INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Exploring Sources of Racial Bias in Automatic Speech Recognition through the Lens of Rhythmic Variation

Li-Fang Lai[1], *Nicole Holliday*[1]

[1]Department of Linguistics and Cognitive Science, Pomona College, USA
{li-fang.lai, nicole.holliday}@pomona.edu

Abstract

Although studies have shown that one issue of bias in modern automatic speech recognition (ASR) technologies is degraded performance for African American English (AAE) speakers, the mechanism by which systems fail for AAE speakers is still not well-understood. The present study aims to offer insight into this issue by examining whether errors are driven by rhythmic variation in ethnolects. We computed seven quantitative measures of speech rhythm in a reading task as produced by AAE and General American English (GAE) speakers and related these metrics to word error rates. The results confirmed racial bias against AAE speakers with higher error rates when AAE speakers produced more variable durations in vowel sounds. Rhythmic variation, on the other hand, is not a contributing factor for the errors in GAE. The result calls for interdisciplinary collaboration between linguists and ASR builders to add timing components of speech to the system to ensure fairness in artificial intelligence for currently underserved groups.

Index Terms: automatic speech recognition, African American English, rhythmic variation, racial bias, fairness in artificial intelligence

1. Introduction

As ASR technologies have become an increasingly important part of everyday life in the U.S., researchers have begun to voice their concerns about unfairness in artificial intelligence. One focal topic of discussion is algorithmic bias as the systems consistently returned degraded performance across regional varieties of AAE as compared to GAE [1-5]. Although it is widely acknowledged that ASR models can benefit from training data that consist of diverse speaker groups, linguists can also contribute positively to improved ASR systems by using what we know about sociolinguistic variation to help address recognition errors. For instance, Martin and Tang [3] highlighted the need for ASR systems to consider AAE morphosyntactic features in the language models as they observed that habitual "be" and its surrounding words were more error-prone than non-habitual "be" and its surrounding words. Koenecke et al., [2] and Wassink et al., [5] found that the misrecognitions in AAE were mainly driven by segmental variation, suggesting that the errors lie with the acoustic model. Martin and Wright [4] noted that errors were jointly triggered by morphosyntactic and phonological features of AAE.

Considering that acoustic modeling for ASRs currently uses very little speech production knowledge aside from spectral features [6], this study seeks to explore whether other aspects of speech production, such as timing characteristics of speech, help account for the performance gap between ethnolects of American English. In particular, we focus on differences in word error rate and rhythmic properties between AAE and GAE.

The remainder of the introduction describes approaches to speech rhythm analysis (section 1.1). We then review literature on American English to discuss ethnolectal differences in rhythmic structure (section 1.2) and the impact of rhythmic variation on ASR performance (section 1.3) to provide context for this study's aims (section 1.4).

1.1. Quantitative analysis of speech rhythm

Research on speech rhythm often describes the world's languages as stress-timed, syllable-timed, and mora-timed based on their duration measurements, though there is substantial variation within these categories [7-13]. The metrics frequently used for statistical measurement of durational variability in speech are summarized in Table 1. In principle, for all metrics but %V, a larger value represents greater durational variability.

Table 1: *Rhythm metrics.*

Metrics	Definition
%V	The percentage of the total duration of vocalic intervals within an utterance [14]
ΔV	The standard deviation of the duration of vocalic intervals within an utterance [14]
ΔC	The standard deviation of the duration of consonantal intervals within each utterance [14]
VarcoV	ΔV normalized for speaking rate [15]
VarcoC	ΔC normalized for speaking rate [15]
nPVI-V	Durational variability in successive pairs of vocalic intervals which normalized for speaking rate [16]
rPVI-C	Durational variability in successive pairs of consonantal intervals based on the raw values [17]

Acoustic analyses show that more stress-timed languages like English, German, and Dutch exhibit greater durational variability than more syllable-timed languages such as French, Spanish, and Mandarin Chinese. The values obtained from Japanese, a mora-timed language, are similar to those from syllable-timed languages [14, 15, 17]. These timing properties differ between languages, to the extent that English, Korean, Mandarin, and Spanish can be distinguished by machine based on language-specific rhythm features alone [18].

