AnglistikVoices:
A student-sourced L2 English speech dataset

Akhilesh Kakolu Ramarao, Anna Sophia Stein
kakolura@hhu.de, anna.stein@hhu.de

1. Problem

1. Limited L2 English accented speech corpora:

- ➔ **Notable exceptions**: Wildcat [1] and ArtieBias [2] corpus
- ➔ **Issues**:
  - ◆ Lack detailed linguistic profile
  - ◆ Predominantly crowd-sourced, affecting quality
- ➔ Few state-of-the-art speech to text models tested on L2 English speech [3]

2. Educational Gap in AI and Technology:

- ➔ **Highlighted by**: EU Regulations like AI ACT [4]
- ➔ **Importance**: Transparency and education in AI
- ➔ **Impact**: Inclusion of minority accents and groups

2. Approach

Create a corpus in a seminar setting to tackle both problems simultaneously:

- ➔ Designed after the CARE (Collaborative, Active, Research-focused, Educational) approach [5]
- ➔ Focused on building an English L2 speech dataset and evaluating the current speech recognition systems against it
- ➔ Students gain experience in conducting phonetic/phonology laboratory experiments and a basic understanding of automatic speech recognition (ASR)

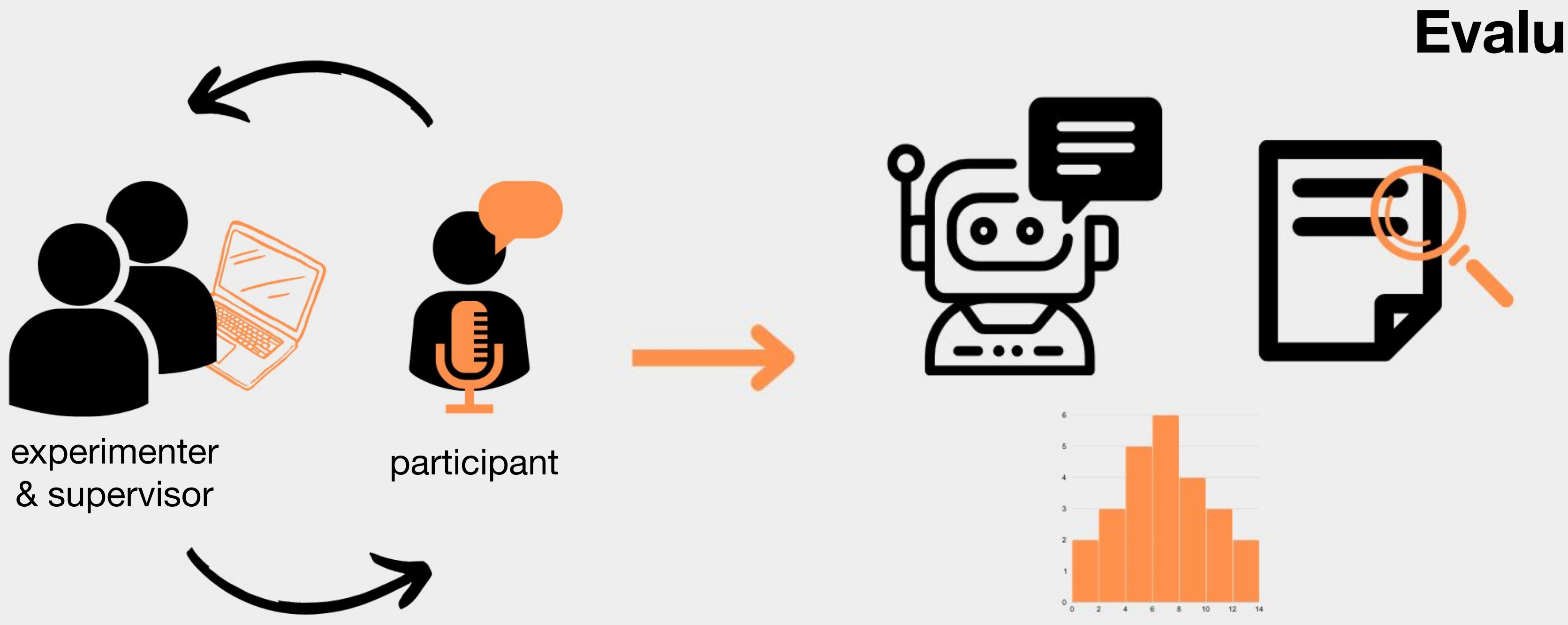
Course setting:

