2nd International Workshop on East-Asian Language Resources & Evaluation (Oriental COCOSDA'99), 129-132, May 13-14, 1999, Taipei

A CANTONESE ACCENT CHINESE SPEECH CORPUS

Shuqing Li, Fang Zheng, Mingxing Xu, Zhanjiang Song, Ditang Fang
Speech Lab, Department of Computer Science & Technology
Tsinghua University, Beijing, 100084, China
+86-10-62772001, fangdt@mail.tsinghua.edu.cn

ABSTRACT

To meet the needs of research in the robustness of the continuous Chinese speech recognition systems, we have established a Cantonese accent Chinese speech corpus (*CACSC*) as the first of a series of Chinese speech corpora with different accents. *CACSC* contains 25 Giga Bytes utterances uttered by 104 males and 100 females. The sampling was undertaken at 16KHz rate with 16bit-width data precision through a standard SoundBlaster of a personal computer under ordinary office environment. *CACSC* is mainly based on the standard Chinese, known as Mandarin, with light Cantonese accents. The establishment of *CACSC* offers a testing bed for robust speech recognition of a certain regional accent. This paper is to describe how CACSC was established and its features in details.

Keywords: Speech corpus, Cantonese accent, continuous speech recognition

1. INTRODUCTION

In order for the research of the Speech Recognition and Understanding (SRU) in the large-vocabulary speaker-independent Chinese dictating systems, we need to establish a large Chinese speech corpus with speech materials covering a broad range plus speech with many different accent backgrounds. This will help to solve the problem of many variables present in Chinese due to different speaker origins, and also to adapt to the feature differences in speaker-independent speech recognition, so as to improve the robustness of the speech recognition systems.

The speech corpus we will describe here is a Chinese speech corpus with Cantonese accent which is the first of a series of Chinese speech corpora with regional accents to be established, such as Yue (Canton Province and Hong Kong), Wu (Jiangsu Province & Zhejiang Province), Min (Fujian Province and Taiwan Province). In this paper, we will describe the project of establishing this first corpus and its features in greater details.

The following aspects will be covered in this paper: the design of the materials, the speakers, the recording condition of the corpus, and the format of the data packages.

2. REQUIREMENTS

The Chinese speech corpus with Cantonese accent we are recording needs to have a minimum amount of speech materials (sentences) and a minimum number of speakers, so that it may provide sufficient information for training and test sets. The speech materials are based on sentences. It needs to cover most of the co-articulations in the continuous Chinese speech. It also needs to be balanced among all syllables. The speakers are with Cantonese accent, but sentences are not spoken in Cantonese. They use the read speech style to speak a complete (semantics) sentence with punctuation marks pronounced. Speakers are required to speak clearly at a normal speed and a natural volume, and they don't have to be exactly the same.

The recording conditions are very similar to where the speech recognition system is being used, like in quiet offices or computer rooms.

3. ESTABLISHING THE SPEECH CORPUS AND ITS FEATURES

3.1. Speech Materials

The design of the speech materials are based on the following demands:

- The pre-defined Chinese sentences should be with all Chinese syllables well balanced. Almost all syllables should be covered while more frequently used syllables may appear more often. This is to reduce the amount of sampling data and number of speakers while there are still enough training data for the least frequently seen syllables.
- The R-suffix Chinese syllables are considered to appear in the read Chinese sentences. The R-suffix syllable is a normal Chinese syllable followed by a suffix who causes a retroflexion of the preceding vowel, typical of the pronunciation of standard Chinese and of some dialects.
- The reading style is sentence by sentence with punctuation marks read out.

Part of the materials is taken from the "863" test speech corpus, i.e. the first group of speech materials used by the speech corpus for speech recognition and speech synthesis supported by the national "863" plan in 1997. This part includes 1,560 sentences chosen from the People's Daily of the years of 1993 and 1994, it covers most Chinese di-phone and tri-phone and the basic sentence forms. There are 397 toneless syllables appearing in the sentences, but 21 syllables are not present. The three relatively most frequently used syllables are

"LIA", "O", "ZHEI".

