# NORMALIZATION FOR SPEECHRATE IN NATIVE AND NONNATIVE SPEECH

Hans Rutger Bosker[1] and Eva Reinisch[2]

[1]Max Planck Institute for Psycholinguistics, Nijmegen;

[2]Institute of Phonetics and Speech Processing, Ludwig Maximilian University Munich
hansrutger.bosker@mpi.nl and evarei@phonetik.uni-muenchen.de

## ABSTRACT

Speech perception involves a number of processes that deal with variation in the speech signal. One such process is normalization for speechrate: local temporal cues are perceived relative to the rate in the surrounding context. It is as yet unclear whether and how this perceptual effect interacts with higher level impressions of rate, such as a speaker's nonnative identity. Nonnative speakers typically speak more slowly than natives, an experience that listeners take into account when explicitly judging the rate of nonnative speech. The present study investigated whether this is also reflected in implicit rate normalization. Results indicate that nonnative speech is implicitly perceived as faster than temporally-matched native speech, suggesting that the additional cognitive load of listening to an accent speeds up rate perception. Therefore, rate perception in speech is not dependent on syllable durations alone but also on the ease of processing of the temporal signal.

**Keywords:** speech perception, speechrate, implicit processing, nonnative speech, cognitive load.

## 1. INTRODUCTION

There is great diversity in the speed at which people speak: we find rate variation between languages [19], between [28] and within [20] individual speakers of a language, and even within a single utterance [14]. One means by which listeners deal with this kind of variability is rate normalization, involving the interpretation of local temporal cues relative to the rate cues in the surrounding context. For instance, the stop voicing contrast in English (e.g., /g/ vs. /k/) is mainly cued by temporal properties, namely duration of voice onset time (VOT). The perception of a stop such as /g/ or /k/ may hence be shifted by presenting it in a fast or a slow context. The effect is contrastive: relative to a fast context, the VOT sounds long, hence the sound is likely to be perceived as voiceless - especially if at a normal rate the voicing would be ambiguous [17]. Similar effects have been found for the perception of vowel duration [26], stress [24], word segmentation [25], and even the perception of larger morphophonological units such as function words [11].

Previous literature suggests that rate normalization is an early perceptual or even general auditory process since it can be elicited by non-speech contexts [10], it has been found in non-human perception [30], and rate normalization generalizes across speakers [17]. Recent evidence from eye-tracking experiments further indicates that contextual rate information influences the perception of temporal target cues as soon as these are available [26]. However, it is as yet unclear whether rate normalization may also be affected by or interacts with "higher level" impressions of rate. For example, [23] observed that casual speech including segmental deletions is perceived as faster than utterances with all segments realized, even when the overall sentence durations were matched (and hence the sentence containing deletions had a lower number of realized segments/syllables per unit time than the fully pronounced sentence). One possible explanation was the commonly experienced association between segmental deletions and fast speech influencing the way listeners normalize for perceived speechrate. The present study assessed a related phenomenon for higher-level influences on rate perception: the nonnative identity of the speaker.

Nonnative speakers typically speak at a lower rate than native speakers [8] due to incomplete mastery of the L2 or a lack of automaticity in L2 speech production [9, 27]. Hence, nonnative speech is (perceived as) less fluent [7], less appropriate [15], more accented [16], and less comprehensible [16] than native speech. Given the resulting trouble with understanding nonnative speech, listeners prefer to listen to it at a slower rate than native speech [16]. Put in different words, nonnative speech that matches native speech in overall duration is perceived as faster than native speech [2]. All the aforementioned studies on the perception of nonnative rate targeted *explicit* perception. That is, they all investigated the subjective impression listeners had of certain speech samples (e.g., speed ratings). It is unclear, however, whether the nonnative identity of the speaker may also influence online speech processing when listeners normalize for rate in a given context. The present study tested this hypothesis.

For this purpose, we recorded native and nonnative speakers of Dutch producing a set of sentences that we matched in overall duration. All sentences were manipulated to end in a Dutch minimal pair duration continuum ranging from short *tak* /tɑk/ "branch" to long *taak* /ta:k/ "task". Additional spectral cues to the vowels were set to an ambiguous value (see Methods for details; cf. [26]). If rate normalization occurs only with regard to the number of syllables realized in a certain period of time, no difference would be expected between temporally-matched native and nonnative speech. If listeners bring to bear their prior experiences with the typically slow speech of nonnative speakers, then nonnative speech may be perceived as slower than temporally-matched native speech. Given the early nature of rate normalization, it may be unlikely that these late subjective impressions will interfere with rate perception. However, listening to accented speech requires more cognitive effort than listening to native speech [13, 29]. An increase in cognitive load is known to speed up time perception [5], making nonnative speech potentially sound faster (cf. explicit rate perception: [2]).

