L2-ARCTIC: A Non-Native English Speech Corpus

Guanlong Zhao[1], Sinem Sonsaat[2], Alif Silpachai[2], Ivana Lucic[2]
Evgeny Chukharev-Hudilainen[2], John Levis[2] and Ricardo Gutierrez-Osuna[1]

[1]Department of Computer Science and Engineering, Texas A&M University, United States
[2]Department of English, Iowa State University, United States
{gzhao, rgutier}@tamu.edu, {sonsaat, alif, ilucic, evgeny, jlevis}@iastate.edu

Abstract

In this paper, we introduce L2-ARCTIC, a speech corpus of non-native English that is intended for research in voice conversion, accent conversion, and mispronunciation detection. This *initial* release includes recordings from ten non-native speakers of English whose first languages (L1s) are Hindi, Korean, Mandarin, Spanish, and Arabic, each L1 containing recordings from one male and one female speaker. Each speaker recorded approximately one hour of read speech from the Carnegie Mellon University ARCTIC prompts, from which we generated orthographic and forced-aligned phonetic transcriptions. In addition, we manually annotated 150 utterances per speaker to identify three types of mispronunciation errors: substitutions, deletions, and additions, making it a valuable resource not only for research in voice conversion and accent conversion but also in computer-assisted pronunciation training. The corpus is publicly accessible at https://psi.engr.tamu.edu/l2-arctic-corpus/.

Index Terms: speech corpus, voice conversion, accent conversion, mispronunciation detection

1. Introduction

Voice conversion (VC) [1] aims to transform utterances from a source speaker to make them sound as if a target speaker had uttered them. The closely related problem of accent conversion (AC) [2] goes a step further, mixing the source speech's linguistic content and accent with the target speaker's voice quality to create utterances with the target's voice but the content and pronunciation of the source speaker. When teaching a second language (L2), accent conversion can be used to create a "golden speaker," a synthesized voice that has the learner's voice quality but with a native speaker's accent (e.g., prosody, intonation, pronunciation) [3]. Several studies [4, 5] have suggested that having such a "golden speaker" to imitate can be beneficial in pronunciation training. Furthermore, in addition to providing language learners with a suitable voice to mimic, detecting mispronunciations is also a critical component for providing useful feedback to the learners in computer-assisted pronunciation training [6].

To train and evaluate voice and accent conversion systems designed for non-native speakers, one needs high-quality parallel recordings from the source and target speakers. Likewise, to develop and benchmark mispronunciation detection algorithms, detailed phoneme level annotations on pronunciation errors (e.g., phone substitution, additions, and deletions) are required. However, existing non-native English corpora (e.g., Speech Accent Archive [7] and IDEA [8]) do not fulfill these requirements (refer to section 2 for a detailed discussion.)

To fill this gap, we have built a non-native English speech corpus that contains ten non-native speakers of English in the initial release. The end goal for this corpus is to include 20 speakers from five different native languages: Hindi, Korean, Mandarin, Spanish, and Arabic. For *each* speaker, the corpus contains the following data:

- Speech recordings: over one hour of prompted recordings of phonetically-balanced short sentences
- Word level transcriptions: orthographic transcription and forced-aligned word boundaries for each sentence
- Phoneme level transcriptions: forced-aligned phoneme transcription for each sentence
- Manual annotations: a selected subset of utterances (~150), including 100 sentences produced by all speakers and 50 sentences that include phonemes likely to be difficult according to each speaker's L1, all annotated with corrected word and phone boundaries; phone substitution, deletion, and addition errors are also tagged

The dataset is hosted on an online archive and is freely available to the research community for non-commercial use. To the best of our knowledge, L2-ARCTIC is the first openly available corpus of its kind.

2. The need for a new L2 English corpus

A number of voice conversion studies [9-12] have relied on the Carnegie Mellon University (CMU) ARCTIC speech corpus [13] and, more recently, the Voice Conversion Challenge (VCC) dataset [14]. However, little attention has been paid to voice conversion between non-native speakers of English, in part due to the lack of high-quality speech recordings from those speakers, despite 80% of the English speakers in the world being non-native [15]. For example, CMU ARCTIC only has a few accented English speakers[1], either native speakers of different English dialects or highly proficient non-native speakers, whereas the VCC dataset was recorded solely by professional voice talents who are native English speakers. Therefore, these standard corpora are not suitable for either voice conversion between non-native speakers nor accent conversion tasks.

