustic unit than a syllable) indexing approaches for English [15], [24], [29]. When the syllable pairs separated by n syllables $P(n)$, $n=1, 2, 3, 4$, were additionally used as indexing terms with results shown in Table VI, it can be found that in general the retrieval performance can be further improved for n up to three probably due to the reasons mentioned in Section V-A. The improvements become relatively limited especially when the syllable pairs were separated by more syllables, probably because too many noisy terms also introduced unavoidable interferences in such cases. From Tables V and VI, the best combination of indexing features for the experimental task here is $S(N)$, $N=1{\sim}3$, plus $P(n)$, $n=1{\sim}3$, in general, which includes six types of indexing terms. This combination was thus used in all the following tests.

TABLE V
FOUR SETS OF RETRIEVAL RESULTS WITH RESPECT TO DIFFERENT COMBINATIONS OF OVERLAPPING SYLLABLE SEGMENTS $S(N)$ (TQ: TEXT QUERIES; SQ: SPEECH QUERIES; TD: TEXT DOCUMENTS; SD: SPEECH DOCUMENTS)

Average Precision	S(N), N=1	S(N), N=1~2	S(N), N=1~3	S(N), N=1~4	S(N), N=1~5
TQ/TD	0.4743	0.9656	0.9695	0.9695	0.9695
SQ/TD	0.4137	0.8898	0.8941	0.8940	0.8940
TQ/SD	0.3456	0.7009	0.7036	0.7034	0.7034
SQ/SD	0.3120	0.6583	0.6620	0.6620	0.6620

TABLE VI
FOUR SETS OF RETRIEVAL RESULTS WHEN THE OVERLAPPING SYLLABLE SEGMENTS $S(N)$, $N=1{\sim}3$, WERE COMBINED WITH SYLLABLE PAIRS SEPARATED BY n SYLLABLES, $P(n)$, FOR DIFFERENT n (TQ: TEXT QUERIES; SQ: SPEECH QUERIES; TD: TEXT DOCUMENTS, SD: SPEECH DOCUMENTS)

Average Precision	S(N), N=1~3 P(n), n=1	S(N), N=1~3 P(n), n=1~2	S(N), N=1~3 P(n), n=1~3	S(N), N=1~3 P(n), n=1~4
TQ/TD	0.9711	0.9726	0.9740	0.9742
SQ/TD	0.8946	0.8967	0.8982	0.8977
TQ/SD	0.7081	0.7142	0.7148	0.7128
SQ/SD	0.6681	0.6731	0.6739	0.6720

VII. COMPARING THE DISCRIMINATING CAPABILITIES OF SYLLABLE-LEVEL FEATURES WITH CHARACTER- AND WORD-LEVEL INFORMATION

Here, discriminating capabilities of syllable-level features in the Mandarin spoken document retrieval task were compared to those at the character- and word-levels. The fusion of information from these three different levels of knowledge will be investigated in the next section. For the character- and word-based approaches, the indexing features and feature vectors were constructed in exactly the same way and the information retrieval model was also exactly the same as those described in Section V, except that the units for indexing were characters and words instead of syllables. The combination of indexing features $S(N)$, $N=1{\sim}3$, and $P(n)$, $n=1{\sim}3$, was used here for evaluation and the retrieval results are presented in Table VII. Also listed in Table VII for reference, as the first and the second numbers in the parentheses, are the results obtained using only the indexing terms, $S(N)$, $N=1$, or the combination $S(N)$, $N=1{\sim}2$, i.e., with mono-character or mono-word, and overlapping segments of two characters or two words, respectively, for each case. The results of the syllable-based indexing approach in the first column of Table VII are simply copied from those in Tables V (first two columns) and VI (third column).

First, it is easy to see from the first two columns of Table VII that the retrieval performance for the character-based indexing approach (second column) is in close parallel to that of the syllable-based approach (first column) but always worse (0.8811 versus 0.8982 for SQ/TD, 0.6988 versus 0.7148 for TQ/SD, 0.6515 versus 0.6739 for SQ/SD), except when both queries and documents have perfect transcriptions, for which the character-based approach is slightly better (0.9778 versus 0.9740 for TQ/TD). This is reasonable taking into account the various considerations for the syllable-level information to be used for retrieval as mentioned above. For example, personal names or other out-of-vocabulary words including those with flexible structures may be recognized as wrong characters, but the syllables may directly carry the correct information, and so on. When both the queries and the documents have perfect transcriptions (for TQ/TD), however, such issues automatically disappear in most cases and the character-level information is much more precise (every syllable is shared by more than one homonym characters). Note that even in such case of TQ/TD, the character-based approach is only very slightly better than the syllable-based approach. The strong discriminating capabilities of the syllable-level features are apparently verified here. Also, because every character is pronounced as a monosyllable, there exists a clear syllable/character correspondence relationship, and the time span of a syllable segment $S(N)$ or the time separation between a syllable pair $P(n)$ are exactly the same for those of a corresponding character segment $S(N)$ or a corresponding character pair $P(n)$ with the same N or n. This may be the reason why the average precisions for the two approaches are in close parallel for the

