STANDARDIZATION OF SPEECH CORPUS

Ai-jun Li [*] ***and Zhi-gang Yin***

Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China, 100732
Email: [*] liaj@cass.org.cn, *yinzhg@cass.org.cn*

ABSTRACT

Speech corpus is the basis for analyzing the characteristics of speech signals and developing speech synthesis and recognition systems. In China, almost all speech research and development affiliations are developing their own speech corpora. We have so many different kinds numbers of Chinese speech corpora that it is important to be able to conveniently share these speech corpora to avoid wasting time and money and to make 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) will be exemplified to illuminate the standardization of speech corpus production.

Keywords: Phonetics, Speech corpus, Data standardization, Interoperability

1 INTRODUCTION

The speech corpus, the collection of speech signals and its annotation, metadata, and documents, is the basis for both analyzing the characteristics of speech signals 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. 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 a 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, is speech corpus very important. For phonetic research, speech corpus can provide diverse and accurate data to help researchers find the rules of languages. For ASR, in order to "train" the system to "understand" any of the speakers' voices, a speech corpus with a great capacity is necessary. Taking advantage of the statistical data of a speech corpus, the ASR system can transform speech signals into text strings by using phonological, linguistic, and stochastic analysis. That is why ASR can "understand" human's voice.

Because of the importance of speech corpora in China, corpora production has received long term support from 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 (Yin, 2006).

With the development of speech corpus technology, a new problem has appeared: on the one hand, many corpora have been established, and much money and time have been put into their technology; on the other hand, these corpora are difficult to share among different affiliations. The main reason for this problem is the lack of general specifications for corpus collection, annotation, and distribution. In order to solve this problem,

standardization research on speech corpus is necessary and specifications should be stipulated.

2 STANDARDIZATION RESEARCH OF SPEECH CORPUS

Standardization of speech corpus includes many aspects as described below.

2.1 Legal considerations

Speech corpora and their production must abide by the laws of the nation. These legal documents should be prepared: a property rights statement of the corpora, agreement with the speakers, agreement with the users, etc.

2.2 Standardization of speech corpus collection procedures

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

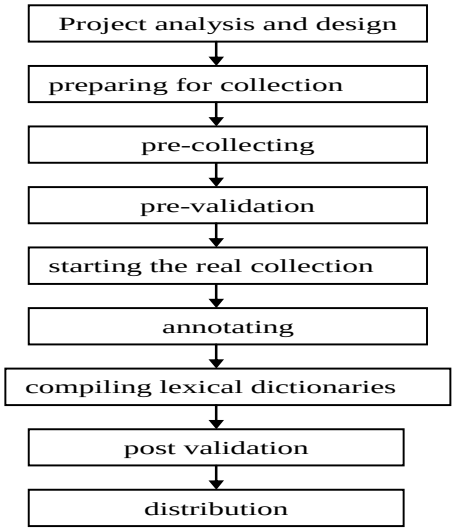

Fig 1： flow chart of the production of corpus

Figure 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, and some can be skipped in a specific task; in fact, an additional step can be introduced by the producer (Li, et al., 2006).

1) Project analysis and design: analyze the speech corpus project and draft its blue print. The specifications of the corpus will be set: corpus size, quantity of speakers, speech style, recording equipment, etc.
2) Preparing for collection: prepare for the corpus according to the blue print: design the input prompts, prepare hardware and software, raise money and organize staff, find speakers, etc.
3) Pre-collecting: if the speech corpus is very large and complicated, pre-collecting a few samples is absolutely necessary. It can find problems and improve the plan, thus avoiding possible mistakes in the 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 and examining whether or not it has reached the criteria. This is employed to accept or reject the corpus.
9) Distribution: distributing the speech corpus which passes the post validation

2.3 Standardization of the speech corpus

Not only should the procedure for producing a corpus, but also the corpus itself should be standardized. Table 1 shows the major specifications in producing a speech corpus (LI & Zu, 2006).

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 detailed information or script (prompt) organization of reading 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 for recording equipment, environmental conditions, recording platform and 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 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 section, RASC863, a Regional Accented Speech Corpus funded by the National 863 Project, will be used to illustrate the above-mentioned standardization.

There are 10 dialect families in China: 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 natural that Putonghua (Standard Chinese, hereafter SC), which is phonetically based on Beijing Mandarin, has been chosen as the communicative spoken language among people from different dialectal regions. However, people with different dialectal backgrounds typically speak SC with a certain degree of accent because of the influence of their mother tongue dialect. This kind of influence can be phonetic, lexical, and/or syntactical.

In the recent years, with the development of the ASR techniques, collecting accented spontaneous speech corpora has become 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: Chongqing, Shanghai, Guangzhou, Xiamen, Taiyuan, Changsha, NanChang, Wenzhou, Luoyang, and

Nanjing. However, only the data for first four regions has been distributed by ChinesLDC (http://www.ChineseLdc.org). Therefore, the following introduction will focus on only these four regions.

The corpus consists of spontaneous speech, read speech, and selected dialectical words. For spontaneous speech, each speaker was asked to select a topic or use one from our prepared topic sheet with a variety of 160 topics and then give a 4-5 minute spontaneous speech on this topic. Also, each speaker was asked to answer 15 elected spontaneous questions. The read speech consisted of 2200 phonetically balanced sentences and 460 frequently used sentences in daily life domain. For each dialectal region, we prepared those words or phrases frequently used in daily life and that 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 age, gender, and educational background) were recruited for the project.