Among the non-native English corpora, the Speech Accent Archive [7] and IDEA [8] cover a wide range of native languages and speakers. However, each speaker only recorded a short paragraph (Speech Accent Archive) or a short free speech task (IDEA), and most of the recordings have strong

[1] JMK: Canadian accent; AWB: Scottish accent; KSP: Indian accent

background noise, making them ill-suited for voice/accent conversion. The Wildcat [16], LDC2007S08 [17], and NUFAESD [18] datasets have a limited number of recordings for each non-native speaker, and have restricted access – LDC2007S08 requires a fee, while Wildcat and NUFAESD are limited to designated research groups.

As for corpora for mispronunciation detection, the CU-CHLOE [19] and College Learners' Spoken English Corpus (COLSEC) [20] only contain speech and error tags from Chinese learners of English, and CU-CHLOE is (to our knowledge) not publicly available. The ISLE Speech Corpus [21] contains mispronunciation tags and is open for academic access, but it only focuses on a limited group of English learners (German and Italian). SingaKids-Mandarin [22] has a rich set of speech data, but it only focuses on mispronunciation patterns in Singapore children's Mandarin speech. In fact, most existing mispronunciation detection systems use their private datasets, which makes it difficult to compare experimental results across different publications [19, 23-25].

To overcome the insufficiencies outlined above, we constructed (and are now releasing) L2-ARCTIC to provide an open corpus for voice conversion between accented speakers, accent conversion, and mispronunciation detection. Zhao et al. [26] have performed a preliminary evaluation on voice/accent conversion tasks using a subset of the speakers in L2-ARCTIC. Using a joint-density GMM with MLPG and global variance compensation [9] (128 mixtures, ~5 min of parallel training data) as the voice conversion system, they obtained 2.5 Mean Opinion Score (MOS) on the converted speech, which was also rated as similar to the target voice. Furthermore, an accent-conversion algorithm based on frame-alignment using posteriorgrams was able to generate speech that was perceived as similar to a non-native target voice but markedly less accented (98% preference compared to non-native speech). This manuscript presents preliminary results on a new task: mispronunciation detection**.**

3. Corpus curation procedure

This initial release of L2-ARCTIC contains English speech of speakers from five different first languages: Hindi[1] [27], Korean, Mandarin, Spanish, and Arabic. We chose these L1s because each one has a distinct foreign/non-native accent in English and provides unique challenges. **Indian** speakers of English typically have native-like English fluency but use segmental and suprasegmental features in ways that are distinct from American English. Thus, Indian speakers have both advantages in approaching pronunciation changes (e.g., familiarity and comfort with English) and disadvantages (comfort with their English variety makes it particularly difficult to adjust their speech to salient differences with American English.) **Korean** learners of English have a large number of high functional load consonant and vowel difficulties (errors with many minimal pairs). Prosodically, Korean and English employ suprasegmental systems that have little overlap [28, 29]. **Mandarin** (Chinese/Putonghua) learners of English have difficulty with a range of consonant and vowel sounds and in producing correct English stress, intonation, and juncture [30-32]. **Spanish** learners of English may have difficulties distinguishing a number of high functional load contrasts in English [33, 34]. Spanish is also a five-vowel language, and Spanish learners find the more complex English vowel system especially challenging. Like English, Spanish uses both word stress and nuclear stress for emphasis but, because it does not use the unstressed vowel schwa, realizes stress differently. Finally, **Arabic** also has significantly fewer vowels than English, and while Arabic has word stress, it does not use stress in the same way that English does [35, 36]. In the future, we may also include speakers from other L1s if we find them to be useful to the research community.

3.1. Participants

For this initial release, we recruited two speakers (one male and one female) for each of the L1s, for a total of ten speakers. Speakers were recruited from Iowa State University's student body; their age range was from 22 to 43 years, with an average of 29 years (std: 6.9.) Demographic information of the speakers is summarized in Table 1. The proficiency level of English was measured using TOEFL iBT scores [37].

