Understanding Racial Disparities in Automatic Speech Recognition: the case of habitual "be"

Joshua L. Martin[1]*, Kevin Tang*[1]

[1]Department of Linguistics, University of Florida, Gainesville, FL, U.S.A., 32611-5454

joshua.martin@ufl.edu, tang.kevin@ufl.edu

Abstract

Recent research has highlighted that state-of-the-art automatic speech recognition (ASR) systems exhibit a bias against African American speakers. In this research, we investigate the underlying causes of this racially based disparity in performance, focusing on a unique morpho-syntactic feature of African American English (AAE), namely habitual "be", an invariant form of "be" that encodes the habitual aspect. By looking at over 100 hours of spoken AAE, we evaluated two ASR systems – DeepSpeech and Google Cloud Speech – to examine how well habitual "be" and its surrounding contexts are inferred. While controlling for local language and acoustic factors such as the amount of context, noise, and speech rate, we found that habitual "be" and its surrounding words were more error prone than non-habitual "be" and its surrounding words. These findings hold both when the utterance containing "be" is processed in isolation and in conjunction with surrounding utterances within speaker turn. Our research highlights the need for equitable ASR systems to take into account dialectal differences beyond acoustic modeling.

Index Terms: speech recognition, racial bias, racial disparities, syntactic features, error analysis, fair machine learning, natural language processing, speech-to-text, African American English

1. Introduction

Linguistic discrimination has adversely affected the lives of marginalized populations for centuries, especially racially minoritized groups in the United States [1]. However, in spite of extensive research on linguistic discrimination between humans, little has been done to investigate the ways that automatic speech recognition (ASR) systems may be inheriting the same sorts of linguistic biases as their creators. This concern was highlighted most recently by Koenecke, et al. [2] who found that the average word error rate (WER) for white American speakers was significantly lower as compared to an average WER for African American speakers (19% vs 35%) among five prominent ASR systems from such companies as Google, Amazon, and Apple.

Building from this ground-breaking study and others that highlight similar issues [3, 4], this paper seeks to probe the underlying causes of racial disparity in ASRs. Koenecke, et al. [2] show that part of its roots lies in the mismatches between the acoustic data that the ASR systems were trained on, namely speech of white Americans, and the acoustic signal that the systems were tested on, namely speech of African Americans. Keeping the negative impact of the acoustic model constant, we seek to extend this analysis by investigating whether the morpho-syntactic differences between African American English (AAE) and Standardized American English (SAE) may also play a significant role in these performance gaps.

In this study, we have chosen to examine how ASRs handle habitual "be". Habitual "be" is an invariant form of "be" that is distinct in AAE and encodes the habitual aspect (e.g., "I **be** in my office by 7:30," meaning, "I **am usually** in my office by 7:30. [5]). Habitual "be" is a helpful case study as it is homophonous with other forms of uninflected "be" found in both AAE and SAE and allows a clear point of comparison for how ASRs handle instances of the word "be" that are habitual and those that are non-habitual. Because it is probable that the *language model* component of most ASRs have not been trained on AAE linguistic structures, utterances containing habitual "be" may be incorrectly inferred as another construction using "be"; for example, "I **be** in my office by 7:30," (habitual) could be inferred as "**I'll be** in my office by 7:30" (future). This incorrect inference of habitual "be" as non-habitual "be" is likely the result of habitual "be" having a lower probability of occurrence than non-habitual "be" given its contexts.

To this end, our two main research questions are, (1) how error prone is habitual "be" in ASR inferences, and (2) how does the occurrence of habitual "be" affect the error level of its local context in ASR inferences? In addition to these main inquiries, we also consider variables such as amount of context, domino effect, speech rate and signal to noise ratio, and what part they play in these processes.

2. Methods

2.1. Audio Data

All audio data fed to the ASR models in our analysis were drawn from the Corpus of Regional African American Language (CORAAL) [6]. CORAAL contains over 100 sociolinguistic interviews with African American speakers, totaling to over 105 hours of audio and including a rich variety of interviewees that vary widely by age, socio-economic background, gender identity, and urban/ruralness.

3,635 instances of the word "be" were collected from transcripts of the interviews. Each instance of "be" was hand-tagged as habitual/non-habitual and instances of "be" from non-Black interviewers were filtered out, resulting in 376 instances of habitual "be" and 2,974 instances of non-habitual "be".