In order that most of the syllables should appear in the materials, we design 80 additional sentences as a common part of the materials to be used in the establishment of *CACSC* for the syllable balance purpose. In the 80 sentences, the 21 unseen syllables, especially the above three syllables, will appear more frequently and repeatedly. So totally there are three groups by 600 sentences. In each group, 520 sentences are taken from the "863" test speech corpus and 80 sentences of our own. Each speaker is asked to read one of the three groups of sentences.

3.2. Recording Condition

The recording condition of corpus is well designed.

Sentences should be read in Mandarin pronunciation with light Cantonese accents.

Close-talking microphones and stand SoundBlasters should be used.

Sufficient silences at the start and the end of each speech segment should be kept.

Utterances should be recorded under ordinary office environments.

The ADC specification is 16KHz sampling rate and 16bit data width.

Speakers should be covered according to the actual proportions of the age (teenager, youngster, middle-aged and elder people) and education (junior high school, senior high school, college, university) distributions.

According to the above requirement, the actual recording conditions are as follows.

Environment: similar to ordinary laboratory conditions. There are 4 recording systems in one room; 3 or 4 people are reading the pre-defined groups of sentences at the same time; no sound proof is setup.

Hardware platforms: PC (Pentium-II), hard drive, CD-ROM, CD-ROM writer.

Microphone: SHURE/CD1-14 microphone, direction-oriented.

Sound Card: Sound Blaster - 16.

Sampling Rate: 16 Khz.

Sampling Accuracy: 16 bits.

Speaking Speed and Volume: Natural, no special requirements.

User Interface: Displayed on the screen is the sentence to be recorded, with the Pinyin string, punctuation marks. The correct pronunciation may be played back on demand. The correct pronunciation auto-detection and alarm on background noises and volume magnitude are available. Speakers can re-record, modify and go forward and backward. The system saves speech data in WAV format upon finish of recording of every sentence with general test capability.

3.3. Speakers

Also according to the recording conditions, speakers are well selected. The following is detailed information of *CACSC*.

Total number of speakers: 204, 104 males and 100 females.

Age: 17-25.

Occupation: mostly college students.

Accent background: all speakers were born in Canton, they spoke Cantonese before college. They speak Mandarin with a noticeable Cantonese accent. A test was performed to each speaker so that they meet our needs. They were required to speak out 50 pairs of confusing syllables such as zh-z, ch-c, sh-s, in-ing. Only those who confused these syllable pairs were asked to stay for recording.

Grouping: we divided the speakers into three groups, for both males and females. There are 30-40 people in each group. Speakers in Group 1 read sentences No. 1-520 and the 80 common sentences, Group 2 No. 521-1,040 and 80 common sentences while Group 3 No. 1,041-1,560 and 80 common sentences.

3.4. Directory and File Name Convention

Data files are stored in a directory for each speaker. The directory has a *Gddd*-style name, where *G* means gender ('M'=male and 'F'=female), and *ddd* is a base-10 speaker ID. Actually, the base-10 speaker ID ranges '000'-'499' for males and '500'-'999' for females.

Under each speaker's directory, there are 6 sub-directories named '1' to '6'. In each sub-directory, there are data of 100 sentences.

Each data file has a name like *ddd* ^ @@@@, where:

ddd is the base-10 speaker ID, ranging from 000 to 499 for males and from 500 to 999 for females;

^ is an English letter, indicating the speaker's accent background, 'A' for Canton and Hong Kong, 'B' for Jiangsu & Zhejiang, 'C' for Fujian and Taiwan, and 'D' for northern China area.

@@@@ is a base-10 number, an id number for each sentence, ranging from 1 to 1800.

The data file should contain the following information.

The Background Noise Level (BNL).

The Recording Date (RD).

The Microphone Specification (MS) including frequency response range, impedance, sensitivity, and so on.

The Analog Digital Conversion Information (ADCI) including sampling rate, precision, channel number, channel noise level, data format, and so on.

The Speaker Information (SI) including name, gender, age, accent, education

Chinese Text Information (CTI) including GB codes, pinyin, *etc.*

The Data Package File Information (DPFI) including length of file header, length of file, length of data segment, number of sentences, and the data body.

