INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

LAIX Corpus of Chinese Learner English: Towards a Benchmark for L2 English ASR

Yanhong Wang[1], *Huan Luan*[1], *Jiahong Yuan*[2], *Bin Wang*[1], *Hui Lin*[1,3]

[1] LAIX Inc.
[2] Baidu Research
[3] Shanghai Key Laboratory of Artificial Intelligence in Learning and Cognitive Science
{laura.wang, lisa.luan, engine.wang, hui.lin}@liulishuo.com

Abstract

This paper introduces a corpus of Chinese Learner English containing 82 hours of L2 English speech by Chinese learners from all major dialect regions, collected through mobile apps developed by LAIX Inc. The LAIX corpus was created to serve as a benchmark dataset for evaluating Automatic Speech Recognition (ASR) performance on L2 English, the first of this kind as far as we know. The paper describes our effort to build the corpus, including corpus design, data selection and transcription. Multiple rounds of quality check were conducted in the transcription process. Transcription errors were analyzed in terms of error types, rounds of reviewing, and learners' proficiency levels. Word error rates of state-of-the-art ASR systems on the benchmark corpus were also reported.

Index Terms: speech recognition corpus, L2 English, Chinese learners

1. Introduction

In the mobile internet era, voice interaction between human and machine becomes natural and has been changing our day-to-day lives, including how we learn a language. There are growing interests of using ASR technologies in second language (L2) acquisition with many applications such as computer assisted pronunciation training (CAPT), interactive dialogues with the computer and speech-enabled multiple-choice exercises, and spoken dialogue systems for touring and interactive practice with corrective feedbacks. Recognizing L2 speech is those applications is extremely challenge because speakers are still in the process of learning the second language. The L2 speech not only tends to contain more hesitations and disfluencies, it also deviates from the naive speech in many aspects such as phonemes, syllables, prosody, word forms and the order of words. Building a high standard benchmark dataset for L2 speech is therefore crucial for the research and development of L2 ASR systems.

Although there are a number of standard datasets available for L1 English, no dataset is widely accepted for evaluating ASR performance on L2 English, especially spontaneous speech. The LAIX Corpus of Chinese Learner English was therefore created for this purpose. We hope the corpus will set a standard for the academia and industry to evaluate the performance of ASR for Chinese L2 English, and will be the first step towards a benchmark for L2 English ASR with different L1 languages.

There have been many important L1 English benchmark data sets available for ASR development, for example, the TIMIT [1] database for small-scale speech and speaker recognition, the Switchboard database [2] for large-scale recognition of telephone conversions, the WSJ database [3] for large-scale continuous speech recognition on the domain of broadcast news, and the Librispeech database [4] of 1000 hours of speech derived from 8000 audiobooks.

A number of corpora of accented English are available. The CSLU Foreign Accented English corpus [5] consists of spontaneous speech in English by native speakers of 22 different languages. The Wildcat corpus [6] contains both scripted and spontaneous speech recordings from native and foreign accented English. The GMU corpus [7] presents a large set of reading samples from a variety of language backgrounds. The AESOP project [8] collected L2 English speech data from Asian countries to derive a set of core properties common to all varieties of Asian English, as well as to discover features that are particular to individual L2 varieties. Finally, CUHK [9] designed, collected and annotated a corpus to elicit supra-segmental information from Chinese learners of English. Note that most of these corpora are designed for accented English with read-aloud speech and with L2 speakers that are more on the high-end of English proficiency level. In this study, we designed a benchmark corpus with a wider coverage of proficiency level (from low to high) and less constrained speech (e.g., spontaneous speech).

For Chinese Learner English, there is no widely accepted standard benchmark for evaluating the performance of ASR, particularly due to the great varieties of Chinese dialects. A large number of speakers from many regions in China have to be recruited to create such a benchmark dataset. Thanks to the LAIX English learning mobile apps we are able to collect speech data from millions of users under the data privacy agreement. The LAIX corpus contains 9,208 speakers from 384 cities from all major dialect regions in China, in total of 82 hours. It represents a full range of Chinese learners' English and therefore can serve as a benchmark dataset for the development of ASR for L2 English by Chinese learners, which will be the first of this kind reported in the literature.

