RESEARCH ARTICLE

An oscillating computational model can track pseudo-rhythmic speech by using linguistic predictions

Sanne ten Oever[1,2,3]*, Andrea E Martin[1,2]

[1]Language and Computation in Neural Systems group, Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands; [2]Donders Centre for Cognitive Neuroimaging, Radboud University, Nijmegen, Netherlands; [3]Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, Netherlands

Abstract Neuronal oscillations putatively track speech in order to optimize sensory processing. However, it is unclear how isochronous brain oscillations can track pseudo-rhythmic speech input. Here we propose that oscillations can track pseudo-rhythmic speech when considering that speech time is dependent on content-based predictions flowing from internal language models. We show that temporal dynamics of speech are dependent on the predictability of words in a sentence. A computational model including oscillations, feedback, and inhibition is able to track pseudo-rhythmic speech input. As the model processes, it generates temporal phase codes, which are a candidate mechanism for carrying information forward in time. The model is optimally sensitive to the natural temporal speech dynamics and can explain empirical data on temporal speech illusions. Our results suggest that speech tracking does not have to rely only on the acoustics but could also exploit ongoing interactions between oscillations and constraints flowing from internal language models.

***For correspondence:**
sanne.tenoever@mpi.nl

Competing interests: The authors declare that no competing interests exist.

Funding: See page 21

Preprinted: 07 December 2020
Received: 03 March 2021
Accepted: 16 July 2021
Published: 02 August 2021

Reviewing editor: Anne Kösem, Lyon Neuroscience Research Center, France

Introduction

Speech is a biological signal that is characterized by a plethora of temporal information. The temporal relationship between subsequent speech units allows for the online tracking of speech in order to optimize processing at relevant moments in time (***Jones and Boltz, 1989***; ***Large and Jones, 1999***; ***Giraud and Poeppel, 2012***; ***Ghitza and Greenberg, 2009***; ***Ding et al., 2017***; ***Arvaniti, 2009***; ***Poeppel, 2003***). Neural oscillations are a putative index of such tracking (***Giraud and Poeppel, 2012***; ***Schroeder and Lakatos, 2009***). The existing evidence for neural tracking of the speech envelope is consistent with such a functional interpretation (***Luo et al., 2013***; ***Keitel et al., 2018***). In these accounts, the most excitable optimal phase of an oscillation is aligned with the most informative time point within a rhythmic input stream (***Schroeder and Lakatos, 2009***; ***Lakatos et al., 2008***; ***Henry and Obleser, 2012***; ***Herrmann et al., 2013***; ***Obleser and Kayser, 2019***). However, the range of onset time difference between speech units seems more variable than fixed oscillations can account for (***Rimmele et al., 2018***; ***Nolan and Jeon, 2014***; ***Jadoul et al., 2016***). As such, it remains an open question how is it possible that oscillations can track a signal that is at best only pseudo-rhythmic (***Nolan and Jeon, 2014***).

Oscillatory accounts tend to focus on the prediction in the sense of predicting 'when', rather than predicting 'what': oscillations function to align the optimal moment of processing given that timing is predictable in a rhythmic input structure. If rhythmicity in the input stream is violated, oscillations must be modulated to retain optimal alignment to incoming information. This can be achieved through phase resets (***Rimmele et al., 2018***; ***Meyer, 2018***), direct coupling of the acoustics to

oscillations (**Poeppel and Assaneo, 2020**), or the use of many oscillators at different frequencies (**Large and Jones, 1999**). However, the optimal or effective time of processing stimulus input might not only depend on when you predict something to occur, but also depend on what stimulus is actually being processed (**Ten Oever et al., 2013**; **Martin, 2016**; **Rosen, 1992**; **deen et al., 2017**).

What and when are not independent, and certainly not from the brain's-eye-view. If continuous input arrives to a node in an oscillatory network, the exact phase at which this node reaches threshold activation does not only depend on the strength of the input, but also depend on how sensitive this node was to start with. Sensitivity of a node in a language network (or any neural network) is naturally affected by predictions in the what domain generated by an internal language model (**Martin, 2020**; **Marslen-Wilson, 1987**; **Lau et al., 2008**; **Nieuwland, 2019**). We define internal language model as the individually acquired statistical and structural knowledge of language stored in the brain. A virtue of such an internal language model is that it can predict the most likely future input based on the currently presented speech information. If a language model creates strong predictions, we call it a strong model. In contrast, a weak model creates no or little predictions about future input (note that the strength of individual predictions depends not only on the capability of the system to create a prediction, but also on the available information). If a node represents a speech unit that is likely to be spoken next, a strong internal language model will sensitize this node and it will therefore be active earlier, that is, on a less excitable phase of the oscillation. In the domain of working memory, this type of phase precession has been shown in rat hippocampus (**O'Keefe and Recce, 1993**; **Malhotra et al., 2012**) and more recently in human electroencephalography (**Bahramisharif et al., 2018**). In speech, phase of activation and perceived content are also associated (**Ten Oever and Sack, 2015**; **Kayser et al., 2016**; **Di Liberto et al., 2015**; **Ten Oever et al., 2016**; **Thézé et al., 2020**), and phase has been implicated in tracking of higher-level linguistic structure (**Meyer, 2018**; **Brennan and Martin, 2020**; **Kaufeld et al., 2020a**). However, the direct link between phase and the predictability flowing from a language model has yet to be established.

The time of speaking/speed of processing is not only a consequence of how predictable a speech unit is within a stream, but also a cue for the interpretation of this unit. For example, phoneme categorization depends on timing (e.g., voice onsets, difference between voiced and unvoiced phonemes), and there are timing constraints on syllable durations (e.g., the theta syllable **Poeppel and Assaneo, 2020**; **Ghitza, 2013**) that affect intelligibility (**Ghitza, 2012**). Even the delay between mouth movements and speech audio can influence syllabic categorizations (**Ten Oever et al., 2013**). Most oscillatory models use oscillations for parsing, but not as a temporal code for content (**Panzeri et al., 2015**; **Kayser et al., 2009**; **Mehta et al., 2002**; **Lisman and Jensen, 2013**). However, the time or phase of presentation does influence content perception. This is evident from two temporal speech phenomena. In the first phenomena, the interpretation of an ambiguous short /ɑ/ or long vowel /a:/ depends on speech rate (in Dutch; **Reinisch and Sjerps, 2013**; **Kösem et al., 2018**; **Bosker and Reinisch, 2015**). Specifically, when speech rates are fast the stimulus is interpreted as a long vowel and vice versa for slow rates. However, modulating the entrainment rate effectively changes the phase at which the target stimulus – which is presented at a constant speech rate – arrives (but this could not be confirmed in **Bosker and Kösem, 2017**). A second speech phenomena shows the direct phase-dependency of content (**Ten Oever and Sack, 2015**; **Ten Oever et al., 2016**). Ambiguous /da/-/ga/ stimuli will be interpreted as a /da/ on one phase and a /ga/ on another phase. This was confirmed in both a EEG and a behavioral study. An oscillatory theory of speech tracking should account for how temporal properties in the input stream can alter what is perceived.

In the speech production literature, there is strong evidence that the onset times (as well as duration) of an uttered word is modulated by the frequency of that word in the language (**O'Malley and Besner, 2008**; **Monsell, 1991**; **Monsell et al., 1989**; **Powers, 1998**; **Piantadosi, 2014**) showing that internal language models modulate the access to or sensitivity of a word node (**Martin, 2020**; **Hagoort, 2017**). This word-frequency effect relates to the access to a single word. However, it is likely that during ongoing speech internal language models use the full context to estimate upcoming words (**Beattie and Butterworth, 1979**; **Pluymaekers et al., 2005a**; **Lehiste, 1972**). If so, the predictability of a word in context should provide additional modulations on speech time. Therefore, we predict that words with a high predictability in the producer's language model should be uttered relatively early. In this way, word-to-word onset times map to the predictability level of that word within the internal model. Thus, not only the processing time depends on the predictability of a

word (faster processing for predictable words; see ***Gwilliams et al., 2020***; ***Deacon et al., 1995***, and ***Aubanel and Schwartz, 2020*** showing that speech time in noise matters), but also the production time (earlier uttering of predicted words).

Language comprehension involves the mapping of speech units from a producer's internal model to the speech units of the receiver's internal model. In other words, one will only understand what someone else is writing or saying if one's language model is sufficiently similar to the speakers (and if we speak in Dutch, fewer people will understand us). If the producer's and receiver's internal language model have roughly matching top-down constrains, they should similarly influence the speed of processing (either in production or perception; ***Figure 1A–C***). Therefore, if predictable words arrive earlier (due to high predictability in the producer's internal model), the receiver also expects the content of this word to match one of the more predictable ones from their own internal model (***Figure 1C***). Thus, the phase of arrival depends on the internal model of the producer and the expected phase of arrival depends on the internal model of the receiver (***Figure 1D***). If this is true, pseudo-rhythmicity is fully natural to the brain, and it provides a means to use time or arrival phase as a content indicator. It also allows the receiver to be sensitive to less predictable words when they arrive relatively late. Current oscillatory models of speech parsing do not integrate the constraints flowing from an internal linguistic model into the temporal structure of the brain response. It is therefore an open question whether the oscillatory model the brain employs is actually attuned to the temporal variations in natural speech.

Here, we propose that neural oscillations can track pseudo-rhythmic speech by taking into account that speech timing is a function of linguistic constrains. As such we need to demonstrate that speech statistics are influenced by linguistic constrains as well as showing how oscillations can be sensitive to this property in speech. We approach this hypothesis as follows: First, we demonstrate that in natural speech timing depends on linguistics predictions (*temporal speech properties*). Then, we model how oscillations can be sensitive to these linguistic predictions (*modeling speech tracking*). Finally, we validate that this model is optimally sensitive to the natural temporal properties in speech and displays temporal speech illusions (*model validation*). Our results reveal that tracking of speech needs to be viewed as an interaction between ongoing oscillations as well as constraints flowing from an internal language model (***Martin, 2016***; ***Martin, 2020***). In this way, oscillations do not have to shift their phase after every speech unit and can remain at a relatively stable frequency as long as the internal model of the speaker matches the internal model of the perceiver.

Results

Temporal speech properties

Word frequency influences word duration

To extract the temporal properties in naturally spoken speech we used the Corpus Gesproken Nederlands (CGN; [Version 2.0.3; 2014]). This corpus consists of elaborated annotations of over 900 hr of spoken Dutch and Flemish words. We focus here on the subset of the data of which onset and offset timings were manually annotated at the word level in Dutch. Cleaning of the data included removing all dashes and backslashes. Only words were included that were part of a Dutch word2vec embedding (github.com/coosto/dutch-word-embeddings; ***Nieuwenhuijse, 2018***; needed for later modeling) and required to have a frequency of at least 10 in the corpus. All other words were replaced with an <unknown> label. This resulted in 574,726 annotated words with 3096 unique words. Two thousand and forty-eight of the words were recognized in the Dutch Wordforms database in CELEX (Version 3.1) in order to extract the word frequency as well as the number of syllables per word (later needed to fit a regression model). Mean word duration was 0.392 s, with an average standard deviation of 0.094 s (***Figure 2—figure supplement 1***). By splitting up the data in sequences of 10 sequential words, we could extract the average word, syllable, and character rate (***Figure 2—figure supplements 2*** and ***3***). The reported rates fall within the generally reported ranges for syllables (5.2 Hz) and words (3.7 Hz; ***Ding et al., 2017***; ***Pellegrino and Coupé, 2011***).

We predict that knowledge about the language statistics influences the duration of speech units. As such we predict that more prevalent words will have on average a shorter duration (also reported in ***Monsell et al., 1989***). In ***Figure 2A***, the duration of several mono- and bi-syllabic words are listed with their word frequency. From these examples, it seems that words with higher word frequency

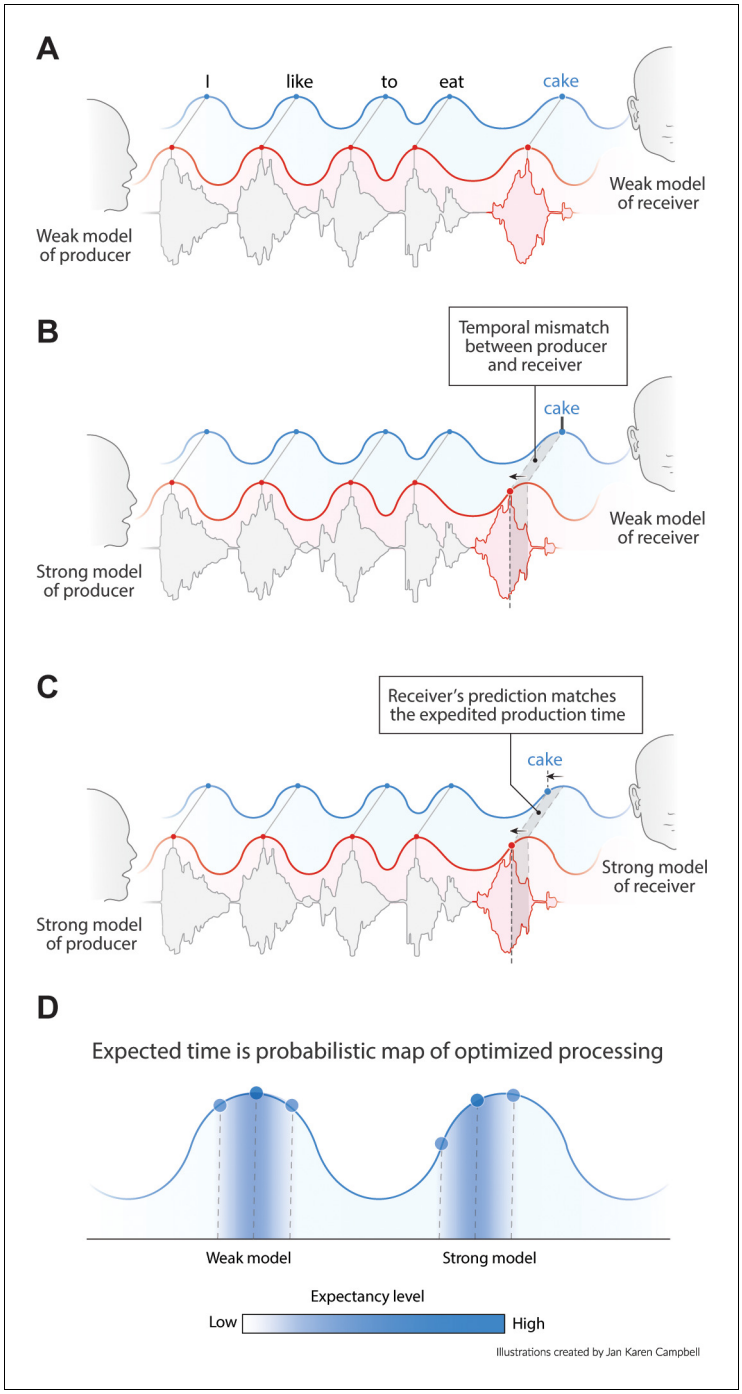

Figure 1. Proposed interaction between speech timing and internal linguistic models. (**A**) Isochronous production and expectation when there is a weak internal model (even distribution of node activation). All speech units arrive around the most excitable phase. (**B**) When the internal model of the producer does not align with the model of the receiver temporal alignment and optimal communication fails. (**C**) When both producer and receiver have a strong internal model, speech is non-isochronous and not aligned to the most excitable phase, but fully expected by the brain. (**D**) Expected time is a constraint distribution in which the center can be shifted due to linguistic constraints.

