INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

An entrained rhythm's frequency, not phase, influences temporal sampling of speech

Hans Rutger Bosker[1,2], *Anne Kösem*[1,2]

[1]Max Planck Institute for Psycholinguistics, PO Box 310, 6500 AH, Nijmegen, Netherlands
[2]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, Netherlands
HansRutger.Bosker@mpi.nl, A.Kosem@donders.ru.nl

Abstract

Brain oscillations have been shown to track the slow amplitude fluctuations in speech during comprehension. Moreover, there is evidence that these stimulus-induced cortical rhythms may persist even after the driving stimulus has ceased. However, how exactly this neural entrainment shapes speech perception remains debated. This behavioral study investigated whether and how the *frequency* and *phase* of an entrained rhythm would influence the temporal sampling of subsequent speech.

In two behavioral experiments, participants were presented with slow and fast isochronous tone sequences, followed by Dutch target words ambiguous between *as* /ɑs/ "ash" (with a short vowel) and *aas* /a:s/ "bait" (with a long vowel). Target words were presented at various phases of the entrained rhythm. Both experiments revealed effects of the frequency of the tone sequence on target word perception: fast sequences biased listeners to more long /a:s/ responses. However, no evidence for phase effects could be discerned.

These findings show that an entrained rhythm's frequency, but not phase, influences the temporal sampling of subsequent speech. These outcomes are compatible with theories suggesting that sensory timing is evaluated relative to entrained *frequency*. Furthermore, they suggest that *phase* tracking of (syllabic) rhythms by theta oscillations plays a limited role in speech parsing.

Index Terms: neural entrainment, phase-locking, temporal sampling, speech parsing, rate normalization, speech rate.

1. Introduction

Speech is a sensory signal that has strong rhythmic features [1]. The brain has been shown to track these features by phase-locking endogenous neural oscillations to the slow amplitude fluctuations in the speech signal [2-4]. However, it remains debated whether this neural tracking of speech is merely an epiphenomenon of speech processing or a causal factor contributing to successful speech comprehension.

Recent neurocognitive models of speech perception [5-7] suggest that this neural tracking is in fact instrumental for speech comprehension. Stimulus-induced neural entrainment provides a way to temporally organize the incoming speech signal. Endogenous theta oscillations (in the 3-8 Hz range) are thought to adjust their phase and frequency to the syllabic rhythm of speech. Entrained theta oscillations would control neuronal excitability in primary auditory areas [8]. The periodic excitation and inhibition imposed by theta oscillations would consequently sample the input signal at the syllabic temporal granularity [4, 5]. Hence, oscillatory models of speech processing posit the existence of neural rhythms that flexibly adapt to speech input to optimize comprehension.

It has been suggested that stimulus-induced cortical rhythms may persist even after stimulation has ceased [9-11]. The first prediction that follows from this is that the *frequency* of neural entrainment would influence subsequent perception. If the entrained neural oscillations sample speech information into syllabic units, then the frequency of neural entrainment should define the sampling frequency of the future speech segments. For instance, entrainment at a high theta frequency (e.g., 7 Hz) would raise the cortical 'sampling frequency'. If the following speech segment is suddenly presented at a slower rate than the preceding rhythm, entrainment will potentially lead to 'oversampling' the target speech item, inducing overestimation of the target's duration [12]. Similarly, entrainment to a lower frequency (e.g., 4 Hz) would lower the cortical 'sampling frequency', thus 'undersampling' a target stimulus presented at a faster rate than the preceding rhythm (i.e., resulting in underestimation of target duration). In line with this hypothesis, behavioral studies have reported that preceding speech rates influence the comprehension of subsequent words [12-14] and a recent MEG study suggests that changes in the frequency of entrainment in temporal regions is capable of modulating subsequent perception [15].

Additionally, if neural entrainment indeed guides temporal sampling of speech, a second prediction is that the perception of speech sounds would be dependent on the *phase* of entrainment. Phase effects have been observed in auditory detection tasks [11, 16, 17]; for instance, auditory perception of near-threshold click trains is inhibited in low excitability phases [18]. Also, studies have suggested that the phase of theta oscillations play a role in phonemic processing [19, 20]. Here, we predict that, if the target is presented in a low excitability phase of the entrained rhythm, part of the stimulus information would be inhibited (relative to a high excitability phase), inducing underestimation of the target's duration.

