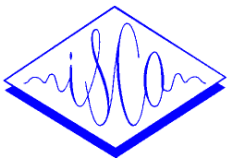

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

# Data Collection and Processing in a Chinese Spontaneous Speech Corpus IIS_CSS

*JunLan Feng XianFang Wang and LiMin Du*
Institute of Acoustics, Chinese Academy Of Sciences,
17 Zhongguancun Rd, Beijing 100080, China
fengjl@iis.ac.cn wangxf@iis.ac.cn dulm@iis.ac.cn

## Abstract

In this paper we report on the first phase of the speech corpus ISS_CSS collection for purposes of the CEST(Chinese-English speech translation) project. The corpus is intended to provide training material for speaker independent spontaneous Chinese speech recognition and automatic dialogue management over the telephone line. This paper describes the collection measures, processing methods, annotation and contents of this corpus. It consists of two parts: human-human dialogues and human-machine dialogues. Presently, the corpus has finished 10-hour speech and the associated annotation. Finally, we will present our collecting plan in the future.

## 1. Introduction

CEST project [1] is a collaborative research project between CAS(Chinese Academy of Sciences) and AT&T. This project focus on tour-related domains. In the first stage, we select two kinds of telephone-based information services as specific domains: travel information retrieval and hotel front-desk services. Making this selection is based on a balance between the following considerations: (i) They urgently needs automatic speech translation technique to relieve inconvenience for tourists. (ii) The dialogues in these two domains are wide-range and can provide large amounts of speech phenomenon for spontaneous speech recognition research. (iii) Collecting real data of the above two scenarios is relatively easier than other tour-related domains, such as taxi. In addition, collecting such a large scale of Chinese spontaneous speech corpus is valuable for large vocabulary Chinese spontaneous speech recognition research, because so far there is no a large scale corpus available.

The collection procedures, preprocessing methods, contents and transcription behind a corpus directly impact the performance of the resulting systems. This paper adopted a popular "dialogue oriented" collection paradigm. The detailed collection procedures will be described in section 2. Section 3 and Section 4 will respectively introduce the preprocessing methods and the transcription. The statistical results about the completed portion of this corpus will be given in section 5.

## 2. Collection Procedures

Due to various limits by laws and the cost to invest in negotiating with some public service organizations and companies, collecting real speech data sometimes becomes a tough task though the speech on telephone line every day amounts to a striking number. After many efforts, a feasible scheme has been selected; it is illustrated in figure 1.

In figure(1), by Computerfone(CF) or Dialogue Card (DC), speech on the telephone line is input into the computer and recorded as wave files (Microsoft RIFF WAV). The signal format is 16 bit, 8kHZ. CF/DC divides this framework into two parts, human and devices on the right side stay inside our center, the left side including involved people and organizations is the outside part. In the first stage, human-human dialogues between a caller outside and an inside agent or between an inside caller

and a real information center. Inside callers consist of employees in our laboratory and visitors who have real needs or would like to contribute to this corpus. Outside callers include all people who need helpful travel information or volunteer to make contributions to this research project. We encourage more calls by providing some small gifts to callers. By these two ways, in two months, we have collected about 300 human-human dialogues including 8-hour speech. During this process, we gradually build a test spoken dialogue system by integrating the general dialogue management framework proposed by us in [2] and "Wizard of Oz" (WoZ) [3] technique. Using this system, we can collect human-machine dialogues, which more directly contribute to the resulting spoken language system, and provide opportunities to examine and improve the system by experiences in a real life. Currently two-hour human-machine dialogues have been collected. It is expected that 40-hour speech corpus can be finished in about one year.

## 3. Data processing

In our strategy, all sounds in a dialogue have been recorded and saved as a file. A dialogue averages about 2.5 minutes, such a large file is not convenient for annotation and other further processing for speech recognition. Hence, we use off-line signal processing technique to determine the beginning point and the end point of speech by a self-adaptation silence detection algorithm. In this method, a dialogue will be divided into several or dozens of utterances which average time is 3 seconds; silence segments between two utterances will be deleted in this stage.

