



# Transcribing Diverse Voices: Using Whisper for ICE corpora

Andreas Weilinghoff<sup>1</sup>

<sup>1</sup>Department of English and American Studies, University of Koblenz, Germany

weilinghoff@uni-koblenz.de

## Abstract

The precise transcription of speech data is crucial yet work-intensive in the field of sociolinguistics. Although recent advancements in end-to-end ASR (e.g. Whisper) offer great potential across various disciplines, these models have rarely been tested for sociolinguistic corpus transcription. This study addresses this gap by harnessing all Whisper models for the re-transcription of classic sociolinguistic reference corpora of non-standard varieties: ICE Nigeria and ICE Scotland. Employing WER metrics, the study utilizes linear mixed-effects modelling to determine significant factors affecting transcription accuracy. The results show that Whisper can manage both varieties, though it is slightly less accurate for Nigerian English. An increased model size reduces WER and boosts robustness, though accuracy varies by sound file. While Whisper proves useful for corpus transcription work overall, challenges such as speaker diarization, hallucinations and idealized transcriptions persist.

**Index Terms:** speech recognition, Whisper, corpus linguistics, Nigerian English, Scottish English

## 1. Introduction

The transcription of sound data is an essential yet time-consuming and labor-intensive part of almost all sociolinguistic research projects. This is especially true for corpus linguistics, as only well-transcribed and carefully annotated corpora provide a reliable basis for subsequent analyses. Since the successful implementation of transformer models with attention mechanisms [1], numerous fields, including automatic speech recognition (ASR), have experienced significant improvements. A striking, open-source example of this is OpenAI Whisper [2].

A central question arising from these developments is how these new ASR models can be used effectively for corpus transcription work. How can these models be integrated into different transcription and annotation workflows, and what are the opportunities they offer and challenges they pose? Furthermore, how do these models perform across varying conditions that typically influence ASR transcription accuracy?

### 1.1. Influences on ASR accuracy

A crucial factor that influences ASR performance is audio quality as speech recognition systems work better on high-quality audio with low background noise [3]. Another important factor is the type or structure of speech: well-structured read speech usually poses less difficulty for ASR systems than uncontrolled conversational speech between different people [3]. In this context, another challenging task is speaker diarization, [4] which is especially important for sociolinguistics as data sources often comprise interview data with two or several speakers. The

transcription of conversational speech is thus crucial for sociolinguistics but poses a significant challenge for many ASR systems [5]. Additionally, the gender of speakers affects ASR accuracy, with many studies reporting better performance on female voices [6, 7, 8]. Language variety also impacts performance; most systems are trained on standard varieties like General American (GA) or Standard Southern British English (SSBE) and many previous studies have shown that commercial ASR systems perform worse on non-native speakers [9] as well as regional [8] and racial [10] minority groups. This means that, for instance, ASR systems have lower word error rates (WER) when transcribing GA speech than African American Vernacular English [10] and they also tend to work better with speakers from the South of England (SSBE) than with accents from, for instance, Newcastle, Liverpool, Bradford or Belfast [8]. This can, in turn, lead to further bias [8].

### 1.2. Evaluating Whisper for Sociolinguistic Corpus Transcription work

Whereas previous investigations evaluated Whisper for native and non-native accents [9] and different standard reference corpora from the field of Natural Language Processing (NLP) [2, 5], few studies have yet investigated how Whisper can be effectively used for the transcription of sociolinguistic corpora. For this purpose, parts of the spoken components of ICE Nigeria [11] and ICE Scotland [12] were re-transcribed with the different Whisper models and the resulting transcriptions were then compared to the manual reference transcriptions via WER metrics. In terms of inferential statistics, the analysis applies linear mixed-effects modelling to find out what significantly influences the ASR transcription performance. The analysis takes different acoustic, technical parameters as well as sociolinguistic parameters into account and focuses on the following research questions:

- (RQ 1) What is the transcription accuracy of different Whisper models for the sociolinguistic corpora ICE Nigeria and ICE Scotland?
- (RQ 2) What has a significant influence on the ASR performance?

