Multi-accented Mandarin Database Construction and Benchmark Evaluations

Xiang Yan, Lei He, Pei Ding, Rui Zhao, and Jie Hao

Toshiba (China) Research and Development Center
5/F Tower W2, Oriental Plaza, No.1 East Chang An Avenue,
Dong Cheng District, Beijing 100738
{yanxiang, helei, dingpei, zhaorui, haojie}@rdc.toshiba.com.cn

Abstract. In this paper, we describe the designing, recording and checking procedures of a multi-accented Mandarin speech database, and present benchmark evaluation of this database. The database was recorded in 6 cities in China, containing 1200 speakers’ accented Mandarin speech of continuous digits, isolated words and sentences. In total, 520k utterances (572.5 hours) were collected. We perform the intra-accent and cross-accent evaluations, together with the evaluation of a multi-accented acoustic model trained from the whole database. The database is a phonetically rich, gender-balanced and accent-balanced database, which could serve as the basic material for accented Mandarin recognition research, and it could also be used for creating real automatic speech recognition products for users with different accents.

Keywords: Mandarin, speech database, multi-accented

1 Introduction

Chinese is a language with many dialects. Though the standard Mandarin is regarded as the only standard spoken language in China, people in different areas of China often speak their own dialects in their daily lives. The dialects affect the speakers’ pronunciation and speaking style even when they are speaking standard Mandarin. As a result, the speech recognizer trained on standard Mandarin speech always mismatches the accented speech and the performance drops dramatically.

Many methods have been proposed to deal with this problem. One simple way is to collect dialectally accented speech and train the recognizer based on it. Another trend is to adapt the recognizer that is originally trained from unaccented speech. The adapting techniques include acoustic model adaptation like MLLR and MAP, as well as the lexicon adaptation, in which the standard Mandarin pronunciation lexicon is adapted to accented Mandarin pronunciation [1]. For all the attempts mentioned above, an accented Mandarin speech database serves as the basis for accented Mandarin speech analyzing, training, adapting and/or testing.

From this point of view, we regard the construction of a phonetically abundant, gender balanced and accent balanced multi-accented Mandarin speech database as the very first step towards solving accent problem in Mandarin ASR. This database can
serve as the basis of accent Mandarin recognition research, and it can also be used for creating automatic speech recognition products on accent-independent and speaker-independent levels.

In the past three years, we constructed a multi-accented Mandarin speech database. The database was collected in six cities in China; each of which represents one accent. In total, 1200 speakers (600 male and 600 female) were recruited, and 102.8 hours continuous digits, 87.2 hours isolated words and 382.5 hours sentences were collected. The database is balanced in accent; that is to say, the speech data for each accent is of the comparable size.

This paper is arranged as follows. Section 2 explains the recording cities' selection criteria. The prompt sheets design is described in Section 3. Section 4 presents the database recording procedure, together with the database information. Section 5 shows the data check procedure. In Section 6, intra-accent and cross-accent benchmark evaluations are presented; and a multi-accented acoustic model is trained and evaluated.

2 Recording cities selection

Spoken Chinese comprises many dialects that can be classified in seven main dialect regions. They are North (Mandarin), Wu, Yue, Min, Xiang, Gan and Kejia [2] (See Fig. 1). Table 1 shows the percentage of population and the representative cities of these dialect regions.

When choosing recording cities, the following criteria were adopted:
1) Chosen cities should cover various dialects
2) Dialects with the larger percentage of population should have more representative cities to be chosen
3) Cities should have relative high level of economic development

According to the above criteria, six cities have been selected. They are: Tianjin (North Dialect), Dalian (North Dialect), Chengdu (North Dialect), Shanghai (Wu Dialect), Guangzhou (Yue Dialect) and Xiamen (Min Dialect). Beijing is not selected since the dialect of Beijing is the base of standard Mandarin, so that we do not consider it as a dialectally accent Mandarin.