Speech rhythm differs within languages as well. Contributing factors for rhythmic variation in American English include geographic location, communicative setting, and speaker ethnicity. We review several studies on ethnolectal differences in rhythmic production in the following section.

10.21437/Interspeech.2023-159

1.2. Ethnolectal rhythmic variation

Although the ways in which rhythmic structure varies with ethnic groups are not extensively studied, previous studies have identified rhythmic variation across ethnolects of American English. For example, Carter [19] found that the mean PVI scores for the Hispanic English speakers fell well below those for the European American/White speakers, indicating more syllable timing for Hispanic English than for White English.

In cross-variety comparisons between Native American English and White English, Coggshall [20] reported that Eastern Cherokee English speakers had smaller median PVI scores (more syllable-timed) than European American English speakers. Lumbee English speakers displayed a rather different pattern: the older Lumbee had median PVI values similar to European Americans. Younger speakers, on the other hand, had shifted towards a more syllable-timed rhythm.

For AAE, Thomas and Carter [21] measured PVI values for recordings produced by African American and European American speakers. The results showed that speech from African Americans born before the American Civil War (1861-1865) was characterized by less variable durations, suggesting that it was more syllable-timed at that point in history. Present-day AAE, however, exhibits highly varied durations and has become more stress-timed, with no significant difference AAE and White English. Research has also shown that AAE rhythm varies with region and may show stylistic differences. For the former, Gilbert et al. [22] found more dynamic rhythmic patterns (greater nPVI-V scores) in West Coast AAE speech and rap flows than those produced by East Coast hip hop artists. With respect to style shift, Nielsen [23] noted that speaker's rhythm in dialogue is more stress-timed (higher mean PVIs) compared to non-dialogue narrative discourse.

1.3. Rhythmic variation and ASR performance

Given that rhythm is an integral element of human speech, researchers have also sought to investigate the effect of rhythmic variation on ASR performance. Lai and Holliday [24] for instance, argued that utterances produced with greater VarcoV values are correlated with higher WER in a reading task produced by AAE speakers. The relationship between rhythm and ASR performance, however, appears less evident in a reading task produced by GAE speakers [25]. This strand of research, though still in its incipient stages, carries great potentials to inform us why systems fail systematically for certain ethnolects and groups of speakers.

1.4. Research aims

The overarching goal of this study is to explore whether the rhythmic production in different ethnolects is involved in determining speech recognition performance. To that end, we test whether the ASR system exhibits similar bias against AAE speakers as has been reported in previous work. If so, we investigate whether AAE is rhythmically distinct from GAE and subsequently examine potential effects of ethno-rhythmic variation on ASR performance.

2. Methods

The data analyzed here were collected as part of larger projects investigating prosodic variation in American English. All participants completed a sociolinguistic interview, followed by multiple scripted speech tasks including reading passages, sentences, and isolated words. The present study analyzed data from all speakers reading the first paragraph of the Rainbow Passage [26], in order to control for possible influences of segmental and phrase-level features and make comparisons between speakers.

2.1. Participants

Two samples of participants were recruited for this study. First, we recruited 19 native speakers of General (White) American English (9 females and 10 males, ages 19-61) residing in Pennsylvania and Tennessee for an online speech production study in 2021. The experimenter and the participants met in Zoom [27] and the audio recordings were created using Zencastr [28]. We also recruited 20 high school students (12 females and 8 males, ages 15-19) identified as Black speakers of AAE from Pennsylvania from 2021 to 2022. The audio recordings took place in their schools.

In both cases, the study was conducted by an experimenter who shared the same racial background with the participant to control in part for potential effects of ethnolinguistic accommodation [29].