## 2. Data and Method

The data for this study was retrieved from the International Corpus of English (ICE) project, initiated in 1990 with the main goal of gathering materials for comparative analyses of English worldwide. The ICE corpora encompass both scripted and unscripted monologues and public and private dialogues across diverse speech situations. These corpora, featuring speech from male and female speakers of various ages, are ideal for this anal-

ysis due to their consistent format and high-quality reference transcriptions.

## 2.1. Data selection

For the present study, I retrieved data from the spoken components of ICE Nigeria [11] and ICE Scotland [12] to test how the Whisper models deal with a postcolonial outer-circle variety (Nigerian English) and an inner-circle variety (Scottish English) that is neither SSBE nor GA. From each corpus, I extracted 60 sound files from 12 different speech categories (broadcast discussions (bdis), broadcast news (bnew), broadcast talks (btal), broadcast transactions (btran), commentaries (com), cross examinations (cr), demonstrations (dem), legal cross-examinations (leg), class lessons (les), non-broadcast talks (nbtal), parliamentary debates (parl), unscripted speeches (unsp)). I selected these categories because they are represented in both corpora, allowing for a direct comparison. Another reason for the selection of the files is the availability of the corresponding manual reference transcriptions, which are essential for evaluating the accuracy of the ASR output. This setup ensures that the study covers a wide range of speech situations and styles, providing a comprehensive assessment of Whisper’s capabilities across Nigerian English and Scottish English. An overview of the final dataset can be found in Table 1.

Table 1: Main dataset of the study separated for ICE Nigeria and ICE Scotland.

Corpus	Total duration	Total word count
ICE Nigeria	13:05:47 h	94,499
ICE Scotland	11:50:31 h	111,418

Overall, the dataset comprises 24:56:18 hours of audio and the reference transcripts comprise 205,917 words in total. While the duration is a bit longer for ICE Nigeria than for ICE Scotland, the reference transcripts of the latter include more words.

## 2.2. Data preparation and re-transcription with Whisper

The extracted sound files were annotated for the following metadata: file duration, number of speakers, gender of speakers and recording quality. The recording quality was assessed subjectively (“good” vs. “bad”) and supported by signal-to-noise (SNR) ratio measurements.

In the following step, the selected sound files of ICE Nigeria and ICE Scotland were re-transcribed with all available Whisper models at the time of writing (tiny, base, small, medium, large\_v2 and large\_v3). I employed an AMD EPYC 7402 processor for the transcription process. To focus solely on the transcription accuracy, the Whisper outputs of the different models were extracted as plain text files (.txt) without any timestamps or speaker labels. In the next step, all original transcriptions of the ICE corpora were retrieved and all speaker labels were removed from these transcripts to maintain only the plain text, aligning the format with that of the Whisper-generated files.

The WERs for the Whisper transcriptions were calculated against the ICE reference transcriptions using the library werpy [13] in Python. The werpy library was chosen due to its capabilities in handling text normalization and alignment, which are crucial for accurate WER calculations. A custom Python script was developed to automate this process, running through the directories to process each file. The script recorded the filenames,

word counts for both ICE reference and Whisper text files, the specific Whisper model used, and the computed WER rates and exported that to a table in a .csv format.

## 2.3. Statistical Analysis

For descriptive statistics, I used the R package ggplot2 [14] to generate boxplots and jitterplots, plotting WER on the  $y$ -axis and categorizing it by corpus and model on the  $x$ -axis. I also calculated the average WER and standard deviations for the different models, providing insights into the transcription accuracy of the Whisper models for the two corpora (RQ 1).

To determine which variables significantly influence the WER (RQ 2), I used methods of inferential statistics. Following the approach of Graham and Roll [9], I applied linear mixed-effects modelling in R with the lme4 [15] and lmerTest [16] packages to investigate which variables significantly influence the performance of Whisper. The dependent variable is the calculated WER, and fixed factors include the corpus, the text category, the Whisper model, the recording quality, the speaker number, the gender and the file duration. The coding for gender had to be simplified into a three-level category with all-male, all-female and mixed-gender sound files. As the variable speaker number includes many different levels (minimum: 1 speaker in a sound file; maximum: 21 speakers in a sound file), I collapsed it into a binary variable to distinguish between monologue and dialogue data. As the sound files generally include different speakers, I treated the individual sound files as random factors. An overview of the random and fixed factors can be found in Table 2.

