Language, Information and Computation(PACLIC 11), 1996, 289-296

A Mandarin Voice Organizer Based On a Template-Matching Speech Recognizer

Jhing-Fa Wang, Jyh-Shing Shyuu, and Chung-Hsien Wu
Institute of Information Engineering
National Cheng Kung University
wangjf@server2.iie.ncku.edu.tw

Abstract

On the observation of current available voice organizers, all of them accept only voice commands or word-based commands. Using natural spoken language to operate organizer is still a difficult problem. In this paper, a template-based speech recognizer which accepts near(constrained) spoken language is proposed. Since the template-based recognizer is a domain-dependent speech recognition system, representing and matching of sentence templates become the main tasks of the recognizer. We use finate state networks(FSNs) to represent the sentence templates and propose a vowel-based, syllable-scoring method to match a correct template. By replacing the template sets, this method can be easily applied to other domains. Besides, two main functions, voice recording and voice message query, are implemented on our organizer using a fast CELP encoder/decoder to compress/decompress the voice data in realtime. Experimental results shows that the collected 31 sentence templates can greatly improve the voice interface between the user and the voice organizer.

1. Introduction

A voice organizer uses the voice input and output to process personal data. For example, for recording voice data or querying a voice message from the voice database, a voice organizer must first recognize the input voice and then respond the user with voice output. Hence, a voice organizer includes at least the following voice processing modules. First, a voice recognizer is used to decode the input speech into operating commands. Secondly, the voice output module responds the user with the information that he wants. Thirdly, a voice storage module is used to save the large amount of voice data.

On the observation of current available voice organizers, all of them accept only a few voice commands or word-based commands(Huang 94). Using natural spoken language to operate voice organizer is still a difficult problem. In this paper, we proposed a template-matching speech recognizer that allows the user to operate the voice organizer using near natural spoken language. The templates that we define here are those sentence patterns which are frequently used to operate the voice organizer. Among so many sentence patterns, they are classified into several so called "templates". The voice recognizer then match the input voice based on these templates. Hence, the input styles are more flexiable than those of voice-command based voice organizer.

In the following Sections, Section 2 describes the system block diagram of our proposed voice organizer. Section 3 go through details of the templated-matching speech recognizer. Section 4 describes the voice response module and voice compression module. The experimental results are given in Section 5. Finally, a conclusion remark is given in Section 6.

2. The System Block Diagram

The overall system block diagram is shown in Fig. 1. First, the input voice is recognized into syllable-lattice by the acoustic signal recognizer. The syllable-lattice is then decoded into operating commands by the template-matching speech recognizer based on pre-created sentence templates. The operating command is dispatched to either query voice message or to record voice data. For recording voice data, a CELP encoder is used to compress the input voice to reduce the storage size of the orginal voice signal and stores the compressed voice to voice database. For querying voice message from the voice database, a CELP decoder is applied to decode the compressed voice and then respond to the user.

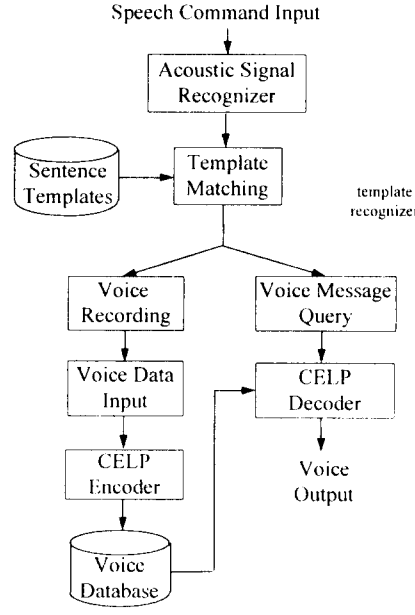


Fig. 1 System Block Diagram of the Voice Organizer

3. The Template-Matching Speech Recognizer

On the observation of the input sentence patterns of a voice organizer, several sentence templates were collected. Each of these templates is represented by a branchable finate state network(FSN) as shown in Fig.2.

