TOWARDS A MANDARIN VOICE MEMO SYSTEM

Hsin-min Wang[1], Bor-shen Lin[2], Berlin Chen[1,2], and Bo-ren Bai[2]

[1] Institute of Information Science, Academia Sinica
[2] Department of Electrical Engineering, National Taiwan University
Taipei, Taiwan, Republic of China
e-mail: whm@iis.sinica.edu.tw

ABSTRACT

Using voice memos in stead of text memos is believed to be more natural, convenient, and attractive. This paper presents a working Mandarin voice memo system that provides functions of automatic notification and voice retrieval. The main techniques include the content-based spoken document retrieval approach and the date-time expression detection and understanding approach. Extensive preliminary experiments were performed and encouraging results were demonstrated.

1. INTRODUCTION

Using voice memos in stead of text memos is believed to be more natural, convenient, and attractive because it is definitely much easier for people to speak memos than to write down memos or to type memos into computers using a keyboard. Furthermore, users are more likely to record detailed information if all they need to do is just speak. However, it's far more difficult to retrieve these voice memos than to retrieve the text ones. With advances in the speech recognition technology, voice retrieval of spoken documents has become feasible [1][2]. However, even if a general content-based spoken document retrieval technology is available, it is certainly not enough for a voice memo system. In general, the contents of memos mainly include both date-time expressions and what-to-do expressions. For example, in a Mandarin memo "明天晚上六點我要跟王先生吃晚飯(I will have dinner with Mr. Wang at 6:00pm tomorrow)", "明天晚上六點 (6:00pm tomorrow)" belongs to a date-time expression while "我要跟王先生吃晚飯 (I will have dinner with Mr. Wang)" is a what-to-do expression. That is, users can retrieve their memos using either date-time expression queries such as "明天晚上六點 (6:00pm tomorrow)" and "我明天晚上六點有事嗎(Is there anything I need to do at 6:00pm tomorrow)", or other natural language queries such as "我什麼時候要跟王先生吃晚飯(When will I have dinner with Mr. Wang)" and simple queries that probably contain only several keywords related to the things to do such as "跟王先生吃晚飯 (have dinner with Mr. Wang)". Thus, in addition to the content-based spoken document retrieval technology, date-time expression detection and understanding [3][4] is also necessary for automatic notification and voice retrieval using speech queries containing date-time expressions. This paper presents the overall architecture of our working system for Mandarin voice memo retrieval and key techniques that have been developed, including a content-based spoken document retrieval approach and a date-time expression detection and understanding approach.

The rest of this paper is organized as follows. The overall architecture of the proposed approach for Mandarin voice memo retrieval is introduced in section 2. The date-time expression detection and understanding approach and the content-based spoken document retrieval approach are described in sections 3 and 4 respectively. Finally the concluding remarks are made in section 5.

2. OVERALL ARCHITECTURE OF THE PROPOSED APPROACH FOR MANDARIN VOICE MEMO RETRIEVAL

Figure 1 shows the block diagram of the proposed Mandarin voice memo system. Given a new voice memo, it is first added to the voice memo database D. The keyword spotter [5] is then applied to construct a keyword graph, based on which the date-time expression part can be understood and transcribed into a year/month/date/time type of knowledge representation, and the date-time representation will be added to the date-time representation database D_T. The date-time representation database plays two roles here. The first one is the target database for retrieving using queries containing date-time expressions, while the second one is for automatic notification. Then, the large vocabulary continuous speech recognizer is applied to the residual speech segments, i.e., the what-to-do expression part, to construct a syllable lattice with the corresponding acoustic score for each syllable candidate. Based on the syllable lattice, the feature extraction module then extracts the desired feature vector to represent the what-to-do part for retrieval. The whole process can be performed off-line for all voice memos in database D to form the corresponding feature vector database D_V, and the date-time representation database D_T, both will be the target databases to be physically retrieved. Since the most time consuming speech recognition process is performed off-line in advance, and all information necessary for retrieval are stored in the feature vector database and date-time representation database, this architecture is very adequate for fast voice memo retrieval.