For each instance of "be", audio clips for both the utterance and speaker turn in which it occurred were extracted using Parselmouth [7]. Utterance here refers to a single segment of speech delimited by pauses and transcribed in CORAAL as a single transcript line. Speaker turn refers to the entire set of a speaker's contiguous utterances before another interlocutor begins speaking. Both utterance and turn were extracted because ASR systems differ in their level of signal decoding. Some work on a single-pass, essentially transcribing audio to text as they go; others have a multi-pass system, whereby they transcribe as they go and then return for more passes over a larger context to determine the most probable inference, correcting themselves in the process.

2.2. ASR Models

Two ASR models were chosen for examination, one commercial and one open source. The commercial model, Google Cloud Speech [8] was selected both for its acclaimed state-of-the-art performance and inclusion in Koenecke, et al. [2]. Researchers behind the model report a WER of 6.7% [9], and Koenecke, et al. found it had a slightly lower error rate for African American speakers relative to other ASRs tested in the study.

Alongside a commercial option, an open source option was also incorporated into our study as commercial options tend to be less forthcoming with information about the ASR's acoustic models, language models, and training data. DeepSpeech [10] was selected because of its prominence and wide-spread use. DeepSpeech boasts an overall WER of 7.5% [11], and version 0.6 (used here) is trained on 3816 hours of transcribed audio taken from spoken English corpora Common Voice English [12], LibriSpeech [13], Fisher [14], and Switchboard [15] and 1700 hours taken from transcribed NPR programs [11].

2.3. Dependent Variables

2.3.1. Accuracy of "Be"

In order to determine the accuracy of the various ASR systems in recognizing each instance of "be", a process of deciding correctness using semi-automatic annotation was devised:

Step 1: if ASR inferences did not contain the word "be", they were judged incorrect. If they did, they were passed on to the next step. **Step 2**: a dependency parser from spaCy [16] was applied to examine syntactic dependencies to the left and right of "be" in the intended speech and the corresponding ASR inference. If each set of dependencies matched, the inference was labeled correct. For remaining instances, if the text of the word immediately to the left and to the right of "be" matched in the intended speech and ASR inference, the inference was deemed correct. **Step 3**: for inferences not labeled in the first two steps, a hand-coding scheme was used to decide correctness. For instances of habitual "be", if the "be" within the ASR inference maintained habitual aspect, the inference was judged correct. For non-habitual instances, if the grammatical type of "be" matched between intended speech and inference, the inference was labeled correct. **Step 4**: All remaining inferences were deemed incorrect.

2.3.2. Word Error Rate

Word Error Rater (WER) is a common measure used to determine the accuracy of ASR systems. WER is calculated by comparing all possible alignments of transcripts of intended speech with corresponding ASR inferences. Each alignment is a sequence of substitutions, deletions, and insertions (SDI) that equalize the word length of the inference with the intended speech. The number of SDI is then summed for each alignment, and the one with the lowest total (i.e., cost) is chosen as the optimal alignment. WER is then calculated by taking the total of SDI for this alignment and dividing it by the number of words in the intended speech.

For our study, we utilized the Wagner-Fischer algorithm [17, 18] to determine WER for full utterances and speaker turns. Additionally, we ascertained WER for the remaining portion of each utterance preceding and following "be" both in isolation and embedded within its larger speaker turn.

2.4. Variables of Interest

2.4.1. Habituality

Habitual "be" is unique in that it encodes the habitual aspect (see Sec. 1 Introduction), while all other forms of uninflected "be" which occur in AAE and SAE do not encode grammatical aspect in and of themselves. These other types of uninflected "be" appear both alone and in conjunction with other verbs in various syntactic constructions that are distinctly non-habitual, such as:

- auxiliary "be" in progressive constructions (e.g., "I will **be** going there tomorrow.")
- auxiliary "be" in passive constructions (e.g., "She should **be** given an award.")
- copula or auxiliary "be" preceded by verbal complements (e.g., "He wanted to **be** a lawyer.")
- copula or auxiliary "be" preceded by a modal (e.g., "They might **be** in the house.")
- imperative "be" (e.g., "**Be** quiet!")

2.4.2. Amount of Context

It has been shown that the incorporation of context into ASR processing can decrease WER and improve ASR accuracy [19, 20, 21]. Given this, we considered how the amount of context surrounding the occurrence of "be" may affect each ASR's ability to correctly process "be". Additionally, while many language models only consider preceding context [22, 23], both preceding and following contexts of 'be' within the utterance were included in our study in order to measure local effects before and after 'be'.