High quality transcription is of great importance to an ASR benchmark dataset. With regard to human transcription agreement on L1 English, [10] observed word level disagreement of less than 5% on two spontaneous speech corpora, even with high noise levels in the signal. Human agreement in the Buckeye corpus ([11, 12]) was found to be around 98% on a word-token basis. To our knowledge there have been no systematic studies of human transcription agreement on spontaneous English speech by Chinese learners. In this study, we present a novel multi-stage approach that can significantly improve inter-transcriber agreement.

http://dx.doi.org/10.21437/Interspeech.2020-1677

The remainder of this paper is organized as follows: Section 2 introduces the data sources; Section 3 describes the transcription procedure and evaluation; Section 4 reports the experiments' result; and Section 5 summarizes the results.

2. Data sources

The data were selected from two sources in Liulishuo mobile Apps: IELTS and Chatbot, based on factors such as speaker gender, pronunciation scores, speaking topics, etc. The datasets are summarized in Table 1 and described in detail in the following sections.

Table 1: *The statistics of IELTS and Chatbot datasets.*

Test set	Number of speakers	Number of utterances	Vocabulary size	Total duration
IELTS	2,860	3,942	8,345	53 hours
Chatbot	6,348	30,457	4,209	29 hours

2.1. IELTS

Utterances in the IELTS dataset were collected from The IELTS Liulishuo APP, a speaking test simulator on mobile devices that is designed to help users to improve their performance on the IELTS spoken English proficiency test. The IELTS test [13] consists of three parts. In part 1, speakers are asked a series of questions on familiar topics such as home, family, weather, films, work, studies, etc. Utterances in part 1 usually last for 20 to 30 seconds. In part 2, speakers are given a particular topic to talk about, including some points to be included in the talk. Speakers have one minute to plan and make notes, and then speak for one to two minutes without being interrupted. In part 3, speakers are asked about opinions on a range of issues related to the topic in part 2. Utterances in part 3 usually last for one minute.

After several steps of data filtering, the final IELTS dataset contains 3,942 utterances from 1,436 female and 1,424 male speakers. Each speaker has less than 10 utterances. Among the 3,942 utterances, 1,933 utterances are from female speakers, and 2,009 utterances are from male speakers. 1,319 utterances belong to part 1 of the test, 1,366 utterances belong to part 2, and 1,257 utterances belong to part 3. There are 184 to 215 utterances on each of the topics, as listed in Table 2, to ensure the diversity of the vocabulary used.

Table 2: *IELTS topic coverage.*

Topics of IELTS corpus	
sports & entertainment	advertising & marketing
media & news & fame	accommodation & community
travel & transportation	animals & plants
study & work	food & health
others	shopping & fashion
history & time	family & relationships
arts & personal taste	economy & management
computers & technology	law & crime & punishment
nature & weather & season	language & linguistics
building & engineering	environment & pollution
character & personality	

In the official IELTS speaking test, the speaking examiner tests on four criteria: fluency and coherence, lexical resource, grammatical range and accuracy and pronunciation. All these four criteria weight equally. Higher scores in IELTS mean higher proficiency. The distribution of the speaking scores in the dataset is consistent with the actual score distribution of IELTS Liulishuo APP users and real IELTS test takers in China, which is given in Table 3.

Table 3: *Actual score distribution of IELTS Liulishuo APP.*

score	percent	score	percent
1.0	4.00%	4.5	11.00%
1.5	2.00%	5.0	17.00%
2.0	2.00%	5.5	23.00%
2.5	3.00%	6.0	12.00%
3.0	4.00%	6.5	6.00%
3.5	5.00%	7.0	4.00%
4.0	8.00%	7.5	0.00%

2.2. Chatbot

Utterances in the Chatbot dataset were collected from The English Liulishuo App, an engaging and fun English learning mobile app which has advanced automatic assessment of English, professional English training courses, comprehensive learning materials, and well-designed leveled games.

The Liulishuo Chatbot has 27 speaking scenes. Users make conversations with the bot on a pre-selected scene. After several steps of data filtering, the final Chatbot dataset contains 30,457 utterances from 3,307 female and 3,041 male speakers. Each speaker has less than 10 utterances. Among the 30,457 utterances, 16,763 utterances are from female speakers and 13,694 from male speakers. Each utterance lasts for approximately three to five seconds. The number of utterances on each scene is shown in Table 4.