## 2. METHOD

### 2.1. Participants

Native Dutch participants ($N = 45$) with normal hearing were recruited from the MPI participant pool. None reported native knowledge of German.

**Table 1:** Number of realized syllables and total carrier duration (ms) of each carrier split by carrier rate. All values were identical across native and nonnative speech.

| | Normal | | Fast | |
|---|---|---|---|---|
| | *n* syll | total dur | *n* syll | total dur |
| 1. | 16 | 3275 | 16 | 2418 |
| 2. | 18 | 3285 | 17 | 2528 |
| 3. | 17 | 3141 | 16 | 2418 |
| 4. | 17 | 2946 | 17 | 2335 |

### 2.2. Design

Two native and two Austrian nonnative female speakers of Dutch were recorded producing four different sentences in Dutch:

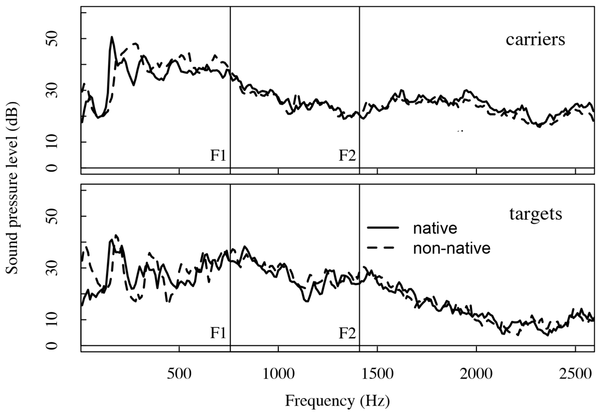
1. *Vervolgens keek Gijs eens goed om zich heen en zei hij het woordje...*
"Then Gijs looked around and said the word..."
2. *Uiteindelijk wist Huib de oplossing niet en gokte hij op het woordje...*
"In the end, Huib did not know the solution and guessed the word..."
3. *Gisteren twijfelde Fleur eerst nog even en koos ze het woordje...*
"Yesterday Fleur hesitated for a while and chose the word..."
4. *Tenslotte liep Frederieke de trein uit en zei ze het woordje...*
"Finally Frederike left the train and said the word...".

All sentences were produced multiple times ending in either *tak* /tɑk/ "branch" or *taak* /ta:k/ "task", spoken at the speaker's habitual rate and at a fast rate. From these sentences, carriers were excised that included all speech up to target onset (the nearest positive-going zero-crossing before the /t/ burst). Carriers did not favor any of the target words semantically and did not contain any /ɑ/ or /a:/-vowels. From these carriers, one fast and one slow token per sentence per speaker was selected such that they contained the same number of syllables and lacked silent pauses (> 150 ms). Using PSOLA in PRAAT [6], carriers were time aligned across speakers as to match the average native duration of that particular carrier at that particular rate, achieving temporal matching across native and nonnative speakers (see Table 1).

To further exclude any interference from spectral contrast effects of the carriers [12], the Long-Term Average Spectra (LTAS) of the native and nonnative time aligned carriers were inspected. Figure 1 shows that little to no difference in the regions of the targets' F1 and F2 could be found.

One *taak* target was excised for each of the four speakers. Because the Dutch /ɑ/-/a:/ vowel contrast is also cued by spectral characteristics, F1 and F2 values of the /a:/ vowels were manipulated to be ambiguous between /ɑ/ and /a/ [1]: $F1 = 758\ Hz$; $F2 = 1410\ Hz$ (see bottom panel of Figure 1). These formant values fell within the range of all speakers' F1 and F2 values. The spectral manipulations were based on Burg's LPC method (implemented in PRAAT), with the source and filter models estimated automatically from the selected vowels. The

