Developing a Benchmark for Pronunciation Feedback: Creation of a Phonemically Annotated Speech Corpus of isiZulu Language Learner Speech

Alexandra O'Neil, Nils Hjortnæs, Zinhle Nkosi, Thulile Ndlovu, Zanele Mlondo, Ngami Phumzile Pewa, and Francis Tyers

Indiana University and University of KwaZulu Natal

Table of Contents

Language Learning and Pronunciation

Language learners often struggle with pronunciation while learning new languages

- Current speech-feedback tools use proprietary software to provide word-level feedback
- Phoneme-level feedback is helpful for improving students' pronunciation (Engwall 2012)
- Corpora annotated for mistakes are required to test the effectiveness of pronunciation feedback tools (Phan, Grósz, and Kurimo 2023)

Why isiZulu?

- Large phonemic inventory, including phonemic aspiration contrast for stops, implosives, and 15 distinct clicks (Canonici 1989; Doke 1961)
- Requiring a model to distinguish between more phonemes holds a model to a more robust standard and phonemic categories can be condensed as needed
- Provides a case study for how this type of corpus could be made for another less-resourced language

Research Goal

Develop a corpus that can be used to test pronunciation feedback tool and contribute:

- The methodology and considerations for creating this type of corpus
- An open-source corpus that can be used by other researchers

Second Language Acquisition

- Perceptual understanding of a language is linked to production (pronunciation) of a language (Williams 1979)
- Learners must notice their mistakes before fixing them (Schmidt 1990)
- Existing corpora are usually written in the field (Granger 2011), which doesn't support mispronunciation detection

Automatic Speech Recognition (ASR)

- 3 available corpora for isiZulu Speech Recognition:
  - NCHLT isiZulu Speech Corpus with 56 hours of speech data (Barnard et al. 2014)
  - African Speech Technology isiZulu Speech Corpus of 399 transcribed phone calls (Roux, Louw, and Niesler 2004)
  - Lwazi Multilingual corpus with 525 minutes of audio (Barnard, Davel, and Van Heerden 2009)
- Use of ASR for Computer-Assisted Pronunciation Training (CAPT) is an emerging research topic (Phan, Grósz, and Kurimo 2023; Peng et al. 2021; Li et al. 2023)

Sentence Selection

- 803 sentences from *"Elementary Zulu: A Course of Elementary Lessons in the Zulu Language: Intended Chiefly for Beginners and Junior"* (W 1921)
- Orthography was standardized using regular expressions and archaic terms were removed
- Information about learning order is preserved in the corpus
- 20 common learner phrases and 8 complex phoneme phrases were added in consult with isiZulu teacher (831 total in corpus)

Recording

- Sentences were randomized and split into blocks of 50, 8 complex phrases were left in final block
- Students sight-read sentences in the block and errors and false starts are preserved in the recordings
- Audacity was used to record, split, and easily rename files
- Default audio processing settings had to be turned off on laptops to prevent automatic dampening of clicks

Processing

- Clips were normalized to facilitate playback at a consistent level for teachers
- Further processing was limited to ensure no accidental filtering of speech sounds, particularly clicks

Organization

- Recording on separate tracks allowed for export of individual sentence clips and easy renaming (see paper appendix for details)
- Filename preserves elicitation order, learning sequence, and speaker ID
  - Example: 101-9-001.wav

Participants

- 12 students (8:4) and 3 teachers (2:1) from the University of KwaZulu-Natal were recorded
- Completed 1-3 semesters of isiZulu
- L1 speakers of English, siSwati, and isiXhosa

ID	L1	Gender	Age	Semesters
1	English	F	19	3
2	English	F	19	3
3	English	F	20	3
4	siSwati	F	18	1
5	siSwati	F	19	1
6	isiXhosa	M	21	1
7	isiXhosa	M	21	1
8	English	F	19	3
9	English	F	18	1
10	English	M	19	1
11	English	F	19	3
12	English	M	19	3
101	isiZulu	F	45	–
102	isiZulu	F	33	–
103	isiZulu	M	30	–

Annotation Layers

Awaphume amahhashi.

	a	w	a	ph	u	m	e		a	m	a	hh	a	sh	i
Sounds:				x											
Tones:										x					
Insertion left:	<<	<<	<<	<<	<<	<<	<<	<<	<<	<<	<<	<<	<<	<<	<< >>

