**Statistical Learning and Prosodic Bootstrapping Differentially Affect Neural Entrainment during Speech Segmentation**

Stefan Elmer§\*1,5, Seyed Abolfazl Valizadeh§1,2,3, Toni Cunillera4 & Antoni Rodriguez-Fornells5,6,7

1Auditory Research Group Zurich (ARGZ), Division Neuropsychology, Institute of Psychology, University of Zurich, 8050, Zurich, Switzerland

2Department of Internal Medicine, University Hospital, University of Zurich, 8091, Zurich, Switzerland

3University Research Priority Program, “Dynamics of Healthy Aging”, University of Zurich, , 8050, Zurich, Switzerland

4Department of Cognition, Development and Educational Psychology, 08035, Barcelona, University of Barcelona

5Department of Cognition, Development and Educational Psychology, Campus Bellvitge, University of Barcelona, 5L’Hospitalet de Llobregat, 08097, Barcelona, Spain

6Cognition and Brain Plasticity Group, Bellvitge Biomedical Research Institute, L’Hospitalet de Llobregat, 08097, Barcelona, Spain

7Institució Catalana de Recerca i Estudis Avançats, ICREA, 08010, Barcelona, Spain

§Shared first authorship

\*Corresponding author

Short title: Neural entrainment and word learning

Keywords: EEG, inter-trial coherence, event-related potentials, flat speech, prosody, word learning, frequency-tagging

Email addresses:

SE: s.elmer@psychologie.uzh.ch

SAV: seyed.valizadeh2@uzh.ch

TC: tcunillera@ub.edu

ARF: arfornells@gmail.com

Correspondence to:

Stefan Elmer, Department of Psychology, Division Neuropsychology, University of Zurich, Binzmühlestrasse 14/25, 8050 Zurich, Switzerland

Date: October 30, 2019

**ABSTRACT**

Neural oscillations constitute an intrinsic property of functional brain organization that facilitates the tracking of linguistic units at multiple time scales through brain-to-stimulus alignment. This ubiquitous neural principle has been shown to facilitate speech segmentation based on statistical regularities. However, it is unknown whether neural entrainment is likewise initiated when prosodic cues are available. Furthermore, there is no common agreement on whether speech segmentation is mediated by a transition of neural synchronization from syllabic rate to word frequency or whether the two time scales are concurrently tracked. Using inter-trial coherence (ITC) analyses and event-related potentials (ERP), we showed that statistical learning is mediated by parallel neural entrainment to words and syllables, with a close relationship between ITC and ERPs. However, when additional stress cues can be used to segment speech, neural entrainment at both time scales is disrupted. These results highlight distinct neural computations underlying statistical learning and prosodic bootstrapping.

**INTRODUCTION**

Speech is a hierarchically organized acoustic signal composed of linguistic units at different time scales, such as phonemes, syllables and words 1. However, unlike literary language, speech constitutes a continuous signal without reliable gaps between single words 2. Hence, one of the main challenges of learning new words is to segment speech, recognize word boundaries and extract word forms, especially when no lexicon is available for word recognition 3-6. Currently, at least two processes have been proposed to facilitate speech segmentation, namely statistical learning and prosodic bootstrapping 7-10. Statistical learning refers to the ability to extract statistical regularities from the speech signal, and relies on the fact that transitional probabilities between adjacent syllables are higher within words than at the word boundaries 10,11. Otherwise, prosodic bootstrapping consists of using prosodic cues like rhythm, intonation and lexical stress to infer speech structure and to detect word boundaries 9,12,13. Previous behavioral studies have demonstrated that both statistical learning 14 and prosodic cues 15,16 can be used to segment speech and extract word forms from continuous acoustic signals. It is noteworthy to mention that EEG 8,17,18 and MRI 18,19 studies on speech segmentation have also convincingly shown that statistical learning and prosodic bootstrapping rely on distinct neural circuits.

The similarity between speech and neural oscillations is that both signals fluctuate in a rhythmic fashion over time 20. Drawing on this compliance, it has been proposed that the temporal alignment of neural oscillations with the speech signal at multiple time scales constitutes a fundamental principle governing linguistic structure building, speech segmentation and word form recognition 1,21,22. In this context, it is important to emphasize that low-frequency oscillations in the delta and theta frequency range have repeatedly been shown to be crucial for tracking syllables and words 1,20,23. On this background, Buiatti and colleagues 24 used a frequency-tagging approach and computed fast Fourier transforms (FFT) to quantify mean neural synchronization to syllabic rate and word frequency while participants were exposed to either structured or random streams of flat speech. FFT analyses across multiple trials revealed neural alignment to the word frequency in the structured condition that positively correlated with the percentage of correctly recognized words. However, neural synchronization at the syllabic rate was only discernible in the random condition, suggesting that during word learning adjacent syllables are bound together to recognize single word units. In a more recent EEG study, Batterink and colleagues 17 compared inter-trial coherence (ITC) ratio of word frequency to syllable frequency between structured and random sequences of flat speech. Results showed that in structured streams ITC ratio was generally higher and increased across blocks. Furthermore, this effect was mainly driven by a linear increase in ITC at the word frequency, and accompanied by a decrease in ITC at the syllable frequency as a function of exposure. However, surprisingly, in both conditions ITC ratio predicted task performance.

Previous EEG studies on speech segmentation almost exclusively focused on statistical learning and evaluated neural entrainment across epochs consisting of multiple word units 17,24. Although these studies validated the suitability of frequency-tagging approaches to tackle the neural principles underlying speech segmentation and word learning, some fundamental questions have not yet been systematically addressed and clarified. In fact, it is currently unknown (1) whether neural synchronization and entrainment likewise operate if additional prosodic cues can be used to segment speech. There is also no common agreement on (2) whether speech segmentation and word form recognition are generally mediated by a neural transition from syllabic rate to word frequency, or whether the two time scales are concurrently tracked17,20,25. Furthermore, (3) notwithstanding that event-related potential (ERP) studies have shown that P200 and N400 responses 4,8,17,18,26 constitute valid indices of statistical learning and prosodic bootstrapping, it is unclear whether these two ERPs share a common neural basis with neural synchronization to the syllabic rate and to the word frequency. To address these questions, we used EEG and evaluated ITC at the word and syllable frequency in structured and random sequences of flat and stressed speech. In addition, we evaluated event-related potentials (ERPs) in time windows overlapping with the P200 and N400 components, and assessed the functional compliance between these two ERP manifestations and ITC at the syllabic rate and ITC at the word frequency.

**RESULTS**

**Behavioral data**

Separate one-sample t-tests against chance level (50%) yielded significance for both the flat (t(29) = 10.833, p < .001) and stressed conditions (t(19) = 4.927, p < .001). Accordingly, these behavioral results testify that the participants were able to segment speech and to learn the new words based on statistical learning and prosodic bootstrapping.

**Event-related potentials**

For the sake of completeness, and to provide comparability with previous EEG studies showing that P200 and N400 responses faithfully mimic speech segmentation based on statistical regularities and prosodic cues 4,8,17,18, we analyzed mean amplitudes of these two ERPs (Figure 1b and 1d) in the time window of 170-250 (P200) and 350-550 (N400) ms. With this purpose in mind, we computed separate 2 x 4 ANOVAs for the flat and stressed conditions, and compared structured and random sequences across the four blocks. Moreover, we used canonical correlation analyses to determine possible relationships between P200 and N400 manifestations, and ITC at the syllabic rate and ITC at the word frequency.

**C:\Users\Stefan Elmer\Desktop\ITC_word_learning\Analyses_2019_absolute_values\Manuskript_ITC\Figures\New_Figure_defdef.tif**

* Please insert Figure 1 at about here –

Figure 1: Grand average event-related potentials of the structured and random sequences of flat (b) and stressed (d) speech. In the flat condition (b) the red line depicts brain responses to structured sequences, whereas the orange line represents the random sequences. In the stressed condition (d) brain responses to structured sequences are shown in blue, whereas random sequences are represented in green. Figures 1a and 1c show ITC differences at the syllabic rate and at the word frequency between structured and random sequences of flat (a) and stressed (c) speech.

**P200 component**

The statistical analysis of the flat speech condition by means of a 2 x 4 ANOVA revealed a main effect of “block” (F(3, 87) = 27.165, p < .001), whereas the main effect of “sequence” (F(1,29) = .058, p = .811) and the “sequence x block” interaction (F(3, 87) = .638, p = .593) did not reach significance.Post-hoc t-tests (one-tailed, Bonferroni-corrected p value for 6 tests, p < .008) indicated an overall increase in mean P200 amplitudes across the four blocks, irrespective of sequence type (Figure 1b and Table 1).

