STANDARDIZATION OF SPEECH CORPUS

Li Ai-jun, Yin Zhi-gang

Institute of Linguistics, Chinese Academy of Social Sciences

ABSTRACT

Speech corpus is the basis for both analyzing the characteristics of speech signal and developing speech synthesis and recognition systems.

In mainland China, almost all of the speech research and development affiliations are developing their own speech corpora. We have so many different kinds and a large number of Chinese speech corpora that it is rather important to be able to conveniently share these speech corpora to avoid wasting of time and money and to make the research work more efficient.

The primary goal of this research is to find a standard scheme which can make the corpus be established more efficiently and be used or shared more easily.

A huge speech corpus on 10 regional accented Chinese, RASC863 (a *Regional Accent Speech Corpus funded by National 863 Project，4 out of 10 regional data have been distributed until this time*) will be exemplified to illuminate the standardization of speech corpus production.

Key words: phonetics, speech corpus, standardization, production, specification

1. Introduction

Speech corpus, the collection of speech signal, its annotation, metadata and documents, is the basis for both analyzing the characteristics of speech signal and developing speech synthesis and recognition systems.

Speech corpus based technology has been widely used in people's lives, although it is still a strange concept for many people. An example is the automatic broadcasting system for traffic information. In this kind of system, the sound is not pronounced by actual speakers, but synthesized by TTS (text to speech) system based on a speech corpus.

Not only for TTS technology, but also for ASR (Automatic Speech Recognition) and phonetic research, speech corpus is very important. For phonetic research, speech corpus can provide diverse and accurate data to help researchers to find the rules of languages. For ASR, In order to "train" the system to "understand" any of the speakers' voice, great capacity speech corpus is necessary. Taking advantage of the statistic data of speech corpus, ASR system can transform the speech signal to text strings by using phonological, linguistic and stochastic analysis. That is why ASR can "understand" human's voice.

Owing to the importance of speech corpus, in China, corpus production has got a long term support of various national funds such as the 863 Hi-tech Project and 973 Development Program of China and the National Science Foundation of China. Many speech research and development affiliations have developed their own speech corpora in recent years.[1]

With the developing of speech corpus technology, the new problem appears: on the one hand, so many corpora have been established and so much money and time have been put into; on the other hand, these corpora are hard to share among different affiliations.

A main reason of this problem is the lack of general specifications on corpus collection, annotation and distribution. In order to solve the problem, standardization research of speech corpus is necessary and specifications should be stipulated.

2. Standardization research of speech corpus

Standardization of speech corpus includes many aspects as decrided below.

1). Legal correlated

Speech corpus and its production must abide the law of nation. These legal documents should be prepared: property right statement of the corpora, agreement with the speakers, agreement with the users, etc.

2) Standardization of collection procedure of speech corpus

Although speech corpus collection is only a procedure, it decides the quality and the efficiency. So the production procedure of speech corpus should be standardized, just as the ISO system for industry.

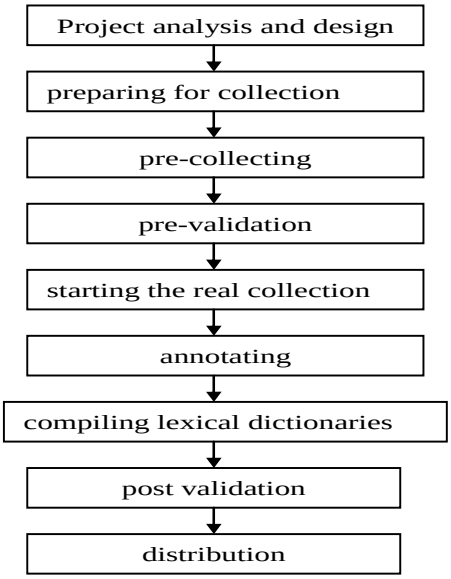

Fig 1: flow chart of the production of speech corpus (after [4])

Fig 1 shows the general procedure of producing a speech corpus. It is unnecessary to follow all of these steps. Some of them can be carried on simultaneously such as collecting and annotating, some can be skipped in a specific task; in fact, an additional step can be introduced by the producer. [2]

1) Project analysis and design: the task of this step is to analysis speech corpus project and draft out the blue print. The specification of corpus will be drawn: the size of corpus, quantity of speakers, speech style, recording equipments, etc.

