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

ABSTRACT

We present SELL-CORPUS, a multiple accented speech corpus for L2 English learning in China, aiming at the potential research of multiple accented acoustic model, mispronunciation detection and pronunciation assessment for future nationwide oral English tests. Our corpus contains 31.6 hour speech recordings contributed by 389 volunteer speakers, including 186 males and 203 females. Our corpus covers seven major regional dialects and provides a baseline for Chinese multiple accented automatic speech recognition system. We released our speech corpus to the public for academic research. To the best of our knowledge, it is the first open-source English speech corpus that accounts for the accents of all major Chinese regional dialects.

Index Terms— English speech corpus, Chinese dialects, Automatic speech recognition, Second language learning, English pronunciation assessment

1. INTRODUCTION

English is spoken and learned by 1.75 billion people worldwide, including 20% native speakers and 80% non-native speakers [1]. For non-native English speakers, their mother tongues have a significant influence on their second language (L2) pronunciation [2]. Because phonetic and syllabic structures of non-English language vary from or even have dramatic difference from English language, L2 speakers need repetitive practice in order to pronounce English words correctly. 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.

Computer-assisted language learning systems provide L2 learners an effective means to improve their speaking skills without the presence of human teachers [3]. Many automatic mispronunciation detection and pronunciation assessment tools were developed [4, 5]. However, for automatic speech recognition (ASR) systems, the performance is often significantly reduced when a speaker's accent is different from that in the training set [6]. Therefore, ASR systems trained with the speech corpus from native English speakers are generally not well suitable for L2 speakers.

Therefore, a few accented English speech corpora and techniques were proposed, by accounting for language accents. The ISLE speech corpus contains German and Italian English learners speech data [7]. In Asia, the NICT JLE Corpus was designed for Japanese English learners [8]. The work in [9] discussed the influence of Indian language (Gujarati and Tamil) on Indian English learners. A few Chinese English corpora were built during the past ten years, such as SWECCL [10] and COLSEC [11]. However, those corpora did not consider the influence of dialects. SHEFCE is a Cantonese-English bilingual speech corpus [12] for L2 English speakers in Hong Kong. A corpus covering a few dialects in China was reported in ESCCL [13], but it is not publicly available. Accounting for language accents, classic methods train an accent-specific pronunciation dictionary or train accent-dependent models on multi-accented speech data. Recently, deep learning demonstrated the success in ASR and many research reported promising performance [14, 15]. However, those solutions typically require a large amount of L2 speech data.

In order to tolerate accents, especially multiple 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. Our target users are 0.4 billion English L2 learners in China [16], who speak Chinese with/without dialectal accents. We recruited more than 500 online volunteers and selected 389 qualified speakers, including 186 males and 203 females. We classified them into seven major regional dialects according to the place where they lived and learned English. Each volunteer speaker is asked to read and record phonetically balanced sentences retrieved from free online publication. Each speech recording has its

This work was supported by the NSFC (No.61631166002, No.61572196). Xinyu Zhang is the corresponding author. Email: xyzhang@sei.ecnu.edu.cn

word-level transcription that can be used in multi-accented acoustic model training. Our English speech corpus provides a baseline for a Chinese multiple accented ASR system. We test our corpus using a few training methods and observe significant performance improvement against the existing corpus. We have released our speech corpus to the public for academic research, which is available for download at `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. This corpus will help promote English learning and teaching, and even multiple dialect accented research using ASR techniques.

2. CORPUS CONSTRUCTION

Our speech corpus aims at providing a baseline of English ASR for nationwide oral English learning and test. Therefore, we need a large number of utterances covering a wide range of regional dialects.

