university of koblenz
continue discovering

Transcribing Diverse Voices: Using Whisper for ICE corpora

Interspeech 2025 - Rotterdam
JProf Dr Andreas Weilinghoff

01 Intro and previous research

01 Intro

university of koblenz
continue discovering

‘Traditional‘ approach in English Linguistics (and other disciplines):

RQs → data collection → manual data transcription → data annotation & processing → data analysis → results

- work-intensive 🛠️
- time-consuming 🕰️

→ **costly** 💸

- tedious 😫

01 Intro

university of koblenz
continue discovering

... the higher the audio quality

... the more structured the speech

... the more ‘standard‘ the speech

... the less speakers involved

... the better

spoken corpus data

sociolinguistic data

World Englishes

(Jurafsky and Martin 2023: 331)

01 Intro – OpenAI Whisper

OpenAI Whisper

Radford et al. 2022

- end-to-end transformer architecture with encoder and decoder blocks
- trained on 680,000 hours of speech via unsupervised learning
- multilingual in 96 languages
- open source

→ "Whisper" appears 56 times in presentation titles at Interspeech 2025

Python script
whisper_to_textgrid.py
(Weilinghoff 2023)

01 Research aims and research questions

→ identify strengths/weaknesses of Whisper for spoken corpus transcription

→ integrate Whisper efficiently in spoken corpus data transcription workflows

RQ1 What is the **transcription accuracy** of different Whisper models for the corpora ICE Nigeria & ICE Scotland?

RQ2 Which variables have a **significant influence on ASR performance**?

01 Previous research – ASR in general

"Speech is easier to recognize if the speaker is speaking the same dialect or variety that the system was trained on" (Jurafsky and Martin 2023: 331)

- ASR bias towards → non-native speakers (e.g. Knill et al. 2018; Graham and Roll 2024)
  → regional accents (Tatman 2017; Markl 2022)
  → racial minority groups (Koenecke et al. 2020)

- influence of gender → better performance for female speakers (Adda-Decker and Lamel 2005; Goldwater et al. 2010)

01 Previous research - Whisper

university of koblenz
continue discovering

Whisper evaluation: (Graham and Roll 2024)

- L1 varieties: → best performance on L1 North American English

  → worse performance on British and Australian accents

  (some L2 Swedish and German accents better than some British accents; e.g. Leeds)

- worse performance on L2 varieties overall; higher English experience and pronunciation accuracy lead to better ASR performance
- worse performance on male speakers
- worse performance on spontaneous speech

02 Data and Method

02 Data and Method

university of koblenz
continue discovering

RQ1 What is the **transcription accuracy** of different Whisper models for the corpora ICE Nigeria & ICE Scotland?

ICE Nigeria (Wunder et al. 2008)

- postcolonial outer-circle variety
- compilation 2007-2013
- manually transcribed spoken component (time-aligned)

Extraction:
- 60 sound files | 12 speech categories

→ 13:05:47 hours | 94,499 words

ICE Scotland (Schützler et al. 2017)

- inner-circle variety (not GA or SSBE)
- compilation 2014-2020
- manually transcribed spoken component (time-aligned)

Extraction:
- 60 sound files | 12 speech categories

→ 11:50:31 hours | 111,418 words

02 Data and Method

corpus	file_name	file_duration	word_count
ICE Nigeria	bdis_01	00:12:47	2143
ICE Nigeria	bdis_02	00:07:46	1165
ICE Nigeria	bdis_03	00:03:23	587
ICE Nigeria	bdis_04	00:07:58	1296
ICE Nigeria	bdis_05	00:01:16	201
ICE Nigeria	bnew_01	00:05:24	555
ICE Nigeria	bnew_02	00:09:07	1143
ICE Nigeria	bnew_03	00:16:27	1473
ICE Nigeria	bnew_04	00:15:24	1231
ICE Nigeria	bnew_05	00:12:54	887
ICE Nigeria	btal_01	00:08:17	1056
ICE Nigeria	btal_02	00:02:51	503
ICE Nigeria	btal_03	00:01:46	193
ICE Nigeria	btal_04	00:08:59	1198
ICE Nigeria	btal_05	00:04:28	708
ICE Nigeria	leg_02	00:23:27	3979
ICE Nigeria	leg_04	00:15:59	2352
ICE Nigeria	leg_11	00:06:19	1212
ICE Nigeria	leg_08	00:02:44	586
ICE Nigeria	leg_09	00:03:59	790
ICE Nigeria	nbtal_01	00:16:55	1536
ICE Nigeria	nbtal_02	00:06:11	521
ICE Nigeria	nbtal_03	00:21:40	2346
ICE Nigeria	nbtal_04	00:26:56	3409
ICE Nigeria	nbtal_05	00:19:25	2391
ICE Nigeria	parl_01	00:07:53	1069
ICE Nigeria	parl_02	00:07:47	1089
ICE Nigeria	parl_03	00:11:16	1350
ICE Nigeria	parl_04	00:16:21	2012
ICE Nigeria	parl_05	00:12:06	2327
…	…	…	…