2.2. Data processing

2.2.1. Segmentation

The speech stream was split into individual intonational phrases (IPs), signaled by a perceivable following pause longer than 100 ms [30-33]. IPs with disfluencies/hesitation, self-correction, or background noise were eliminated, yielding 496 (279 AAE + 217 GAE) IPs for subsequent analysis.

Then, each IP was segmented into individual consonantal or vocalic phonemes using both auditory impression and visual inspection of the waveforms and spectrograms in Praat [34]. Following standard segmentation criteria [14, 21, 30], sequences of consonants or vowels were transcribed as a single interval and intervocalic rhotic /r/s were treated as part of the vocalic interval. A Praat script [35] was used to extract the interval durations. The seven rhythm metrics (%V, ΔV, ΔC, VarcoV, VarcoC, nPVI-V and rPVI-C) were computed for each IP in Python [36].

2.2.2. Evaluation of ASR performance

The audio recordings were processed by Deepgram [37], an end-to-end deep learning speech recognition platform that generates accurate and human-readable transcripts. The ASR transcript was aligned and segmented into IPs based on the ground-truth transcript. The word error rate (WER), defined as the number of word substitutions, deletions, and insertions in the system output, divided by the total number of words in the ground truth [38], was calculated for each IP using the `pywer` package [39] in Python [36].

3. Analyses and results

Three sets of analyses were conducted: the effect of ethnicity on ASR performance (section 3.1), rhythmic variation within ethnolects (section 3.2), and the link between ethno-rhythmic variation and ASR performance (section 3.3). We discuss the statistical analysis and the results in more detail in each section.

3.1. Ethnic bias in ASR performance

To test the effect of ethnolect on ASR performance, we fitted a linear mixed-effect model (LMM) for `WER`, with

`ethnolect` as a fixed effect and `speaker` as a random intercept using the `lme4` [40] and the `lmerTest` [41] packages in R [42].

The result revealed a significant effect of ethnolect on ASR performance as the WER for AAE speakers is higher than that of GAE speakers (β = 8.39, t = 3.72, p < .001, R^2 = 0.42, Figure 1).

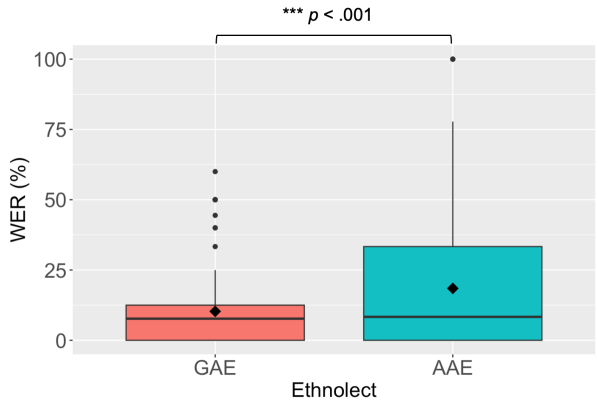

Figure 1: *WER by ethnolect (diamonds = means).*

3.2. Ethnolectal differences in speech rhythm

Next, we tested whether the speech timing is ethnolinguistically-distinct. To address issues of collinearity between the rhythm measures, we fitted separate linear mixed-effect models for `%V`, `ΔC`, `ΔV`, `VarcoC`, `VarcoV`, `nPVI-V` and `rPVI-C`, with `ethnolect` as a fixed effect and a random intercept by `speaker` in R [42].

The analyses showed that when normalized for speaking rate, AAE speakers produced smaller VarcoV values (i.e., less variable durations in vocalic intervals) than GAE speakers (β = -6.46, t = -3.52, p = <.001, R^2 = -0.32). The analyses additionally returned marginal significance of ethnolect on speech rhythm as AAE speakers produced smaller VarcoC (β = -3.17, t = -2.0, p = .052, R^2 = -0.23) and nPVI-V (β = -4.12, t = -1.93, p = .055, R^2 = -0.17) scores than GAE speakers (Figure 2). No significant differences were found between the two groups in terms of their %V (p = .243), ΔC (p = .784), ΔV (p = .485), and rPVI-C (p = .187) values.

