The Research Project of Man-Computer Dialogue System in Chinese

Guan Dinghua Chu Min Zhang Quan Liu Jian and Zhang Xiangdong*

Institute of Acoustics, Chinese Academy of Sciences, Beijing, 100080, China

E-mail: mchu@plum.ioa.ac.cn

ABSTRACT

This paper gives a brief introduction about the five-year research project of "Man-Computer Dialogue System in Chinese", which was supported by the Chinese Academy of Sciences. The project is carried out in two steps. In the first step, research works undertook by several research groups separately on the core area such as speech recognition, speech synthesis, language understanding and dialogue organizing module. And in the second step, all techniques are assembled together to form a demo dialogue system of traveling information inquiry system. The current state of all above core areas and some evaluation results are discussed in the first part of this paper and the framework of the traveling information inquiry system is presented in the second part.

1. INTRODUCTION

With the development of Computer Science, the need for dialogue systems between human beings and computers is growing rapidly. During the past five years, Institute of Acoustics, Institute of Automation and University of Science and Technology have conducted cooperatively on a research project of "Man-Computer Dialogue System in Chinese", which was supported by the Chinese Academy of Sciences.

Comparing with European language, Chinese (always refers to Mandarin in this paper) has several special characteristics that make it to be processed differently. a) Chinese is a syllable- based language (one Chinese character corresponds to a monosyllable) and it has a small syllable set (there are only 412 syllables if tone does not be taken into account). b) There are at least 60,000 commonly used words in Chinese, and a new word emerges when several characters are often used together. c) There has no obvious boundary for words in Chinese text. These characters cause some facilities in processing and also bring special problems.

The five-year project is carried out in two stages. In the first stage, research works undertook by several research groups separately on the core area such as speech recognition, language understanding and dialogue model, speech synthesis, speech database construction and system performance evaluation. And in the second stage, all techniques are assembled together to form a demo dialogue system of traveling information inquiry system.

The current state of core techniques and some evaluation results are given in section 2. And the framework of the traveling inquiry system is given in the section 3. Section 4 gives a brief summary.

2. THE CORE TECHNIQUES

2.1 Chinese Continuous Speech Recognition

Continuous speech recognition is an important component of a dialogue system. Though the traditional LPC cepstrum parameters are by far the best recognition parameter, it is an all-pole model, which is not suitable for modeling most of Chinese Initial Part(unvoiced). An Initial-Final-Transition Model of Chinese syllables is constructed for voiced/unvoiced segmenting. And a hybrid procedure of the Initial Part (unvioced segment) Classification and the continuous voiced segments recognition is developed in our recognizer [1]. Figure 1 is the framework of the recognizer.

During recognition procedure, continuous speech stream is first segmented into unvoiced segment, continuous voiced segment and silence segment. Then an Initial Part (unvoiced segment) Classifier and a continuous voiced segments recognizer are used to recognize unvoiced and voiced segment separately. At last, Diagonal Sort Algorithm is used to combine the recognition results of these two parts.

Sub Feature Space Classification (SFSC) algorithm[2] is developed for unvoiced segment. The basic idea of SFSC algorithm is to divide the duration, zero cross and energy 3-D space into 64 sub-spaces. Initials are first classified into these sub-spaces in time domain. Then, the distances between these Initial Parts and the centers of those sub-spaces are calculated, from which the Initial Parts classified sequence is produced. This method could restrict the confusion of the acoustic parameter feature distribution. Totally 81 Initial

models are used. The recognition rate is 73.30% for top one candidate and 98.96% for top five candidates.

As to the continuous voiced segment, a new segmentation algorithm [3] for cutting continuous voiced speech into syllables is proposed. The algorithm is based on the VQ-distance between two consecutive speech frames represented by frequency domain coefficients, LPC cepstrum coefficients and its regression coefficients. According to the distribution of these coefficients, the different weights are chosen for different order coefficients in reverse proportion with their covariance. And two Neighbourhood-Codebooks of each order co-efficient used for segmentation are trained by using the unified threshold. In recognition, Multi-Codebook DHMM algorithm is adopted, coupling with this new segmentation algorithm. After vector quantization, a sequence of VQ codes which represents the input speech is obtained, and impossible separate-points are kicked out. The recognition rate is 82.3% for top one candidate and 98.6% for top five candidates.

After the Initial sequences and final sequences are obtained, Diagonal Sort Algorithm is used to combine the two kinds of sequence to form syllable sequences. In a formal test, the accuracy rate of Chinese syllables is 72.27% for the top one candidates and 96.3% for the top ten candidates. (The test set contained more than 1000 sentences including 5789 Chinese words.)