corpus	file_name	file_duration	word_count
ICE Scotland	bdis_01 (s1)	00:08:53	470
ICE Scotland	bdis_02	00:20:45	3030
ICE Scotland	bdis_03	00:06:00	1115
ICE Scotland	bdis_04	00:13:58	2964
ICE Scotland	bdis_05	00:11:56	2914
ICE Scotland	bnew_01	00:02:14	159
ICE Scotland	bnew_02 (s1)	00:02:48	93
ICE Scotland	bnew_03 (s1)	00:01:39	96
ICE Scotland	bnew_04 (s1)	00:03:36	179
ICE Scotland	bnew_05	00:01:47	305
ICE Scotland	btal_01	00:02:37	415
ICE Scotland	btal_02	00:02:34	453
ICE Scotland	btal_03	00:03:24	473
ICE Scotland	btal_04	00:02:52	379
ICE Scotland	btal_05	00:07:51	934
ICE Scotland	leg_01	00:19:08	2033
ICE Scotland	leg_02	00:22:32	2168
ICE Scotland	leg_03	00:02:29	324
ICE Scotland	leg_04	00:10:39	1333
ICE Scotland	leg_05	00:05:04	713
ICE Scotland	nbtal_01	00:21:55	3040
ICE Scotland	nbtal_02	00:30:00	4835
ICE Scotland	nbtal_03	00:11:17	1739
ICE Scotland	nbtal_04	00:04:45	713
ICE Scotland	nbtal_05	00:02:31	387
ICE Scotland	parl_01	00:20:54	3782
ICE Scotland	parl_02	00:20:09	3427
ICE Scotland	parl_03	00:11:31	1776
ICE Scotland	parl_04	00:25:21	4178
ICE Scotland	parl_05	00:36:08	5900
…	…	…	…

→ different varieties
→ different file sizes
→ different speech forms
→ monologues and dialogues
→ different speaker groups
→ different quality

02 Data and Method

RQ1 What is the **transcription accuracy** of different Whisper models for the corpora ICE Nigeria & ICE Scotland?

- retrieval of audio files and reference transcriptions (→ plain .txt)
- re-transcription of files with Whisper models (tiny, base, small, medium, large_v2, large_v3) via AMD EPYC 7402 processor

- normalization and comparison of manual reference transcription and Whisper transcriptions via **Word Error Rate (WER)** using werpy library (Armstrong 2024) via Python script

$$WER = \frac{S + D + I}{N}$$

https://www.andreas-weilinghoff.com/#code

02 Data and Method

university of koblenz
continue discovering

RQ2 Which variables have a **significant influence on ASR performance**?

- annotation for metadata (corpus, text category, model, sound quality, speaker number, gender, file duration)

- following approach of Graham and Roll (2024):

→ linear mixed effects modelling of WER with lme4 (Bates et al. 2015) and lmerTest (Kuznetsova et al. 2017) packages in R (R core team 2024)

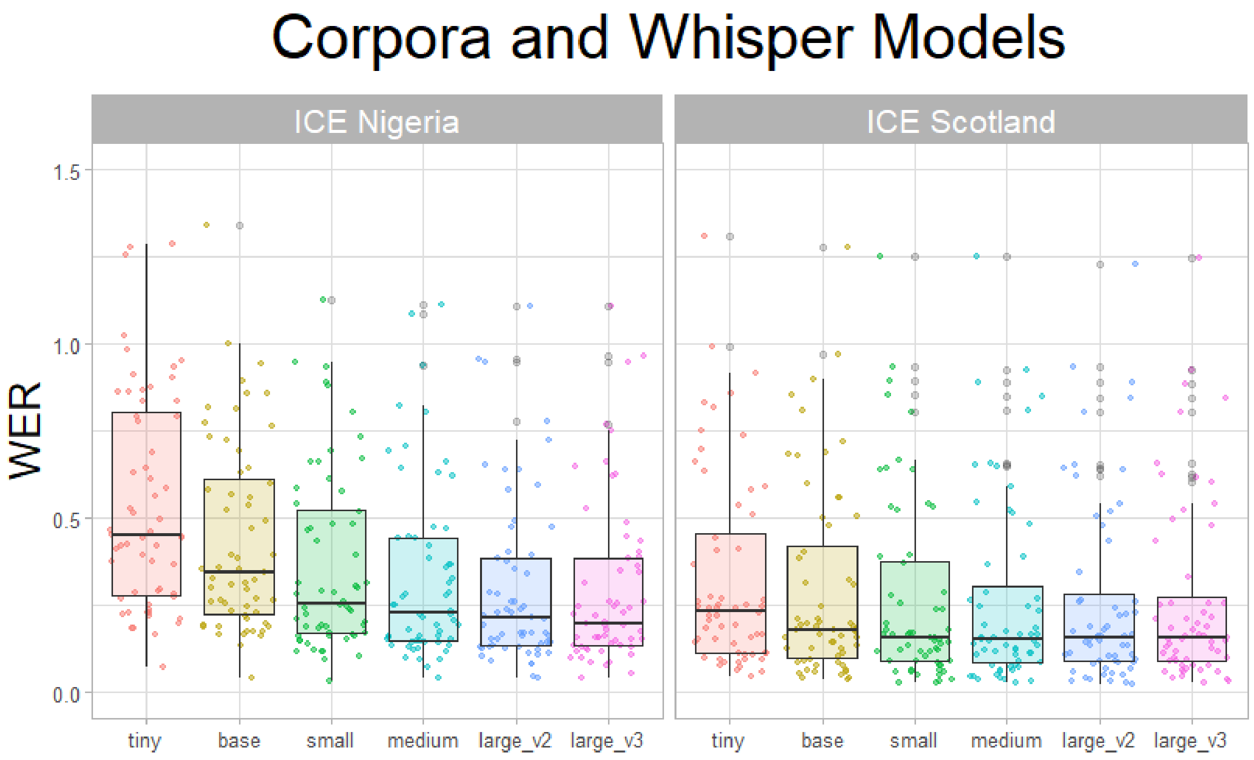
RANDOM FACTORS	TYPE	LEVELS
sound file	categorical	120 individual sound files
FIXED FACTORS	**TYPE**	**LEVELS**
corpus	categorical	ICE Nigeria, ICE Scotland
text category	categorical	bdis, bnew, btal, btran, com, cr, dem, leg, les, nbtal, parl, unsp
model	categorical	tiny, base, small, medium, large_v2, large_v3
quality_2	categorical	okay, bad
speaker number binary	categorical	mono, poly
gender	categorical	female, male, mixed
file duration (min)	numerical	1-48

03 Findings

03 Findings

university of koblenz
continue discovering

RQ1 What is the **transcription accuracy** of different Whisper models for the corpora ICE Nigeria & ICE Scotland?