TABLE VII
FOUR SETS OF RETRIEVAL RESULTS FOR COMPARISON OF THE SYLLABLE-, CHARACTER-, AND WORD-BASED INDEXING APPROACHES. THE FIRST AND SECOND NUMBERS IN THE PARENTHESES ARE THE RESULTS WHEN ONLY $S(N)$, $N = 1$ AND ONLY $S(N)$, $N = 1{\sim}2$, WERE USED RESPECTIVELY, WHILE THE NUMBERS IN FRONT OF THE PARENTHESES ARE THOSE WHEN $S(N)$, $N = 1{\sim}3, P(n)$, $n = 1{\sim}3$ ARE ALL USED TOGETHER

Average Precision	Syllable-based	Character-based	Word-based
TQ/TD	0.9740 (0.4743, 0.9656)	0.9778 (0.7680, 0.9604)	0.9027 (0.8804, 0.9003)
SQ/TD	0.8982 (0.4137, 0.8898)	0.8811 (0.6671, 0.8676)	0.7755 (0.7489, 0.7683)
TQ/SD	0.7148 (0.3456, 0.7009)	0.6988 (0.5577, 0.6872)	0.6160 (0.5988, 0.6138)
SQ/SD	0.6739 (0.3120, 0.6583)	0.6515 (0.5136, 0.6429)	0.5549 (0.5386, 0.5534)

cases of SQ/TD, TQ/SD, and SQ/SD. Furthermore, when only $S(N)$, $N = 1$, was used (first numbers in the parentheses), the character-based approach give much better results than the syllable-based approach (0.7680 versus 0.4743 for TQ/TD, 0.6671 versus 0.4137 for SQ/TD, 0.5577 versus 0.3456 for TQ/SD, 0.5136 versus 0.3120 for SQ/SD), apparently because mono-characters are much more precise and monosyllables cause much more ambiguities as mentioned above. However, when $S(N)$, $N = 2$, was added, very significant improvements were obtained in both cases (the second numbers versus the first numbers in the parentheses of the first two columns), and the results of the two approaches become in close parallel, obviously because of the plenty information carried by segments of two syllables or two characters, including the large number of frequently used bi-character (or bi-syllabic) words. Also, comparing the second numbers in the parentheses for the first two columns, it can be found that the syllable-based approach significantly outperforms the character-based approach as long as $S(N)$, $N = 2$, can be included. This again verified the strong discriminating capabilities of the syllable-level features.

For the word-based approach in the last column, on the other hand, the results for $S(N)$, $N = 1$ (the first numbers in the parentheses), are the best among the three columns (0.8804 for TQ/TD, 0.7489 for SQ/TD, 0.5988 for TQ/SD, 0.5386 for SQ/SD). The words obviously provide the most precise information than any subword units such as characters or syllables, if only one unit can be used. However, when more indexing terms can be included, the results for the word-based approach become always the worst in all cases, and in fact significantly worse than the corresponding syllable- or character-based approaches, even for the case of perfect transcriptions for both queries and documents (TQ/TD). The out-of-vocabulary problem and the very flexible wording structure mentioned previously are two of the many possible reasons. In addition, a Chinese word can be composed of one to several characters (or syllables), so the time span for a word segment $S(N)$ or the time separation for a word pair $P(n)$ can be much longer as compared to the corresponding syllable or character segments $S(N)$ or pairs $P(n)$ even for the same N or n, which may naturally introduce more interfering noisy indexing terms. From these experimental results, we can conclude that the subword-based (syllable- and character-based) approaches are better than the word-based approach for the Mandarin spoken document retrieval task, though many research results have indicated that the word-based approach is very useful in such tasks for western languages such as English [24], [35].