2.4.3. Domino Effect

Along with acoustic models, ASRs function on language models which, at their most fundamental levels, compute the probability of a word given its adjacent contexts. If a word is misinferred, it can cause other words to be misinferred as well, creating a *domino effect* in the rest of the inference. Take the intended phrase, "melting ice", and its misinferred version, "cooking rice". If the intended context of "ice" ("melting") has a lower probability, in terms of the acoustic and language models, than the potential candidate word ("cooking"), the subsequent intended word ("ice") following the more likely, but incorrect, candidate ("cooking") would have a *lower* language model probability than the other subsequent candidate word ("rice"). As a result, "ice" would also be misinferred.

For this reason, we sought to examine the effect that the incorrect inference of other words would have on the error rate of "be" within a context. We hypothesized that the more words that are misinferred within the context of "be", the higher the error rate of "be". This effect could potentially be a result of either the preceding or following context of "be" or both. Conversely, it is also possible that the accuracy of "be" could affect the error rate of other words occurring within its local context. Thus, we also hypothesized that if "be" were misinferred, then error rates for words surrounding "be" would be higher. This effect could potentially affect either the preceding or following context of "be" or both.

2.4.4. Speech Rate

Faster rates of speech [24, 25] and sometimes very slow rates of speech [24, 26] have been shown to correlate with higher error

rates. We hypothesized, then, that the higher the speech rate, the higher the error rate will be for ASR inferences. Speech rate was calculated as the number of syllables per second in each utterance where utterance is defined as a breath group (see Sec. 2.1) following common practice in phonetic research [27].

2.4.5. Signal to Noise Ratio

A frequent and complex issue in ASR performance is the amount of noise contained within input audio. The speech-to-noise ratio (SNR) of a given speech signal is a commonly used measure that determines the amount of noise that is included in the signal [28]. To understand how noise may have impacted ASR inferences in our study, we included an examination of the SNR levels of each audio file. We hypothesized that the higher the SNR level, the higher the error rate will be for both "be" and its surrounding context. The SNR of a signal without a reference noise file was estimated with the Waveform Amplitude Distribution Analysis (WADA-SNR) [29] with the Matlab implementation by Ellis [30].

3. Experiment

3.1. Experiment 1: How error prone is habitual "be"?

Here, we assessed how error prone habitual "be" would be within ASR inferences and which variables of interest had the most significant impacts. Mixed effects logistic regression (*glmer* in the *lme4* library [31]) was used. For each of the two ASRs (DeepSpeech and Google) with the two signal types (utterance and turn), a model predicting the accuracy of "be" was fitted with the fixed effects outlined in Section 2.4 as well as speaker as a random intercept. Categorical variables were sum-coded and continuous variables were z-score normalized, except for word count which was also log-transformed (base-10).

3.2. Experiment 2: How error prone are words surrounding habitual "be"?

Here, we turned our attention to the words surrounding habitual "be" and assessed whether they would be more error prone than those surrounding non-habitual "be", while also investigating which variables of interest had the most significant impacts. Linear mixed-effects regression (*lmer*) was used. For each of the two ASRs (DeepSpeech and Google) with the two signal types (utterance and turn), two models were fitted to analyze the WER separately for words preceding "be" and following "be". The model structure is similar to that of Experiment 1 with a few differences: 1) the accuracy of "be" was included as a fixed effect, and 2) the WER and the word count variables of the same direction as the dependent variable were excluded.

4. Results

4.1. Descriptive statistics

Our results show that both DeepSpeech and Google were much less able to correctly infer habitual "be" than all other types of non-habitual "be", both at the utterance and turn level, as displayed in Table 1. At the utterance level, Google was 2.18 times less capable of correctly inferring habitual "be" than non-habitual "be", while DeepSpeech was 4.92 times less capable. For turns, Google was 2.70 times less capable of correctly inferring habitual "be" than non-habitual "be", while DeepSpeech was 3.31 times less capable.

Further, as shown in Table 2, words surrounding habitual "be" in DeepSpeech were 1.22 times more likely to be erroneous than words surrounding non-habitual "be" in utterances, and 1.27 more likely in turns. For Google, words surrounding habitual "be" were, on average, 1.43 times more likely to be erroneous in utterances and 1.56 times more likely in turns. Overall then, both the inferences of habitual "be" and the words surrounding them are much more error prone.

