Do Non-Native Speakers Use Context Speaking Rate in Spoken Word Recognition?

Melissa M. Baese-Berk[1], *Tuuli H. Morrill*[2], *Laura C. Dilley*[3]

[1] University of Oregon, Eugene, OR, United States of America
[2] George Mason University, Fairfax, VA, United States of America
[3] Michigan State University, East Lansing, MI, United States of America

mbaesebe@uoregon.edu, tmorrill@gmu.edu, ldilley@msu.edu

Abstract

Context speaking rate is an important cue in spoken-word recognition in a speaker's native language [1], [2]. Native speakers entrain to the context rate; when they encounter ambiguous regions of speech, native speakers perceive fewer words and/or syllables when the surrounding speaking material is presented a relatively slow rate than when presented with a relatively fast context speaking rate. In the present study, we ask whether non-native speakers are able to use context speaking rate in the same way. We present results from an experiment examining whether non-native speakers show similar patterns to native speakers when determining the number of words being spoken. Results suggest that while non-native speakers are sensitive speaking rate when they hear ambiguous regions of speech, they only show such sensitivity when the speech is relatively slow. When the speech is fast, they do not demonstrate context speaking rate effects. This suggests that some aspects of the context speaking rate effect may be closely tied to proficiency, while other aspects may demonstrate more language-general patterns.

Index Terms: spoken word recognition, speaking rate, non-native perception

1. Introduction

Timing information plays a critical role in perception at many levels. For example, several studies have demonstrated that native listeners can show very robust perception of speech, even when spectral cues are greatly reduced or missing, leaving only timing information intact [3], [4]. This work suggests that timing plays a crucial role in speech perception for native speakers. However, it is unclear whether non-native speakers use timing cues in the same robust fashion that native speakers do. In fact, we have relatively little understanding of how non-native speakers perceive prosodic cues in their non-native language. In the present study, we ask whether non-natives are sensitive to context speaking rate in perception, and if so, whether their sensitivity to these cues differs from native speakers.

Native speakers show sensitivity to context speaking rate in a variety of ways. For example, boundaries between phonological categories shift as a function of the surrounding speaking rate [5]-[7]. A similar effect emerges for perception of geminate and singleton consonants [8]. In addition, speaking rate variation which occurs on words as a function of prosodic structure (e.g., at phrasal boundaries) modulates lexical segmentation and competition for those words relative to lexical competitors [9]. In addition to these local or "proximal' effects of phrase-boundary related speaking rate on lexical segmentation and phoneme perception, distal (i.e., non-local) context speaking rate in the vicinity of words influences perception of those words for native speakers. Dilley & Pitt [2] demonstrated that function words can perceptually appear and disappear as a function of distal context speaking rate. In an utterance like *Anyone must be a minor or child* the function word could be highly coarticulated, so that e.g., *minor or* sounds like [maɪnɚ:]. This coarticulation could lead the utterance to be perceived as either ending with *minor or child* or *minor child*. Dilley & Pitt demonstrated a *lexical rate effect*, such that the function word (e.g., *or*) could disappear or appear as a function of the distal context speaking rate. That is, the function word "disappeared" in perception in a significant number of cases when the distal speech rate was made relatively slow compared to the proximal speech rate of the function-word containing utterance portion, even though the acoustic properties of that portion were identical across comparable conditions. These various rate normalization effects may stem in part from mechanisms of general auditory and timing perception, including entrainment [10], [11].

This lexical rate effect has been demonstrated across multiple studies to be a powerful factor influencing perception across varying conditions, including differences in linguistic rhythm [12], and distinct proximal acoustic environments varying in intensity, fundamental frequency, and duration [13]. Further, this effect emerges for both utterance-level context speaking rate and global speaking rate [14]. However, these studies all examined perception of English by native speakers. Thus far, only one study has examined this disappearing word effect in a language other than English or with non-native speakers. Russian speakers show a similar context speaking rate effect when presented with ambiguous stretches of speech [1]. This study also provides some evidence that highly proficient non-native speakers of Russian demonstrate similar context speaking rate effects as native speakers of Russian. However, in part because the system of vowel reduction in English is quite different than that of Russian, it is unclear whether non-native speakers of English will show the same speaking rate effects.