Table 1: *Demographic information of the speakers*

Speaker	L1	Gender	TOEFL iBT
HKK	Korean	M	114
YDCK	Korean	F	110
BWC	Mandarin	M	80
LXC	Mandarin	F	86
YBAA	Arabic	M	100
SKA	Arabic	F	79
EBVS	Spanish	M	70
NJS	Spanish	F	110
RRBI	Hindi	M	91
TNI	Hindi	F	99

3.2. Recording the corpus

To create the corpus, we used the 1,132 sentences in the CMU ARCTIC prompts. There were multiple reasons to choose these sentences. First, the ARCTIC prompts are phonetically balanced (100%, 79.6%, and 13.7% coverage for phonemes, diphones, and triphones, respectively), are open source, and can produce around one hour of edited speech. Second, the ARCTIC corpus itself has proven to work well with speech synthesis [38] and voice conversion tasks [9-11, 39]. Finally, the ARCTIC prompts are challenging for non-native English speakers so they can elicit potential pronunciation problems.

The speech was recorded in a quiet room at Iowa State University (ISU). We used a Samson C03U microphone and Earamble studio microphone pop filter for recordings; the microphone was placed 20 cm from the speaker to avoid air puffing. During each recording session, a linguist guided the L2 speaker through the process, asking the speaker to re-record a sentence if the production contained significant disfluency or deviated from the prompt. All speakers were instructed to speak in a natural manner. The speech was sampled at 44.1 kHz and saved as a WAV file.

Once the recording was finished, we removed repetitions and false starts, performed amplitude normalization, and segmented the utterances into individual WAV files. All of the above were done in Audacity [40]. The utterances were carefully trimmed to remove the leading and trailing silence and non-speech sounds such as lip smacks.

[1] Hindi is an Indo-Aryan language that is both an L1 and a language of wider communication. Thus Hindi speakers in the corpus may use Hindi as an L2, speaking another Indian language as an L1. Educated Indian English is a stable contact variety of English.

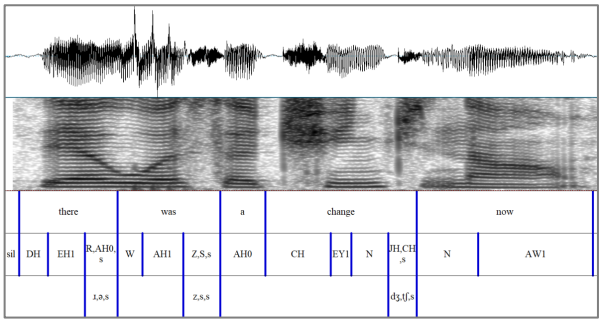

Figure 1: *A TextGrid with manual annotations. Top to bottom: speech waveform, spectrogram, words, phonemes and error tags, comments from the annotator*