3.5. Data Size and Storage

Each speaker speaks for about 4 hours (about 100 to 140 MBytes) and the total data size is about 28 GB. There are about 4 to 5 speakers' data stored in one CD-ROM, totally about 41 CDs' (with 25Giga Bytes' speech data) corpus has been established.

The label for each CD has the following form:

^AG -- ddd ~ ddd

where:

^: is sentence group number, ‘1’, ‘2’, ‘3’,
A: is accent background letter, ‘A’, ’B’, ‘C’, ‘D’,
G: is gender letter, ‘F’ or ‘M’, and
ddd ~ ddd: is the speaker ID range.

4. CONCLUSIONS

The Chinese Mandarin speech corpus with Cantonese accent we have recorded is already very general and complete, but still there is a lot of room for improvements.

1.Age of the speakers: originally we planned to have: young teenagers (voice-changing period) under 16, young-aged (senior high school, college and graduate students) 16 to 30, middle-aged 31 to 60 and seniors 61 and older; four age groups. But in this phase as we do the recording in Beijing, we are mostly having people from only the second group. We will add more speakers of other groups later when we go to Canton.

2.The number of speakers we have is 204 which is not very high. Since if we required the minimum number of times for each syllable to be recorded, we need at least 200 speakers for male and female. Obviously, our current number is still a little low.

3.Dialect background: considering the percentage of Chinese-speaking population around the world and significant difference in spoken Chinese in different regions, in addition to Chinese Mandarin with accent, we also need to record Chinese speech corpus in different dialects.

4.Our current speech corpus data is only a continuous-speech waveform data in .WAV file format. Only digitalization is done. To put in use for speech recognition research, we still need to do a lot work of dividing and labeling the syllables.

In conclusion, the establishment of a benchmark speech corpus is very necessary and helpful for research in speech recognition. But recording a speech corpus takes a lot of human and other recourses. For large scale speech corpus with many accents or in many dialects, it is an even bigger challenge.

If this work can be organized and coordinated under a joint effort among the institutions that are conducting speech recognition research and in a standardized format, each takes part of the recording work and make it available for download on the network so everyone can share the database data, it will definitely be a very valuable venture. If there was a coordinator who can organize the whole work, divide it into several parts, and distribute each part to one institution like us, this work can be done more efficiently and perfectly in a standardized format. After all recording task finished, this coordinator or organization can collect all data together and put it on the web for shareing. Isn’t a greatly valuable venture? Should this become a viable plan, we will contribute with our best efforts.

REFERENCES

[1] Ditang Fang, "The Robustness of Speech Recognition and Establishing the Speech Corpus", *Computing Application and Software*, 1994.

[2] Fei Qu, Taiyi Huang and Xijun Zhang, "Design of Speech Materials for the Chinese Speech Corpus", *the 4th National Conference on Man Machine Speech Communication*, Beijing, 1996.

[3] Shuqing Li, Ditang Fang and Shan Qing, "A Large Chinese Speech Corpus", *the 4th National Conference on Man Machine Speech Communication*, Beijing, 1996.

[4] Jianfen Cao, "Fundamentals of Modern Speech", *People's Education Press*, 1990.

[5] Keynote Presentation for the '98 Intel International Speech Forum, Beijing, 1998.

Appendix I: The Chinese syllable table.

