

# THE DESIGN AND IMPLEMENTATION OF A CONCEPT PROTOTYPE FOR BEIJING 2008 OLYMPIC GAMES

J. Zhang<sup>1</sup>, J. Liu<sup>1</sup>, M. Li<sup>1</sup>, J. Pan<sup>1</sup>, J. Han<sup>1</sup>, L. Tuo<sup>1</sup>, B. Sun<sup>1</sup>, J. Wang<sup>1</sup>, Y. Yan<sup>1,2</sup>

<sup>1</sup>Institute of Acoustics, Chinese Academy of Sciences, Beijing 100080, P. R. China

<sup>2</sup>Oregon Health & Science University, Oregon, USA

[jzhang, jliu, mli, jpan, jhan, ltuo, bsun, jwang, yyan}@hccl.ioa.ac.cn](mailto:{jzhang, jliu, mli, jpan, jhan, ltuo, bsun, jwang, yyan}@hccl.ioa.ac.cn)

## ABSTRACT

*This paper presents current research activities for developing a Concept Prototype for the Beijing 2008 Olympic Games (CPBOG) at the Institute of Acoustics, Chinese Academy of Sciences. The System is to provide live sports, touring, traveling information for visitors to Beijing, China during 2008 Olympic periods. Our in house up-to-date speech recognition, speech understanding, multimedia processing, text to speech and dialog manager techniques are used in the attempt to develop a user-friendly multi-modal multi-lingual spoken language system. Certain key technologies are presented and some realization problems are discussed.*

## 1. INTRODUCTION

The arrival of computer half a century ago opened a wide range of possible applications, and the research on speech recognition began almost simultaneously with the birth of computer. Since then, making computer understand human voice has been the dream of speech researchers around the world. Recent advances in hardware/software have made automatic speech recognition one of the *potentially* most pervasive user interface components. As mobile devices linked by wired and wireless networks become an increasingly important and integral part of people's daily life, how to make the access and interaction easier is becoming more and more important. One of the biggest challenges for an ideal device is how to make data entry, device manipulation and information retrieval easier since keyboard may not be readily available anytime anywhere. Nowadays, voice-enabled electronic products such as cellular phone, dictation machine, voice portal system, speech communicator [1-2], have already been put into practice. There are still some critical issues that need to be dealt with before these systems can be widely accepted and used in real world environments.

The aim of our research is to develop an embedded spoken language system that can run on a handheld device and provide live sports, touring, traveling information for guests visiting Beijing, China during 2008 Olympic periods. The commonly used spoken language system architecture is being extended to serve the design need. As shown in Fig. 1, a multi-lingual support component is attached to the language generation module. So the system can return to

the user in different languages. When activated, this function can be used as an automatic interpreter for bilingual human-human conversation. For example, when turned on, once the system understands the user question, instead of trying to give answers as normal spoken language system does, it plays back the question that can be in any pre-set languages. This way, a user can use it to communicate with others for simple query in addition to those services that will be provided by 2008 digital Olympic servers. In order to achieve this long-term goal, first attempt was made in our lab to develop a concept prototype (some on-line services are not yet available and the up-to-date technologies are still not robust enough). The purpose for this effort is to find critical challenges and shortcomings of today's technology and shed light on future directions.

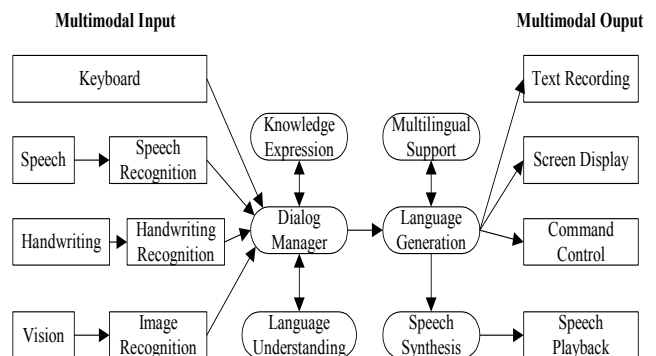


Fig. 1. Architecture of Concept Prototype System

Speech recognition and understanding is an important research field especially in Chinese language world because of the difficulties of inputting Chinese with keyboard and the varieties of accents and dialects. Founded in 2002 at Institute of Acoustics, Chinese Academy of Sciences, our lab is conducting applied research for next generation human computer interface technologies. The current research and development efforts are centered at multi-modal, multi-lingual spoken language systems. It is our hope that by integrating capabilities such as speech recognition, natural language understanding, speech synthesis, facial animation, expression/lip recognition, dialogue management, language translation, information retrieval, and summarization, human machine interaction