Our understanding of how non-native speakers utilize timing information is quite limited. While many studies have examined how native language background influences non-native speech perception [15]-[17], these studies have primarily focused on segmental features. Some studies have examined supra-segmental features; however, these have primarily focused on word-level features, such as stress [18]. In the current study, we examine perception of suprasegmental features across an entire utterance. We ask whether non-native speakers utilize context speaking rate to the same extent that native speakers do.

2. Methods

2.1. Participants

The first group of participants consisted of college-aged, native Mandarin speakers (n = 16) who spoke English as their second language. Participants were currently attending a US university for intensive English education or for completion of their undergraduate degrees. Participants had been in the US between 1 month and 2 years. Proficiency was roughly matched with all students being in the same levels of oral and written communication ESL classes, which is the highest level offered by the university. The second participant group (n = 16) consisted of college-aged, native English speakers, currently enrolled in a US university. No participants reported any speech or hearing difficulties.

2.2. Materials

Materials for this study were generated from utterances used to create the materials for the study by Dilley & Pitt [2]. Sentences were created with an ambiguous region of speech that was expected to be highly reduced and co-articulated. The sentence context was grammatically consistent with both presence or absence of the function word; for example, in *Anyone must be a minor or child,* the latter portion could be interpreted as either *minor or child* or *minor child.* The original recordings of these stimuli from Dilley & Pitt [2] were used for the present study.

Using the Pitch-Synchronous Overlap and Add (PSOLA) algorithm in Praat [19], we manipulated the original recordings of stimuli from Dilley and Pitt [2] for use in one of four experimental conditions. The stimuli were divided into two regions. The target region was defined as the critical function word (e.g., *or*), plus the preceding syllable and the following phoneme. The context region was defined as the remainder of the utterance prior to and following the target region. Four conditions were created in which these two regions were manipulated differently (see Figure 1 for a schematic of the rate manipulations). For the Normal Rate condition, the entire utterance was presented at the original (i.e., normal) spoken rate. In the Slowed Context condition, the context was expanded by a factor of 1.9, and the target region was presented at the original rate. In the Target Compressed condition, the target region was compressed by a factor of 0.6, while the context was presented at the original rate. In the Target+Context Compressed condition, the target and context were both compressed by a factor of 0.6.

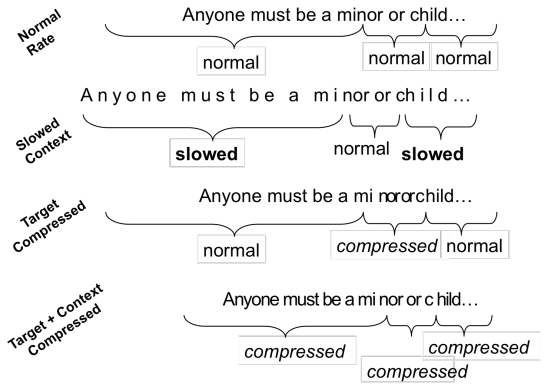

Figure 1: *Schematic of rate manipulations.*

2.3. Procedure

The stimulus presentation was very similar to the procedure described by Dilley & Pitt [2]. We piloted the transcription task used for that experiment with a separate group of non-native listeners (n = 10). In this task, participants heard an utterance and were asked to transcribe it. Transcription was completely free as participants were not given any information outside of the auditory stimulus during the task. However, after piloting the task, we discovered that many participants made a very large number of errors on non-target sentences. Several additional participants were unable to complete the task, likely due to the relatively low frequency of many of the target words in the utterances. It appeared that these participants did not know several of the words (e.g., "leisure" and "harbor"). The participants who failed to complete the experiment were not included in the brief description of the free transcription data reported briefly below.

