Summary

The L2-ARCTIC corpus is a speech corpus of non-native English that is intended for research in voice conversion, accent conversion, and mispronunciation detection. In total, the corpus contains 26,867 utterances from 24 non-native speakers with a balanced gender and L1 distribution. Most speakers recorded the full CMU ARCTIC set. The total duration of the corpus is 27.1 hours, with an average of 67.7 minutes (std: 8.6 minutes) of speech per L2 speaker. On average, each utterance is 3.6 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 238,702 word segments, giving an average of around nine (9) words per utterance, and over 851,830 phone segments (excluding silence).

Human annotators manually examined 3,599 utterances, annotating 14,098 phone substitutions, 3,420 phone deletions, and 1,092 phone additions.

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. We provide a list of those special cases in section **Notes**.

About the suitcase corpus (added on March 12, 2020)

This portion of the L2-ARCTIC corpus involves spontaneous speech. We include recordings and annotations from 22 of the 24 speakers who recorded the sentences. Speakers SKA and ASI did not participate in this task. Each speaker retold a story from a picture narrative used in applied linguistics research on comprehensibility, accentedness, and intelligibility. The pictures are generally known as the suitcase story. Each retelling of the narrative was done after looking over the story and asking the researchers questions about what was happening. Few participants had questions. The annotations were carried out by two research assistants trained in phonetic transcription. Each did half of the transcriptions, then checked the other half done by the other research assistant. Finally, all transcriptions were checked by John Levis, a co-PI for the project. This project was funded by National Science Foundation award 1618953, titled “Developing Golden Speakers for Second-Language Pronunciation.”

The total duration of this subset is 26.1 minutes, with an average of 1.2 minutes (std: 41.5 seconds) per speaker. Using the manual annotation results, we estimate a speech to silence ratio of 2.3:1 across the whole dataset. The dataset contains around 3,083 word segments, giving an average of 140 words per recording, and around 9,458 phone segments (excluding silence). The manual annotations include 1,673 phone substitutions, 456 phone deletions, and 90 phone additions.

Manual annotations

Each manually annotated TextGrid file will always have a "words" and "phones" tier while some of them may have an additional tier that contains comments from the annotators. Each phone segment was tagged in the "phones" tier following the conventions below,

- **Correctly pronounced**: The existing forced-alignment label was unchanged
- **Phone substitution error**: We changed the label to `CPL,PPL,s`, where `CPL` is the correct phoneme label (i.e., what should have been produced), `PPL` is the perceived phoneme label (i.e., what was actually produced), and `s` stands for substitution. If the perceived phoneme label was hard to judge, we used the `err` tag as its phoneme label, i.e., tagged it as `CPL,err,s`. If the perceived phoneme sounded like a deviation of the standard American English pronunciation, we marked it using a "deviation" symbol `*`. For

example, if the correct label for a phone segment is `AH`, and the speaker pronounced it as an `AO` with a foreign accent, we mark this error as `AH,AO*,s`

- **Phone addition error**: we created an empty interval in the "phones" tier, adjusted its boundaries, and changed its label to `sil,PPL,a`, where `sil` stands for silence, and `a` stands for addition (insertion). If the perceived phoneme label was not in the American English phoneme set, we used the `err` and `*` tags as noted above
- **Phone deletion error**: We found a silent segment where that phone segment should be, and annotated it as `CPL,sil,d`, where `d` stands for deletion

The "phones" tier will only contain ASCII symbols while the comments may contain UTF8 symbols (e.g., IPA notations).

Directory structure

Each speaker's data is organized in its subdirectory under the root folder; the root folder also contains a README.md file (this file), a README.pdf file (this files converted to PDF), a LICENSE file, and a PROMPTS file (containing the original text prompts from CMU ARCTIC). Each speaker's directory is structured as follows:

- `/wav`: Containing audio files in WAV format, sampled at 44.1 kHz
- `/transcript`: Containing orthographic transcriptions, saved in TXT format
- `/textgrid`: Containing phoneme transcriptions generated from forced-alignment, saved in TextGrid format
- `/annotation`: Containing manual annotations, saved in TextGrid format

The suitcase corpus is stored in a subfolder named `suitcase_corpus` under the root folder and follows a similar directory structure, except that we did not include the forced-aligned TextGrid files, since the manual annotations contain more accurate alignments. All files in the suitcase corpus are named by speaker codes.

File summary and speaker information

Speaker	Gender	Native Language	# Wav Files	# Annotations
ABA	M	Arabic	1129	150
SKA	F	Arabic	974	150
YBAA	M	Arabic	1130	149
ZHAA	F	Arabic	1132	150
BWC	M	Chinese	1130	150
LXC	F	Chinese	1131	150
NCC	F	Chinese	1131	150
TXHC	M	Chinese	1132	150
ASI	M	Hindi	1131	150
RRBI	M	Hindi	1130	150
SVBI	F	Hindi	1132	150

Speaker	Gender	Native Language	# Wav Files	# Annotations
TNI	F	Hindi	1131	150
HJK	F	Korean	1131	150
HKK	M	Korean	1131	150
YDCK	F	Korean	1131	150
YKWK	M	Korean	1131	150
EBVS	M	Spanish	1007	150
ERMS	M	Spanish	1132	150
MBMPS	F	Spanish	1132	150
NJS	F	Spanish	1131	150
HQTV	M	Vietnamese	1132	150
PNV	F	Vietnamese	1132	150
THV	F	Vietnamese	1132	150
TLV	M	Vietnamese	1132	150
Total			**26867**	**3599**