Table 1: *Accuracies of habitual "be" and non-habitual "be"*

	Utterance		**Turn**	
	DeepSpeech	*Google*	*DeepSpeech*	*Google*
Non-habitual	39.21%	71.58%	49.32%	76.90%
Habitual	7.97%	32.71%	14.88%	28.53%
Bias ratio (Non-hab/Hab)	4.92	2.18	3.31	2.70

Table 2: *WERs of words surrounding habitual "be" and non-habitual "be"*

	Utterance		**Turn**	
	DeepSpeech pre, post	*Google* pre, post	*DeepSpeech* pre, post	*Google* pre, post
Non-habitual	0.60, 0.56	0.37, 0.35	0.46, 0.48	0.28, 0.29
Habitual	0.73, 0.68	0.55, 0.48	0.59, 0.61	0.43, 0.45
Bias ratio (Hab/Non-hab)	1.22, 1.22	1.47, 1.39	1.27, 1.27	1.53, 1.56

4.2. Statistical Analysis: Experiment 1

4.2.1. DeepSpeech

Table 3 summarizes the fixed effects of DeepSpeech's accuracy of "be" with utterance and turn-level input. The results were similar across each. All variables were significant, except for speech rate in the turn condition and SNR. The key variable, habituality, was significant in the negative direction ($\beta_{utt.}$: -1.676, β_{turn}: -1.366) suggesting that habitual "be" is more error prone than non-habitual "be". The word count variables were significant suggesting an effect of context size on the accuracy of "be". However, the two word count variables act in opposite direction with word count following "be" having a negative effect on accuracy ($\beta_{utt.}$: -0.098, β_{turn}: -0.097), a surprising finding. WER has a consistent negative effect on accuracy from both sides of "be". With both the word count and WER variables, those preceding "be" have a stronger effect than those following "be" by comparing their absolute β values. This suggests that the negative effect of the language model has on accuracy is asymmetrical. Speech rate has a negative effect on accuracy ($\beta_{utt.}$: -0.157) at the utterance-level.

4.2.2. Google

Table 4 summarizes the fixed effects of Google's accuracy of "be" with both utterance and turn-level input. While the effect of habituality remains significant in the negative direction ($\beta_{utt.}$: -1.279, β_{turn}: -1.966), the other variables of interest behave differently with Google compared to the DeepSpeech. Word count following "be" is only significant at the turn level, while word count preceding "be" is only significant at the utterance level. WER has a consistent negative effect on accuracy from both sides of "be"; however, unlike DeepSpeech, WER following "be" has a *stronger* effect than WER preceding "be".

Table 3: *Fixed effects summaries of DeepSpeech's accuracy of "be" with utterance- and turn-level input*

	Utterance			Turn		
	β	*z-value*	sig.	β	*z-value*	sig.
Habituality	**-1.676**	-7.274	***	**-1.366**	-7.508	***
Word count (pre-"be")	**0.531**	10.077	***	**0.114**	2.355	*
Word count (post-"be")	**-0.098**	-1.890	·	**-0.097**	-1.960	*
WER (pre-"be")	**-1.193**	-21.205	***	**-0.844**	-16.631	***
WER (post-"be")	**-0.926**	-16.939	***	**-0.738**	-14.111	***
Speech Rate	**-0.157**	-2.725	**	-0.075	-1.424	n.s.
SNR	-0.024	-0.432	n.s.	0.076	1.334	n.s.

*** (p <0.001), ** (p <0.01), * (p <0.05), · (p <0.1) and n.s. (p >0.1)

Speech rate was not significant. SNR was only nominally significant at the turn level.

Table 4: *Fixed effects summaries of Google's accuracy of "be" with utterance- and turn-level input*

	Utterance			Turn		
	β	*z-value*	sig.	β	*z-value*	sig.
Habituality	**-1.279**	-8.063	*******	**-1.970**	-12.575	***
Word count (pre-"be")	**0.330**	5.695	***	0.085	1.576	n.s.
Word count (post-"be")	-0.084	-1.445	n.s.	**-0.301**	-5.314	***
WER (pre-"be")	**-0.988**	-17.385	***	**-0.759**	-14.559	***
WER (post-"be")	**-1.011**	-17.461	***	**-0.851**	-15.515	***
Speech Rate	-0.093	-1.553	n.s.	-0.005	-0.089	n.s.
SNR	-0.016	-0.267	n.s.	**0.105**	1.693	·

