ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken
Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

DEVELOPMENT OF A CHINESE TELEPHONY CONVERSATIONAL CORPUS FOR SPEECH PROCESSING

LIU Yi[1], *Pascale Fung*[1], *Shudong HUANG*[2], *Christopher Cieri*[2], *ZHAI Lufeng*[1], *CHEN Benfeng*[1]

[1]Human Language Technology Center, Department of Electrical and Electronic Engineering,
University of Science and Technology (HKUST), Hong Kong
{eeyliu,pascale,lfzhai,bfchen}@ee.ust.hk
[2] Linguistic Data Consortium, University of Pennsylvania, U.S.
{shudong,ccieri@ldc.upenn.edu}

ABSTRACT

This paper describes the development of the EARS (Effective, Affordable, Reusable Speech-to-text) Chinese corpus, a telephony conversational speech database for speech processing. The EARS database is the first of its kind collected for Mandarin Chinese telephony spontaneous speech. The purpose of developing this EARS Chinese corpus is to collect Mandarin conversations between either strangers or friends, which cover a wide range of topics, over landline and cellular channels. All the speech data are annotated with standard Chinese character transcription as well as specific mark-ups for spontaneous speech. This corpus will be used for conversational and spontaneous Mandarin speech recognition tasks, under the DAPRA EARS framework. This paper introduces the design, development, structure, and initial phonetic analysis of the first 50-hour collection of this corpus. Additional 300 to 500 hours of data will be collected and transcribed between 2004 and 2005.

1. INTRODUCTION

Speech database is the fundamental resource for automatic speech recognition (ASR) and speech synthesis [8]. Nowadays, almost all state-of-the-art ASR systems are based on statistical models which need to be trained on a large amount of data recorded from a large number of speakers with different ages, accents, speaking styles, speaking modes, etc. The recognition accuracy of ASR systems relies heavily on the acoustic and language models, which are trained from selected speech database using statistical algorithms. Therefore, a high quality database with sufficient acoustic and style coverage, in a natural setting, of a large variety of speakers, is essential for the next step in ASR research.

There have been tremendous efforts in creating different types of corpus for English and other major western languages [8]. The Linguistic Data Consortium (LDC), established in 1992, is a center for supporting and coordinating corpora development activities. LDC has released more than 200 corpora in more than 20 languages to over 750 worldwide organizations. For Mandarin Chinese database, there exist several versions with different content for different usage. For example, the 863 Putonghua database is used for microphone read speech processing [14]. HKU96 and HKU99 are also a microphone read speech corpora focusing on the coverage of the phonetic characteristics in continuous speech [13]. The Hub4NE Broadcast News corpus provided by LDC is a collection of Broadcast News speech recorded from mainland China, Taiwan and US television and radio stations. The MAT telephony speech database consists of spontaneous Mandarin speech collected in Taiwan [10]. The CASS corpus, which is created to cover most of the phonetic variations in spontaneous Mandarin speech [6] from lecture speech. The CallHome Mandarin corpus, which consists of conversational speech between family members and friends over long-distance telephone. These speech databases have been providing the infrastructure for Mandarin Chinese speech processing as well as Chinese phonetic research.

The focus in ASR research has gradually shifted from read speech to spontaneous and conversational speech recently [7]. In particular, telephony conversational speech recognition is the latest pursuit by the ASR community since this type of speech is commonly used in daily life, and a lot of practical speech applications are based on telephony conversational speech. SWITCHBOARD is a typical telephony conversational speech corpus is widely used for English [3]. The Switchboard corpus contains a large amount of different types of speakers with sufficient speech data for acoustic model training and testing used in English ASR research. However, for Mandarin Chinese speech research, there is a dearth of high quality telephony conversational speech data. CallHome is a telephony conversational speech database. However, this database only contains two hours of data and all the conversational topics are related to family and school life. Moreover, there are many cross-talks since several family members might talk in the background at the same time. It is difficult to generate robust acoustic models for conversational speech recognition using the CallHome database. MAT is another Chinese telephony spontaneous speech corpus. Its disadvantage is that all the recorded speakers are from Taiwan, mostly with a strong Taiwanese accent. Such a database is not applicable to ASR systems for the majority of Mandarin speakers who are from mainland China. Therefore, it is desirable to develop a Mandarin telephony spontaneous conversational corpus from mainland Chinese speakers. This corpus should be analogous to the English EARS corpus and can be used efficiently for Mandarin telephony conversational speech recognition.