Phoneme set

Index	ARPAbet	Example	Annotation	Type
1	AA	odd	AA D	vowel
2	AE	at	AE T	vowel
3	AH	hut	HH AH T	vowel
4	AO	ought	AO T	vowel
5	AW	cow	K AW	vowel
6	AX	discus	D IH S K AX S	vowel
7	AY	hide	HH AY D	vowel
8	B	be	B IY	stop
9	CH	cheese	CH IY Z	affricate
10	D	dee	D IY	stop
11	DH	thee	DH IY	fricative
12	EH	Ed	EH D	vowel
13	ER	hurt	HH ER T	vowel

Index	ARPAbet	Example	Annotation	Type
14	EY	ate	EY T	vowel
15	F	fee	F IY	fricative
16	G	green	G R IY N	stop
17	HH	he	HH IY	aspirate
18	IH	it	IH T	vowel
19	IY	eat	IY T	vowel
20	JH	gee	JH IY	affricate
21	K	key	K IY	stop
22	L	lee	L IY	liquid
23	M	me	M IY	nasal
24	N	knee	N IY	nasal
25	NG	ping	P IH NG	nasal
26	OW	oat	OW T	vowel
27	OY	toy	T OY	vowel
28	P	pee	P IY	stop
29	R	read	R IY D	liquid
30	S	sea	S IY	fricative
31	SH	she	SH IY	fricative
32	T	tea	T IY	stop
33	TH	theta	TH EY T AH	fricative
34	UH	hood	HH UH D	vowel
35	UW	two	T UW	vowel
36	V	vee	V IY	fricative
37	W	we	W IY	semivowel
38	Y	yield	Y IY L D	semivowel
39	Z	zee	Z IY	fricative
40	ZH	seizure	S IY ZH ER	fricative

Notes

As we mentioned before, 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, we list all those special cases here,

- ABA: arctic_a0158, arctic_b0013, and arctic_b0398 were not recorded
- ASI: arctic_b0013 was not recorded; did not record the suitcase narrative
- BWC: arctic_b0345 was removed because the quality was bad, arctic_b0013 was not recorded
- EBVS: passing beyond arctic_b0408, we only recorded those L1-dependent sentences
- HJK: arctic_b0013 was not recorded
- HKK: arctic_b0013 was not recorded
- LXC: arctic_b0013 was not recorded
- NCC: arctic_b0013 was not recorded
- NJS: arctic_b0013 was not recorded
- RRBI: arctic_a0298 and arctic_b0013 were not recorded
- SKA: passing beyond arctic_b0358, we only recorded those L1-dependent sentences; did not record the suitcase narrative
- TNI: arctic_b0013 was not recorded
- YBAA: arctic_a0094 and arctic_b0013 were not recorded
- YDCK: arctic_b0013 was not recorded
- YKWK: arctic_b0013 was not recorded

Tips

- The rule of thumb is that as long as there exists an audio file for a sentence, there will be an accompanying orthographic transcription and a forced-aligned phoneme transcription
- The forced-alignment was generated by the Montreal Forced Aligner v1.0.0
- For words that were not included in the pronunciation dictionary we used, they were replaced by `<unk>` in the forced-aligned phoneme transcriptions
- The pronunciation error tags may include whitespace characters, and you may need to remove them (e.g., through a regular expression) before using the tags
- All audio files were sampled at 44.1 kHz. Thus, you may need to resample them for your application
- All TextGrid files were encoded in `UTF8`

Helpful toolkits

Here are some tools we have used to access TextGrid files,

- Praat: Visualizing and modifying TextGrid files in a GUI
- mPraat: Read/write TextGrid files in Matlab
- TextGridTools: Read/write TextGrid files in Python

Citation

For more details about the corpus, please refer to our Interspeech'18 paper. We encourage you to cite this paper if you used L2-ARCTIC in your publication or product.

```
@inproceedings{zhao2018l2arctic,
  author={Guanlong {Zhao} and Sinem {Sonsaat} and Alif {Silpachai} and
Ivana {Lucic} and Evgeny {Chukharev-Hudilainen} and John {Levis} and
Ricardo {Gutierrez-Osuna}},
  title={L2-ARCTIC: A Non-native English Speech Corpus},
  year=2018,
```

```
  booktitle={Proc. Interspeech},
  pages={2783--2787},
  doi={10.21437/Interspeech.2018-1110},
  url={http://dx.doi.org/10.21437/Interspeech.2018-1110}
}
```

Revision history

- 03/12/2020: `v5.0`, add the suitcase corpus, which contains un-scripted speech and corresponding annotations from 22 of the 24 speakers
- 06/06/2019: `v4.0`, refined annotations; changed most of the `err` tags to more specific error types
- 04/03/2019: `v3.0`, add 4 Vietnamese speakers to the corpus
- 09/28/2018: `v2.0`, add 10 new speakers to the corpus
- 03/26/2018: `v1.0`, the initial release

Disclaimer

We may recommend the use of software, information, products, or websites that are owned or operated by other parties. We offer or facilitate this recommendation by hyperlinks or other methods to aid your access to the third-party resource. While we endeavor to direct you to helpful, trustworthy resources, We cannot endorse, approve, or guarantee software, information, products, or services provided by or at a third-party resource or track changes in the resource. Thus, we are not responsible for the content or accuracy of any third-party resource or for any loss or damage of any sort resulting from the use of, or for any failure of, products or services provided at or from a third party resource. We recommend these resources on an “as is” basis. When you use a third-party resource, you will be subject to its terms and licenses and no longer be protected by our privacy policy or security practices, which may differ from the third policy or practices or other terms. You should familiarize yourself with any license or use terms of, and the privacy policy and security practices of, the third party resource, which will govern your use of that resource.