Table 4: *Chatbot's speaking scene list.*

Chatbot's speaking scene list	
592: Daily life	613: Halloween
2556: A little girl named Amy	153: Hobbies
206: How to introduce family	897: Job interview
552: Conversations in bar	774: A cat named Kitty
2106: Work	204: Hip-hop star Kong Lingqi
2915: Detective Charlie	2889: Sam's love story
2135: Escort little chicken home	1585: New York
765: Childhood	2191: Boyfriend
1492: A superstar's born	746: Chat with AI robot
286: Friend	1138: Study with AI teacher Tony
1553: Travelling frog	2131: In the airport
1233: Self introduction	745: Donald Trump

The users' English proficiency was assessed by a placement test in the Liulishuo App. The Placement Test sorts individual users into one of eight levels, and provides content customized to their level. Level 1, beginner, can use basic English to communicate simple facts and personal information with others. Level 2, elementary, can use simple sentences to discuss topics related to daily life and ask for information. Level 3, lower intermediate, can talk about events in the past, present and future and use more varied and complex language to describe themselves. Level 4, intermediate, can logically describe causes and effects of events. Level 5, upper intermediate, can communicate in a relaxed and fluent manner, and begin to discuss abstract concepts and to back their opinions with reasoning. Level 6, advanced 1, can understand or discuss both concrete and abstract topics, and can participate in any technical discussions in relative specialization. Level 7, advanced 2, can use language flexibly and effectively for social, academic and professional purposes, and can communicate in complex sentences with complex

vocabulary and phrases. Level 8, mastery, can understand with ease virtually everything heard or read. The distribution of the PT level in the dataset is consistent with the actual PT level distribution of the App users, as listed in Table 5.

Table 5: *Actual PT level distribution of English Liulishuo.*

level	percent
1	23.05%
2	31.00%
3	35.00%
4	7.00%
5	3.00%
6	0.05%
7	0.00%
8	0.00%

3. Data transcription

All utterances were manual transcribed and inspected. Raw transcriptions were transcribed by 20 experienced transcribers, and then we had reviewers check the transcriptions for multiple rounds.

3.1. Transcription guidelines

The transcribers followed the following guidelines:

- Incomplete words: the first portion of text is the speaker's actual pronunciation; the text in the square brackets is the transcriber's best guess as to what was intended, such as "ci[city]";
- Fillers: excepted for already existed modal particles (e.g. "oh"), only “err”, “ah” and “um” can be used as fillers. Choose the one that is most similar in acoustic terms.
- Speech errors: as to incorrect use of plural endings, past participle, singular and derivations, the first portion of text is the incorrect form, and the text in the square brackets is the root word, such as "sheeps[sheep]"
- Unintelligible words: labeled as "<false_word>"
- Undistinguishable words: labeled as "<?>"
- Breath: labeled as "<breath>"
- Laughter: labeled as "<laughter>"
- Cough: labeled as "<cough>"
- Chinese: labeled as "<Chinese>"
- Other person's voice: labeled as "<others>"

Transcribers and reviewers further obtained the prompts of the different test tasks so that they were able to familiarize themselves with the context of the responses.

3.2. Procedure

After transcribers finished raw transcriptions, we had several rounds of reviewing. The transcription time was approximately 9.6 times real-time (xRT) for the first pass and 4-5.7 times real-time (xRT) for each of the following rounds of reviewing.

The IELTS test set has five rounds of reviewing, and the Chatbot test set has four rounds. There are nine reviewers in total, five for the IELTS test set and four for the Chatbot test set. The reviewers are all proficient in English, and they are experienced in linguistic annotation tasks and familiar with our transcription guidelines. Prior to the reviewing, we asked them to participate in three rounds of transcription tests. We compared their results and computed each reviewer's WER. After training and discussion, the reviewers' WER results were below a pre-defined threshold of 10%.

The first two rounds of reviewing were run in parallel, followed by a third round of reviewing. In the third round the reviewer checked all transcriptions based on the difference (i.e., deletion, insertion, and substitution) between the first two rounds of transcriptions.

For IELTS, in the fourth-round reviewer reviewed all transcriptions whereas the last reviewer (the fifth round) randomly checked 30% of all transcriptions. For Chatbot, the fourth-round reviewer randomly selected 30% to 100% of all transcriptions for checking, dependent on the transcription quality.

3.3. Evaluation

To evaluate transcription accuracy, the transcripts after the last round reviewing were taken as the gold standard. Tags, punctuation marks and time stamp information were removed to produce clean reference transcripts. Uppercase letters were converted to lowercase.

