STANDARDIZATION OF SPEECH CORPUS

Li Ai-jun, Yin Zhi-gang

Phonetics Laboratory,

Institute of Linguistics,

Chinese Academy of Social Sciences

1. Introduction

- Speech corpus (database) is the collection of speech signal, its annotation and documents.
- Speech corpus is the basis for both phonetic research and developing speech synthesis and recognition systems.

- Speech synthesis---TTS (Text to Speech)

  example: “Welcome to 20th International CODATA Conference”

  standard chinese (mandarin) SiChuan dialect Chind’s voice

- Speech recognition---ASR (Automatic Speech Recognition)

  IBM VIAVOICE, MICROSOFT OFFICE XP……

- phonetic research

Importance of standardization research

- In China, many speech research and development affiliations are developing their own speech corpora.

863, 973, the National Science Foundation of China …

it is very important to be able to conveniently share these speech corpora to avoid waste of time and money and to make the research work more efficiency.

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

2. Standardization research of speech corpus

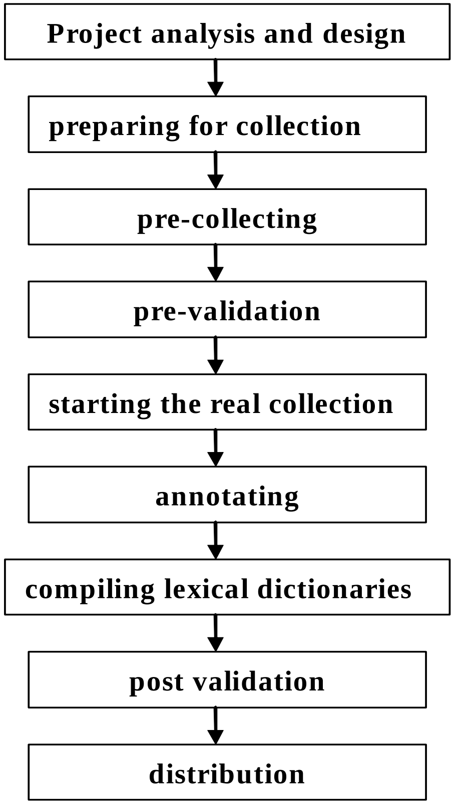
- 1). Legal correlated
- 2) Standardization of collection procedure of speech corpus
- 3). Standardization of speech corpus

1). Legal correlated

- Legal documents of speech corpus:

  property right statement of the corpora (database), agreement with the speakers, agreement with the users, …

2) Standardization of collection procedure of speech corpus



Project analysis and design

preparing for collection

pre-collecting

pre-validation

starting the real collection

annotating

compiling lexical dictionaries

post validation

distribution

Fig 1： the collection procedure of speech corpus

3). Standardization of speech corpus

- *Specification of speakers:* Describing the speaker’s features;
- *specification of corpus design:* Describing the corpus organization and contents;
- *specification of recording:* Describing the recording technical specifications and the recording platform ;
- *specification of annotation:* Describing the annotation conventions;
- *specification of validation:* Setting explicit the criteria that the corpus should fulfill. Giving an overview of the features to be checked;
- *specification of distribution:* Describing the distribution plan, principles and the storage medium.

3. Detailed Specifications exemplified by RASC863

- a speech corpus example --- RASC863
(Regional Accented Speech Corpus funded by National 863 Project)

- RASC863 is a speech corpus with four regional accents, namely Chongqing, Shanghai, Guangzhou and Xiamen.
- 800 speakers (200 * 4)
- 70GB

RASC863 - **1. Specification of speakers**

- Specification of speakers describes the number of speakers to be recorded for each language and their characterizations.

  age, education level, gender , dialectal coverage

- Sometimes it has to describe the speaking styles.

  read speech, answering speech, command/control speech, descriptive speech, non-prompted speech, spontaneous speech, neutral vs. emotional speech and dialogue.

RASC863-the distribution of speakers

Items	Levels	Male	female
Age/gender	16-30 (y) 31-45 (y) Older than 50 (y)	45 45 10	45 45 10
Education	Junior high school Senior high school Undergraduate/ graduated	5 15 80	5 15 80
Accent category	L1 L2 L3	5 70 25	5 70 25

RASC863 – **2. Specification of corpus design**

- The aim of speech corpus design is to determine what to be recorded and to get the necessary script.
- The RASC863 prompt sheet for each speaker:

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)

RASC863-**3.Specification of recording**

- Usually the specification of recording contains recording guide, technical parameters, recording procedures, recording log files, etc.

- Hardware: notebook,

usb sound card (M-Audio usbpre),

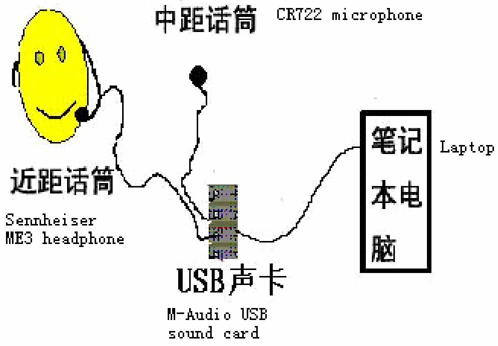
Microphone (sennheiser earphone,

CR 722 capacitor microphone)

Software: Cooledit Pro 2.0

YYSRecorder

M-Audio USB



RASC863-**4.Specification of corpus structure**

- Corpus structure related to the corpus internal organization structure, the file naming rules and the storage media for distribution

- In RASC863, each recorded sound corresponds to a metadata file and a wave file.
- The metadata file describes the detailed information related to this recorded sound file
- 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

……

RASC863-5.Specification of annotation

- Specification of annotation describes the annotation format, rules, tools, consistency criterion.
- Speech corpus annotation includes speech to characters transcription, segmental annotation and prosodic annotation.
- if there are more than one transcribers transcribing or annotating simultaneously, their annotation consistency should be checked first.
- Rasc863: C-ToBI3.0, SAMPA-C

RASC863-**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.

RASC863-7.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.

- ***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

- Some speech corpus of Phonetics Laboratory (www.ChineseLDC.org)

supported by 863, 973, the National Science Foundation of China and the National Science Foundation of America

Ongoing research: man-machine conversation based on spontaneous speech ……

the specifications of corpus will be extended and perfected

THANK YOU!