3 Text design

In order to collect speech data with as much information about phonetic variations as possible, we designed the prompt sheets to cover more monophones, inter-syllable biphones and triphones.
Table 1. Percentage of population and representative cities of the main dialects of China

<table>
<thead>
<tr>
<th>Dialect</th>
<th>Percentage of Populations</th>
<th>Representative cities</th>
</tr>
</thead>
<tbody>
<tr>
<td>North (Mandarin)</td>
<td>73%</td>
<td>Beijing, Tianjin, Dalian, Chengdu, Jinan, Nanjing</td>
</tr>
<tr>
<td>Wu</td>
<td>7.2%</td>
<td>Shanghai</td>
</tr>
<tr>
<td>Yue</td>
<td>4%</td>
<td>Guangzhou</td>
</tr>
<tr>
<td>Min</td>
<td>5.7%</td>
<td>Xiamen, Fuzhou</td>
</tr>
<tr>
<td>Xiang</td>
<td>3.2%</td>
<td>Changsha</td>
</tr>
<tr>
<td>Gan</td>
<td>3.3%</td>
<td>Nanchang</td>
</tr>
<tr>
<td>Kejia</td>
<td>3.6%</td>
<td>Meixian</td>
</tr>
</tbody>
</table>

3.1 Continuous digits

The continuous digits prompt sheets contain 4-8 digits with average length of 6 digits. They were automatically generated to ensure the uniform distribution of all digits in the prompt sheets. In sum, 3772 and 1000 continuous digits prompt sheets were generated as training set and test set texts respectively.

Since digit “ONE” has two pronunciations in Chinese, namely, “yi1” and “yao1”, we introduced both pronunciations in the prompt sheets.
3.2 Isolated word

For isolated word recording, training set was selected from a large phoneme abundant lexicon. The training set contains the words composed of 2 to 4 characters and they are selected to cover more monophones, inter-syllable biphones and triphones. In sum, 700 isolated words were selected.

The test set contains 500 isolated words with length ranging from 1 to 7 characters. None of the isolated words in the test set has ever been presented in the training set.

3.3 Sentence

As for sentence recording, prompt sheets were generated from a large newspaper text corpus, which contained 10M Chinese characters. In total, 4725 sentences were selected, ranging from 4 to 30 characters long with the average length of 14.7 characters.

The test set contains 375 distinct sentences, with length ranging from 4 to 29 characters. None of the sentences in the test set has ever been presented in the training set.

Compared with the commonly used 863 database (provided by the China 863 high-tech projects), it can be seen that our training set isolated word and sentence prompt sheets have a better speech unit coverage than that of the 863 database. (See Table 2)

<table>
<thead>
<tr>
<th>Unit Coverage (%)</th>
<th>Presented Database</th>
<th>863 Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>monophone</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>inter-syllable biphone</td>
<td>96.00</td>
<td>93.94</td>
</tr>
<tr>
<td>triphone</td>
<td>84.81</td>
<td>48.08</td>
</tr>
</tbody>
</table>

3.4 Pinyin annotation

Pinyin annotations were also provided for each prompt sheet as the reference of the standard pronunciation. The pinyin annotations were automatically generated by an annotation tool first, and then manually checked.
4 Database recording

4.1 Speakers

For each city, 200 native speakers (100 male and 100 female) were recruited for data collection. In total, 1200 speakers participated in the recording. Each speaker has lived in the corresponding city since his or her birth or early childhood. The ages of the speakers range from 17 to 33, with the average age of 20-21.

The speakers were randomly divided into two groups: 180 speakers (90 male and 90 female) as training set group and 20 speakers (10 male and 10 female) as test set group. The two groups of speakers read training set and test set prompt sheets respectively.

After recording, the following information was collected from each speaker: 1) name and gender; 2) age and birthplace; 3) recording date.

4.2 Recording

All recordings were performed in a quiet office environment, with a SHURE SM-58 directional microphone, using a SoundBlaster PIC128D sound card. The speech was recorded at 16kHz sampling rate and 16 bits resolution.

Two people monitored the recording as a team. One controlled the prompt display, checked pronunciation, volume and beginning/end silence length, and the other supervised the whole process of recording.

Speakers were asked to read as naturally and fluently as possible. If an utterance was mispronounced, or if the beginning and/or end silence length was less than 500ms, the whole utterance was read again.

4.3 Database information

A total of 520k utterances (572.5 hours) have been collected. The detailed information is shown in Table 3.

<table>
<thead>
<tr>
<th>Table 3. Database information</th>
</tr>
</thead>
<tbody>
<tr>
<td># of utterances</td>
</tr>
<tr>
<td>continuous digits</td>
</tr>
<tr>
<td>isolated-word</td>
</tr>
<tr>
<td>sentence</td>
</tr>
<tr>
<td>size (in hour)</td>
</tr>
</tbody>
</table>
5 Data check

The database check step can guarantee the quality of the database. Two people checked the database separately, and then their check results were merged. We checked the following aspects:

1) Whether the labels that consist of phonetic transcriptions correspond to the speech data.
2) Whether the speech is complete and the volume is suitable.
3) Is there any undesirable noise in the speech?