The evaluation of the stressed condition yielded main effects of “sequence” (F(1, 22) = 126.774, p < .001) and “block” (F(3, 66) = 5.747, p = .001). The “sequence x block” interaction did not reach significance (F(3, 66) = 2.458, p = .071). As visible in Figure 1d, the main effect of “sequence” was related to increased P200 amplitudes in structured compared to the random streams of stressed speech. Furthermore, post-hoc t-test (one-tailed, Bonferroni-corrected p value for 6 tests, p < .008) revealed that the main effect of “block” was associated with decreased P200 amplitudes from block 1 to block 3 and 4, regardless of sequence (Figure 1d and Table 1).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Condition | Sequences | Contrast | Degrees of freedom | t-value | p-value |
| Flat | All sequences | Block 1 vs. Block 2 | 29 | -2.570 | 0.008\* |
|  |  | Block 1 vs. Block 3 | 29 | -6.055 | < 0.001\* |
|  |  | Block 1 vs. Block 4 | 29 | -6.829 | < 0.001\* |
|  |  | Block 2 vs. Block 3 | 29 | -3.261 | 0.0015\* |
|  |  | Block 2 vs. Block 4 | 29 | -5.496 | < 0.001\* |
|  |  | Block 3 vs. Block 4 | 29 | -3.524 | < 0.001\* |
| Stressed | All sequences | Block 1 vs. Block 2 | 22 | 1.688 | 0.053 |
|  |  | Block 1 vs. Block 3 | 22 | 3.666 | < 0.001\* |
|  |  | Block 1 vs. Block 4 | 22 | 2.821 | 0.005\* |
|  |  | Block 2 vs. Block 3 | 22 | 2.442 | 0.011 |
|  |  | Block 2 vs. Block 4 | 22 | 1.501 | 0.074 |
|  |  | Block 3 vs. Block 4 | 22 | -0.434 | 0.334 |

Table 1: Post-hoc comparisons of the significant main effects of “block” in the omnibus ANOVAs for P200 amplitudes. \* Depicts significance after correction for multiple comparisons.

* Please insert Table 1 at about here -

**N400 component**

The statistical analysis of the flat condition yielded main effects of “sequence” (F(1, 29) = 52.268, p < .001) and “block” (F(3, 87) = 15.010, p < .001) as well as a significant quadratic “sequence x block” interaction (F(1, 29) = 4.594, p = .041). As visible in Figure 1b, the quadratic interaction between “sequence” and “block” originated from a U-shaped devolution of the N400 component over time in structured sequences of flat speech, whereas the main effect of “sequence” was related to larger N400 responses in structured compared to random streams (Figure 1b). Post-hoc t-tests (one-tailed, Bonferroni-corrected p value for 6 tests, p < .008) computed to disentangle the “sequence x block” interaction showed a significant reduction in N400 amplitudes from block 2 to 4 (t(29) = -4.595, p < .001) and from block 3 to 4 (t(29) = -3.819, p < .001). Furthermore, post-hoc t-tests (one-tailed, Bonferroni-corrected p value for 6 tests, p < .008) revealed that the main effect of “block” originated from overall reduced N400 amplitudes in block 4 compared to the first three blocks (block 1\_4: t(29) = -5.277, p < .001; block 2\_4: t(29) = -5.493, p < .001; block 3\_4: t(29) = -5.664, p < .001, Figure 1b), irrespective of sequence type. All other comparisons did not reach significance (Table 2).

The analysis of the stressed condition yielded a main effect of “block” (F(3, 66) = 3.543, p = .019), whereas the main effect of “sequence” (F(1, 22) = 1.007, p = .326) and the “sequence x block” interaction (F(3, 66) = 1.313, p = .278) did not reach significance. According to post-hoc t-tests (one-tailed, Bonferroni-corrected p value for 6 tests, p < .008), the main effect of “block” was associated with increased N400 responses in block 2 compared to block 1, regardless of sequences (block 1\_2: t(22) = 2.675, p = .007, Figure 1d). All others comparisons did not reach significance (Table 2).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Condition | Sequences | Contrast | Degrees of freedom | t-value | p-value |
| Flat | Structured | Block 1 vs. Block 2 | 29 | 2.318 | 0.014 |
|  |  | Block 1 vs. Block 3 | 29 | 1.254 | 0.110 |
|  |  | Block 1 vs. Block 4 | 29 | -2.540 | 0.009 |
|  |  | Block 2 vs. Block 3 | 29 | -1.537 | 0.068 |
|  |  | Block 2 vs. Block 4 | 29 | -4.595 | < 0.001\* |
|  |  | Block 3 vs. Block 4 | 29 | -3.819 | < 0.001\* |
| Flat | All sequences | Block 1 vs. Block 2 | 29 | 1.260 | 0.109 |
|  |  | Block 1 vs. Block 3 | 29 | 0.333 | 0.370 |
|  |  | Block 1 vs. Block 4 | 29 | -5.277 | < 0.001\* |
|  |  | Block 2 vs. Block 3 | 29 | -1.072 | 0.146 |
|  |  | Block 2 vs. Block 4 | 29 | -5.493 | < 0.001\* |
|  |  | Block 3 vs. Block 4 | 29 | -5.664 | < 0.001\* |
| Stressed | All sequences | Block 1 vs. Block 2 | 22 | 2.675 | 0.007\* |
|  |  | Block 1 vs. Block 3 | 22 | 2.512 | 0.010 |
|  |  | Block 1 vs. Block 4 | 22 | 1.736 | 0.048 |
|  |  | Block 2 vs. Block 3 | 22 | -0.555 | 0.292 |
|  |  | Block 2 vs. Block 4 | 22 | -1.296 | 0.104 |
|  |  | Block 3 vs. Block 4 | 22 | -0.690 | 0.249 |

Table 2: Post-hoc comparisons of the significant main effects of “block” in the omnibus ANOVAs for N400 amplitudes. \* Depicts significance after correction for multiple comparisons.

* Please insert Table 2 at about here -

**General neural synchronization to pertinent speech units**

In line with previous work 4,17, ITC analyses yielded clear maxima at the word and syllable frequency (Figure 2a and 2b). To first testify the overall neural alignment to pertinent speech units, we performed separate univariate ANOVAs for the flat and stressed conditions with the within-subject factor “sequence” (structured and random). In this context, we separately evaluated ITC at the word frequency, word learning index (WLI), ITC at the syllabic rate, and syllable learning index (SLI, see methods). Analyses of the flat condition (Figure 3) revealed that ITC at the word frequency (F(1, 29) = 39.84, p < .001) and the WLI (F(1, 29) = 7.92, p = .009) were increased in structured compared to random sequences. In contrast, the SLI was higher in random compared to the structured sequences (F(1, 29) = 5.01, p = .033), whereas ITC analyses at the syllabic rate did not reveal significant differences between structured and random streams (F(1, 29) = 3.43, p = .074). These results suggest that statistical learning is mediated by neural synchronization to word units, whereas word frequency and syllabic rate are possibly concurrently tracked. The latter assumption is consistent with the significant correlation we revealed between ITC at the syllabic rate and ITC at the word frequency (Pearson’s r, one-tailed, r = .315, p = .045, Figure 4a) in structured sequences of flat speech. Furthermore, based on the ITC and ERP results, we correlated mean ITC at the word frequency and the WLI with mean N400 amplitudes. Correlation analyses (Pearson’s r, one-tailed, Bonferroni-corrected p-value for 2 tests, p < .025) yielded a significant negative relationship between ITC at the word frequency and mean N400 amplitudes (r = -.604, p < .001, Figure 4b), whereas the correlation between WLI and mean N400 responses did not reach significance (r = -.050, p = .396).

Interestingly, the statistical analyses of the stressed condition (Figure 3) did not reveal significant differences between structured and random sequences with respect to ITC at the word frequency (F(1, 22) = 2.32, p = .142), WLI (F(1, 22) = 1.42, p = .246), ITC at the syllabic rate (F(1, 22) = .332, p = .57) or SLI (F(1, 22) = 1.065, p = .313). However, since the structured and random conditions elicited clear ITC peaks corresponding to the syllabic rate and the word frequency (Figure 2b) but did not differ, results suggest that (1) neural alignment to syllabic units constitutes an ubiquitous neural principle underlying speech processing (see also the results of the flat condition), whereas (2) neural synchronization to the word frequency might be an epiphenomenon of prosodic cues rather than genuinely reflecting speech segmentation and word form extraction.

C:\Users\Stefan Elmer\Desktop\ITC_word_learning\Analyses_2019_absolute_values\Manuskript_ITC\Figures\Figure_def_frequency_tagging.tif

* Please insert Figure 2 at about here -

Figure 2. a = mean ITC values in the flat condition for structured (red) and random (orange) sequences. b = mean ITC values in the stressed condition for structured (blue) and random (green) sequences. c = ITC values across the four blocks of the flat condition for structured (red) and random (orange) sequences. d = ITC values across the four blocks of the stressed condition for structured (blue) and random (green) sequences.

C:\Users\Stefan Elmer\Desktop\ITC_word_learning\Analyses_2019_absolute_values\Manuskript_ITC\Figures\Figure_1_def.tif

* Please insert Figure 3 at about here –

Figure 3. Single-subject data and violin plots with density distribution and mean for the flat (left) and stressed (right) conditions and structured (red and blue) and random (orange and green) sequences. a = ITC at the word frequency, b = ITC at the syllable frequency, c = word learning index (WLI), d = syllable learning index (SLI). \* = p < .05, \*\* = p < .01, \*\*\* = p < .001.

