Perceptual salience of tonal speech errors

Zifeng Liu, Ioana Chitoran, Giuseppina Turco

▶ **To cite this version:**

Zifeng Liu, Ioana Chitoran, Giuseppina Turco. Perceptual salience of tonal speech errors. Speech Prosody 2024, Speech Prosody 2024, ISCA, pp.427-431, 2024, 10.21437/SpeechProsody.2024-87 . hal-04671541

HAL Id: hal-04671541

https://u-paris.hal.science/hal-04671541v1

Submitted on 15 Aug 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Speech Prosody 2024
2-5 July 2024, Leiden, The Netherlands

Perceptual salience of tonal speech errors

Zifeng Liu[1,2], *Ioana Chitoran*[1,2], *Giuseppina Turco*[1,3]

[1]Université Paris Cité, France
[2]Clillac-ARP, France
[3]CNRS Laboratoire de Linguistique Formelle, France
zifeng.liu@etu.u-paris.fr, ioana.chitoran@u-paris.fr, giuseppina.turco@cnrs.fr

Abstract

The present study examines the perceptual salience of tonal speech errors compared to segmental errors (consonant and vowel). Tonal errors are observed less often than segmental errors. We thus hypothesize that tone errors are more easily ignored during transcription tasks because tones may have lower perceptual salience relative to segments. We test this hypothesis in Mandarin, via a number reconstruction task. Sixty-nine Mandarin native listeners heard sequences of numbers in which one number was altered by substituting either its vowel, consonant, or tone. They were asked to identify which number that was. Mandarin listeners identified the original number most accurately when consonants were substituted, and were the least accurate when vowels were substituted. For tone substitution, the accuracy was lower than for consonant substitution, but not significantly different from vowel substitutions. Reaction times to identify a number with tone substitution were comparable to those for other types of substitutions. The results show that, contrary to our hypothesis, tone errors are not perceptually less salient than segmental errors. Specifically, tone errors are as salient as vowel errors and more salient than consonant errors, suggesting a similar phonological status shared by tone, vowel and consonant in constraining word selection.

Index Terms: tone, speech error, speech perception, word reconstruction, V-bias

1. Introduction

Speech errors have been considered an important source of evidence for the psychological reality of linguistic units [1]. Studies in this domain have long analyzed speech errors as categorical misplacement of linguistic units in the serial ordering of segments. Recent articulatory and acoustic studies on speech errors revealed that errors are not simply categorical substitutions, but also gradient and partial intrusions from simultaneously activated units [2]–[4], supporting a dynamic view of speech planning. Since our study is concerned with a comparison of the distribution of the three types of errors, rather than with the fine-grained details of their realization, we assume here a categorical view.

Studies of speech errors in tone languages have focused primarily on the production of errors and collected speech errors by orthographic transcription. The reliability of this methodology has been questioned in the past for the perceptual bias in collecting speech errors [3], [5]. Constrained by the human perceptual system, transcribers/listeners tend to overlook or correct unconsciously particular kinds of errors. Such perceptual bias has been attested in a series of experimental studies that asked participants to shadow or detect consonant or vowel mispronunciations [5], [6]. The perceptual bias is thus demonstrated at the segment level only. To the best of our knowledge, no study of speech errors in tone languages probes the potential perceptual bias in detecting tonal errors. However, tonal errors are reported less often than segmental errors in tone languages [6], [8]. This may suggest that tonal errors are more easily overlooked in perception compared to other types of speech errors. The goal of the current study is thus to investigate whether tonal errors are perceptually less salient than segmental errors.

1.1. Disparity between tonal errors and segmental errors

All the studies on tonal errors investigate the comparability between tonal and segmental errors in terms of two aspects: 1) Can tonal errors be characterized contextually as phonological anticipation, perseveration or exchange, similar to segmental errors? 2) Are tonal errors comparable to segmental errors in terms of frequency of occurrence?

The answers to these questions further lead to two opposite accounts about the role of lexical tones in speech production. One account assumes that tones are units of speech production, and so they are selected in the same way as segments in phonological encoding [7]. As a consequence, tonal errors are supposed to behave similarly to segmental errors, which are mainly due to contextual substitutions, including anticipation, perseveration or exchange. Another account [8] presupposes that tones are not involved in phonological encoding, rather tones are inherent to the metrical frame, which means they are not actively selected. This explains why tonal errors are rare compared to segmental errors [8].