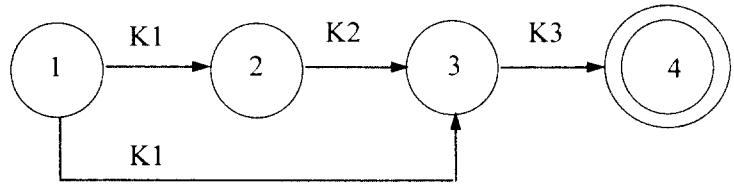


Fig.2 A FSN of a Keyword Sequence$K_1K_2K_3$

A state transition in the network accepts a keyword set and then go to the next state. When the final state is reached, the input keyword sequence is accepted by the FSN. For example, the following two sentence patterns are both accepted by the FSN in Fig.2.

A(K1) GOOD(K2) STUDENT (K3)
A(K1) STUDENT(K3)

It is noted that Ki in Fig.2 represents a keyword set rather than a single keyword. The keyword sets we used are shown below.

K1 = {APPEND，QUERY，DELETE，SETUP，END}
K2 = {ONE,TWO,...，NINETY-NINE}
K3 = {YEAR}
K4 = {MONTH}
K5 = {DAY}
K6 = {MORNING，NOON，AFTERNOON}
K7 = {TO}

The general form of a template can be summarized as follows,

(FUNCTIONS) + (DATE) + (TO) + (DATE)

The 'FUNCTIONS' keyword set represents K1. The 'DATE' keyword set consists of K2,K3,K4,K5, and K6. The 'To' keyword set is K7. For example, the following two sentence patterns are accepted by our voice organizer.

查詢(QUERY) 六(SIX) 月(MONTH) 一(ONE) 日(DAY) 到(TO) 六(SIX) 月(MONTH) 六(SIX) 日(DAY)
新增(APPEND) 七(SEVEN) 月(MONTH) 五(FIVE) 日(DAY) 下午(AFTERNOON)

The actual template format we used is formulated as follows,

K1 A1B1 K2 A2B2 · · · Kn AnBn

where

K1，K2,...,Kn are keyword sets
Ai=1: keyword Ki must exist

Ai=0: keyword Ki can be skipped
Bi=1 : Ki is a final state
Bi=0 : Ki is not a final state.

For example, Fig. 1. is formulated as follows,

K1 10 K2 00 K3 11

The template-matching recognizer starts with matching the keyword sets from the syllable-lattice. If there are more than one accepted templates, the acoustic score of the keyword sequence is used to determine the most likely one. To improve the keyword recognition rates, a vowel-based Mandarin word recognizer is proposed(J.F. WANG 94). The combination of the vowel candidate lists is first used to preselect possible keywords, and a wighting function which is based on the syllable score is applied to each preselected keyword. The acoustic score for the keyword sequence is then obtained by summing the syllable score of each keyword. Fig. 3 is an example of the vowel-based keyword recognizer.

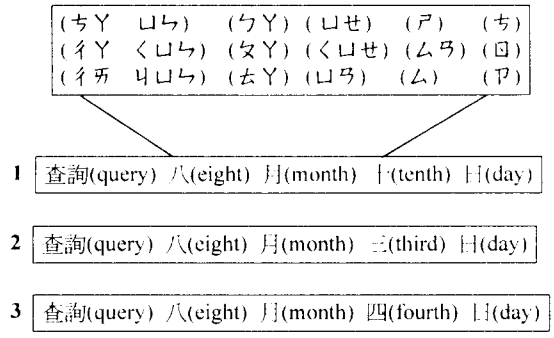
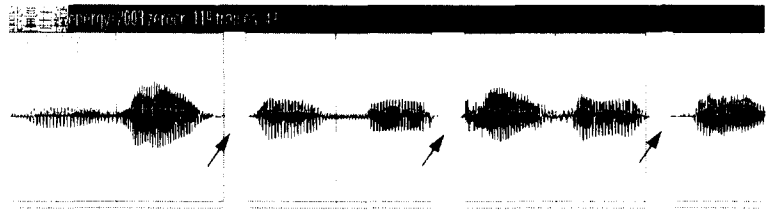