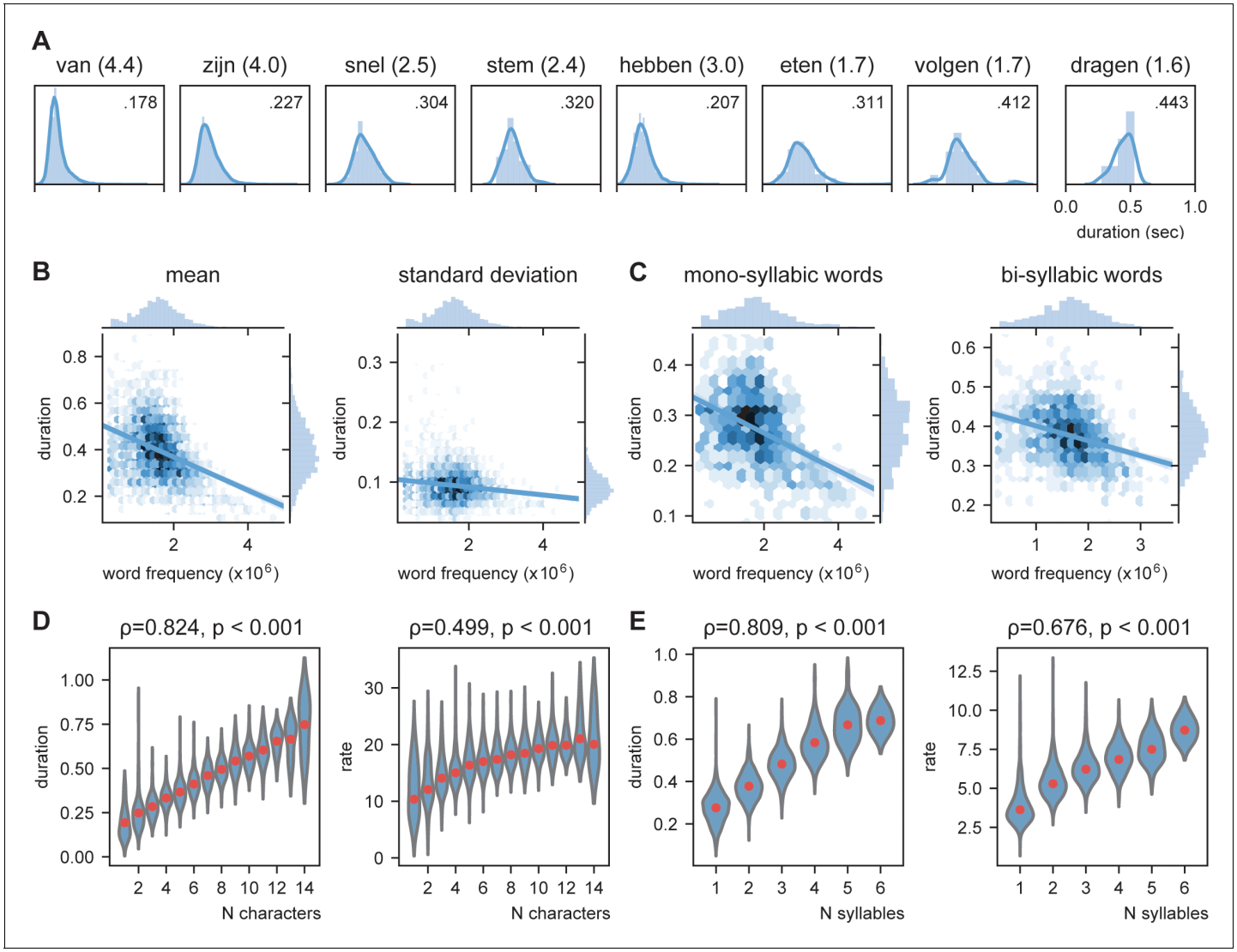

Figure 2. Word frequency modulates word duration. (**A**) Example of mono- and bi-syllabic words of different word frequencies in brackets (van=from, zijn=be, snel=fast, stem=voice, hebben=have, eten=eating, volgend=next, toekomst=future). Text in the graph indicates the mean word duration. (**B**) Relationship between word frequency and duration. Darker colors mean more values. (**C**) same as (**B**) but separately for mono- and bi-syllabic words. (**D**) Relationship character amount and word duration. The longer the words, the longer the duration (left). The increase in word duration does not follow a fixed number per character as duration as measured by rate increases. (**E**) same as (**D**) but for number of syllables. Red dots indicate the mean.
The online version of this article includes the following figure supplement(s) for figure 2:

Figure supplement 1. Distribution of mean duration (**A**) and of average rate (**B**).
Figure supplement 2. Distribution of mean duration split up for word length (in characters).
Figure supplement 3. Distribution of mean duration split up for syllable length.

generally have a shorter duration. To test this statistically we entered word frequency in an ordinary least square regression with number of syllables as control. Both number of syllables (coefficient = 0.1008, t(2843) = 75.47, p<0.001) as well as word frequency (coefficient = −0.022, t(2843) = −13.94, p<0.001) significantly influence the duration of the word. Adding an interaction term did not significantly improve the model (F (1,2843) = 1.320, p=0.251; ***Figure 2B,C***). The effect is so strong that words with a low frequency can last three times as long as high-frequency words (even within mono-syllabic words). This indicates that word frequency could be an important part of an internal model that influences word duration.

The previous analysis probed us to expand on the relationship between word duration and length of the words. Obviously, there is a strong correlation between word length and mean word duration

(number of characters 0.824, p<0.001; number of syllables: ρ = 0.808, p<0.001; for number of syllables already shown above; ***Figure 2D,E***). In contrast, this correlation is present, but much lower for the standard deviation of word duration (number of characters: ρ = 0.269, p<0.001; number of syllables: ρ = 0.292, p<0.001). Finding a strong correlation does not imply that for every time unit increase in the word length, the duration of the word also increases with the same time unit, i.e., bi-syllabic words do not necessarily have to last twice as long as mono-syllabic words. Therefore, we recalculated word duration to a rate unit considering the number of syllables/characters of the word. Thus, a 250 ms mono- versus bi-syllabic word would have a rate of 4 versus 8 Hz, respectively. Then we correlated character/syllabic rate with word duration. If word duration increases monotonically with character/syllable length, there should be no correlation. We found that the syllabic rate varies between 3 and 8 Hz as previously reported (***Figure 2E***, right; ***Ding et al., 2017***; ***Pellegrino and Coupé, 2011***). However, the more syllables there are in a word, the higher this rate (ρ = 0.676, p<0.001). This increase was less strong for the character rate (ρ = 0.499, p<0.001; ***Figure 2D***, right).

These results show that the syllabic/character rate depends on the number of characters /syllables within a word and is not an independent temporal unit (***Ghitza, 2013***). This effect is easy to explain when assuming that the prediction strength of an internal model influences word duration: transitional probabilities of syllables are simply more constrained within a word than across words (***Thompson and Newport, 2007***). This will reduce the time it takes to utter/perceive any syllable which is later in a word. In the current model, we focus on words (based on the availability of word2-vec embedding used to calculate contextual predictabilities based on a RNN) instead of syllables, so we will not test this prediction for syllables, but instead we can investigate the effect of transitional probabilities and other statistical regularities flowing from internal models across words (see next section and [***Jadoul et al., 2016***] for statistical regularities in syllabic processing).

Word-by-word predictability predicts word onset differences

The brain's internal model likely provides predictions about what linguistic features and representations, and possibly about which specific units, such as words, to expect next when listening to ongoing speech (***Martin, 2016***; ***Martin, 2020***). As such, it is also expected that word-by-word onset delays are shorter for words that fit the internal model (i.e., those that are expected; ***Beattie and Butterworth, 1979***). To investigate this possibility, we created a simplified version of an internal model predicting the next word using recurrent neural nets (RNN). We trained an RNN to predict the next word from ongoing sentences (***Figure 3A***). The model consisted of an embedding layer (pretrained; github.com/coosto/dutch-word-embeddings), a recurrent layer with a tanh activation function, and a dense output layer with a softmax activation. To prevent overfitting, we added a 0.2 dropout to the recurrent layers and the output layer. An Adam optimizer was used at a 0.001 learning rate and a batch size of 32. We investigated four different recurrent layers (GRU and LSTM at either 128 or 300 units; see ***Figure 3—figure supplement 1***). The final model we use here includes a LSTM with 300 units. Input data consistent of 10 sequential words (label encoding) within the corpus (of a single speaker; shifting the sentences by one word at a time), and an output consisted of a single word. A maximum of four unknown labeled words (words not included in the word2vec estimations. Four was chosen as it was <50% of the words) was allowed in the input, but not in output. Validation consisted of a randomly chosen 2% of the data.

The output of the RNN reflects a probability distribution in which the values of the RNN sum up to one and each word has its own predicted value (***Figure 3A***; see ***Figure 3—figure supplement 2*** for differences across words and sentence position). As such we can extract the predicted value of the uttered word and relate the RNN prediction with the stimulus onset delay relative to the previous word. We entered word prediction in a regression using the stimulus onset difference between the current word in the sentence and the previous word (i.e., onset difference of words). We added the control variables bigram (using the NLTK toolbox based on the training data only), frequency of previous word, syllable rate (rate of the full sentence input), and mean duration of previous word (all variables that can account for part of the variance that affects the duration of the last word). We only used the test data (total of 7361 sentences, excluding all words in which the previous word (W-1) was not present in Celex. 4837 sentences). Many of the variables were skewed to the right; therefore, we transformed the data accordingly (see ***Table 1***; results were robust to changes in these transformation).

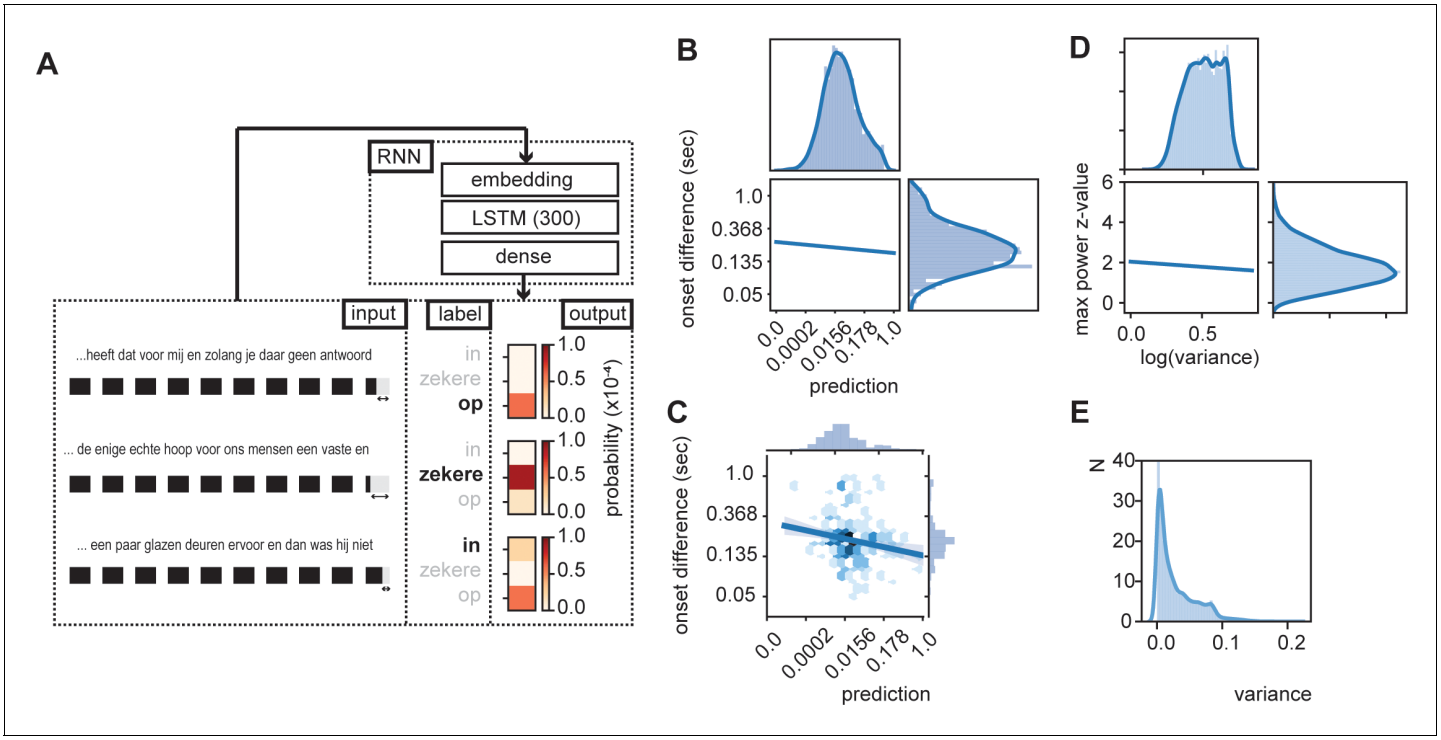

Figure 3. RNN output influence word onset differences. (**A**) Sequences of 10 words were entered in an RNN in order to predict the content of the next word. Three examples are provided of input data with the label (bold word) and probability output for three different words. The regression model showed a relation between the duration of last word in the sequence and the predictability of the next word such that words were systematically shorter when the next word was more predictable according to the RNN output (illustrated here with the shorted black boxes). (**B**) Regression line estimated at mean value of word duration and bigram. (**C**) Scatterplot of prediction and onset difference of data within ± 0.5 standard deviation of word duration and bigram. Note that for (**B**) and (**C**), the axes are linear on the transformed values. (**D**) Regression line for the correlation between logarithm of variance of the prediction and theta power. (**E**) None-transformed distribution of variance of the predictions (within a sentence). Translation of the sentences in (**A**) from top to bottom: '... that it has for me and while you have no answer [on]', '... the only real hope for us humans is a firm and [sure]', '... a couple of glass doors in front and then it would not have been [in]'.