Table 2: Overview of random and fixed factors used for linear mixed-effects modelling.

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	categorical	good, bad
speaker number	categorical	mono, poly
gender	categorical	female, male, mixed

I applied a stepwise regression procedure with a backward selection for model building. I started with a full model including all variables and, as far as token numbers allowed, all possible interactions and I used the generic step function of the lmerTest package [16] to exclude the factors and interactions which returned no significant effects. The  $R^2$  values for the models were generated via the r.squareGLMM() function from the MuMin package [17] and the best model fit was found via maximum likelihood ratio tests. For a better interpretation of the model estimates, I applied the procedure by Tanner et al. [18]) and took the exponent of the model parameter’s value. For example, an estimate of  $e^{0.19} = 1.2$  represents a WER increase of 20% when compared to the intercept [18].

### 3. Results

The results are divided into two subsections, each addressing one research question. Subsection 3.1 presents the results for the accuracy of the different Whisper models for the two reference corpora ICE Scotland and ICE Nigeria (RQ 1). Subsection 3.2 provides the results on which variables significantly influence transcription accuracy (RQ 2).

#### 3.1. Transcription accuracies of different Whisper models

An overview of the average WERs of the different Whisper models for the overall dataset and separated for the two corpora can be found in Table 3.

Table 3: Average WER and standard deviation of Whisper models across the dataset grouped by corpus.

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

The average values show that, as expected, larger Whisper models tend to produce outputs with lower WERs than smaller Whisper models. The transcript accuracy generally improves from the models tiny, base, small, medium to large\_v2; only the models large\_v2 and large\_v3 show comparable WER values. For ICE Nigeria, large\_v3 (mean WER: 0.29) is marginally better than large\_v2 (mean WER: 0.30), but this is not the case for ICE Scotland (large\_v2: 0.26; large\_v3: 0.26). Overall, Whisper generally produces transcripts with lower WERs for ICE Scotland than for ICE Nigeria. The WER values are always higher for ICE Nigeria across all Whisper models. The highest average WER can be found for ICE Nigeria with the model tiny (0.54). This WER value is relatively high as it specifies that more than half of the words in the ASR transcript do not match those in the ICE reference transcription. In contrast, the lowest average WER of 0.26 can be found for the large models on ICE Scotland data. Apart from the mean WER, Table 3 also shows relatively high standard deviations across the whole dataset. The highest standard deviation can be found for the model tiny, and the value decreases with the model size across both corpora. For ICE Scotland, the standard deviation is even higher than the mean WER for the models medium and large\_v2. The lowest standard deviations can be found for the large models (v2 and v3) for ICE Nigeria with a value of 0.24. Although the standard deviation generally decreases as the Whisper model sizes increase (from tiny to large\_v3), the reduction in standard deviation is less pronounced than the differences observed in the WER values. This indicates that, although the larger Whisper models show less variability in ASR accuracy than the smaller models, the differences in WER across different files remain relatively strong. Furthermore, the decline in standard deviation is much less pronounced for ICE Scotland (model tiny: 0.28; model large\_v3: 0.26) than for ICE Nigeria (model tiny: 0.30; model large\_v3: 0.24). This observation implies that the improvement in consistency of ASR performance across model sizes is more substantial for the ICE Nigeria dataset than for the ICE Scotland dataset. More details about the ASR accuracy for

both corpora is provided in Figure 1.

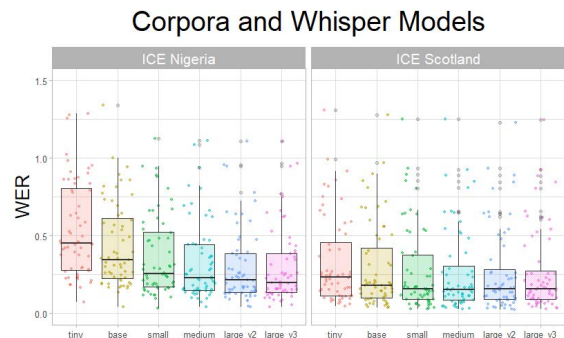


Figure 1: Boxplots and jitterplots of WERs separated for the Whisper models and corpora.