If any one of the above problems occurred in a speech, the speech file was tagged “unusable”.

6 Benchmark evaluations and analysis

6.1 Experiment setup

Digit acoustic model adopts the general triphone structure. As for isolated word and sentence speech, 101 right-context-dependent Initials and 38 context-independent Finals model structure is adopted. Three state left-to-right hidden Markov model (HMM) with 16 Gaussian mixture components is used. The acoustic features are 39-dimensional Mel frequency cepstral coefficients (MFCC_E_D_A_Z). HTK V3.3 [3] is used for feature extraction, training and testing.

6.2 Intra-accent evaluation

For intra-accent evaluation, each accent test set is decoded using the acoustic model trained on its own accent training set. We regard it as the upper bounds for each accent sub-databases’ performance. The evaluation results are shown in Table 4. The figures in “continuous digits” line mean Sentence Error Rate (SER), and the figures in “isolated word” and “sentence” lines mean Word Error Rate (WER).

<table>
<thead>
<tr>
<th></th>
<th>Chengdu</th>
<th>Dalian</th>
<th>Guangzhou</th>
<th>Shanghai</th>
<th>Tianjin</th>
<th>Xiamen</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>continuous</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>digits</td>
<td>12.40</td>
<td>8.75</td>
<td>12.20</td>
<td>10.50</td>
<td>7.90</td>
<td>12.00</td>
<td>10.63</td>
</tr>
<tr>
<td>isolated</td>
<td>1.00</td>
<td>0.25</td>
<td>0.25</td>
<td>1.00</td>
<td>0.50</td>
<td>0.55</td>
<td>0.59</td>
</tr>
<tr>
<td>word sentence</td>
<td>36.02</td>
<td>28.79</td>
<td>38.84</td>
<td>34.07</td>
<td>36.68</td>
<td>42.66</td>
<td>36.18</td>
</tr>
</tbody>
</table>
6.3 Cross-accent evaluation

For cross-accent evaluation, each accent test set is decoded using the acoustic models trained on other five accents training sets respectively. Then the evaluation results are averaged across the accents. The evaluation results are shown in Table 5. The figures in “continuous digits” line mean average SER and the figures in “isolated word” and “sentence” lines mean average WER.

Table 5. Cross-accent evaluation results (Figures in %)

<table>
<thead>
<tr>
<th>Accent</th>
<th>Chengdu</th>
<th>Dalian</th>
<th>Guangzhou</th>
<th>Shanghai</th>
<th>Tianjin</th>
<th>Xiamen</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isolated</td>
<td>1.57</td>
<td>2.4</td>
<td>1.88</td>
<td>2.07</td>
<td>2.75</td>
<td>2.23</td>
<td>2.15</td>
</tr>
<tr>
<td>Sentence</td>
<td>32.42</td>
<td>45.19</td>
<td>47.02</td>
<td>44.8</td>
<td>45.4</td>
<td>47.8</td>
<td>43.77</td>
</tr>
</tbody>
</table>

Table 6 shows intra-accent and cross-accent evaluation results, with the relative SER or WER increase of cross-accent evaluation against intra-accent evaluation in the rightmost column.

The performance drops dramatically in the cross-accent evaluations due to the accent mismatch, which implies the large difference between Mandarin accents. However, the impact of the accent mismatch differs from task to task. It can be seen from Table 6 that the continuous digits and isolated word recognition tasks are more vulnerable to the accent mismatch.

Table 6. Comparison of intra-accent and cross-accent evaluation results. (Figures in %)

<table>
<thead>
<tr>
<th></th>
<th>Intra-accent</th>
<th>Cross-accent</th>
<th>Relative SER/WER increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous digits</td>
<td>10.63</td>
<td>22.40</td>
<td>111</td>
</tr>
<tr>
<td>Isolated word</td>
<td>0.59</td>
<td>2.15</td>
<td>264</td>
</tr>
<tr>
<td>Sentence</td>
<td>36.18</td>
<td>43.77</td>
<td>21</td>
</tr>
</tbody>
</table>

Besides, we find another interesting phenomenon.