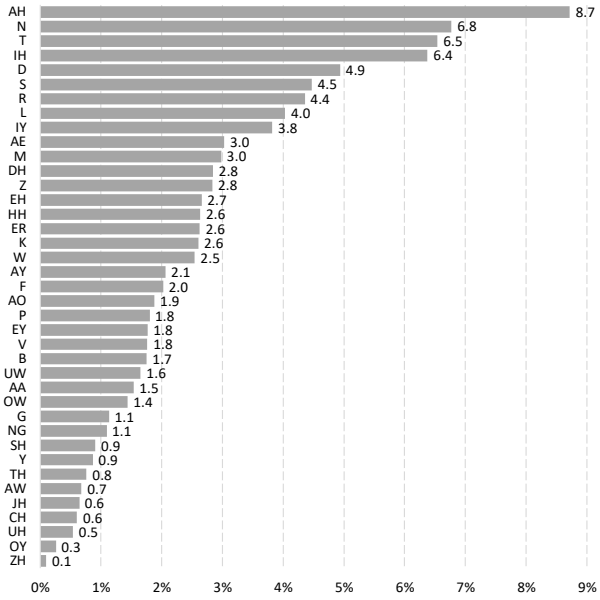

Figure 2: *Phoneme distribution of the corpus*

3.3. Corpus annotations

Our corpus provides orthographic transcriptions at the word level. We used the Montreal forced-aligner [41] to produce phonetic transcriptions in PRAAT's TextGrid format [42], which contains word and phone boundaries (Figure 1). Further, we performed manual annotations on a selected subset of sentences for each speaker. For all the speakers, we annotated a common set of 100 sentences. In addition, we annotated 50 sentences that included phoneme difficulties that were L1-dependent. In the end, the corpus contains up to 150 curated phonetic transcriptions per speaker[1]. Those transcriptions contain manually-adjusted word and phone boundaries, correct phoneme labels, mispronunciation error tags (phone additions, deletions, and substitutions), and comments from the annotators. To facilitate computer processing, we used the ARPAbet phoneme set for the phonetic transcriptions as well as the error tags. In the comment part of the transcriptions, however, annotators were allowed to use IPA symbols. To ensure high-quality annotations, we developed automated scripts to check the annotation consistency and then asked human annotators to fix problems. The annotators (N=3) were PhD students in the Applied Linguistics and Technology program at ISU. They were experienced in transcribing speech samples of native or non-native English speakers.

[1] Some speakers did not read all sentences, and a few sentences were removed for some speakers since those recordings did not have the required quality.

4. Corpus statistics

In total, the dataset contains 11,026 utterances, with most speakers recording the full ARCTIC set (1,132 utterances.)[1] The total duration of the corpus is 11.2 hours, with an average of 67 minutes (std: 9 minutes) of speech per L2 speaker. On average, each utterance is 3.7 seconds in duration. The pause before and after each utterance is generally no longer than 100 ms. Using the forced alignment results, we estimate a speech to silence ratio of 7:1 across the whole dataset. The dataset contains over 97,000 word segments, giving an average of around nine words per utterance, and over 349,000 phone segments (excluding silence). The phoneme distribution is shown in Figure 2.

Human annotators manually examined 1,499 utterances, annotating 5,199 phone substitutions, 1,048 phone deletions, and 497 phone additions. Figure 3 (a) shows the top-20 most frequent phoneme substitution tags in the corpus. The most dominant substitution errors were "Z→S," (voicing) "DH→D," (fricative→stop) "IH→IY," and "OW→AO" (use of a tense vowel for lax, and vice versa.) Each contains English phoneme distinctions that lead to common substitution errors for varied English learners. Figure 3 (b) shows the phone deletion errors in the annotations. In our sample group, the most frequent phoneme deletions were "D," "T," and "R," almost always in non-initial position. Many non-native speakers of English do not pronounce the American English phoneme "R" in postvocalic position (e.g., in *car* and *farm.*) "T" and "D" often occur as word endings and in consonant clusters both within and across words, where they were often omitted. Figure 3 (c) shows the phone addition errors in the annotations. The ones that stood out were "AH," "EH," "R," "AX (schwa)," "G", and "IH." The vowel additions simplify complex syllable structures with consonant clusters and so may serve to make the word more pronounceable. Table 2 provides a breakdown of pronunciation errors by L1s. Although others have used L1 to predict L2 pronunciation errors [33, 34, 43], such predictions are often inaccurate when applied to individual learners. Thus, this list is meant to start a discussion of the types of errors that actually occur in L2-ARCTIC.

Table 2: *Most frequent errors by native language; the top-5 error occurrences are listed in descending order*

L1	Substitutions	Deletions	Additions
Hindi	DH→D, Z→S, W→V EY→EH, TH→T	R, D, T ER, HH	R, AH, S, Y AA
Korean	DH→D, Z→S, IH→IY OW→AO, EH→AE	D, T, R HH, K	AX, IH, AH, S Y
Mandarin	Z→S, DH→D, IH→IY N→NG, V→F	D, T, R L, N	AH, AX, IH N, R
Spanish	Z→S, IH→IY, DH→D AE→AA, AH→AO	D, T, AH Z, IH	EH, AX, AH IH, IY
Arabic	P→B, OW→AO R→ERR, DH→Z, Z→S	T, R, D AH, IH	G, AH, IH AX, EH