can be a graceful experience for ordinary users. The existing infrastructure [3] in speech processing and understanding area makes it possible for us to implement a multi-modal multi-lingual system for people to retrieve live information via telephone line, internet, cellular phone, etc. Like modern speech communicator systems [1,4], the implementation of our Concept Prototype for Beijing 2008 Olympic Games (CPBOG) involves techniques in speech recognition, speech understanding, text-to-speech (TTS) and dialog manager (DM), etc. The CPBOG system is to provide traffic, sports, and touring information for people whose native language is English or Chinese, and provide certain real-time language translation functions for simple questions. So in addition to general servers like speech recognizer, TTS and DM, there are also servers taking care of machine translation and GPS location based services in our system.

This paper will present our system and discuss some implementation issues. The system architecture and overall functions are introduced in the second section. Then some key components: speech recognition server, dialog manager and speech synthesis server are discussed in more detail in following sections. The conclusion is given in the end.

## 2. SYSTEM ARCHITECTURE

The prototype system was programmed in C++ to run in Windows environments. However, with minor modifications, it is portable to UNIX like environments. The system main menu is shown in Fig. 2. The functionalities of each server are described as follows:



Fig. 2. Main Menu of CPBOG System

### 1. Multimodal Input

These input servers provide information via keyboard, speech and handwriting recognition, or even image recognition. The multimodal input is to facilitate

different persons to communicate with the prototype conveniently according to their own personal condition at the given time. Also it is to potentially improve the efficiency and accuracy of the input. For example, speech and image (e.g., lip movements and face expression) may be integrated later as different information sources that may be useful for machine understanding. Also handwriting recognition is widely accepted in China for its convenience in speeding up Chinese input before speech recognition techniques become mature and show decent performance under low SNR or cross-talking conditions. In our implementation, a continuous-density HMM based speech recognizer is used. The output of the recognizer is sent to the dialog manager server. The prototype can accept speech in English or Chinese. The language type can be set in advance by users as it is now and will be automatically identified in the future.

### 2. Dialog Manager

Dialog manager (DM) is an indispensable server to take charge of task flow control and schedule. It acts as an agent between an end user and application servers [4]. Dialog Manager has several functions. It controls what to do in the next step according to previous events. It needs to resolve ambiguities in the interpretation of recognizer outputs. According to the final meaning understanding, the input content is decided and next step is taken. If the input content is meaningless or undetermined, then DM will ask the language generator server to generate utterances such as "I don't understand" or "Would you please say it again". If the input content is to retrieve some information from website or online multimedia database, it will search and try to find the corresponding resources and the language generation server will prompt this information to the end user in an appropriate way. In order to determine if user's keyword is correctly recognized, confidence measure (CM) techniques are exploited and if the CM score is poor for a given input, then DM may direct other servers to ask the user to input again.

### 3. Language Generation

According to the user's request, the system may generate English prompts or Chinese prompts. If the end user uses English, then the system will prompt in English. To feedback text information to the user, the language generation server will send its text content to the speech synthesizer server and the waveform from speech synthesizer is played back to the user via multi-modal output server. As described earlier, this module was extended with a multilingual component, so in addition to the above said functions, it can also be used as an interpreter when multi-lingual TTS is attached.

#### 4. Multimodal Output

Multimodal output is to provide all kinds of information to the user in the best presentation form based on user preference. The user may ask for information which is in the file formats of *mpeg*, *mov*, *wav*, *text*, *jpg* and so on. So a particular server named “effect server” is specially designed for this purpose. The server can play and display file content in all these file formats, either in sound or visual forms.

### 3. DATA COLLECTION AND SCENARIO DESIGN

As the very first step to develop the system, we first collected movies, articles, maps, etc. concerning touring and sports information. After analyzing and summarizing these collected materials, we designed about 10 scenarios in this presented prototype. All possible questioning and answering sentence patterns are proposed from friends and students. The most typical sentence patterns were recorded both in English and Chinese. 60 speakers uttered 100 sentences each. The recorded speech is in 8 KHz sample rate and under ordinary classroom environment. The speech database is used to adapt our general purpose acoustic model.