Whisper model	ICE Nigeria		ICE Scotland	
	mean WER	std dev	**mean WER**	std dev
tiny	**0.54**	0.30	**0.32**	0.28
base	**0.45**	0.30	**0.29**	0.27
small	**0.36**	0.26	**0.27**	0.27
medium	**0.33**	0.25	**0.26**	0.27
large_v2	**0.30**	0.24	**0.26**	0.27
large_v3	**0.29**	0.24	**0.26**	0.26

03 Findings

RQ2

Which variables have a **significant influence on ASR performance**?

Extremely high R^2 values for best model:

```
wer ~ (model * corpus) + (model * quality_2) + text_category + speaker_number_binary +
gender_simplified + (1 | file_name)
```

Marginal R^2: 0.72

Conditional R^2: 0.95

03 Findings

university of koblenz
continue discovering

RQ2 Which variables have a **significant influence on ASR performance**?

```
wer ~ (model * corpus) + (model * quality_2) + text_category + speaker_number_binary + gender_simplified + (1 | file_name)
```

Significant factors:

model → 10%, 21%, 25%, 30% decrease of WER with model size

corpus → 11% decrease of WER for ICE Scotland

quality → 24% decrease of WER for good quality audio

text_category → increase of WER for text categories: *com, cr, dem, leg, les, unsp*

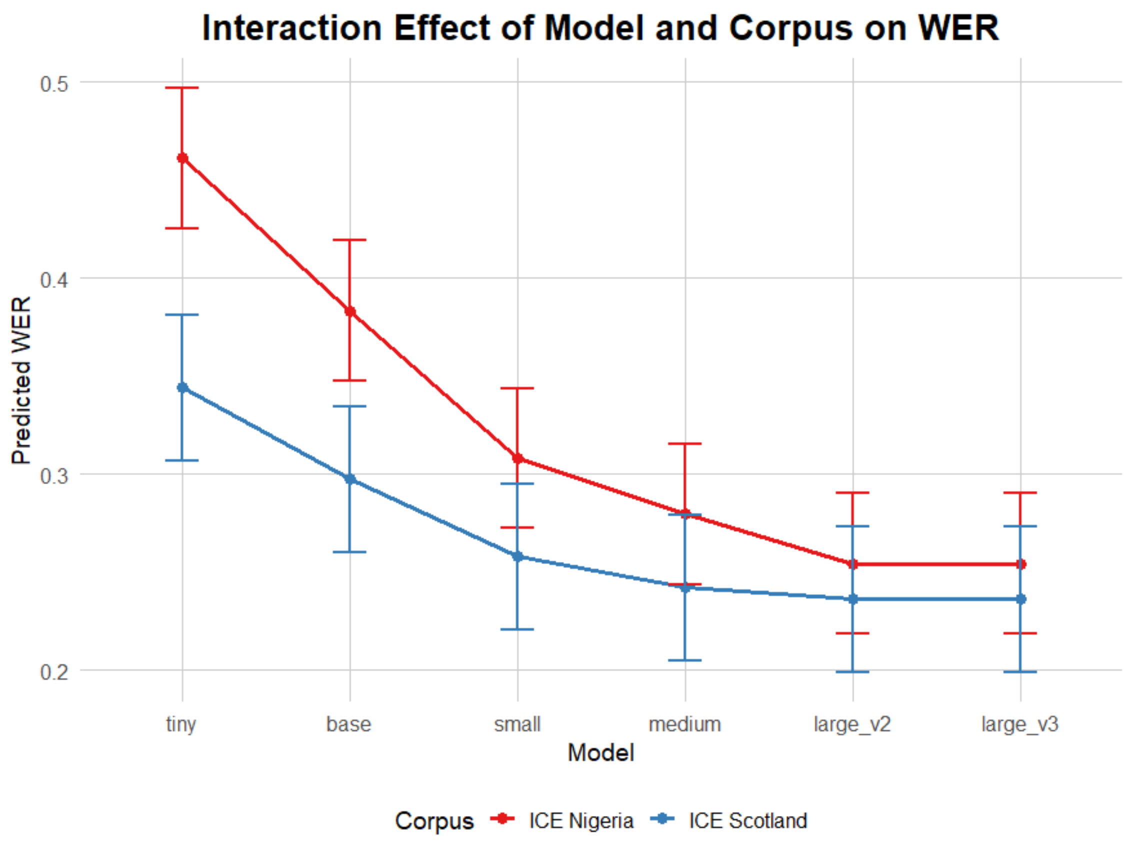
speaker_number → 19% increase of WER for audio files with several speakers