VIII. FUSION OF SYLLABLE-, CHARACTER- AND WORD-LEVEL INFORMATION

Although the syllable-based indexing features have been shown to provide very strong discriminating capabilities in Mandarin spoken document retrieval, the character- and word-level information does bring extra knowledge which does not exist in the syllable-level information. For example, the ambiguities caused by different homonym characters sharing the same syllable can be clarified by the characters, and the words carry much more semantic information than the syllables. It is therefore believed that a proper fusion of syllable-, character-, and word-level information would be helpful for Mandarin spoken document retrieval. To study the properties of such information fusion, the relevance measure between the query and document can be modified as

$$R\left(\vec{q}, \vec{d}\right) = w_s R_S\left(\vec{q}, \vec{d}\right) + w_c R_C\left(\vec{q}, \vec{d}\right) + w_w R_W\left(\vec{q}, \vec{d}\right),$$
$$w_s + w_c + w_w = 1 \quad (6)$$

which is simply the weighted sum of the relevance scores for the syllable-, character- and word-level indexing features, $R_S(\vec{q}, \vec{d})$, $R_C(\vec{q}, \vec{d})$, and $R_W(\vec{q}, \vec{d})$, respectively, as used in the above. The weighting parameters, w_s, w_c and w_w, were obtained empirically at the moment. The results of using indexing features $S(N)$, $N = 1{\sim}3$, and $P(n)$, $n = 1{\sim}3$, all together are shown in the last column of Table VIII ($S+C+W$, syllable plus character plus word). For comparison purposes, the results for the fusion of any two out of the three levels of information are also listed in the first three columns of Table VIII ($S + C$, syllable plus character, similarly $S + W$ and $C + W$).

Several interesting observations can be made from Table VIII. First, comparing the first, second, and last columns ($S + C$, $S + W$ and $S + C + W$) of Table VIII with the best results using the syllable-level information only in Table VII (first column), it is clear that in all cases adding either the character- or word-level information or both to the syllable-level information always gives better results. In other words, the extra knowledge brought by the character- and word-level information is apparently useful. Second, for cases with speech recognition errors in either queries or documents or both (SQ/TD, TQ/SD, SQ/SD), the fusion of character- and word-level information but not including the syllable-level information ($C + W$, the third column in Table VIII) gives significantly worse results than using the syllable-level information alone (first column in Table VII) (0.8788 versus 0.8982 for SQ/TD, 0.7003 versus 0.7148 for TQ/SD, 0.6513 versus 0.6739 for SQ/SD). This further verified that the discriminating capabilities of the syllable-level indexing features are very often stronger than those of the character- and word-level indexing features for the task studied here. The only exception is for the perfect transcription case (TQ/TC) (0.9803 versus 0.9740), in which the complete correct characters and words do provide more precise information than the syllables only. Third, the discriminating functions of syllable- and character-level information are apparently additive. The results in the first column of Table VIII $(S + C)$ is significantly better than using either the syllable-level information alone (first column

TABLE VIII
FOUR SETS OF RETRIEVAL RESULTS FOR FUSION OF THE SYLLABLE-(S), CHARACTER-(C), AND WORD-BASED (W) INDEXING APPROACHES. "$S + C$" MEANS SYLLABLE-LEVEL FEATURES PLUS CHARACTER-LEVEL FEATURES, AND SO ON

Average Precision	S+C	S+W	C+W	S+C+W
TQ/TD	0.9795	0.9758	0.9803	0.9797
SQ/TD	0.9045	0.9006	0.8788	0.9022
TQ/SD	0.7260	0.7213	0.7003	0.7267
SQ/SD	0.6829	0.6780	0.6513	0.6814

of Table VII) or the character-level information alone (second column of Table VII) in all four cases TQ/TD, SQ/TD, TQ/SD, SQ/SD. Fourth, similar additive discriminating functions can be more or less found, but not prominent, for the fusion of syllable- and word-level information (second column of Table VIII) as compared to each individual information (first and third columns of Table VII) and the fusion of character- and word-level information (third column of Table VIII) as compared to each individual information (second and third columns of Table VII). But, in both cases, the additional gain brought by the word-level information is relatively limited, and in fact for the latter case the extra word-level information did degrade the performance for the character-level information alone for SQ/TD and SQ/SD cases (comparing $C + W$ in the third column of Table VIII with the character-based alone in the second column of Table VII, 0.8788 versus 0.8811 for SQ/TD and 0.6513 versus 0.6515 for SQ/SD). Finally, comparing the first and the last columns in Table VIII ($S + C$ versus $S + C + W$) again indicates the same situation, i.e., adding the extra word-level information to the syllable- and character-level information yields only negligible improvements, if not degrading the performance. All of these again verified that the subword-level (syllable- or character-level) information is more useful in Mandarin spoken document retrieval.