The online version of this article includes the following figure supplement(s) for figure 3:

Figure supplement 1. Recurrent neural network evaluation.

Figure supplement 2. RNN prediction distributions.

All predictors except word frequency of the previous word showed a significant effect (***Table 1***). The variance explained by word frequency was likely captured by the mean duration variable of the previous word, which is correlated to word frequency. The RNN predictor could capture more variance than the bigram model, suggesting that word duration is modulated by the level of predictability within a fuller context than just the conditional probability of the current word given the previous

Table 1. Summary of regression model for logarithm of onset difference of words.

Variable	Trans	B	β	SE	t	p	VIF
Intercept	x	0.9719		0.049	19.764	<0.001	
RNN prediction	$x^{(1/6)}$	−0.3370	−0.0862	0.047	−7.163	<0.001	1.5
Bigram	log(x)	−0.0118	−0.0316	0.005	−2.424	0.015	1.8
Word frequency W-1	x	0.0049	0.0076	0.009	0.546	0.585	2.0
Mean duration W-1	log(x)	1.1206	0.7003	0.022	50.326	<0.001	2.0
Syllable Rate	x	−0.1033	−0.2245	0.004	−23.014	<0.001	1.0

Model R^2 = 0.542. Trans = transformation, W-1 = previous word, B = unstandardized coefficient, β = standardized coefficient, SE = standard error, t = t value, p = p value, VIF = variance inflation factor.

word (***Figure 3B,C***). Importantly, it was necessary to use the trained RNN model as a predictor; entering the RNN predictions after the first training cycle (of a total of 100) did not results in a significant predictor ($t(4837) = -1.191$, $p=0.234$). Also adding the predictor word frequency of the current word did not add significant information to the model ($F(1, 4830) = 0.2048$, $p=0.651$). These results suggest that words are systematically lengthened (or pauses are added. However, the same predictors are also significant when excluding sentences containing pauses) when the next word is not strongly predicted by the internal model. We also investigate whether RNN predictions have an influence on the duration of the word that has to be uttered. We found no effect on the duration (Supporting ***Table 1***).

Sentence isochrony depends on prediction variance

In the previous section, we investigated word-to-word onsets, but did not investigate how this influences the temporal properties within a full sentence. In a regular sentence, predictability values change from word-to-word. Based on the previous results, it is expected that overall sentences with a more stable predictability level (sequential words are equally predictable) should be more isochronous than sentences in which the predictability shifts from high to low. This prediction is based on the observation that when predictions are equal the expected shift is the same, while for varying predictions, temporal shifts vary (***Figure 3B,C***).

To test this hypothesis, we extracted the RNN prediction for 10 subsequent words. Then we extracted the variance of the prediction across those 10 words and extracted the word onset itself. We created a time course at which word onset were set to 1 (at a sampling rate of 100 Hz). Then we performed an fast Fourier transform (FFT) and extracted z-transformed power values over a 0–15 Hz interval. The power at the maximum power value with the theta range (3–8 Hz) was extracted. These max z-scores were correlated with the log transform of the variance (to normalize the skewed variance distribution; ***Figure 3E***). We found a weak, but significant negative correlation ($r = -0.062$, $p<0.001$; ***Figure 3D***) in line with our hypothesis. This suggests that the more variable the predictions within a sentence, the lower the peak power value is. When we repeated the analysis on the envelope, we did not find a significant effect.

Materials and methods

Speech Tracking in a Model Constrained Oscillatory Network

In order to investigate how much of these duration effects can be explained using an oscillator model, we created the model Speech Tracking in a Model Constrained Oscillatory Network (STiMCON). STiMCON in its current form will not be exhaustive; however, it can extract how much an oscillating network can cope with asynchronies by using its own internal model illustrating how the brain's language model and speech timing interact (***Guest and Martin, 2021***). The current model is capable of explaining how top-down predictions can influence the processing time as well as provide an explanation for two known temporal illusions in speech.

STiMCON consists of a network of semantic nodes of which the activation A of each level in the model l is governed by:

$$A_{l,T} = C_{l-1\rightarrow l} * A_{l-1,T} + C_{l+1\rightarrow l} * A_{l+1,T} + inhib(Ta) + osc(T) \quad (1)$$

in which C represents the connectivity patterns between different hierarchical levels, T the time in a sentence, and Ta the vector of times of an individual node in an inhibition function (in milliseconds). The inhibition function is a gate function:

$$inhib(Ta) = \begin{cases} -3 * BaseInhib, Ta \\ 3 * BaseInhib, 20 \leq Ta < 100 \\ BaseInhib, Ta > 100 \end{cases} \quad (2)$$

in which BaseInhib is a constant for the base level of inhibition (negative value, set to -0.2). As such nodes are by default inhibited, as soon as they get activated above threshold (activation threshold set at 1) Ta sets to zero. Then, the node will have suprathreshold activation, which after 20 ms returns to increased inhibition until the base level of inhibition is returned. These values are set to

reflect early excitation and longer lasting inhibition, which are only loosely related to neurophysiological time scales. The oscillation is a constant oscillator:

$$osc(T) = Am * e^{2\pi i\omega T + i\varphi} \quad (3)$$

in which Am is the amplitude of the oscillator, ω the frequency, and φ the phase offset. As such we assume a stable oscillator which is already aligned to the average speech rate (see ***Rimmele et al., 2018***; ***Poeppel and Assaneo, 2020*** for phase alignment models). The model used for the current simulation has one an input layer (l−1 level) and one single layer of semantic word nodes (l level) that receives feedback from a higher level layer (l+1 level). As such only the word (l) level is modeled according to ***Equation 1–3*** and the other levels form fixed input and feedback connection patterns. Even though the feedback influences the activity at the word level, it does not cause a phase reset as the phase of the oscillation does not change in response to this feedback.

Language models influence time of activation

To illustrate how STiMCON can explain how processing time depends on the prediction of internal language models, we instantiated a language model that had only seen three sentences and five words presented at different probabilities (I eat cake at 0.5 probability, I eat nice cake at 0.3 probability, I eat very nice cake at 0.2 probability; ***Table 2***). While in the brain the prediction should add up to 1, we can assume that the probability is spread across a big number of word nodes of the full language model and therefore neglectable. This language model will serve as the feedback arriving from the l+1-level to the l-level. The l-level consists of five nodes that each represent one of the words and receives proportional feedback from l+1 according to ***Table 2*** with a delay of 0.9*ω seconds, which then decays at 0.01 unit per millisecond and influences the l-level at a proportion of 1.5. The 0.9*ω was defined as we hypothesized that onset time would be loosely predicted around on oscillatory cycle, but to be prepared for input slightly earlier (which of course happens for predictable stimuli), we set it to 0.9 times the length of the cycle. The decay is needed and set such that the feedback would continue around a full theta cycle. The proportion was set empirically such to ensure that strong feedback did cause suprathreshold activation at the active node. The feedback is only initiated when supra-activation arrives due to l−1-level bottom-up input. Each word at the l−1-level input is modeled as a linearly function to the individual nodes lasting length of 125 ms (half a cycle, ranging from 0 to 1 arbitrary units). As such, the input is not the acoustic input itself but rather reflects a linear increase representing the increasing confidence of a word representing the specific node. φ is set such that the peak of a 4 Hz oscillation aligns to the peak of sensory input of the first word. Sensory input is presented at a base stimulus onset asynchrony of 250 ms (i.e., 4 Hz).

When we present this model with different sensory inputs at an isochronous rhythm of 4 Hz, it is evident that the timing at which different nodes reach activation depends on the level of feedback that is provided (***Figure 4***). For example, while the /I/-node needs a while to get activated after the initial sensory input, the /eat/-node is activated earlier as it is pre-activated due to feedback. After presenting /eat/, the feedback arrives at three different nodes and the activation timing depends on the stimulus that is presented (earlier activation for /cake/ compared to /very/).

Table 2. Example of a language model.
This model has seen three sentences at different probabilities. Rows represent the prediction for the next word, e.g., /I/ predicts /eat/ at a probability of 1, but after /eat/ there is a wider distribution.

	I	Eat	Very	Nice	Cake
I	0	1	0	0	0
eat	0	0	0.2	0.3	0.5
very	0	0	0	1	0
nice	0	0	0	0	1
cake	0	0	0	0	0

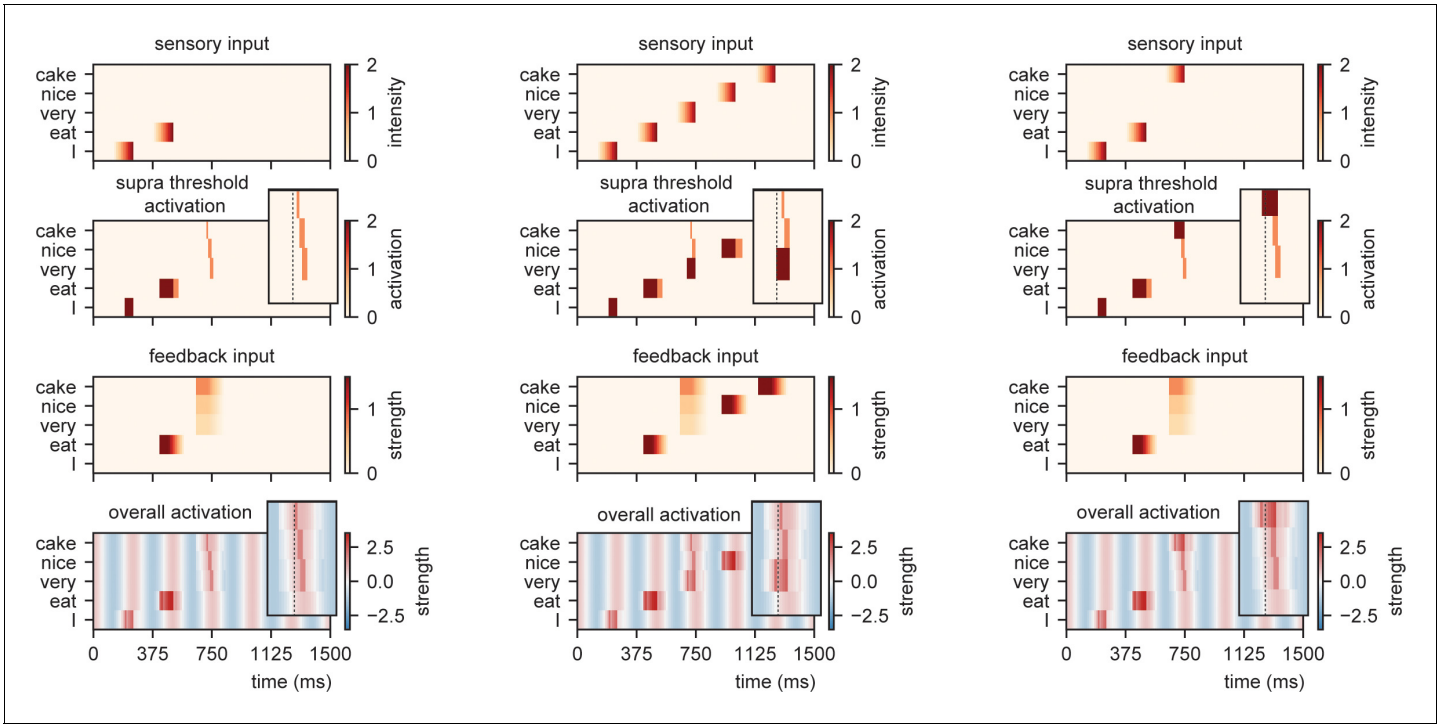

Figure 4. Model output for different sentences. For the supra-threshold activation dark red indicates activation which included input from I+1 as well as I1, orange indicates activation due to I+1 input. Feedback at different strengths causes phase dependent activation (left). Suprathreshold activation is reached earlier when a highly predicted stimulus (right) arrives, compared to a mid-level predicted stimulus (middle).

Time of presentation influences processing efficiency

To investigate how the time of presentation influences the processing efficiency, we presented the model with /I eat XXX/ in which the last word was varied in content (***Figure 5A***; either /I/, /very/, /nice/, or /cake/), intensity (linearly ranging from 0 to 1), and onset delay (ranging between −125 and +125 ms relative to isochronous presentation). We extracted the time at which the node matching the stimulus presentation reached activation threshold first (relative to stimulus onset and relative to isochronous presentation).