C:\Users\Stefan Elmer\Desktop\ITC_word_learning\Analyses_2019_absolute_values\Manuskript_ITC\Figures\Figure_3_defdef.tif

* Please insert Figure 4 at about here -

Figure 4. Correlation analyses for structured sequences of flat speech. a = Correlation between mean ITC at the syllable frequency and word frequency. b = Correlation between mean ITC at the word frequency and mean N400 amplitude. c, d, e, f = Correlations between mean ITC at the word frequency and mean N400 amplitude in the first (c), second (d), third (e) and fourth (f) block. n.s. = not significant after correction for multiple comparisons.

**Neural entrainment as a function of exposure across blocks**

Neural entrainment was quantified by analyzing ITC values at the word and syllable frequency across the four blocks. Accordingly, we performed separate univariate ANOVAs, with the within-subject factor “blocks” for structured and random sequences of flat and stressed speech. Analyses of the structured sequences of flat speech (Figure 5) revealed main effects of “block” for both ITC at the word (F(3, 87) = 4.656, p = .005) and syllable frequency (F(3, 87) = 6.795, p < .001), whereas in the random condition statistical analyses yielded significance only for ITC at the syllable frequency (syllable frequency: F(3, 87) = 16.958, p < .001; word frequency: F(3, 87) = 2.156, p = .099).

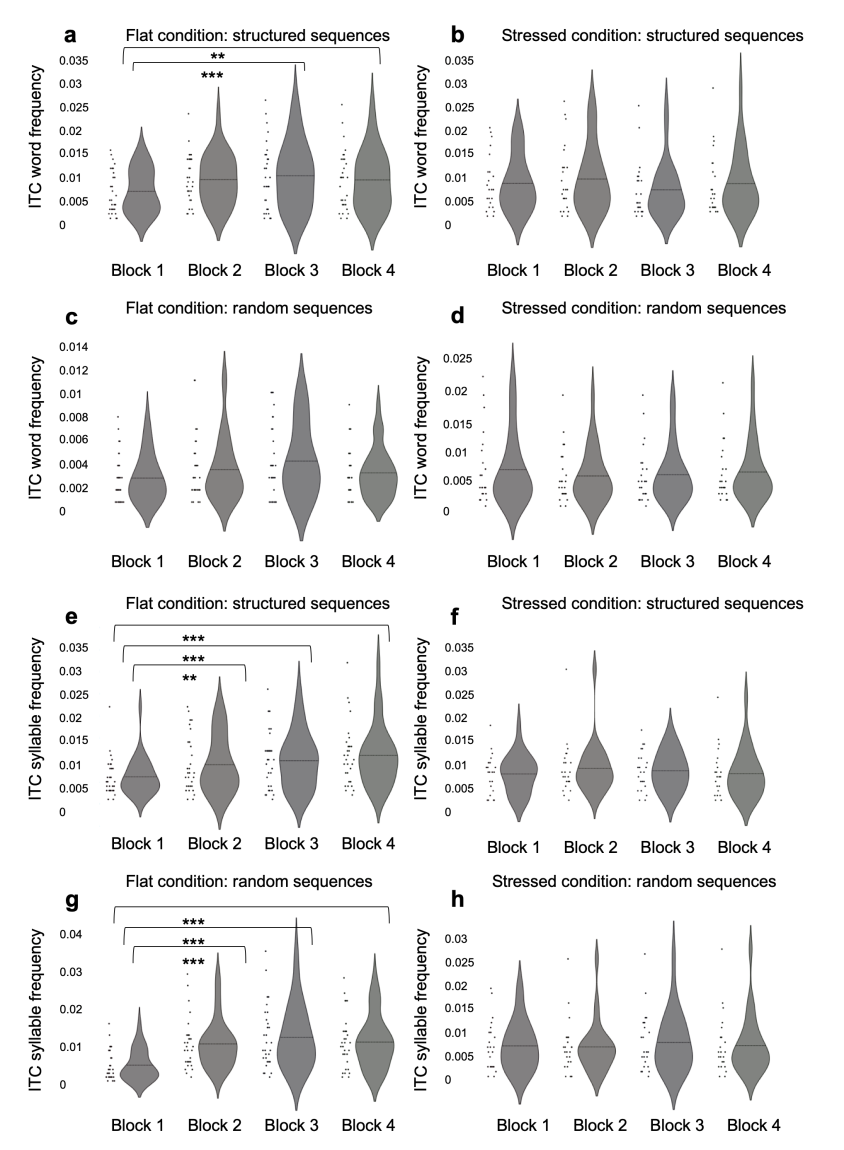
Post-hoc comparisons (one-tailed, Bonferroni-corrected p value for 6 tests, p < .008) of the structured sequences of flat speech revealed neural entrainment to the word frequency from the first to the third and fourth blocks (block 1\_3: t(29) = -3.416, p = .001; block 1\_4: t(29) = -2.754, p = .005). Furthermore, neural entrainment to the syllabic rate increased from the first to the second, third and fourth block (block 1\_2: t(29) = -2.654, p = .006; block 1\_3: t(29) = -3.312, p = .001; block 1\_4: t(29) = -3.62, p < .001). In a similar way, post-hoc analyses of the random sequences of flat speech revealed neural entrainment to the syllabic rate in the second, third and fourth block compared to the first one (block 1\_2: t(29) = -6.113, p < .001; block 1\_3: t(29) = -5.823, p < .001; block 1\_4: t(29) = -5.177, p < .001). All other comparisons did not reach significance (Table 3). Taken together, these results suggest that during flat speech neural oscillations generally entrain to the syllabic rate, whereas speech segmentation based on statistical regularities is mediated by neural entrainment to the word frequency.

Based on the parallel dynamic ITC (word frequency) and N400 changes we observed across the blocks of structured sequences of flat speech, we performed additional correlation analyses and assessed possible relationships between these two electrophysiological parameters separately for each block. Correlation analyses (Pearson’s r, one-tailed, Bonferroni-corrected p value for 4 tests, p < .012, Figure 4c-f) consistently yielded significant negative relationships between ITC at the word frequency and N400 amplitudes in the first (r = -.612, p < .001), second (r = -.634, p < .001) and fourth (r = -.482, p = .003) but not in the third block (r = -.359, p = .026). In a similar way, drawing on the main effects of “block” we revealed for ITC at the syllable frequency and P200 responses in both structured and random sequences of flat speech, we correlated these neural markers across the four blocks (Pearson’s r, one-tailed, Bonferroni-corrected p value for 4 tests, p < .012). In structured sequences of flat speech, correlation analyses revealed a significant positive relationship between ITC at the syllabic rate and mean P200 amplitude in the fourth block (block 4: r = .677, p < .001; block 1: r = .232, p = .109; block 2: r = .266, p = .078; block 3: r = .075, p = .347, Figure 6a). Analyses of the random sequences of flat speech highlighted consistent positive relationships between ITC at the syllable frequency and mean P200 amplitude in the second, third and fourth block (block 2: r = .522, p = .002; block 3: r = .652, p < .001; block 4: r = .534, p = .001; block 1: r = .034, p = .429, Figure 6b-d).

Statistical analyses of the stressed condition (Figure 5) did not reveal significant effects of “block” neither for structured (word frequency: F(3, 66) = 1.144, p = .338; syllable frequency: F(3, 66) = 1.266, p = .293) nor for random (word frequency: F(3, 66) = .298, p = .827; syllable frequency: F(3, 66) = .512, p = .675) sequences. Accordingly, these results suggest that stress cues disrupt neural entrainment to pertinent speech units.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Frequency | Condition | Contrast | Degrees of freedom | t-value | p-value |
| Word | Structured flat speech | Block 1 vs. Block 2 | 29 | -2.440 | 0.010 |
|  |  | Block 1 vs. Block 3 | 29 | -3.416 | 0.001\* |
|  |  | Block 1 vs. Block 4 | 29 | -2.754 | 0.005\* |
|  |  | Block 2 vs. Block 3 | 29 | -0.748 | 0.230 |
|  |  | Block 2 vs. Block 4 | 29 | 0.189 | 0.425 |
|  |  | Block 3 vs. Block 4 | 29 | 1.003 | 0.162 |
| Syllable | Structured flat speech | Block 1 vs. Block 2 | 29 | -2.654 | 0.006\* |
|  |  | Block 1 vs. Block 3 | 29 | -3.312 | 0.001\* |
|  |  | Block 1 vs. Block 4 | 29 | -3.620 | < 0.001\* |
|  |  | Block 2 vs. Block 3 | 29 | -0.965 | 0.172 |
|  |  | Block 2 vs. Block 4 | 29 | -1.625 | 0.058 |
|  |  | Block 3 vs. Block 4 | 29 | -1.290 | 0.104 |
| Syllable | Random flat speech | Block 1 vs. Block 2 | 29 | -6.113 | < 0.001\* |
|  |  | Block 1 vs. Block 3 | 29 | -5.823 | < 0.001\* |
|  |  | Block 1 vs. Block 4 | 29 | -5.177 | < 0.001\* |
|  |  | Block 2 vs. Block 3 | 29 | -1.749 | 0.045 |
|  |  | Block 2 vs. Block 4 | 29 | -0.493 | 0.313 |
|  |  | Block 3 vs. Block 4 | 29 | 1.206 | 0.119 |

Table 3: Post-hoc comparisons of the significant main effects of “block” in the omnibus ANOVAs for ITC. \* Depicts significance after correction for multiple comparisons.