As could be seen in Table 1, dialects of Chengdu, Dalian and Tianjin belong to the same main dialect region, that is, North Dialect. Generally speaking, these dialects are more mutually intelligible than dialects from different main dialect regions, such as dialects of Tianjin and Guangzhou. Based on this fact, we may have the intuition that the accented speech of these three cities should have less difference, and we may expect for a lower relative WER increase of the cross-accent evaluations against the intra-accent evaluations in Chengdu, Dalian and Tianjin accent sub-databases.

However, experiments on Chengdu, Dalian and Tianjin sub-databases show that the relative WER increase are 267% and 26% for isolated word and sentence recognition tasks respectively, which are very close to those of all six accents’ evaluation, namely, 264% and 21%.
So we could conclude that, the greater mutually intelligibility between dialects does not guarantee smaller performance degradation in cross-accented speech recognition. This also demonstrates that our selection of three cities (Chengdu, Dalian and Tianjin) of North dialect to perform recording is not redundancy, but necessity.

6.4 Multi-accented acoustic model

We also trained a multi-accented acoustic model using all the accents’ training sets of the database. Table 7 shows the multi-accented acoustic model’s performance in decoding each accent test set. The figures in “continuous digits” line mean SER and the figures in “isolated word” and “sentence” lines mean WER.

Table 7. Multi-accented acoustic model evaluation results (Figures in %)

<table>
<thead>
<tr>
<th></th>
<th>Chengdu</th>
<th>Dalian</th>
<th>Guangzhou</th>
<th>Shanghai</th>
<th>Tianjin</th>
<th>Xiamen</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous digit</td>
<td>10.05</td>
<td>10.70</td>
<td>14.20</td>
<td>10.55</td>
<td>14.80</td>
<td>15.30</td>
<td>12.60</td>
</tr>
<tr>
<td>Isolated word</td>
<td>1.20</td>
<td>0.60</td>
<td>0.60</td>
<td>1.30</td>
<td>1.10</td>
<td>0.65</td>
<td>0.91</td>
</tr>
<tr>
<td>Sentence</td>
<td>38.93</td>
<td>34.08</td>
<td>42.23</td>
<td>36.88</td>
<td>42.96</td>
<td>45.59</td>
<td>40.11</td>
</tr>
</tbody>
</table>

Table 7 shows intra-accent and multi-accented evaluation results, with the relative SER or WER increase of multi-accented evaluation again intra-accent evaluation in the rightmost column. Compared with intra-accent evaluation, the multi-accented performance degrades. However, the degradation is much smaller than that of the cross-accent evaluation, especially for continuous digits (relative SER increase: 11% vs. 111%) and isolated word (relative WER increase: 54% vs. 264%) recognition tasks.

Thus, training a multi-accented acoustic model using large, accent-balanced speech could, to some extent, mitigate the accent problem in Mandarin speech recognition, especially for continuous digits and isolated word recognition tasks.

Table 8. Comparison of intra-accent and multi-accented evaluations. (Figures in %)

<table>
<thead>
<tr>
<th></th>
<th>Intra-accent SER/WER</th>
<th>Multi-accent SER/WER</th>
<th>Relative SER/WER increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous digits</td>
<td>10.63</td>
<td>12.60</td>
<td>19</td>
</tr>
<tr>
<td>Isolated word</td>
<td>0.59</td>
<td>0.91</td>
<td>54</td>
</tr>
<tr>
<td>Sentence</td>
<td>36.18</td>
<td>40.11</td>
<td>11</td>
</tr>
</tbody>
</table>
7 Conclusion

In this paper, we described the construction of a multi-accented Mandarin speech database: the criteria of selecting recording cities, the features of the prompt sheets, the procedures of the speech recording and data checking.

We also benchmarked the database in terms of intra-accent and cross-accent, and a multi-accented acoustic model was trained by using the whole database and then tested. The experiments showed that:

1) Performance drops in cross-accent evaluations compared with intra-accent, and the performance degradations differ from task to task. Continuous digits and isolated words recognition tasks are more sensitive to accent mismatch.

2) The greater mutually intelligibility in dialects, for example, dialects belong to the same main dialect region, does not guarantee a smaller performance degradation in cross-accent evaluations.

3) Multi-accented acoustic model training can mitigate the accent problem in continuous digits and isolated word recognition tasks.

References