3.3.1. Results of IELTS

Table 6 shows WERs on IELTS across multi-stage transcription. The highest WER is 11.39% and the lowest is 0.46%. As we can expect, there is some variation among transcriber performance and the reviewing process reduces WERs greatly.

The WER of raw transcription (prior to reviewing) is 11.39%, which is much higher than that of native speech previously reported (5% WER or less). The WER dropped to 5.57% after first round of reviewing and to 8.74% after the second round of reviewing (we should note that the first two rounds of reviewing were run in parallel, it was not that round 2 was based on round 1). The large difference between round 1 and round 2 called for another round of reviewing. After the reviewing of round 3, WER dropped significantly to 1.04%. It further dropped to 0.46% after round 4. The results reveal that a multi-stage transcription protocol can dramatically reduce transcription errors. From our experiment, three rounds of reviewing should be necessary and sufficient, and further reviewing can only slightly reduce transcription errors.

Table 6: *WERs based on using the transcription of round 5 as the gold standard on IELTS during reviewing process.*

IELTS's WER results during reviewing process		
raw	round5	11.39
round1	round5	5.57
round2	round5	8.74
round3	round5	1.04
round4	round5	0.46

In Figure 1, we take a closer look at transcription WERs for different IELTS scores. We can see that WERs decrease with the increase of the IELTS score. For raw transcription, there is a great decrease in WER from 19.07% to 11.6% as the score increases from 1-1.5 to 2-2.5. After that, WERs are gradually reduced as the score increases. The same trend can also be found for the first two rounds of reviewing. That is, utterances of lower proficiency speakers are more difficult to

transcribe than those of high proficiency speakers. However, after two rounds of reviewing, proficiency level has little effect on human transcription accuracy in further reviewing, as shown in Figure 1 for round 3 and round 4.

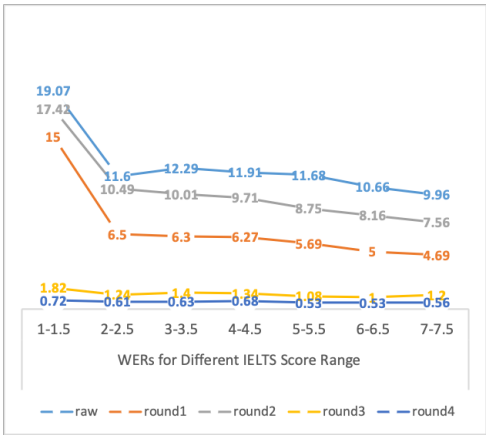

Figure 1: *WERs for different IELTS score range.*

In Table 7, we compare the percentage of transcription errors and that of incomplete words. We can see that, as also shown in Figure 1, proficiency level has a great influence on the percentage of word transcription errors, the lower the proficiency level, the higher chance of word transcription errors. However, with the increase of the IELTS score, the percentage of incomplete words falls only slightly from 2.72% to 1.45%. Speaker's proficiency level doesn't have a large impact on the occurrences of incomplete words.

Table 7: *Percentage of word errors and incomplete words of different IELTS score range.*

Score	Percentage of Errors	Percentage of Incomplete Words
1-1.5	19.07%	2.72%
2-2.5	11.60%	2.21%
3-3.5	12.29%	2.57%
4-4.5	11.91%	2.63%
5-5.5	11.68%	2.30%
6-6.5	10.66%	1.89%
7-7.5	9.96%	1.45%

3.3.2. Results of Chatbot

The transcription accuracies on Chatbot are listed in Table 8, calculated against the results after the last round of reviewing. The WER was greatly improved from 17.91% to 0.67% after three rounds of reviewing. As listed in Table 5, more than half of the speakers in the Chatbot dataset were on a PT level between beginner and elementary. Moreover, the selected Chatbot utterances only last three to five seconds and appear in no context. These explain why WER in the raw transcriptions are very high (nearly 18%). Nonetheless, it dropped greatly to 0.67% after three rounds of reviewing.