Reading Material Selection: The recording materials are retrieved from Project Gutenberg [17], a collection of free digital eBooks. To avoid the speaking fatigue while reading a very long utterance, we maintain the length of the utterance within 40 to 130 characters. Every 50 utterances are grouped together as a set. A few random sets of utterances are assigned to volunteer speakers, which avoids the utterance overlapping of recording materials between speakers. All volunteers are non-native English speakers and have distinct accents inherited from their Chinese dialects. In addition, their oral English skills show great variation. We use a dictionary containing about 4,000 common vocabulary words to filter out the utterances if their words are not included in the dictionary. Eventually, the recording materials in our corpus contain 11,000 utterances in total. These utterances are phonetically balanced, with 100%, 88.09% and 31.21% coverage for phonemes, diphones and triphones, respectively. This guarantees the performance of our corpus.

Volunteer Speaker Distribution: 389 qualified volunteer speakers range from 18 to 30 year old. Their mother tongues cover all seven regional dialects: Mandarin (north and southwest regions), Wu language, Cantonese, Gan dialect, Minnan dialect, Xiang dialect and Hakka. Considering that Mandarin accents vary widely across the north and southwest of China, we further briefly divide Mandarin into north Mandarin and southwest Mandarin according to their accent resemblance. The population distribution for these dialectal regions is shown in Fig. 1.

In addition, all these volunteer speakers have not ever received any professional training in native English pronunciation. These volunteer speakers come from different geographical regions, where they lived for a long time an learned English while speaking one of major dialects. Their oral English retain apparent accents inherited from their Chinese dialects. Their oral proficiency in English also varies widely. Note that, the oral proficiency difference reflects the unbalance of English education in China. However, we do not account for education background and English proficiency in our corpus. The recordings by all the volunteer speakers are mixed together, only retaining their gender and dialect tags. The statistics of volunteer speakers is given in Table 1.

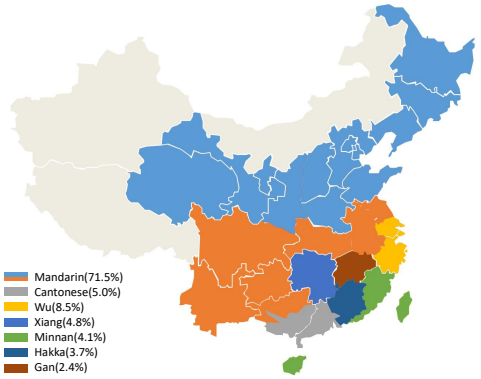

Fig. 1. Geographical distribution of major Chinese dialects.

Data Recording & Cleaning: We design a mobile APP, as shown in Fig. 2 to efficiently record and collect speech data. We use self-reported questionnaires to collect the volunteer speakers' information, including gender, hometown city and dialect.

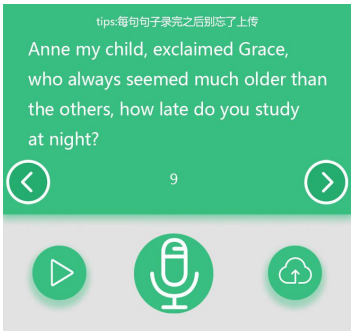

Fig. 2. Mobile APP used to record and collect speech data.

The mono channel recordings are sampled at 16kHz. We examine all the recordings and filter out some items exhibiting unacceptable mispronunciation. Note that, we allow mispronunciations as long as they are understandable. Especially, we allow the existence of slight, or even strong accents.

Some recordings may include background noises and popping sounds (i.e., pronunciation of aspirated plosives). Therefore, we used a noise suppression library, RNNoise [18] to filter out these audio noises and popping sounds. Note that RNNoise may introduce slight artifacts during noise suppression. In addition, we use SoX [19] examine and filter out extra silences.

Dialects	Mandarin (N/S)	Cantonese (i.e.,Yue)	Wu	Xiang	Minnan	Hakka (i.e., Kejia)	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.