0-7803-8678-7/04/$20.00 ©2004 IEEE

Supported by the DARPA EARS project [4], we develop a telephone/cellphone channel Mandarin conversational speech database. The speech data in the corpus are natural spontaneous telephone conversations between two native Mandarin speakers, who might or might not know each other previously. The length of each conversation is between six to ten minutes. Each conversation focuses on one topic. The topics of conversations are selected from a wide range of domains relevant to Chinese culture. The speakers in the database are gender balanced and are from various regional and age groups. In order to cover all possible variations in telephony speech, the recorded data are over the public telephone network and include fixed-line, cell phone and IP telephones. In addition, the recorded speech data are selected and manually annotated with standard Chinese character transcription as well as specific spontaneous speech mark-ups. This paper describes the 50-hour data initially collected Mandarin telephony conversational speech data. This 50 hours of speech data and its transcription has been prepared and disseminated to the EARS community, while additional 300 to 500 hours of data will be collected and transcribed between 2004 and 2005.

The paper is organized as follows: Section 2 summarizes the database design in detail. Section 3 describes the corpus transcribing procedure and principles. Section 4 analyses the database from a phonetic viewpoint.

2. DATABASE DESIGN

The design of the entire 200 hours corpus, scheduled to be completed by 2005, about 1200 pairs of fluent speakers (native speakers preferred) of Mandarin will be recruited to talk via the public telephone/cellphone networ. One call per speaker is preferred. 10 minutes conversational speech per speaker pair will be recorded and saved as 8KHz and 8Bit, A-law waveform. For a conversation, speech from each speaker is recorded and saved to a wave file separately over a single channel. The two wave files are then merged into a single 8KHz, 8Bit, U-law SPH format file automatically by a self-developed merging tool at a post-process step. A website with MySQL database is also setup to register speaker and speech information such as gender, age, dialect, secondary language, birth place, education level, phone type, background noise, etc.

2.1. Topic Information

The conversation between two speakers is recorded at any given session. In order to create a flexible, natural and creative conversational exchange, the conversational topics are not "enforced" i.e., the call is included as long as there is continuous conversation on any topic. There are altogether 39 topics in the collection. Unlike the "CALL HOME/CALL FRIEND" conversations, which tended to be all about family and school, our topics cover a range of domains relevant to the mainland Chinese culture, as illustrated in Table 1.

2.2. Speaker Information

In addition to recording, all speakers need to provide a paper-based form of speaker information. This form includes information of speaker gender, age, dialect, language (multi-choice), birth place, cities the speaker lived in for more than half a year (multiple-choice), education level, noise, and phone type. Before real recording begins, questions related to this set of speaker information are asked and all the telephony speech answers from speakers are recorded and each category of information is saved to a separate wave form file. This section will summarize this speaker information for the collected 50 hours speech data.

Topic ID	Topic
0	TV sports program
1	Marriage and life
2	Comedy and joke
3	Suppose you are given 1 million to go abroad
4	Your career life
5	Suppose the time tube exists
...	...

Table 1: Samples of 39 definitions of topics

2.2.1. Speaker Gender Balance and Age Coverage

This set of data consists of 588 speakers, of which 271 are male and 317 are female. The recruited speakers cover an age range between 16 and 32 (Figure 1). Special attention of gender balance and age coverage is paid during recruiting.