* Please insert Figure 5 at about here -

Figure 5. Single-subject data and violin plots with density distribution and mean for the four blocks. ITC at the word frequency (a, b, c, d) and ITC at the syllable frequency (e, f, g, h) are shown for the flat (a, c, e, g) and stressed conditions (b, d, f, h) as well as for structured (a, b, e, f) and random sequences (c, d, g, h). \*\* = p < .01, \*\*\* = p < .001.

**C:\Users\Stefan Elmer\Desktop\ITC_word_learning\Analyses_2019_absolute_values\Manuskript_ITC\New_Figures_tables_with_ERP\Figure_6.tif**

* Please insert Figure 6 at about here –

Figure 6: Significant correlations between mean P200 amplitudes and ITC at the syllable frequency in the flat condition. a = block 4 of structured sequences, b = block 2 of random sequences, c = block 3 of random sequences, d = block 4 of random sequences.

**DISCUSSION**

In the present EEG study, we examined the neural computations governing speech segmentation based on statistical learning and prosodic bootstrapping while participants learned new words embedded in continuous speech streams. To maximize statistical power, in a first ITC analysis we compared overall neural synchronization to word frequency and syllabic rate between structured and random sequences. This approach aimed at testing whether neural synchronization to words and syllables likewise operates under statistical learning and prosodic bootstrapping conditions. However, unlike previous studies 17,24, we focused on the neural alignment between single words instead of evaluating ITC across multiple adjacent units. In a second ITC analysis, we then analyzed neural entrainment across blocks and examined whether speech segmentation and word learning are generally mediated by a neural transition from syllabic rate to word frequency, or whether the two time scales are concurrently tracked. Furthermore, for reasons of completeness and for the sake of comparability with previous studies 4,8,17,18, we also evaluated mean P200 and N400 amplitudes as well as possible relationships between these two ERPs and ITC at the syllabic rate and ITC at the word frequency.

**General neural synchronization to pertinent speech units**

Results of the flat speech condition confirmed previous findings showing increased neural synchronization to the word frequency in structured compared to random sequences 4,17,24. Such an alignment of neural oscillations to words may constitute the neural basis of speech segmentation based on statistical regularities. However, in contrast to Buiatti and colleagues 24, we also noticed clear ITC peaks at the syllabic rate in both structured and random sequences (Figure 2) and ITC at the syllable frequency did not differ between the two streams. Therefore, our results do not support the assumption that speech segmentation and word form recognition are mediated by neural suppression of low-level speech features, but rather suggest that pertinent speech units are concurrently tracked. This argument has previously already been proposed by other authors 20,27 in the context of sentence processing, and would also explain the significant positive relationship we revealed between ITC at the syllabic rate and ITC at the word frequency (Figure 4a). Furthermore, previous work has shown that speech segmentation is facilitated when multiple cues are available in one or even more sensory modalities 28,29. This perspective is also grounded in the parallel neural entrainment to syllabic rate and word frequency we revealed across blocks in the structured sequences of flat speech. Furthermore, based on the fact that in flat speech word forms can only be inferred by taking into account transitional probabilities between adjacent syllables, our results suggest that the concurrent tracking of syllables and words constitutes an important foundation of word learning. Interestingly, we also found negative correlations between ITC at the word frequency and mean N400 amplitude (Figure 4b-f). Since the N400 component oscillates in the range of 1-2 Hz 8 which roughly corresponds to the word frequency (1.43 Hz), we may speculate that phase synchronization across multiple words constitutes a computational mechanism underlying N400 generation.

Surprisingly, in the stressed condition we did not reveal significant differences between structured and random sequences, neither in terms of neural synchronization to the syllabic rate nor in terms of alignment to the word frequency. However, this does not mean that there is no neural alignment to syllables and words. In fact, as visible in Figure 2, both structured and random sequences were associated with two clear peaks corresponding to these two time scales. In this context, it is important to mention that the main difference between structured sequences of flat and stressed speech is that in the latter case statistical learning and prosodic bootstrapping interact. This implies that in the stressed condition both transitional probabilities between adjacent syllables and prosodic cues can be used to recognize word boundaries. Nevertheless, since in the stressed condition ITC at the word frequency did not differ between structured and random sequences, these results might suggest that neural synchronization at the word frequency was an epiphenomenon of stress cues rather than reflecting true neural alignment to word units. In particular, it is conceivable that the acoustic salience of stressed syllables may have captured the attention of the participants, making neural entrainment to words superfluous. Accordingly, in the presence of stress cues brain-to-stimulus synchronization at the word level may not be initiated or blocked, possibly because speech segmentation is directly mediated by specific subregions of the auditory cortex involved in pitch processing 30-33 and auditory attention 34,35. Such an argument would be consistent with a previous ERP study 8 using exactly the same paradigm, and showing that the comparison between structured and random sequences of stressed speech elicited a P200 component mainly originating from the primary and secondary auditory cortices. Finally, it is important to mention that the comparable ITC peaks we noticed at the syllable frequency in structured and random sequences of stressed speech are in line with the results of the flat conditions, and substantiate the conclusion that neural alignment to syllabic units constitutes an intrinsic neural principle underlying speech processing 20,36,37.

**Neural entrainment as a function of exposure across blocks**

Separate ITC analyses of the structured and random sequences across the four blocks only revealed neural entrainment in the flat speech condition. In particular, in structured streams ITC at the word frequency and at the syllabic rate conjointly increased across blocks, whereas in random sequences ITC only increased at the syllable frequency. In contrast, in the stressed condition ITC at the syllabic rate and at the word frequency did not linearly increase across the four blocks neither in structured nor in random sequences. These results suggest that speech segmentation based on statistical learning is mediated by neural entrainment to word units. However, when additional prosodic cues can be used to segment speech and recognize word boundaries, neural entrainment to words and syllables is not initiated.

The neural entrainment to word units we revealed in the third and fourth blocks of structured sequences of flat speech is in line with previous results of Batterink and colleagues 17 showing a linear increase in ITC across three blocks of approximately four minutes each. Recently, also Henin and co-workers 25 who collected intracranial recordings in 23 patients found that phase coherence at the word frequency emerged after only about four minutes of exposure. Furthermore, in accordance with previous studies showing a contribution of the left dorsal stream to speech segmentation based on statistical regularities 18,19, Henin and colleagues demonstrated that neural entrainment to word units was particularly pronounced in two cluster of electrodes located over the supratemporal plane and the inferior frontal gyrus 25. Finally, the rapid brain-to-stimulus alignment we observed at the word frequency is also compatible with a previous ERP study of Cunillera and colleagues 18 who compared structured and random sequences and found N400 manifestations as early as in the second block.

In contrast to the results of Buiatti and colleagues 24 as well as of Batterink and co-workers 17, our data did not support the idea of a neural suppression or a linear decrease in ITC at the syllabic rate in structured sequences of flat speech. In fact, we rather observed a parallel increase in ITC at the word and syllable frequency, leading to suggest that during statistical learning these two time scales are concurrently tracked. Such a parallel alignment to basic speech elements and learned higher-order word units would be in agreement with the assumption that single words can only be recognized based on transitional probabilities between adjacent syllables 10,11 or chunk of syllables 38,39. However, this argument is weakened by the fact that neural entrainment to the syllable frequency also appeared in random sequences of flat speech. Therefore, results are interpreted as evidence that neural entrainment to the syllabic rate constitutes an intrinsic and automatic mechanism involved in parsing the speech signal into units of appropriate granularity rather than reflecting a genuine neural principle of statistical learning 20.

Finally, and most importantly, no previous studies have examined neural entrainment mechanisms at the intersection between statistical learning and prosodic bootstrapping in comparison to a pure statistical learning condition. Therefore, our study provides important insights into the influence of stress cues on ITC dynamics. In this context, it is noteworthy to mention that we did not find any evidence for neural entrainment to the word frequency in structured sequences of stressed speech. Therefore, our data clearly indicate that when prosodic cues can be used to recognize word boundaries, neural entrainment becomes unnecessary for segmenting speech. However, the most surprising aspect was that stress cues on every third syllable also disrupted neural entrainment to syllables in both structured and random sequences. Since this was not the case in the flat speech condition, results emphasize that stress cues generally disrupt neural entrainment to pertinent speech units, possibly because pitch is processed at a lower hierarchical level and dependent upon the functional integrity of the auditory cortex 40-42. This would be in line with previous studies showing that prosodic bootstrapping elicited a P200 component originating from the auditory cortex, whereas statistical learning is mediated by the left dorsal stream 8,18,19. In any case, future studies combining high temporal and spatial resolution are needed to more carefully determine the neural substrate underlying neural entrainment at multiple time scales in statistical learning and prosodic bootstrapping conditions. In conclusion,ourresults demonstrate that speech segmentation based on statistical regularities relies on the concurrent tracking of words and syllabic units, whereas the presence of additional prosodic cues disrupts neural entrainment at both time scales. Accordingly, statistical learning and prosodic bootstrapping rely on distinct neural circuits that might possibly be located in the auditory cortex and the dorsal stream.