Figure 5B shows the output. When there is no feedback (i.e., at the first word /I/ presentation), a classical efficiency map can be found in which processing is most optimal (possible at lowest stimulus intensities) at isochronous (in phase with the stimulus rate) presentation and then drops to either side. For nodes that have feedback, input processing is possible at earlier times relative to isochronous presentation and parametrically varies with prediction strength (earlier for /cake/ at 0.5 probability, then /very/ at 0.2 probability). Additionally, the activation function is asymmetric. This is a consequence of the interaction between the supra-activation caused by the feedback and the sensory input. As soon as supra-activation is reached due to the feedback, sensory input at any intensity will reach supra-activity (thus at early stages of the linearly increasing confidence of the input). This is why for the /very/ stimulus activation is still reached at later delays compared to /nice/ and /cake/ as the /very/-node reaches supra-activation due to feedback at a later time point. In regular circumstances, we would of course always want to process speech, also when it arrives at a less excitable phase. Note, however, that the current stimulus intensities were picked to exactly extract the threshold responses. When we increase our intensity range above 2.1, nodes will always get activated even on the lowest excitable phase of the oscillation.

When we investigate timing differences in stimulus presentation, it is important to also consider what this means for the timing in the brain. Before, we showed that the amount of prediction can influence timing in our model. It is also evident that the earlier a stimulus was presented the more time it took (relative to the stimulus) for the nodes to reach threshold (more yellow colors for earlier delays). This is a consequence of the oscillation still being at a relatively low excitability point at

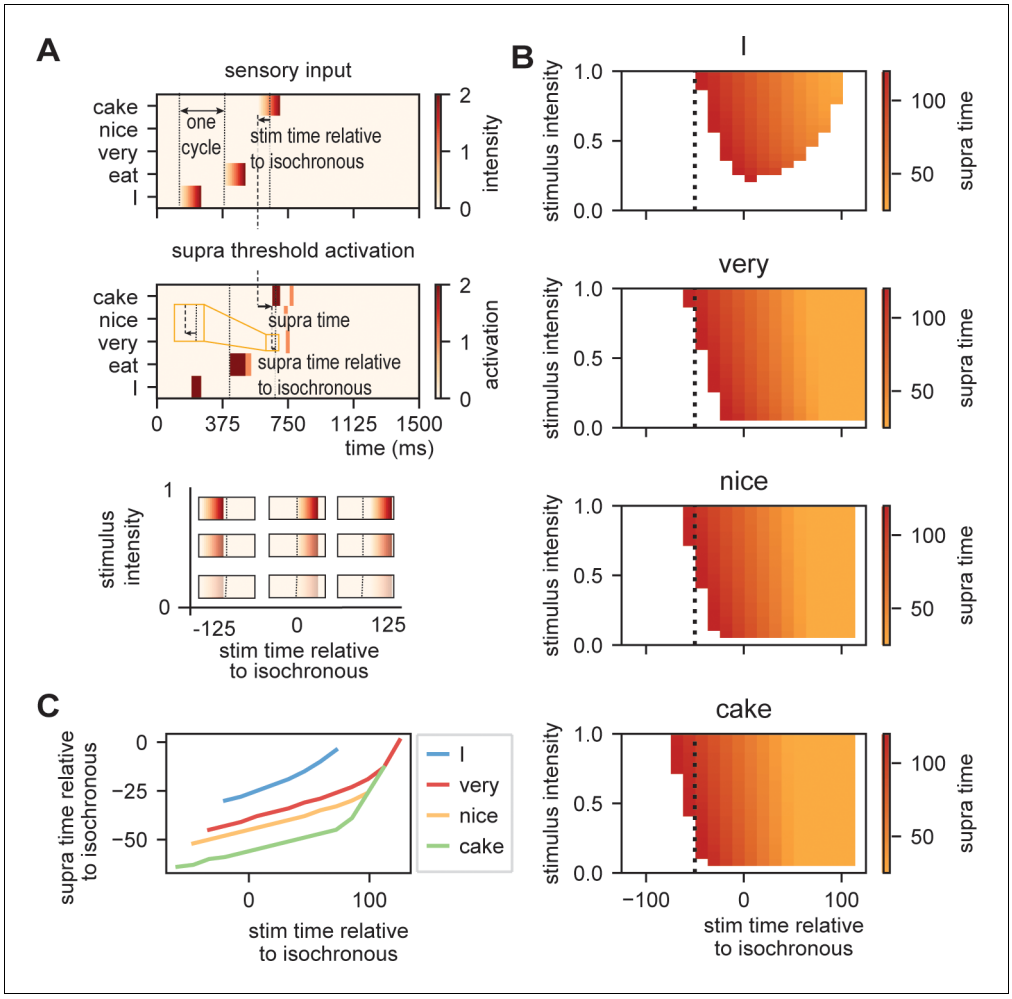

Figure 5. Model output on processing efficiency. (**A**) Input given to the model. Sensory input is varied in intensity and timing. We extract the time of activation relative to stimulus onset (supra-time) and relative to isochrony onset. (**B**) Time of presentation influences efficiency. Outcome variable is the time at which the node reached threshold activation (supra-time). The dashed line is presented to ease comparison between the four content types. White indicates that threshold is never reached. (C) Same as (B), but estimated at a threshold of 0.53 showing that oscillations regulate feedforward timing. Panel (C) shows that the earlier the stimuli are presented (on a weaker point of the ongoing oscillation), the longer it takes until supra-threshold activation is reached. This figure shows that timing relative to the ongoing oscillation is regulated such that the stimulus activation timing is closer to isochronous. Line discontinuities are a consequence of stimuli never reaching threshold for a specific node.

stimulus onset for stimuli that are presented early during the cycle. However, when we translate these activation threshold timing to the timing of the ongoing oscillation, the variation is strongly reduced (***Figure 5C***). A stimulus timing that varies between 130 ms (e.g., from −59 to +72 in the /cake/ line; excluding the non-linear section of the line) only reaches the first supra-threshold response with 19 ms variation in the model (translating to a reduction of 53–8% of the cycle of the ongoing oscillation, i.e., a 1:6.9 ratio). This means that within this model (and any oscillating model) the activation of nodes is robust to some timing variation in the environment. This effect seemed weaker when no prediction was present (for the /I/ stimulus this ratio was around 1:3.5. Note that when determining the /cake/ range using the full line the ratio would be 1:3.4).

Top-down interactions can provide rhythmic processing for non-isochronous stimulus input

The previous simulation demonstrate that oscillations provide a temporal filter and the processing at the word layer can actually be closer to isochronous than what can be solely extracted from the stimulus input. Next, we investigated whether dependent on changes in top-down prediction, processing within the model will be more or less rhythmic. To do this, we create stimulus input of 10 sequential words at a base rate of 4 Hz to the model with constant (***Figure 6A***; low at 0 and high at 0.8 predictability) or alternating word-to-word predictability. For the alternating conditions, word-to-word predictability alternates between low and high (sequences which word are predicted at 0 or 0.8 predictability, respectively) or shift from high to low. For this simulation, we used Gaussian sensory input (with a standard deviation of 42 ms aligning the mean at the peak of the ongoing oscillation; see ***Figure 6—figure supplement 1*** for output with linear sensory input). Then, we vary the onset time of the odd words in the sequence (shifting from −100 up to +100 ms) and the stimulus intensity (from 0.2 to 1.5). We extracted the overall activity of the model and computed the FFT of the created time course (using a Hanning taper only including data from 0.5 to 2.5 s to exclude the onset responses). From this FFT, we extracted the peak activation at the stimulation rate of 4 Hz.

The first thing that is evident is that the model with no content predictions has overall lowest power, but has the strongest 4 Hz response around isochronous presentation (odd word offset of 0 ms) at high stimulus intensities (***Figure 6B–D***) following closely the acoustic input. Adding overall high predictability increases the power, but also here the power seems symmetric around zero. The spectra of the alternating predictability conditions look different. For the low to high predictability condition, the curve seems to be shifted to the left such that 4 Hz power is strongest when the predictable odd stimulus is shifted to an earlier time point (low–high condition). This is reversed for the high–low condition. At middle stimulus intensities, there is a specific temporal specificity window at which the 4 Hz power is particularly strong. This window is earlier for the low–high than the high–low alternation (***Figure 6C,D*** and ***Figure 6—figure supplement 2***). The effect only occurs at specific middle-intensity combination as at high intensities the stimulus dominates the responses and at low intensities the stimulus does not reach threshold activation. These results show that even though stimulus input is non-isochronous, the interaction with the internal model can still create a potential isochronous structure in the brain (see ***Meyer et al., 2019***; ***Meyer et al., 2020***). Note that the direction in which the brain response is more isochronous matches with the natural onset delays in speech (shorter onset delays for more predictable stimuli).

Model validation

STiMCON's sinusoidal modulations of RNN predictions is optimally sensitive to natural onset delays

Next, we aimed to investigated whether STiMCON would be optimally sensitive to speech input timings found naturally in speech. Therefore, we tried to fit STIMCON's expected word-to-word onset differences to the word-to-word onset differences we found in the CGN. At a stable level of intensity of the input and inhibition, the only aspect that changes the timing of the interaction between top-down predictions and bottom-up input within STiMCON is the ongoing oscillation. Considering that we only want to model for individual words how much the prediction $(C_{l+1\rightarrow l} * A_{l+1,T})$ influences the expected timing we can set the contribution of the other factors from ***Equation (1)*** to zero remaining with the relative contribution of prediction:

$$C_{l+1\rightarrow l} * A_{l+1,T} = top\ down\ influence\ = -osc(T) \tag{4}$$

We can solve this formula in order to investigate the expected relative time shift (T) in processing that is a consequence of the strength of the prediction (ignoring that in the exact timing will also depend on the strength of the input and inhibition):

$$relative\ time\ shift = \frac{1}{2\pi\omega}\left(\arcsin\left(\frac{C_{l+1\rightarrow l} * A_{l+1,T}}{-Am}\right) - \varphi\right) \tag{5}$$

ω was set as the syllable rate for each sentence, and Am and φ were systematically varied. We fitted a linear model between the STiMCON's expected time and the actual word-to-word onset

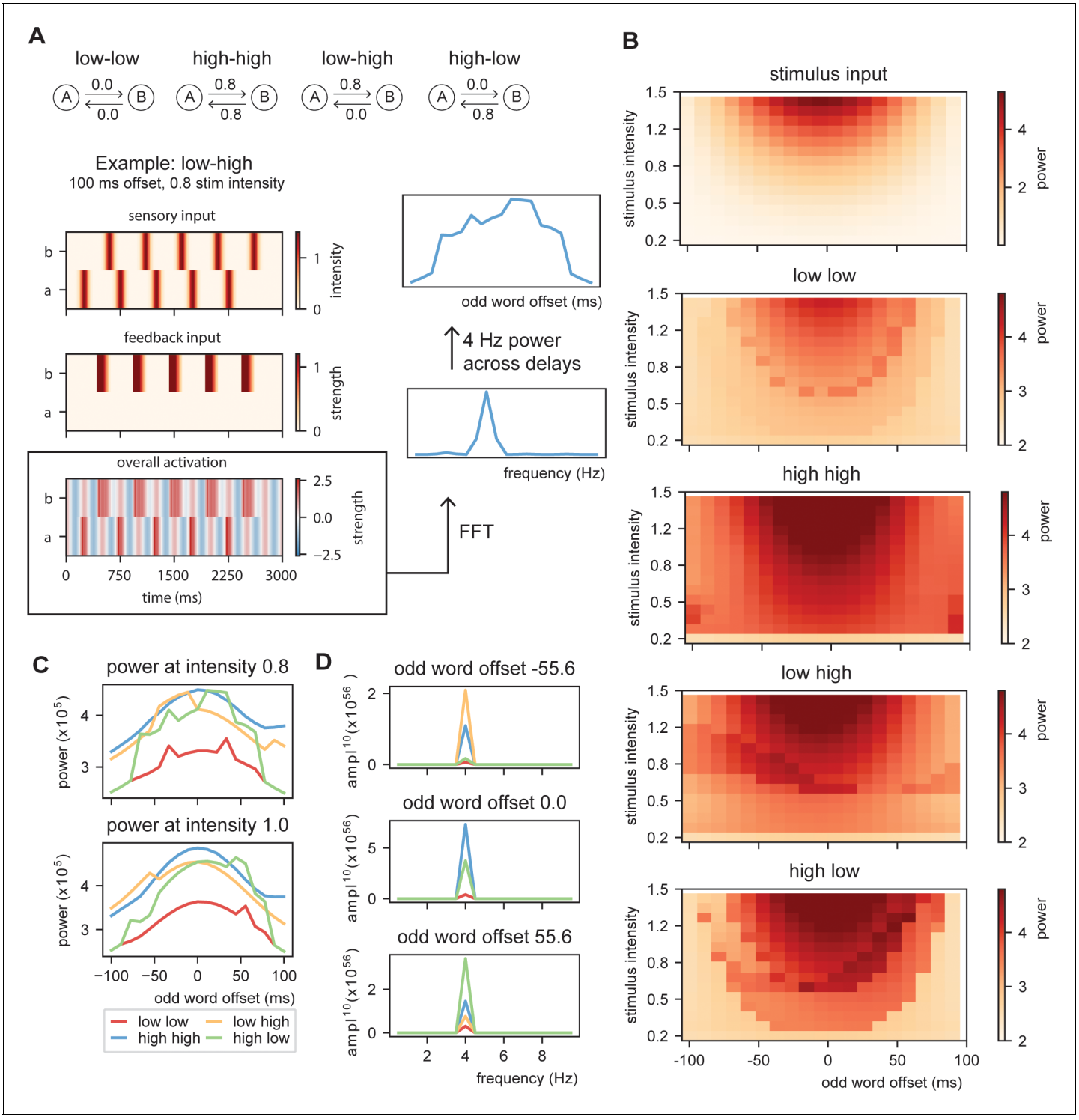

Figure 6. Model output on rhythmicity. (**A**) We presented the model with repeating (**A, B**) stimuli with varying internal models. We extracted the power spectra and peak activity at various odd stimulus offsets and stimulus intensities. (**B**) Strength of 4 Hz power depends on predictability in the stream. When predictability is alternated between low and high, activation is more rhythmic when the predictable odd stimulus arrives earlier and vice versa. (**C**) Power across different internal models at intensity of 0.8 and 1.0 (different visualization than **B**). (**D**) Magnitude spectra at three different odd word offsets at 1.0 intensity. To more clearly illustrate the differences, the magnitude to the power of 20 is plotted.