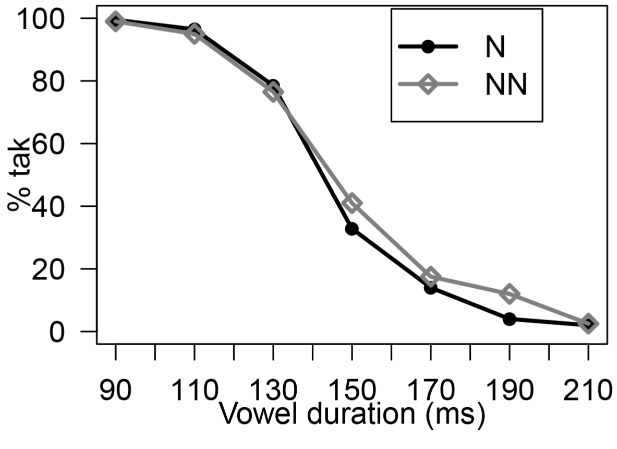
**Figure 1:** Long-Term Average Spectra (LTAS) of the carriers and the targets, split by Nativeness.



formant values in the filter models were inspected and adjusted to result in the desired formant values. Finally, the source and filter models were recombined and the new targets were adjusted to have the same overall amplitude as the original targets. Based on these spectrally ambiguous targets, duration continua were created for each speaker using PSOLA, ranging from 90 ms to 210 ms in steps of 20 ms. Closure and burst durations of /t/ and /k/ were matched as well.

A pretest was run to check the categorization functions of the target duration continua. Participants ($N = 23$) listened to all steps of the continua from the four speakers *in isolation* and indicated whether they heard *tak* or *taak*. Figure 2 shows a slightly higher *%tak* for nonnative speech. However, Generalized Linear Mixed Models indicated no statistical difference between native and nonnative targets, hence our matching procedures succeeded.

**Figure 2:** Categorization of native and nonnative target words in isolation in the pretest.



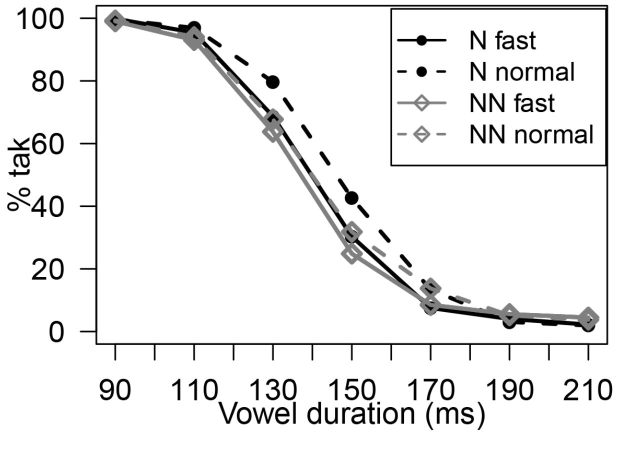
### 2.3. Procedure

For the main experiment, the target continua were spliced back onto the four carriers of the respective speaker. Thus, a stimulus set of 224 tokens (4 speakers x 4 carriers x 2 rates x 7 continuum steps) was created. All carrier-target combinations were presented to participants twice. Participants' task was to indicate whether the sentence-final target word was *tak* or *taak*. Sentences were presented in random order but blocked by Nativeness. Whether participants first heard the native block or the nonnative block was counter-balanced across participants. One complete experimental session lasted approximately 1 hour, with three breaks at each quartile of the trials. After the experiment, participants listened to the 32 carriers (i.e., without the sentence-final targets) and rated them for perceived accentedness and perceived speed (7-point scales: higher rating indicated higher accentedness/rate).

## 3. RESULTS

Trials without a response were excluded from the data ($N = 20$; =0.1%). Average categorization data are represented in Figure 3. A Generalized Linear Mixed Model (GLMM; [21]) as implemented in the lme4 library [4] in R [22] tested the binomial responses for fixed effects of Continuum Step (continuous predictor rescaled around the median), Rate (categorical predictor with fast rate as its intercept), and Nativeness (categorical predictor with native speech as its intercept), and their interactions, with crossed random effects of Participants and Carriers. Only by-participant random slopes for Continuum Step, Nativeness, and their interaction were included because models with more complex random