We selectively design about 10 different scenarios that include ‘ask\_current\_position’, ‘ask\_nearby\_stadium’, ‘ask\_athlete\_info’, ‘ask\_ticket\_info’, etc. The system will simulate a complete process that a guest may go through when he or she visits a new city. A typical scenario is given in Fig. 3 in which the system tells the user that according to real-time traffic information, due to the heavy traffic along the original suggested routine, new routine may be taken instead.

Since the final goal of designing this prototype is to provide an affordable, usable, mobile device for people to use in their journey. The software part of the system should be small enough to install in some hand-held devices. To meet the resource and hardware limitations of such kind of devices, we paid special attention to recognition engine and TTS speech database compression.

### 4. SPEECH RECOGNITION

We have developed an ASR engine based on our own Large Vocabulary Continuous Speech Recognition (LVCSR) technologies. It provides a set of functions that encapsulate the recognition technology. To efficiently reduce acoustic model size and computational complexity for CPBOG system, we use Bhattacharyya distance measure [5] as a criterion to quantize the mean and variance vectors of Gaussian mixtures. To minimize the quantization error, the front-end feature vector was separated into multiple streams (such as MFCCs, delta-MFCCs and delta-delta MFCCs)

and then the modified K-means clustering algorithm was applied to each stream [6].

Our speech recognition engine is implemented to support a finite state grammar. The state transitions in this diagram have corresponding ASR-API functions or internal events supported by the recognition engine. Some API functions can be called only in some specific states of the finite state machine. All state transitions result in notification to the application of this state machine. Key features of our recognition engine include:

- Real-time speaker-independent continuous speech recognition with accuracy up to 96%.
- Task-independent acoustic model.
- Flexible user-defined vocabulary.
- Keyword spotting technology.
- Multi-threaded and asynchronous/synchronous operation modes supported.

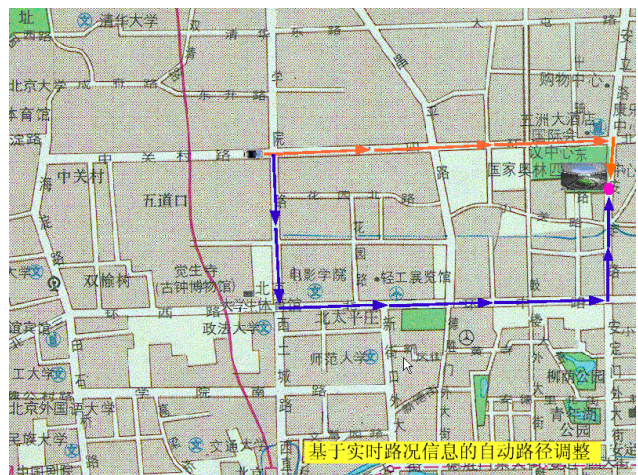


Fig. 3. Dynamic Routine Adjustment According to Real-Time Traffic Information

### 5. DIALOG MANAGER

Successful completion of a task requires that, at the end of a session, the two agents (user and system) involved agree on a particular result. It also requires that, between them, they have some understanding of how to go about completing the task. This in turn requires that two types of knowledge be captured in systems that successfully support this task: a representation for the task domain specific information that needs to be created and a representation that captures the structure of the activity needed to complete the task. It has been suggested that properly devised data structures that reflect the constraints imposed by the task domain can be used to interpret user inputs and guiding dialogue [7-9]. It is also desirable to capture the structure of the activity that takes place and thereby capture whatever expertise may be involved.

Our target is to provide a spoken dialog manager to be used in the CPBOG spoken dialog system. Through our empirical studies of dialog process in purposeful task domain such as traveling plan, meeting scheduler and so on, we believe it can be modeled as an approximating procedure with contributions from user and computer as input and converging on task completion. Our approach is based on a generalization of the form-filling paradigm. We use a hierarchical task structure model to describe task domain specific information. It is a general-purpose task structure model that can be referred as domain independent at least for a class of tasks. We can see the model is a hierarchical structure in contrast to the conventional flat slot structure. It contains information slots at various levels abstraction. The main merit of the task structure model is that we can use some mathematical objects such as order structure and set structure to describe the operational semantics of the lowest level object. This may benefit the domain-independent design of the dialogue management architecture.