*** (p <0.001), ** (p <0.01), * (p <0.05), · (p <0.1) and n.s. (p >0.1)

4.3. Statistical Analysis: Experiment 2

4.3.1. DeepSpeech

Table 5 summarizes the fixed effects of DeepSpeech's WER of words preceding and following "be" with both utterance and turn-level input. Habituality is significant in the positive direction, suggesting that habitual "be" increases the WER of words; however, it only affects post-"be" WER ($\beta_{utt.}$: 0.069, β_{turn}: 0.126). Accuracy of "be" has a negative effect on WER of words before and after "be" for both utterance and turn levels. Word count is significant only when preceding "be" ($\beta_{utt.}$: 0.020, β_{turn}: 0.016), suggesting that the preceding context has a stronger effect on the accuracy of the following words than the following context has on the preceding words. WER on the opposite side of "be" has a significant effect but only at the turn level (β_{pre}: 0.025, β_{post}: 0.037). Speech Rate is significant across conditions in a positive direction – the higher speech rate, the higher the WER. SNR had no effect on WER.

4.3.2. Google

Table 6 summarizes the fixed effects of Google's WER of words preceding and following "be" with both utterance and turn-level input. The overall patterns are similar to those of DeepSpeech with two notable differences: 1) habituality is significant in the positive direction only at the utterance level for WER of both sides of "be" (β_{pre}: 0.091, β_{post}: 0.041) and 2) word count preceding "be" is not significant.

Table 5: *Fixed effects summaries of DeepSpeech's pre-"be" and post-"be" WER with utterance- and turn-level input*

	Utterance		Turn	
	Pre-"be"	Post-"be"	Pre-"be"	Post-"be"
	β	β	β	β
Habituality	$0.012^{n.s.}$	$\mathbf{0.069}^{***}$	$-0.032^{n.s.}$	$\mathbf{0.126}^{***}$
"be"-accuracy	$\mathbf{-0.308}^{***}$	$\mathbf{-0.311}^{***}$	$\mathbf{-0.246}^{***}$	$\mathbf{-0.298}^{***}$
Word count (pre-"be")	—	$\mathbf{0.020}^{**}$	—	$\mathbf{0.016}^{\cdot}$
Word count (post-"be")	$-0.005^{n.s.}$	—	$0.006^{n.s.}$	—
WER (pre-"be")	—	$0.004^{n.s.}$	—	$\mathbf{0.037}^{***}$
WER (post-"be")	$0.005^{n.s.}$	—	$\mathbf{0.025}^{***}$	—
Speech Rate	$\mathbf{0.038}^{***}$	$\mathbf{0.036}^{***}$	$\mathbf{0.028}^{***}$	$\mathbf{0.059}^{***}$
SNR	$-0.001^{n.s.}$	$0.007^{n.s.}$	$0.005^{n.s.}$	$0.012^{n.s.}$

*** (p <0.001), ** (p <0.01), * (p <0.05), · (p <0.1) and n.s. (p >0.1)

Table 6: *Fixed effects summaries of Google's pre-"be" and post-"be" WER with utterance- and turn-level input*

	Utterance		Turn	
	Pre-"be"	Post-"be"	Pre-"be"	Post-"be"
	β	β	β	β
Habituality	$\mathbf{0.091}^{***}$	$\mathbf{0.041}^{\cdot}$	$0.020^{n.s.}$	$0.033^{n.s.}$
"be"-accuracy	$\mathbf{-0.322}^{***}$	$\mathbf{-0.348}^{***}$	$\mathbf{-0.244}^{***}$	$\mathbf{-0.366}^{***}$
Word count (pre-"be")	—	$0.008^{n.s.}$	—	$0.010^{n.s.}$
Word count (post-"be")	$0.008^{n.s.}$	—	$-0.010^{n.s.}$	—
WER (pre-"be")	—	$0.011^{n.s.}$	—	$\mathbf{0.068}^{***}$
WER (post-"be")	$0.009^{n.s.}$	—	$\mathbf{0.045}^{***}$	—
Speech Rate	$\mathbf{0.038}^{***}$	$\mathbf{0.040}^{***}$	$\mathbf{0.029}^{***}$	$\mathbf{0.050}^{***}$
SNR	$-0.006^{n.s.}$	$-0.006^{n.s.}$	$0.002^{n.s.}$	$0.002^{n.s.}$

*** (p <0.001), ** (p <0.01), * (p <0.05), · (p <0.1) and n.s. (p >0.1)

5. Conclusions

Based on our results, habitual "be" and its surrounding words seem much more error prone than non-habitual "be". Across the board, habituality was a significant predictor of these errors. The effect of preceding context and word error was higher than in the other direction, likely due to language models which only take into account the context of preceding words. Lastly, acoustic factors did not have a consistent effect in our experiments.