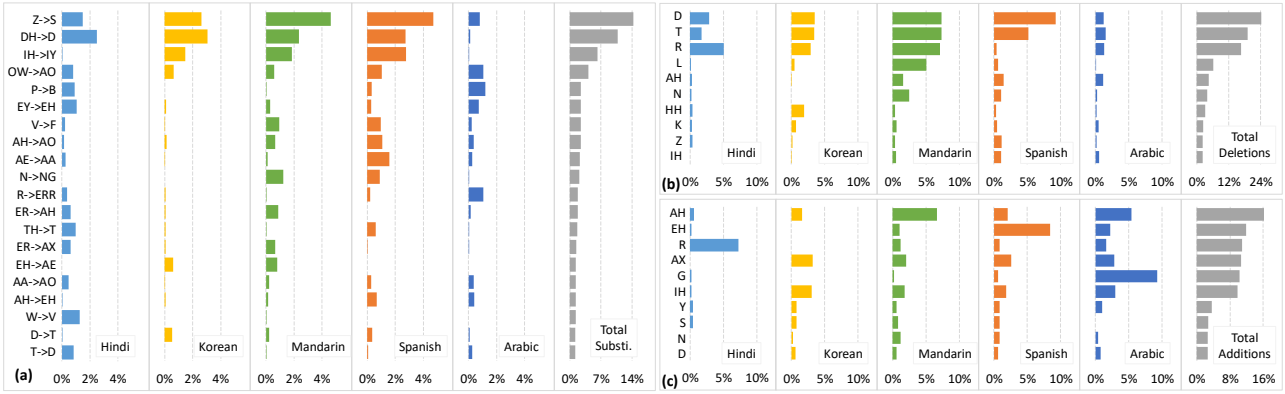

Figure 3: *L1-dependent pronunciation error distributions and the aggregated results. (a) Substitutions (b) Deletions (c) Additions. Errors with low frequencies were omitted; all the values are the percentages with respect to the total number of each error type (i.e., normalized universally); "ERR" means an erroneous pronunciation that is not in the ARPAbet phoneme set.*

5. Mispronunciation detection evaluation

This section provides initial results on mispronunciation detection using the 10 speakers that we have currently released. Our implementation is based on the conventional Goodness of Pronunciation (GOP) method as defined in [44]. The acoustic model we used was a triphone model (tri6b) as defined by Kaldi's Librispeech training script [45]. It is a GMM trained with 960 hours of native English speech [46], and contains 150,000 Gaussian mixtures. Three-state left-to-right HMMs were used for non-silent sounds. The Kaldi implementation does not have a fixed number of Gaussians for each HMM state. The Word Error Rate (WER) of this acoustic model was around 8% on clean speech when combined with a 3-gram language model.

We used the phone-independent thresholding variation of the GOP method to make the classification decisions, i.e., if the GOP score of a phone segment was higher than a threshold $\boldsymbol{P}$, then it was accepted as a correct pronunciation, otherwise it was rejected as an error. As a preliminary result, we only focused on substitution errors since the GOP is not suited for detecting additions and deletions.

Two hundred and six (206) utterances were withheld to determine the search range of the phoneme-independent detection threshold. The remaining 1,293 utterances were used as the testing set. In the testing data, excluding the additions and deletion tags, there are 41,353 phone samples in total, where 4,415 (10.7%) were tagged as substitution errors. We set the log GOP threshold between -16 and 0 and made the step size 0.1. For each experiment condition, we computed the detection precision rate as N_{TN}/N and the recall rate as N_{TN}/N_{errors}, where N_{TN} is the number of correctly predicted substitution errors, N is the total number of segments predicted as substitution errors, and N_{errors} is the total number of substitution errors in the testing set. The Precision-Recall curve is shown in Figure 4. When we set the threshold to -4.2 (in log scale), the precision equals recall (0.29). From this result, we can see that the dataset is quite challenging, because it contains speech data from different L1 backgrounds and recorded by speakers with a wide range of pronunciation challenges. This GOP implementation is open source and is available online[1].

[1] https://github.com/guanlongzhao/kaldi-gop

6. Conclusion

This paper has presented L2-ARCTIC, a new non-native English speech corpus designed for voice conversion, accent conversion, and mispronunciation detection tasks. Each speaker in L2-ARCTIC produced sufficient speech data to capture their voice identity and accent characteristics. Detailed annotations on mispronunciation errors are also included. Thus, it is possible to use this corpus to develop and evaluate mispronunciation detection algorithms. To the best of our knowledge, L2-ARCTIC is the first of its own kind, and we believe it fills gaps in both voice/accent conversion and pronunciation training.