The online version of this article includes the following figure supplement(s) for figure 6:

Figure supplement 1. Power at 4 Hz using linearly increasing sensory input.

Figure supplement 2. Example of overall activation at threshold 0.8 (Gaussian shaped input).

differences. This model was similar to the model described in the section *Word-by-word predictability predicts word onset differences* and included the predictor syllable rate and duration of the previous word. However, as we were interested in how well non-transformed data matches the natural onset timings, we did not perform any normalization besides ***Equation (5)***. As this might involve violating some of the assumptions of the ordinary least square fit, we estimate model performance by repeating the regression 1000 times fitting it on 90% of the data (only including the test data from the RNN) and extracting R^2 from the remaining 10%.

Results show a modulation of the R^2 dependent on the amplitude and phase offset of the oscillation (***Figure 7A***). This was stronger than a model in which transformation in ***Equation (5)*** was not applied ($R^2$for a model with no transfomation was 0.389). This suggests that STiMCON expected time durations matches the actual word-by-word duration. This was even more strongly so for specific oscillatory alignments (around -0.25π offset), suggesting an optimal alignment phase relative to the ongoing oscillation is needed for optimal tracking (***Giraud and Poeppel, 2012***; ***Schroeder and Lakatos, 2009***). Interestingly, the optimal transformation seemed to automatically alter a highly skewed prediction distribution (***Figure 7B***) toward a more normal distribution of relative time shifts (***Figure 7C***). Note that the current prediction only operated on the word node (to which we have the RNN predictions), while full temporal shifts are probably better explained by word, syllabic, and phrasal predictions.

STiMCON can explain perceptual effects in speech processing

Due to the differential feedback strength and the inhibition after suprathreshold feedback stimulation, STiMCON is more sensitive to lower predictable stimuli at phases later in the oscillatory cycle. This property can explain two illusions that have been reported in the literature, specifically, the observation that the interpretation of ambiguous input depends on the phase of presentation (***Ten Oever and Sack, 2015***; ***Kayser et al., 2016***; ***Ten Oever et al., 2020***) and on speech rate (***Bosker and Reinisch, 2015***). The only assumption that has to be made is that there is an uneven base prediction balance between the ways the ambiguous stimulus can be interpreted.

The empirical data we aim to model comprises an experiment in which ambiguous syllables, which could either be interpreted as /da/ or /ga/, were presented (***Ten Oever and Sack, 2015***). In

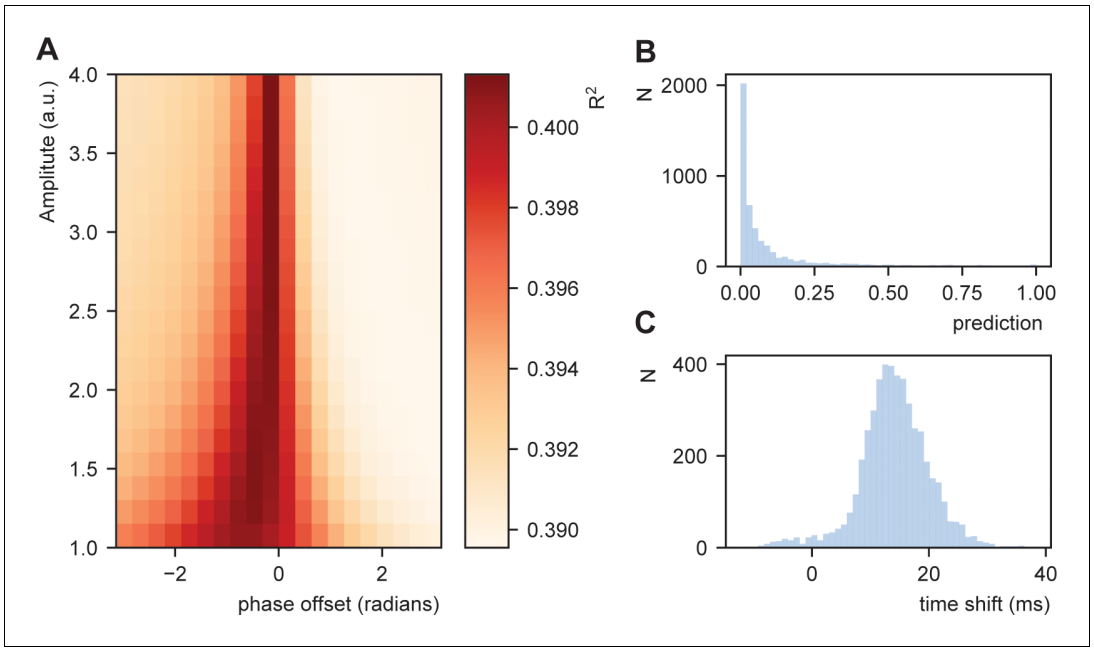

Figure 7. Fit between real and expected time shift dependent on predictability. (**A**) Phase offset and amplitude of the oscillation modulate the fit to the word-to-word onset durations. (**B**) Histogram of the predictions created by the deep neural net. (**C**) Histogram of the relative time shift transformation at phase of -0.15π and amplitude of 1.5.

one of the experiments in this study, broadband simuli were presented at specific rates to entrain ongoing oscillations. After the last entrainment stimulus, an ambiguous /daga/ stimulus was presented at different delays (covering two cycles of the presentation rate at 12 different steps), putatively reflecting different oscillatory phases. Dependent on the delay of stimulation participants perceived either /da/ or /ga/, suggesting that phase modulates the percept of the participants. Besides this behavioral experiment, the authors also demonstrated that the same temporal dynamics were present when looking at ongoing EEG data, showing that the phase of ongoing oscillations at the onset of ambiguous stimulus presentation determined the percept (***Ten Oever and Sack, 2015***).

To illustrate that STiMCON is capable of showing a phase (or delay) dependent effect, we use an internal language model similar to our original model (***Table 2***). The model consists of four nodes (N1, N2, Nda, and Nga). N1 and N2 represent nodes responsive to two stimulus S1 and S2 that function as entrainment stimuli. N1 activation predicts a second unspecific stimulus (S2) represented by N2 at a predictability of 1. N2 activation predicts either da or ga at 0.2 and 0.1 probability, respectively. This uneven prediction of /da/ and /ga/ is justified as /da/ is more prevalent in the Dutch language as /ga/ (***Zuidema, 2010***), and it thus has a higher predicted level of occurring. Then, we present STiMCON (same parameters as before) with /S1 S2 XXX/. XXX is varied to have different proportion of the stimulus /da/ and /ga/ (ranging from 0% /da/ to 100% /ga/ in 12 times steps; these reflect relative proportions that sum up to one such that at 30% the intensity of /da/ would be at max 0.3 and of /ga/ 0.7) and is the onset is varied relate to the second to last word. We extract the time that a node reaches suprathreshold activity after stimulus onset. If both nodes were active at the same time, the node with the highest total activation was chosen. Results showed that for some ambiguous stimuli, the delay determines which node is activated first, modulating the ultimate percept of the participant (***Figure 8A***, also see ***Figure 8—figure supplement 1A***). The same type of simulation can explain how speech rate can influence perception (***Figure 8—figure supplement 1B***; but see ***Bosker and Kösem, 2017***).

To further scrutinize on this effect, we fitted our model to the behavioral data of ***Ten Oever and Sack, 2015***. As we used an iterative approach in the simulations of the model, we optimized the model using a grid search. We varied the parameters of proportion of the stimulus being /da/ or /ga/ (ranging between 10:5:80%), the onset time of the feedback (0.1:0.1:1.0 cycle), the speed of the feedback decay (0:0.01:0.1), and a temporal offset of the final sound to account for the time it takes to interpret a specific ambiguous syllable (ranging between −0.05:0.01:0.05 s). Our first outcome variable was the node that show the first suprathreshold activation (Nda = 1, Nga = 0). If both nodes were active at the same time, the node with the highest total activates was chosen. If both nodes had equal activation or never reached threshold activation, we coded the outcome to 0.5 (i.e., fully ambiguous). These outcomes were fitted to the behavioral data of the 6.25 Hz and 10 Hz presentation rate (the two rates showing a significant modulation of the percept). This data was normalized to have a range between 0 and 1 to account for the model outcomes being binary (0, 0.5, or 1). As a second outcome measure, we also extracted the relative activity of the /da/ and /ga/ nodes by subtracting their activity and dividing by the summed activity. The activity was calculated as the average activity over a window of 500 ms after stimulus onset and the final time course was normalized between 0 and 1.

For the first node activation analysis, we found that our model could fit the data at an average explained variance of 43% (30% and 58% for 6.25 Hz and 10 Hz, respectively; ***Figure 8C,D***). For the average activity analysis, we found a fit with 83% explained variance. Compared to the original sinus fit, this explained variance was higher for the average activation analysis (40% for three parameter sinus fit [amplitude, phase offset, and mean]). Note that for the first node activation analysis, our fit cannot account for variance ranging between 0–0.5 and 0.5–1, while the sinus fit can do this. If we correct for this (by setting the sinus fit to the closest 0, 0.5, or 1 value and doing a grid search to optimize the fitting), the average fit of the sinus is 21%. Comparing the fits of the rectified sinus versus the first node activation reveals an average Akaike information criterion of the model and sinus fits of −27.0 and −24.1, respectively. For the average activation analysis, this was −41.5 versus −27.8, respectively. This overall suggests that the STiMCON model has the better fit. Thus, STiMCON does better than a fixed-frequency sinus fit. This is a likely consequence of the sinus fit not being able to explain the dampening of the oscillation later (i.e., the perception bias is stronger for shorter compared to longer delays).

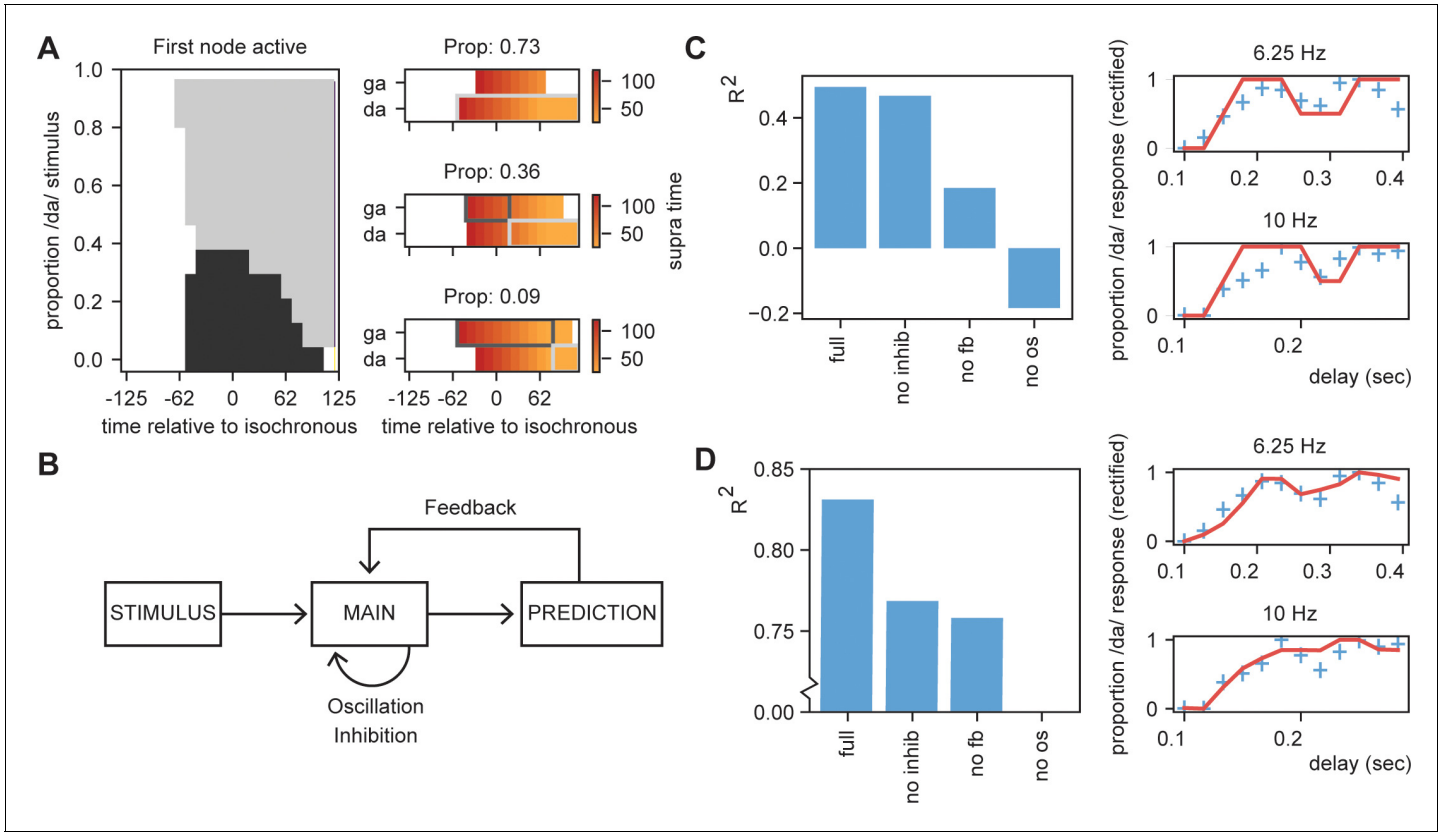

Figure 8. Results for /daga/ illusions. (**A**) Modulations due to ambiguous input at different times. Illustration of the node that is active first. Different proportions of the /da/ stimulus show activation timing modulations at different delays. (**B**) Summary of the model and the parameters altered for the empirical fits in (**C**) and (**D**). (**C**). R2 for the grid search fit of the full model using the first active node as outcome variable, a model without inhibition (no inhib), without uneven feedback (no fb), or without an oscillation (no os). The right panel shows the fit of the full model on the rectified behavioral data of ***Ten Oever and Sack, 2015***. Blue crossed indicate rectified data and red lines indicate the fit. (**D**) is the same as (**C**) but using the average activity instead of the first active node. Removing the oscillation results in an R^2 less than 0.