0:00 / 0:05

Submit

Skip

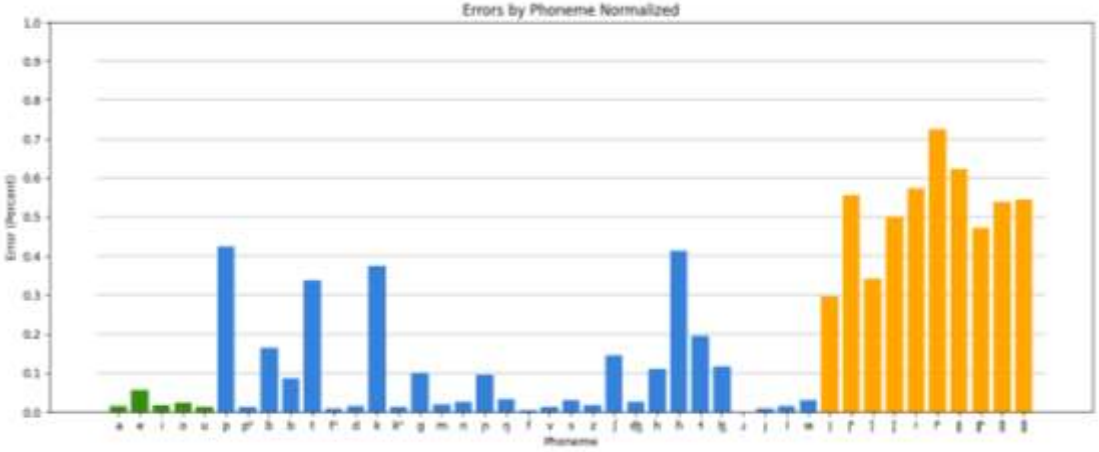
Mispronunciation Detection

Common errors marked by teachers:

- Unaspirated plosives were often mispronunced
- Vowels had the lowest error rate with e being the most problematic
- Clicks were the most commonly mispronounced sounds

Future analysis by SLA research and breaking down by participant background would be interesting, but isn't the purpose of this corpus

Mispronunciation Detection



Inter-Annotator Agreement

- Corpus contained 9,626 sentence recordings with 169,074 phonemes
- At least two annotations are available for 71,370 phonemes
- Calculated using Krippendorf's Alpha (Krippendorff 2011) as we had partial annotation
- IAA was calculated to be .585

Conclusion and Next Steps

- Corpus is published on South African Centre for Digital Language Resources (SADiLaR) website
- Code for annotation tool available on project's Github
- Next step: Follow methodology of other CAPT projects to use the corpus to evaluate current methods of automatic phoneme mispronunciation detection

References I

Artstein, Ron and Massimo Poesio (2008). "Survey Article: Inter-Coder Agreement for Computational Linguistics". In: *Computational Linguistics* 34.4, pp. 555–596. DOI: 10.1162/coli.07-034-R2. URL: https://aclanthology.org/J08-4004.

Barnard, Etienne, Marelie Davel, and Charl Van Heerden (2009). "ASR corpus design for resource-scarce languages". In: ISCA.

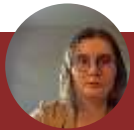
Barnard, Etienne et al. (2014). "The NCHLT speech corpus of the South African languages". In: Workshop Spoken Language Technologies for Under-resourced Languages (SLTU).

Canonici, Noverino N. (1989). *Imisindo Yesizulu: A Simple Introduction to Zulu Phonology*. Durban: Department of Zulu Language and Literature, University of Natal.

Doke, Clement M. (1961). *Textbook of Zulu Grammar*. 6th edition (1st edition 1927). Cape Town: Longmans.

Engwall, Olov (2012). "Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher". In: *Computer Assisted Language Learning* 25.1, pp. 37–64.

Granger, Sylviane (2011). "How to use foreign and second language learner corpora". In: *Research methods in second language acquisition: A practical guide*, pp. 5–29.



References II

Krippendorff, Klaus (2011). "Computing Krippendorff's alpha-reliability". In.

Landis, J Richard and Gary G Koch (1977). "The measurement of observer agreement for categorical data". In: *biometrics*, pp. 159–174.

Li, Jing et al. (2023). "Enhancing Whisper Model for Pronunciation Assessment with Multi-Adapters". In: *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1955–1959. DOI: 10.1109/APSIPAASC58517.2023.10317374.

Peng, Linkai et al. (Aug. 2021). "A Study on Fine-Tuning wav2vec2.0 Model for the Task of Mispronunciation Detection and Diagnosis". en. In: *Interspeech 2021*. ISCA, 4448–4452. DOI: 10.21437/Interspeech.2021-1344. URL: https://www.isca-archive.org/interspeech_2021/peng21e_interspeech.html.

Phan, Nhan, Tam´as Gro´sz, and Mikko Kurimo (2023). "CaptainA-A mobile app for practising Finnish pronunciation". In: *The 24rd Nordic Conference on Computational Linguistics*.

Roux, J. C., P. H. Louw, and T. R. Niesler (May 2004). "The African Speech Technology Project: An Assessment". In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Ed. by Maria Teresa Lino et al. Lisbon, Portugal: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2004/pdf/445.pdf.

References III

Schmidt, Richard W (1990). "The role of consciousness in second language learning1". In: *Applied linguistics* 11.2, pp. 129–158.

W, M. F. (1921). *Elementary Zulu: A Course of Elementary Lessons in the Zulu Language: Intended Chiefly for Beginners and Junior Pupils*. en. Google-Books-ID: Wvw0AQAAMAAJ. Juta.

Williams, Lee (1979). "The modification of speech perception and production in second-language learning". In: *Perception & Psychophysics* 26.2, pp. 95–104.

Questions?