2) Preparing for collection: preparing for corpus according to the blue print: designing the prompts, preparing hardware and software, raising money and organizing staff, finding speakers, etc.

3) Pre-collecting: if the speech corpus is very large and complicated, pre-collecting a few of samples is absolutely needed. It can find problems and improve the plan, avoid the possible misplay of formal collection.

4) Pre-validation: evaluating the pre-collected corpus and improving the blue print.

5) Starting the real collection.

6) Annotating: annotating the speech corpus.

7) Compiling lexical dictionaries.

8) Post validation: evaluating the speech corpus, examine whether it reach the criteria or not.
It is employed to accept or reject the corpus.

9) Distribution: distributing the speech corpus which passes the post validation

3). Standardization of speech corpus

Not only the procedure of producing a corpus, but also the corpus should be standardized. Table 1 shows the major specifications in producing a speech corpus. [3]

Specifications	Comments
Specification of speakers	Describing the speaker's features such as age, gender, educational background, voice quality, language and accent.
Specification of corpus design	Describing the corpus organization and contents. For instance, the detail information or scripts (prompts) organization of read and spontaneous speech, dialogues or monologues, elicited spontaneous speech (answering questions, etc.), expressive speech. Introduction to the phonetic or linguistic coverage and the algorithm used for selecting the corpus scripts.
Specification of recording	Describing the recording technical specifications on recording equipments, environmental conditions, the recording platform and the data storage strategy such as sampling rate, speech wave format….
Specification of annotation	Describing the annotation conventions of sound to characters transcription, phonetic annotation or other information such as syntactic annotation.
Validation Criteria	Setting explicit the criteria that the corpus should fulfill. Giving an overview of the features to be checked and the criteria employed to accept or reject the corpus.
Specification of distribution	Describing the distribution plan, principles and the storage medium.

Table 1. Specifications of speakers and corpus collection

3. Detailed Specifications exemplified by RASC863

In this chapter, RASC863，a Regional Accented Speech Corpus funded by National 863 Project, will be used to illustrate the above-mentioned standardization.

There are 10 dialect families in China, namely Guan (Mandarin), Jin, Wu, Hui, Xiang, Gan, Kejia (Hakka), Yue (Cantonese), Min, and Ping. It is well known that Chinese dialects differ greatly from each other and are not mutually intelligible. Thus it is quite naturally that Putonghua (Standard Chinese, hereafter SC), which is phonetically based on Beijing Mandarin, has been chosen as the communicative spoken language between people from different dialectal regions. However, people with different dialectal backgrounds typically speak SC with a certain degree of accent due to the influence of their mother tongue dialect. And this kind of influence could be phonetic, lexical, and/or syntactical.

In the recent years, with the development of the ASR techniques, collecting accented spontaneous speech corpora becomes an urgent demand in the field of speech technology, as well as in the field of phonetic sciences. Funded by the National 863 High-Tech Project, we collected a speech corpus with 10 representative regional accents, namely Chongqing, Shanghai, Guangzhou,Xiamen, Taiyuan, Changsha, NanChang, Wenzhou, Luoyang, and Nanjing.However, only the data of first four regions have been distributed by ChinesLDC. (http://www.ChineseLdc.org) So the following introduction will only focus on these four regions.

The corpus consists of spontaneous speech, read speech and selected dialectical words. For the spontaneous speech, each speaker was asked to select a topic or from our prepared topic sheet with a variety of 160 topics and then to give a 4-5 minute spontaneous speech on this topic. Besides, each speaker was asked to answer 15 elicted spontaneous questions. The read speech consists of 2200 phonetically balanced sentences, and 460 frequently used sentences in

daily life domain. For each dialectal region, we prepared those words or phrases that are frequently used in daily life and are different from Standard Chinese, and each speaker was asked to read 15 dialectal words. 800 speakers (200 from each region; balanced in terms of the age, gender, and educational background) were recruited in the project.