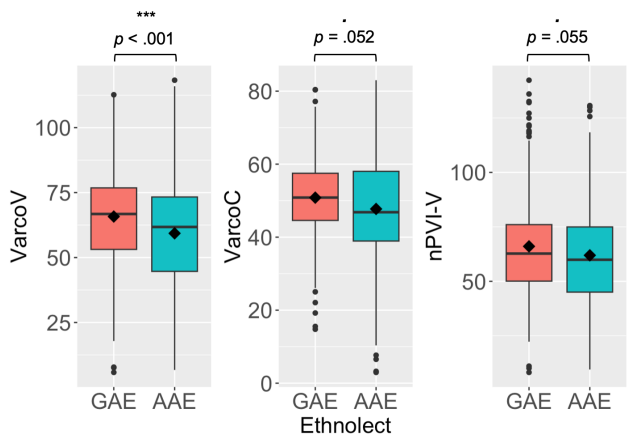

Figure 2: *Rhythmic variation by ethnolect (diamonds = means).*

3.3. Ethno-rhythmic variation and ASR performance

Since the results from section 3.2 only revealed ethnolectal differences in VarcoV, VarcoC, and nPVI-V, our analysis here will focus on the three metrics. We fitted separate simple linear models to predict WER based on each rhythm measure as produced by different ethnicities (formula: `lm(WER ~ VarcoV/VarcoC/nPVI-V*ethnolect)`). In each test, influential data points were removed using the `influencePlot()` function in the `car` package [43]. The data points for each test are: VarcoV = 492, VarcoC = 491, and nPVI-V = 491. In each of the models, the IPs were treated as independent observations as we did not find significant correlation of WER between successive pairs of IPs (r = 0.09).

The analyses returned a significant interaction effect between rhythm metric and ethnicity on recognition performance (Figure 3), in which the WER increased when AAE speakers produced greater VarcoV scores (F(3, 488) = 16.20, p < .001, R^2 = 0.45). GAE speakers exhibited a reverse trend as their WERs *decreased* when they produced larger VarcoVs. No significant interaction effects were found for VarcoC (p = .967) and nPVI-V (p = .092).

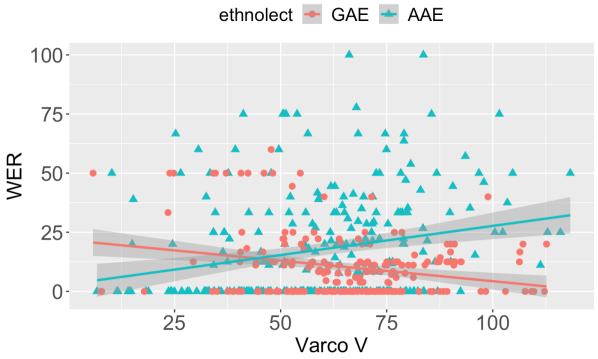

Figure 3: *Interaction plot for WER based on VarcoV and ethnicity.*

4. Discussion

The present study aims to examine whether differences in rhythmic variation between ethnolects impact speech recognition performance. Three sets of analysis were conducted to answer this question. First, we tested whether there is a performance gap between the AAE and GAE samples. The result echoes previous work [1-5] by showing bias against AAE speakers – the system generated significantly higher WERs for AAE than for GAE (Figure 1). Recent research has also begun to leverage sociolinguistic knowledge to probe the mechanism by which these systems fail for AAE speakers. For instance, in-depth analyses of recognition errors [2, 5] have come to a similar conclusion that phonological features characteristic of AAE may play a bigger role than morphosyntactic structure in triggering misrecognitions.