Figure 1 highlights several key trends that could already be observed in Table 3: ASR accuracy increases with model size, and the WER is not only generally lower but also more consistent for ICE Scotland compared to ICE Nigeria. However, the jitter plots further reveal considerable variation in WER across individual files, which likely contributes to the high standard deviations observed in Table 3. While some transcripts exhibit exceptionally low WERs, nearing zero, others surpass the value of 1.0. The plot further indicates that, although many WER values are under 0.25, there are significant outliers at much higher levels. For ICE Scotland in particular, there are some outliers at the top which can be found across different Whisper models. These observations underline that WER seems to be highly dependent on individual files, with some transcripts achieving extremely high accuracy and others likely to be useless.

#### 3.2. Significant influences on transcription accuracy

In order to find out what has a significant influence on the WER, I fitted several linear mixed-effects models on the WER. Due to a lack of token numbers for several factors, I could only include a limited number of interactions. In all models, the fixed factor file duration did not show significant effects. This means that the transcription accuracy does not depend on the length of a file. Whisper can work with longer and shorter sound files, and the file length does not have a significant influence on ASR accuracy. An overview of the best model (formula:  $\text{wer}(\text{model} * \text{corpus}) + (\text{corpus} * \text{quality}) + (\text{model} * \text{quality}) + \text{text\_category} + \text{speaker\_no\_binary} + \text{gender} + (1 | \text{file\_name}))$ ) is shown in Table 4.

As expected from the plots (see Figure 1), all Whisper models show significant effects. The estimates reveal that the WER of the model base is roughly 9% lower than when compared to the intercept (model: tiny), which corresponds to the average percentage values. Similarly, the WER decrease is roughly 20% for the model small, 25% for the model medium, 29% for the model large\_v2 and 30% for large\_v3 when compared to the intercept. This corroborates the observation that the WER generally decreases with increasing model size. The fixed factor quality\_2 also shows significant effects, and the estimate reveals that the WER decreases by 25% in good-quality sound files. Hence, Whisper can handle good-quality sound files better than bad-quality sound files, which is in line with previous studies. Apart from that, also the text categories broadcast news (bnew), commentaries (com), cross examinations (cr), class

Table 4: Best linear mixed-effects model fit (marginal  $R^2 = 0.65$ ; conditional  $R^2 = 0.96$ )

Model Fit Statistics				
AIC	BIC	logLik	deviance	df.resid
-1663.4	-1503.1	866.7	-1733.4	685
Scaled Residuals				
Min	1Q	Median	3Q	Max
-3.6191	-0.3717	-0.0134	0.3188	6.2057
Random Effects				
Groups	Name	Variance	Std dev	
file_name	(Intercept)	0.025513	0.15973	
Residual		0.002678	0.05175	
Fixed Effects				
Effect	Estimate	t value	Pr(>  t )	Signif. code
(Intercept)	0.2021	2.463	0.01518	*
modelbase	-0.0993	-8.783	$< 2e^{-16}$	***
modelsmall	-0.2092	-18.499	$< 2e^{-16}$	***
modelmedium	-0.2562	-22.656	$< 2e^{-16}$	***
modellarge_v2	-0.2961	-26.187	$< 2e^{-16}$	***
modellarge_v3	-0.3005	-26.570	$< 2e^{-16}$	***
corpusICE Sco	-0.0805	-1.081	0.28199	
quality_good	-0.2516	-5.067	1.34e-06	***
text_categorybnew	0.3848	4.989	2.08e-06	***
text_categorybtal	0.1246	1.449	0.15008	
text_categorybtran	0.0681	0.881	0.38031	
text_categorycom	0.2401	2.928	0.00408	**
text_categorycr	0.4159	5.708	8.44e-08	***
text_categorydem	0.2172	2.604	0.01038	*
text_categoryleg	0.1569	2.025	0.04512	*
text_categoryles	0.2482	3.079	0.00257	**
text_categorynbtal	0.1503	1.747	0.08312	.
text_categoryparl	-0.0221	-0.303	0.76203	
text_categoryunsp	0.2738	3.092	0.00247	**
speaker_no_poly	0.2842	5.564	1.63e-07	***
gender_allmale	0.1243	2.798	0.00599	**
gender_multiple	0.1467	2.474	0.01477	*