To test these two predictions, we adopted the 'rate normalization' paradigm, typically used to investigate effects of surrounding speech rate on the perception of subsequent words [12-14, 21, 22]. Participants were presented with fast and slow isochronous tone precursors, after which Dutch target words followed that were ambiguous between *as* /ɑs/ "ash" (with a short vowel) and *aas* /a:s/ "bait" (with a long vowel). If perception is sensitive to the *frequency* of entrainment, target word categorization should show a higher percentage of long /a:s/ responses after fast tone precursors. If perception is sensitive to the *phase* of entrainment, categorization data should reveal a 'perceptual oscillation' [23], with higher percentages of long /a:s/ responses in high excitability than low excitability phases.

http://dx.doi.org/10.21437/Interspeech.2017-73

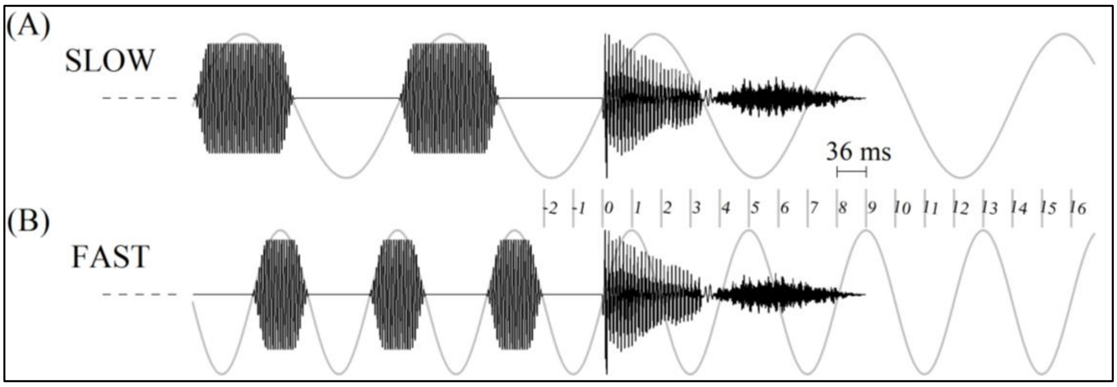

Figure 1: *Schematic diagram of slow and fast tone precursors and following target words, with hypothetical entrained oscillators overlaid. The various time-points of vowel onset in the two experiments, spaced 36 ms apart, are indicated (-2 to 9 in Experiment 1; -2 to 16 in Experiment 2).*

2. Experiment 1

2.1. Method

2.1.1. Participants

Native Dutch participants (N = 19) with normal hearing were recruited from the Max Planck Institute's participant pool. They gave informed consent as approved by the Ethics Committee of the Social Sciences department of Radboud University (project code: ECSW2014-1003-196). One participant's data were excluded because this participant only reported hearing /a:s/.

2.1.2. Materials

The stimuli in the experiment consisted of isochronous tone precursors followed by target words (see Figure 1). Two different precursors were created in Praat [24]. The 'fast' precursor consisted of 16 sine tones (each 71 ms long including 20 ms rise-and-decay times) presented at an isochronous rate of 7 Hz. The 'slow' precursor consisted of 8 sine tones (each 143 ms long including 20 ms rise-and-decay times) presented at an isochronous rate of 3.5 Hz (i.e., twice as slow). Rates were selected to fall within the range of typical speech rates. Fundamental frequency of the tones was fixed at 440 Hz, avoiding spectral masking of target vowel formants.