For our primary data, a different experiment was designed using the same stimuli but with a two-alternative forced choice task (2AFC). For this task, participants viewed the first portion of the sentence. They then heard the entire utterance and were asked to select between one of two completions of the utterance, where one contained the critical function word and the other did not, similar to the task in Heffner et al. [13]. For example, participants were given *Anyone must be a* _____, and the participants had to determine whether they heard *minor or child* or *minor child*.

Participants in the 2AFC task heard fifty experimental trials and seventy filler trials. Participants were randomly assigned to one of four lists differing in item-condition pairing. The target stimuli were divided among the conditions in roughly equal proportions across the four lists.

2.4. Analysis

We used logistic mixed effects models implemented in R [20] to analyze the data. For each observation, we coded whether or not the listener reported hearing a function word. This binary code was used as the dependent variable. Fixed effects included in the model were language background, experimental condition, and the interactions between the two factors. Both factors were contrast coded. Random effects included were the maximal effects allowing for the model to converge and included random slopes for participants and items. Significance of each predictor was determined via model comparison.

3. Results

Figure 2 displays the proportion of cases where function words were reported for each participant in the 2AFC task. From this figure, it is clear that native speakers report fewer function words when the context is relatively slow compared to the target, as they did on the free transcription task original presented in Dilley & Pitt [2]. This is similar to the findings in Heffner et al. [13], as well.

It is also clear that some differences emerge between the language groups. Non-native speakers demonstrate a similar effect of context speaking rate when comparing the unaltered and Slowed Context conditions. However, the difference between the Target Compressed and Target+Context Compressed conditions was much greater for the native speakers than the non-native speakers. Further, it appears that

function word reports from the non-native speakers are closer to chance across many conditions than the native speakers, suggesting an attenuation of the effect overall.

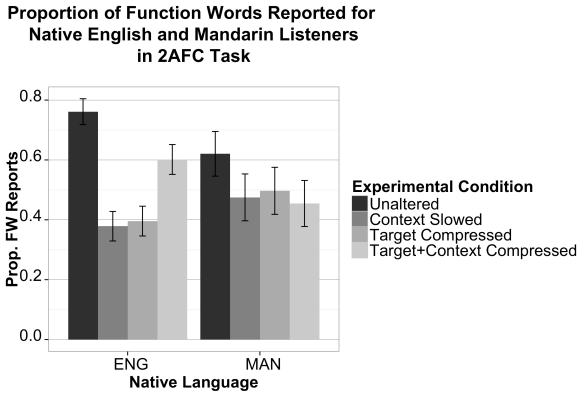

Figure 2: *Proportion of function words reported for both language groups in all experimental conditions for the two-alternative forced choice task.*

The results of the mixed effects model support these observations. The results of the model are summarized in Table 1. Overall, the main effect of native language was a significant predictor of model fit ($\chi^2 = 41.745$, $p < .001$). Further, the comparison between the Unaltered and Context-Slowed condition was a significant predictor of model fit. ($\chi^2 = 13.921$, $p < .001$). The comparison between the cases where the target is relatively fast (i.e., Context Slowed and Target Compressed) and the cases where the target is presented at the same rate as the context (i.e., Unaltered and Target+Context Compressed) is also a significant predictor ($\chi^2 = 60.56$, $p < .001$). However, the comparison between the Unaltered condition and the Target+Context Compressed condition does not emerge as a significant predictor of model fit ($\chi^2 = 0.1179$, $p > 0.7$). These results suggest that listeners were more likely to report function words when the context speaking rate was relatively slow compared to the target. Further, overall, there was no significant difference between cases where the entire utterance was sped and the entire utterance was unaltered.

The results of the interactions in the model show a similar pattern. The interaction between native language and the cases where the target is relatively fast (i.e., context slowed and target compressed) vs. the cases where the target is spoken at the same rate as the context is significant ($\chi^2 = 40.481$, $p < .001$). The interaction between native language and the Unaltered vs. Target+Context Compressed cases is also significant ($\chi^2 = 13.938$, $p < .001$). However, the interaction between native language and the Unaltered vs. Context Slowed comparison is not significant ($\chi^2 = .04$, $p = .98$). These interactions suggest that while there is no overall difference between cases where the context and target regions are presented at the same rate, this is modulated by native language. Non-native speakers are less likely to report function words in the case where the target and context are sped up than when the speaking rate is unaltered. In cases where the context speaking rate is slowed, both native English and native Mandarin speakers are less likely to report having heard function words.