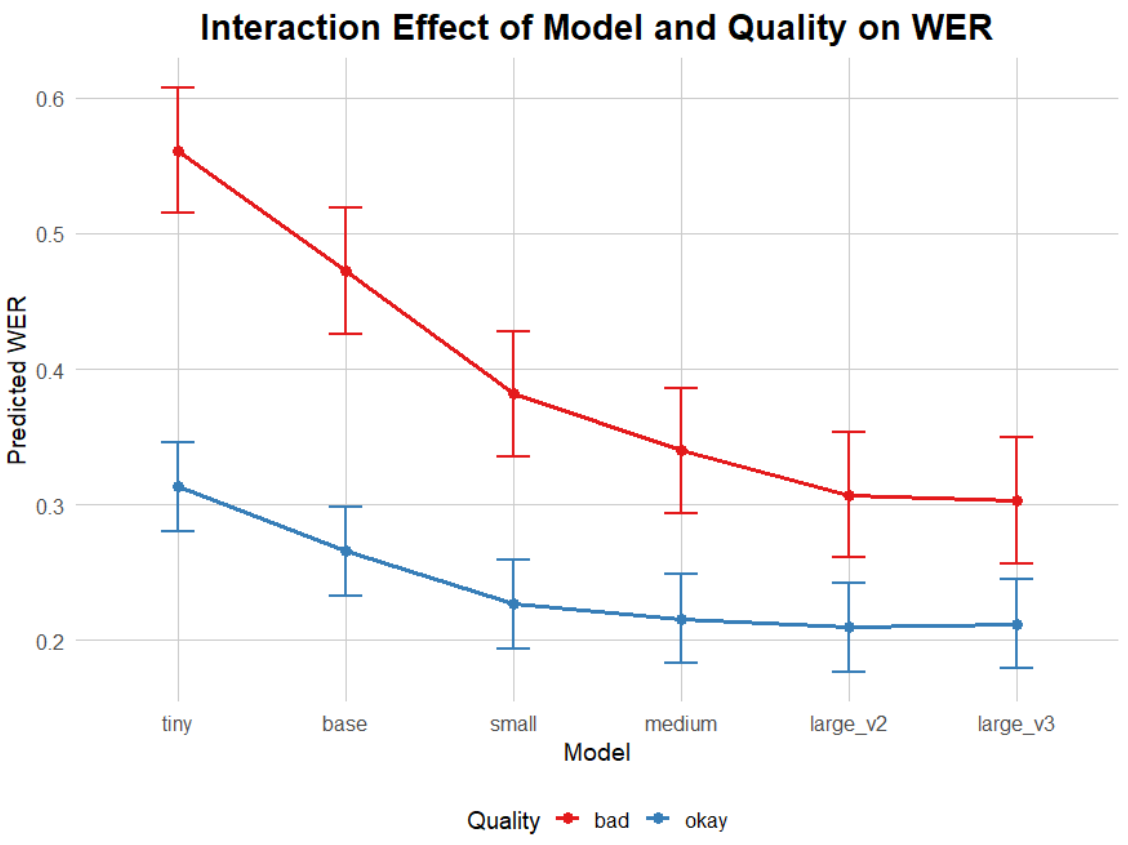
gender → 8% increase of WER for audio files with male speakers

03 Findings

university of koblenz
continue discovering

RQ2 Which variables have a **significant influence on ASR performance**?





04 Discussion

04 Discussion

university of koblenz
continue discovering

Hallucinations for specific files across different models

→ bad quality audio
→ long periods of silence
→ speaker overlaps / interruptions
→ switch to Nigerian Pidgin English

Whisper small	Whisper base
At UJ, he cautioned supervisors on the need to follow the enumerators and ensure that proper enumeration is affected. It doesn't mean once we are a supervisor, you find yourself, you wait linearly, looking for help now. Look at some people walking, are they doing the right thing? That's what you think. That's what you think. That's what you think. That's what you think. That's what you think. That's what you think. That's what you think.	At UJ, he cautioned supervisors on the need to follow the enumerators and ensure that proper enumeration is affected. It doesn't move walls here, it's super special, you find yourself with Legally, you can go out now. Look at some people walking, are they doing the right thing? That's what you do. That's what you do. That's what you do. That's what you do. That's what you do. That's what you do. That's what you do. That's what you do.

