forum acusticum 2026
8-12 September · Graz, Austria

KEEPING UP WITH THE RHYTHM: TEMPORAL DYNAMICS OF SPEECH RATE NORMALIZATION

Azal Le Bagousse[1], Matthieu Fraticelli[1], Christopher Heffner[2], Paige Tuttösi[3,4], Léo Varnet[1]

[1] *Laboratoire des systèmes perceptifs, École normale supérieure, PSL University, CNRS, 75005 Paris, France*
[2] *University at Buffalo, Department of Communicative Disorders and Sciences, United States*
[3] *Simon Fraser University, School of Computing Science, Canada*
[4] *Simon Fraser University, Department of Linguistics, Canada*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

Keywords: *speech rate normalization, reverse correlation, psycholinguistics*

Speech comprehension relies on normalization mechanisms that help listeners cope with the natural variability in speech production. In speech rate normalization, for instance, the perception of an ambiguous target phoneme can be influenced by the rate of the preceding context. While previous research has demonstrated that both proximal and distal contexts contribute to the effect, they provide only limited insight into the the temporal structure of the rate integration window.

This study uses a prosodic reverse correlation approach to assess how 100-millisecond segments within the context sentence contribute to the speech normalization effect. Eight pairs of ambiguous sentences were used, each participant categorizing one pair across 300 trials. On each trial, every segment was randomly time-compressed or stretched. Segment-based correlations between compression factors and responses reveal how speech rate across different segments influences perception.

The results demonstrate a broad temporal integration window, with significant effects observed up to 1 second before the target. A detailed analysis of the correlation patterns across the eight sentences challenges the idea that loudness, lexical stress, or phonetic similarity alone drive speech rate integration. The presence of negative correlation coefficients further suggests that more complex mechanisms are involved in the temporal integration of speech rate.

Corresponding author: *leo.varnet@cnrs.fr.*

1. Introduction

An essential characteristic of the speech signal is the semi-regular rhythmic patterns contained in its slow temporal modulations, which are critical for comprehension [1], [2]. To reliably decode incoming sounds, the brain must synchronize with the low-frequency information in the speech envelope, enabling the parsing of continuous signals into intonational groups, words, syllables and phonemes [2], [3], [4]. However, spontaneous speech is inherently dynamic, with variations in speech rate and irregularities relative to any core rhythm frequency. Temporal modulations vary over time in both local mean frequency and regularity [5]. Speech often includes silences, breaks and restarts [6]. This challenges the assumption of a narrow, stable rhythmic frequency for neural synchronization. Instead, it is likely that the brain must form a time-varying estimation of the current speech rate.

A striking illustration of the brain's ability to track speech rate in real time is contextual speech rate normalization [4], [7]. In such experiments, the perception of an ambiguous target word is influenced by the speech rate of the preceding context. For instance, the ambiguous sequence of phonemes /biːnɒlɪʤ/ can be perceived as "bee knowledge" or "bean knowledge" depending on whether the context sentence "Bailey has much..." is sped up or slowed down [8]. Behavioral studies have explored the temporal extent of this effect, by manipulating separately the "proximal" rate and "distal" rate [8], [9], [10]. It has been demonstrated that distal speech rate 200 ms before the target can influence perception [11], and distal effects have been observed up to 400 ms before the target [8]. Interestingly, some authors have proposed that overall sentence structure, lexical stress patterns, and similarity to the target phoneme also modulate the strength of the distal effect [9], [10], [12].

This study aimed to obtain a more fine-grained insight on the temporal structure of the integration window for speech rate estimation. For this purpose, a prosodic reverse correlation paradigm was employed, similar to the approach used by Osses et al. [13] and Ponsot et al. [14]. The context sentence was divided into arbitrary 100-ms-long segments, whose speech rate was randomly varied by stretching/compression. Correlating segment-specific speech rate manipulations with participants' perceptual responses yields a "temporal kernel" that estimates the influence of each segment on perceived speech rate.