Target words were created by recording a female native Dutch speaker producing the minimal pair *as* /ɑs/ "ash" - *aas* /a:s/ "bait". Because the Dutch /ɑ/-/a:/ contrast is cued by both spectral and temporal characteristics (/ɑ/ has a shorter duration as well as lower formant values than /a:/; [25]), spectral and temporal manipulations were performed in order to create ambiguous target vowels. Using one long /a:/ vowel, a *spectral F2 continuum* was created based on Burg's LPC method in Praat, with the source and filter models estimated automatically from the selected vowel. The formant values in the filter models were inspected and adjusted to result in a fixed ambiguous F1 value (810 Hz) and one of 3 desired F2 values (1400 - 1500 Hz in steps of 50 Hz; all falling with the speaker's natural range). Then, the source and filter models were recombined resulting in three different vowel tokens which were all given a fixed ambiguous duration of 130 ms using PSOLA. After the vowels were adjusted to have the same overall amplitude as the original vowel, they were combined with one single /s/ token (200 ms long) to form 3 manipulated target words.

2.1.3. Procedure

Stimulus presentation was controlled by Presentation software (Version 16.5; Neurobehavioral Systems, Albany, CA, USA), with auditory stimuli presented over headphones at a comfortable level in sound-attenuating booths. A trial started with the presentation of a fixation cross in the middle of the screen. After 330 ms, a tone precursor was presented, followed by a target word. The onset of the target word was varied, relative to the entrained rhythm, using 11 different time-points, spaced 36 ms apart (see Figure 1). One time-point (time point 0) was the 'expected' time point, with the target word presented isochronous with the preceding rhythm. Two time-points were 'early' (71 and 36 ms before time-point 0) and eight were 'late' (after time-point 0). All precursor and target combinations appeared equally often at all time-points (in random order), repeated four times (total number of trials = 264; 2 precursors x 3 vowels x 11 time-points x 4 repetitions).

At target offset, the fixation cross was replaced by two response options *as* and *aas* on the left and right side of the screen (position counterbalanced across participants), and participants were instructed to indicate by button press which target word they had heard ("1" for the left word; "0" for the right word). After participants' response (or timeout after 5 seconds), a blank screen was presented for 660 ms, after which the next trial started immediately.

2.2. Results

Categorization data of Experiment 1, calculated as the percentage of /a:/ responses (% /a:/), are presented in Figure 2 for each precursor and each time-point, averaging over the three different vowel tokens.

A generalized linear mixed model with a logistic linking function as implemented in the lme4 library [26] in R [27] tested participants' binomial responses (1 = /a:/; 0 = /ɑ/) for fixed effects of Vowel F2 (continuous linear predictor, scaled and centered around the mean) and Precursor Rate (categorical predictor, with 'fast' mapped onto the intercept), with random effects of Participants. Interaction terms were not included because they did not significantly improve the model's fit. By-participant random slopes for all fixed effects were included in the model. This model revealed a significant effect of Vowel F2 (β = .679, z = 6.491, p < .001; the higher the F2, the more /a:/ responses) and Precursor Rate (β = -.320, z = -2.703, p = .007; fewer /a:/ responses after 'slow' precursor than after 'fast' precursor).

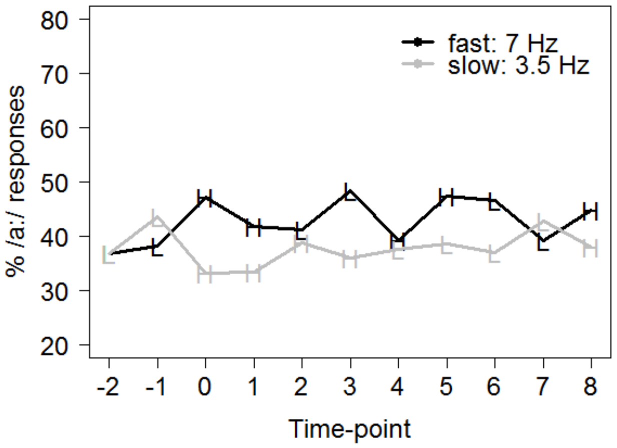

Figure 2: *Categorization data from Experiment 1 for each time-point, split by precursor rate. H = high excitability phase; L = low excitability phase.*

If categorization would have been sensitive to entrained phase, we would predict different categorization of trials with targets in high excitability phases vs. low excitability phases. Because the precise phase of target presentation depended on the preceding precursors (see Figure 2), phase effects were quantified in two separate models: one for fast trials and one for slow trials. Both models tested participants' binomial responses for fixed effects of Vowel F2 and Phase (categorical predictor with two levels: H vs. L; see Figure 2), and their interaction, with random effects of Participants and by-participant random slopes for all fixed effects. Neither model revealed an effect of Phase, providing no evidence for differences in categorization of trials with targets in high excitability phases vs. low excitability phases.