lessons (les), demonstrations (dem), legal cross-examinations (leg) and unscripted speeches (unsp) all show significant effects with positive effect sizes. The estimates are particularly high for the categories broadcast news (0.3848) and cross examinations (0.4159), which underlines a strong and significant increase in WER when compared to the intercept (= broadcast discussions). This means that some text types pose less difficulty for the ASR system than others. Another observation is that there is a significant influence of the fixed factor speaker number. The variable shows significant effects and positive effect sizes. When compared to the intercept which incorporates exclusively monologue data, the WER tends to be roughly 28% higher for files with several speakers. This finding corresponds with the general observation that ASR systems can deal better with structured monologue data than with less structured conversational speech between speakers [3]. Significant effects are also observed for the variable gender. Files with male speakers or speakers of multiple genders have significantly higher WERs than files with exclusively female speakers. It is not surprising that the WER increases in files with multiple genders, as these files, by definition, must represent conversational speech with several speakers. The variable speaker number has already indicated that WER is significantly higher in conversational data with several speakers, and this generally includes files with multiple genders. However, it is interesting to note that the ASR system handles female speech better than male speech in the provided dataset. This largely corresponds to the observations made by Markl [8]. Another finding is that the variable corpus does not show significant effects. This means that while the WER is generally lower for ICE Scotland than for ICE Nigeria

(see Table 3 and Figure 1), the overall difference in Whisper’s accuracy between the two varieties does not reach statistical significance overall.

Another striking observation is the relatively high  $R^2$  values. The marginal  $R^2$  value for the best model is 0.65, which shows that the model can explain 65% of the variation in the dataset based on the fixed factors. The conditional  $R^2$  value, which represents both fixed and random effects, is very high at 0.96. This means that the whole model accounts for 96% of the variation in the dataset. Such high  $R^2$  values are unusual in sociolinguistic datasets, but one has to keep in mind that the target variable is the evaluation metric WER and not a specific sociolinguistic variable. Furthermore, the high conditional  $R^2$  value reveals that the individual files tend to have a very strong influence on Whisper’s performance. As seen in Figure 1, there are some files with an extremely low WER and others with a very high WER. This further underlines that Whisper’s performance is strongly dependent on the sound file itself.

## 4. Discussion and Conclusion

Upon closely examining the transcriptions and sound files, certain patterns emerge. Whisper occasionally hallucinates, leading to unusual repetitions of phrases that do not match the actual audio. These errors are most common in segments with poor audio quality, extended silences, speaker overlaps, interruptions, or switches to Nigerian Pidgin English.

Another factor contributing to the relatively high WER is that Whisper automatically “corrects” utterances, resulting in idealized transcriptions (e.g. hesitations and false starts are often omitted). Since manual reference transcripts and sociolinguistic transcripts typically include these features, their absence in Whisper’s output increases the WER due to discrepancies between the idealized Whisper transcription and the more accurate manual reference that reflects actual spoken content. This discrepancy poses significant challenges for sociolinguistic corpus transcription. The hallucinations and idealized transcriptions significantly contribute to the high conditional  $R^2$  values because these errors often occur consistently across different models and individual files. This regularity is why these specific files disproportionately influence the WER and overall model accuracy.

Another challenge in sociolinguistic corpus transcription is speaker diarization and the occasionally inaccurate timestamps produced by Whisper. Although these timestamps can be enhanced with WhisperX [19] and while this system also incorporate diarization capabilities via pyannote [4], speaker diarization continues to pose a significant challenge for sociolinguistic transcription work. Therefore, speaker diarization, manual checks, including corrections for hallucinations and potential adjustments in idealized transcriptions, remain crucial. Despite these issues, Whisper remains a valuable tool for transcription.