A recent study by Tuttösí et al. [15] used a similar approach to investigate speech normalization in L1 and L2 speakers, focusing on vowel laxness/tenseness constrasts (e.g. "peel" vs. "pill"). While their study had a different objective, it demonstrated the efficacy of reverse correlation approach in revealing the temporal profile of acoustic context effects. They observed that the temporal contour of rate information intake in context sentences was remarkably stable across individuals, in both French and English, with distal effects extending up to 800 ms before the target word. This study also manipulated pitch contours; however, pitch normalization effects were found to be less robust and more language-dependent than temporal contours. A follow-up study confirmed the generalizability of these temporal kernels to other sentences [16].

2. Materials and methods

2.1. Participants

Participants were recruited via the Prolific platform and paid £4.5 for 30 minutes of participation. The experiment was hosted on the MindProbe server. Participants were native English speakers (US/UK) with no reported hearing deficits. For this study, a total of 189 sets of data were collected across 8 different conditions.

Additionally, 12 "easy trials" were introduced randomly within the main experiment (see Screening phase). Only participants who achieved at least 75% accuracy (9/12 correct) on these trials were included in the final analysis, resulting in a total pool of 152 participants (see redistribution in Table 1).

2.2. Sentences and conditions

Stimuli were directly adapted from Heffner et al. [8], focusing on the three segmentation contrasts that exhibited the strongest effects in the original study. The original audio stimuli from that study were natural recordings made by 6 native speakers of American English (3 male, 3 female) who were naive to the study's purpose (see [8]).

In the first 3 conditions of the present study (C1 to C3), test stimuli consisted in a contextual sentence "Bailey has much ..." followed by the ambiguous sequence of phonemes /biːnɒlɪdʒ/ that could be interpreted either as "bee knowledge" or "bean knowledge" (see Fig.1). /n/ is the target phoneme since the segmentation of the ambiguous sentence depends on the perceived duration of this phoneme. Condition C4 was based on the same target word, preceded with a different context sentence: 'Jessica said /biːnɒlɪdʒ/' (C4). The remaining four conditions (C5-C8) used two different ambiguous targets: /laɪmaɪl/ ('lime aisle/mile', /m/ being the target phoneme) and /pleɪsʌmtaɪm/ ('place/play sometime', /s/ being the target phoneme). These ambiguous targets were included into different sentences: 'You can find it by the /laɪmaɪl/' (C5), 'Jessica often said /laɪmaɪl/' (C6), 'David asked if I'd seen his /pleɪsʌmtaɪm/' (C7) and 'Jessica frequently said /pleɪsʌmtaɪm/' (C8).

The following table lists the condition-specific sample-sizes:

Table 1: Number of recruited (R) and included (I) participants per condition (C)

C	1	2	3	4	5	6	7	8	**Total**
R	13	10	11	32	32	30	30	31	**189**
I	10	9	11	25	24	25	20	28	**152**

2.3. Screening phase

A screening procedure was used to ensure appropriate audio setup and to confirm that participants could reliably perform the auditory task. First, a binaural Huggins pitch headphone screening was conducted [17]. The procedure uses a local phase inversion between the left and right channels around 600 Hz. The resulting pitch is only audible through binaural integration so it requires a true stereo signal delivered to each ear and ensures proper headphone use. Participants would only be included if the correct audio setup was used. Then, an "easy-trial" task follows. Easy trials consisted of versions of the original sentence where the target phoneme had been compressed or stretched, resulting in unambiguous sentences. Since these manipulations clearly bias the perception toward long ('bean', 'mile', 'place') or short ('bee', 'aisle', 'play') words, they served as an objective measure of participant attention in the task. Twelve easy trials (6 slowed, 6 sped-up) were used in the screening phase, and participants would be excluded if not performing perfectly (12/12 score). Those control trials were also inserted randomly throughout the main experiment as an attention checker.

2.4. Experimental protocol

The study followed a between-subjects experimental design, with eight separate conditions. In each condition, participants completed 300 trials of a one-interval two-alternative task.