In addition to segmental variation, recent work by Lai and Holliday [24] points out another promising direction for ASR research, arguing that for AAE speakers, timing properties of speech also impact recognition performance. The present study extends this line of research by studying the rhythmic structure produced by AAE and GAE speakers and relating their rhythmic production to ASR performance. Unlike Thomas and Carter [21] who reported that contemporary AAE and GAE spoken in North Carolina are rhythmically indistinguishable,

our analyses reveal that the two groups of speakers patterned differently because AAE speakers exhibited smaller VarcoV, VarcoC, and nPVI-V scores than GAE speakers (Figure 2).

These differences in prosodic rhythm may be explained through the phonological properties characteristic of AAE. For example, although both groups of speakers were recorded reading the same material, diphthongs such as /aɪ, aʊ, ɔɪ/ are more likely to be monophthongized in AAE than in GAE speech. Given that monophthongs are acoustically shorter than diphthongs [44], monophthongal /aɪ, aʊ, ɔɪ/ may therefore lessen durational variability in vocalic intervals, leading to smaller VarcoVs and nPVI-Vs in AAE. Smaller VarcoC scores for AAE speakers may have to do with syllable complexity – AAE speakers show final consonant cluster reduction more often than GAE speakers. Simpler syllable structure may then decrease their VarcoC values [15].

Lastly, we examined whether there is a link between rhythmic production and ASR performance. The analysis revealed an interesting interaction effect in which WER increased when AAE speakers produced more variable vowel durations (larger VarcoV values). This is in line with [24], showing that VarcoV is a more reliable predictor of recognition performance, at the least in read speech. This relationship, however, is not seen in GAE, which indicates that recognition errors in GAE are not driven by speech timing. Although our analysis did not reveal clear tendencies between VarcoC and ASR performance and between nPVI-V and ASR accuracy as produced by different ethnic groups, these interaction effects may become relevant when we include interview speech in future analysis, since variation in the timing properties is more exaggerated in natural spontaneous speech than in read speech [30].

5. Conclusions

This study presents an initial inquiry into the relationship between ethnolinguistically-conditioned rhythmic variation and ASR performance. The analysis lends support to the argument that rhythmic structure can be an area of challenge for the machines because speech produced with more variable vowel durations is more error-prone, especially for AAE speakers. Given that rhythm is an integral part of speech production which is influenced by factors pertaining to speaker ethnicity, geographic location, and communicative setting, we believe adding timing components to acoustic models is a crucial step to improve the fairness of ASR systems.

The present study, coupled with [21, 22], can also serve as a starting point for cross-regional comparison of AAE rhythm as AAE is not monolithic across Black communities. The results can generate insights into AAE prosody, a severely under-theorized area in both linguistics and ASR literature. We are also interested in exploring whether creaky voice/vocal fry, a type of phonation produced with irregular vocal fold vibration [45, 46] and is characterized by specific spectral structure [47], may have an impact on ASR performance, especially in an interaction with rhythmic properties. The combined results on rhythm and voice quality can establish a much-needed empirical basis for programmers and ASR architects to fine-tune and optimize their models, ensuring that voice technology is accessible to AAE users. Broadly, these results can also benefit other speaker groups, such as Southern English speakers who are also highly variable in their vowel durations [30].

6. Acknowledgements

We would like to thank our participants for sharing their time and voices with us. We would also like to thank Carly Danielson, Aleah Combs, Razan Osman, Yendi Guindo, Tyeirra Lynch, and Emelia Benson Meyer for conducting interviews and managing data. Funding for this study was provided in part by the Center for Language Science at the Pennsylvania State University to Li-Fang Lai and by Spencer Foundation grant #202100096 to Nicole Holliday and Sabriya Fisher.

7. References

[1] R. Tatman and C. Kasten, “Effects of talker dialect, gender & race on accuracy of Bing speech and YouTube automatic captions,” *Proceeding of Interspeech 2017*, 2017, pp. 934–938.