**P200 component**

For reasons of consistency and comparability with other studies, ERP analyses were restricted to two specific components that have been shown to be sensitive to speech segmentation based on statistical regularities and prosodic cues 4,8,17. With this purpose in mind, we focused on the P200 and N400 components, and separately evaluated mean amplitudes across the four blocks of the flat and stressed conditions. In line with previous work 8, structured sequences of stressed speech elicited larger P200 responses than random sequences, whereas a comparable P200 modulation was not observed in the flat condition. However, P200 analyses also documented new findings that might be particularly interesting for the integration of ERPs and ITCs into a common electrophysiological framework of speech segmentation that will be introduced afterwards. In particular, we noticed that P200 amplitudes conjointly increased with ITC at the syllabic rate over time in structured and random sequences of flat speech. In contrast, in the stressed condition this was not the case and P200 amplitudes generally decreased in the last two blocks compared to the first one.

In the context of an EEG study carried out on a subgroup of participants also included in the present work, Cunillera and colleagues 8 identified the P200 component as a distinctive electrophysiological marker of speech segmentation in structured stressed sequences. Drawing on the fact that the main cortical sources underlying this specific ERP were estimated in the bilateral auditory cortex 8, the generally increased P200 responses we revealed in structured compared to random sequences of stressed speech support the conjecture that statistical information and prosodic cues interact at early stages of auditory processing. In particular, since in structured sequences syllables with a higher pitch can be predicted based on statistical information, our results substantiate the assumption that in the presence of statistical regularities prosodic cues grab attention more effectively than in random sequences. This perspective is also compatible with previous EEG studies showing that the P200 component is modulated by auditory attention 43,44 and sensitive to pitch 45,46 and prosody 47,48. Nevertheless, it is noteworthy to mention that in the flat condition P200 responses and ITC at the syllabic rate conjointly increased across the four blocks (Figure 1 and 2), and ITC at the syllable frequency correlated with mean P200 amplitudes (Figure 6) in both structured (block 4) and random sequences (block 2, 3 and 4). Such a compliance of ITCs and ERPs is particularly interesting for two reasons. First, because it introduces the idea that both EEG parameters are anchored on a common neural mechanism, and second because it suggests that the tracking of syllabic units over time is related to P200 manifestations. However, since in the stressed condition a similar progression was not visible and P200 amplitudes generally decreased in the last two blocks compared to the first one, we might speculate that the presence of additional prosodic cues induced neural adaptation in the auditory cortex as reflected by lower P200 amplitudes 49,50.

**N400 component**

The results of the flat condition replicated previous EEG findings showing increased N400 amplitudes in structured compared to random sequences 4,8,17,18. Furthermore, as previously already reported by Cunillera and colleagues 18, the N400 component was characterized by a U-shaped response pattern with smallest amplitudes in the last block. The fast emergence and configuration of the N400 component in the first three blocks of structured sequences is interpreted as a marker of speech segmentation reflecting the codification and strengthening of episodic memory traces for linguistic representations or novel words 4,8,17,18. This argument is reinforced by the significant correlations we observed between mean N400 amplitudes and ITC at the word frequency (Figure 4). Otherwise, the intrinsic meaning of decreased N400 amplitudes in the fourth block is somewhat unclear. In fact, reduced N400 amplitudes have previously been associated with an optimized access to verbal memory as a function of learning 51,52. Nevertheless, based on our ERP and ITC data it is necessary to envisage an alternative interpretation. Notably, as visible in Figures 2c, in structured sequences neural alignment to the syllable frequency significantly increased after the first block, and in the fourth block it reached almost the same level as that of random sequences. Furthermore, by subtracting the neural alignment to the syllable frequency in random sequences from structured sequences (Figure 1a), we obtained a U-shaped function that roughly coincides with the time course of the N400 component (Figure 1b). This observation is be rooted in the notion that the reduced N400 amplitude we revealed in the fourth block might be somehow related to increased syllabic tracking. This argument is not only supported by the positive correlation we revealed between ITC at the syllable frequency and P200 amplitudes in the fourth block of structured sentences, but also by the steady increase of both parameters over time. A possible explanation might be that learning based on statistical principles follows a logarithmic function 53, where the increment of learning is the strongest at the beginning and saturates with deliberate practice. Hence, we speculate that a change in attentional focus from words to syllables in the last block might constitute a strategy to further improve learning. Finally, the statistical results of the stressed condition brought to light a main effect of “block” that originated from increased N400 amplitudes in block 2 compared to block 1. However, since this effect was independent of sequence type and the N400 did not further increase in the successive blocks, this result might be attributable to an increased recruitment of verbal memory functions which are necessary to analyze statistical regularities and memorize pseudowords.

**An electrophysiological framework of speech segmentation**

Based on the results of this study, we propose an electrophysiological framework of speech segmentation and word learning that integrates ERPs and neural oscillations. The starting point of our framework relies on the following main results of the flat (a) and stressed (b) conditions:

Flat condition: (1a) ITC at the word frequency and N400 responses were larger in structured compared to random sequences, and both parameters showed a dynamic pattern across the blocks. (2a) ITC at the syllable frequency and P200 responses did not differ between structured and random sequences but generally increased across the blocks. (3a) In structured sequences ITC at the word frequency correlated with N400 responses, whereas in random sequences ITC at the syllable frequency correlated with P200 responses. (4a) In structured sequences ITC at the syllabic rate was positively related to ITC at the word frequency.

Stressed condition: (1b) ITC at the word frequency and N400 responses did not differ between structured and random sequences, and ITC at the word frequency did not increase over time. However, N400 responses were generally enlarged in block 2 compared to block 1, irrespective of sequence type. (2b) ITC at the syllable frequency did not differ between structured and random sequences and did not change across the blocks. However, P200 responses were larger in structured compared to random sequences and generally decreased from block 1 to block 3 and 4, irrespective of sequence type.

Taking into account the main results of the flat condition, we deduced that ITC at the word frequency and N400 responses were sensitive to speech segmentation and word form extraction, whereas ITC at the syllabic rate and the P200 component reflected the tracking of syllabic units. Furthermore, the relationships we observed between neural oscillations and ERPs may indicate that ITC at the word frequency and N400 responses, as well as ITC at the syllabic rate and P200 responses, share a common neural origin. From a dynamic perspective, we argue that in the absence of prosodic cues speech segmentation is mediated by concurrent tracking of syllabic units and word forms over time. The parallel neural entrainment at these two time scales, in association with the correlation we revealed between ITC at the syllabic rate and ITC at the word frequency, is compatible with the assumption that word boundaries are most likely recognized by taking into account statistical regularities between adjacent syllables. Such a nesting relation between ITC at the syllabic rate and ITC at the word frequency might also explain the U-shaped N400 learning curve that is often observed when participants have to segment structured sequences of flat speech 8,18. In fact, the U-shaped time course of the N400 component with reduced amplitudes in the fourth block showed a striking similarity to the difference wave of ITC at the syllable frequency between structured and random sequences (Figures 1a and 1b). This similarity, together with the positive correlation we noticed between ITC at the syllabic rate and P200 amplitudes in the last block of structured sequences of flat speech, leads to suggest that a stronger attentional orientation to syllabic units might have contributed to further promote learning achievement. This perspective is rooted in the general principle that experiential learning follows a logarithmic function 53, and that the learning effect becomes smaller as a function of experience.

The additional examination of a stressed speech condition allowed us a more accurate specification of the neural principles governing speech segmentation based on the interaction between statistical learning and prosodic bootstrapping. Notably, in the stressed condition we did not reveal differences between structured and random sequences in terms of N400 responses, ITC at the word frequency or ITC at the syllabic rate. Moreover, these electrophysiological indices did not systematically increase across the blocks. Consequently, based on these results, we deduced that in the presence of prosodic cues syllables and words are not concurrently tracked and neural entrainment to pertinent speech units is not initiated or even blocked. Since structured sequences of stressed speech were generally associated with larger P200 responses originating from the auditory cortex 8 compared to random sequences, we propose that pitch and predictive coding interact at early stages of auditory processing and facilitate speech segmentation, possibly through the allocation of attentional resources. Such a perspective would not only be in agreement with a previous EEG study that used exactly the same paradigm 8 but also with data showing that attention is tightly coupled to predictions 54 and that predictive coding and pitch processing interact in the auditory cortex 55. Anyway, this proposition could be directly tested by comparing structured and random sequences of stressed speech under both active and passive listening conditions. Finally, it is noteworthy to emphasize that a contribution of the auditory cortex to pitch processing and the segmentation of stressed speech would also go hand in hand with the sequence-unspecific neural adaptation we determined as a function of exposure (P200). Certainly, the involvement of the auditory cortex in speech segmentation does not preclude verbal memory functions, and this would be compatible with the overall increased N400 responses we found in block 2 compared to block 1, regardless of sequence type.