- ➔ 2nd year Bachelor-level seminar in the English studies department
- ➔ Once per week over the course of one semester (14 weeks)
- ➔ Using both laboratory and classroom settings
- ➔ No technical background was assumed

3. Methods

Corpus creation

- ➔ Groups of three students
- ➔ Each student takes each role once
- ➔ Participants records stimuli from ArtieBias corpus
- ➔ The experiment was conducted in a phonetics lab using Audacity [6]
- ➔ Manual sentence-level alignment



Evaluation

- ➔ Students transcribed audio using ASR models hosted on Huggingface
- ➔ Model transcriptions were manually evaluated using Word Error Rate as a metric and analysed for potential source of errors
- ➔ Student's findings replicate the findings of [3] that modern ASR models struggle with different

4. Corpus

- ➔ 20 speakers, 1200 stimuli (sentences), 60 sentences per participant
- ➔ Around 150 minutes of read L2 English speech
- ➔ License: CC-BY-4-SA
- ➔ Metadata collected:
  - ◆ **Ages**: 19-30
  - ◆ **Gender**: 14 female, 6 prefer not to say
  - ◆ **Highest level of education**: A-levels, BA, diploma
  - ◆ **Native languages**: Albanian, German, Vietnamese, Lingala, French, Romania, Greek, Russian, Kannada
  - ◆ **Other languages**: English, Spanish, French, Japanese, German, Mandarin, Italian, Hindi
  - ◆ **Ages of acquisition for each language**: 2-21
  - ◆ **Other data**: Primary source of English education, secondary/ other sources, Scores on official English tests (TOEFL, OOPT, …), time spent in English-speaking country, country where they grew up