Given a speech query, exactly the same procedure discussed above is first applied to obtain the corresponding feature vector V_q or date-time representation T_q. Then, the matching module performs feature vector matching between V_q and all V_d's in D_V or date-time representation matching between T_q and all T_d's in D_T. Finally, the decision module chooses the voice memos that most fit the speech query as the results.

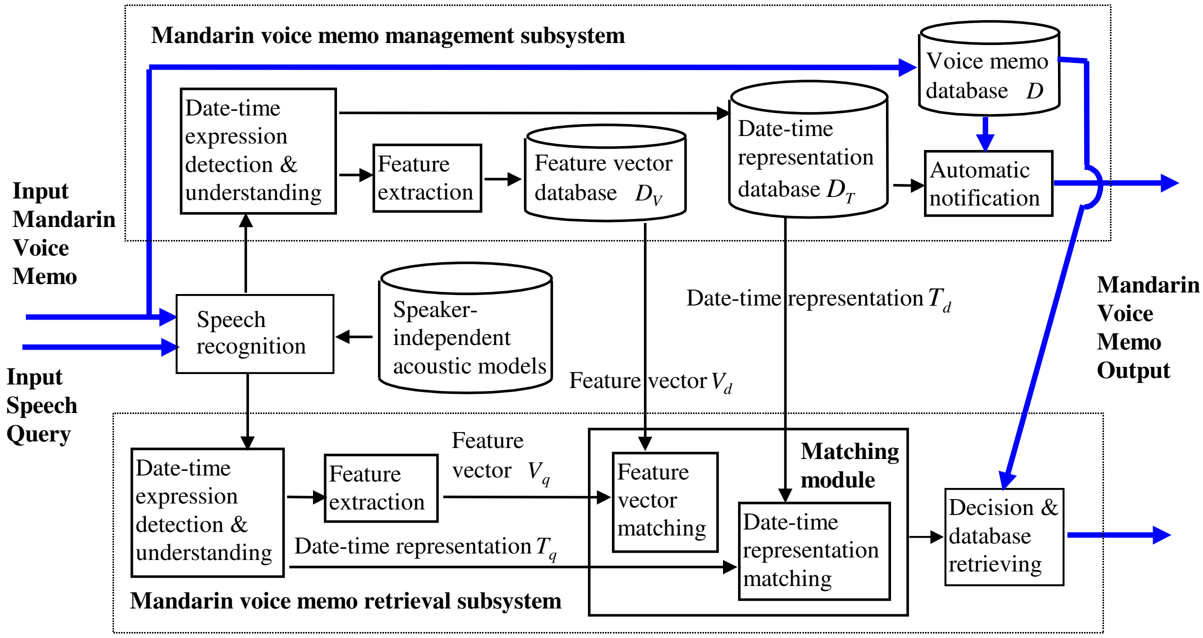

Figure 1: The block diagram of a Mandarin voice memo system.

3. DATE-TIME DETECTION AND UNDERSTANDING

The keyword set for date-time expression used here consists of a total of 375 keywords, such as "今天" ("today"), "星期一" ("Monday"), "早上" ("morning"), and so on. The kernel of the date-time detection and understanding approach is a hierarchical tag-graph search scheme [6]. Each keyword for date-time expressions is first assigned a "semantic tag". Grammar rules are then developed based on these semantic tags. A layering algorithm is then used to construct the "hierarchy" for all the semantic tags and associated grammar rules. For example, considering the grammar rule "HOUR: NUMBER + O'CLOCK" for date-time expressions, the tag "HOUR", is automatically promoted to a layer higher than the tags "NUMBER" and "O'CLOCK", because the knowledge regarding "HOUR" should be determined after those for "NUMBER" and "O'CLOCK". According to the layers that the tags are assigned to, all the grammar rules can then be used to construct a set of grammar trees for all the different layers. Now, for the input utterances the keyword graph is used as the initial graph in a bottom-up search algorithm, in which all the higher-layer tag graphs are generated hierarchically one by one. During this bottom-up search, the lower-layer tags are "merged" into higher-layer tags. A best first search, which searches only on such target-tags as "DATE", "TIME" or "TIME-RANGE", is further applied to find the top-N tag sequences on the tag-graph hierarchy under the constraints of the date-time domain knowledge. These top-N tag sequences with associated parsing trees are then sent to a semantic transcription module, in which knowledge correctness is checked and phrase verification is done, and finally transcribed into associated semantic slots for speech understanding.