Fig. 3 an example of the vowel-based keyword recognizer

4. The Voice Compression Module and Voice Response Module

After the input voice command has been recognized, user begins to record his voice data or query specific voice message. In order to reduce the storage size of voice data, the CELP coder is adopted. The CELP coder is a 4.8KBPS speech coder. This means that the compression ratio is about 13:1 for 8KHz sampling of speech signals(Federal Standard 1016, 1991). Traditionally, complex analysis-by-synthesis architecture makes the CELP encoder require additional hardware to compress the speech signal in realtime(Kleijn, 1990). To reduce the computational cost, three arithmetic reduction methods are proposed so that the CELP coder can be run in realtime using only software. They are:(1) two look-up tables for replacing cosine operations when transferring line

spectrum pair(LSP) to linear predictive coefficients(LPC), (2) taking advantage of the relationship between two contiguous filtered overlapped codewords to comply with circular addressing technique for codebook searching, (3) predicting the dot production of the filtered stochastic codeword from the previous one and the first two dimensions of the codeword vector. In addition, silences between two prosodic segments or paused portion between sentences as shown in Fig.4 are recorded only by their duration instead of processing by CELP encoder. The computational complexity can be further reduced by 10%. For querying a voice message from the voice database, the CELP decoder is used to decode the compressed voice. During the decoding process, the decoder identifies the output speech is silence or voice data based on tagged voice database.



waveform: "你爲甚麼不回家(Why don't you go home)"

Fig. 4 Silence compression between two prosodic segments

5. Experiments

For the experiment, we have applied the proposed system to a voice time-scheduler. The user can input time, date, and other keyword sequence using near natural spoken language to operate the time-scheduler. There are 5 main functions in the scheduler and they are summarized as follows.

1. APPEND: allow users to record their voice into the voice database
2. QUERY: query voice messages from the voice database
3. DELETE: delete one voice message from the voice database
4: SETUP: setup the system
5. END: close the application

All of these functions are controlled by voice.We have collected and constructed 31 sentence templates for the scheduler(See Appendix). The VenusDictate large vocabulary speech recognizer(J.F. Wang 94) is used to recognize keyword sets. However, the lexicon in VenusDictate is replaced by the keyword sets as mentioned in Section 3. A new user is asked to give 15-20 miniutes of speech data to train the acoustic model before he operates the scheduler.

In the on-line testing, we found that the accuracy of template-matching is highly depends on the recognition rates of keywords. In general cases, a template that contains no confusing keyword gives satisfactory recognition results. However, it is known that the Mandarin digits of '4' and '10' are very difficult distinguished in their pronunciations. In this case, the syllable score becomes the judgement to match the correct template.

For the voice compression, the computational cost is the key point for the voice organizer. The CELP coder can give a realtime response on a Pentium computer without DSP hardware. Hence, the compressing procedure is almost finished while the user stop recording his voice.

6. Conclusion

In this paper, a template-based voice organizer is proposed. The user can use near natural spoken language to operate the voice organizer rather than using specific voice commands. To achieve this requirement, we create 31 templates and use an FSN to express each template. A single template can accept more than one keyword sequence. Hence, the input pattern is more flexiable than voice-command based voice organizer. Besides, a fast CELP coder is designed to compress the voice data so that the storage size of voice data can be greatly reduced. In words, our proposed system provides user a convenient voice interface for voice recording and voice message querying.

References

Jhing-Fa Wang, Jyh-Shing Shyuu, and Chung-Hsien Wu. 94: A Robust and Huge Vocabulary Mandarin Speech Recognition System with a Friendly Human-Interface. Proceedings of 1994 Global Cooperative Software Development Conference: 45-55.

Lin-Shan Lee, et al. 1995: Golden Mandarin (III) - A User-Adaptive Prosodic Segment-Based Mandarin Dictation Machine for Chinese Language with Very Large Vocabulary. IEEE, Internation Conference on Acoustic, Speech and Signal Processing : 57-60.

Eng-Fong Huang, Chian-Hung Chen and Hsiao-Chuan Wang. 1994: New Search Algorithm for Fast Syllable Hypothesization and Lexical Access in a Large- Vocabulary Mandarin Polysyllable Word Recognizer. Computer Processing of Chinese and Oriental Languages, Vol.8, No. 2: 211-225.

Yuqing Gao, Hsiao-Wuen Hon, Zhiwei Lin, Gareth Loudon, S. Yogananthan and Baosheng Yuan. 1995: TANGERINE: A Large Vocabulary Mandarin Dictation SYSTEM. IEEE, Internation Conference on Acoustic, Speech and Signal Processing: 77-80.