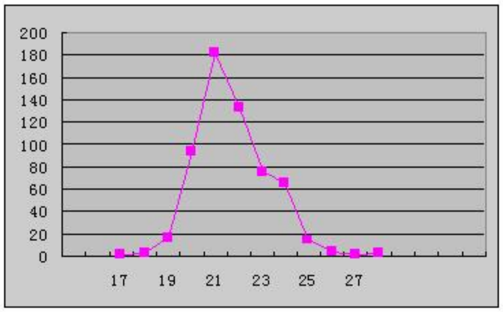


Figure 1: Speaker age distribution of 50 hours speech data (Speaker Number vs. Speaker Age)

Birth Place	Speaker Number	Birth Place	Speaker Number
Guangdong	168	Guangxi	4
Jilin	137	Shandong	4
Shanghai	34	Beijing	3
Hunan	32	Gansu	3
Anhui	27	Guizhou	3
Liaoning	25	Wuhan	3
Sichuan	24	Xizang	3
Henan	19	Hainan	2
Jiangsu	17	Hebei	2
Jiangxi	13	Chongqing	1
Shenzhen	13	Neimenggu	1
Heilongjiang	12	Ningxia	1
Hubei	12	Qinghai	1
Fujian	11	Tianjin	1
Shanxi	11	Unkown	1

Table 2: Speaker birthplace distribution of 50 hours speech data

2.2.2. Speaker Accent

All recorded speech in the database is in standard Mandarin or with acceptable accent only. Strongly accented speech and regional dialect speech are not included in the database. 410 speakers (about 70%) are standard Mandarin speakers. The information of speaker birthplace is collected and summarized in Table 2 as auxiliary and objective factor to judge speaker accents.

2.3. Speaking style and rate

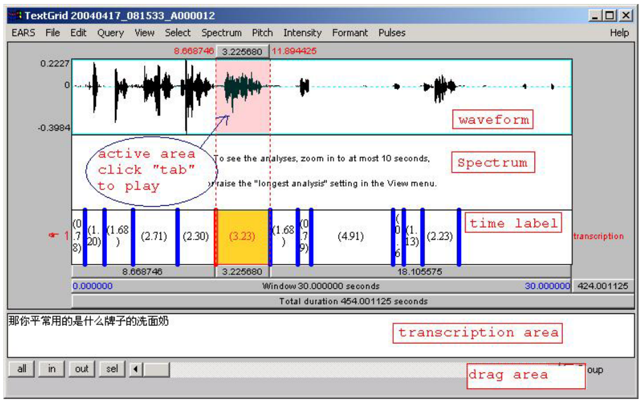
Speakers are encouraged to speak naturally, and are discouraged from imitating broadcast news, use too much emotion, or speak too fast or too slow. Each speaker makes about 70 to 90 utterances per 10 minute conversation under normal speaking rate, and most utterances are about 8 to 10 words. Ideally, the maximum number of Chinese word should be less than 20 for acoustic model training. Table 3 shows the statistical information on the average speaking rate.

Average sentence number per 10 min speech per speaker	80.7
Average sentence length	4.5 s
Average character number per sentence	13.4
Average speaking speed	0.336s per character

Table 3: Speaking rate information

3. TRANSCRIPTION PRINCIPLES

The goal of transcription is to provide an accurate, verbatim transcript of the entire database, which is time-aligned with the audio file in sentence level [5]. Currently, speech files are transcribed using standard simplified Chinese orthography in GB code according to what the transcribers hear. Additional features of audio signal and speech are annotated with specific mark-ups for spontaneous speech. At the later stage, characters will be converted into Pin Yin according to a standard pronunciation lexicon.



To see the analyses, zoom in to at most 10 seconds, or raise the "longest analysis" setting in the View menu.

那你平常用的是什么牌子的洗面奶

Figure 2. Main user interface of the of transcription tool

As a good user interface is conducive to good transcriptions of a large amount of speech, a tool for segmenting, labeling and transcribing speech is developed with a user-friendly interface (Figure 2). Using this tool, transcribers segment the speech to sections shorter than 10s according to the wave they see and what they hear. Each transcription corresponding to a speech file is saved with two formats: XML and TextGrid [9], which are flexibly designed and can be easily transferred to any other format.

3.1. Orthography and Spelling

In spontaneous speech, the pronunciation of individual Chinese characters/words is often not standard. Take a word “喜欢” as an example, some speakers from Guangdong say “xi fan” instead of “xi huan”. In such a case, annotators write the exact character and words as they hear, assuming the pronunciation is standard. Here the standard Chinese characters “喜欢” are used and not some alternate characters that represent the non-standard pronunciation (such as 喜翻).