The context sentence was divided into 100-ms segments, with the latest segment ending 150 ms before the target phoneme (see Fig.1). The target

word itself remained unchanged throughout the experiment, while the context was manipulated from trial to trial. The speech rate of each segment was modified by stretching or compressing its duration by a random factor. At each trial, new random factors were drawn, independently for each segment, from a normal distribution (mean = 1.5 and SD = 1, bounded to ± 1 SD). These values were used to resynthesize the ambiguous stimulus with randomly compressed or stretched segments while preserving the f0, using the WORLD toolbox [18]. Contrary to previous implementation of the prosodic reverse correlation approach [13], [15], f0 contours were maintained as originally recorded because the focus of the present study was on the effect of speech rate alone.

For each trial, participants were required to choose between two potential interpretations of the target (e.g., 'bee knowledge' or 'bean knowledge'). Each participant tested in a given condition was presented the same set of 300 stimuli (i.e., with the same random seeds), in a random order.

2.5. Analysis

For each participant, the stretching/compression factor of each segment (> 1 streching, < 1 compression) was correlated with the binary perceptual responses. The segment-specific correlation coefficient are reported as a temporal kernel, revealing the influence of each segment's speech rate on target perception – and therefore the dynamics of speech rate estimation throughout the context sentence. Positive correlation coefficients indicate that the more stretched a segment is the more it induces a bias towards short target-word answers.

3. Results and discussion

Overall, compressing the context sentence led participants to perceive the ambiguous target phoneme as longer, while stretching the context sentence led them to perceive the target as shorter, resulting in positive correlation coefficients. This finding aligns with the existing literature on speech rate normalization and successfully replicates the results reported by Heffner et al. [8] on the same stimuli. However, the reverse correlation approach employed in this study provides a more detailed insight into the temporal dynamics underlying both proximal and distal context effects.

The first three conditions were designed to assess whether the specific set of 300 randomly generated stimuli influenced the measured kernel. For this purpose, 3 distinct sets were generated for the same sentence, each corresponding to a separate condition. While minor variations were observed among the three resulting kernels, an ANOVA revealed no significant effect of condition (Fig.1), confirming that the findings are not dependent on the particular set of 300 random stimuli used.

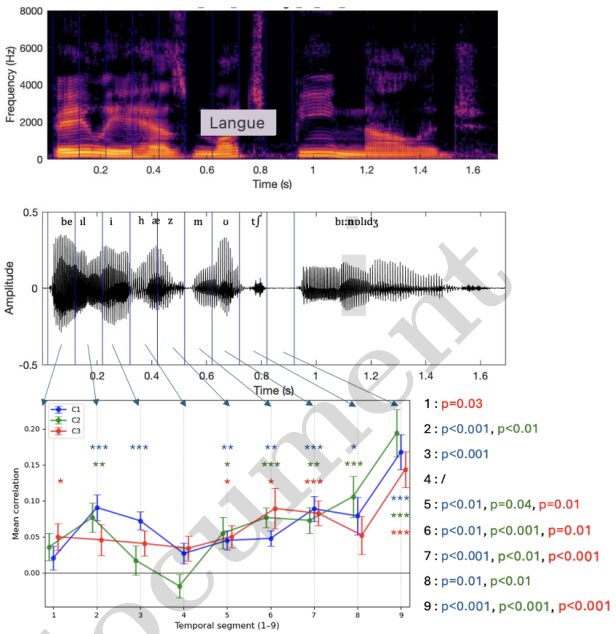

Fig. 1: Results for conditions C1, C2 and C3 (blue, green, red). Top: spectrogram of the original stimulus with with segments edges marked as blue vertical lines. Middle: waveform of the original stimulus, with the same blue edges. Bottom: temporal kernels averaged across participants (correlation between "bee knowledge" response and stretching factor applied to each temporal segment). *:p<0.05, **:p<0.01, ***:p<0.001.

Given the absence of significant differences between C1, C2, and C3, the data from these conditions were pooled for further analysis. Fig.2 presents the mean effect of segment compression/stretching on "bee knowledge"/"bean knowledge" perception in the sentence "Bailey has much /biːnɒlɪdʒ/". Separate t-tests conducted on each segments indicated that almost all values were significantly different from zero (p<0.05, without correction for multiple comparisons), except for Segment 4. The effect was the most pronounced for the final segment, closest to the target phoneme. Correlation coefficients demonstrated a general decline in the magnitude of the effect with increasing distance from the target phoneme (from segment 9 to segment 4). This pattern aligns with the well-documented observation that proximal contextual effects are typically stronger than distal ones [8].