IX. Improved Syllable-Level Indexing Features From Syllable Lattices

The syllable-based indexing approach discussed above has shown its strong discriminating capabilities for Mandarin spoken document retrieval tasks. However, the size of such syllable-based indexing terms can be huge if constructed from syllable lattices with multiple syllable candidates for each utterance segment which may include a syllable. In all the above experiments, in order to reduce the computation requirements, only the top one syllable candidates were used in constructing the syllable-based indexing terms, and as a result some information in the syllables lattice were inevitably lost. For example, the accuracies for the top one syllables used in the above experiments were 89.2% and 73.37%, respectively, for speech queries and speech documents, as shown in Table III. If the top five candidates in the syllable lattices can all be used to construct the indexing features, the inclusion rates for correct syllables can be as high as 92.28% and 84.78%, respectively, for speech queries and speech documents. However, including all the top five syllable candidates not only increases the computation requirements, but the wrong syllable candidates also introduce interfering indexing terms in the retrieval processes. It is therefore desirable to develop techniques to effectively condense the syllable lattices and select the most discriminating indexing terms, while deleting those syllable candidates or indexing terms which may not be helpful. In this section, three techniques along this direction are presented. They are the syllable-level utterance verification technique [31], the deletion of low-frequency indexing terms, and the stop terms.

A. Syllable-Level Utterance Verification (SUV)

When the number of syllable candidates for each utterance segment which may include a syllable (or the depth of the syllable lattices) is increased from one to m, the number of syllable segments $S(N)$ and syllable pairs separated by n syllables is increased from one to m^N and m^2, respectively. Although one of them may be exactly correct and provide the right information, the other $m^N - 1$ or $m^2 - 1$ indexing terms all carry one or more wrong syllables, and therefore are noisy terms and inevitably cause interferences in the retrieval processes. The situation becomes even worse when either m or N is larger. A relatively simple syllable-level utterance verification technique was thus used here. The basic idea is that any occurrence of the indexing terms with an acoustic confidence measure $c_t(i)$ in (2) below a pre-assigned threshold is simply deleted. The threshold can be different when constructing different types of indexing features. The results for such a simple verification approach are list in the first three columns of Table IX, which is for SQ/SD only. As can be found in the table, when using top one syllable candidates only (the first column), the total number of indexing terms for the six types of indexing terms discussed previously [$S(N), N = 1{\sim}3$, and $P(n), n = 1{\sim}3$] was 7.96 × 10^5. If all the top five syllable candidates were included (the second column), the total number of indexing terms was tremendously increased to 4.52 × 10^9 while the retrieval performance was slightly improved from 0.6739 to 0.6781. Apparently, the extra indexing terms did bring both useful information and noisy interferences. But when the indexing terms with lower acoustic confidence measures were deleted (the third column), the retrieval result was further improved to 0.6826 while the number of indexing terms reduced to 3.30 × 10^6.

B. Deletion of Low-Frequency Indexing Terms (DLF)

The contemporary newswire text corpus collected for language model training as described in Section IV-A can also provide very good information cues in identifying and filtering out the infrequent indexing terms. In other words, it is assumed here the statistical characteristics of syllables in the contemporary newswire text corpus were similar to that of the spoken document collection to be retrieved, and low-frequency indexing terms very often include some wrong syllables, thus, can be deleted. The statistical distributions of the indexing terms used here, $S(N), N = 1{\sim}3$, and $P(n), n = 1{\sim}3$, in the newswire text corpus were therefore calculated as the reference for pruning. Taking the indexing terms $S(N), N = 2$, for example, an specific indexing term composed of the segment of two syllables (s_k, s_l) was deleted if the ratio of the frequency counts of the segment (s_k, s_l) to the total of frequency counts

TABLE IX
RETRIEVAL RESULTS AND NUMBERS OF SYLLABLE-LEVEL INDEXING TERMS FOR THE SQ/SD CASE WHEN IMPROVED TECHNIQUES FOR GENERATING MULTIPLE INDEXING HYPOTHESES WERE INCORPORATED (SUV: SYLLABLE-LEVEL UTTERANCE VERIFICATION; DLF: DELETION OF LOW FREQUENCY INDEXING TERMS; ST: STOP TERMS)

		Syllable-based (*S(N)*, *N*=1~3, *P(n)*, *n*=1~3)				
		Top1	Top5	Top5 +SUV alone	Top5 +SUV+DLF	Top5 +SUV+ DLF +ST
SQ/SD	Average Precision	0.6739	0.6781	0.6826	0.6875	0.6901
	Indexing Term Size	$7.96\text{x}10^5$	$4.52\ \text{x}10^9$	$3.30\text{x}10^6$	$9.24\text{x}10^5$	$9.15\ \text{x}10^5$

of all possible segments of two syllables in the contemporary newswire text corpus is less than a pre-assigned value r_o. The pruning threshold r_o was different for different types of indexing terms. The results can be found in the fourth column of Table IX. When the deletion of low-frequency indexing terms (DLF) was additionally applied, the total number of indexing terms was reduced from 3.30×10^6 to 9.24×10^5, or only 27.23% of the original size, and in particular the number of the indexing terms $S(3)$ was actually dramatically reduced from 1.89×10^6 to 4.18×10^4. However, the average precision was improved from 0.6826 to 0.6875.