"a",	"ai",	"an",	"ang",	"ao",
"ba",	"bai",	"ban",	"bang",	"bao",
"bei",	"ben",	"beng",	"bi",	"bian",
"biao",	"bie",	"bin",	"bing",	"bo",
"bu",	"ca",	"cai",	"can",	"cang",
"cao",	"ce",	"cei",	"cen",	"ceng",
"cha",	"chai",	"chan",	"chang",	"chao",
"che",	"chen",	"cheng",	"chi",	"chong",
"chou",	"chu",	"chua",	"chuai",	"chuan",
"chuang",	"chui",	"chun",	"chuo",	"ci",
"cong",	"cou",	"cu",	"cuan",	"cui",
"cun",	"cuo",	"da",	"dai",	"dan",
"dang",	"dao",	"de",	"dei",	"den",
"deng",	"di",	"dia",	"dian",	"diao",
"die",	"ding",	"diu",	"dong",	"dou",
"du",	"duan",	"dui",	"dun",	"duo",
"e",	"ei",	"en",	"eng",	"er",
"fa",	"fan",	"fang",	"fei",	"fen",
"feng",	"fiao",	"fo",	"fou",	"fu",
"ga",	"gai",	"gan",	"gang",	"gao",
"ge",	"gei",	"gen",	"geng",	"gong",
"gou",	"gu",	"gua",	"guai",	"guan",
"guang",	"gui",	"gun",	"guo",	"ha",
"hai",	"han",	"hang",	"hao",	"he",
"hei",	"hen",	"heng",	"hm",	"hng",
"hong",	"hou",	"hu",	"hua",	"huai",
"huan",	"huang",	"hui",	"hun",	"huo",
"ji",	"jia",	"jian",	"jiang",	"jiao",
"jie",	"jin",	"jing",	"jiong",	"jiu",
"ju",	"juan",	"jue",	"jun",	"ka",
"kai",	"kan",	"kang",	"kao",	"ke",
"kei",	"ken",	"keng",	"kong",	"kou",
"ku",	"kua",	"kuai",	"kuan",	"kuang",
"kui",	"kun",	"kuo",	"la",	"lai",
"lan",	"lang",	"lao",	"le",	"lei",
"leng",	"li",	"lia",	"lian",	"liang",
"liao",	"lie",	"lin",	"ling",	"liu",
"lo",	"long",	"lou",	"lu",	"luan",
"lv",	"lue",	"lun",	"luo",	"ma",
"mai",	"man",	"mang",	"mao",	"me",
"mei",	"men",	"meng",	"mi",	"mian",
"miao",	"mie",	"min",	"ming",	"miu",
"mo",	"mou",	"mu",	"n",	"na",
"nai",	"nan",	"nang",	"nao",	"ne",

"nei", "nen", "neng", "ng", "ni",
"nian", "niang", "niao", "nie", "nin",
"ning", "niu", "nong", "nou", "nu",
"nuan", "nv", "nue", "nun", "nuo",
"o", "ou", "pa", "pai", "pan",
"pang", "pao", "pei", "pen", "peng",
"pi", "pian", "piao", "pie", "pin",
"ping", "po", "pou", "pu", "qi",
"qia", "qian", "qiang", "qiao", "qie",
"qin", "qing", "qiong", "qiu", "qu",
"quan", "que", "qun", "ran", "rang",
"rao", "re", "ren", "reng", "ri",
"rong", "rou", "ru", "rua", "ruan",
"rui", "run", "ruo", "sa", "sai",
"san", "sang", "sao", "se", "sen",
"seng", "sha", "shai", "shan", "shang",
"shao", "she", "shei", "shen", "sheng",
"shi", "shou", "shu", "shua", "shuai",
"shuan", "shuang", "shui", "shun", "shuo",
"si", "song", "sou", "su", "suan",
"sui", "sun", "suo", "ta", "tai",
"tan", "tang", "tao", "te", "tei",
"teng", "ti", "tian", "tiao", "tie",
"ting", "tong", "tou", "tu", "tuan",
"tui", "tun", "tuo", "wa", "wai",
"wan", "wang", "wei", "wen", "weng",
"wo", "wu", "xi", "xia", "xian",
"xiang", "xiao", "xie", "xin", "xing",
"xiong", "xiu", "xu", "xuan", "xue",
"xun", "ya", "yan", "yang", "yao",
"ye", "yi", "yin", "ying", "yo",
ong", "you", "yu", "yuan", "yue",
"yun", "za", "zai", "zan", "zang",
"zao", "ze", "zei", "zen", "zeng",
"zha", "zhai", "zhan", "zhang", "zhao",
"zhe", "zhei", "zhen", "zheng", "zhi",
"zhong", "zhou", "zhu", "zhua", "zhuai",
"zhuan", "zhuang", "zhui", "zhun", "zhuo",
"zi", "zong", "zou", "zu", "zuan",
"zui", "zun", "zuo"

Appendix II: 21 Syllables unseen in '863' materials.

"cei", "chua", "den", "ei", "eng",
"fiao", "hm", "hng", "kei", "lia",
"lo", "n", "ng", "nou", "nun",
"o", "rua", "shei", "tei", "yo",
"zhei"