To further inspect oscillatory activity in the behavioral responses, the power spectra of the vowel categorization curves of each individual participant were computed using the Fourier transform. The average power spectra (Figure 3) reveal that there was little evidence for a peak in power around the entrained frequency (3.5 Hz for slow; 7 Hz for fast).

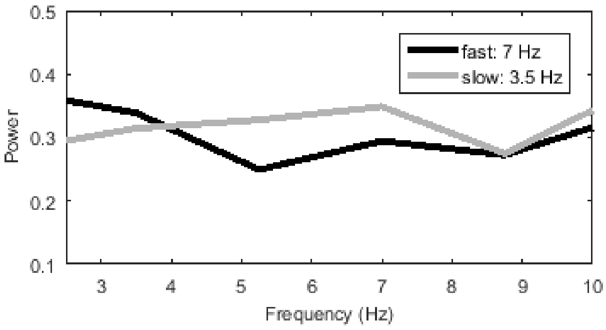

Figure 3: *Average power spectra of the categorization curves from Experiment 1, split by precursor rate.*

3. Experiment 2

Experiment 2 was identical to Experiment 1, aiming to test effects of the frequency and phase of an entrained rhythm on vowel perception, only this time making use of longer precursors, and testing more time-points.

3.1. Method

3.1.1. Participants

Native Dutch participants (N = 23) with normal hearing were recruited from the Max Planck Institute's participant pool. They gave informed consent as approved by the Ethics Committee of the Social Sciences department of Radboud University (project code: ECSW2014-1003-196).

3.1.2. Materials

Experiment 2 used the materials from Experiment 1, with three adjustments. First, the tone precursors were extended (21 instead of 16 tones in the 'fast' precursor; 12 instead of 8 tones in the 'slow' precursor) to provide more opportunity for entrainment to the precursor rhythm prior to target onset. Second, the rate of the 'slow' precursor was adjusted to avoid harmonics at the frequency of the 'fast' precursor (4 Hz instead of 3.5 Hz, with tone durations of 125 ms). Third, in order to present slightly more ambiguous vowels, a vowel continuum was used with 3 vowels ranging in F2 from 1450-1550 Hz in steps of 50 Hz, and a fixed duration of 120 ms.

3.1.3. Procedure

Experiment 2 adopted the procedure from Experiment 1, with three adjustments. First, more time-points were used (19 instead of 11) to improve the frequency resolution in the time-frequency analysis. Second, in order to increase the 'expectedness' of the target, target words were presented at the 'expected' time-point (time-point 0) four times as often as other time-points. Because this led to a relatively large number of trials, each participant was tested in two sessions (separated by approximately one week). Third, in Experiment 2, each trial started exactly 7000 ms after the onset of the previous trial. That is, a trial started with a fixation cross and after 1000 ms the auditory stimulus was presented, followed by the response screen. The next trial was initiated exactly 7000 ms after the previous trial, irrespective of whether a participant entered a response or not.

3.2. Results

Categorization data of Experiment 2 are presented in Figure 4 (<2% missing responses excluded).

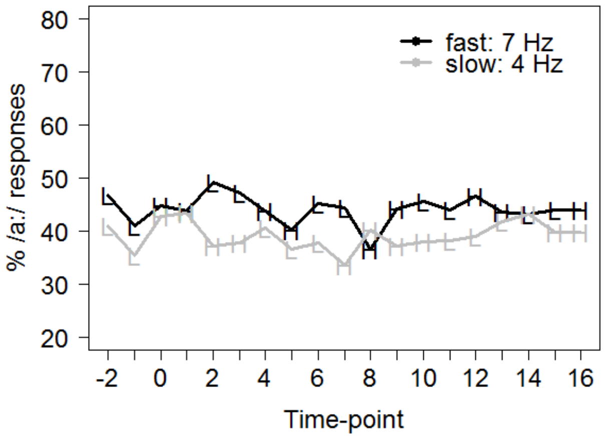

Figure 4: *Categorization data from Experiment 2 for each time-point, split by precursor rate. H = high excitability phase; L = low excitability phase.*

A generalized linear mixed model (identical in structure to the one introduced in Experiment 1) revealed a significant effect of Vowel F2 (β = .670, z = 7.465, p < .001; the higher the F2, the more /a:/ responses) and a marginally significant effect of Precursor Rate (β = -.259, z = -1.835, p = .067; fewer /a:/ responses after 'slow' precursor than after 'fast' precursor).