Several studies provide supporting evidence for the first account. According to one of the first studies on Mandarin tonal errors [7], the patterns of tonal errors were similar to those of segment errors. All of the tonal errors were explained as phonological movements like anticipation, perseveration or exchange from neighboring tones, suggesting that tonal errors resulted from the mismatched selection of target tones in speech planning. Subsequent studies of tonal errors from other Chinese languages [9]–[12] also reported comparable patterns between tonal and segmental errors, echoing the claim that tones behave similarly to segments in speech planning.

Despite these converging findings from different tonal languages, [8] pointed out that tonal errors are fairly rare compared to segmental errors. In the corpus of [8], only 24 tonal errors were identified among 987 speech errors. More importantly, the reported tonal errors could not be interpreted only as phonological errors; rather they might be due to errors from other speech processes that did not involve tone selection. Therefore, [8] proposed that lexical tones in Mandarin are not represented and processed similarly to segments. Alternatively,

10.21437/SpeechProsody.2024-87

tones are inherent to the metrical frame, while only segments need to be selected in speech planning and then inserted in this metrical frame. The absence of a tone selection mechanism explains why tonal errors are rare. Results from laboratory speech [13] sustain the proposal of [8]. They used a tongue twister paradigm to elicit tonal and segmental errors in Mandarin. Even under such extreme conditions, the rate of tonal errors was lower than that of segmental errors (3503 segmental errors vs. 1372 tonal errors).

Regardless of the lack of consensus on the phonological nature of tones, all the above-mentioned studies of speech errors reflect the pattern that tones are more resistant to errors than segments. The absolute number of tonal errors is far from comparable to segmental errors.

1.2. Consonant vs. Vowel Asymmetries in Word Reconstruction

Phonological contrasts (i.e., vowel and consonant) are not processed equally in auditory word recognition. [13] explored this asymmetry between consonants and vowels using the word reconstruction task. In this task, after hearing a non-word, native English participants were asked to form a real word by changing either the consonants or vowels of the non-word. They were more accurate and faster when changing the vowel of the non-word than when changing the consonants. When they were only allowed to change the consonants to form a real word, they made more errors and the response times were slower than when they could only change the vowels. Thus, the lexical access of English listeners relies more on consonants than on vowels (dubbed as C-bias). This trend has been repeatedly detected in different languages through different experimental paradigms (see [14] for a recent review).

However, the story is not as neat when lexical tone, the third phonological unit, comes into play. [15] extended the word reconstruction task to Mandarin to investigate the universality of the C-bias. Interestingly, the C-bias disappeared in Mandarin. In contrast to previous studies, Mandarin speakers tended to change the tone when any single sound change was required to turn a non-word into a real word. When only one type of sound had to be changed, vowels yielded the lowest accuracy and the longest reaction times. Changing tones was the most accurate and fastest, followed by changes in consonants. In other words, the vowels, but not the consonants or tones, mostly constrained lexical processing in Mandarin.

As the first study on word reconstruction in Mandarin, the results of [15] provided critical insights into the processing bias among vowels, consonants, and tones in Mandarin. Counter-intuitively, as native speakers of a tone language, Mandarin listeners depend least on lexical tones to turn non-words into real words. Contrary to the findings of [14], the lexical access of Mandarin listeners relies mostly on vowels rather than consonants (V-bias). As identifying speech errors involves reconstructing meaningful words from nonsense words, and considering the relative rarity of reported tonal errors, it is still unclear if this processing bias is subject to the performance of native-speaker transcribers in detecting tonal errors.

The present study used a Number Reconstruction Task to investigate the impact of processing bias on perceiving different types of speech errors. We restricted the task to a subset of the lexicon by using digit numbers as stimuli. We turned these numbers into non-number words by substituting their vowel, consonant, and tone to create different types of speech errors.

Our first prediction, in line with previous work, is the following: if tonal errors are less salient than segmental errors, native Mandarin listeners are more likely to ignore them, and will reconstruct the numbers with tone substitutions with higher accuracy and faster reaction times than when reconstructing the same numbers modified by consonant and vowel substitution. Alternatively, if tonal errors are perceived comparably to segmental errors, we expect that listeners' performance on numbers with tone substitution will not be significantly different from segmental substitutions. We still expect that numbers with vowel substitution will be reconstructed more slowly and less accurately than those with consonant substitution.