R.Billi, G. Massia and F. Nesti. 1986: Word Preselection for Large Vocabulary Speech Recognition. IEEE Internation Conference on Acoustic, Speech and Signal Processing: 65-68.

Frank K. Soong, and Biing-Hwang Juang. 1984: Line Spectrum Pair (LSP) and Speech Data Compression. IEEE Internation Conference on Acoustic, Speech and Signal Processing: 1101-1104.

Federal Standard 1016, Telecommunications.1991: Analog to Digital Conversion of Radio Voice by 4,800 bit/second Code Excited Linear Prediction (CELP), National Communications System, Office of Technology and Standards, Washington, DC 20305-2010.

Kleijn, Krasinski, and Ketchum. 1990: Fast Method for the CELP Speech Coding Algorithm. IEEE Trans. ASSP: 1330-1342.

Appendix

R00 8 K1 00 NO 10 K2 10 NO 10 K3 10 NO 10 K4 10 K5 11
R01 6 K1 00 NO 10 K3 10 NO 10 K4 10 K5 11
R02 2 K1 00 K5 11
R03 6 K1 00 NO 10 K2 10 NO 10 K4 10 K5 11
R04 6 K1 00 NO 10 K2 10 NO 10 K3 10 K5 11
R05 7 K1 00 NO 10 K2 10 NO 10 K3 10 NO 10 K4 11
R06 5 K1 00 NO 10 K2 10 NO 10 K3 11
R07 5 K1 00 NO 10 K2 10 NO 10 K4 11
R08 4 K1 00 NO 10 K2 10 K5 11
R09 5 K1 00 NO 10 K3 10 NO 10 K4 11
R10 4 K1 00 NO 10 K3 10 K5 11
R11 4 K1 00 NO 10 K4 10 K5 11
R12 3 K1 00 NO 10 K2 11
R13 3 K1 00 NO 10 K3 11
R14 3 K1 00 NO 10 K4 11
R15 16 K1 00 NO 10 K2 10 NO 10 K3 10 NO 10 K4 10 K5 10 K6 10 NO 10 K2 10 NO 10 K3 10 NO 10 K4 10 K5 11
R16 14 K1 00 NO 10 K2 10 NO 10 K3 10 NO 10 K4 10 K5 10 K6 10 NO 10 K3 10 NO 10 K4 10 K5 11
R17 10 K1 00 NO 10 K3 10 NO 10 K4 10 K6 10 NO 10 K3 10 NO 10 K4 11
R18 14 K1 00 NO 10 K2 10 NO 10 K3 10 NO 10 K4 10 K5 10 K6 10 NO 10 K3 10 NO 10 K4 10 K5 11
R19 12 K1 00 NO 10 K2 10 NO 10 K3 10 NO 10 K4 10 K5 10 K6 10 NO 10 K4 10 K5 11
R20 10 K1 00 NO 10 K2 10 NO 10 K3 10 NO 10 K4 10 K5 10 K6 10 K5 11
R21 8 K1 00 NO 10 K3 10 NO 10 K4 10 K6 10 NO 10 K4 11
R22 11 K1 00 NO 10 K3 10 NO 10 K4 10 K5 10 K6 10 NO 10 K3 10 NO 10 K4 11

R23 9 K1 00 NO 10 K3 10 NO 10 K4 10 K5 10 K6 10 NO 10 K4 11
R24 10 K1 00 NO 10 K3 10 NO 10 K4 10 K5 10 K6 10 NO 10 K4 10 K5 11
R25 8 K1 00 NO 10 K3 10 NO 10 K4 10 K6 10 NO 10 K4 11
R26 9 K1 00 NO 10 K3 10 NO 10 K4 10 K6 10 NO 10 K4 10 K5 11
R27 7 K1 00 NO 10 K4 10 K6 10 NO 10 K4 10 K5 11
R28 8 K1 00 NO 10 K4 10 K5 10 K6 10 NO 10 K4 10 K5 11
R29 6 K1 00 NO 10 K4 10 K6 10 NO 10 K4 11
R30 7 K1 00 NO 10 K4 10 K5 10 K6 10 NO 10 K4 11