## 4. Annotations

Annotation of spontaneous speech is a tough work. The difficulties result from almost unlimited speech events in spontaneous speech. In order to classify all these events into finite special codes and language symbols, drawing a uniform and detailed criterion for various spontaneous items is first necessary. In addition, an appropriate software tool will be helpful to improve the transcribing efficiency and quality of annotations.

### 4.1 The Criteria of annotation

For a speech corpus, the contents of its annotations in a corpus determine its usability and usability range. Though we intended to firstly focus our research on mandarin, and low noise environment spontaneous speech, these conditions can not be controlled and met while collecting speech in real life. So, in order to facilitate our research, in IIS_CSS corpus, annotation is composed of two different parts. One part contains the general information of an utterance, such as dialects, mood, environment noise and the speaker's sex.. The second part includes texts corresponding to each speech segments.

First, we will introduce the criteria of labeling general information for ISS_CSS. We classify dialects in Chinese speech into ten distinct types besides mandarin, transcribers select one of them for each utterance. Annotating dialects can help us to select utterances to formulate a right training set or test set and to control the regional distribution of speakers in this corpus. Mood means the speaker's evident aptitude. When transcribing, transcribers subjectively assess and make a choice of several listed items. Environment noise has been quantified to three levels; transcribing this information serves to study on robust speech recognition. Labeling the speaker's sex facilitates us to balance the proportion of males and females in corpus.

The second part, text annotation is the main body of the annotation task. Transcribing written language is easy because human has gained a high ability to map speech segments to words. his native language But for spontaneous speech including many speech events which can not be represented by common Chinese characters, it becomes difficult. To a certain extent, this diffculty results from our daily habit to ignore instead to classify these events. Another main reason comes from these events themselves has no evident acoustic stability. Considering the limits of human's ability to distinct

between non-written language speech events[4][5] and some experiences of other successful corpus, we compressed almost unlimited such special speech events into seven symbols which were listed in table 1.

ISS_CSS corpus aims to provide speech material for acoustic modeling and textual material for language modeling and dialogue management models. So we first transcribe common speech with Chinese characters which will be used to train language models and automatically create dialogue framework, and then automatically convert them to sequences of Pinyin (Chinese syllables) which serve to train acoustic models. Due to some Chinese characters have different pronunciations depending on context, transcribers are required to check their associated Pinyin. Thus, the whole label symbol set consists of Chinese characters, Pinyin, special label codes.

Another two remained problems are about mispronunciation and number. Mispronunciation often occurs in daily life. For example the speaker probably read Chinese character "**山**", which correct pronunciation is "shan1", as "san2". The last number is the tone . For this condition, we transcribe its associated speech segment to "**山**(san2)" to record text and pronunciation. Arabia representation of numbers is a natural method, but it can not map to a single pronunciation. So, transcribers are required to transcriber all numbers with Chinese characters.

### 4.2 annotation cost

Annotations were performed by trained transcribers, usually our employees, who have assisted speech research for no shorter than one year and have good working records. In our experience, the annotation of an entire call of approx. 4 minutes speech took about 60 minutes.

An annotation tool software has been developed, in which speech segments can be selected freely and replayed; label codes listed in a box can be added by double click; necessary time stamps can be provided by simple click on button "BeginStamp" or "EndStamp". All these work together aim to reduce transcribers' workload and the possibility to make errors. As soon as a sequence of Chinese characters, label codes and time stamps has been finished for an utterance, it will be automatically transferred to a sequence of PinYin (Chinese syllables) and label codes by a automatic pronunciation notation module. Since some Chinese characters have several pronunciations, sometimes transcribers' need to select proper pinyin-labels for these special Chinese characters.

In order to assure the validity of annotation, we arrange three-step labeling procedure for each utterance, each step is completed by different transcribers. The first transcriber only annotates texts for each utterance. The second annotate the general information and check the common Chinese characters. The third one check the label codes for special speech events. In each step, the transcriber's code is recorded.

## 5. Statistical Results

From the transcription, some figures have been computed in order to evaluate the coverage of this corpus. Currently the corpus available is one forth of the entire corpus foreseen in front of project requirements. The corpus contains 350 calls, 4600 utterances, 10-hour speech. According to our record , 147 speakers (93 males, 54 females) has contributed their speech. Currently only 5 percent of utterances fall into special dialects. 85 percent of speech were spoken in office-like environment. The occurrence counts of numbers, and spontaneous speech events were listed in Table 1.