[2] A. Koenecke *et al.*, “Racial disparities in automated speech recognition,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[3] J. L. Martin and K. Tang, “Understanding racial disparities in automatic speech recognition: The case of habitual ‘be’,” *Proceedings of Interspeech 2020*, 2020, pp. 626–630.

[4] J. L. Martin and K. E. Wright, “Bias in automatic speech recognition: The case of African American language,” *Applied Linguistics*, Article amac066, 2022.

[5] A. B. Wassink, C. Gansen, and I. Bartholomew, “Uneven success: Automatic speech recognition and ethnicity-related dialects,” *Speech Communication*, vol. 40, pp. 50–70, 2022.

[6] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, “Speech production knowledge in automatic speech recognition,” *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

[7] K. L. Pike, *The Intonation of American English*, 2nd ed. Ann Arbor: University of Michigan Press, 1946.

[8] B. Bloch, “Studies in Colloquial Japanese IV Phonemics,” *Language*, vol. 26, no. 1, pp. 86–125, 1950.

[9] M. S. Han, “The feature of duration in Japanese,” *Onsei no kenkyuu*, vol. 10, pp. 65–80, 1962.

[10] D. Abercrombie, *Studies in Phonetics and Linguistics*. Oxford: Oxford University Press, 1965.

[11] D. Abercrombie, *Elements of General Phonetics*. Edinburgh: Edinburgh University Press, 1967.

[12] P. Ladefoged, *A Course in Phonetics*. New York: Harcourt Brace Jovanovich, 1975.

[13] R. F. Port, J. Dalby, and M. O’Dell, “Evidence for mora timing in Japanese,” *The Journal of the Acoustical Society of America*, vol. 81, no. 5, pp. 1574–1585, 1987.

[14] F. Ramus, M. Nespor, and J. Mehler, “Correlates of linguistic rhythm in the speech signal,” *Cognition*, vol. 73, no. 3, pp. 265–292, 1999.

[15] V. Dellwo, “Rhythm and speech rate: A variation coefficient for deltaC,” in *Language and language-processing*, P. Karnowski and I. Szigeti, Eds. Frankfurt am Main: Peter Lang, 2006, pp. 231–241.

[16] E. L. Low, E. Grabe, and F. Nolan, “Quantitative characterizations of speech rhythm: Syllable-timing in Singapore English,” *Language and Speech*, vol. 43, no. 4, pp. 377–401, 2000.

[17] E. Grabe and E. L. Low, “Durational variability in speech and the rhythm class hypothesis,” in *Laboratory Phonology 7*, C. Gussenhoven and N. Warner, Eds. Berlin: Mouton de Gruyter, 2002, pp. 515–546.

[18] H. Kim and J.-S. Park, “Automatic language identification using speech rhythm features for multi-lingual speech recognition,” *Applied Sciences*, vol. 10, no. 7, Article 2225, 2020.

[19] P. M. Carter, “Quantifying rhythmic differences between Spanish, English, and Hispanic English,” in *Theoretical and Experimental Approaches to Romance Linguistics: Selected Papers from the 34th Linguistic Symposium on Romance Languages*, R. S. Gess and E. J. Rubin, Eds. Amsterdam: Benjamins., 2005, pp. 63–75.

