SELL-CORPUS: AN OPEN SOURCE MULTIPLE ACCENTED CHINESE-ENGLISH SPEECH CORPUS FOR L2 ENGLISH LEARNING ASSESSMENT

Yu Chen, Jun Hu and Xinyu Zhang*
Shanghai Key Laboratory of Trustworthy Computing, Shanghai, China
School of Computer Science & Software Engineering, East China Normal University, China

Introduction

English is spoken and learned by 1.75 billion people worldwide. For non-native English speakers, their mother tongues have a significant influence on their second language (L2) pronunciation. Most of L2 learners have severe problems with their pronunciations. Some pronunciations are obviously incorrect and cannot be understood at all. Some pronunciations exhibit slight/strong accents and may be understood with various degrees of tolerance.

The situation in China is even more severe since there is a shortage of qualified English teachers and students' mispronunciation cannot be immediately pointed out and corrected.

In order to tolerate accents, non-native accented English speech corpus to cover all major dialectal regions in China. We selected 389 qualified speakers and classified them into seven major regional dialects according to the place where they lived and learned English. The population distribution for these dialectal regions is shown in Figure 1. We have released our speech corpus to the public for academic research, which is available for downloadable http://www.mosrecognition.org/self-corpus. To the best of our knowledge, it is the first open-source English speech corpus that covers all major regional dialects.

Corpus Statistics

We design a mobile APP, as shown in Figure 2 to efficiently record and collect speech data. The monophone recordings are sampled at 16kHz. We use self-reported questionnaires to collect the volunteer speakers' information, including gender, hometown city and dialect.

Figure 3: Geographical distribution of major Chinese dialects.

Computer-assisted language learning systems using automatic speech recognition (ASR) systems provide L2 learners an effective means to improve their speaking skills without the presence of human teachers. However, for ASR, the performance is often significantly reduced when a speaker's accent is different from that in the training set. Therefore, ASR systems trained with the speech corpus from native English speakers are generally not well suitable for L2 speakers.

Our corpus is contributed by 389 volunteer speakers, and it consists of 31.6 hour recordings in total, including 16.7 hours by male volunteers and 14.9 hours by female volunteers. Our corpus consists of a training set, a development set and a test set. The training set has 10,519 speech recordings contributed by 347 volunteer speakers. The development set has 873 speech recordings by 21 speakers. The test set has 709 speech recordings by 21 speakers. Table 1 list the statistics of our corpus.

Table 1: Statistics on speakers' gender, utterances and recording hours in our corpus.

<table>
<thead>
<tr>
<th>Gender</th>
<th># of female</th>
<th># of male</th>
<th>Total</th>
<th>Duration (hrs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>female</td>
<td>86</td>
<td>87</td>
<td>173</td>
<td>16.7</td>
</tr>
<tr>
<td>male</td>
<td>193</td>
<td>192</td>
<td>385</td>
<td>10.0</td>
</tr>
</tbody>
</table>

Our corpus is expected to not only help construct ASR system for future nationwide open English tests, but also can be used for academic research like multiple accentuated acoustic model and pronunciation assessment.

Conclusions

We present a multiple accented speech corpus for English learning in China. We trained a few baseline models to understand the benefits of our corpus. We have released our speech corpus to the public and it is the first open-source English speech corpus that covers all major regional Chinese dialects. Our corpus is expected to not only help construct ASR system for future nationwide open English tests, but also can be used for academic research like multiple accented acoustic model and pronunciation assessment.

References