A simplified sentence level grammar for DM is shown in Fig. 4. In this case, the scenario presented is that a user may get lost and ask the system “where am I”. This kind of question can be expressed in several ways in English or Chinese. In this example, some logical operators are used. The sign ‘|’ means ‘or’ computation, and ‘[’, ‘]’ means ‘optional’. So people may ask in sentences like “tell me where am I” or “would you please tell me where am I”.

```
[*1] would you please tell me where am i 请告诉我现在的位置
<ask_current_pos_cn> ::= [请][问|告诉]我现在的 <pos_cn>
<ask_current_pos_en> ::= [would you please] [tell me] where am i
<pos_cn> = 地方 | 位置 | 方位
<*ask_current_pos:void:1001> ::= <ask_current_pos_cn> | <ask_current_pos_en>
```

Fig. 4. Typical Sentence Grammar for DM

## 6. SPEECH SYNTHESIS AND MACHINE TRANSLATION

Our Embedded Speech Synthesis Engine is designed specially for mobile devices such as cellular speech applications. It is modeled on Chinese full syllable and certain particular units. Speech compression and encoding algorithm has also been integrated. The engine has been optimized in speech database size and naturalness suitable for mobile applications with limited resources. The speech database can also be accustomed according to varieties of resources availability and application requirements so that the overall performance will be optimal in certain applications. Our speech synthesis server takes only 100 K RAM and 320 K ROM and can run on machine with 25 MIPS CPU. It is in Chinese female voice and supports 8 KHz or 16 KHz sampling rates.

For English TTS, we pre-recorded all possible words. During the synthesis, we find corresponding waveform for

each word and concatenate them. Since current machine translation techniques are not mature enough to provide decent performance when it is required to translate domain independent sentences from one language (e.g., Chinese) to another language (e.g., English), we use a specific method to deal with our domain specific applications. All questioning and answering sentences are strictly structured in grammar, so that a user may only say certain sentences in certain ways. For this kind of bilingual translation, the accuracy is near 100%. For example, a caller may dial into the system in English via cellular phone and a user in another end will hear it in Chinese (speech from the caller has been translated). Similarly, the end user’s Chinese speech will be translated into English by the system and played back to the caller.

## 7. CONCLUSIONS AND FUTURE WORK

We have developed a concept prototype for Beijing Olympic 2008. Although some of the components (e.g., Vision and Image Recognition) are not yet integrated, it has functioned as we expected. We will use this task as a driven force for our embedded multi-modal multi-lingual spoken language system research.

## References

- [1] J. Zhang, W. Ward, B. Pellom, X. Yu and K. Hacioglu, Improvements in Audio Processing and Language Modeling in the CU Communicator, Eurospeech, 3:2209--12 Denmark, 2001.
- [2] Pellom, B., Ward, W., Hansen, J., Kacioglu, K., Zhang, J. P., etc., University of Colorado Dialog Systems for Travel and Navigation. Human Language Technology Conference (HLT-2001), San Diego, March 2001.
- [3] Yonghong Yan, Speech Technology: When will the dream turn into a reality? China-Japan Joint Conference on Acoustics, Nanjing China, Nov. 2002.
- [4] Victor W. Zue and James R. Glass, Conversational Interfaces: Advances and Challenges, Proceedings of the IEEE, Vol. 88, No. 8, pp. 1166-1180, Aug. 2000.
- [5] Brain Mak, Etienne Barnard, “Phone Clustering Using the Bhattacharyya Distance”, In Proc. of ICSLP, vol. 4, pp. 2005-2008, 1996.
- [6] Y.Linde, A.Buzo and R.M.Gray, “An Algorithm for Vector Quantizer design”. IEEE Trans. Comm., vol. COM-26, pp. 702-710.
- [7] Denecke, M. and Waibel, A. Dialogue strategies guiding users to their communicative goals. Proceedings of Eurospeech 97, Sept. 1997, Rhodes, Greece.
- [8] Wright, J., Gorin, A. and Abella, A. Spoken language understanding within dialogs using a graphical model of task structure. Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP), Dec. 1998, Sydney, Australia.
- [9] Veldhuijzen Van Zanten, G. Adaptive mixed-initiative dialogue management. Interactive Voice Technology for Telecommunications Applications, IVTTA-98. Proceedings. IEEE 4th Workshop, 1998.