AnglistikVoices: an L2 English speech dataset for educational and technological advancement in speech technology

Akhilesh Kakolu Ramarao and Anna Sophia Stein

Heinrich Heine University

Speech science plays a pivotal role in the development of speech technology. Recently, there is a surge in large-scale Automatic Speech Recognition (ASR) models (e.g., Whispers and Wav2Vec) claiming to be the “State of the Art”. These advances, however, have created new challenges for educators and researchers in terms of inclusivity and diversity of accents. While there has been a significant effort to provide diverse sets of speech data for ASR development, notable performance discrepancies persist in different underrepresented L1-Englishes [1, 2], and L2-accented Englishes [3]. Such insufficient development of ASR excludes technological access for individuals of marginalized groups and accents. **Data representativeness**: There is a lack of speech corpora that focus on L2 accented speech, with the notably exception of the Wildcat Corpus [4]. Another significant initiative is the ArtiBias Corpus [5], which is a cleaned, expert-transcribed subset of the popular Mozilla Common Voice dataset [6]. Nonetheless, these speech corpora lack detailed linguistic profiles of the speakers and, in addition, are mostly crowd-sourced, leading to qualitative differences in comparison to lab-recorded speech. Furthermore, creating a corpus only addresses part of the issue in biased ASR. To fully address the biases, we must understand the detailed phonetic and phonological make-up of the different accents that led to uneven performance differences. **Gap in education**: The need for technological education is underlined by the recently passed European Union’s General Data Protection Regulation (GDPR) and the EU AI ACT which emphasizes the importance of transparency, accountability, and appropriate education on AI among users. The CARE (Collaborative, Active, Research-focused, Educational) approach, proposed by [7] provides a framework for research-based teaching and was originally introduced for active-learning seminars in phonology and phonetics. This holistic framework engages students in collaborative research projects, encompassing data collection, analysis, and hypothesis testing while fostering collabo-ration among students and faculty and simulating real research. **Our study**: To address both the educational and the technological need for understanding and developing unbiased ASR, we present the creation of the **AnglistikVoices** corpus built using the CARE approach, currently containing around 150 mins and around 1200 sentences of read L2 English speech.

Methods: Following the success of the CARE approach, this corpus was developed during a 14-weeks bachelor-level course at the English and American studies department at [ANON]. After being introduced to the basic ASR concepts, the students conducted recordings at the in-house lab facility in groups, acting as both experimenters, and participants and evaluated their own recorded speech against different ASRs. As a participant, they read 60 stimuli from the Arti-bias corpus [5], while the experimenter recorded the sentences. We setup the Huggingface spaces to host whisper.en (tiny, medium) [1] models and DeepSpeech [8] models (v0.7.3, v0.9.3), which enabled students to generate transcriptions from their own recordings. The students not only used quantitative measures (such as WER, CER and so on) but also investigated how the ASR errors are predictable from the phonetic/phonological features of their L1 language.

Conclusion: Our research contributes towards demystifying the black box of ASR models highlighting the source of the errors using L1 and L2 phonetics and phonology. We provide a proof of concept of how to simultaneously tackle both educational and technological concerns in speech technology using research-based teaching. In the future, we aim to expand this corpus in other classes in the area of phonetics/phonology and technology using this educational framework.

References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learn-ing*. PMLR, 2023, pp. 28492–28518.

[2] C. Graham and N. Roll, "Evaluating openai's whisper asr: Performance analysis across diverse accents and speaker traits," *JASA Express Letters*, vol. 4, no. 2, 2024.

[3] S. Hollands, D. Blackburn, and H. Christensen, "Evaluating the Performance of State-of-the-Art ASR Systems on Non-Native English using Corpora with Extensive Language Background Variation," in *Proc. Interspeech 2022*, 2022, pp. 3958–3962.

[4] A. R. Bradlow, R. E. Baker, A. Choi, M. Kim, and K. J. Van Engen, "The wildcat corpus of native-and foreign-accented english," *Journal of the Acoustical Society of America*, vol. 121, no. 5, p. 3072, 2007.

[5] J. Meyer, L. Rauchenstein, J. D. Eisenberg, and N. Howell, "Artie bias corpus: An open dataset for detecting demographic bias in speech applications," in *Proceedings of the twelfth language resources and evaluation conference*, 2020, pp. 6462–6468.

[6] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[7] C. Bjorndahl and M. Gibson, "The care approach to incorporating undergraduate research in the phonetics/phonology classroom," *Language*, vol. 98, no. 1, pp. e1–e25, 2022.

[8] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.