2. Methods

2.1. Participants

Sixty-nine native speakers of Mandarin (Age: 18-37; mean=24; SD=3.6) were recruited via social media. All the participants were from Mainland China and self-reported using Mandarin as their L1 and dominant daily language. Before the experiment, all the participants read and signed a consent form. The study was approved by the Research Ethics Committee of Université Paris Cité (n° IRB 00012021-130).

2.2. Materials

Four one-digit numbers (0, 3, 6, 9), which are all monosyllabic and high-frequency words in Mandarin, were used as test words. Each number bears one of four Mandarin lexical tones. Each number undergoes three substitutions of tone, onset consonant, and vowel to create different types of errors (Table 1). The experimental material was grouped into four conditions: a) no error condition (N), where all the numbers were pronounced correctly; b) tone error condition (T), where the tone of the target number was modified; c) consonant error condition (C), where the consonant of the target number was modified; d) vowel error condition (V), where the vowel of the target number was modified. The stimuli were split into 8 subsets and counterbalanced to make sure that each participant would be exposed to only one condition for one stimulus.

All the experimental materials were generated by the Amazon text-to-speech software Amazon Polly [16].

Table 1: *Manipulation of numbers in each condition, substitution (error) in bold.*

Target number	Tone error (T)	Consonant error (C)	Vowel error (V)
0 [lɪŋ35]	[lɪŋ**51**]	[**j**ɪŋ35]	[l**a**ŋ35]
3 [san55]	[san**21**]	[**ʂ**an55]	[s**ə**n55]
6 [ljou51]	[ljou**35**]	[**n**jou51]	[lj**a**ʊ51]
9 [tɕjou21]	[tɕjou**55**]	[**ɕ**jou21]	[tɕj**a**ʊ21]

The substitution followed the "one-feature change" criterion: only substitute either place or manner of articulation for onsets; only substitute either register or direction (rising or falling) for contour tones; only substitute either backness, height, or roundness for vowels. As numbers 6 and 9 in Mandarin contain a diphthong, another diphthong was used to complete the substitution for these numbers.

Each number undergoing one type of substitution was then grouped with other unmodified numbers to create a string of 4 digits. In order to avoid the case where the error appears at the very beginning or end of the string (i.e., in the first or last number of the string), an additional number (either 4, 5, 8, or 10) was added at both edges of the string acting as a frame. The frame number was added in such a way that it did not share the same tone or trigger tone sandhi with its adjacent number. The length of the string in all stimuli was of 6 digits (Table 2).

Table 2: *Example stimuli for 6 as the target number, Number 4 as frame.*

Condition	
No Error (NoE)	4 3 0 6 [ljou 51] 9 4
Tone (T)	4 3 0 [ljou **35**] 9 4
Consonant (C)	4 3 0 [**n**jou 51] 9 4
Vowel (V)	4 3 0 [lj**ɑʊ** 51] 9 4

The position of the target number was varied throughout the string so that participants cannot predict its position as the task progresses. There was only one error per string. In half of the stimuli, the error was in an "early" position (the first two positions of the string). In the other half it was in a "late" position (the last two positions in the string).

2.3. Procedure

The experiment was programmed with jsPsych [17] and was conducted online, including a practice and a test session. Instructions were presented in Mandarin before each session. Participants were randomly assigned to one subset of stimuli, resulting in 8 participants for each subset of stimuli. Participants were asked to wear their headsets and to do the experiment in a quiet room. During each trial, the stimulus was displayed both visually and audibly. The number undergoing substitution was not shown on the screen and was replaced by an underscore. Under the visually displayed stimulus, a dial pad was presented. The task for participants was to click on the missing number on the dial pad once they heard it (Figure.1). Each trial lasted 4500 milliseconds, and participants had to decide within this timing; if they did not, the experiment moved on to the next trial. At the end of each trial, participants were instructed to click on the cross button at the center of the screen. This served to recalibrate the position of the cursor.

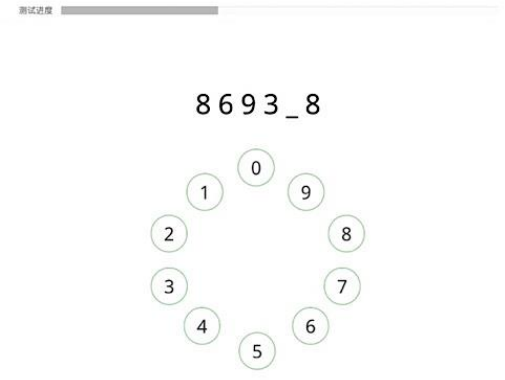

Figure 1: *An example of trial from the number reconstruction task.*

3. Results

A total of 1656 responses were collected (24 trials * 69 participants). After screening, 96 responses were rejected (6%, 4 participants) due to self-reported hearing disorders (3 participants) and below threshold accuracy (< 75%) during practice trials (1 participant). The remaining 1560 responses were used for the statistical analysis. Responses from those who successfully selected the original digit number were coded as 'Correct', and those who selected wrongly or failed to decide within the time limit were coded as 'Incorrect'.