To retrieve the semantic meaning of tag sequence, each tag in grammar rules is first attached with a symbol that represents its semantic meaning. The tag "WEDNESDAY", for example, is attached with a symbol "3" which represents index of this weekday while the tag "明天"("tomorrow") is attached with a symbol "+1" which represents the date by reference of today. Then each tag in the sequence with attached symbol is further transcribed into semantic slots, such as year, month, date, and time. Based on this approach, not only the semantically simple tags such as "tomorrow" can be expressed and transcribed, but the complicated ones, such as "mothers' day", can be interpreted correctly. During semantic transcription, date-time knowledge correctness such as range of value is also checked to reject those sequences with knowledge inconsistencies. Furthermore, phrase verification can be applied to reject those sequences with low confidences.

	Voice Memos		Simple Queries	
	No Veri.	With Veri.	No Veri.	With Veri.
Ins.	4.92%(10)	1.65%(3)	0.00%(0)	0.00%(0)
Del.	10.34%(21)	12.09%(22)	2.94%(7)	4.72%(11)
Sub.	25.61%(52)	14.29%(26)	8.82%(21)	9.87%(23)
Accuracy	59.11%	71.98%	88.24%	85.41%
Phrase no.	203	182	238	233

Table 1: Results of key-phrase detection and understanding.

3.1 Experimental results

The experiments discussed in this section were performed based on 102 voice memos, each of which contains both what-to-do and date-time expressions, and 100 simple speech queries, each of which contains only date-time expressions. Practically, the speech queries can contain other phrases such as "please show me", "I'd like to know", and so on, in addition to the date-time expression part. Such quasi-natural-language queries are in fact very similar to voice memos in the structure except that they are usually much shorter than voice memos. Thus, only voice memos and simple queries were used for evaluation here. Table 1 shows the accuracy of key-phrase detection and understanding. Note that, here the date-time expression part of each utterance (voice memo and speech query) may contain several phrases, such as "DATE", "TIME-RANGE" and "TIME", and each phrase may contain several semantic slots, e.g. the phrase "DATE" may contain year, month and date. A total of 203 and 238 phrases were found from the 102 voice memos and 100 speech queries, respectively. For voice memos, without phrase verification, the average phrase accuracy is only 59.11%. It was found that for those voice memos containing only one date-time phrase the accuracy is actually very poor, while for those with three date-time phrases the accuracy is much better. On average, the date-time expression part accounts for only 22.82% of the voice memos in length, i.e., the what-to-do expression is about three times longer than the date-time expression. Thus, those voice memos with only one date-time phrase are very likely to be inserted by the fake date-time phrases spotted from the very long what-to-do expression part. When phrase verification was applied to filter out the date-time phrases with lower scores, the phrase spotting accuracy was improved to 71.98% at a rejection rate of 10.34%. For simple queries that have no what-to-do part, the phrase accuracy was 88.24%, which is much better than that of voice memos as expected. However, the accuracy degraded to 85.41% with phrase verification applied. It is because most of the errors occurred in simple queries are from the highly acoustically confusing phrases that can not be rejected by phrase verification reliably.

The experimental results reported here are actually not good enough, a dialogue control automata for generating response to the user for confirmation or correction is therefore under study.