2.2 Chinese Speech Synthesis

Research on the naturalness of synthetic Chinese has shown that fundamental frequency, duration, co-articulation and intensity are important factors. Among them the fundamental frequency and duration affect the naturalness most. Since the TD-PSOLA algorithm can modify the pitch, duration and intensity of speech segment in time domain with little distortion, it is adopted in our Chinese TTS system[4]. The system can produce rather natural speech, which can be accepted by most listeners. The outline of the TTS system is given in figure 2. The main function of the text analyzer is to segment text into words and phrases. It helps a lot on naturalness when the analyzer has the ability to conjoin monosyllabic words with neighboring word and to cut long sentences into breathing groups[5]. The text analyzer also provides structure information for the prosody model. Constructing a proper prosody model is essential for improving the naturalness of synthetic speech. Chinese phoneticians and phonologists have achieved many valuable results on the variation of prosodic features of Chinese. Though these results can not be used directly to form the prosody rules for synthetic systems, they give light to our prosody analysis. On the basis of the existent results and acoustic analysis of CCTV news broadcast speech that is carried out in our laboratory, a prosody model for broadcast style speech is constructed. The prosody model contains a pitch model and a duration model, which work on syllable level, word level and sentence level separately. Though, in human speech, pitch and duration act together to form the overall prosody, and effect each other, they are process in two independent sub-model in order to simplify the model. More details about the prosody model can be found in [6]. The output of the prosody model is the duration and pitch contour of all syllables in the synthetic speech. Waveforms getting from the syllable dictionary are modified by TD-PSOLA algorithm to make them carrying the target duration and pitch contour.

In a formal evaluation that is organized by the government, the intelligibility of the TTS system is 94.1% and the naturalness reach 3.4 in MOS scale.

2.3 Language understanding

Though a statistical language model can performs well in speech recognition, it cannot understand the meaning of the sentence. So it can not be used in a dialogue system. A natural language understanding system has been constructed under the conception of Hierarchical Network of Concepts (HNC) [7], which defines relationships between concepts by hierarchical networks. The understanding system undertakes two important functions: one is to generate correct input question sentences from the results of speech recognizer, even though these results contain lots of fuzzy syllable; the other is to generate answers for input questions.

The HNC theory treats human language as sets of concepts and connections between concepts that formed a hierarchical network. The initial motive of the theory is to simulate the language perceptive patterns of the brain of human being. A set of symbols is designed to represent the categories of concepts and their internal relationships. These symbols enable computers to process human languages. In HNC theory, the cognitive structure is composed of local and global association venation and the expression of association venation is a basic problem in linguistic deep structure (i.e. in semantic layer). Generally, there are two ways for human to understand natural language. One starts from words that compose a sentence, and the other starts from the holistic structure of the sentence and the context. The first one is the local association venation, and the second one is the global association venation. Of course, when human understand the language, the two associations is not completely separate, they are interrelated, intercurrent and interactional. When a

computer understands human languages, it also needs to use the two associations. The prime idea of HNC is to help computer understanding language by founding two associations venation. In fact, HNC tries to use semantic analysis to substitute conventional syntax analysis and statistic processing.

Various knowledge, which take a very important role in human understanding of languages, is organized into three layers in HNC structure. They are the conceptual knowledge layer, the linguistic knowledge layer, and the common sense and professional knowledge layer, each of which should have a supporting knowledge database. The knowledge databases for Mandarin are in constructing and small databases for travelling information have been finished. These databases are used in our travelling information inquiry system.

After the understanding system is utilized for the results of continuous speech recognizer, correct recognition rate improves much. The first candidate correct rate for the original recognizer is 72.2%. After understanding, Chinese characters correct rate reaches 90.6 % . In most situations, the understanding system generates an answer for the inquiry, even if it obtained an in complete sentence.

3. THE FRAMEWORK THE TRAVELING INFORMATION INQUIRY SYSTEM

Continuous speech recognition, speech synthesis, language understanding are assembled to form a dialogue system on traveling information inquiry which has a vocabulary of 1000 words and can handle 89 kinds of sentence patterns. The framework of the system is in Figure 3. No formal testing results available for the dialogue system now.

4. SUMMARY

This paper gives a brief introduction of research project "", which is of "Man-Computer Dialogue System in Chinese", which was supported by the Chinese Academy of Sciences. The current state of the core techniques, such as speech recognition, speech synthesis and language understanding, and some testing results are discussed. And the framework of a demo dialogue system on travelling information inquiry, which is a tight assembling of the core techniques, is also presented. By now, no formal testing results available for the dialogue system. The present speech recognizer is speaker dependent. So it can not be utilized in a real dialogue system. Great effort has been taken to make it speaker independent. The current TTS system can produce news broadcast style speech, while a dialogue system prefers spoken language that contains more variations in speaking styles, intonations and emotions. These are the key points of our research in the future. Robust of the dialogue manage module essential for a real dialogue system since it often encounters informal and incomplete questions.

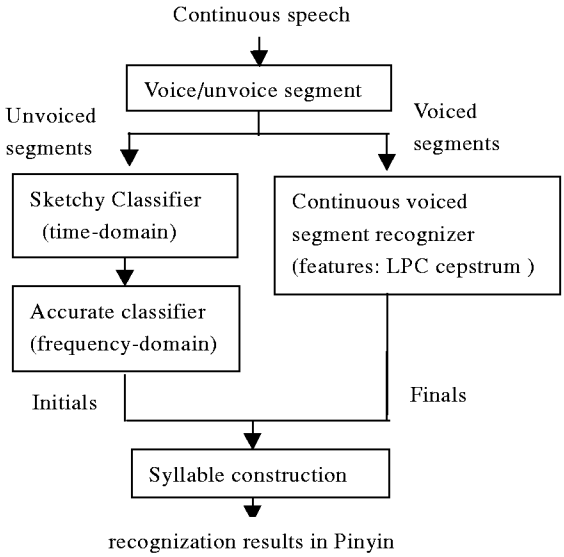

Figure 1: The framework of the Chinese continuous speech recognition.