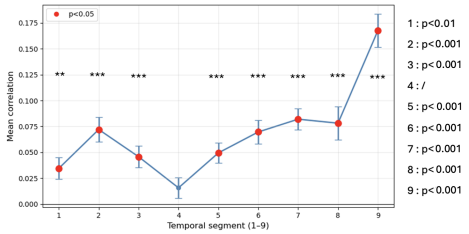

Fig. 2: Average temporal kernel across participants in conditions C1, C2 and C3. Red dots indicate weights significantly different from zero at the group level.

However, the temporal kernel also exhibited a distinct peak around segment 2, corresponding approximately to vowel /eɪ/ in 'Bailey'. This reflects a long-distance speech-normalization effect occurring more than 800 ms before the target phoneme – much further than the distal effect typically observed in previous studies [8]. The underlying mechanisms driving the enhanced weighting of these distal effects remain unclear. Previous research [9], [10], [12] has proposed several hypotheses to account for these effects, including (1) an increased contribution of loud segments, (2) phonetic similarity to the target, or (3) the presence of a lexical stress. Conditions C4 to C8 were included to further examine these potential explanations.

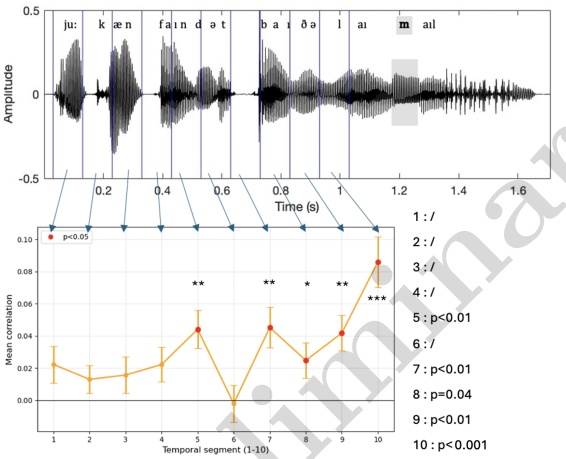

Fig. 3: Same caption as Fig.2 for C5 ('You can find it by the lime-aisle/mile')

In the sentence "You can find it by the /laɪmaɪl/" (C5, Fig.3), the temporal kernel shows a strong increase towards the end of the contextual sentence, similar to the results observed in conditions C1-C3. The final segment exhibited the strongest contribution, further indicating that listeners relied more heavily on the information closest to the target. Yet, again, some segments were found to be weighted more heavily than their temporal distance from the target would predict: segment 5 (/md/) and 7 (/tb/). In the sentence "David asked if I'd seen his /pleɪsʌmtaɪm/" (C7, Fig.4), the weighting pattern was irregularly distributed across segments, with multiple regions contributing positively to perception. Due to substantial inter-participant variability in this condition, no segment weights reached statistical significance. Nevertheless, segments corresponding to /ɪd/, /s/ and /aɪd/ were all associated with relatively stronger weights, with a pronounced peak on segment 12, which aligns with the position of the last vowel of the context sentence (/ɪ/).

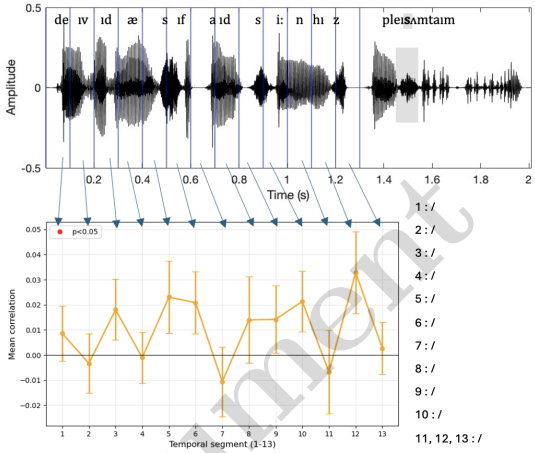

Fig. 4: Same caption as Fig.2 for C7 ('David asked if I'd seen his place/play-sometime')