## 6. Conclusion

This paper reports our progress in collecting corpus for CEST project. Collecting procedures, data processing, transcription has been explained in detail. This corpus will be completely finished by 2000, November in our plan. In the future, the transcription will be expanded to serve for translation. Further more, we will continually collect other domain-specific corpus and close-speaking corpus.

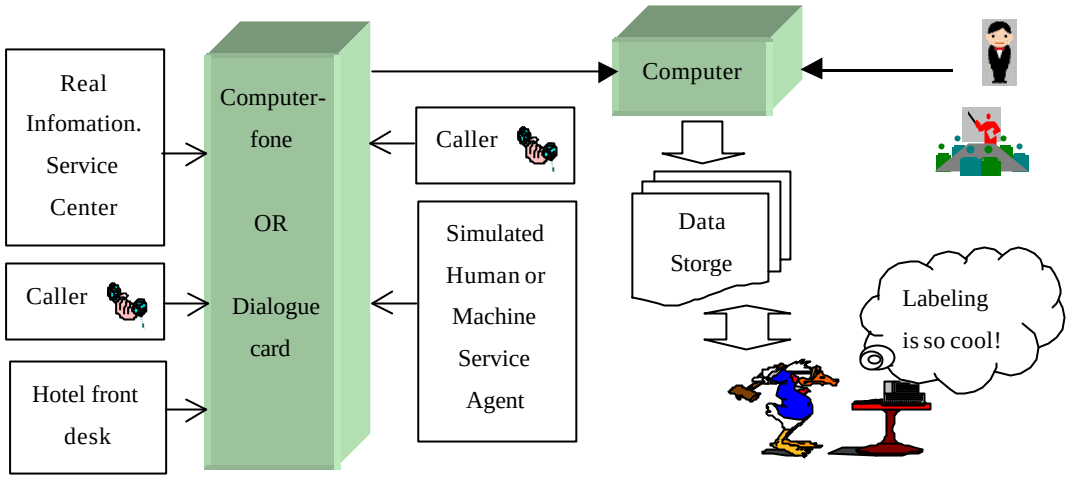


Figure 1: The framework of data collecting

| Special Events | Count | Explaination |
|---|---|---|
| Numbers | 700 | Numbers |
| Filled pauses | 5900 | Short non-silence disfluencies, such as [um],[uh] [eh] [ou] |
| Hesitation | 300 | Short silence in the context of disfluencies |
| Laugh | 109 | Laughter |
| Breath | 98 | Breath |
| Bksound | 2300 | The caller speak in a evident noise environment. |
| MutiSound | 570 | The caller' and the service agent speak at same time. |
| Barge_in | 68 | The speakers barge in the system's prompt. |
| Echo | 30 | The machine's echo prompt. |
| Noise | 2000 | Non-speech Noise and background speech noise |

Table 1: Statistical results of IIS_CSS corpus (8-hour human-human dialogue, 2-hour human-machine dialogues)

## 7. Referances

[1] "A Chinese-English Speech Translation Prototypes System CEST-CAS1.0", Limin Du, Junlan Feng, Yi SONG, Jinchen SUN, Qun Liu, the proceeding of ICSPAT1999

[2] " A General Architecture for Task-Oriented Spoken Dialogue System", JunLan Feng and LiMin Du, the proceeding of ICSPAT1999.

[3] "The Acquisition of a Speech Corpus for Limited Domain Translation", Demetrio Aiello, Loredana Cerrato, Cristina Delogu, Andrea Di Carlo Fondazione Ugo Bordoni pp2223-2226, EuroSpeech1999.

[4] "Modeling Disfluency and BackGround Events in ASR for A Natural Language Understanding Task", R.C. Rose and G.Riccardi, pp 1709~1712, ICASSP99

[5] "Two Swedish SpeechDat Databases - Some Experiences and Results, Kjell Elenius, pp2243-2246, EuroSpeech 1999