04 Discussion

university of koblenz
continue discovering

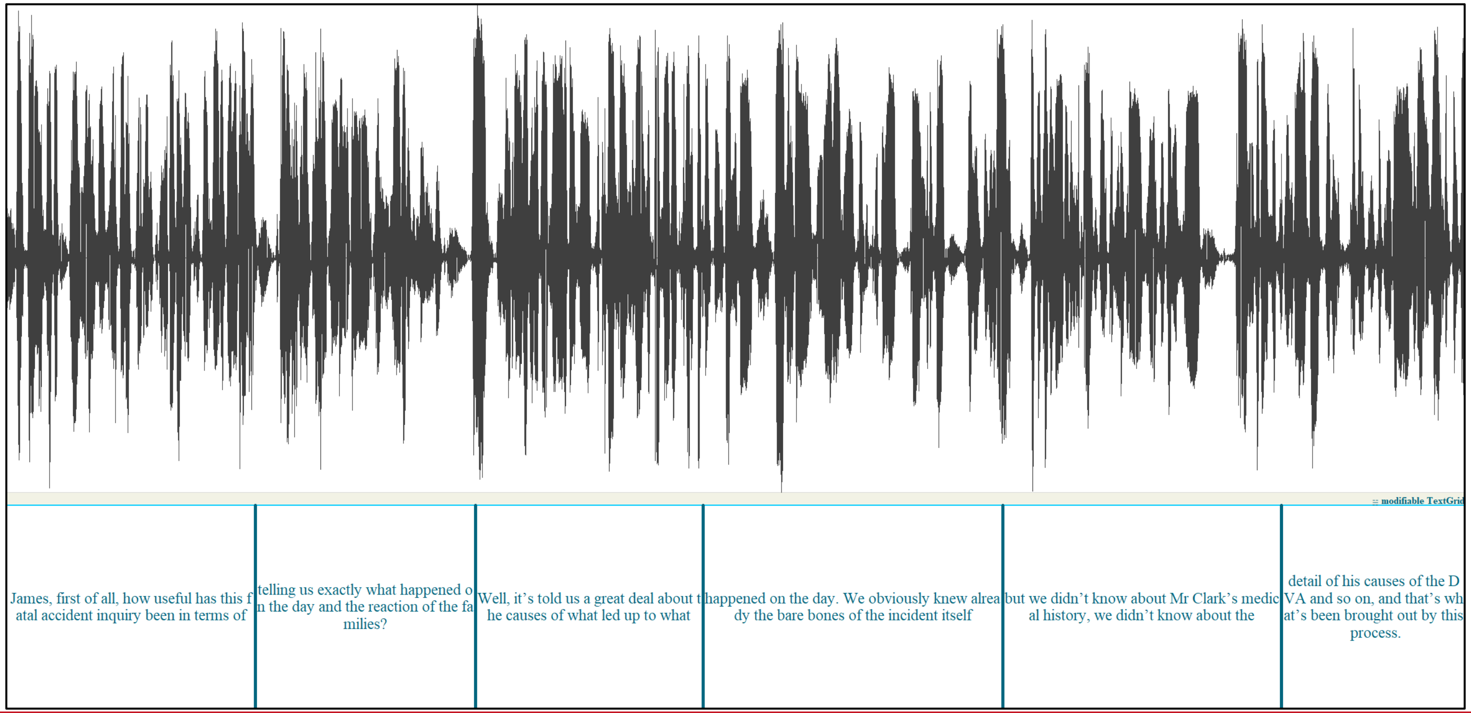
- **idealized instead of verbatim transcripts**

→ problematic for close transcription
→ increase in WER
→ CrisperWhisper (Wagner et al. 2024) as alternative?

Human transcription	Whisper transcription
the position that as as of just now is that I- I've obviously spent time yesterday covering matters which substantially were not in the note of argument now and that took a little bit more time erm a- and so erm erm today I I think I can probably go quite a lot faster	The position as of just now is that I've obviously spent time yesterday covering matters which substantially were not in the note of argument, and that took a little bit more time, and so today I think I can probably go quite a lot faster

04 Discussion – speaker diarization

- **lack of speaker diarization (Radford et al. 2022, p. 3)**
- limited capabilities of WhisperX (Bain et al. 2023) and pyannote (Bredin, 2020) or Whisper and NVIDIA NeMo (Ashraf, 2024)



04 Discussion

- some human reference transcripts worse than Whisper transcripts

→ increase in WER

(Semi-)automatic approach:

‘Traditional‘ approach in English Linguistics (and other disciplines):

RQ → data collection → ASR data transcription + manual checks / manual data transcription → data annotation → data analysis → results

manual checks:
- → hallucinations
- → idealized transcriptions
- → speaker diarization

05 Conclusion

05 Conclusion

RQ1 What is the **transcription accuracy** of different Whisper models for the corpora ICE Nigeria & ICE Scotland?

- best accuracy for models large_v2 & large_v3 (also most robust models)
- worse results for ICE Nigeria than for ICE Scotland overall

→ accent bias (outer circle variety)

→ Whisper more robust than other systems

(recording quality of ICE Nigeria worse)

05 Conclusion

RQ2 Which variables have a **significant influence on ASR performance**?