The online version of this article includes the following figure supplement(s) for figure 8:

Figure supplement 1. Explaining speech timing illusions.

Finally, we investigated the relevance of the three key features of our model for this fit: inhibition, feedback, and oscillations (***Figure 8B***). We repeated the grid search fit but set either the inhibition to zero, the feedback matrix equal for both /da/ and /ga/ (both 0.15), or the oscillation at an amplitude of zero. Results showed for both outcome measures that the full model showed the best performance. Without the oscillation, the models could not even fit better than the mean of the model ($R^2 < 0$). Removing the feedback had a negative influence on both the outcome measures, dropping the performance. Removing the inhibition reduced performance for both outcome measures, but more strongly on the average activation compared to the first active node model. This suggest that all features (with potentially to a lesser extend the inhibition) are required to model the data, suggesting that oscillatory tracking is dependent on linguistic constrains flowing from the internal language model.

Discussion

In the current paper, we combined an oscillating computational model with a proxy for linguistic knowledge, an internal language model, in order to investigate the model's processing capacity for onset timing differences in natural speech. We show that word-to-word speech onset differences in natural speech are indeed related to predictions flowing from the internal language model (estimated through an RNN). Fixed oscillations aligned to the mean speech rate are robust against

natural temporal variations and even optimized for temporal variations that match the predictions flowing from the internal model. Strikingly, when the pseudo-rhythmicity in speech matches the predictions of the internal model, responses were more rhythmic for matched pseudo-rhythmic compared to isochronous speech input. Our model is optimally sensitive to natural speech variations, can explain phase-dependent speech categorization behavior (***Ten Oever and Sack, 2015***; ***Thézé et al., 2020***; ***Reinisch and Sjerps, 2013***; ***Ten Oever et al., 2020***), and naturally comprises a neural phase code (***Panzeri et al., 2015***; ***Mehta et al., 2002***; ***Lisman and Jensen, 2013***). These results show that part of the pseudo-rhythmicity of speech is expected by the brain and it is even optimized to process it in this manner, but only when it follows the internal model.

Speech timing is variable, and in order to understand how the brain tracks this pseudo-rhythmic signal, we need a better understanding of how this variability arises. Here, we isolated one of the components explaining speech time variation, namely, constraints that are posed by an internal language model. This goes beyond extracting the average speech rate (***Ding et al., 2017***; ***Poeppel and Assaneo, 2020***; ***Pellegrino and Coupé, 2011***) and might be key to understanding how a predictive brain uses temporal cues. We show that speech timing depends on the predictions made from an internal language model, even when those predictions are highly reduced to be as simple as word predictability. While syllables generally follow a theta rhythm, there is a systematic increase in syllabic rate as soon as more syllables are in a word. This is likely a consequence of the higher close probability of syllables within a word which reduces the onset differences of the later uttered syllables (***Thompson and Newport, 2007***). However, an oscillatory model constrained by an internal language model is sensitive to these temporal variations, it is actually capable of processing them optimally.

The oscillatory model we here pose has three components: oscillations, feedback, and inhibition. The oscillations allow for the parsing of speech and provide windows in which information is processed (***Giraud and Poeppel, 2012***; ***Ghitza, 2012***; ***Peelle and Davis, 2012***; ***Martin and Doumas, 2017***). Importantly, the oscillation acts as a temporal filter, such that the activation time of any incoming signal will be confined to the high excitable window and thereby is relatively robust against small temporal variations (***Figure 5C***). The feedback allows for differential activation time dependent on the sensory input (***Figure 5B***). As a consequence, the model is more sensitive to higher predictable speech input and therefore active earlier on the duty cycle (this also means that oscillations are less robust against temporal variations when the feedback is very strong). The inhibition allows for the network to be more sensitive to less predictable speech units when they arrive later (the higher predictable nodes get inhibited at some point on the oscillation; best illustrated by the simulation in ***Figure 8A***). In this way, speech is ordered along the duty cycle according to its predictability (***Lisman and Jensen, 2013***; ***Jensen et al., 2012***). The feedback in combination with an oscillatory model can explain speech rate and phase-dependent content effects. Moreover, it is an automatic temporal code that can use time of activation as a cue for content (***Mehta et al., 2002***). Note that previously we have interpreted the /daga/ phase-dependent effect as a mapping of differences between natural audio-visual onset delays of the two syllabic types onto oscillatory phase (***Ten Oever et al., 2013***; ***Ten Oever and Sack, 2015***). However, the current interpretation is not mutually exclusive with this delay-to-phase mapping as audio-visual delays could be bigger for less frequent syllables. The three components in the model are common brain mechanisms (***Malhotra et al., 2012***; ***Mehta et al., 2002***; ***Buzsáki and Draguhn, 2004***; ***Bastos et al., 2012***; ***Michalareas et al., 2016***; ***Lisman, 2005***) and follow many previously proposed organization principles (e.g., temporal coding and parsing of information). While we implement these components on an abstract level (not veridical to the exact parameters of neuronal interactions), they illustrate how oscillations, feedback, and inhibition interact to optimize sensitivity to natural pseudo-rhythmic speech.

The current model is not exhaustive and does not provide a complete explanation of all the details of speech processing in the brain. For example, it is likely that the primary auditory cortex is still mostly modulated by the acoustic pseudo-rhythmic input and only later brain areas follow more closely the constraints posed by the language model of the brain. Moreover, we now focus on the word level, while many tracking studies have shown the importance of syllabic temporal structure (***Giraud and Poeppel, 2012***; ***Ghitza, 2012***; ***Luo and Poeppel, 2007***) as well as the role of higher order linguistic temporal dynamics (***Meyer et al., 2019***; ***Kaufeld et al., 2020b***). It is likely that predictive mechanisms also operate on these higher linguistic levels as well as on syllabic levels. It is

known, for example, that syllables are shortened when the following syllabic content is known versus producing syllables in isolation (**Pluymaekers et al., 2005a**; **Lehiste, 1972**). Interactions also occur as syllables part of more frequent words are generally shortened (**Pluymaekers et al., 2005b**). Therefore, more hierarchical levels need to be added to the current model (but this is possible following **Equation (1)**). Moreover, the current model does not allow for phase or frequency shifts. This was intentional in order to investigate how much a fixed oscillator could explain. We show that onset times matching the predictions from the internal model can be explained by a fixed oscillator processing pseudo-rhythmic input. However, when the internal model and the onset timings do not match, the internal model phase and/or frequency shift are still required and need to be incorporated (see e.g. **Rimmele et al., 2018**; **Poeppel and Assaneo, 2020**).

We aimed to show that a stable oscillator can be sensitive to temporal pseudo-rhythmicities when these shifts match predictions from an internal linguistic model (causing higher sensitivity to these nodes). In this way, we show that temporal dynamics in speech and the brain cannot be isolated from processing the content of speech. This is in contrast with other models that try to explain how the brain deals with pseudo-rhythmicity in speech (**Giraud and Poeppel, 2012**; **Rimmele et al., 2018**; **Doelling et al., 2019**). While some of these models discuss that higher-level linguistic processing can modulate the timing of ongoing oscillations (**Rimmele et al., 2018**), they typically do not consider that in the speech signal itself the content or predictability of a word relates to the timing of this word. Phase resetting models typically deal with pseudo-rhythmicity by shifting the phase of ongoing oscillations in response to a word that is offset to the mean frequency of the input (**Giraud and Poeppel, 2012**; **Doelling et al., 2019**). We believe that this cannot explain how the brain uses what/when dependencies in the environment to infer the content of the word (e.g., later words are likely a less predictable word). Our current model does not have an explanation of how the brain can actually entrain to an average speech rate. This is much better described in dynamical systems theories in which this is a consequence of the coupling strength between internal oscillations and speech acoustics (**Doelling et al., 2019**; **Assaneo et al., 2021**). However, these models do not take top-down predictive processing into account. Therefore, the best way forward is likely to extend coupling between brain oscillations and speech acoustics (**Poeppel and Assaneo, 2020**), with the coupling of brain oscillations to brain activity patterns of internal models (**Cumin and Unsworth, 2007**).

In the current paper, we use an RNN to represent the internal model of the brain. However, it is unlikely that the RNN captures the wide complexities of the language model in the brain. The decades-long debates about the origin of a language model in the brain remains ongoing and controversial. Utilizing the RNN as a proxy for our internal language model makes a tacit assumption that language is fundamentally statistical or associative in nature, and does not posit the derivation or generation of knowledge of grammar from the input (**Chater, 2001**; **McClelland and Elman, 1986**). In contrast, our brain could as well store knowledge of language that functions as fundamental interpretation principles to guide our understanding of language input (**Martin, 2016**; **Martin, 2020**; **Hagoort, 2017**; **Martin and Doumas, 2017**; **Friederici, 2011**). Knowledge of language and linguistic structure could be acquired through an internal self-supervised comparison process extracted from environmental invariants and statistical regularities from the stimulus input (**Martin and Doumas, 2019**; **Doumas et al., 2008**; **Doumas and Martin, 2018**). Future research should investigate which language model can better account for the temporal variations found in speech.

A natural feature of our model is that time can act as a cue for content implemented as a phase code (**Lisman and Jensen, 2013**; **Jensen et al., 2012**). This code unravels as an ordered list of predictability strength of the internal model. This idea diverges from the idea that entrainment should align to the most excitable phase of the oscillation with the highest energy in the acoustics (**Giraud and Poeppel, 2012**; **Rimmele et al., 2018**). Instead, this type of phase coding could increase the brain representational space to separate information content (**Lisman and Jensen, 2013**; **Panzeri et al., 2001**). We predict that if speech nodes have a different base activity, ambiguous stimulus interpretation should dependent on the time/phase of presentation (see **Ten Oever and Sack, 2015**; **Ten Oever et al., 2020**). Indeed, we could model two temporal speech illusions (**Figure 8**, **Figure 8—figure supplement 1**). There have also been null results regarding the influence of phase on ambiguous stimulus interpretation (**Bosker and Kösem, 2017**; **Kösem et al., 2016**). For the speech rate effect, when modifying the time of presentation with a neutral entrainer (summed sinusoidals with random phase), no obvious phase effect was reported (**Bosker and**

Kösem, 2017). A second null result relates to a study where participants were specifically instructed to maintain a specific perception in different blocks which likely increases the pre-activation and thereby the phase (*Kösem et al., 2016*). Future studies need to investigate the use of temporal/phase codes to disambiguate speech input and specifically use predictions in their design.

The temporal dynamics of speech signals needs to be integrated with the temporal dynamics of brain signals. However, it is unnecessary (and unlikely) that the exact duration of speech matches with the exact duration of brain processes. Temporal expansion or squeezing of stimulus inputs occur regularly in the brain (*Eagleman et al., 2005*; *Pariyadath and Eagleman, 2007*), and this temporal morphing also maps to duration (*Eagleman, 2008*; *Terao et al., 2008*; *Ulrich et al., 2006*) or order illusions (*Vroomen and Keetels, 2010*). Our model predicts increased rhythmic responses for non-isochronous speech matching the internal model. The perceived rhythmicity of speech could therefore also be an illusion generated by a rhythmic brain signal somewhere in the brain.

When investigating the pseudo-rhythmicity in speech, it is important to identify situations where speech is actually more isochronous. Two examples are the production of lists (*Jefferson, 1990*) and infant-directed speech (*Fernald, 2000*). In both these examples, it is clear that a strong internal predictive language model is lacking either on the producer's or on the receiver's side, respectively. The infant-directed speech also illustrates that a producer might proactively adapt its speech rhythm to the expectations of the internal model of the receiver to align better with the predictions from the receiver's model (*Figure 9B*; similar to when you are speaking to somebody that is just learning a new language). Other examples in which speech is more isochronous is during poems, during emotional conversation (*Hawkins, 2014*), and in noisy situations (*Bosker and Cooke, 2018*). While speculative, it is conceivable that in these circumstances one puts more weight on a different level of hierarchy than the internal linguistic model. In the case of poems and emotional conversation, an emotional route might get more weight in processing. In the case of noisy situations, stimulus input has to pass the first hierarchical level of the primary auditory cortex which effectively gets more weight than the internal model.

Conclusions

We argued that pseudo-rhythmicity in speech is in part a consequence of top-down predictions flowing from an internal model of language. This pseudo-rhythmicity is created by a speaker and expected by a receiver if they have overlapping internal language models. Oscillatory tracking of this signal does not need to be hampered by the pseudo-rhythmicity, but can use temporal variations as a cue to extract content information since the phase of activation parametrically relates to the likelihood of an input relative to the internal model. Brain responses can even be more isochronous to pseudo-rhythmic compared to isochronous speech if they follow the temporal delays imposed by the internal model. This account provides various testable predictions which, we list in *Table 3* and *Figure 9*. We believe that by integrating neuroscientific explanations of speech tracking with linguistic models of language processing (*Martin, 2016*; *Martin, 2020*), we can improve to explain temporal speech dynamics. This will ultimately aid our understanding of language in the brain and provide a means to improve temporal properties in speech synthesis.

Code availability statement

Code for the creation of the main figures is available on GitHub (*Ten Oever & Martin, 2021*; copy archived at swh:1:rev:873a2bf5c79fe2f828e72e14ef74db409d387854).

Acknowledgements

AEM was supported by the Max Planck Research Group and Lise Meitner Research Group 'Language and Computation in Neural Systems' from the Max Planck Society, and by the Netherlands Organization for Scientific Research (grant 016.Vidi.188.029 to AEM). *Figure 1* and *9* were created in collaboration with scientific illustrator Jan-Karen Campbell (http://www.jankaren.com).