Similar to Experiment 1, differences between trials with target onsets in high excitability phases vs. low excitability phases (see Figure 4) were quantified separately for trials with fast vs. slow precursors. Models were built that were identical in structure to the ones presented in Experiment 1; however, neither model revealed an effect of Phase, providing no evidence for differences in target categorization in high vs. low excitability phases. In the same vein, average power spectra of the individual vowel categorization curves (cf. Figure 5) again revealed little evidence for a peak in power around the entrained frequency (4 Hz for slow; 7 Hz for fast).

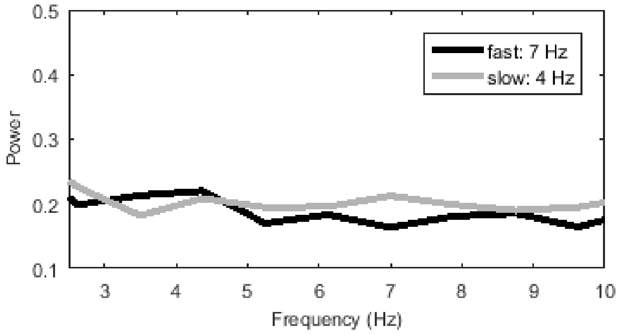

Figure 5: *Average power spectra of the categorization curves from Experiment 2, split by precursor rate.*

4. Discussion

This study involved a behavioral investigation into whether an entrained rhythm's *frequency* and/or *phase* influence temporal sampling of subsequent speech (i.e., when the entraining rhythm has already ceased). Participants were presented with fast and slow tone precursors, followed by target words ambiguous between Dutch *as* /ɑs/ "ash" (with short vowel) and *aas* /a:s/ "bait" (with long vowel). Target word onset was varied such that target vowels occurred in distinct phases of the entrained rhythm.

Results from Experiment 1 revealed an effect of the entrained rhythm's *frequency* on subsequent perception: listeners reported a higher percentage of /a:/ responses after fast tone precursors. However, no evidence was found for effects of entrained *phase*: the hypothesis that perception might fluctuate with entrained phase was not supported.

Experiment 2 mirrored Experiment 1, using an adjusted experimental design (e.g., longer precursors and more time-points). Results from Experiment 2 showed a marginally significant effect of the entrained *frequency* and, again, no evidence for *phase* effects. The fact that the frequency effect in Experiment 2 did not quite reach statistical significance may be explained by the smaller rate difference between fast and slow precursors in Experiment 2 (4 vs. 7 Hz; cf. 3.5 vs. 7 Hz in Experiment 1), and the relatively longer temporal distance between precursor and target in Experiment 2 (max. 643 ms; cf. max. of 357 ms in Experiment 1).

The observed effects of a preceding rhythm's *frequency* are in line with previous findings showing that preceding (syllabic) rhythms influence speech perception [12, 13, 15, 28]. At the same time, the present study provides no clear evidence for *phase* effects on speech perception.

The phase of neural oscillations is thought to control auditory processing, so that an auditory input would be amplified if presented at a certain phase, and suppressed if presented at the opposite phase [8]. Consistent with this hypothesis, an entrained rhythm's phase modulates participants' accuracy in detecting near-threshold tones, either embedded in the rhythm [17, 18, 29, 30] or following it [11]. The phase of entrained oscillations may also impact the early stages of speech processing: the perception of consonants in syllable onsets is dependent on the phase of theta oscillations [19], and specific phonemic features can be decoded from the dynamics of low-frequency brain oscillations [20]. However, recent findings show that the phase of entrained oscillations does not reliably indicate the perceived segmentation of words within ambiguous speech streams [31]. This suggests that, while the phase tracking of syllabic rhythms by theta oscillations could be instrumental for the acoustic and phonemic analysis of speech, it might not play a strong role in speech parsing. The absence of phase effects in the present study also supports this view.

