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 degree 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.

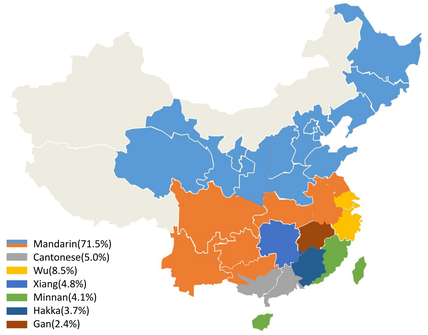

Figure 1: Geographical distribution of major Chinese dialects.

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.

In order to tolerate accents, non-native accented English speech data with orthographic and plausible annotations for mispronunciation is the key. In this paper, we aim to design and build a Chinese-English speech corpus to cover all major regional dialects 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 downloadat http://www.roseducation.org/sell-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 speechdata. The mono channel recordings are sampled at 16kHz. We use self-reported questionnaires to collect the volunteer speakers' information, including gender, hometown cityand dialect.

Figure 2: Mobile APP used to record and collect speech data

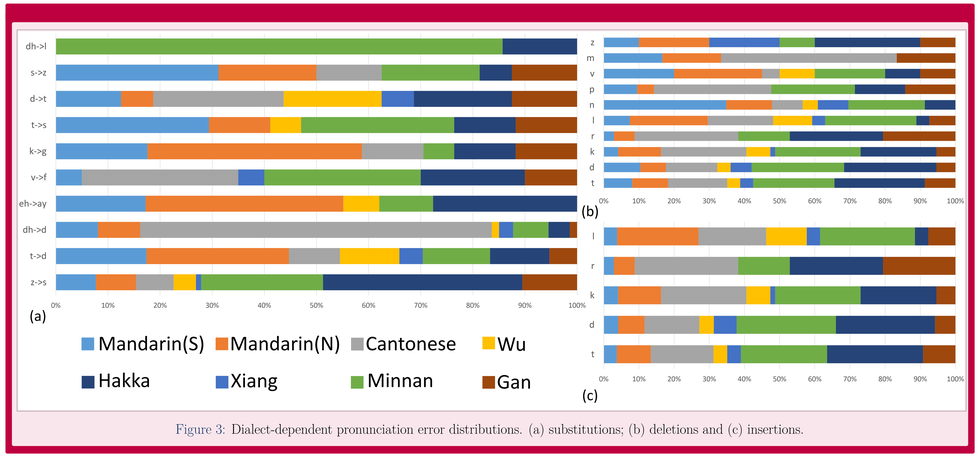

Figure 3: Dialect-dependent pronunciation error distributions. (a) substitutions; (b) deletions and (c) insertions.

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 795 speech recordings by 21 speakers.Table 1 list the statistics of our corpus.

Dialects	Mandarin	Cantonese	Wu	Xiang	Minnan	Hakka	Gan
# of speakers	185	31	108	13	24	10	18
# of male	98	9	39	6	19	10	9
# of female	87	22	69	7	5	0	9
# of utterances	5830	689	3714	398	613	300	643
duration(hrs)	14.8	1.7	9.6	1.0	1.7	0.9	1.9

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

We manually annotated 1600 utterances from the seven major dialectal regions, after performing 230 phoneme insertions, 2018 phoneme substitutions and 2158 phoneme deletions. Figure 3 shows the phoneme error statistics for each dialectal region, while taking the most frequent errors. We observe that each regional dialect exhibits different error frequency. For example, in Minnan dialect, the frequency of phoneme substitution [dh]->[l] appears significantly higher than other regional dialects.

Experiments

We present our baseline experiments of ASR system on SELL-CORPUS using Kaldi toolkit. A statistics of the resulting triphone delta-delta GMM-HMM model (tri1), LADMLLT GMM-HMM model (tri2), SAT(fMLLR) GMM-HMM model(tri3) and TDNN is given in Table 2.

stage & trained models	dev-set	test-set	L2-ARCTIC (BWX,LXC)
1. monophone	62.76	-	-
2. tri1	30.19	31.61	34.13(70.65)
3. tri2(LDA+MLLT)	27.42	27.24	38.46(71.13)
4. tri3(LDA+MLLT+SAT)	17.09	17.76	25.80(67.36)
5. Chain-TDNN	10.00	11.51	19.59(57.94)

Table 2: WER(%) of our ASR system based on SELL-Corpus test data and based on the subsets BWC, LXC in L2-ARCTIC.The results given in parentheses in the fourth column are the WER of the ASR system based on the native English corpus LibriSpeech for speech input with Chinese accents.

Conclusions

We presented 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 oral English tests, but also can be used for academic research like multiple accented acoustic model and pronunciation assessment.

References

[1] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[2] Guanlong Zhao, Sinem Sonsaat, Alif O Silpachai, Ivana Lucic, Evgeny Chukharev-Khudilaynen, John Levis, and Ricardo Gutierrez-Osuna. L2-ARCTIC: A non-native English speech corpus. Technical report, Perception Sensing Instrumentation Lab, 2018.

Contact Information

- Download: http://www.roseducation.org/sell-corpus
- Email: xyzhang@sei.ecnu.edu.cn