All numerals are written out as complete Chinese characters as they are pronounced, not in Arabic numerals, e.g., 一九九七, 一千九百九十七, 幺三三四五六. Acronyms that are normally written as a single word but pronounced as a sequence of individual letters are written in all caps, with each individual letter preceded by a ~ tilde symbol, e.g., ~I~P, ~M~S~N, ~A~C 米兰

3.2. Disfluent Speech

Regions of disfluent speech are particularly difficult to transcribe. Speakers may stumble over their words, repeat themselves, utter partial words, restart phrases or sentences, and use lots of hesitation sounds. Annotators take particular care in sections of disfluent speech to transcribe exactly what is spoken, including the filled pauses, repetitions, partial words and mispronounced words used by the speaker.

Filled pauses are non-lexemes (non-words) that speakers employ to indicate hesitation or to maintain control of a conversation while thinking of what to say next. Each language has a limited set of filled pauses that speakers can employ. Annotators use the standardized spellings for filled pauses. We use only three Chinese filled pauses: %呵, %呃, %唔. All filled pauses are indicated with a % sign preceding the word.

When a speaker breaks off in the middle of the word, annotators transcribe as much of the word as can be made out. A single dash - is used to indicate point at which word was broken off, e.g., 纸-纸包装, 分-分区.

Partial pronunciations of Chinese words are indicated by a PinYin sign with the ~ sign. For example, in the sentence “最想去哪个 sh~什么山啊”, “什” with standard Pinyin “shen” is pronounced as “sh”.

Speaker restarts are indicated with double dash --. Annotators use this convention for cases where a speaker stops short, cutting him/herself off before continuing with the utterance.

体育节目就体--就足球--足球那些喽
那去--进去那买了

A plus symbol + is used for obviously mispronounced words (not regional or non-standard dialect pronunciation). These words are transcribed using the standard character and not try to represent the pronunciation.

那我喜欢+旅游　（not 那我喜翻+旅游）

3.3. Noise

Speaker-produced noise is identified with one of the following five tags: <laugh>, <breath>, <cough>, <lipsmack>, <sneeze>. When there is noticeable background noise (not speaker noise) present during a span of speech, a <noise> tag is used.

4. PHONETIC CHARACTERISTICS

Chinese is written with characters known as hànzi. Each character represents a syllable of spoken Chinese and has a meaning. Chinese is a monosyllabic language. This corpus contains totally 637,579 syllables and covers 408 base syllables.

In Chinese ASR systems, initial and final units are always used as subword units instead of phonemic units. Initials and finals are the smallest natural pronunciation units in Chinese speech, and are different from phonemes or phones in western language. One initial corresponds to one phoneme, while one final may consist of one or several phonemes. There are 21 initials and 37 finals for Putonghua. This database covered all the 58 initial/finals of Putonghua. In these initial/finals, ‘e’ is most frequently used and appears 84076 times. Even the least frequently used ‘vn’ appears as many as 875 times. (Figure 3)