C. Stop Terms (ST)

In word-based information retrieval, a stop word list is usually used to remove the noncontent words. For the syllable-based approach developed here, a similar syllable-based stop term list can be constructed for the indexing terms used here based on the IDF scores in (2). For each type of indexing terms, say $S(N)$, $N = 1{\sim}3$, and $P(n)$, $n = 1{\sim}3$, the M most frequently occurring indexing terms (i.e., with the lowest IDF scores) were taken as the stop terms and removed from the indexing representations. The pre-assigned numbers of M for the stop terms were different for different types of indexing terms. We can see from the last column of Table IX that the retrieval performance was further improved to 0.6901 when the stop term list was applied additionally, although the size of indexing terms is reduced only very slightly.

X. FURTHER RETRIEVAL TECHNIQUES APPLIED ON SYLLABLE-BASED FEATURES

Quite several techniques have been proved to be effective for word-based approaches in retrieving both text and speech documents in western languages [20], [33], but it is relatively less known whether these techniques are equally effective for the Chinese language with syllable-based approaches. In this section, two prevailing techniques in this category are therefore applied on the syllable-based approach and investigated. They are the blind relevance feedback and the term association matrix.

A. Blind Relevance Feedback (BREF)

It has been found that some indexing terms not appearing in the query may still act as useful cues for relevance judgments. For example, the information from the relevant or irrelevant documents selected or deleted in the first stage retrieval can be further used to identify the indexing terms relevant to the user's intention. In this research, a blind relevance feedback procedure was used to reformulate the initial query expression automatically based on the modified Rocchio formula [20], [33]

$$\vec{q}' = \alpha \cdot \vec{q} + \beta \cdot \sum_{\vec{d}_i \in D_r} \vec{d}_i - \gamma \sum_{\vec{d}_j \in D_{irr}} \vec{d}_j \tag{7}$$

where $\vec{q}$ and $\vec{q}'$ are the initial and modified query feature vectors, D_r and D_{irr} are the sets of relevant and irrelevant documents, respectively, and α, β and γ are empirically adjustable weighting parameters. The results will be given in Section X-C.

B. Term Association Matrix (TAM)

The indexing terms co-occurring frequently within the same short passages of documents very often jointly describe some specific events, areas or topics, and thus may have some degree of synonymity association [33]. Based on this assumption, the database of speech documents to be retrieved as described in Section III was automatically divided into a total of 8152 passages based on the silence boundaries. A global association matrix A was then constructed for each type of the indexing terms, in which each entry $a(m, n)$ stands for some kind of association between two specific indexing terms t_m and t_n

$$a(m, n) = \frac{\hat{c}_{m,n}}{c_m + c_n - \hat{c}_{m,n}} \tag{8}$$

where c_m and c_n are, respectively, the total numbers of passages out of the 8152 in the entire database which include the indexing terms t_m and t_n, and $\hat{c}_{m,n}$ is the total number of passages out of the 8152 which include both t_m and t_n. For example, $a(m, n) = 1$ if t_m and t_n always appear in the same passage, and $a(m, n) = 0$ if t_m and t_n never appear in the same passage. The query feature vector is then reformulated by including in the new query expression a limited number of extra indexing terms which have the highest synonymity association to those nonzero indexing terms existing in the original query expression. The results will be given in the following.

C. Experimental Results

The retrieval results for the above schemes applied to the syllable-level indexing features studied here are summarized in the first row of Table X. The baseline result on the left is the highest result in Table IX. It can be found from this row of results that the blind relevance feedback previously found useful in word-based approaches for English is also useful in the syllable-based approach here for the Chinese language (0.6901 improved to 0.7111 for "+BREF alone"). However, the term association matrix only provided relatively limited improvements (0.6901 improved to 0.6962 for "+TAM alone"), probably because the synonymity association really makes good sense only when defined among words with semantics. But the syllables do represent multiple homonym characters with different meanings, therefore the synonymity association defined among syllable-level indexing terms also carry some limited information but some ambiguity as well. When both approaches of the blind relevance feedback and the term association matrix were applied together (0.7148 for

"+TAM+BREF"), the result was slightly better than the case of using the blind relevance feedback alone (0.7111 for "+BREF alone"). This again indicated that the synonymity association is helpful but with only limited effects.