The corpus is released under the CC BY-NC 4.0 license [47] and is available at `https://psi.engr.tamu.edu/l2-arctic-corpus/`. Future work will be focusing on adding ten more speakers to the corpus.

7. Acknowledgments

This work was supported by NSF awards 1619212 and 1623750. We would like to thank the anonymous participants for recording the corpus. We also would like to thank Ziwei Zhou for his assistance with the annotations. We appreciate suggestions from Christopher Liberatore and Shaojin Ding on early versions of this manuscript.

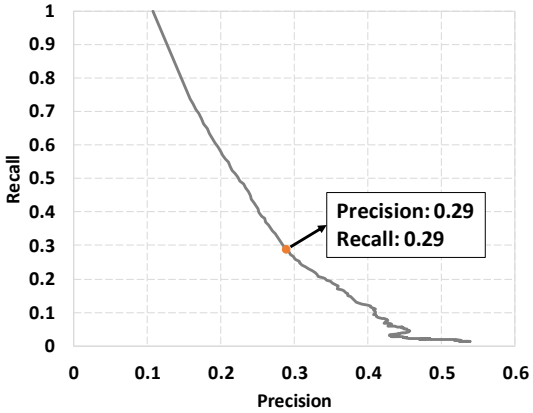

Figure 4: *Precision-Recall curve of a phone-independent GOP system to demo mispronunciation detection on L2-ARCTIC*

8. References

[1] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication,* vol. 88, pp. 65-82, 2017.

[2] S. Aryal and R. Gutierrez-Osuna, "Can Voice Conversion Be Used to Reduce Non-Native Accents?," in *ICASSP*, 2014, pp. 7879-7883.

[3] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech Communication,* vol. 51, no. 10, pp. 920-932, 2009.

[4] K. Probst, Y. Ke, and M. Eskenazi, "Enhancing foreign language tutors–in search of the golden speaker," *Speech Communication,* vol. 37, no. 3, pp. 161-173, 2002.

[5] M. P. Bissiri, H. R. Pfitzinger, and H. G. Tillmann, "Lexical stress training of German compounds for Italian speakers by means of resynthesis and emphasis," in *Australian International Conference on Speech Science & Technology*, 2006, pp. 24-29.

[6] J. Levis, "Computer technology in teaching and researching pronunciation," *Annual Review of Applied Linguistics,* vol. 27, pp. 184-202, 2007.

[7] S. Weinberger. Speech accent archive [Online]. Available: http://accent.gmu.edu

[8] P. Meier. IDEA: International Dialects of English Archive [Online]. Available: http://www.dialectsarchive.com/

[9] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, no. 8, pp. 2222-2235, 2007.

[10] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *ICASSP*, 2015, pp. 4869-4873.

[11] G. Zhao and R. Gutierrez-Osuna, "Exemplar selection methods in voice conversion," in *ICASSP*, 2017, pp. 5525-5529.

[12] Y.-C. Wu, H.-T. Hwang, C.-C. Hsu, Y. Tsao, and H.-M. Wang, "Locally Linear Embedding for Exemplar-Based Spectral Conversion," in *Interspeech*, 2016, pp. 1652-1656.

[13] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004, pp. 223-224.

[14] T. Toda *et al.*, "The Voice Conversion Challenge 2016," in *Interspeech*, 2016, pp. 1632-1636.

[15] J. Jenkins. (2008). *English as a lingua franca*. Available: http://www.jacet.org/2008convention/JACET2008_keynote_jenkins.pdf

[16] K. J. Van Engen, M. Baese-Berk, R. E. Baker, A. Choi, M. Kim, and A. R. Bradlow, "The Wildcat Corpus of native-and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles," *Language and speech,* vol. 53, no. 4, pp. 510-540, 2010.

[17] T. Lander. CSLU: Foreign Accented English Release 1.2 LDC2007S08 [Online]. Available: https://catalog.ldc.upenn.edu/ldc2007s08

[18] T. Bent and A. R. Bradlow, "The interlanguage speech intelligibility benefit," *The Journal of the Acoustical Society of America,* vol. 114, no. 3, pp. 1600-1610, 2003.

[19] K. Li, X. Qian, and H. 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.