**METHODS**

**Participants**

Thirty students of the University of Barcelona took part in the flat speech condition (age range = 19-44 years, mean age = 23.69, SD = 5.68, 17 females), and 23 of them were re-invited to perform the stressed condition (age range = 19-38 years, mean age = 23.18, SD = 4.58, 15 females). All participants were right-handed native Spanish-Catalan speakers with normal hearing and no neurological impairments. The experiment was approved by the local ethical committee of the University of Barcelona, and the participants were paid for their participation.

**Materials and procedure**

The experimental design of this study was the same as the one previously used by Cunillera and colleagues 8. Therefore, in the following paragraph we literally reiterate the description of the stimulus material used in the previous study. “Five words streams (languages) were created for each of the stressed and flat conditions. The word streams had the same structure as the ones created by Saffran et al. 11. Each stream consisted of 4 different trisyllabic nonsense words (pseudowords). Each word was repeated 192 times, resulting in a total of 3840 items per condition (5 languages x 4 words/non-words x 192 repetitions). Words were concatenated to form a text stream which was then transformed into an acoustic stream using the speech synthesizer MBROLA which is based on concatenation of diphones 56. The Cooledit software was used to equate the length of the different streams with millisecond precision, and the exact duration of each stream was 8 min 54 s and 528 ms. Crucially, the streams were constructed without acoustic pauses between single items. Since only 59 syllables can be used for the construction of the five streams, only one syllable was repeated in 2 streams. In all streams the transitional probability across syllables forming a word was 1.0, while syllables spanning word boundaries had a transitional probability of 0.33. The same pool of syllables was used for the construction of the languages in both the stressed and flat conditions, but syllables were concatenated in a different order. The resulting word streams in the flat condition contained no pauses or other acoustic cues indicating word onset. In contrast, in the stressed condition, each stream contained an acoustic indication of word onset. The pitch of the first syllable of each word in the stream was increased by 20 Hz, creating an artificial stress at the beginning of each possible word 15. Although stressed syllables are also characterized by an increase in length, we maintained the duration stable across syllables within a word in order to avoid segmentation based on syllable lengths rather than on pitch. The fact that all syllables across streams were matched in length permits a direct comparison between conditions. Furthermore, 10 different streams (five for each condition, stressed and flat) were created as a baseline (random streams) using the same syllables presented in each word stream but concatenated in random order. Accordingly, each syllable in the stream could be followed by any of the other eleven syllables composing the stream. Thus, the transitional probability across syllables was 0.09. The low transitional probability should create a condition where the extraction of a clearly segmented word should be impossible, as the probability could not be used to identify possible words. In the flat condition, random streams contained no acoustic information. In contrast, in the stressed condition, the first syllable of each syllable triplet was stressed regardless of which syllable fell in the stressed positions. In order to test segmentation of the streams, test items were created. For the stressed and flat conditions items consisted of the four words forming each stream, plus part-words that consisted of the concatenation of the two last syllables of a word and the first one of another, or the last syllable of a word and the first two syllables of another. For non-words conditions, items comprised sixteen trisyllabic sequences selected from the streams. After exposure to each stream, a two alternative forced-choice (2AFC) behavioral test was administered to determine whether the participants were able to identify the words heard previously”

C:\Users\Stefan Elmer\Desktop\ITC_word_learning\Analyses_2019_absolute_values\Manuskript_ITC\Figures\Figure_4_def.tif

* Please insert Figure 7 at about here –

Figure 7. Schematic representation of the experimental design and theoretical framework of brain-to-stimulus alignment.

**EEG data acquisition, pre-processing and ERP analyses**

The scalp EEG was recorded from 29 electrodes located at standard positions using electrocap (Electro-Cap International). Vertical eye movements were monitored with an electrode at the infraorbital ridge of the right eye, and electrode impedances were kept below 3 kΩ. The electrophysiological signal was filtered on-line with a bandpass of 0.01-50 Hz (half-amplitude cutoffs) and digitized at a rate of 250 Hz. All pre-processing steps were performed with the Brain Vision Analyzer software package (version 2.01; Brain Products). In particular, the EEG signal was re-referenced off-line to the mean of the activity at the two mastoid electrodes, data were filtered with a low-pass filter of 30 Hz (including a Notch filter of 50 Hz), and artifacts (eye movements and blinks) were corrected using an independent component analysis 57. Furthermore, an automatic raw data inspection was used to remove remaining artefacts if a voltage gradient criterion of 50 μV/ms or an amplitude criterion of ±100 μV (200 ms before and after the event) was exceeded. Afterwards, each block of the different “languages” was segmented into single epochs of 796 ms (including a pre-stimulus baseline of 100 ms), and baseline correction was performed in the time range from -100 to 0 ms. The single epochs were subjected to two different type of analyses where we evaluated ERPs and ITCs. ERP analyses focused on two specific components that have previously been shown to be sensitive to speech segmentation based on statistical learning and prosodic bootstrapping, namely the P200 and N400 waveforms 4,8,17,18. For the ERP analyses, the single baseline-corrected epochs were averaged separately for the structured and random sequences of flat and stressed speech and for the four blocks. Afterwards, based on the fact that the P200 and N400 components elicited maximal voltage strength at central and anterior electrodes, and for reasons of comparability between ERP and ITC values, we averaged N200 and N400 responses across 6 channels, namely F3, Fz, F4, C3, Cz and C4. Finally, in accordance with previous studies using exactly the same paradigm and stimuli, mean amplitudes were extracted in two time windows overlapping with the P200 (170-250 ms) and N400 (350-550 ms) components8,18. Otherwise, for the ITC analyses the pre-stimulus period was removed, and the single epochs were exported to MATLAB for ITC analyses with home-made scripts.

**ITC and wavelet analysis**

ITC across the whole frequency spectrum was computed for each electrode, structured and random sequences of flat and stressed speech and the four blocks using the following Morlet wavelet transform:

Before transferring the signal to the wavelet domain, we added zero-padding of the same length as the single EEG segments (200 sample points) at the beginning as well as at the end of the single epochs to increase resolution at low frequencies. All ITC analyses were computed using the command “cwt” and home-made MATLAB scripts. Phase information was extracted from the wavelet transfer function, and ITC values corresponding to the word (1.43 Hz) and syllable (4.31 Hz) frequency were calculated by summation of phase angle 58 of all epochs according to the following formulas:

In these equations, N corresponds to the number of trials, whereasdepicts the local phase angle of the signal. Since we were interested in absolute phase-shifting, we computed the absolute value of ITC at the word frequency and at the syllabic rate. Furthermore, based on a previous work of Battering and colleagues 17, we calculated the ratio of ITC at the word frequency to ITC at the syllable frequency (word learning index, WLI) as well as the ratio of ITC at the syllable frequency to ITC at the word frequency (syllable learning index, SLI). Based on the fact that in the structured sequences ITC at the word frequency (1.43 Hz) and in the random sequences ITC at the syllabic rate (4.31 Hz) showed the strongest values at frontal and central electrodes (Figure 7), and in order to increase signal-to-noise ratio, ITC values, WLI and SLI were averaged across six frontal and central electrodes (F3, Fz, F4, C3, Cz and C4) and subjected to statistical analyses. A similar pooling procedure has previously been used by Batterink and colleagues 17.

C:\Users\Stefan Elmer\Desktop\ITC_word_learning\Analyses_2019_absolute_values\Manuskript_ITC\Figures\Figure_6_defdef.tif

- Insert Figure 8 at about here please -

Figure 8. Topographic distribution of mean ITC values at the (a) word frequency for structured sequences and at the (b) syllable frequency for random sequences.

**Statistical analyses**

All analyses were performed using [parametric statistics](https://www.sciencedirect.com/topics/medicine-and-dentistry/parametric-test) implemented in the IBM SPSS Statistics 22 software (SPSS, an IBM company, Armonk, New York, USA). In order to testify that the participants were able to segment speech and learn the words, we performed separate one-sample t-tests for the flat and stressed conditions and tested the percentage of correct responses against chance level (50 %, the behavioral data of 3 participants of the stressed condition are missing). The ERP data were evaluated by means of separate 2 x 4 (2 structured/random sequences x 4 blocks) ANOVAs for the P200 and N400 components and the flat and stressed conditions. Otherwise, ITC data were evaluated using a hierarchical approach. To maximize statistical power, in a first analysis we compared mean ITC (averaged across four blocks) at the syllable and word frequency, WLI and SLI between structured and random sequences of flat and stressed speech. With this purpose in mind, we computed separate univariate ANOVAs with the within-subject factor “sequences” (structured and random). This analysis aimed at testing whether neural synchronization to pertinent speech units likewise operates in statistical learning and prosodic bootstrapping conditions. In a second analysis, we then focused on neural entrainment dynamics as a function of exposure, and evaluated ITC at the syllabic rate and at the word frequency across the four blocks of the flat and stressed conditions for both structured and random sequences. Accordingly, we computed separate univariate analyses with the within-subject factor “block”. This hierarchically lower analysis allowed us to test whether in flat and stressed conditions speech segmentation is generally mediated by a neural transition from syllabic rate to word frequency, or whether syllables and words are concurrently tracked. In the case of neural transitions, ITC at word frequency should increase across the blocks, whereas ITC at the syllable frequency is expected to decrease. In contrast, if syllabic rate and word frequency are concurrently tracked, one would expect a general linear increase in ITC at both time scales over the blocks. Significant effects were further inspected by means of post-hoc t-tests (corrected for multiple comparisons using the Bonferroni procedure). Correlative analyses were computed according to Pearson's r (one-tailed, corrected for multiple comparisons using the Bonferroni procedure).