3.1. Response Accuracy

As shown in Figure 2, compared to NoE, participants reconstructed the original number most accurately when the consonant was substituted (94.6%), followed by tone substitution (81%), and then vowel substitution (74.4%).

No-Error trials (390 out of 1560, 25%) were excluded from the statistical analysis because this condition is not necessary for comparing the accuracy among different types of errors. After filtering, 1170 responses were analyzed in the statistical model. A Generalized Linear Mixed-Effect Model [18] was fitted to inspect accuracy (correct, incorrect) as a function of the three types of errors (tone, consonant, vowel). Participant and item were added as random effects. Taking T as reference, the accuracy scores were significantly lower than C (β=2.01, SD=0.39, z=5.07, p<0.005). Meanwhile, there was no significant difference in the accuracy scores of T and V (p>0.05).

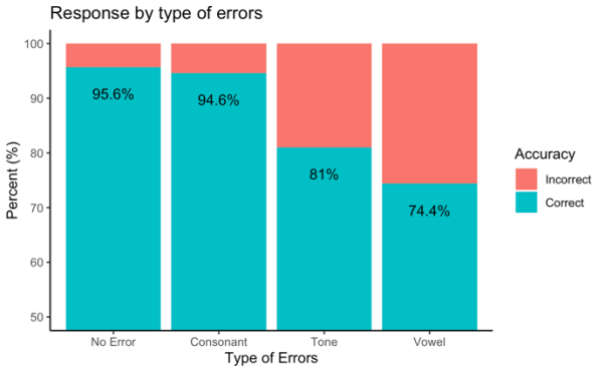

Figure 2: *The accuracy scores (y-axis) by types of errors (x-axis), taking No-Error Condition as baseline.*

3.2. Response Times

Response Times (RTs) were recorded from the beginning of the produced string of numbers. Each audio file of stimuli was automatically annotated and aligned using a script in Praat [19], and the segmentation was corrected manually by the first author. Recall that the target number shifted between two positions in the string, resulting in a time lag between the beginning of the stimulus and the onset of the target number. To retrieve the exact RTs when participants heard the target number, the time lag between the beginning of the stimulus and the onset of the target number was subtracted from the recorded RTs.

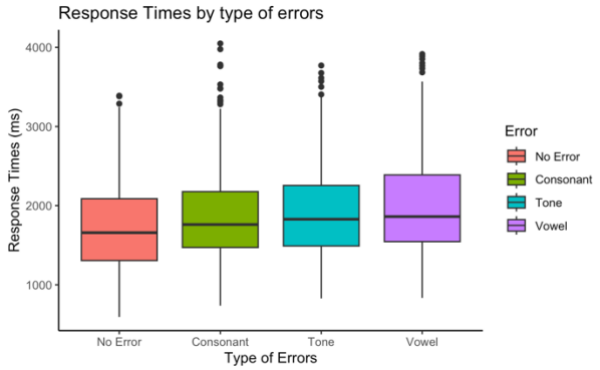

Figure 3: *The calibrated RTs (y axis) of each type of error (x axis).*

Only RTs for correct responses were analyzed (1348 out of 1560 responses, 86.4%). Among the three error conditions (Figure. 3), participants took the longest time to reconstruct the number when vowels were substituted, while they were faster when consonants were substituted (C: mean=1884.6, SD=601.3; V: mean=2022.3, SD=657.4). RTs for the T fell in between (T: mean=1926.4, SD=601.9). RTs were log-shift transformed according to [20]. Outliers were removed using median absolute deviation [21] (39 responses, 2.9%). No-Error Trials (357 out of 1309 responses, 27.2%) were also excluded for the same reason as in the accuracy analysis. After all selections, 952 responses were included in the statistical model. A Linear Mixed-Effect Model fitted log-shift transformed RTs as a function of types of errors, including by-participant and by-item random intercepts. Results from the model revealed that RTs in T are significantly faster than V (β=0.068, SD=0.31, t=2.17, p<0.05), but not significantly slower than C.