Data Annotations: Each recording in our speech corpus contains a word-level orthographic transcription. We manually inspect and clean all recordings, by inserting, substituting, or deleting mismatching characters. We select 8 data-sets from the seven major dialectal regions and each data-set contains about 200 utterances, contributed by at least five volunteer speakers. We make use of the CMU pronouncing dictionary [20] to generate phonetic transcriptions corresponding to these recordings. We use P2FA [21] for automatic forced alignment; then the temporal boundary of words and phones are manually adjusted using PRAAT [22]. We use IPA (International Phonetic Alphabet) to assist the process of manual annotation, as shown in Fig. 3.

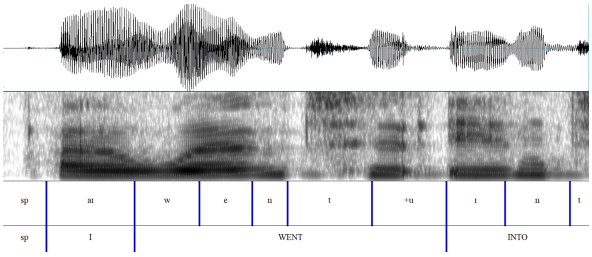

Fig. 3. Manual annotations assisted by PRAAT.

3. CORPUS STRUCTURE AND STATISTICS

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 2 summarizes these three sets used in our corpus.

Our corpus is contributed by 389 volunteer speakers, including 186 males and 203 females. 276 speakers contribute less than fifty recordings each. The rest 113 contribute fifty recordings each, and among them, 6 speakers have 150 recordings each. The number of volunteer speakers varies in seven dialect regions, but we guarantee the sufficient recordings from any regional dialect. Our corpus consists of 31.6 hour recordings in total, including 16.7 hours by male volunteers and 14.9 hours by female volunteers. Tables 1 and 2 list the statistics of our corpus.

data-set	duration(hrs)	male (hrs)	female(hrs)
training	27.2	14.0	13.2
development	2.3	1.4	0.9
test	2.1	1.3	0.8

Table 2. Statistics on speakers' gender and recording hours.

We manually annotated 1600 utterances, after performing 230 phoneme insertions, 2018 phoneme substitutions and 2158 phoneme deletions. Fig. 4 shows the phoneme error statistics for each dialectal region, while taking the most frequent errors. The common insertion errors occur to [t], [d], [k], [r] and [l]. The common substitution errors occur to [z]→[s],[t]→[d],[dh]→[d],[eh]→[ay] and [v]→[f]. The common deletion errors occur to [t], [d], [k], [r] and [l]. 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. This also implies that a regional dialect has a significant impact on speech pronunciation in English learning.

4. EXPERIMENT

In this section, we present our baseline experiments of ASR system on SELL-CORPUS using Kaldi toolkit [23].

A monophone model is first trained using 2,000 utterances selected from the train set. The input features are 13 dimensional Mel-frequency cepstral coefficients (MFCC). Based on the training alignments of mono-trained model with 3,000 utterances, a triphone GMM-HMM model is trained using delta and delta-delta features, denoted as tri1. The GMM is further trained by linear discriminant analysis (LDA) and MLLT transform with 5,000 utterances, denoted as tri2. A speaker-adapted model [24, 25] is trained using feature-space maximum likelihood linear regression (fMLLR) to transfer the features with all of train set data based on the alignments using the tri2 model. We denote the fMLLR trained model as tri3. A time-delayed neural network (TDNN) is then trained using MFCC features and i-vector features. The TDNN is built using 6 hidden layers in total, consisting of 1024 hidden units in each layer. Our corpus is available for download at `http://www.roseducation.org/sell-corpus`.