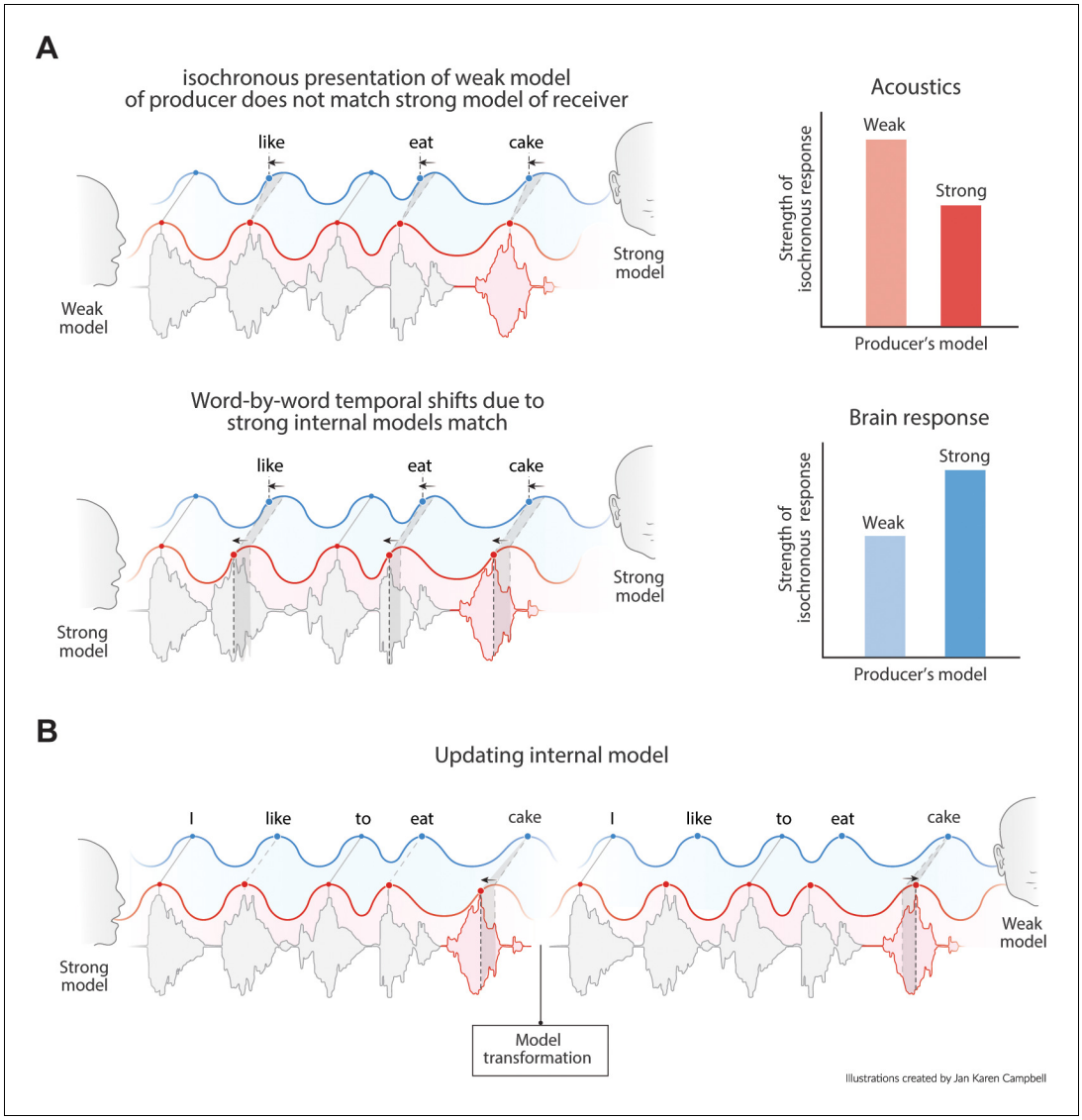

Figure 9. Predictions of the model. (**A**) Acoustics signals will be more isochronous when a producer has a weak versus a strong internal model (top right). When the producer's strong model matches the receiver's model, the brain response will be more isochronous for less isochronous acoustic input. (**B**) When a producer realizes the model of the receiver is weak, it might transform its model and thereby their speech timing to match the receiver's expectations.

Table 3. Predictions from the current model.

When there is a flat constraint distribution over an utterance (e.g., when probabilities are uniform over the utterance), the acoustics of speech should naturally be more isochronous (***Figures 9A*** and ***3D,E***).
If speech timing matches the internal language model, brain responses should be more isochronous even if the acoustics are not (***Figure 9A***).
The more similar the internal language models of two speakers, the more effective they are in 'entraining' each other's brain.
If speakers suspect their listener to have a flatter constraint distribution than themselves (e.g., the environment is noisy, or the speakers are in a second language context), they adjust to the distribution by speaking more isochronous (***Figure 9B***).
One adjusts the weight of the constraint distribution to a hierarchical level when needed. For example, when there is noise, participants adjust to the rhythm of primary auditory cortex instead of higher order language models. As a consequence, they speak more isochronous.
The theoretical account provides various predictions that are listed in this table.

Additional information

Funding

Funder	Grant reference number	Author
Max Planck Society	MaxPlanck Research Group	Andrea E Martin
Nederlandse Organisatie voor Wetenschappelijk Onderzoek	016.Vidi.188.029	Andrea E Martin
Max Planck Society	Lise Meitner Research Group	Andrea E Martin

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Sanne ten Oever, Conceptualization, Data curation, Formal analysis, Visualization, Methodology, Writing - original draft, Writing - review and editing; Andrea E Martin, Conceptualization, Resources, Supervision, Funding acquisition, Validation, Writing - review and editing

Author ORCIDs

Sanne ten Oever https://orcid.org/0000-0001-7547-5842
Andrea E Martin https://orcid.org/0000-0002-3395-7234

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.68066.sa1
Author response https://doi.org/10.7554/eLife.68066.sa2

Additional files

Supplementary files

- Supplementary file 1. Summary of regression model for logarithm of word duration.
- Transparent reporting form

Data availability

Data used in the dataset relate to the corpus gesproken nederlands. Information about this dataset can be found here: http://lands.let.ru.nl/cgn/. Access to the dataset can be requested here: https://taalmaterialen.ivdnt.org/download/tstc-corpus-gesproken-nederlands/. Data regarding the simulations in Figure 8 are based on data from Ten Oever & Sack (2015). As this data regards a closed database owned by Maastricht University it is not openly available. However, the data is available upon request without any restrictions via sanne.tenoever@mpi.nl or datamanagement-fpn@maastrichtuniversity.nl.

References

Arvaniti A. 2009. Rhythm, timing and the timing of rhythm. *Phonetica* **66**:46–63. DOI: https://doi.org/10.1159/000208930, PMID: 19390230

Assaneo MF, Rimmele JM, Sanz Perl Y, Poeppel D. 2021. Speaking rhythmically can shape hearing. *Nature Human Behaviour* **5**:71–82. DOI: https://doi.org/10.1038/s41562-020-00962-0, PMID: 33046860

Aubanel V, Schwartz J-L. 2020. The role of isochrony in speech perception in noise. *Scientific Reports* **10**:1–12. DOI: https://doi.org/10.1038/s41598-020-76594-1

Bahramisharif A, Jensen O, Jacobs J, Lisman J. 2018. Serial representation of items during working memory maintenance at letter-selective cortical sites. *PLOS Biology* **16**:e2003805. DOI: https://doi.org/10.1371/journal.pbio.2003805, PMID: 30110320

Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, Friston KJ. 2012. Canonical microcircuits for predictive coding. *Neuron* **76**:695–711. DOI: https://doi.org/10.1016/j.neuron.2012.10.038, PMID: 23177956

Beattie GW, Butterworth BL. 1979. Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech* **22**:201–211. DOI: https://doi.org/10.1177/002383097902200301
Bosker HR, Cooke M. 2018. Talkers produce more pronounced amplitude modulations when speaking in noise. *The Journal of the Acoustical Society of America* **143**:EL121–EL126. DOI: https://doi.org/10.1121/1.5024404, PMID: 29495684
Bosker HR, Kösem A. 2017. An entrained rhythm's frequency, not phase, influences temporal sampling of speech. Interspeech. DOI: https://doi.org/10.21437/Interspeech.2017-73
Bosker HR, Reinisch E. 2015. *Normalization for Speechrate in Native and Nonnative Speech. 18th International Congress of Phonetic Sciences (ICPhS 2015)*: International Phonetic Association.
Brennan JR, Martin AE. 2020. Phase synchronization varies systematically with linguistic structure composition. *Philosophical Transactions of the Royal Society B: Biological Sciences* **375**:20190305. DOI: https://doi.org/10.1098/rstb.2019.0305
Buzsáki G, Draguhn A. 2004. Neuronal oscillations in cortical networks. *Science* **304**:1926–1929. DOI: https://doi.org/10.1126/science.1099745, PMID: 15218136
Chater M. 2001. *Connectionist Psycholinguistics*. Greenwood Publishing Group.
Cumin D, Unsworth CP. 2007. Generalising the Kuramoto model for the study of neuronal synchronisation in the brain. *Physica D: Nonlinear Phenomena* **226**:181–196. DOI: https://doi.org/10.1016/j.physd.2006.12.004
Deacon D, Mehta A, Tinsley C, Nousak JM. 1995. Variation in the latencies and amplitudes of N400 and NA as a function of semantic priming. *Psychophysiology* **32**:560–570. DOI: https://doi.org/10.1111/j.1469-8986.1995.tb01232.x, PMID: 8524990
deen V, Kochs S, Smulders F, De Weerd P. 2017. Learned interval time facilitates associate memory retrieval. *Learn Memory* **24**:158–161. DOI: https://doi.org/10.1101/lm.044404.116
Di Liberto GM, O'Sullivan JA, Lalor EC. 2015. Low-Frequency cortical entrainment to speech reflects Phoneme-Level processing. *Current Biology* **25**:2457–2465. DOI: https://doi.org/10.1016/j.cub.2015.08.030, PMID: 26412129
Ding N, Patel AD, Chen L, Butler H, Luo C, Poeppel D. 2017. Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews* **81**:181–187. DOI: https://doi.org/10.1016/j.neubiorev.2017.02.011, PMID: 28212857
Doelling KB, Assaneo MF, Bevilacqua D, Pesaran B, Poeppel D. 2019. An oscillator model better predicts cortical entrainment to music. *PNAS* **116**:10113–10121. DOI: https://doi.org/10.1073/pnas.1816414116, PMID: 31019082
Doumas LA, Hummel JE, Sandhofer CM. 2008. A theory of the discovery and predication of relational concepts. *Psychological Review* **115**:1–43. DOI: https://doi.org/10.1037/0033-295X.115.1.1, PMID: 18211183
Doumas LA, Martin AE. 2018. Learning structured representations from experience. *Psychology of Learning and Motivation* **69**:165–203. DOI: https://doi.org/10.1016/BS.PLM.2018.10.002
Eagleman DM, Tse PU, Buonomano D, Janssen P, Nobre AC, Holcombe AO. 2005. Time and the brain: how subjective time relates to neural time. *Journal of Neuroscience* **25**:10369–10371. DOI: https://doi.org/10.1523/JNEUROSCI.3487-05.2005, PMID: 16280574
Eagleman DM. 2008. Human time perception and its illusions. *Current Opinion in Neurobiology* **18**:131–136. DOI: https://doi.org/10.1016/j.conb.2008.06.002, PMID: 18639634
Fernald A. 2000. Speech to infants as hyperspeech: knowledge-driven processes in early word recognition. *Phonetica* **57**:242–254. DOI: https://doi.org/10.1159/000028477, PMID: 10992144
Friederici AD. 2011. The brain basis of language processing: from structure to function. *Physiological Reviews* **91**:1357–1392. DOI: https://doi.org/10.1152/physrev.00006.2011, PMID: 22013214
Ghitza O. 2012. On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum. *Frontiers in Psychology* **3**:238. DOI: https://doi.org/10.3389/fpsyg.2012.00238, PMID: 22811672
Ghitza O. 2013. The theta-syllable: a unit of speech information defined by cortical function. *Frontiers in Psychology* **4**:138. DOI: https://doi.org/10.3389/fpsyg.2013.00138, PMID: 23519170
Ghitza O, Greenberg S. 2009. On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica* **66**:113–126. DOI: https://doi.org/10.1159/000208934, PMID: 19390234
Giraud AL, Poeppel D. 2012. Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience* **15**:511–517. DOI: https://doi.org/10.1038/nn.3063, PMID: 22426255
Guest O, Martin AE. 2021. How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science* **16**:789–802. DOI: https://doi.org/10.1177/1745691620970585, PMID: 33482070
Gwilliams L, King J-R, Marantz A, Poeppel D. 2020. Neural dynamics of phoneme sequencing in real speech jointly encode order and invariant content. *bioRxiv*. DOI: https://doi.org/10.1101/2020.04.04.025684
Hagoort P. 2017. The core and beyond in the language-ready brain. *Neuroscience & Biobehavioral Reviews* **81**:194–204. DOI: https://doi.org/10.1016/j.neubiorev.2017.01.048, PMID: 28193452
Hawkins S. 2014. Situational influences on rhythmicity in speech, music, and their interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**:20130398. DOI: https://doi.org/10.1098/rstb.2013.0398
Henry MJ, Obleser J. 2012. Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. *PNAS* **109**:20095–20100. DOI: https://doi.org/10.1073/pnas.1213390109, PMID: 23151506

Herrmann B, Henry MJ, Grigutsch M, Obleser J. 2013. Oscillatory phase dynamics in neural entrainment underpin illusory percepts of time. *Journal of Neuroscience* **33**:15799–15809. DOI: https://doi.org/10.1523/JNEUROSCI.1434-13.2013, PMID: 24089487

Jadoul Y, Ravignani A, Thompson B, Filippi P, de Boer B. 2016. Seeking temporal predictability in speech: comparing statistical approaches on 18 world languages. *Frontiers in Human Neuroscience* **10**:586. DOI: https://doi.org/10.3389/fnhum.2016.00586, PMID: 27994544

Jefferson G. 1990. List construction as a task and resource. *Interaction Competence* **63**:92. DOI: https://doi.org/10.1016/j.pragma.2006.07.008

Jensen O, Bonnefond M, VanRullen R. 2012. An oscillatory mechanism for prioritizing salient unattended stimuli. *Trends in Cognitive Sciences* **16**:200–206. DOI: https://doi.org/10.1016/j.tics.2012.03.002, PMID: 22436764