As pointed out by an anonymous reviewer, two of the consonant errors in the task coincide with dialectal variants of Mandarin, and participants from Southern/Northeastern China may be familiar with these substitutions. We conducted another series of models to investigate the effect of region (S vs. N) on reconstructing errors. Results on accuracy and RTs did not show a significant effect of region, nor an interaction between region and errors.

4. Discussion & Conclusions

The current study aims to compare the perceptual salience of tonal errors with segmental errors using a Number Reconstruction Task. From the accuracy scores, we can see that the accuracy in tone substitutions is significantly lower than in consonant substitutions, but not significantly different from vowel substitutions, suggesting that tonal errors are as salient as vowel errors and more salient than consonant errors. The results in RTs are consistent with the accuracy scores. The vowel stands out as the most important unit, requiring the most processing cost to reconstruct vowel errors. Tone errors require the second most amount of time and they are significantly faster than vowel errors but not significantly different from consonant errors, highlighting a specific role of tone in error perception compared to segments. In general, our results in accuracy scores and RTs do not fully support the hypothesis that tonal errors are perceptually less salient, and thus may be more easily ignored, than segmental errors.

The results in the vowel substitutions (i.e., participants reconstructed the numbers with vowel substitution with the lowest accuracy and the slowest RTs) are in line with the findings of [15], but contrary to previous findings in other languages [13], [14], [22] in which a C-bias is consistently reported. The current study reveals a V-bias instead of a C-bias for Mandarin speakers in lexical processing. In other words, the information carried by vowels is vital for Mandarin speakers to recognize words successfully. When vowels were substituted, Mandarin speakers were most likely to fail in identifying target words. A similar V-bias is reported in toddlers learning Cantonese, also a tone language [23]. As tones are mainly carried by vowels, it is reasonable to argue that C-bias is reversed by the additional tonal information loaded onto the vowels, resulting in a V-bias in tone languages.

The finding from [15], that tones were reconstructed most accurately and fastest, does not align with the results of the current study. On the one hand, the stimuli used in [15] are nonwords formed by tonal accidental gaps, meaning illegal tone combinations with phonotactically legal segments. Real words may be more easily reconstructed from such nonwords by changing their tones rather than their consonants or vowels. In the task of [15], when participants were asked to turn these tone accidental gap nonwords into real words, the most efficient and natural way to do so was to change their tone. Thus, the result of [15] can alternatively be interpreted as a stimuli-driven bias. On the other hand, the digit numbers used in the current study constrained participants' selection to a limited subset of their lexicon, and the candidates (i.e., numbers) of this subset are all high-frequency words. It is hence possible that to some extent, this specific experimental design reduced the lexical effect during word reconstruction, while emphasizing the weight of the phonological role of segments and tones. In other words, when lexical frequency was no longer a reliable cue for lexical access, participants relied more on contrastive information to process the task, enhancing the role of tones.

Building upon the previous discussion, the current study confirms the V-bias in Mandarin word reconstruction, but we cannot attribute the rarity of tonal errors to this processing bias. Mandarin listeners perceive tonal errors at least as easily as segmental errors when errors are categorical. However, the extent to which the errors created by "one-feature" phonemic substitution reflect the true nature of different types of errors remains unclear. Future studies should investigate the salience of tonal errors with naturalistic error data. Alternatively, errors created in a more fine-grained acoustic scale serve as excellent follow-up stimuli to cross-verify the current results, especially considering that the gradient, dynamic nature of tonal errors has not yet been explored.

In sum, our results have shown that substituting tones affects word identification as much as substituting vowels, and hinders it even more than substituting consonants. Thus, tonal errors are not perceptually less salient than other types of errors, indicating that tones, as contrastive units, share a similar phonological role with vowels and consonants.

5. Acknowledgments

This research has been funded by the China Scholarship Council (CSC202108070106) and by the Labex EFL (ANR-10-LABX-0083-LabEx EFL) to Université Paris Cité.

6. References

[1] V. A. Fromkin, *Speech errors as linguistic evidence*, vol. 77. Walter de Gruyter, 2013.