The detail specifications of RASC863 are as follows:

1) Specification of speakers

Specification of speakers describes the number of speakers to be recorded for each dialectal accent and their characterizations. Sometimes it describes the speaking styles. Speaker characterization concerns the distribution of age, education level, gender, and the dialectal coverage aspired to. 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 the speech can be described in different ways according to task, topic, or simply in text description.

Table 2 illustrates the distribution of speakers' ages, genders, accent degrees, and educational backgrounds for each region. In accent category, L1-L3 stand for the three major accent degree levels from better to worse, and A and B stand for two sub-levels.

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. Speakers' distribution for each region

2) Specification of corpus design

The aim of speech corpus design is to determine what is to be recorded and to get the necessary script. Whether a corpus needs a designated script before collection is determined by the corpus type and corpus content (LI, et al., 2004). 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

Table 3. Prompt sheet for each speaker

3) Specification of recording

Usually the specification of recording contains a 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. 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 a YYSRecorder was used in recording other speeches. All speech data were sampled in 16KHz and 16 bit. A metadata file, as exemplified in Table 4 for instance, was generated for each sound file. Positions of the microphone and speaker are shown in Figure 2.

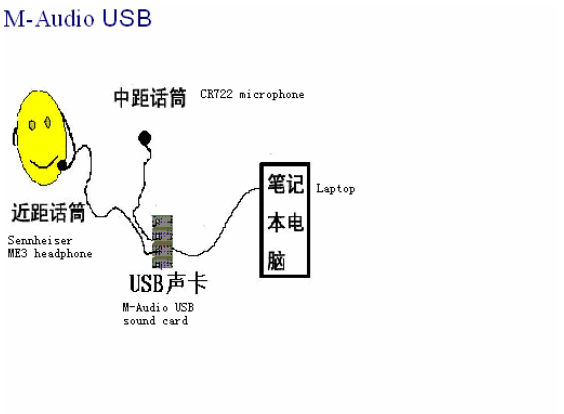

Figure 2. Recording system mounting diagram

4) Specification of corpus structure

Corpus structure relates 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, the normal data structure is shown in Table 5. File naming should be done based on file content. Care was taken to be sure that the length of the name follows the ISO-9960 file attributes; otherwise, it will be difficult to record the final corpus data on a CD or DVD (Li & Zu, 2006). In RASC863, each recorded sound corresponds to a metadata file and a wave file that are automatically generated by recording software. The metadata file describes the detailed information related to this recorded sound file as shown in Table 4.

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. Metadata for each sound file

//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 5. A typical corpus structure

5) Annotation Specification

Speech corpus annotation includes speech-to-characters transcription, segmental annotation, and prosodic annotation. Specification of annotation describes the annotation format, rules, tools, and consistency criteria. Sometimes, if there is more than one transcriber transcribing or annotating simultaneously, their annotation consistency should be checked first. In the RASC863 project, Chinese character transcription as well as paralinguistic and non-linguistic labeling have 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 phonetic annotation. C-ToBI3.0 and SAMPA-C annotation systems were used in prosodic annotation and segmental annotation (Li & Zu, 2006).

6) Legal agreement

The agreement between producer and speaker, often called the speaker agreement, in which the usage of the recorded speech data and even some of the speaker's information, is very important. 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) Validation and distribution specification

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 (Li & Zu, 2006).
Corpus distribution can be made through a distribution organization or the corpus production affiliation itself. The producer should provide the information about the corpus to the distributor and users. Finally, the legal agreement between producer, distributor, and user should be signed before formal distribution.

4 DISCUSSION AND CONCLUSION

RASC863 is an attempt to standardize speech corpus research. RASC 863, supported by funding from many national and international organizations, such as the 863 Hi-tech Project, 973 Development Program of China, the National Science Foundation of China, the National Science Foundation (USA), and the phonetic laboratory of the Institute of Linguistics, Chinese Academy of Social Sciences, has constructed a great deal of high quality speech corpus in recent years (for more information, see http://chineseldc.org). In this progress, we realized the importance of standardized research for promoting the sharing of resources and the improvement of phonetics.
Currently, we are focusing on a multi-model speech corpus collection and speech-act-annotation oriented towards a man-machine interactive mode especially from speech perspectives. The specifications of the corpus should be extended to higher levels.

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

5 REFERENCES

Li, A., Yin, Z., Wang, T., Fang, Q., & Hu, F. (2004) *RASC863 - A Chinese Speech Corpus with Four Regional Accents.* ICSLT-o-COCOSDA: New Delhi, India.

Li, A. & Zu, Y. (2006) Corpus Design and Annotation for Speech Synthesis and Recognition. In *Advances in Chinese Spoken Language Processing.* Lee, C., Li, H., Lee, L., Wang, R., & Huo, Q. (eds.) World Scientific Publishing Co. (in press)

Schiel, F. & Draxler, C. (2003) Production and validation of speech corpora. Bastard Verlag: Munchen, Erstausgabe.

Yin, Z. (2006) The introduction of speech corpus research and establishment. *The newspaper of CASS.*