The detail specifications of RASC863 will be described as followings.

1). Specification of speakers

Specification of speakers describes the number of speakers to be recorded for each diactal accent and their characterizations. Sometimes it has to describe the speaking styles.

Speaker characterization concerns the distribution of age, education level, gender and the dialectal coverage aspired.

The speaking styles of speakers can be read speech, answering speech, command/control speech, descriptive speech, non-prompted speech, spontaneous speech, neutral vs. emotional speech and dialogue. The content of speaking can be described in different ways according to task, topic or simply in text description.

Table 2 describes the distribution of speakers' ages, genders, accent degrees and educational backgrounds for each region.

items	Levels	Male	Female
Age/gender	16-30 (y)	45	45
	31-45 (y)	45	45
	Older than 50 (y)	10	10
Education	Junior high school	5	5
	Senior high school	15	15
	Undergraduate/ graduated	80	80
Accent category	L1-A	0	0
	L1-B	5	5
	L2-A	35	35
	L2-B	35	35
	L3-A	20	20
	L3-B	5	5

Table 2 200 Speakers' distribution for each region

这里没有解释 L1-A….

2) Specification of corpus design

The aim of speech corpus design is to determine what to be recorded and to get the necessary script. Whether a corpus needs designated script before collection or not is determined by the corpus type and corpus content. [2]

The RASC863 prompt sheet for each speaker is shown in Table 3

Items	Speech style	Content
0	Spontaneous	4 to 5 minutes
1-15	Spontaneous	15 question answers
16-388	Read	23 common sentences
36-50	Read	15 dialectal words
51-165	Read	110 phonetically balanced sentences (<30 syllables each)

Table3: Prompt sheet for each speaker

3). Specification of recording

Usually the specification of recording contains recording guide, technical parameters, speaker recruiting plan

and approach, recording procedures, recording log files, pre-validation, etc.

Speech data of RASC863 were recorded directly into a laptop computer via an external USB M-audio sound card. And a Sennheiser earphone and a CR 722 capacitor microphone (20-20000Hz) were used simultaneously to acquire the audio signal. For each recording session, the acoustic environment and background noise (in db) was recorded.

The software Cooledit Pro 2.0 was used in recording the 4-5 minute spontaneous speech, and YYSRecorder was used in recording other speeches. All speech data were sampled in 16KHz and 16 bit.

A meta data file, as exemplified in table 4 for instance, was generated for each sound file.

Microphone and the speaker's positions are shown in Fig.2

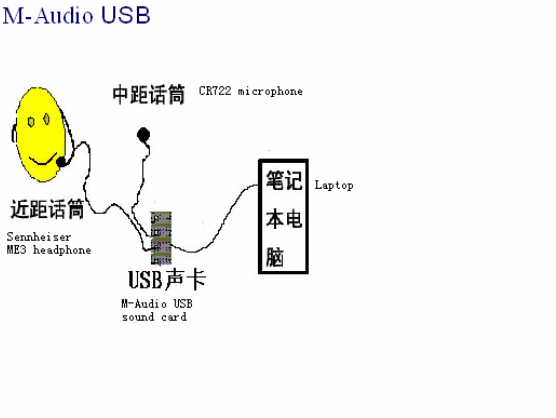

Fig. 2 Recording system mounting diagram

4) Specification of corpus structure

Corpus structure related to the corpus internal organization structure, the file naming rules and the storage media for distribution. Usually the released data format must be given. At present, a normal data structure is as shown in table3. File naming should be given based on the file content. Be sure that the length of the name has to follow the ISO-9960 file attributes; otherwise, it will be trouble to record the final corpus data on a CD or DVD.[3]

In RASC863, each recorded sound corresponds to a meta data file and a wave file which are automatically generated by recording software. The metadata file describes the detailed information related to this recorded sound file as shown in Table4.