model → the larger the model, the better the performance

corpus → better performance for ICE Scotland

quality → the better the audio, the better the results

text_category → better results for scripted speech

speaker_number → better results for monologue data

gender → better performance for (only) female speaker data

06 Outlook

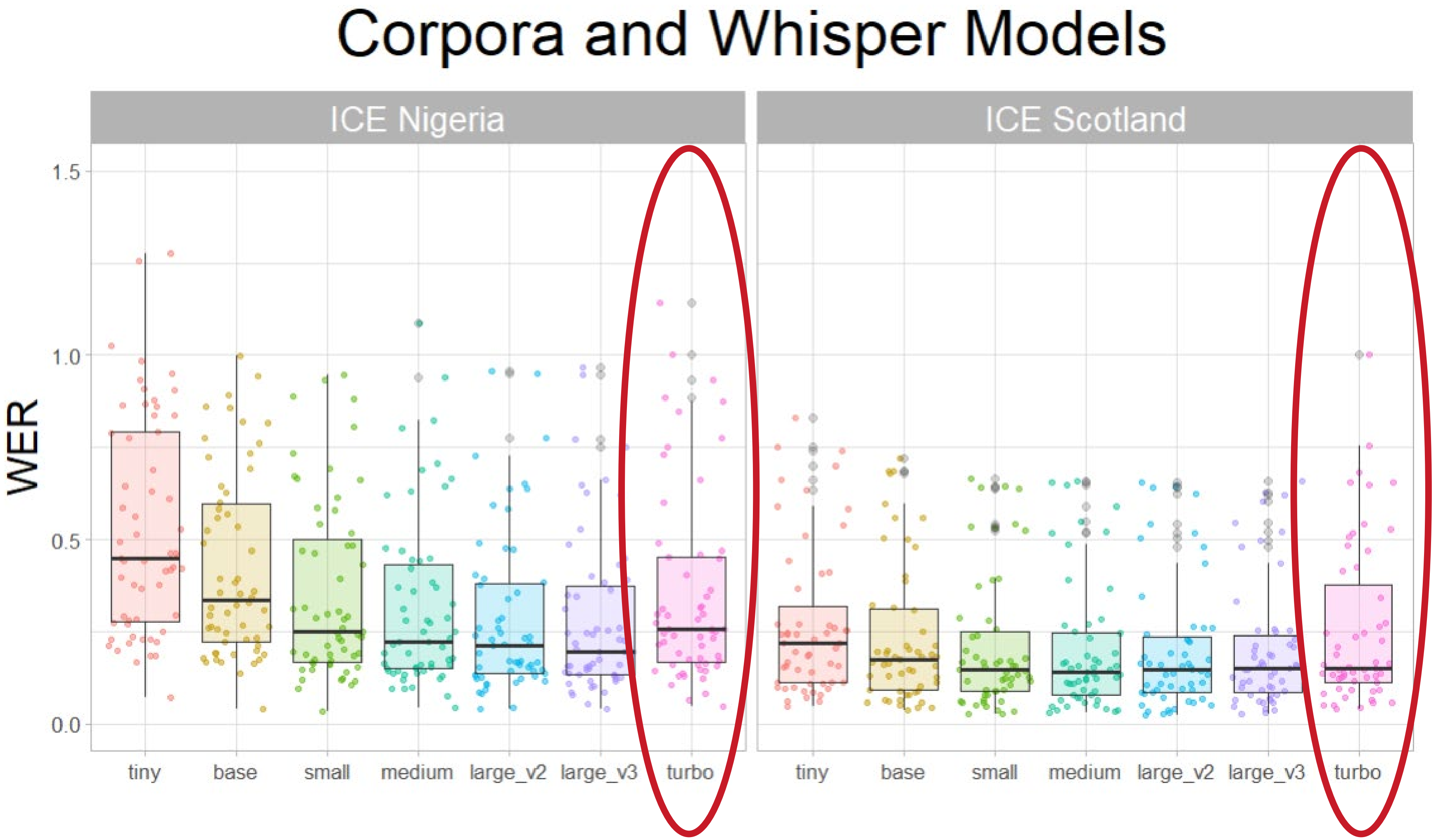
06 Outlook

NEXT STEPS

- extend dataset (more data, other varieties, model turbo)
- integrate more precise acoustic parameters into analysis and modelling
- use other evaluation metrics than WER
- compare human transcribers and Whisper more closely
- …

06 Outlook

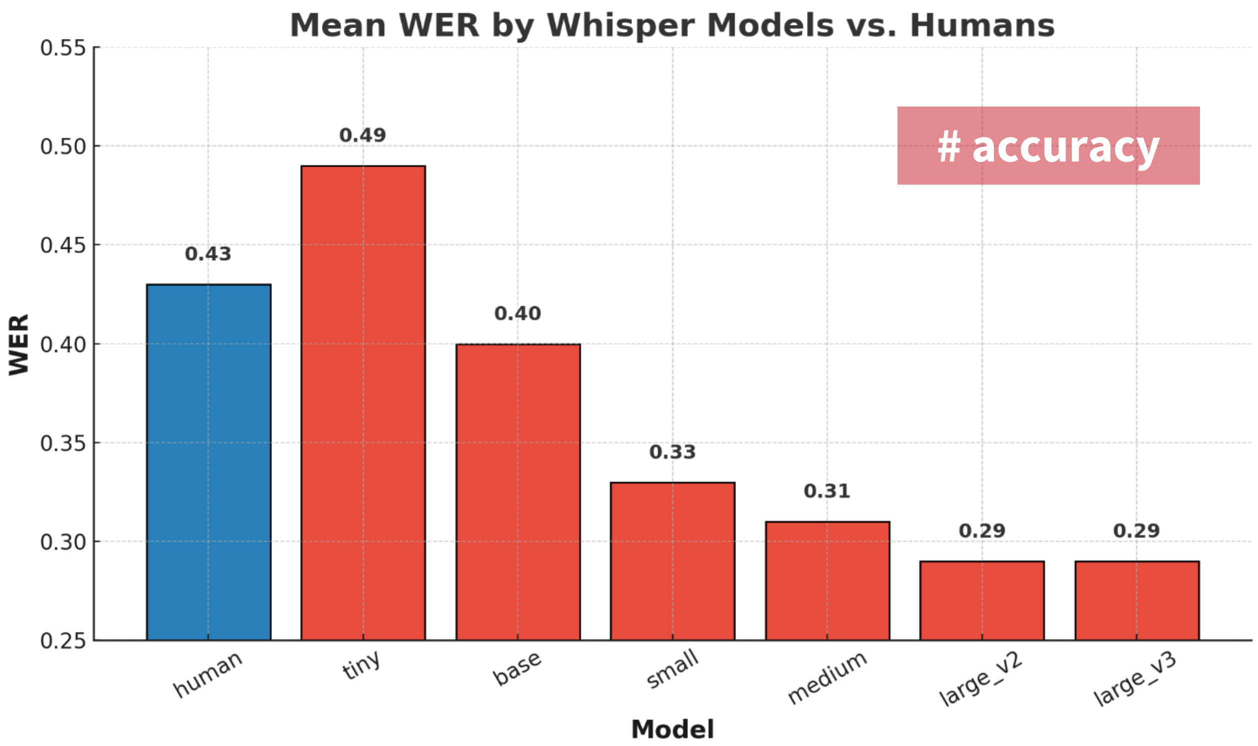
university of koblenz
continue discovering

