Development of a Chinese speech corpus covering inter-dialect phonological differences *

◎ Xuebin Ma†, Akira Nemoto‡, Nobuaki Minematsu†, Feng Shi‡(†Univ. of Tokyo,‡Nankai Univ.)

1 Introduction

Nowadays, speech corpora are always regarded as the most important infrastructure of modern spoken language technologies and many corpora have been built by some companies and institutions. In China, there are also many Mandarin corpora developed, such as the corpora built under the 863 project, some in-house corpora modified from other ones, etc. However, only several dialectal or accented Mandarin corpora are available to public which cover individual dialects. Besides, these corpora are developed under different conditions and can hardly be combined and used in study of different dialects.

In this paper, the development of a Chinese speech corpus is presented, which is built intended to cover the inter-dialect phonological differences and be used in dialect-based speaker classifications. In the following section, the corpus design is described. In section 3, the data collection is described. In section 4, the process of annotation and verification of the corpus is addressed. At last, a conclusion about this corpus is given.

2 Corpus Design

Chinese dialects are mainly grouped into 7 large dialect regions (GuanHua, Wu, Xiang, Gan, Kejia, Yue, Min) traditionally [2] and each region also has many sub-dialects. For example, Guanhua (Mandarin) has 7 sub-dialects. All the dialects and sub-dialects are different in varying degrees, grammatically, lexically, phonologically and phonetically, so people from different dialect regions always have some difficulties in oral communication. Since 1956, standard Mandarin has been popularized officially and almost every dialect speaker began to learn Mandarin. However, affected by their native dialects, many of them speak Mandarin with regional accents.

Although there are so many differences among the dialects, they still share more or less common features: same written characters, similar sound systems, same phonological features, etc. For example, every character is pronounced as mono-syllable with the same phonological structure constructed by a tone, an initial and a final. In fact, all these dialects are developed from the same root, Middle Chinese, which refers to the Chinese spoken language during the period from 6th to 10th century. During this period of time, the system of historical Chinese phonology was matured and also described by some rhyme books, such as Qieyuan (切韻) and Guangyun (広韻). According to these books, some clues are found that phonology of Chinese dialects are all developed from the historical Chinese phonology, and the complex relationships among these dialects can also be well explained by comparing their phonological features. Nowadays, for this purpose, some handbooks are written by linguists comparing different dialectal pronunciations and some widely used written characters are also listed, which are used to check the dialectal features of different phonological units, initials, finals or tones, separately.

Table 1 Examples of the reading materials

characters	爬, 辣, 架, 花, 刮, 河, 色
words	爬行, 辣妹, 架子, 桂花, 緑色
sentences	五星紅旗是中華人民共和国的国旗

As the object of this corpus is automatic classification of speakers based on their dialects, sub-dialects of Mandarin or accented Mandarin, this corpus is designed to be composed by three parts: Chinese dialects, sub-dialects of Mandarin and accented Mandarin. Then, trying to cover the inter-dialect phonological differences, a list of 116 written characters in [3], which is used to check the dialectal pronunciations of finals, is adopted as the reading materials for the syllable part of this corpus. Based on these characters, 80 two-syllable words are used as the reading materials for the second part. At last, 5 sentences trying to cover as many final units as possible and 5 sentences composed only by voiced phonemes are adopted as the reading materials for the sentence part. Table 1 is an example of these materials.

* 方言的音韻特徴を考慮した中国語の音声コーパスの開発。馬学彬 †, 根本晃 ‡, 峯松信明 †, 石鋒 ‡ († 東京大学, ‡ 中国南開大学)。

3 Data Collection

Two kinds of subjects participated in the recording. The first kind of subjects are 19 native dialect speakers who are from Yue, Xiang, Min, Gan and Hakka dialect regions. They were selected after their language backgrounds were checked to ensure that they were brought up in the same dialect regions and their parents are the same native dialect speakers. All of them speak Mandarin with accents of varying degrees affected by their native dialects. During reading, they were asked to read the materials in their native dialects and accented Mandarin, separately. The second kind of subjects are 26 speakers from five sub-dialects regions of Mandarin. 6 of them, 3 boys and 3 girls, are 11 or 12 years old, others are all adults. They were asked to read the materials in their sub-dialects of Mandarin. During the recording, all the materials were read three times.

All recordings were carried out in a quite room with a supervisor, so the data are all expected to be clean. Before the recoding, all the reading materials were checked by the speakers. The recording equipments included a high quality microphone fixed on the table, a linear PCM recorder of Sony company. The data was sampled at 48 kHz.

4 Post-process and Verification

All the data are split automatically into separate files by the pause among them, except sentences are split manually. After that, every file is labeled phonetically and manually by linguistic students using Wavesurfer. After checking the spectrum and raw file, every syllable of accented Mandarin or sub-dialects of Mandarin is labeled into two parts, initial and final, with transcriptions developed from Chinese Pinyin. Fig. 1 is an example of the labeling of a sentence. However, this transcription system cannot be used in the annotation of dialectal utterances because their phonological features are different from Mandarin significantly. Then using some handbooks [5], which list the real dialectal pronunciation of many common used written characters with IPA symbols, a new transcription system for every dialect is built and applied in the annotation of dialectal utterances. An experienced phonetician, the second author, labels these dialectal utterances with the new transcription systems after checking their

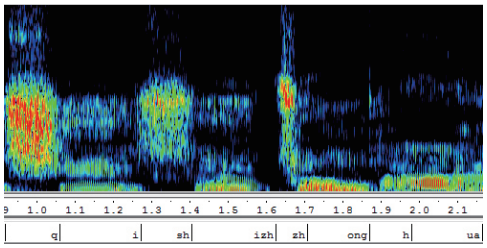

Fig. 1 An example of the annotation

dialectal pronunciation by referring to these handbooks.

The verifications are carried out in three stages. At the first stage of data splitting, the automatically split files are checked to ensure each utterance is properly stored as a separate file. Meanwhile, noise files are all discarded. Stage two of verification is performed during the annotation carried out by the linguistic students, who checked every file by looking at its spectrum. At last, for the data of every dialect, a native speaker is asked to verify the annotations to resolve all the marked problems and correct any mistakes. This stage can also guarantee the high degree of accuracy in the annotation.

5 Conclusions

In this paper, the entire development process of a Chinese speech corpus covering inter-dialect phonological differences is described. This corpus can be divided into three parts, Chinese dialects, sub-dialects of Mandarin and accented Mandarin. Every part covers rich phonetic contents, including syllables, words and sentences, which are adopted to show the inter-dialect phonological differences among Chinese dialects. Most participants of the recording are young students at the age of about 20, except 6 children. All the recoding was done in a supervised environment and annotation was carried out by students from a linguistic laboratory. Verifications were also performed to guarantee the correctness and accuracy of the data and annotation.

References

[1] http://www.glossika.com/en/dict/

[2] Yuan Jiahua et al, HanYu FangYan GaiYao, Language & Culture Press, 2000

[3] Institute of Linguistics of Chinese Academy of Social Sciences, Hanyu DiaoCha ZiBiao, The Commercial Press, 2007

[4] Richard VanNess Simmons et al, Handbook for Lexicon Based Dialect Fieldwork, Zhonghua Book Company, 2006

[5] Linguistic Lab. of Beijing Univ, HanYu FangYan ZiHui, Language & Culture Press, 2003