[20] H. Yang and N. Wei, "Construction and data analysis of a Chinese learner spoken English corpus," ed: Shanhai Foreign Languse Eduacation Press, 2005.

[21] W. Menzel *et al.*, "The ISLE corpus of non-native spoken English," in *Proceedings of LREC 2000: Language Resources and Evaluation Conference, vol. 2*, 2000, pp. 957-964.

[22] N. F. Chen, R. Tong, D. Wee, P. X. Lee, B. Ma, and H. Li, "SingaKids-Mandarin: Speech Corpus of Singaporean Children Speaking Mandarin Chinese,"in *Interspeech*, 2016, pp. 1545-1549.

[23] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication,* vol. 67, pp. 154-166, 2015.

[24] Y.-B. Wang and L.-s. Lee, "Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning," *IEEE/ACM Transactions on Audio, Speech and Language Processing,* vol. 23, no. 3, pp. 564-579, 2015.

[25] H. Huang, H. Xu, X. Wang, and W. Silamu, "Maximum F1-score discriminative training criterion for automatic mispronunciation detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 23, no. 4, pp. 787-797, 2015.

[26] G. Zhao, S. Sonsaat, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Accent Conversion Using Phonetic Posteriorgrams," in *ICASSP*, 2018.

[27] P. Pramod, "Indian English Pronunciation," in *The Handbook of English Pronunciation*, M. Reed and J. Levis, Eds.: Wiley Blackwell, 2015, pp. 301-319.

[28] S.-A. Jun, "Prosody in sentence processing: Korean vs. English," *UCLA Working Papers in Phonetics,* vol. 104, pp. 26-45, 2005.

[29] M. Ueyama and S.-A. Jun, "Focus realization of Japanese English and Korean English intonation," *UCLA Working Papers in Phonetics,* pp. 110-125, 1996.

[30] J. Anderson-Hsieh, R. Johnson, and K. Koehler, "The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentais, prosody, and syllable structure," *Language learning,* vol. 42, no. 4, pp. 529-555, 1992.

[31] M. C. Pennington and N. C. Ellis, "Cantonese speakers' memory for English sentences with prosodic cues," *The Modern Language Journal,* vol. 84, no. 3, pp. 372-389, 2000.

[32] J. Chang, "Chinese speakers," *Learner English,* vol. 2, pp. 310-324, 1987.

[33] J. Morley, "Teaching American English Pronunciation," ed: JSTOR, 1993.

[34] B. Smith, *Learner English: A teacher's guide to interference and other problems*. Cambridge University Press, 2001.

[35] M. Benrabah, "Word-stress–a source of unintelligibility in English," *IRAL-International Review of Applied Linguistics in Language Teaching,* vol. 35, no. 3, pp. 157-166, 1997.

[36] K. De Jong and B. A. Zawaydeh, "Stress, duration, and intonation in Arabic word-level prosody," *Journal of Phonetics,* vol. 27, no. 1, pp. 3-22, 1999.

[37] Y. Cho and B. Bridgeman, "Relationship of TOEFL iBT® scores to academic performance: Some evidence from American universities," *Language Testing,* vol. 29, no. 3, pp. 421-442, 2012.

[38] H. Zen *et al.*, "The HMM-based speech synthesis system (HTS) version 2.0," in *SSW*, 2007, pp. 294-299.

[39] D. Erro, E. Navas, and I. Hernaez, "Parametric voice conversion based on bilinear frequency warping plus amplitude scaling," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 21, no. 3, pp. 556-566, 2013.

[40] *Audacity®*. Available: http://www.audacityteam.org/

[41] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: trainable text-speech alignment using Kaldi," in *Interspeech*, 2017, pp. 498-502.

[42] P. P. G. Boersma, "Praat, a system for doing phonetics by computer," *Glot international,* vol. 5, 2002.

[43] M. Munro, "How well can we predict L2 learners' pronunciation difficulties?," *CATESOL Journal,* vol. 30, no. 1, pp. 267-282, 2018.

[44] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication,* vol. 30, no. 2, pp. 95-108, 2000.

[45] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition & Understanding*, 2011.

[46] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206-5210.

[47] *Creative Commons Attribution-NonCommercial 4.0 International Public License*. Available: https://creativecommons.org/licenses/by-nc/4.0/legalcode