The concurrent *presence* of frequency effects and the *absence* of phase effects may be accounted for by alternative functions of neural oscillations in sensory processing. In addition to the absolute phase, which refers to the excitability state of the neural network, the relative phase of firing may encode information about the content of sensory events [32]. In particular, low-frequency entrained oscillations could provide a temporal metric used to evaluate sensory timing [33]. During speech listening, neural oscillations would create a temporal reference frame on which the duration of syllabic events is estimated by measuring the relative phase distance between the syllable onset and offset. This model would predict bigger phase distances (leading to overestimation of the item's duration) for faster frequencies of entrainment compared to slower frequencies, which is what was observed in this study.

5. Conclusions

This study showed effects of a preceding rhythm's frequency on subsequent temporal sampling of speech: an entrained rhythm with a relatively high frequency biased perception of subsequent speech sounds towards longer segments. No evidence was observed for effects of entrained phase on speech perception, in contrast to previously observed phase effects in sound processing. These findings suggest different roles of neural entrainment at distinct stages of the speech processing hierarchy.

6. Acknowledgements

This research was supported by a Gravitation grant from the Dutch Government to the Language in Interaction Consortium. We would like to thank Ronald Fischer for technical assistance and Annelies van Wijngaarden for allowing us to record her voice for the speech materials in this study. Thanks also to Anne Hoek, Carlien Alferink, Milou Huijsmans, and Nikki van Gasteren for testing participants.

7. References

[1] N. Ding, A. Patel, L. Chen, H. Butler, C. Luo, and D. Poeppel, "Temporal modulations in speech and music," *Neuroscience and Biobehavioral Reviews,* Online version, 2017.

[2] H. Luo and D. Poeppel, "Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex," *Neuron,* vol. 54, pp. 1001-1010, 2007.

[3] K. B. Doelling, L. H. Arnal, O. Ghitza, and D. Poeppel, "Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing," *Neuroimage,* vol. 85, pp. 761-768, 2014.

[4] J. E. Peelle, J. Gross, and M. H. Davis, "Phase-locked responses to speech in human auditory cortex are enhanced during comprehension," *Cerebral Cortex,* vol. 23, pp. 1378-1387, 2013.

[5] A.-L. Giraud and D. Poeppel, "Cortical oscillations and speech processing: emerging computational principles and operations," *Nature Neuroscience,* vol. 15, pp. 511-517, 2012.

[6] O. Ghitza, "Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm," *Frontiers in Psychology,* vol. 2, 2011.

[7] J. E. Peelle and M. H. Davis, "Neural oscillations carry speech rhythm through to comprehension," *Frontiers in psychology,* vol. 3, 2012.

[8] P. Lakatos, A. S. Shah, K. H. Knuth, I. Ulbert, G. Karmos, and C. E. Schroeder, "An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex," *Journal of Neurophysiology,* vol. 94, pp. 1904-1911, 2005.

[9] E. Spaak, F. P. de Lange, and O. Jensen, "Local entrainment of alpha oscillations by visual stimuli causes cyclic modulation of perception," *The Journal of Neuroscience,* vol. 34, pp. 3536-3544, 2014.

[10] A. Kösem and V. Van Wassenhove, "Distinct contributions of low-and high-frequency neural oscillations to speech comprehension," *Language, Cognition and Neuroscience,* Online Version, pp. 1-9, 2016.

[11] G. Hickok, H. Farahbod, and K. Saberi, "The rhythm of perception: entrainment to acoustic rhythms induces subsequent perceptual oscillation," *Psychological Science,* vol. 26, pp. 1006-1013, 2015.

[12] H. R. Bosker, "Accounting for rate-dependent category boundary shifts in speech perception," *Attention, Perception & Psychophysics,* vol. 79, pp. 333-343, 2017.