Across the three sentences in conditions C4 ('Jessica said /biːnɒlɪdʒ/'), C6 ('Jessica often said /laɪmaɪl/') and C8 ('Jessica frequently said /pleɪsʌmtaɪm/'), consistent patterns emerge in the temporal weighting of acoustic segments. Some vowels, such as the /a/ in 'Jessica' and the /ɛ/ in 'said', were systematically associated with higher positive weights, suggesting that listeners relied more strongly on these segments (Fig.5,6,7). Unexpectedly, C4 also exhibited a significant negative weight on the last segment (Fig.5). One possible explanation would be related to the presence or absence of co-articulation on the /d/ and /b/ consonants, which may modulate the role of these segments in speech rate estimation.

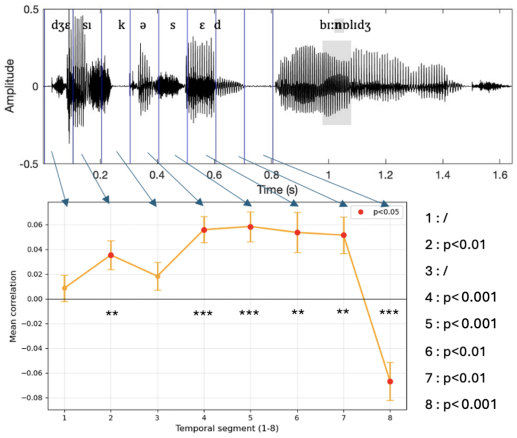

Fig. 5: Same caption as Fig.2 for C4 ('Jessica said bean/bee-knowledge')

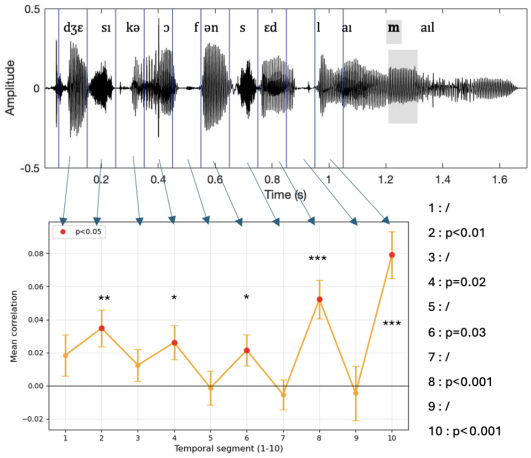

Fig. 6: Same caption as Fig.2 for C6 ('Jessica often said lime-aisle/mile')

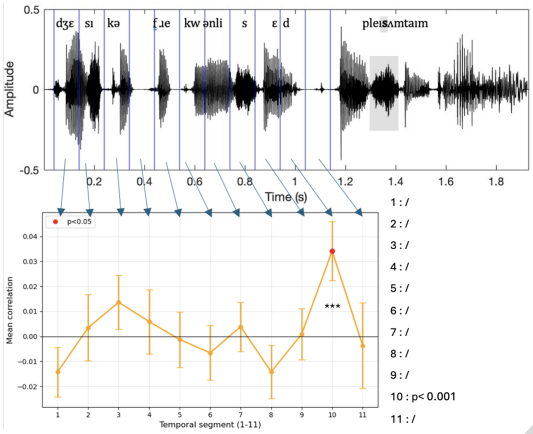

Fig. 7: Same caption as Fig.2 for C8 ('Jessica frequently said place/play-sometime')

Comparing the kernels obtained for condition C1 to C8 offers critical insights into the hypotheses outlined earlier. First, loudness does not appear to be a primary determinant of the contribution of a segment to the estimation of overall speech rate. This is evident in several conditions where silent or quasi-silent segments (such as segment 9 in C1-C3, segment 7 in C5, segment 8 in C4) were heavily weighted, a finding that is reminiscent of previous studies indicating that inserting proximal silent or noisy segments overrides distal normalization effects [19]. More generally, the loudest segments were not consistently associated with the strongest perceptual weights.