This study assessed the accuracy of various Whisper models via WER on the sociolinguistic corpora ICE Nigeria and ICE Scotland. While WER is higher for Nigerian English than Scottish English, the gap narrows across the larger models. Consistent with prior research, results show that better sound quality, scripted speech, monologues and female voices yield more accurate transcriptions. However, the lack of speaker diarization, hallucinations and idealized transcriptions pose challenges for sociolinguistic corpus transcription. Future research on addressing these issues will be crucial in further enhancing the ASR’s capabilities for transcribing diverse voices.

## 5. References

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017. [Online]. Available: <https://arxiv.org/pdf/1706.03762.pdf>
- [2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022. [Online]. Available: <https://arxiv.org/abs/2212.04356>
- [3] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2023. [Online]. Available: [https://web.stanford.edu/~jurafsky/slp3/ed3book\\_jan72023.pdf](https://web.stanford.edu/~jurafsky/slp3/ed3book_jan72023.pdf)
- [4] H. Bredin, "Pyannote.audio: Neural building blocks for speaker diarization," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 7124–7128.
- [5] S. O'Connor Russell, I. Gessinger, A. Krason, G. Vigliocco, and N. Harte, "What automatic speech recognition can and cannot do for conversational speech transcription," *Research Methods in Applied Linguistics*, vol. 3, no. 3, 2024. [Online]. Available: <https://doi.org/10.1016/j.rmal.2024.100163>
- [6] M. Adda-Decker and L. Lamel, "Do speech recognizers prefer female speakers?" in *Proceedings of INTERSPEECH*, Lisbon, Portugal, 2005, International Speech Communication Association, Baixas, France.
- [7] S. Goldwater, D. Jurafsky, and C. Manning, "Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, no. 3, pp. 181–200, 2010.
- [8] N. Markl, "Language variation and algorithmic bias: Understanding algorithmic bias in British English Automatic Speech Recognition," in *Proceedings of the 2022 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022)*. Seoul, South Korea: Association for Computing Machinery, June 2022, pp. 521–534.
- [9] C. Graham and N. Roll, "Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits," *The Journal of the Acoustical Society of America*, 2024. [Online]. Available: <https://doi.org/10.1121/10.0024876>
- [10] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in Automated Speech Recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.
- [11] E. Wunder, H. Voormann, and U. Gut, "The ICE Nigeria corpus project: Creating an open, rich and accurate corpus," *International Computer Archive of Modern and Medieval English (ICAME) Journal*, vol. 34, pp. 78–88, 2008.
- [12] O. Schützlér, U. Gut, and R. Fuchs, "New perspectives on Scottish Standard English: Introducing the Scottish component of the International Corpus of English," in *Perspectives on Northern Englishes*, S. Hancil and J. C. Beal, Eds. Mouton de Gruyter, 2017, pp. 273–302.
- [13] R. Armstrong, "werpy - Word Error Rate for Python," <https://github.com/analyticsinmotion/werpy>, 2024, [Computer software].
- [14] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. [Online]. Available: <https://ggplot2.tidyverse.org>
- [15] D. Bates, M. Mälcher, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.
- [16] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.
- [17] K. Bartón, "Mumin package (version 1.46.0)," <https://cran.r-project.org/web/packages/MuMIn/MuMIn.pdf>, 2022, [Computer software].
- [18] J. Tanner, M. Sonderegger, J. Stuart-Smith, and S. D. Consortium, "Vowel duration and the voicing effect across English dialects," *Toronto Working Papers in Linguistics*, vol. 41, no. 1, 2019. [Online]. Available: <https://doi.org/10.33137/twpl.v41i1.32769>
- [19] M. Bain, J. Huh, T. Han, and A. Zisserman, "WhisperX: Time-accurate speech transcription of long-form audio," 2023. [Online]. Available: <https://www.robots.ox.ac.uk/~vgg/publications/2023/Bain23/bain23.pdf>