[13] E. Reinisch and M. J. Sjerps, "The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context," *Journal of Phonetics,* vol. 41, pp. 101-116, 2013.

[14] H. R. Bosker, "How our own voice influences our perception of others," *Journal of Experimental Psychology: Learning, Memory, and Cognition,* Online Version, 2017.

[15] A. Kösem, H. R. Bosker, A. Meyer, O. Jensen, and P. Hagoort, "Neural entrainment reflects temporal predictions guiding speech comprehension," presented at the Eighth Annual Meeting of the Society for the Neurobiology of Language (SNL 2016), London, UK, 2016.

[16] M. J. Henry and J. Obleser, "Frequency modulation entrains slow neural oscillations and optimizes human listening behavior," *Proceedings of the National Academy of Sciences,* vol. 109, pp. 20095-20100, 2012.

[17] M. J. Henry, B. Herrmann, and J. Obleser, "Entrained neural oscillations in multiple frequency bands comodulate behavior," *Proceedings of the National Academy of Sciences,* vol. 111, pp. 14935-14940, 2014.

[18] L. Riecke, E. Formisano, C. S. Herrmann, and A. T. Sack, "4-Hz transcranial alternating current stimulation phase modulates hearing," *Brain Stimulation,* vol. 8, pp. 777-783, 2015.

[19] S. ten Oever and A. T. Sack, "Oscillatory phase shapes syllable perception," *Proceedings of the National Academy of Sciences,* vol. 112, pp. 15833-15837, 2015.

[20] G. M. Di Liberto, J. A. O'Sullivan, and E. C. Lalor, "Low-frequency cortical entrainment to speech reflects phoneme-level processing," *Current Biology,* vol. 25, pp. 2457-2465, 2015.

[21] H. R. Bosker, E. Reinisch, and M. J. Sjerps, "Cognitive load makes speech sound fast but does not modulate acoustic context effects," *Journal of Memory and Language,* vol. 94, pp. 166-176, 2017.

[22] J. C. Toscano and B. McMurray, "The time-course of speaking rate compensation: effects of sentential rate and vowel length on voicing judgments," *Language, Cognition and Neuroscience,* vol. 30, pp. 529-543, 2015.

[23] R. VanRullen, "Perceptual cycles," *Trends in Cognitive Sciences,* vol. 20, pp. 723-735, 2016.

[24] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," 2016.

[25] P. Escudero, T. Benders, and S. C. Lipski, "Native, non-native and L2 perceptual cue weighting for Dutch vowels: The case of Dutch, German, and Spanish listeners," *Journal of Phonetics,* vol. 37, pp. 452-465, 2009.

[26] D. Bates, M. Maechler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software,* vol. 67, pp. 1-48, 2015.

[27] R Development Core Team, "R: A Language and Environment for Statistical Computing [computer program]," 2012.

[28] H. Quené and R. Port, "Effects of timing regularity and metrical expectancy on spoken-word perception," *Phonetica,* vol. 62, pp. 1-13, 2005.

[29] B. Zoefel and R. VanRullen, "Selective perceptual phase entrainment to speech rhythm in the absence of spectral energy fluctuations," *The Journal of Neuroscience,* vol. 35, pp. 1954-1964, 2015.

[30] P. Lakatos, G. Musacchia, M. N. O'Connel, A. Y. Falchier, D. C. Javitt, and C. E. Schroeder, "The spectrotemporal filter mechanism of auditory selective attention," *Neuron,* vol. 77, pp. 750-761, 2013.

[31] A. Kösem, A. Basirat, L. Azizi, and V. van Wassenhove, "High-frequency neural activity predicts word parsing in ambiguous speech streams," *Journal of Neurophysiology,* vol. 116, pp. 2497-2512, 2016.

[32] C. Kayser, R. A. Ince, and S. Panzeri, "Analysis of slow (theta) oscillations as a potential temporal reference frame for information coding in sensory cortices," *PLoS Computational Biology,* vol. 8, p. e1002717, 2012.

[33] A. Kösem, A. Gramfort, and V. van Wassenhove, "Encoding of event timing in the phase of neural oscillations," *NeuroImage,* vol. 92, pp. 274-284, 2014.