Assessing the role of phonetic similarity to the target is more challenging. Although the set of sentence was not designed to test this factor, the available evidence suggest that phonetic similarity is not a critical determinant: for instance, the ambiguous target /pleɪsʌmtaɪm/ in C8 (Fig.7) resulted in kernels with weak weights on segments containing the phoneme /s/.

Finally, the temporal kernels measured in the different conditions do not consistently align with the pattern of lexical stress. While, in conditions C1-3, the stressed syllables "Bai" and "much" are associated with significant weights, the three conditions C4, C6 and C8 show a strong correlation in the final /ə/ in "Jessica" although this vowel does not bear lexical stress.

Taken together, these results challenge the three primary hypotheses proposed to explain the weighting of distal segments in speech rate integration. Neither loudness, lexical stress, nor phonetic similarity to the target phoneme emerged as consistent or dominant factors in shaping the temporal kernels observed across conditions. The lack of alignment with these hypotheses suggests that the mechanisms underlying speech rate normalization are more complex and/or likely involve additional, as-yet-unidentified factors. To further clarify how listeners integrate speech rate information, this analysis will be extended to additional sentences in future research.

A major finding of this study is the extended duration of the speech rate integration window. While distal context effects have been extensively explored in the past, the results have been inconsistent, with some studies finding significant distal effects while other failed to do so. This inconsistency may stem from the variable definitions of what constitutes a "distal" region, as noted by Heffner et al. [8], since experimental procedures have so far typically consisted in manipulating the speech rate of a predefined distal region, which inherently relies on arbitrary temporal boundaries. The reverse correlation approach offers a distinct advantage in this regard as it does not require to isolate a region based on a priori criteria, but instead reveals the continuous shape of the temporal integration window over time. This allowed us to reveal that distal context effects can be both relatively strong and temporally localized – perhaps explaining why previous studies manipulating broader segments yielded inconsistent results. In C1-C4, segments located around 800 ms before the target phoneme significantly influenced its perception. In C6, significant weights were observed up to 1 second prior to the target phoneme (/s/ in 'Jessica'). Such long-distance speech-normalization effects are substantially larger to what has been typically reported in the literature. However, similar findigs have been previously documented by Tuttösì et al [15], suggesting that such effects may be relatively common in speech and that listeners may build a representation of the speaker's rate over a much longer time window than previously assumed.

Finally, the case of the final silent segment in conditions C1-C4 is particularly surprising: despite this segment being essentially silent (see Fig.1 and Fig.5) and containing similar phonetic information in the two sentences (the closure preceding the release of /b/ in /biːnplɪdʒ/), it was significantly larger than

zero in C1-C3, but significantly lower than zero in C4. Such negative weights might suggest a double comparison mechanism, with the duration of the target phoneme /n/ in C4 evaluated relative to the speech rate in a preceding segment, which is assessed relative to the duration of the /b/ closure silence.

4. Acknowledgments

The authors are grateful to B. Jiang for conducting a pilot version of this experiment. This study was supported by the ANR-DFG grant DRhyADS (ANR-22-FRAL-0003) to L. Varnet and A. Tavano, and the EUR Frontiers in Cognition (ANR-17-EURE-0017).

References

[1] L. Varnet, M. C. Ortiz-Barajas, R. G. Erra, J. Gervain, and C. Lorenzi, "A cross-linguistic study of speech modulation spectra," *The Journal of the Acoustical Society of America*, vol. 142, no. 4, p. 1976, 2017. DOI: 10.1121/1.5006179

[2] D. Poeppel and M. F. Assaneo, "Speech rhythms and their neural foundations," *Nature Reviews Neuroscience*, vol. 21, no. 6, pp. 322–334, 2020. DOI: 10.1038/s41583-020-0304-4

[3] A.-L. Giraud and D. Poeppel, "Speech Perception from a Neurophysiological Perspective," in *The Human Auditory Cortex*, 43, Springer, 2012, pp. 225–260.