**ACKNOWLEDGMENTS**

This research was supported by research grants from the Spanish Government (MCYT) to A.R.F. (with EC Funds FEDER SEJ2005-06067/PSIC). Furthermore, A.R.F. was supported by the Ramon y Cajal research program.

### AUTHORS CONTRIBUTION

ARF, SE and TC planned the study, TC performed the EEG measurements and SAV wrote the MATLAB scripts and computed ITC analyses. SE and SAV analyzed the EEG data and performed the statistical analyses. SE, SAV, TC and ARF contributed to the interpretation of the data and wrote the manuscript.

**COMPETING INTERESTS**

The authors declare no competing interests.

**CODE AVAILABILITY**

The codes used for computing ITC analyses are available from SAV upon reasonable request.

**DATA AVAILABILITY**

The EEG data of this study are available from TC upon reasonable request.

**FIGURE LEGENDS**

Figure 1: Grand average event-related potentials of the structured and random sequences of flat (b) and stressed (d) speech. In the flat condition (b) the red line depicts brain responses to structured sequences, whereas the orange line represents the random sequences. In the stressed condition (d) brain responses to structured sequences are shown in blue, whereas random sequences are represented in green. Figures 1a and 1c show ITC differences at the syllabic rate and at the word frequency between structured and random sequences of flat (a) and stressed (c) speech.

Figure 2. a = mean ITC values in the flat condition for structured (red) and random (orange) sequences. b = mean ITC values in the stressed condition for structured (blue) and random (green) sequences. c = ITC values across the four blocks of the flat condition for structured (red) and random (orange) sequences. d = ITC values across the four blocks of the stressed condition for structured (blue) and random (green) sequences.

Figure 3. Single-subject data and violin plots with density distribution and mean for the flat (left) and stressed (right) conditions and structured (red and blue) and random (orange and green) sequences. a = ITC at the word frequency, b = ITC at the syllable frequency, c = word learning index (WLI), d = syllable learning index (SLI). \* = p < .05, \*\* = p < .01, \*\*\* = p < .001.

Figure 4. Correlation analyses for structured sequences of flat speech. a = Correlation between mean ITC at the syllable frequency and word frequency. b = Correlation between mean ITC at the word frequency and mean N400 amplitude. c, d, e, f = Correlations between mean ITC at the word frequency and mean N400 amplitude in the first (c), second (d), third (e) and fourth (f) block. n.s. = not significant after correction for multiple comparisons.

Figure 5. Single-subject data and violin plots with density distribution and mean for the four blocks. ITC at the word frequency (a, b, c, d) and ITC at the syllable frequency (e, f, g, h) are shown for the flat (a, c, e, g) and stressed conditions (b, d, f, h) as well as for structured (a, b, e, f) and random sequences (c, d, g, h). \*\* = p < .01, \*\*\* = p < .001.

Figure 6: Significant correlations between mean P200 amplitudes and ITC at the syllable frequency in the flat condition. a = block 4 of structured sequences, b = block 2 of random sequences, c = block 3 of random sequences, d = block 4 of random sequences.

Figure 7. Schematic representation of the experimental design and theoretical framework of brain-to-stimulus alignment.

Figure 8. Topographic distribution of mean ITC values at the (a) word frequency for structured sequences and at the (b) syllable frequency for random sequences.

**TABLE LEGENDS**

Table 1: Post-hoc comparisons of the significant main effects of “block” in the omnibus ANOVAs for P200 amplitudes. \* Depicts significance after correction for multiple comparisons.

Table 2: Post-hoc comparisons of the significant main effects of “block” in the omnibus ANOVAs for N400 amplitudes. \* Depicts significance after correction for multiple comparisons.

Table 3: Post-hoc comparisons of the significant main effects of “block” in the omnibus ANOVAs for ITC. \* Depicts significance after correction for multiple comparisons.

**REFERENCES**

1 Ding, N., Melloni, L., Zhang, H., Tian, X. & Poeppel, D. Cortical tracking of hierarchical linguistic structures in connected speech. *Nat Neurosci* **19**, 158-+, doi:10.1038/nn.4186 (2016).

2 Lehiste, I. An Acoustic-Phonetic Study of Internal Open Juncture. *Phonetica* **5**, S3-S54 (1960).

3 Assaneo, M. F. *et al.* Spontaneous synchronization to speech reveals neural mechanisms facilitating language learning. *Nat Neurosci* **22**, 627-+, doi:10.1038/s41593-019-0353-z (2019).

4 Batterink, L. J. & Paller, K. A. Statistical learning of speech regularities can occur outside the focus of attention. *Cortex* **115**, 56-71, doi:10.1016/j.cortex.2019.01.013 (2019).

5 Kuhl, P. K. Early language acquisition: Cracking the speech code. *Nat Rev Neurosci* **5**, 831-843, doi:10.1038/nrn1533 (2004).

6 Rodriguez-Fornells, A., Cunillera, T., Mestres-Misse, A. & de Diego-Balaguer, R. Neurophysiological mechanisms involved in language learning in adults. *Philos T R Soc B* **364**, 3711-3735, doi:10.1098/rstb.2009.0130 (2009).

7 Brent, M. R. Speech segmentation and word discovery: a computational perspective. *Trends Cogn Sci* **3**, 294-301, doi:Doi 10.1016/S1364-6613(99)01350-9 (1999).

8 Cunillera, T., Toro, J. M., Sebastian-Galles, N. & Rodriguez-Fornells, A. The effects of stress and statistical cues on continuous speech segmentation: An event-related brain potential study. *Brain Res* **1123**, 168-178, doi:10.1016/j.brainres.2006.09.046 (2006).

9 Jusczyk, P. W., Houston, D. M. & Newsome, M. The beginnings of word segmentation in English-learning infants. *Cognitive Psychol* **39**, 159-207, doi:DOI 10.1006/cogp.1999.0716 (1999).

10 Saffran, J. R., Newport, E. L. & Aslin, R. N. Word segmentation: The role of distributional cues. *J Mem Lang* **35**, 606-621, doi:DOI 10.1006/jmla.1996.0032 (1996).

11 Saffran, J. R., Aslin, R. N. & Newport, E. L. Statistical learning by 8-month-old infants. *Science* **274**, 1926-1928, doi:DOI 10.1126/science.274.5294.1926 (1996).

12 Cutler, A. & Norris, D. The Role of Strong Syllables in Segmentation for Lexical Access. *J Exp Psychol Human* **14**, 113-121, doi:Doi 10.1037/0096-1523.14.1.113 (1988).

13 Norris, D., McQueen, J. M. & Cutler, A. Merging information in speech recognition: Feedback is never necessary. *Behav Brain Sci* **23**, 299-+, doi:Doi 10.1017/S0140525x00003241 (2000).

14 Mattys, S. L., White, L. & Melhorn, J. F. Integration of multiple speech segmentation cues: A hierarchical framework. *J Exp Psychol Gen* **134**, 477-500, doi:10.1037/0096-3445.134.4.477 (2005).

15 Johnson, E. K. & Jusczyk, P. W. Word segmentation by 8-month-olds: When speech cues count more than statistics. *J Mem Lang* **44**, 548-567, doi:10.1006/jmla.2000.2755 (2001).

16 Thiessen, E. D. & Saffran, J. R. When cues collide: Use of stress and statistical cues to word boundaries by 7-to 9-month-old infants. *Dev Psychol* **39**, 706-716, doi:10.1037/0012-1649.39.4.706 (2003).

17 Batterink, L. J. & Paller, K. A. Online neural monitoring of statistical learning. *Cortex* **90**, 31-45, doi:10.1016/j.cortex.2017.02.004 (2017).

18 Cunillera, T. *et al.* Time course and functional neuroanatomy of speech segmentation in adults. *Neuroimage* **48**, 541-553, doi:10.1016/j.neuroimage.2009.06.069 (2009).

19 Lopez-Barroso, D. *et al.* Word learning is mediated by the left arcuate fasciculus. *P Natl Acad Sci USA* **110**, 13168-13173, doi:10.1073/pnas.1301696110 (2013).

20 Giraud, A. L. & Poeppel, D. Cortical oscillations and speech processing: emerging computational principles and operations. *Nat Neurosci* **15**, 511-517, doi:10.1038/nn.3063 (2012).