//Root
DATA : speech data
// subdirectories may be added such as
Male/Female
Recording session
Speech types (read, spontaneous…)
…
ANNOT: annotation data
META: metadata about corpus itself
Specs: specifications of corpus
Prom: prompt files
DOC: documents
LEX: lexicon or its statistic files
TOOLS: recording, analysis or annotation tools

Table 3 A typical corpus structure

Session ID
Speaker ID
Date of Recording
Recording place
Speaking style
***** acoustic and technical description
Recoding sound name
Environmental Conditions
Microphones
Sampling rate
Bits per sample
***** Annotation part
Annotation Convention
***** Annotation should be like this
Orthographic annotation
Prosodic Annotation
Segmental Annotation
***** Or like this
Corresponding annotation file

Table 4 Meta data for each sound file

5) Specification of annotation

Speech corpus annotation includes speech to characters transcription, segmental annotation and prosodic annotation. Specification of annotation describes the annotation format, rules, tools, consistency criterion. Sometimes, if there are more than one transcribers transcribing or annotating simultaneously, their annotation consistency should be checked first.

In RASC863 project, the Chinese Character transcription, as well as the paralinguistic and non-linguistic labeling, has been made for the spontaneous part. Additionally, phonetic annotation has been made for read speech for 80 speakers, 20 from each dialectal region. The speech software Praat was employed for the phonetic annotation. C-ToBI3.0 and SAMPA-C annotation system were used in prosodic annotation and segmental annotation.[3]

6) Legal agreement

A very important thing is about the agreement between the producer and the speaker, often called speaker agreement, in which the usage of the recorded speech data or even some of the speaker's information should be clearly demonstrated. Other aspects, such as whether the speech data can be distributed or copied unlimitedly, should also be described in the agreement. Before recording, every speaker should sign the agreement.

7) Specification of validation and distribution

Corpus validation criterion is the final validation after the pre-validation and the finishing of the whole corpus production. It can check the quality of corpus and provide the reference criterion to users.[3]

Corpus distribution can be made through a distribution organization or the corpus production affiliation itself. The producer should provide the information about corpus to distributor and users. And legal agreement between producer, distributor and user should be signed before formal distribution.

Discussion and conclusion

RASC863 is an available attempt to standardization research of speech corpus. Besides RASC 863, supported by many national and international funds, such as the 863 Hi-tech Project, 973 Development Program of China, the National Science Foundation of China and the National Science Foundation of America, phonetic laboratory of Institute of Linguistics, Chinese Academy of Social Sciences, has established a lot of high quality speech corpus in recent years（more information please see http：//www.chineseldc.org）. In this progress, we realized the importance of standardization research for prompting the share of resources and the improvement of phonetics.

Currently, we are focusing on multi-model speech corpus collection and speech act annotion orienting to man-machine interactive mode especially from speech perspectives. The specifications of corpus should be extended to more higher levels.

In the future, we hope more and more affiliations could participate in this work. Only in this way, can the speech corpora be established more efficiently and be used or shared more easily.

References

[1] Yin Zhigang，“The introduction of speech corpus research and establishment”，The newspaper of CASS, 2006

[2] Li Aijun , Zhigang Yin, Tianqing Wang, Qiang Fang, Fang Hu (2004), RASC863 - A Chinese Speech Corpus with Four Regional Accents, ICSLT-o-COCOSDA, New Delhi, India.

[3]Aijun LI, Yiqing Zu,(2006) Corpus Design and Annotation for Speech Synthesis and Recognition, as a chapter in Advances in Chinese Spoken Language Processing, edited by Chin-Hui Lee, Haizhou Li, Lin-shan Lee, Ren-Hua Wang, Qiang Huo, World Scientific Publishing Co.(in progress)

[4] Florian Schiel and Christoph Draxler, Production and validation of speech corpora, Bastard Verlag Munchen, Erstausgabe, (2003).