[2] J. Alderete, M. Baese-Berk, K. Leung, and M. Goldrick, “Cascading activation in phonological planning and articulation: Evidence from spontaneous speech errors,” *Cognition*, vol. 210, p. 104577, 2021.
[3] S. A. Frisch and R. Wright, “The phonetics of phonological speech errors: An acoustic analysis of slips of the tongue,” *J. Phon.*, vol. 30, no. 2, pp. 139–162, Apr. 2002, doi: 10.1006/jpho.2002.0176.
[4] L. Goldstein, M. Pouplier, L. Chen, E. Saltzman, and D. Byrd, “Dynamic action units slip in speech production errors,” *Cognition*, vol. 103, no. 3, pp. 386–412, 2007.
[5] A. Cutler, “The reliability of speech error data,” *Linguistics*, vol. 19, no. 7–8, pp. 561–582, 1981.
[6] J. Alderete and M. Davies, “Investigating Perceptual Biases, Data Reliability, and Data Discovery in a Methodology for Collecting Speech Errors From Audio Recordings,” *Lang. Speech*, vol. 62, no. 2, pp. 281–317, Jun. 2019, doi: 10.1177/0023830918765012.
[7] I.-P. Wan and J. Jaeger, “Speech errors and the representation of tone in Mandarin Chinese,” *Phonology*, pp. 417–461, 1998.
[8] J.-Y. Chen, “The representation and processing of tone in Mandarin Chinese: Evidence from slips of the tongue,” *Appl. Psycholinguist.*, vol. 20, no. 2, pp. 289–301, 1999.
[9] I.-P. Wan, “Mandarin speech errors into phonological patterns/汉语音韵语误之分类型态,” *J. Chin. Linguist.*, pp. 185–224, 2007.
[10] I.-P. Wan, “On the phonological organization of Mandarin tones,” *Lingua*, vol. 117, no. 10, pp. 1715–1738, 2007.
[11] J. H. -c Liu and H. S. Wang, *Speech Errors of Tone in Taiwanese*. 2007.
[12] J. Alderete, Q. Chan, and H. H. Yeung, “Tone slips in Cantonese: Evidence for early phonological encoding,” *Cognition*, vol. 191, p. 103952, 2019.
[13] B. Van Ooijen, “Vowel mutability and lexical selection in English: Evidence from a word reconstruction task,” *Mem. Cognit.*, vol. 24, pp. 573–583, 1996.
[14] T. Nazzi and A. Cutler, “How consonants and vowels shape spoken-language recognition,” *Annu. Rev. Linguist.*, vol. 5, pp. 25–47, 2019.
[15] S. Wiener and R. Turnbull, “Constraints of tones, vowels and consonants on lexical selection in Mandarin Chinese,” *Lang. Speech*, vol. 59, no. 1, pp. 59–82, 2016.
[16] “Text to Speech Software – Amazon Polly – Amazon Web Services.” Amazon, 2022. Accessed: Mar. 28, 2023. [Online]. Available: https://aws.amazon.com/polly/
[17] J. R. de Leeuw, R. A. Gilbert, and B. Luchterhandt, “jsPsych: Enabling an Open-Source Collaborative Ecosystem of Behavioral Experiments,” *J. Open Source Softw.*, vol. 8, no. 85, p. 5351, May 2023, doi: 10.21105/joss.05351.
[18] D. Bates, M. Mächler, B. Bolker, and S. Walker, “Fitting Linear Mixed-Effects Models using lme4.” arXiv, Jun. 23, 2014. Accessed: Dec. 27, 2023. [Online]. Available: http://arxiv.org/abs/1406.5823
[19] P. Boersma and D. Weenink, “Praat: doing Phonetics by Computer.” 2023. Accessed: Jan. 01, 2024. [Online]. Available: https://www.fon.hum.uva.nl/praat/
[20] Z. J. Burchill and T. F. Jaeger, “How reliable are standard reading time analyses? Hierarchical bootstrap reveals substantial power over-optimism and scale-dependent Type I error inflation,” *J. Mem.*
[21] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, “Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median,” *J. Exp. Soc. Psychol.*, vol. 49, no. 4, pp. 764–766, 2013.
[22] A. Cutler, N. Sebastián-Gallés, O. Soler-Vilageliu, and B. Van Ooijen, “Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons,” *Mem. Cognit.*, vol. 28, pp. 746–755, 2000.
[23] H. Chen, D. T. Lee, Z. Luo, R. Y. Lai, H. Cheung, and T. Nazzi, “Variation in phonological bias: Bias for vowels, rather than consonants or tones in lexical processing by Cantonese-learning toddlers,” *Cognition*, vol. 213, p. 104486, 2021.