**Figure 3:** Categorization of native (N) and nonnative (NN) target words in fast and normal carriers.



effects structures failed to converge. This GLMM revealed, firstly, a significant effect of Continuum Step, confirming the temporal contrast between /ɑ/ and /a:/ (i.e., the longer the duration of the target vowel, the fewer *tak* responses; $\beta = -4.329$, $z = -19.462$, $p < 0.001$). Secondly, a main effect of Rate confirmed that listeners normalized for speechrate in the carriers (i.e., speech at normal rate led to more *tak* responses; $\beta = 0.578$, $z = 7.579$, $p < 0.001$). Thirdly, a main effect of Nativeness indicated that nonnative speech led to fewer *tak* responses ($\beta = -0.314$, $z = -2.007$, $p = 0.045$). Finally, the interaction between Rate and Nativeness indicated that the effect of Rate was smaller in nonnative speech ($\beta = -0.288$, $z = -2.757$, $p = 0.006$). A comparable model with the nonnative data as its intercept still found an effect of Rate, indicating a mere decrease of the effect of Rate in the nonnative data (i.e., not absence).

We also tested for presentation order effects within the two experimental blocks by adding the predictor Trial Number and its interaction with Nativeness to the GLMM. Fixed effects of Continuum Step, Rate, Nativeness, and Rate x Nativeness remained, but one additional effect of Trial Number ($\beta = -0.123$, $z = -3.156$, $p = 0.002$) was observed. This shows that, as the experiment progressed, participants reported fewer *tak* responses. Finally, we also investigated effects of Block Order but found no effects.

Linear Mixed Models [3] of the explicit speed ratings revealed no statistically significant difference between the perceived rate of the native and nonnative carriers (Normal rate: $M_{native} = 3.150; M_{nonnative} = 3.158$; Fast rate: $M_{native} = 5.575; M_{nonnative} = 5.436$). The accent ratings revealed a significant difference between native and nonnative speech (averaged across rates: $M_{native} = 1.203; M_{nonnative} = 5.189$).

## 4. DISCUSSION

In our study we set out to test whether the implicit processing of speechrate as evidenced in a classical rate normalization paradigm may be influenced by higher-level information about the speaker's nonnative identity. The main experiment indicated fewer *tak* responses for nonnative speech than for temporally-matched native speech. Note that in the pretest, where targets were presented in isolation, more *tak* responses were obtained for nonnative speech. This suggests that, given the contrastive nature of rate normalization, nonnative speech is implicitly perceived as faster than temporally-matched native speech. Although the explicit rate judgments of the carriers did not reveal a difference between native and nonnative speech (possibly due to small sample size; 32 ratings per participant), earlier studies of explicit rate perception do show nonnative speech to be perceived as faster than temporally-matched native speech (e.g., [2]).

The factor responsible for the difference in perceived rate between native and nonnative speech may be cognitive load: listening to accented speech is cognitively effortful [29], as evidenced by poorer comprehension and worse intelligibility [16]. The psychophysical literature, in turn, indicates that an increase in cognitive load may speed up people's time perception [5], possibly explaining why foreign-accented speech sounds fast.

Cognitive load may also explain why we found an effect of Trial Number, indicating a small but gradual decrease in *tak* responses as the experiment progressed. Listening repeatedly to the same few sentences may have caused fatigue, straining participants' attention and memory spans, increasing cognitive load, and as a consequence leading to a faster perception of the carrier sentences.

Finally, not only did we find nonnative speech to be perceived as faster than native speech, we also found rate normalization effects to be reduced in nonnative speech. Apparently, listeners have more difficulty tracking the rate of less intelligible speech. This observation is in line with studies on neural entrainment of brain oscillations to the syllabic rate of speech. These studies (e.g., [18]) argue that listeners show reduced phase-locking to speech that is less intelligible (e.g., noise-vocoded degraded speech). The same principle may apply to the alignment of neural oscillations to nonnative speech rate, reducing rate normalization in nonnative speech.

Concluding, we find that nonnative speech may be perceived as faster than temporally-matched native speech. We argue that the additional cognitive load of listening to accented speech speeds up time perception, affecting temporal speech contrasts. Our results could not be explained by spectral or temporal differences in the targets or carriers, are in line with explicit judgments of native and nonnative speech rate, and are supported by the observed presentation order effect. Thus, we conclude that rate perception is not dependent on objective syllable durations alone, but is also affected by the relative ease of processing of the speech signal.

## 5. REFERENCES

[1] Adank, P., Van Hout, R., Smits, R. 2004. An acoustic description of the vowels of northern and south-

ern standard dutch. *The Journal of the Acoustical Society of America* 116(3), 1729–1738.

[2] Anderson-Hsieh, J., Koehler, K. 1988. The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning* 38(4), 561–613.