4. CONTENT-BASED SPOKEN DOCUMENT RETRIEVAL

Unlike the text documents, the spoken documents such as voice memos can't be retrieved by directly comparing them with the speech queries. Both the speech queries and the spoken documents must be transcribed into some kind of content features such as phone strings, texts, keywords and so on using speech recognition techniques, based on which the similarity between the speech queries and the spoken documents can then be measured. To choose appropriate content features is therefore very important. In Mandarin Chinese, the combination of the 416 syllables gives almost unlimited number of monosyllabic or polysyllabic words. The special monosyllabic structure of the Chinese language makes it possible to compare the spoken documents and the speech queries directly at the syllable level. Due to limited number of syllables, each syllable is usually shared by many homonym characters. So a single syllable in principle dose not carry enough information for information retrieval, especially when it is used alone. On the other hand, the combination of several syllables very often corresponds to unique, or very few, polysyllabic words, which will be much more powerful in retrieving. But there can be huge number of combinations of several syllables, which is difficult to handle. The simplest case of combination of several syllables is apparently the syllable pair. In fact, in Chinese language, most frequently used polysyllabic words are bi-syllabic, i.e., they are pronounced as a syllable pair. Therefore, syllable pairs certainly carry plurality of linguistic information, which may be useful for information retrieval. However, a characteristic feature of the Chinese language is the very flexible wording structure. So the same concept appearing in the query and in the desired relevant spoken documents may be represented by two words in different forms, which certainly gives different syllable pairs. However, very often in such cases the two words representing the same concept do have some syllables in common. Furthermore, due to the difficulties in speech recognition, the inevitable recognition errors such as deletions, insertions, and substitutions occurring in both the speech queries and the spoken documents certainly seriously disturb the syllable pair information in the syllable lattices as well. To deal with these problems, the feature vector adopted here contains both the syllable information and the syllable pair information. The basic idea is that even if a single syllable does not carry too much information due to large number of homonym characters sharing the same syllable, it may bring extra information if used in addition to the syllable pairs, especially when the syllable pair information may be disturbed or even destroyed due to the reasons mentioned above.

For each voice memo *d*, by searching through the syllable lattice of the what-to-do expression part, all acoustic scores of syllables and adjacent syllable pairs in the syllable lattice can be extracted to form the feature vector V_d ,

$$\begin{aligned} V_d = (&as(s_1),...,as(s_i),...,as(s_{416}),\\ &as(s_1,s_1),...,as(s_i,s_j),...,as(s_{416},s_{416})) \end{aligned} \tag{1}$$

where $as(s_i)$ is the acoustic score of the syllable s_i , and $as(s_i,s_j)$ is the acoustic score of the syllable pair (s_i,s_j) . Similarly, for any speech query, a feature vector V_q can be obtained via exactly the same procedure but on-line in real-time.

4.1 Retrieving process

Given the voice memo database *D* and its corresponding feature vector database D_V and a speech query *q*, the retrieving problem is now reduced to a search process to find the voice memo *d** (or the top *n* voice memos) in the voice memo database *D* that is most relevant to the query. Here, a Cosine measure which is often used in information retrieval [7] is used to estimate the similarity between a query *q* and the voice memos *d*:

$$Sim(d,q) = \cos(V_d,V_q) = \frac{V_d \cdot V_q}{|V_d||V_q|}, for d \in D \tag{2}$$

Voice memos with higher $\cos(V_d,V_q)$ values will thus be ranked and selected as the results.

4.2 Experimental results

The example speech database to be retrieved in the following experiments consists of 500 Mandarin spoken documents for Chinese news, which were produced by 5 male speakers. On average, each document contains about 100 characters (i.e., 100 syllables), while the individual length ranges from 44 to 269 characters. On the other hand, 160 speech queries produced by 4 male speakers were used for testing. 80 of them are simple queries each containing only one key phrase for some news items without any irrelevant words. An example key phrase is

“中央研究院(Academia Sinica)”. The other 80 queries are in the quasi-natural-language mode including some irrelevant words in addition to the key phrases. For example, “有沒有關於中央研究院的新聞？(Is there any news about Academia Sinica?)”. For the 500 spoken documents to be retrieved, those documents relevant to each of the 160 queries were identified manually in advance for performance evaluation purposes. The number of documents relevant to each query ranges from 1 to 20.

Though the speech database used here for testing is not a real voice memo database. The database scale and the document length are actually very similar to that of a voice memo database. The experimental results reported here thus can reflect the situation of using speech queries to retrieve a voice memo database with a content-based spoken document retrieval approach.