[20] E. L. Coggshall, “The prosodic rhythm of two varieties of Native American English,” *University of Pennsylvania Working Papers in Linguistics*, vol. 14, no. 2, Article 2, 2008.
[21] E. R. Thomas and P. M. Carter, “Prosodic rhythm and African American English,” *English World-Wide*, vol. 27, no. 3, pp. 331–355, 2006.
[22] S. Gilbers, N. Hoeksema, K. de Bot, and W. Lowie, “Regional variation in West and East Coast African-American English prosody and rap flows,” *Language and Speech*, vol. 63, no. 4, pp. 713–745, 2020.
[23] R. Nielsen, “The stylistic use of prosodic rhythm in African American English,” *RASK – International journal of Language and Communication*, vol. 37, pp. 301–334, 2013.
[24] L-F. Lai and N. Holliday. “More than spectral features: The role of rhythmic variation in speech recognition,” under review.
[25] L-F. Lai, J. G. van Hell, and J. Lipski, “The role of rhythm and vowel space in speech recognition,” *Proceedings of Speech Prosody 2022*, 2022, pp. 425–429.
[26] G. Fairbanks, *Voice and Articulation Drillbook.*, 2nd ed. New York: Harper & Row, 1960.
[27] Zoom, “One platform to connect,” *Zoom*. https://zoom.us/.
[28] Zencastr, “Podcasting made easy! Start your podcast today,” *zencastr.com*. https://zencastr.com/.
[29] H. Giles, D. M. Taylor, and R. Bourhis, “Towards a theory of interpersonal accommodation through language: Some Canadian data,” *Language in Society*, vol. 2, no. 2, pp. 177–192, 1973.
[30] C. G. Clopper and R. Smiljanić, “Regional variation in temporal organization in American English,” *Journal of Phonetics*, vol. 49, pp. 1–15, 2015.
[31] A. Henderson, F. Goldman-Eisler, and A. Skarbek, “Sequential temporal patterns in spontaneous speech,” *Language and Speech*, vol. 9, no. 4, pp. 207–216, 1966.
[32] A. Wennerstrom and A. F. Siegel, “Keeping the floor in multiparty conversations: Intonation, syntax, and pause,” *Discourse Processes*, vol. 36, no. 2, pp. 77–107, 2003.
[33] C. G. Clopper and R. Smiljanić, “Effects of gender and regional dialect on prosodic patterns in American English,” *Journal of Phonetics*, vol. 39, no. 2, pp. 237–245, 2011.
[34] P. Boersma and D. Weenink, “Praat: doing Phonetics by Computer [Computer program version 6.2.14],” https://www.fon.hum.uva.nl/praat/
[35] D. McCloy, “getDurationPitchFormants.praat [Praat script],” 2011.
[36] Python, “Welcome to Python.org [Computer program version 3.1.1],” *Python.org*, 2022. https://www.python.org/
[37] Deepgram, “Deepgram - Automated speech recognition,” *Deepgram*. https://deepgram.com/
[38] D. Klakow and J. Peters, “Testing the correlation of word error rate and perplexity,” *Speech Communication*, vol. 38, no. 1–2, pp. 19–28, 2002.
[39] J. Nozaki, “pywer [Python package pywer version 0.1.1],” 2021.
[40] D. Bates, M. Maechler, B. Bolker, and S. Walker, “lme4: Linear mixed-effects models using ‘Eigen’ and S4 [R package lme4 version 1.1-31],” 2022.
[41] A. Kuznetsova, P. B. Brockhoff, R. H. B. Christensen, “Tests in linear mixed effects models [R package lmerTest version 3.1-3],” 2022.
[42] R Core Team, “R: A language and environment for statistical computing [Computer program version 4.2.2],” *R-project.org*, 2022. https://www.r-project.org/
[43] J. Fox and S. Weisberg, “Companion to applied regression [R package car version 3.1-1],” 2022.
[44] K. Tsukada, “An acoustic comparison of English monophthongs and diphthongs produced by Australian and Thai speakers,” *English World-Wide*, vol. 29, no. 2, pp. 194–211, 2008.
[45] M. Gordon and P. Ladefoged. “Phonation types: A crosslinguistic overview,” *Journal of Phonetics*, vol. 29, pp. 383–406, 2001.
[46] A. Ní Chasaide and C. Gobl, “Voice source variation,” in *The Handbook of Phonetic Sciences*, W. J. Hardcastle and J. Laver, Eds. Oxford: Blackwell, 1997, pp. 427–461.
[47] T-J. Yoon, J. Cole, and M. Hasegawa-Johnson, “Detecting non-modal phonation in telephone speech,” *Proceedings of Speech Prosody 2008*, 2008, pp. 33–36.