XI. Further Comparison and Fusion With Character- and Word-Level Information

Although it has been shown in the previous two sections that some improved approaches may utilize the discriminating functions of syllable-level indexing features better, it was also found previously that the discriminating functions of character-level indexing features are in close parallel with those of syllable-level indexing features, and the fusion of syllable-, character- and word-level information is helpful. Therefore, in this section, we further applied some of the improved approaches mentioned above on character- and word-level indexing features, and tried to examine the results when different levels of indexing features were integrated. The results for SQ/SD case are listed in the rest part of Table X. The approaches of the blind relevance feedback and the term association matrix were directly applied to character- and word-level indexing features in exactly the same way as described previously, except the syllables were replaced by characters or words. Those approaches used for syllable lattices in Section IX, the utterance verification and some pruning techniques, on the other hand, are not necessarily equally applicable or effective for character-, and word-level indexing features. So they were not used here. The second and third rows of Table X are the results for character- and word-level indexing features individually, where the baseline on the left are the best results from Table VII. The lower rows of Table X are the results for information fusion.

Interesting observations can be made here. First, looking at the first three rows of Table X, it can be found that the blind reference feedback (+BREF alone) is equally effective for the subword-based (the character- and syllable-based) approaches as for the word-based approach (the retrieval performance was improved from 0.6901 to 0.7111 for the syllable-based approach, from 0.6515 to 0.6768 for the character-based, from 0.5549 to 0.5870 for the word-based). Query expansion based on the term association matrix (+TAM alone), on the other hand, is even more effective than the blind relevance feedback for the word-based approach (from 0.5549 improved to 0.5898 versus 0.5870), but achieves only relatively insignificant improvements for the subword-based approaches (from 0.6515 to 0.6573 for the character-based and from 0.6901 to 0.6962 for the syllable-based). One possible reason may be, as mentioned previously, that the word-level indexing terms carry more precise semantic information than the subword-level ones, therefore the word-level synonymity terms can offer much more information cues, while the subword-level synonymity terms cannot. Second, when the term association matrix and the blind relevance feedback were jointly applied (+TAM+BREF), the retrieval performance was only slightly better than using the blind reference feedback alone (+BREF alone). This is exactly the same for all the three cases, the syllable-, character- or word-based, probably because both the blind reference feedback and the term association matrix try to select extra indexing terms based on the given spoken document collection, which are somehow similar. The information cues captured by these two techniques are therefore additive only to a very limited extent. Third, the results for the syllable- and character-based approaches in the first two rows of Table X are again in close parallel, except those for the syllable-based approach (first row) are always about 0.035 to 0.040 higher. This is in good agreement with what was found previously in Table VII. Fourth, from the results for information fusion in the last four rows in Table X ($S + C$, $S + W$, $C + W$, and $S + C + W$), we can see that under all the four experimental conditions (baseline, +BREF alone, +TAM alone, and +TAM+BREF), the phenomena observed previously in Table VIII were almost completely reproduced here. For example, integrating either the character- or word-level information or both with the syllable-level information ($S+C$, $S+W$ or $S+C+W$) achieved better results than using the syllable-level information only, so the syllable-level information is additive with the character- and word-level information. Also, the fusion of character- and word-level information ($C + W$) gave slightly worse results than using the character-level information only. Many of such phenomena have been discussed in detail previously, so not repeated further here. But the strong discriminating capabilities of syllable-level indexing features are again verified here. Finally, with all these improved techniques (+TAM+BREF) and information fusion ($S+C+W$), the best average precision achieved was 0.7274 for the SQ/SD case, which is only slightly better than the best results using syllable-level features only (0.7148 on the right of the first row).

TABLE X

Retrieval Results for the SQ/SD Case for Syllable-, Character- and Word-Based Information and Fusion of Them, When the Two Improved Retrieval Techniques Were Applied (BREF: Blind Relevance Feedback; TAM: Term Association Matrix; "$S + C$" Means Syllable-Level Features Plus Character-Level Features, and So on)

Average Precision		Baseline	+ BREF alone	+ TAM alone	+ TAM+ BREF
SQ/SD	Syllable (S)	0.6901	0.7111	0.6962	0.7148
	Character (C)	0.6515	0.6768	0.6573	0.6781
	Word (W)	0.5549	0.5870	0.5898	0.5922
	S+C	0.6989	0.7210	0.7081	0.7233
	S+W	0.6952	0.7180	0.7073	0.7226
	C+W	0.6513	0.6757	0.6612	0.6772
	S+C+W	0.6988	0.7219	0.7078	0.7274