The testing results are plotted in the recall-precision graph as were often done in information retrieval [8] in Figure 2, in which the recall rate on the horizontal scale is the average percentage of the spoken documents in the database manually identified to be relevant to the desired subject which are actually retrieved correctly, and the precision rate on the vertical scale is the average percentage of the automatically retrieved documents which have been manually identified to be relevant to the desired subject. Apparently all precision rates decrease with increasing recall rates. This is a very natural trend. By including more documents with higher similarity measures in equation (2), the recall rate will be naturally increased since more correct documents will be included, but the precision rate will also be naturally decreased since more irrelevant documents will also be includes. It was found that the quasi-natural-language queries produced roughly 0.09-0.15 degradation in precision rates as compared to those using simple queries. If the non-interpolated average precision rate which is a single number reflecting the retrieving performance often used in information retrieval [8] is used here, it can be found that the average precision rate actually degraded from 0.5580 to 0.4486. Apparently, the noisy syllable pairs constructed from the irrelevant words in quasi-natural-language queries have increased the degree of ambiguity. However, in any case retrieving the spoken documents using quasi-natural-language speech queries is the most natural and attractive and thus highly desired. Therefore better approaches to improve the performance of such cases is very important.

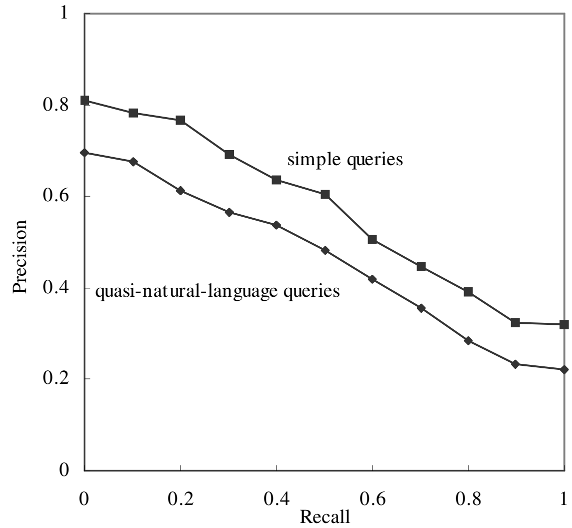

Figure 2: The performance comparison between using simple queries and quasi-natural-language queries.

5. CONCLUDING REMARKS

This paper presents the new framework of Mandarin voice memo retrieval developed in our group. A Mandarin voice memo system is one of the possible initial applications of our long term research projects, including the research project towards voice retrieval of Mandarin speech information and the research project towards natural language understanding. Though both two tasks are very difficult and still remain a long way away with many problems yet unsolved. With some partial solutions and advances in the speech recognition technology, a few initial applications such as a voice memo system are believed to be possible. Though the experimental results reported in this paper are actually not good enough, these preliminary results at least show the potential in this direction.

ACKNOWLEDGMENTS

This work was partially supported by the Republic of China National Science Council under Contract No. NSC 87-2213-E-001-026.

REFERENCES

1. Lin-shan Lee, Bo-ren Bai, and Lee-feng Chien, “Syllable-based Relevance Feedback Techniques for Mandarin Voice Record Retrieval Using Speech Queries”, *ICASSP97*, pp. 1459-1462.
2. K. Sparck Jones, G. J. F. Jones, J. T. Foote, and S. J. Young, “Experiments in Spoken Document Retrieval”, *Information Processing & Management*, Vol. 32, No. 4, pp. 399-417, 1996.
3. T. Kawahara, C. H. Lee, and B. H. Juang, “Key-phrase Detection and Verification for Flexible Speech Understanding”, *ICSLP96*, pp. 681-684.
4. Bor-shen Lin, Hsin-min Wang, and Lin-shan Lee, “A Key-phrase Understanding Framework Integrating Real World Knowledge with Speech Recognition with Initial Application in Voice Memo Systems for Chinese Language”, *TENCON97*, pp. 157-160.
5. Berlin Chen, Hsin-min Wang, Lee-feng Chien, and Lin-shan Lee, “A*-Admissible Key-Phrase Spotting with Sub-Syllable Level Utterance Verification”, *ICSLP98*.
6. Bor-shen Lin, Berlin Chen, Hsin-min Wang, and Lin-shan Lee, “Hierarchical Tag-Graph Search for Spontaneous Speech Understanding in Spoken Dialog Systems”, *ICSLP98*.
7. Salton, G. (1983). Introduction to Modern Information Retrieval, McGraw-Hill, NY.
8. Harman, D. (1995). “Overview of the Fourth Text Retrieval Conference (TREC-4)”, available at “http://trec.nist.gov/pubs/trec4/overview.ps.