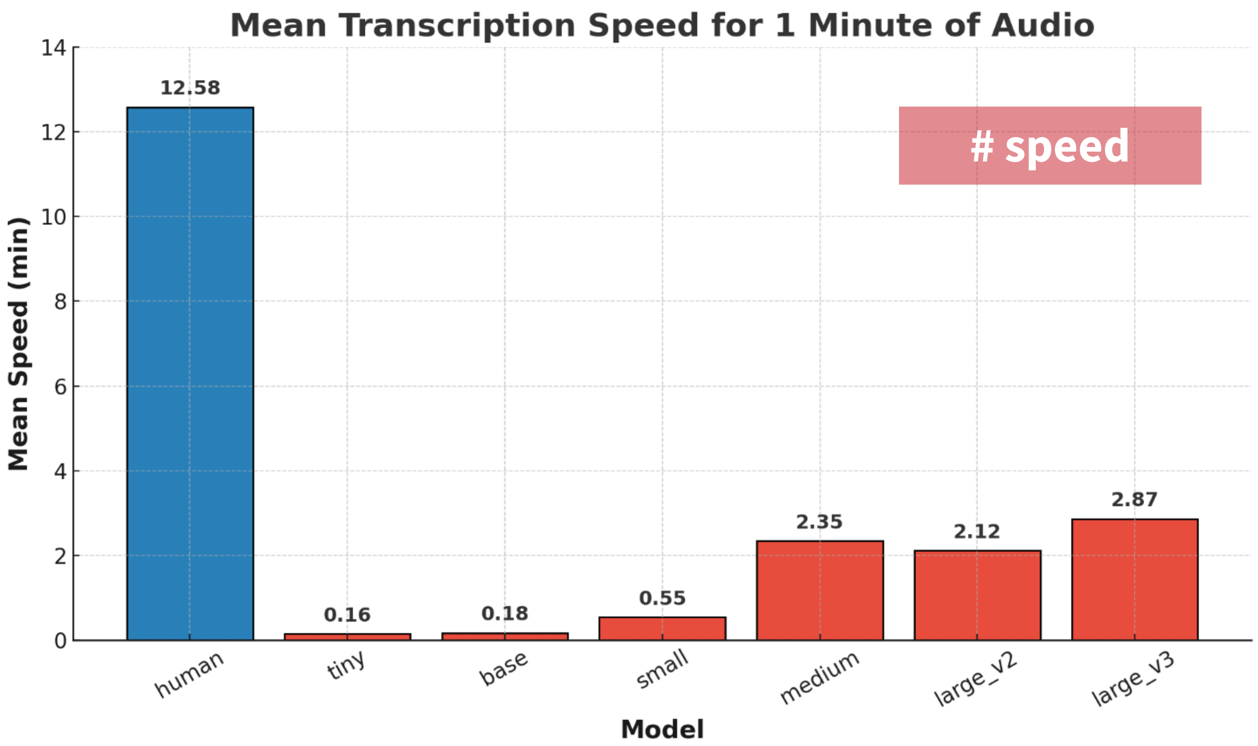


06 Outlook

university of koblenz
continue discovering

RQ3 How does **Whisper compare with trained human transcribers** in terms of accuracy and speed?



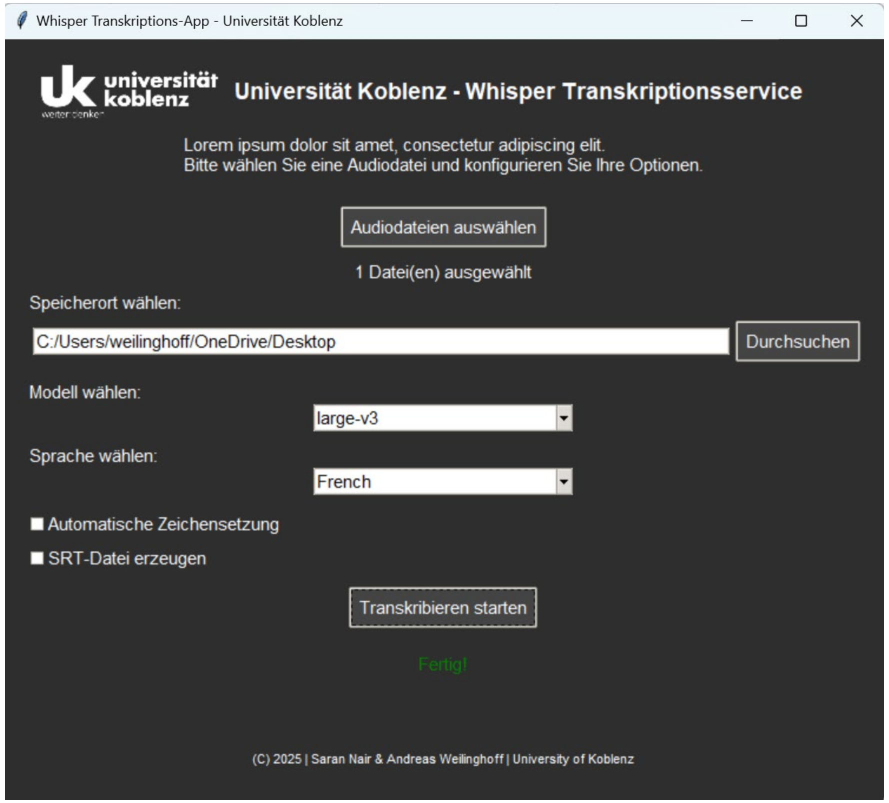


05 Outlook – user-friendliness

→ **Whisper requires command line / progamming knowledge (Python)**
→ projects to increase user-friendliness