XII. Concluding Remarks and Future Work

This paper presents the initial results of a long-term research project toward voice retrieval of Mandarin speech information. Considering the monosyllabic structure of the Chinese language, a whole class of indexing features using syllable-level statistical characteristics was investigated. The experimental results have shown that the overlapping syllable segments with length N can capture the information of polysyllabic words or phrases, while the syllable pairs separated by n syllables can tackle the problems arising from the flexible wording structure, abbreviation, and speech recognition errors. The strong discriminating capabilities of the syllable-based features as compared to the word- or character-based were verified, and good approaches to integrate such capabilities with character- and word-level information to achieve the best retrieval performance were also investigated, including some improved techniques to obtain better indexing features or expanded query expressions. The techniques developed in this paper are currently being applied to the topic detection and tracking databases (TDT-2 and TDT-3) of Mandarin broadcast news in a next-stage project. This project is now in good progress, and initial results very similar to what were found here in this paper are being generated [36].

REFERENCES

[1] G. Salton, *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983.

[2] Yahoo home page. [Online]. Available: http://www.yahoo.com.

[3] Excite home page. [Online]. Available: http://www.excite.com.

[4] Alta Vista home page. [Online]. Available: http://www.altavista.com.

[5] Google home page. [Online]. Available: http://www.google.com.

[6] Openfind home page. [Online]. Available: http://www.openfind.com.tw.

[7] K. F. Lee, *Automatic Speech Recognition: The Development of the SPHINX System*. Norwell, MA: Kluwer, 1989.

[8] J. G. Wilpon, L. R. Rabiner, C. H. Lee, and E. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 1870–1878, Nov. 1990.

[9] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[10] J. T. Foote, G. J. F. Jones, K. Spärck Jones, and S. J. Young, "Talker—Independent keyword spotting for information retrieval," in *Proc. Eur. Conf. Speech Communication Technology*, 1995, pp. 2145–2148.

[11] H. M. Wang and L. S. Lee *et al.*, "Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary using limited training data," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 195–200, Mar. 1997.

[12] J. Kupiec, D. Kimber, and V. Balasubramanian, "Speech-based retrieval using semantic co-occurrence filtering," in *Proc. Human Knowledge Technology Workshop*, 1994, pp. 373–377.

[13] L. F. Chien, S. C. Lin, and L. S. Lee *et al.*, "Internet Chinese information retrieval using unconstrained mandarin speech queries based on a client-server architecture and a PAT-tree-based language model," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, vol. 2, 1997, pp. 1155–1158.

[14] D. A. James, "A system for unrestricted topic retrieval from radio news broadcasts," in *Proc. Int. Conf. Acoustic, Speech, Signal Processing*, 1996, pp. 279–282.

[15] K. Ng and V. Zue, "Subword unit representations for spoken document retrieval," in *Proc. Eur. Conf. Speech Communication Technology*, 1997, pp. 1607–1610.

[16] M. Wechsler, "Spoken document retrieval based on phoneme recognition," Ph.D. dissertation, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland, 1998.

[17] J. Allan, J. Callan, W. Croft, L. Ballesteros, D. Byrd, R. Swan, and J. Xu, "INQUERY does battle with TREC-6," in *Proc. 6th Text Retrieval Conf. (TREC-6)*, 1998.

[18] D. Abberley, S. Renals, G. Cook, and T. Robinson, "The THISL spoken document retrieval system," in *Proc. 6th Text Retrieval Conference (TREC-6)*, 1998.

[19] S. E. Johnson, P. Jourlin, G. L. Moore, K. Spärck Jones, and P. C. Woodland, "The Cambridge University spoken document retrieval system," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1999.

[20] A. Singhal and F. Pereira, "Document expansion for speech retrieval," in *Proc. ACM SIGIR Conference on R&D in Information Retrieval*, 1999.

[21] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, and R. Schwartz, "Speech and language technologies for audio indexing and retrieval," *Proc. IEEE*, vol. 88, pp. 1338–1353, Aug. 2000.

[22] S. Renals, D. Abberley, D. Kirby, and T. Robinson, "Indexing and retrieval of broadcast news," *Speech Commun.*, vol. 32, pp. 5–20, 2000.

[23] P. Jourlin, S. E. Jonson, K. Spärck Jones, and P. C. Woodland, "Spoken document representations for probabilistic retrieval," *Speech Commun.*, vol. 32, pp. 21–36, 2000.