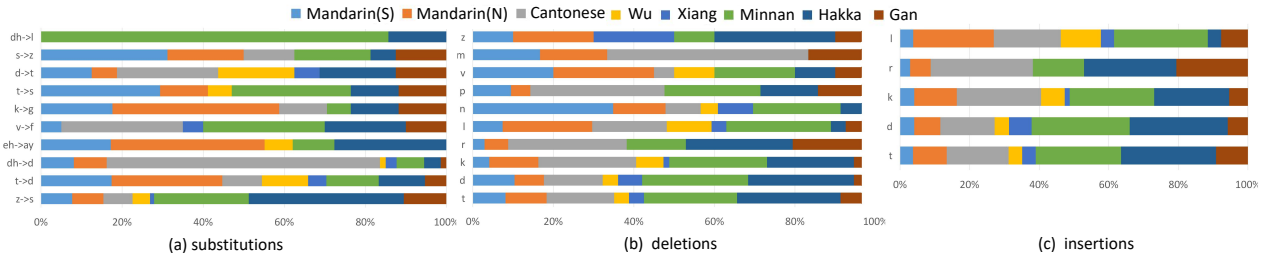

Fig. 4. Dialect-dependent pronunciation error distributions. (a) substitutions; (b) deletions and (c) insertions.

In our ASR system, we use 3-gram language models suggested in LibriSpeech [26]. A statistics of the resulting triphone delta-delta GMM-HMM model (tri1), LAD-MLLT GMM-HMM model (tri2), SAT(fMLLR) GMM-HMM model(tri3) and TDNN is given in Table 3.

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 3. 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 [26] for speech input with Chinese accents.

Based on the models trained above, we can achieve up to 11.51% word error rate (WER) on average in the final TDNN model using our corpus. As shown in Table 3, it was observed that the WER decreases significantly from the model in stage 1 to the one in stage 5. We also present the comparison results while applying the trained models to two subsets BWC and LXC in the Chinese accented corpus L2-ARCTIC[1] [27]. We observe a performance improvement (8.08%) over the results using BWC and LXC in L2-ARCTIC. Moreover, an ASR model trained using our corpus outperforms the native English corpus LibriSpeech [26] for speech input with Chinese accents (see the fourth column of Table 3). This also confirms that accents have significant influence with respect to pronunciation assessment.

By replacing the word-level language model with a phoneme bigram language model, we decode the test-set in a phone-level using the model tri4 and the TDNN model, respectively. We collect the most frequent phone-level decoding prediction errors in insertion, substitution and deletion, as shown in Table 4. While comparing these results with the manual annotation results in Section 3, we find that our training results can also capture most of phonetic errors that are often made by L2 English learners in China.

insertions	substitutions	deletions
[dh],[n],[d], [ah],[t],[ih], [s],[l],[ao],[r]	[z]→[s],[d]→[t], [t]→[d], [iy]→[ih], [ae]→[eh],[ah]→ [ih], [s] →[z], [ih]→[iy], [eh]→ae],[dh]→[l]	[t],[d],[ah],[n], [l],[dh],[ih],[ae], [er],[r]

Table 4. The most frequent (top 10) phoneme mistakes made by volunteer speakers among seven major regional districts.

5. 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. Our corpus has a few limitations. First, we did not include the minorities in Xinjiang, Tibet and Inner Mongolia (light-grey territory in Fig. 1), as their languages are phonetically, morphologically, and syntactically different from Chinese. Second, we only include major regional dialects based on geographical territory suggested by Chinese dialectology. In fact, the realities of speech are more complicated than the standard subdivision of regional dialects. Therefore, further studies and data collection are necessary to bring a deeper understanding on accents and pronunciation errors by L2 speakers. There exists the gender imbalance in our speech collection. We would like to collect more speech data in order to gradually reduce such an imbalance. With the increment of speech recordings, we expect the automatic alignment will be increasingly accurate and manual alignment/adjustment can be avoided.

[1]A joint research at Texas A&M University and Iowa State University, including recordings from twenty non-native speakers of English whose first languages are Hindi, Korean, Mandarin, Spanish, and Arabic.

6. REFERENCES

[1] British Council, "The English Effect," 2013.

[2] Michael Swan, "The influence of the mother tongue on second language vocabulary acquisition and use," *Vocabulary: Description, Acquisition and Pedagogy*, pp. 156–180, 1997.