Whisper Uni Server

05 Outlook – user-friendliness



Whisper Desktop App

https://github.com/Andreas-Weilinghoff/whisper_desktop_app

05 Outlook – user-friendliness

university of koblenz
continue discovering

→ Whisper finetuning for specific varieties

RheinlandPfalz
MINISTERIUM DER JUSTIZ

universität koblenz
weiter:denken

Fine Tuning 'Whisper' for enhanced speech recognition across diverse English accents: Feature analysis, Evaluation, and strategies to reduce hallucinations

Master's Thesis

in partial fulfillment of the requirements for
the degree of Master of Science (M.Sc.)
in Web and Data Science

Finetuning and adapting for legalese

Finetuning for Indian & Scottish English

References

References

Adda-Decker, M., and Lamel, L. (2005). “Do speech recognizers prefer female speakers?,” in *Proceedings of INTERSPEECH*, Lisbon, Portugal (International Speech Communication Association, Baixas, France).

Armstrong, R. (2024). *werpy - Word Error Rate for Python* [Computer software]. https://github.com/analyticsinmotion/werpy

Ashraf, M. (2023). *Speaker Diarization Using OpenAI Whisper*. [Computer software]. https://github. com/MahmoudAshraf97/whisper-diarization

Baevski,. , Zhou, Y., Mohamed, A. & Auli, M. (2020). “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *NeurIPS*, vol. 33, pp. 12449–12460.

Bain, M., Huh, J., Han, T. & Zisserman, A. (2023). *WhisperX: Time-Accurate Speech Transcription of Long-Form Audio*. https://www.robots.ox.ac.uk/~vgg/publications/2023/Bain23/bain23.pdf

Boersma, P., & Weenink, D. (2019). Praat: doing phonetics by computer (Version 6.1.08) [Computer software]. http://www.praat.org/

Desplanques, B., Thienpondt, J. & Demuynck, K. (2020). *ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification.* https://arxiv.org/pdf/2005.07143.pdf

Goldwater, S., Jurafsky, D., & Manning, C. (2010). “Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates,” *Speech Communication* 52(3), 181–200.

Graham, C. & Roll, N. (2024). Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits. *The Journal of the Acoustical Society of America*. https://doi.org/10.1121/10.0024876

Hirschle, J. (2022). *Deep Natural Language Processing. Einstieg in Word Embedding, Sequence-to-Sequence Modelle und Transformer mit Python*. Munich: Hanser Publishing.

IBM (2022). Watson Speech to Text. [Software]. Retrieved from: https://www.ibm.com/cloud/watson-speech-to-text [Date of access: 25 Nov. 2022].

Jurafsky, D. & Martin, J. H. (2023). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* https://web.stanford.edu/~jurafsky/slp3/ed3book_jan72023.pdf

References

Klein, G. (2023). *faster-whisper*. [Computer software]. https://github.com/guillaumekln/faster-whisper

Knill, K., Gales, M., Kyriakopoulos, K., Malinin, A., Ragni, A., Wang, Y., and Caines, A. (2018). “Impact of ASR performance on free speaking language assessment,” in *Proceedings of Interspeech 2018*, Hyderabad, India (International Speech Communication Association, Baixas, France), pp. 1641–1645.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., and Goel, S. (2020). “Racial disparities in automated speech recognition,” *Proc. Natl. Acad. Sci. U.S.A.* 117(14), 7684–7689.

Markl, N. (2022). “Language variation and algorithmic bias: Understanding algorithmic bias in British English automatic speech recognition,” in *Proceedings of 2022 5th ACM Conference on Fairness, Accountability, and Transparency* (FAccT 2022), June 21–24, Seoul (Association for Computing Machinery, New York), pp. 521–534.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C. & Sutskever, I. (2022): Robust Speech Recognition via Large-Scale Weak Supervision. https://arxiv.org/abs/2212.04356

Schützler, O., Gut, U., & Fuchs, R. (2017). New perspectives on Scottish Standard English: Introducing the Scottish component of the International Corpus of English. In S. Hancil & J. C. Beal (Eds.), *Perspectives on Northern Englishes* (pp. 273–302). Mouton de Gruyter.

Tatman, R., and Kasten, C. (2017). “Effects of talker dialect, gender and race on accuracy of Bing speech and YouTube automatic captions,” in *Proceedings of Interspeech, Stockholm, Sweden* (International Speech Communication Association, Baixas, France), pp. 934–938.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I. (2017). Attention Is All You Need. https://arxiv.org/pdf/1706.03762.pdf

Wagner, L., Thallinger, B., & Zusag, M. (2024, August 29). CrisperWhisper: Accurate timestamps on verbatim speech transcriptions (Version 1) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2408.16589

Weilinghoff, A. (2023): whisper_to_textgrid+eaf.py (Version 1.0) [Source code]. https://www.andreas-weilinghoff.com/#code

Weilinghoff, A. and Nair, S. (2025). AI Transcription Desktop App. https://github.com/Andreas-Weilinghoff/whisper_desktop_app

Wunder, Eva-Maria & Voormann, Holger & Gut, Ulrike. (2008). “The ICE Nigeria corpus project: Creating an open, rich and accurate corpus.” *International Computer Archive of Modern and Medieval English (ICAME) Journal*, 34, pp. 78-88.

Thank you very much for your attention!

Uni web: https://uni-ko.de/oUfpi

Private web: andreas-weilinghoff.com

ICAME 47

A CONFLUENCE OF CORPUS RESEARCH
IN THE AGE OF AI

Dates: 26-30 May 2026 | CfP Deadline: 31 October 2025 | Web: wp.uni-koblenz.de/icame47/