	Estimate	Std. Err.	*z* value
Intercept	0.207	0.119	1.74
Native Language	-.107	0.116	-0.96
Unaltered vs. Context Slowed	0.445	0.397	1.122
Relatively Slow vs. Same Rate	-2.887	0.457	-6.309
Unaltered vs. Target+Context Expanded	-0.106	0.328	-0.323
Native Language* Unaltered vs. Context Slowed	-0.007	0.319	-0.021
Native Language* Relatively Slow vs. Same Rate	2.884	0.450	6.413
Native Language* Unaltered vs. Target+Context Expanded	-1.431	0.381	-3.752

Table 1. *Logistic mixed effects regression model predicting proportion of function words reported in each of the training conditions.*

We examined whether a similar pattern of context speaking rate effects emerged for the non-native participants who completed the free transcription task. The results of this task are presented in Figure 3. This figure suggests that non-native listeners may be able to take advantage of context speaking rate even in the conditions where some portion of the utterance is sped, as listeners report more function words in cases where both the target and context are compressed than in cases where only the target is compressed. However, it is important to note that participants who were able to complete this task had a relatively higher level of proficiency than those who completed the 2AFC task.

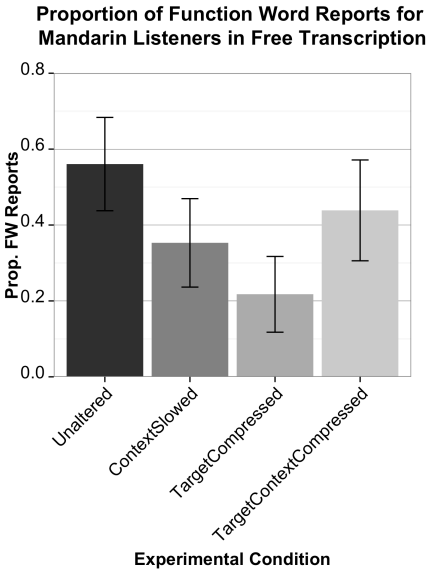

Figure 3: *Proportion of function words reported for native Mandarin listeners in the free transcription task.*

The results of the mixed effects model support these observations. The results of the model are summarized in Table 2. Model comparisons revealed that the comparison between the Unaltered and Context Slowed cases is a significant predictor of model fit ($\chi^2 = 21.041$, $p < .001$). The comparison between cases where the context is relatively slow compared to the context vs. when they are both presented at the same rate was also a significant predictor of model fit ($\chi^2 = 5.371$, $p < .03$). However, the difference between the Unaltered and Target+Context Compressed conditions was not significant ($\chi^2 = 2.267$, $p > 0.13$). Participants reported fewer function words when the target was relatively fast compared to the surrounding context, especially when the context was slowed. However, unlike in the 2AFC task, participants showed no difference between the Unaltered and Target+Context compressed cases, suggesting that higher proficiency non-native listeners may utilize speaking rate in similar ways to native speakers.

	Estimate	Std. Err.	*z* value
Intercept	-0.78	0.4396	-1.17
Unaltered vs. Context Slowed	1.2348	0.5497	2.247
Relatively Slow vs. Same Rate	-3.463	0.8293	-4.176
Unaltered vs. Target+Context Expanded	0.9908	0.6619	1.497

Table 2: *Logistic mixed effects regression model predicting proportion of function words reported in each of the experimental conditions.*

4. Discussion

This study extended understanding of the lexical rate effect first reported in [2]; here, we investigated this phenomenon in both native and non-native English speaker groups. Both groups showed a lexical rate effect: they reported fewer function words in a coarticulated stretch of speech containing a function word when the context was relatively slow than when the context and target regions were presented at the same rate. More importantly, both native and non-native speakers in this task showed similar sensitivity to context speaking rates in several ways. For example, when comparing the Unaltered and Context-Slowed conditions, both native English speakers and native Mandarin speakers demonstrated a similar sensitivity to context speaking rate.