Table 8: *WERs based on using the transcription of round4 as the gold standard on Chatbot during reviewing process.*

Chatbot's WER results during reviewing process		
raw	round4	17.91
round1	round4	8.11
round2	round4	7.54
round3	round4	0.67

4. Experiments

To see how state-of-the-art English ASR systems perform on the benchmark corpus, four ASR systems for English with open API services were evaluated. The evaluation scripts are available at https://github.com/lingochamp/open-asr. We also evaluated LAIX's Kaldi's hybrid system, a TDNN-f contains 17 sub-sampled time-delay neural network layers with low-rank matrix factorization (TDNNF) [14], and trained on our 10K-hours in-house speech of L2 English, using the lattice-free MMI [15] recipe in Kaldi toolkit [16]. Language model is a 3-gram language model (LM) with modified Kneser-Ney smoothing trained using the SRILM toolkit [17]. The performance of all the systems on the LAIX benchmark corpus is presented in Table 9. Note that the comparison is by no means fair, but the results shed some light on the challenging nature of recognizing L2 speech and how different the task is from conventional ASR tasks.

Table 9: *WERs on Chatbot and IELTS of different ASR systems*

API service for English	Chatbot	IELTS
LAIX's kaldi	10.87	12.50
Iflytek	18.04	25.97
Google	53.83	36.68
Baidu	31.25	24.84
Tencent	43.57	27.90

5. Conclusion

In this paper, we introduce a corpus of Chinese Learner English, to serve as benchmark test sets for ASR on L2 English. Besides a detailed description of the corpus, including data and the transcription effort, we carefully compared and evaluated WERs of several rounds of reviewing. We found that the WER in the raw transcription was substantially higher than that for native speech, and three rounds of reviewing were necessary and sufficient to approach the agreement level on native speech transcription. We demonstrated that transcription accuracy was positively correlated to speakers' proficiency level (accuracy is higher for high proficiency speakers), and negatively correlated to the lack of context (accuracy is lower when no context is provided). The LAIX benchmark corpus archive contains 16-bit PCM wave files with 16 kHz sampling frequency (one file for each utterance), the text of the utterance and their corresponding transcription files in Praat TextGrid format. The corpus is available for research purpose upon request.

6. Acknowledgments

The Chatbot and IELTS test sets were designed and reviewed by the members of the LAIX Data Team in the Algorithm Department. We wish to thank all the transcribers and reviewers that took part in the effort of creating the data sets. The reviewers are Yanhong Wang, Yuchan Hong, Zhicheng Zhang, Wei Xin, Zhe Gu, Yiwen Lu, Jing Liu, Li Li and Siqi Xie.

7. References

[1] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren.: DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. NIST(1990).

[2] Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). SWITCHBOARD: telephone speech corpus for research and development. In Proceedings of Acoustics, Speech, and Signal Processing (ICASSP-92). IEEE.

[3] D. B. Paul and J. M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," in Proc. of the DARPA Speech and Natural Language Work- shop, Feb. 1992.

[4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 2015, pp. 5206–5210.

[5] https://catalog.ldc.upenn.edu/LDC2007S08

[6] Van Engen, K. J., Baese-Berk, M., Baker, R. E., Kim, M., & Bradlow, A. R. (2010). The Wildcat Corpus of native- and foreign- accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. Language & Speech, 53, 510–540.

[7] http://accent.gmu.edu

[8] C.-Y. Tseng and T. Visceglia, "AESOP (Asian English Speech Corpus Project) and TWNAESOP," in Proc. Int. Conf. Workshop TEFL Appl. Linguistics, 2010, pp. 60-65.

[9] Li, M., Zhang, S., Li, K., Harrison, A., Lo, W.K., Meng, H., 2011b. Design and collection of an L2 English corpus with a suprasegmental focus for Chinese learners of English, in: Proc. ICPhS.

[10] Deshmukh, N., Duncan, R. J., Ganapathiraju, A. & Picone, J. (1996). Benchmarking Human Performance for Continuous Speech Recognition. Proceedings of ICSLP-96, pp. 2486-2489. Philadelphia, PA, October.

[11] Raymond, W.D., Pitt, M., Johnson, K., Hume, E., Makashay, M., Dautricourt, R. & Hilts, C. (2002). An analysis of transcription consistency in spontaneous speech from the Buckeye corpus. Proceedings of ICSLP-02, pp. 1125-1128. Denver, CO, September.

[12] Pitt M.A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. Speech Communication, 45 (1), pp.89-95.

[13] https://www.ielts.org/about-the-test/test-format

[14] D.Povey,G.Cheng,Y.Wang,K.Li,H.Xu,M.Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks." in Interspeech, 2018, pp. 3743– 3747.

[15] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi." in Interspeech, 2016, pp. 2751–2755.

[16] "Kaldiasr,"http://kaldi-asr.org.

[17] "Srilm - the SRI language modeling toolkit," http://www.speech.sri.com/projects/srilm/.