[4] J. E. Peelle and M. H. Davis, "Neural Oscillations Carry Speech Rhythm through to Comprehension," *Frontiers in Psychology*, vol. 3, 2012. DOI: 10.3389/fpsyg.2012.00320

[5] R. Fusaroli and K. Tylén, "Investigating conversational dynamics: Interactive alignment, Interpersonal synergy, and collective task performance," *Cognitive science*, vol. 40, no. 1, pp. 145–171, 2016.

[6] S. Garrod and M. J. Pickering, "Why is conversation so easy?" *Trends in cognitive sciences*, vol. 8, no. 1, pp. 8–11, 2004.

[7] A. Kösem, H. R. Bosker, A. Takashima, A. Meyer, O. Jensen, and P. Hagoort, "Neural Entrainment Determines the Words We Hear," *Current Biology*, vol. 28, no. 18, 2867–2875.e3, 2018. DOI: 10.1016/j.cub.2018.07.023

[8] C. C. Heffner, R. S. Newman, and W. J. Idsardi, "Support for context effects on segmentation and segments depends on the context," *Attention, Perception & Psychophysics*, vol. 79, no. 3, pp. 964–988, 2017. DOI: 10.3758/s13414-016-1274-5

[9] J. Steffman, "Rhythmic and speech rate effects in the perception of durational cues," *Attention, Perception, & Psychophysics*, vol. 83, no. 8, pp. 3162–3182, 2021. DOI: 10.3758/s13414-021-02334-w

[10] Q. Summerfield, "Articulatory rate and perceptual constancy in phonetic perception," *Journal of Experimental Psychology. Human Perception and Performance*, vol. 7, no. 5, pp. 1074–1095, 1981. DOI: 10.1037//0096-1523.7.5.1074

[11] E. Reinisch, A. Jesse, and J. M. McQueen, "Speaking rate from proximal and distal contexts is used during word segmentation," *Journal of Experimental Psychology. Human Perception and Performance*, vol. 37, no. 3, pp. 978–996, 2011. DOI: 10.1037/a0021923

[12] G. R. Kidd, "Articulatory-rate context effects in phoneme identification," *Journal of Experimental Psychology. Human Perception and Performance*, vol. 15, no. 4, pp. 736–748, 1989. DOI: 10.1037//0096-1523.15.4.736

[13] A. Osses, E. Spinelli, F. Meunier, E. Gaudrain, and L. Varnet, "Prosodic cues to word boundaries in a segmentation task assessed using reverse correlation," *JASA Express Letters*, vol. 3, no. 9, p. 095 205, 2023. DOI: 10.1121/10.0021022

[14] E. Ponsot, J. J. Burred, P. Belin, and J.-J. Aucouturier, "Cracking the social code of speech prosody using reverse correlation," *Proceedings of the National Academy of Sciences*, vol. 115, no. 15, pp. 3972–3977, 2018. DOI: 10.1073/pnas.1716090115

[15] P. Tuttösí et al., "Mmm whatcha say? Uncovering distal and proximal context effects in first and second-language word perception using psychophysical reverse correlation," in *Interspeech 2024*, 2024, pp. 1010–1014. DOI: 10.21437/Interspeech.2024-1296

[16] P. Tuttösí, H. Yeung, Y. Wang, J.-J. Aucouturier, and A. Lim, "You Sound a Little Tense: L2 Tailored Clear TTS Using Durational Vowel Properties," in *Proc. SSW 2025*, 2025, pp. 228–235. DOI: 10.21437/SSW.2025-35

[17] A. E. Milne et al., "An online headphone screening test based on dichotic pitch," *Behavior Research Methods*, vol. 53, no. 4, pp. 1551–1562, 2021. DOI: 10.3758/s13428-020-01514-0

[18] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016. DOI: 10.1587/transinf.2015EDP7457

[19] 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 and Cognitive Processes*, vol. 28, no. 9, pp. 1275–1302, 2013. DOI: 10.1080/01690965.2012.672229