These results are broadly consistent with the findings of Dilley, Morrill and Banzina [1]. They found that, when compared to native Russian speakers, non-native Russian speakers were less able to take advantage of context speaking rate information. However, context speaking rate effects increased when language proficiency increased. While the present study did not expressly examine language proficiency, the findings are consistent with the suggestion that non-native speakers are able to utilize context speaking rate in spoken word recognition under a variety of circumstances.

However, substantial differences were also observed between native and non-native speakers in the present study. Namely, non-native speakers did not show any sensitivity in the 2AFC task to context speaking rate in the conditions where the context or target were compressed (i.e., Target Compressed and Target+Context Compressed). Further, participants reported fewer function words when any portion of the utterance is sped. It is possible that this reduction in context speaking rate sensitivity may be tied to language proficiency. When the utterance is sped, non-native listeners may be unable to fully integrate timing cues to coarticulated function words with the other phonological, lexical, and semantic cues they are exposed to.

We also found that non-native speakers reported fewer function words than native speakers in general in both the 2AFC and free transcription tasks. This may be a result of function words being frequently highly reduced and coarticulated in English [21]. This could result in reduced perceptibility, especially for non-native speakers who may not yet have a full grasp of reduction processes in English (see [22] for lack of reduction in function word production for non-native speakers). Future work could examine whether the overall number of function word reports increases as a function of proficiency. It is also possible that non-native speakers are simply guessing whenever the speaking rate is manipulated. If participants are truly sensitive to speaking rate, they should report more function words if exposed to stimuli where the target is relatively long compared to the context (i.e., a Target Expanded condition).

These results support the hypothesis that entrainment to context speaking rate is possible for both native and non-native speakers, but that this entrainment may be subject to some limitations. While normalization for context speaking rate may be related to general cognitive mechanisms, including entrainment, distinct degrees of entrainment across linguistic groups would suggest a domain-specific (language-related) component to rate normalization, as well. This possibility may be supported by recent findings [23] demonstrating that the lexical rate effect can only be induced when a context is perceived as intelligible speech. Taken together, these results suggest that the mechanisms underlying the lexical rate effect may have both domain-specific, and domain-general components. Perhaps listeners are entraining to an abstract representation of language, rather than the low-level amplitude envelopes of the speech signal.

5. Conclusions

In conclusion, the present study suggests non-native speakers of English do show some sensitivity to context speaking rate. However, this sensitivity may be more robust in cases where the overall speaking rate is not sped relative to typical rates. Across both tasks, e utterance is sped, non-native listeners report fewer instances of having heard a function word in general, and show reduced sensitivity to context speaking rate effects. Further, non-native speakers report having heard fewer function words than native speakers, across all conditions. These results suggest that some aspects of sensitivity to context speaking rate may be susceptible to language experience or proficiency.

6. Acknowledgements

We thank Lily Huston, Misaki Kato, Quinten Konyn, and Paul Olejarczuk for assistance in data collection. This work was partially funded by the Program in Linguistics at George Mason University, a Faculty Research Award from the University of Oregon to MMB and NSF Grant BCS #1431063 to LCD.

7. References

[1] L. C. Dilley, T. H. Morrill, and E. Banzina, "New tests of the distal speech rate effect: examining cross-linguistic generalization," *Frontiers in Psychology*, vol. 4, 2013.

[2] L. C. Dilley and M. A. Pitt, "Altering context speech rate can cause words to appear or disappear," *Psychological Science*, vol. 21, no. 11, pp. 1664–1670, 2010.

[3] R. Remez, P. Rubin, D. Pisoni, and T. D. Carrell, "Speech perception without traditional speech cues," *Science*, vol. 212, no. 4497, pp. 947–949,1981.

[4] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science, New Series*, vol. 270, no. 5234, pp. 303–304, 1995.