[3] Baayen, R. H., Davidson, D. J., Bates, D. M. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4), 390–412.

[4] Bates, D., Maechler, M., Bolker, B. 2012. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-39.

[5] Block, R. A., Hancock, P. A., Zakay, D. 2010. How cognitive load affects duration judgments: A meta-analytic review. *Acta Psychologica* 134(3), 330–343.

[6] Boersma, P., Weenink, D. 2012. Praat: doing phonetics by computer [computer program]. Version 5.3.18.

[7] Bosker, H. R., Quené, H., Sanders, T. J. M., De Jong, N. H. 2014. The perception of fluency in native and non-native speech. *Language Learning* 64, 579–614.

[8] Cucchiarini, C., Strik, H., Boves, L. 2000. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *The Journal of the Acoustical Society of America* 107(2), 989–999.

[9] De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., Hulstijn, J. H. 2013. Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics* 34(5), 893–916.

[10] Diehl, R. L., Walsh, M. A. 1989. An auditory basis for the stimulus-length effect in the perception of stops and glides. *The Journal of the Acoustical Society of America* 85(5), 2154–2164.

[11] Dilley, L. C., Pitt, M. A. 2010. Altering context speech rate can cause words to appear or disappear. *Psychological Science* 21(11), 1664–1670.

[12] Ladefoged, P., Broadbent, D. E. 1957. Information conveyed by vowels. *The Journal of the Acoustical Society of America* 29(1), 98–104.

[13] Mattys, S. L., Davis, M. H., Bradlow, A. R., Scott, S. K. 2012. Speech recognition in adverse conditions: A review. *Language and Cognitive Processes* 27(7-8), 953–978.

[14] Miller, J. L., Grosjean, F., Lomanto, C. 1984. Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica* 41(4), 215–225.

[15] Munro, M. J., Derwing, T. M. 1998. The effects of speaking rate on listener evaluations of native and foreign-accented speech. *Language Learning* 48(2), 159–182.

[16] Munro, M. J., Derwing, T. M. 2001. Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition* 23(4), 451–468.

[17] Newman, R. S., Sawusch, J. R. 2009. Perceptual normalization for speaking rate iii: Effects of the rate of one voice on perception of another. *Journal of phonetics* 37(1), 46–65.

[18] Peelle, J. E., Gross, J., Davis, M. H. 2013. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex* 23(6), 1378–1387.

[19] Pellegrino, F., Coupé, C., Marsico, E. 2011. Across-language perspective on speech information rate. *Language* 87(3), 539–558.

[20] Quené, H. 2013. Longitudinal trends in speech tempo: The case of queen beatrix. *The Journal of the Acoustical Society of America* 133(6), EL452–EL457.

[21] Quené, H., Van den Bergh, H. 2008. Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language* 59(4), 413–425.

[22] R Development Core Team, 2012. R: A language and environment for statistical computing. ISBN 3-900051-07-0.

[23] Reinisch, E. submitted. Segmental deletions make a sentence sound fast: evidence from normalization for speaking rate.

[24] Reinisch, E., Jesse, A., McQueen, J. M. 2011. Speaking rate affects the perception of duration as a suprasegmental lexical-stress cue. *Language and Speech* 54(2), 147–165.

[25] Reinisch, E., Jesse, A., McQueen, J. M. 2011. Speaking rate from proximal and distal contexts is used during word segmentation. *Journal of Experimental Psychology: Human Perception and Performance* 37(3), 978.

[26] Reinisch, E., Sjerps, M. J. 2013. The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. *Journal of Phonetics* 41(2), 101–116.

[27] Segalowitz, N., Hulstijn, J. H. 2005. Automaticity in bilingualism and second language learning. Kroll, J., De Groot, A., (eds), *Handbook of bilingualism: Psycholinguistic approaches* Oxford, UK. Oxford University Press 371–388.

[28] Tsao, Y.-C., Weismer, G. 1997. Interspeaker variation in habitual speaking rate: Evidence for a neuromuscular component. *Journal of Speech, Language, and Hearing Research* 40(4), 858–866.

[29] Van Engen, K. J., Peelle, J. E. 2014. Listening effort and accented speech. *Frontiers in Human Neuroscience* 8(577).

[30] Welch, T. E., Sawusch, J. R., Dent, M. L. 2009. Effects of syllable-final segment duration on the identification of synthetic speech continua by birds and humans. *The Journal of the Acoustical Society of America* 126(5), 2779–2787.