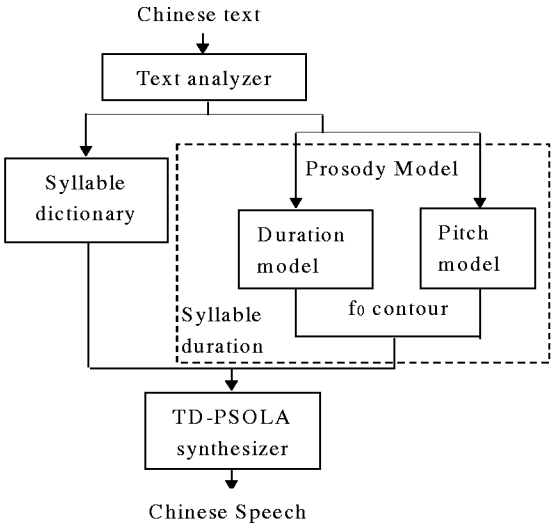

Figure 2 : The outline of the Chinese TTS system

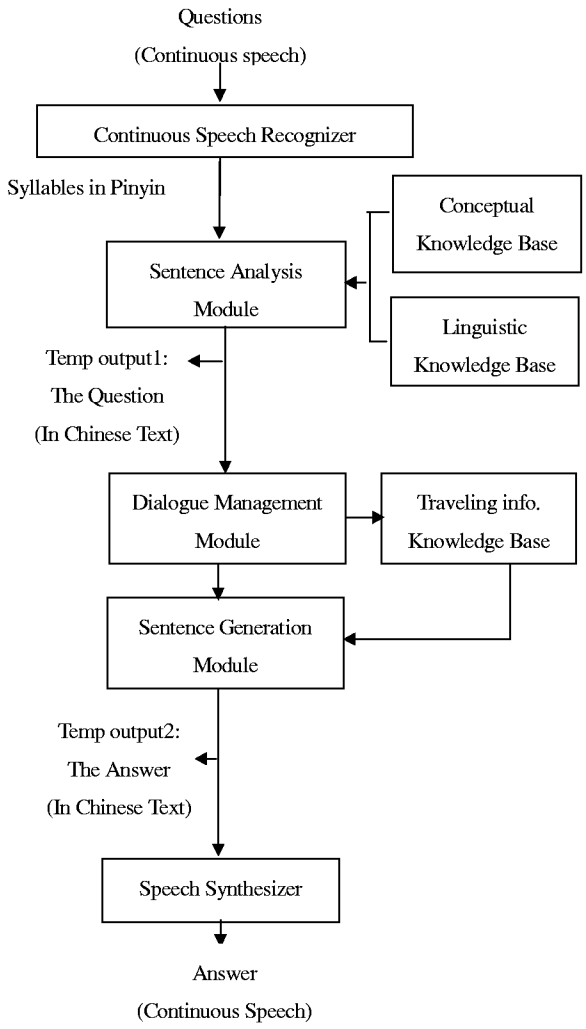

Figure3: The framework of travelling information inquiry system

5. REFERENCES

1. Liu Jian, Zhang Xiangdong, Yu Tiecheng, Dong meixiang, "Speaker-dependent Continuous Chinese Speech Recognition System", *Proceedings of 6th. National Youth Computer Symposium*, 1996, 654-659
2. Zhang Xiangdong, Liu Jian and Yu Tiecheng, "Research on Chinese Initial Part Recognition Algorithms", *Proceedings of 1997 China-Japan Symposium on Advanced Information Technology*, 1997, 221-226
3. Liu Jian, Zhang Yuan and Yu Tiecheng, "A New Segmentation Algorithm for Voiced Chinese Continuous Speech & Its Applications to Recognition", *Proceedings of 1997 China-Japan Symposium on Advanced Information Technology*, 1997, 215-220
4. GUAN Dinghua, CHU Min and Lu Shinan, "A Chinese Text-to-speech System with High Intelligibility and Natralnesss", *Proc. of ICA'95* Trodheim Norway, 1995, Vol.3, 31-34.
5. Tang Difei, Chu Min, Lu Shinan and He lin, "Word segmentation for Chinese TTS system—Lengend Voice", *Proceedings of the first China-Japan workshop on spoken language processing*, 1997, 153-156
6. Chu Min, Tang Difei, Lu Shinan and Guan dinghua, "The prosody model for the Chinese TTS system: Lengend Voice", *Proceedings of the first China-Japan workshop on spoken language processing*, 1997, 111-116
7. Huang Zengyang, "The outline of HNC theory", *Journal of Chinese information processing*, Vol.11, No.4, 1997, 11-20