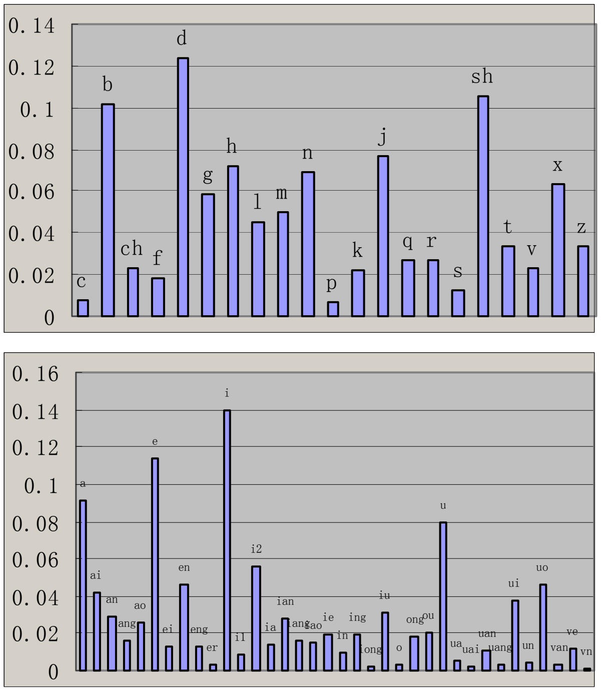


Figure 3 Distribution of standard Chinese Initials and Finals

5. SUMMARY

We describe the design and collection of a new Chinese telephony conversational speech database. This database is a first of its kind that provides naturally spoken conversations between two native Mandarin speakers on a variety of pre-assigned topics. This database is being made available to DARPA EARS sites, to be used to analyze conversational Mandarin, to design and evaluate algorithms for Mandarin speech recognition. It provides waveform speech data， the orthographic character level transcription data and speaker and speech information. At the present stage, the data amounts to approximately 5.5 Gigabytes (GB). More speech data is now continuously being collected. The database will also be phonetically labeled in the future. A relational software environment including recording, transcribing, labeling, and speaker information management has also been developed.

6 REFERENCES

[1]. Chan, C., 1998, “Design considerations of a Putonghua database for speech recognition”, In: *Proceedings of the Conference on Phonetics of the Language in China*, pp. 13–16.

[2]. Huo, Q., Ma, B., 1999, “Training material considerations for task-independent subword modeling: design and other possibilities”, In: *Proceedings of 1999 Oriental COCOSDA Workshop*, pp. 85–88.

[3]. J. Godfrey, E. Holliman and J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development", in *Proceedings of the IEEE ICASSP*, vol. 1, pp. 517-520, San Francisco, CA, USA, March 1992.

[4]. LDC EARS Project Website: http://wave.ldc.upenn.edu/Projects/EARS/

[5]. LDC EARS Project RT-04 Careful Transcription Guidelines: http://www.ldc.upenn.edu/Projects/Transcription/rt-04/RT-04-guidelines-V3.0.pdf

[6]. Li Aijun, Zheng Fang, William Byrne, Pascale Fung, Terri Kamm, Liu Yi, SONG Zhanjiang, Umar Ruhi, Veera Venkataramani, CHEN Xiaoxia, “CASS: A Phonetically Transcribed Corpus Of Mandarin Spontaneous Speech”, *International Conference on Spoken Language Processing*, October, 2000, Beijing, China.

[7]. LIU Yi and Pascale Fung, “Modeling Partial Pronunciation Variations for Spontaneous Mandarin Speech Recognition”, *Journal of Computer Speech and Language*, Vol.17, No.4, pp. 357-379, Oct. 2003

[8]. Tan Lee, W.K. Lo, P.C. Ching and Helen Meng, "Spoken language resources for Cantonese speech processing", in *Speech Communication*, Vol.36, No.3-4, pp.327 - 342, March 2002.

[9]. TextGrid as an objection of PRAAT: http://www.fon.hum.uva.nl/praat/manual/TextGrid.html

[10]. Tseng, C.Y., 1995. A phonetically oriented speech database for Mandarin Chinese. *Proceedings of 1995 International Congress of Phonetic Sciences*, Vol. 3, pp. 326–329.

[11]. Wang Renhua, Xia Deyu, Ni Jinfu, and Liu Bicheng, “USTC95 - a Putonghua Corpus”, In *Proceedings of 1996 International Conference on Spoken Language Processing*, Vol. 3, pp. 1894-1897.

[12]. Wu, Y., 1998, “Chili Mandarin speech corpus”, In: *Newsletter of ISCSLP’98 Special Interest Group: Linguistic Database and Tools*, pp. 1–3.

[13]. Zu, Y.Q., Li, W.X., Ho, M.C., Chan, C., 1996. HKU96 – a Putonghua corpus (CD-ROM version). Department of Computer Science, University of Hong Kong.

[14]. Zu Y.Q., 1997. “Sentence Design for speech synthesis and speech recognition”, In *Proceedings of 5th European Conference of Speech Communication and Technology*, vol2., pp.743-746.