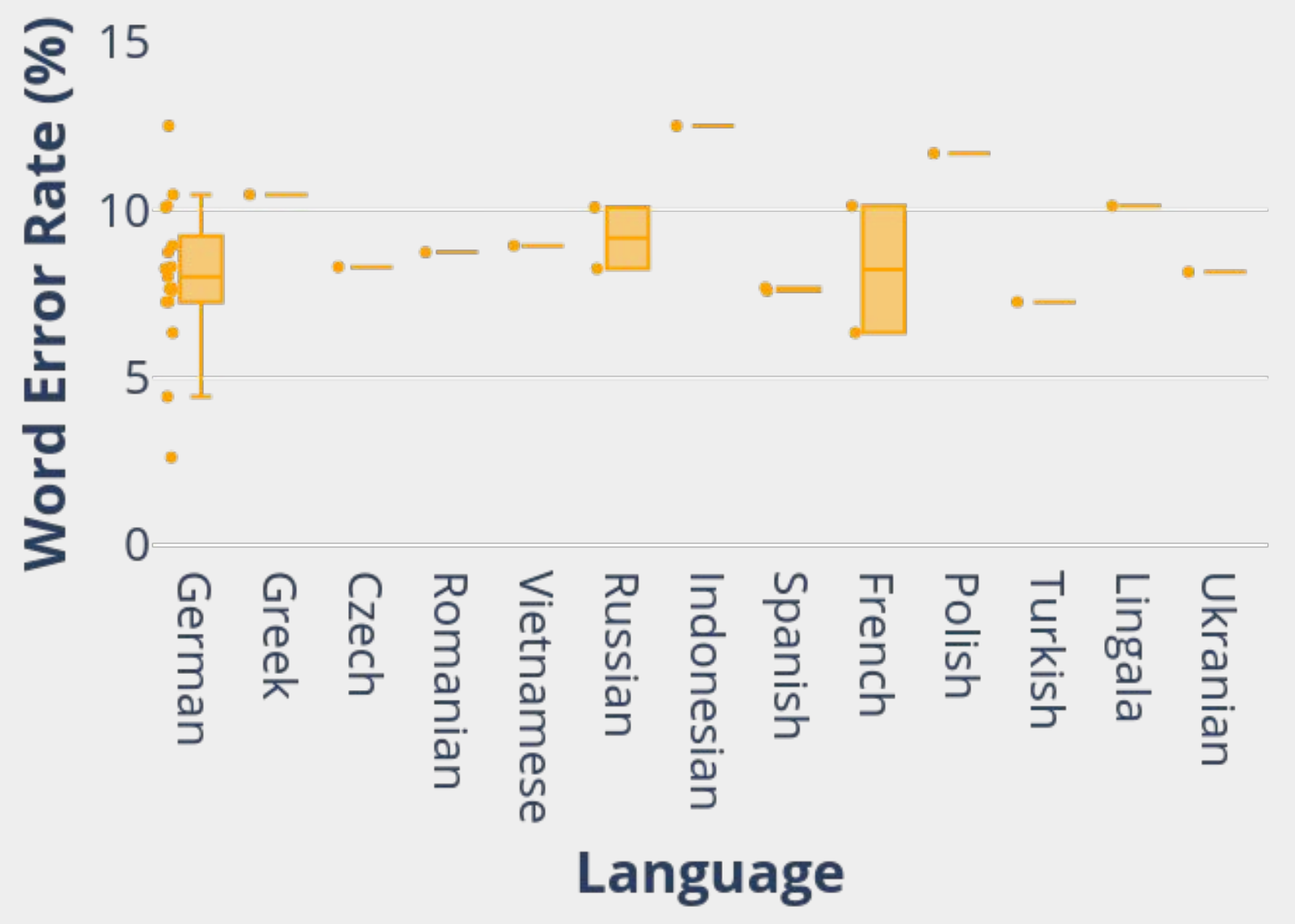
5. Analysis

Model Analysis:

Word Error Rate (WER) is the ratio of incorrectly recognized words to the total number of words spoken

Whisper:

- ➔ version: medium.en
- ➔ Trained on 680,000 hours of multilingual speech fine-tuned on English
- ➔ Whereas, the WER (%) for standard American English is 6.08%



Student Feedback:

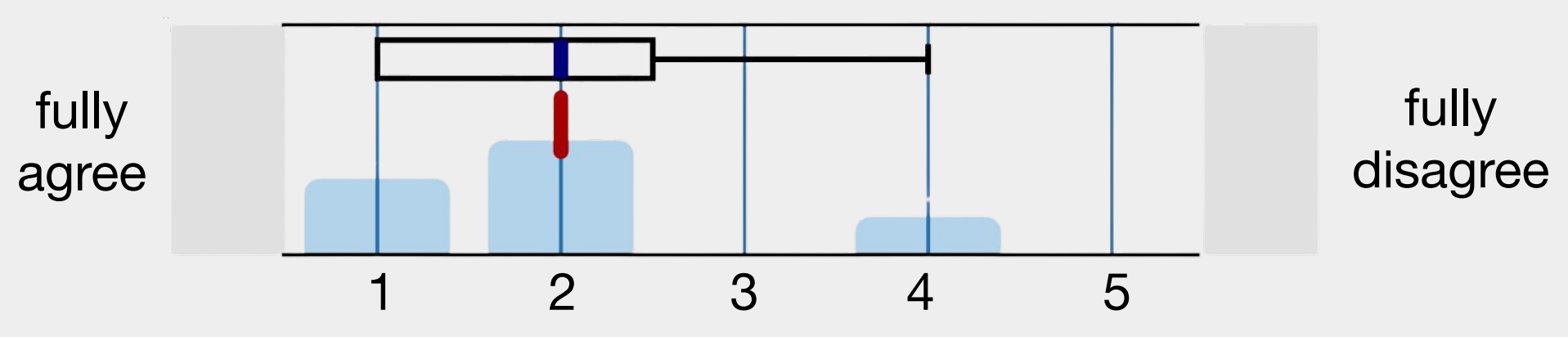
What makes learning in this course work well:

- ➔ "You can only learn something if you come to the seminar"
- ➔ "Lots of practical application and reading and writing papers, exciting topic"

Which suggestions for improvement do you have?

- ➔ "Material about different accents should be bigger focus of the lecture"
- ➔ "Less content and better development of the content that is really relevant"

"The demands made in the course are, in my opinion, attainable."



6. Conclusion

- ➔ Proof of concept for simultaneously tackling both educational and technological concerns in speech technology
- ➔ Students wrote term paper to report and reflect on research results
- ➔ Our teaching approach fosters understanding of technological advancements and limitations of AI
- ➔ State-of-the-art ASR models struggle with accented speech
- ➔ Whisper performs better than Deepspeech across English accents
- ➔ License allows the corpus to be extended in future iterations of this class architecture

Acknowledgements

We would like to thank the following people:

- ➔ Students who participated in the seminar
- ➔ Slamlab (https://slam.phil.hhu.de/) for providing the lab facility
- ➔ Prof. Dr. Kevin Tang and the practice audience at the English Language and Linguistics department at HHU

References

[1] 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.
[2] 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.
[3] 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.
[4] "The Act texts | EU Artificial Intelligence Act." https://artificialintelligenceact.eu/the-act/
[5] 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.
[6] Audacity Team. Audacity. Version 3.2.4, 2024. https://www.audacityteam.org/