[24] G. Ng, R. Wilkinson, and J. Zobel, "Experiments in spoken document retrieval using phoneme N-grams," *Speech Commun.*, vol. 32, pp. 61–77, 2000.

[25] G. J. F. Jones, J. T. Foote, K. Spärck Jones, and S. J. Young, "Video mail retrieval using voice: An overview of the stage 2 system," in *Proc. MIRO Workshop*, 1995.

[26] B. Chen, H. M. Wang, and L. S. Lee, "Retrieval of mandarin broadcast news using spoken queries," in *Proc. Int. Conf. Spoken Language Processing*, 2000.

[27] P. C. Woodland, S. E. Johnson, P. Jourlin, and K. Spärck Jones, "Effects of out of vocabulary words in spoken document retrieval," in *Proc. ACM SIGIR Conference on R&D in Information Retrieval*, 2000.

[28] 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. Int. Conf. Acoustics, Speech, Signal Processing*, 2000.

[29] S. Srinivasan and D. Petkovic,, "Phonetic confusion matrix based spoken document retrieval," in *Proc. ACM SIGIR Conf. R&D Information Retrieval*, 2000.

[30] P. Kenny, R. Hollan, V. N. Gupta, M. Lennig, P. Mermelstein, and D. O'Shaughnessy, "A*—Admissible heuristics for rapid lexical access," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 49–58, Jan. 1993.

[31] B. Chen, H. M. Wang, L. F. Chien, and L. S. Lee, "A*—Admissible keyphrase spotting with sub-syllable level utterance verification," in *Proc. Int. Conf. on Spoken Language Processing*, 1998.

[32] CKIP Group, "Analysis of syntactic categories for Chinese," Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C., CKIP Tech. Rep. 93-05, 1993.

[33] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. New York: ACM Press, 1999.

[34] D. Harman. (1995) Overview of the fourth text retrieval conference (TREC-4. [Online]. Available: http://trec.nist.gov/pubs/trec4/overview.ps.

[35] K. Ng, "Information fusion for spoken document retrieval," in *Proc. Int. Conf. on Acoustic, Speech, Signal Processing*, 2000.

[36] B. Chen, H. M. Wang, and L. S. Lee, "An HMM/N-gram-based linguistic processing approach for Mandarin spoken document retrieval," in *Proc. Eur. Conf. Speech Communication Technology*, 2001.

Berlin Chen was born in 1971. He received the B.S. and M.S. degrees in computer science and information engineering from National Chiao Tung University (NCTU), Hsinchu, Taiwan, R.O.C., in 1994 and 1996, respectively. He received the Ph.D. degree in computer science and information engineering from National Taiwan University (NTU), Taipei, Taiwan, in 2001.

From October 1996 to September 2001, he was with the Chinese Knowledge Information Processing (CKIP) Group, Institute of Information Science, Academia Sinica, Taipei. He is currently a Postdoctoral Researcher with the Speech Processing Laboratory, NTU. His research has been focused on speech recognition and information retrieval.

Hsin-min Wang (S'92–M'95) was born in Taiwan, R.O.C., in 1967. He received the B.S. and Ph.D. degrees in electrical engineering from National Taiwan University (NTU), Taipei, Taiwan, in 1989 and 1995, respectively.

In October 1995, he joined the Institute of Information Science, Academia Sinica, Taipei, as a Postdoctoral Researcher. In November 1996, he was promoted to Assistant Research Fellow. From February 2000 to July 2001, he was an adjunct Assistant Professor with the Department of Electrical Engineering, National Taipei University of Technology. His research interests include speech processing, natural language processing, spoken dialogue processing and multimedia information retrieval.

Dr. Wang is a member of ISCA.

Lin-shan Lee (S'76–M'77–SM'88–F'93) received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA.

He has been a Professor of electrical engineering and computer science at the National Taiwan University, Taipei, Taiwan, R.O.C., since 1982; he was a Department Head from 1982 to 1987. He holds a joint appointment as a Research Fellow of Academia Sinica, Taipei, and was an Institute Director there from 1991 to 1997. His research interests include digital communications and Chinese spoken language processing. He developed several of the earliest versions of Chinese spoken language processing systems in the world, including text-to-speech system, natural language analyzer, and dictation systems.

Dr. Lee was the Guest Editor of a special issue on intelligent signal processing in communications of the IEEE JOURNAL OF SELECTED AREAS IN COMMUNICATIONS in December 1994 and January 1995. He was the Vice President for International Affairs (1996–1997) and the Awards Committee Chair (1998–1999) of the IEEE Communications Society. He has been a member of Permanent Council of International Conference on Spoken Language Processing (ICSLP).