21 Ghitza, O. Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. *Front Psychol* **2**, doi:ARTN 13010.3389/fpsyg.2011.00130 (2011).

22 Panzeri, S., Brunel, N., Logothetis, N. K. & Kayser, C. Sensory neural codes using multiplexed temporal scales. *Trends Neurosci* **33**, 111-120, doi:10.1016/j.tins.2009.12.001 (2010).

23 Elmer, S., Albrecht, J., Valizadeh, S. A., Francois, C. & Rodriguez-Fornells, A. Theta Coherence Asymmetry in the Dorsal Stream of Musicians Facilitates Word Learning. *Sci Rep-Uk* **8**, doi:ARTN 456510.1038/s41598-018-22942-1 (2018).

24 Buiatti, M., Pena, M. & Dehaene-Lambertz, G. Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. *Neuroimage* **44**, 509-519, doi:10.1016/j.neuroimage.2008.09.015 (2009).

25 Henin, S. *et al.* Statistical learning shapes neural sequence representations. *bioRxiv*, doi:<https://doi.org/10.1101/583856> (2019).

26 De Diego Balaguer, R., Toro, J. M., Rodriguez-Fornells, A. & Bachoud-Levi, A. C. Different neurophysiological mechanisms underlying word and rule extraction from speech. *PLoS One* **2**, e1175, doi:10.1371/journal.pone.0001175 (2007).

27 Ding, N. *et al.* Characterizing Neural Entrainment to Hierarchical Linguistic Units using Electroencephalography (EEG). *Front Hum Neurosci* **11**, doi:ARTN 48110.3389/fnhum.2017.00481 (2017).

28 Altmann, G. T. M. Statistical learning in infants. *P Natl Acad Sci USA* **99**, 15250-15251, doi:10.1073/pnas.262659399 (2002).

29 Cunillera, T., Laine, M., Camara, E. & Rodriguez-Fornells, A. Bridging the gap between speech segmentation and word-to-world mappings: Evidence from an audiovisual statistical learning task. *J Mem Lang* **63**, 295-305, doi:10.1016/j.jml.2010.05.003 (2010).

30 Griffiths, T. D. *et al.* Direct Recordings of Pitch Responses from Human Auditory Cortex. *Curr Biol* **20**, 1128-1132, doi:10.1016/j.cub.2010.04.044 (2010).

31 Leaver, A. M. & Rauschecker, J. P. Functional Topography of Human Auditory Cortex. *J Neurosci* **36**, 1416-1428, doi:10.1523/Jneurosci.0226-15.2016 (2016).

32 Schneider, P. *et al.* Structural and functional asymmetry of lateral Heschl's gyrus reflects pitch perception preference. *Nat Neurosci* **8**, 1241-1247, doi:10.1038/nn1530 (2005).

33 Yi, H. G., Leonard, M. K. & Chang, E. F. The Encoding of Speech Sounds in the Superior Temporal Gyrus. *Neuron* **102**, 1096-1110, doi:10.1016/j.neuron.2019.04.023 (2019).

34 Lakatos, P. *et al.* Global dynamics of selective attention and its lapses in primary auditory cortex. *Nat Neurosci* **19**, 1707-1717, doi:10.1038/nn.4386 (2016).

35 Wostmann, M., Herrmann, B., Maess, B. & Obleser, J. Spatiotemporal dynamics of auditory attention synchronize with speech. *P Natl Acad Sci USA* **113**, 3873-3878, doi:10.1073/pnas.1523357113 (2016).

36 Hyafil, A., Fontolan, L., Kabdebon, C., Gutkin, B. & Giraud, A. L. Speech encoding by coupled cortical theta and gamma oscillations. *Elife* **4**, doi:ARTN e0621310.7554/eLife.06213 (2015).

37 Pefkou, M., Arnal, L. H., Fontolan, L. & Giraud, A. L. theta-Band and beta-Band Neural Activity Reflects Independent Syllable Tracking and Comprehension of Time-Compressed Speech. *J Neurosci* **37**, 7930-7938, doi:10.1523/Jneurosci.2882-16.2017 (2017).

38 Perruchet, P., Poulin-Charronnat, B., Tillmann, B. & Peereman, R. New evidence for chunk-based models in word segmentation. *Acta Psychol* **149**, 1-8, doi:10.1016/j.actpsy.2014.01.015 (2014).

39 Perruchet, P. & Vinter, A. PARSER: A model for word segmentation. *J Mem Lang* **39**, 246-263, doi:DOI 10.1006/jmla.1998.2576 (1998).

40 Andermann, M., Patterson, R. D., Vogt, C., Winterstetter, L. & Rupp, A. Neuromagnetic correlates of voice pitch, vowel type, and speaker size in auditory cortex. *Neuroimage* **158**, 79-89, doi:10.1016/j.neuroimage.2017.06.065 (2017).

41 Tang, C., Hamilton, L. S. & Chang, E. F. Intonational speech prosody encoding in the human auditory cortex. *Science* **357**, 797-801, doi:10.1126/science.aam8577 (2017).

42 Zatorre, R. J. & Gandour, J. T. Neural specializations for speech and pitch: moving beyond the dichotomies. *Philos T R Soc B* **363**, 1087-1104, doi:10.1098/rstb.2007.2161 (2008).

43 Rif, J., Hari, R., Hamalainen, M. S. & Sams, M. Auditory attention affects two different areas in the human supratemporal cortex. *Electroencephalogr Clin Neurophysiol* **79**, 464-472, doi:10.1016/0013-4694(91)90166-2 (1991).

44 Rosburg, T., Trautner, P., Elger, C. E. & Kurthen, M. Attention effects on sensory gating - Intracranial and scalp recordings. *Neuroimage* **48**, 554-563, doi:10.1016/j.neuroimage.2009.06.063 (2009).

45 Shahin, A., Bosnyak, D. J., Trainor, L. J. & Roberts, L. E. Enhancement of neuroplastic P2 and N1c auditory evoked potentials in musicians. *J Neurosci* **23**, 5545-5552 (2003).

46 Trainor, L. J., Shahin, A. & Roberts, L. E. Effects of musical training on the auditory cortex in children. *Ann Ny Acad Sci* **999**, 506-513, doi:10.1196/annals.1284.061 (2003).

47 Paulmann, S. & Kotz, S. A. Early emotional prosody perception based on different speaker voices. *Neuroreport* **19**, 209-213, doi:DOI 10.1097/WNR.0b013e3282f454db (2008).

48 Pinheiro, A. P., Vasconcelos, M., Dias, M., Arrais, N. & Goncalves, O. F. The music of language: An ERP investigation of the effects of musical training on emotional prosody processing. *Brain Lang* **140**, 24-34, doi:10.1016/j.bandl.2014.10.009 (2015).

49 Grill-Spector, K., Henson, R. & Martin, A. Repetition and the brain: neural models of stimulus-specific effects. *Trends Cogn Sci* **10**, 14-23, doi:10.1016/j.tics.2005.11.006 (2006).

50 Hyde, K. L., Peretz, I. & Zatorre, R. J. Evidence for the role of the right auditory cortex in fine pitch resolution. *Neuropsychologia* **46**, 632-639, doi:10.1016/j.neuropsychologia.2007.09.004 (2008).

51 Kutas, M. & Hillyard, S. A. Reading Senseless Sentences - Brain Potentials Reflect Semantic Incongruity. *Science* **207**, 203-205, doi:DOI 10.1126/science.7350657 (1980).

52 Vanpetten, C. & Kutas, M. Interactions between Sentence Context and Word-Frequency in Event-Related Brain Potentials. *Mem Cognition* **18**, 380-393, doi:Doi 10.3758/Bf03197127 (1990).

53 Mirman, D., Magnuson, J. S., Estes, K. G. & Dixon, J. A. The link between statistical segmentation and word learning in adults. *Cognition* **108**, 271-280, doi:10.1016/j.cognition.2008.02.003 (2008).

54 Spratling, M. W. Predictive coding as a model of biased competition in visual attention. *Vision Res* **48**, 1391-1408, doi:10.1016/j.visres.2008.03.009 (2008).

55 Kumar, S. *et al.* Predictive Coding and Pitch Processing in the Auditory Cortex. *J Cognitive Neurosci* **23**, 3084-3094, doi:DOI 10.1162/jocn\_a\_00021 (2011).

56 Dutoit, T., Pagel, V., Pierret, N., Bataille, F. & vanderVrecken, O. The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. *Icslp 96 - Fourth International Conference on Spoken Language Processing, Proceedings, Vols 1-4*, 1393-1396 (1996).

57 Jung, T. P. *et al.* Removing electroencephalographic artifacts by blind source separation. *Psychophysiology* **37**, 163-178, doi:Doi 10.1111/1469-8986.3720163 (2000).

58 van Diepen, R. M. & Mazaheri, A. The Caveats of observing Inter-Trial Phase-Coherence in Cognitive Neuroscience. *Sci Rep-Uk* **8**, doi:ARTN 299010.1038/s41598-018-20423-z (2018).