Beyond these findings, our outcomes reveal that many ASR systems may be biased not only against the acoustic aspects of AAE [2], but also morpho-syntactic features as well. Evidence shown here suggests that ASR systems may work much more poorly for speakers who utilize unique grammatical aspects of AAE such as habitual "be".

Finally, our results suggest that word error rate may not be the best measure for ASR accuracy. In traditional WER, all words are equal. However, as shown here, a word with a specific morpho-syntactic category can have a greater impact than what traditional WER captures.

6. Acknowledgements

We would like to thank Dr. Galia Hatav and Dr. James Garner for their discussions on the syntactic features, Dr. Ratree Wayland for her feedback on the paper, and Halee Corbin for her techinical support with the SNR computation.

7. References

[1] J. Baugh, "Linguistic profiling and discrimination," *The Oxford handbook of language and society*, pp. 349–368, 2016.

[2] 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.

[3] R. Tatman and C. Kasten, "Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions." in *INTERSPEECH*, 2017, pp. 934–938.

[4] R. Dorn, "Dialect-specific models for automatic speech recognition of african american vernacular english," in *Student Research Workshop*, 2019, pp. 16–20.

[5] L. J. Green, *African American English: a linguistic introduction*. Cambridge University Press, 2002.

[6] T. Kendall and C. Farrington, "The corpus of regional african american language," *Version*, vol. 6, p. 1, 2018.

[7] Y. Jadoul, B. Thompson, and B. De Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[8] Google cloud speech-to-text. [Online]. Available: https://cloud.google.com/speech-to-text/

[9] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4774–4778.

[10] 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.

[11] R. Morais. Deepspeech 0.6: Mozilla's speech-to-text engine gets fast, lean, and ubiquitous. [Online]. Available: https://hacks.mozilla.org/2019/12/deepspeech-0-6-mozillas-speech-to-text-engine/

[12] 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.

[13] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[14] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text." in *LREC*, vol. 4, 2004, pp. 69–71.

[15] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1992, pp. 517–520.

[16] M. Honnibal and I. Montani, "spacy 2: Natural language understanding with bloom embeddings," *Convolutional Neural Networks and Incremental Parsing*, 2017.

[17] R. A. Wagner and M. J. Fischer, "The string-to-string correction problem," *Journal of the ACM (JACM)*, vol. 21, no. 1, pp. 168–173, 1974.

[18] K. Gorman, "wagnerfischerpp.py," https://gist.github.com/kylebgorman/8034009, 2013.

[19] P. Aleksic, M. Ghodsi, A. Michaely, C. Allauzen, K. Hall, B. Roark, D. Rybach, and P. Moreno, "Bringing contextual information to google speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[20] K. Hall, E. Cho, C. Allauzen, F. Beaufays, N. Coccaro, K. Nakajima, M. Riley, B. Roark, D. Rybach, and L. Zhang, "Composition-based on-the-fly rescoring for salient n-gram biasing," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[21] J. Scheiner, I. Williams, and P. Aleksic, "Voice search language model adaptation using contextual information," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 253–257.

[22] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Computational linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[23] Y. Halpern, K. B. Hall, V. Schogol, M. Riley, B. Roark, G. Skobeltsyn, and M. Baeuml, "Contextual prediction models for speech recognition." in *INTERSPEECH*, 2016, pp. 2338–2342.

[24] M. A. Siegler and R. M. Stern, "On the effects of speech rate in large vocabulary speech recognition systems," in *1995 international conference on acoustics, speech, and signal processing*, vol. 1. IEEE, 1995, pp. 612–615.

[25] E. Fosler-Lussier and N. Morgan, "Effects of speaking rate and word frequency on pronunciations in convertional speech," *Speech Communication*, vol. 29, no. 2-4, pp. 137–158, 1999.

[26] T. Shinozaki and S. Furui, "Error analysis using decision trees in spontaneous presentation speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU'01.* IEEE, 2001, pp. 198–201.

[27] N. H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior research methods*, vol. 41, no. 2, pp. 385–390, 2009.

[28] P. N. Garner, "Snr features for automatic speech recognition," in *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2009, pp. 182–187.

[29] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[30] S. R. Quackenbush, "Objective measures of speech quality," Ph.D. dissertation, Georgia Institute of Technology, 1995.

[31] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.