[3] Silke M Witt and Steve J Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[4] Kun Li, Xiaojun Qian, and Helen Meng, "Mispronunciation detection and diagnosis in L2 English speech using multidistribution deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 193–207, 2017.

[5] Tobias Cincarek, Rainer Gruhn, Christian Hacker, Elmar Nöth, and Satoshi Nakamura, "Automatic pronunciation scoring of words and sentences independent from the non-natives first language," *Computer Speech & Language*, vol. 23, no. 1, pp. 65–88, 2009.

[6] Liu Wai Kat and Pascale Fung, "Fast accent identification and accented speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1999, vol. 1, pp. 221–224.

[7] Eric Atwell, Peter Howarth, and Clive Souter, "The ISLE corpus: Italian and German spoken learners English," *International Computer Archive of Modern and Medieval English*, vol. 27, pp. 5–18, 2003.

[8] Emi Izumi, Kiyotaka Uchimoto, and Hitoshi Isahara, "The NICT JLE Corpus: Exploiting the language learner's speech database for research and education," *International Journal of the Computer, the Internet and Management*, vol. 12, no. 2, pp. 119–125, 2004.

[9] Caroline R Wiltshire and James D Harnsberger, "The influence of Gujarati and Tamil L1s on Indian English: A preliminary study," *World Englishes*, vol. 25, no. 1, pp. 91–104, 2006.

[10] QiuFang Wen, LiFei Wang, and MaoCheng Liang, *Spoken and Written English Corpus of Chinese Learners*, Foreign Language Teaching and Research Press, 2009.

[11] Huizhong Yang and Naixing Wei, "Construction and research of spoken English corpus for Chinese learners," 2005.

[12] Raymond W.M. Ng, Alvin C.M. Kwan, Tan Lee, and Thomas Hain, "SHEFCE: A Cantonese-English bilingual speech corpus for pronunciation assessment," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5825–5829.

[13] Hua Chen, Qiufang Wen, and Aijun Li, "A learner corpus-ESCCL," in *The Speech Prosody Conference*, 2008, pp. 155–158.

[14] Takashi Fukuda, Raul Fernandez, Andrew Rosenberg, Samuel Thomas, Bhuvana Ramabhadran, Alexander Sorin, and Gakuto Kurata, "Data augmentation improves recognition of foreign accented speech," *Interspeech*, pp. 2409–2413, 2018.

[15] Abhinav Jain, Minali Upreti, and Preethi Jyothi, "Improved accented speech recognition using accent embeddings and multitask learning," *Interspeech*, pp. 2454–2458, 2018.

[16] Kingsley Bolton and David Graddol, "English in China today," *English Today*, vol. 28, no. 3, pp. 3–9, 2012.

[17] "Project Gutenberg," `https://www.gutenberg.org`.

[18] "RNNoise," `https://people.xiph.org/~jm/demo/rnnoise`.

[19] "SoX," `http://sox.sourceforge.net`.

[20] "CMUDICT SPHINX 4.0," `https://http://svn.code.sf.net/p/cmusphinx/code/trunk/cmudict/sphinxdict/cmudict_SPHINX_40/`.

[21] Jiahong Yuan and Mark Liberman, "Speaker identification on the SCOTUS corpus," *Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3878–3881, 2008.

[22] Paul Boersma and David J M Weenink, "PRAAT, A system for doing phonetics by computer," *Glot International*, vol. 5, 2002.

[23] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*.

[24] Mark JF Gales and Philip C Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech & Language*, vol. 10, no. 4, pp. 249–264, 1996.

[25] Tasos Anastasakos, John McDonough, Richard Schwartz, and John Makhoul, "A compact model for speaker-adaptive training," in *IEEE International Conference on Spoken Language*, 1996, vol. 2, pp. 1137–1140.

[26] 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)*, 2015, pp. 5206–5210.

[27] 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," Tech. Rep., Perception Sensing Instrumentation Lab, 2018.