Jones MR, Boltz M. 1989. Dynamic attending and responses to time. *Psychological Review* **96**:459–491. DOI: https://doi.org/10.1037/0033-295X.96.3.459, PMID: 2756068

Kaufeld G, Bosker HR, Alday PM, Meyer AS, Martin AE. 2020a. Linguistic structure and meaning organize neural oscillations into a content-specific hierarchy. *bioRxiv*. DOI: https://doi.org/10.1101/2020.02.05.935676

Kaufeld G, Bosker HR, Ten Oever S, Alday PM, Meyer AS, Martin AE. 2020b. Linguistic structure and meaning organize neural oscillations into a Content-Specific hierarchy. *The Journal of Neuroscience* **40**:9467–9475. DOI: https://doi.org/10.1523/JNEUROSCI.0302-20.2020, PMID: 33097640

Kayser C, Montemurro MA, Logothetis NK, Panzeri S. 2009. Spike-phase coding boosts and stabilizes information carried by spatial and temporal spike patterns. *Neuron* **61**:597–608. DOI: https://doi.org/10.1016/j.neuron.2009.01.008, PMID: 19249279

Kayser SJ, McNair SW, Kayser C. 2016. Prestimulus influences on auditory perception from sensory representations and decision processes. *PNAS* **113**:4842–4847. DOI: https://doi.org/10.1073/pnas.1524087113, PMID: 27071110

Keitel A, Gross J, Kayser C. 2018. Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLOS Biology* **16**:e2004473. DOI: https://doi.org/10.1371/journal.pbio.2004473, PMID: 29529019

Kösem A, Basirat A, Azizi L, van Wassenhove V. 2016. High-frequency neural activity predicts word parsing in ambiguous speech streams. *Journal of Neurophysiology* **116**:2497–2512. DOI: https://doi.org/10.1152/jn.00074.2016, PMID: 27605528

Kösem A, Bosker HR, Takashima A, Meyer A, Jensen O, Hagoort P. 2018. Neural entrainment determines the words we hear. *Current Biology* **28**:2867–2875. DOI: https://doi.org/10.1016/j.cub.2018.07.023, PMID: 30197083

Lakatos P, Karmos G, Mehta AD, Ulbert I, Schroeder CE. 2008. Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science* **320**:110–113. DOI: https://doi.org/10.1126/science.1154735, PMID: 18388295

Large EW, Jones MR. 1999. The dynamics of attending: how people track time-varying events. *Psychological Review* **106**:119–159. DOI: https://doi.org/10.1037/0033-295X.106.1.119

Lau EF, Phillips C, Poeppel D. 2008. A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience* **9**:920–933. DOI: https://doi.org/10.1038/nrn2532, PMID: 19020511

Lehiste I. 1972. The timing of utterances and linguistic boundaries. *The Journal of the Acoustical Society of America* **51**:2018–2024. DOI: https://doi.org/10.1121/1.1913062

Lisman J. 2005. The theta/gamma discrete phase code occuring during the hippocampal phase precession may be a more general brain coding scheme. *Hippocampus* **15**:913–922. DOI: https://doi.org/10.1002/hipo.20121, PMID: 16161035

Lisman JE, Jensen O. 2013. The Theta-Gamma neural code. *Neuron* **77**:1002–1016. DOI: https://doi.org/10.1016/j.neuron.2013.03.007

Luo H, Tian X, Song K, Zhou K, Poeppel D. 2013. Neural response phase tracks how listeners learn new acoustic representations. *Current Biology* **23**:968–974. DOI: https://doi.org/10.1016/j.cub.2013.04.031, PMID: 23664974

Luo H, Poeppel D. 2007. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* **54**:1001–1010. DOI: https://doi.org/10.1016/j.neuron.2007.06.004, PMID: 17582338

Malhotra S, Cross RWA, van der Meer MAA. 2012. Theta phase precession beyond the Hippocampus. *Reviews in the Neurosciences* **23**:39–65. DOI: https://doi.org/10.1515/revneuro-2011-0064

Marslen-Wilson WD. 1987. Functional parallelism in spoken word-recognition. *Cognition* **25**:71–102. DOI: https://doi.org/10.1016/0010-0277(87)90005-9, PMID: 3581730

Martin AE. 2016. Language processing as cue integration: grounding the psychology of language in perception and neurophysiology. *Frontiers in Psychology* **7**:120. DOI: https://doi.org/10.3389/fpsyg.2016.00120, PMID: 26909051

Martin AE. 2020. A compositional neural architecture for language. *Journal of Cognitive Neuroscience* **32**:1407–1427. DOI: https://doi.org/10.1162/jocn_a_01552

Martin AE, Doumas LA. 2017. A mechanism for the cortical computation of hierarchical linguistic structure. *PLOS Biology* **15**:e2000663. DOI: https://doi.org/10.1371/journal.pbio.2000663, PMID: 28253256

Martin AE, Doumas LAA. 2019. Predicate learning in neural systems: using oscillations to discover latent structure. *Current Opinion in Behavioral Sciences* **29**:77–83. DOI: https://doi.org/10.1016/j.cobeha.2019.04.008

McClelland JL, Elman JL. 1986. The TRACE model of speech perception. *Cognitive Psychology* **18**:1–86. DOI: https://doi.org/10.1016/0010-0285(86)90015-0, PMID: 3753912

Mehta MR, Lee AK, Wilson MA. 2002. Role of experience and oscillations in transforming a rate code into a temporal code. *Nature* **417**:741–746. DOI: https://doi.org/10.1038/nature00807, PMID: 12066185

Meyer L. 2018. The neural oscillations of speech processing and language comprehension: state of the art and emerging mechanisms. *European Journal of Neuroscience* **48**:2609–2621. DOI: https://doi.org/10.1111/ejn.13748, PMID: 29055058

Meyer L, Sun Y, Martin AE. 2019. Synchronous, but not entrained: exogenous and endogenous cortical rhythms of speech and language processing. *Language, Cognition and Neuroscience* **35**:1089–1099. DOI: https://doi.org/10.1080/23273798.2019.1693050

Meyer L, Sun Y, Martin AE. 2020. "Entraining" to speech, generating language? *Language, Cognition and Neuroscience* **35**:1138–1148. DOI: https://doi.org/10.1080/23273798.2020.1827155

Michalareas G, Vezoli J, van Pelt S, Schoffelen JM, Kennedy H, Fries P. 2016. Alpha-Beta and gamma rhythms subserve feedback and feedforward influences among human visual cortical Areas. *Neuron* **89**:384–397. DOI: https://doi.org/10.1016/j.neuron.2015.12.018, PMID: 26777277

Monsell S, Doyle MC, Haggard PN. 1989. Effects of frequency on visual word recognition tasks: where are they? *Journal of Experimental Psychology: General* **118**:43–71. DOI: https://doi.org/10.1037/0096-3445.118.1.43

Monsell S. 1991. *The Nature and Locus of Word Frequency Effects in Reading*. Routledge.

Nieuwenhuijse A. 2018. Dutch Word2Vec Model. *GitHub*. 4014bf0. https://github.com/coosto/dutch-word-embeddings

Nieuwland MS. 2019. Do 'early' brain responses reveal word form prediction during language comprehension? A critical review. *Neuroscience & Biobehavioral Reviews* **96**:367–400. DOI: https://doi.org/10.1016/j.neubiorev.2018.11.019, PMID: 30621862

Nolan F, Jeon H-S. 2014. Speech rhythm: a metaphor? *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**:20130396. DOI: https://doi.org/10.1098/rstb.2013.0396

O'Keefe J, Recce ML. 1993. Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus* **3**:317–330. DOI: https://doi.org/10.1002/hipo.450030307, PMID: 8353611

O'Malley S, Besner D. 2008. Reading aloud: qualitative differences in the relation between stimulus quality and word frequency as a function of context. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **34**:1400–1411. DOI: https://doi.org/10.1037/a0013084

Obleser J, Kayser C. 2019. Neural entrainment and attentional selection in the listening brain. *Trends in Cognitive Sciences* **23**:913–926. DOI: https://doi.org/10.1016/j.tics.2019.08.004, PMID: 31606386

Panzeri S, Petersen RS, Schultz SR, Lebedev M, Diamond ME. 2001. The role of spike timing in the coding of stimulus location in rat somatosensory cortex. *Neuron* **29**:769–777. DOI: https://doi.org/10.1016/S0896-6273(01)00251-3, PMID: 11301035

Panzeri S, Macke JH, Gross J, Kayser C. 2015. Neural population coding: combining insights from microscopic and mass signals. *Trends in Cognitive Sciences* **19**:162–172. DOI: https://doi.org/10.1016/j.tics.2015.01.002, PMID: 25670005

Pariyadath V, Eagleman D. 2007. The effect of predictability on subjective duration. *PLOS ONE* **2**:e1264. DOI: https://doi.org/10.1371/journal.pone.0001264, PMID: 18043760

Peelle JE, Davis MH. 2012. Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology* **3**:320. DOI: https://doi.org/10.3389/fpsyg.2012.00320, PMID: 22973251

Pellegrino F, Coupé C. 2011. A cross-language perspective on speech information rate. *Language* **87**:539–558. DOI: https://doi.org/10.2307/23011654

Piantadosi ST. 2014. Zipf's word frequency law in natural language: a critical review and future directions. *Psychonomic Bulletin & Review* **21**:1112–1130. DOI: https://doi.org/10.3758/s13423-014-0585-6, PMID: 24664880

Pluymaekers M, Ernestus M, Baayen RH. 2005a. Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica* **62**:146–159. DOI: https://doi.org/10.1159/000090095, PMID: 16391500

Pluymaekers M, Ernestus M, Baayen RH. 2005b. Lexical frequency and acoustic reduction in spoken dutch. *The Journal of the Acoustical Society of America* **118**:2561–2569. DOI: https://doi.org/10.1121/1.2011150, PMID: 16266176

Poeppel D. 2003. The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication* **41**:245–255. DOI: https://doi.org/10.1016/S0167-6393(02)00107-3

Poeppel D, Assaneo MF. 2020. Speech rhythms and their neural foundations. *Nature Reviews Neuroscience* **21**:322–334. DOI: https://doi.org/10.1038/s41583-020-0304-4, PMID: 32376899

Powers DM. 1998. Editor applications and explanations of zipf's law. New Methods in Language Processing and Computational Natural Language Learning.

Reinisch E, Sjerps MJ. 2013. The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. *Journal of Phonetics* **41**:101–116. DOI: https://doi.org/10.1016/j.wocn.2013.01.002

Rimmele JM, Morillon B, Poeppel D, Arnal LH. 2018. Proactive sensing of periodic and aperiodic auditory patterns. *Trends in Cognitive Sciences* **22**:870–882. DOI: https://doi.org/10.1016/j.tics.2018.08.003, PMID: 30266147

Rosen S. 1992. Temporal information in speech: acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **336**:367–373. DOI: https://doi.org/10.1098/rstb.1992.0070, PMID: 1354376

Schroeder CE, Lakatos P. 2009. Low-frequency neuronal oscillations as instruments of sensory selection. *Trends in Neurosciences* **32**:9–18. DOI: https://doi.org/10.1016/j.tins.2008.09.012, PMID: 19012975

Ten Oever & Martin. 2021. STiMCON. *Software Heritage*. swh:1:rev:873a2bf5c79fe2f828e72e14ef74db409d387854. https://archive.softwareheritage.org/swh:1:dir:

cf831eabfe75473deb3aafac084e8af91398ae29;origin=https://github.com/sannetenoever/STiMCON;visit=swh:1:snp:fbce7be5ac6a1486f21dcc28e7a79b952d3e1c92;anchor=swh:1:rev:873a2bf5c79fe2f828e72e14ef74db409d387854

Ten Oever S, Sack AT, Wheat KL, Bien N, van Atteveldt N. 2013. Audio-visual onset differences are used to determine syllable identity for ambiguous audio-visual stimulus pairs. *Frontiers in Psychology* **4**:331. DOI: https://doi.org/10.3389/fpsyg.2013.00331, PMID: 23805110

Ten Oever S, Hausfeld L, Correia JM, Van Atteveldt N, Formisano E, Sack AT. 2016. A 7T fMRI study investigating the influence of oscillatory phase on syllable representations. *NeuroImage* **141**:1–9. DOI: https://doi.org/10.1016/j.neuroimage.2016.07.011, PMID: 27395392

Ten Oever S, Meierdierks T, Duecker F, De Graaf TA, Sack AT. 2020. Phase-Coded oscillatory ordering promotes the separation of closely matched representations to optimize perceptual discrimination. *iScience* **23**:101282. DOI: https://doi.org/10.1016/j.isci.2020.101282, PMID: 32604063

Ten Oever S, Sack AT. 2015. Oscillatory phase shapes syllable perception. *PNAS* **112**:15833–15837. DOI: https://doi.org/10.1073/pnas.1517519112, PMID: 26668393

Terao M, Watanabe J, Yagi A, Nishida S. 2008. Reduction of stimulus visibility compresses apparent time intervals. *Nature Neuroscience* **11**:541–542. DOI: https://doi.org/10.1038/nn.2111, PMID: 18408716

Thézé R, Giraud AL, Mégevand P. 2020. The phase of cortical oscillations determines the perceptual fate of visual cues in naturalistic audiovisual speech. *Science Advances* **6**:eabc6348. DOI: https://doi.org/10.1126/sciadv.abc6348, PMID: 33148648

Thompson SP, Newport EL. 2007. Statistical learning of syntax: the role of transitional probability. *Language Learning and Development* **3**:1–42. DOI: https://doi.org/10.1080/15475440709336999

Ulrich R, Nitschke J, Rammsayer T. 2006. Perceived duration of expected and unexpected stimuli. *Psychological Research Psychologische Forschung* **70**:77–87. DOI: https://doi.org/10.1007/s00426-004-0195-4, PMID: 15609031

Vroomen J, Keetels M. 2010. Perception of intersensory synchrony: a tutorial review. *Attention, Perception, & Psychophysics* **72**:871–884. DOI: https://doi.org/10.3758/APP.72.4.871, PMID: 20436185

Zuidema W. 2010. *A Syllable Frequency List for Dutch*. Taalportaal.