[5] A. M. Libermann, Delattre, L. J. Gerstman, and F. S. Cooper, "Tempo of frequency change as a cue for distinguishing classes of speech sounds.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 52, no. 2, p. 127, 1956.

[6] J. L. Miller and A. M. Libermann, "Some effects of later-occurring information on the perception of stop consonant and semivowel," *Attention, Perception & Psychophysics*, vol. 25, no. 6, pp. 457–465, 1979.

[7] J. R. Sawusch and R. S. Newman, "Perceptual normalization for speaking rate II: Effects of signal discontinuities," *Attention, Perception & Psychophysics*, vol. 62, no. 2, pp. 285–300, 2000.

[8] J. M. Pickett and L. R. Decker, "Time factors in perception of a double consonant," *Language and Speech*, vol. 3, pp. 11–17, 1960.

[9] A. P. Salverda, D. Dahan, M. K. Tanenhaus, K. Crosswhite, M. Masharov, and J. McDonough, "Effects of prosodically modulated sub-phonetic variation on lexical competition," vol. 105, no. 2, pp. 466–476, Nov. 2007.

[10] E. Large and M. R. Jones, "The dynamics of attending: How people track time-varying events," *Psychological Review*, vol. 106, pp. 119–159, 1999.

[11] M. R. Jones and J. D. McAuley, "Time judgments in global temporal contexts," *Attention, Perception & Psychophysics*, vol. 67, no. 3, pp. 398–417, 2005.

[12] T. H. Morrill, L. C. Dilley, J. D. McAuley, and M. A. Pitt, "Distal rhythm influences whether or not listeners hear a word in continuous speech: Support for a perceptual grouping hypothesis," *Cognition*, vol. 131, no. 1, pp. 69–74, 2014.

[13] C. C. Heffner, L. C. Dilley, J. D. McAuley, and M. A. Pitt, "When cues combine: How distal and proximal acoustic cues are integrated in word segmentation," *Language & Cognitive Processes*, vol. 28, no. 9, pp. 1275-1302, 2012.

[14] M. M. Baese-Berk, C. C. Heffner, L. C. Dilley, M. A. Pitt, T. H. Morrill, and J. D. McAuley, "Long-term temporal tracking of speech rate affects spoken-word recognition," *Psychological Science*, vol. 25, no. 8, pp. 1546–1553, 2014.

[15] J. S. Logan, "Training Japanese listeners to identify English /r/ and /l/: A first report," *Journal of the Acoustical Society of America*, vol. 89, no. 2, pp. 874–886, 1991.

[16] C. T. Best, G. W. McRoberts, and E. Goodell, "Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system," *Journal of the Acoustical Society of America*, vol. 109, no. 2, pp. 775–794, 2001.

[17] J. E. Flege, I. R. A. Mackay, and D. Meador, "Native Italian speakers' production and perception of English vowels," *Journal of the Acoustical Society of America*, vol. 106, no. 5, pp. 2973–2987, 1999.

[18] N. Cooper, A. Cutler, and R. Wales, "Constraints of Lexical Stress on Lexical Access in English: Evidence from Native and Non-native Listeners," *Language and Speech*, vol. 45, no. 3, pp. 207–228, 2002.

[19] P. Boersma and D. Weenik, "Praat: doing phonetics by computer." 2015.

[20] R Development Core Team, "R: A language and environment for statistical computing." Vienna, Austria, 2014.

[21] A. Bell, D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory, and D. Gildea, "Effects of disfluencies, predictability, and utterance position on word form variation in English conversation," *Journal of the Acoustical Society of America*, vol. 113, no. 2, pp. 1001–24, 2003.

[22] R. E. Baker, M. Baese-Berk, L. Bonnasse-Gahot, M. Kim, K. J. Van Engen, and A. R. Bradlow, "Word durations in non-native English," *Journal of Phonetics*, vol. 39, no. 1, pp. 1–17, 2011.

[23] M. A. Pitt, C. Szostak, and L. C. Dilley, "Rate dependent speech processing can be speech specific: Evidence from the perceptual disappearance of words under changes in context speech rate," *Attention, Perception and Psychophysics*, pp. 1-12 2015.