Isochronous is beautiful? Syllabic event detection in a neuro-inspired oscillatory model is facilitated by isochrony in speech

Mamady Nabé, Julien Diard, Jean-Luc Schwartz

Interspeech 2022
18-22 September 2022, Incheon, Korea

Isochronous is beautiful? Syllabic event detection in a neuro-inspired oscillatory model is facilitated by isochrony in speech

Mamady Nabé[1,2], *Julien Diard*[1], *Jean-Luc Schwartz*[2]

[1]Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LPNC, 38000 Grenoble, France
[2]Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

mamady.nabe@univ-grenoble-alpes.fr, julien.diard@univ-grenoble-alpes.fr,
jean-luc.schwartz@gipsa-lab.grenoble-inp.fr

Abstract

Oscillation-based neuro-computational models of speech perception are grounded in the capacity of human brain oscillations to track the speech signal. Consequently, one would expect this tracking to be more efficient for more regular signals. In this paper, we address the question of the contribution of isochrony to event detection by neuro-computational models of speech perception. We consider a simple model of event detection proposed in the literature, based on oscillatory processes driven by the acoustic envelope, that was previously shown to efficiently detect syllabic events in various languages. We first evaluate its performance in the detection of syllabic events for French, and show that "perceptual centers" associated to vowel onsets are more robustly detected than syllable onsets. Then we show that isochrony in natural speech improves the performance of event detection in the oscillatory model. We also evaluate the model's robustness to acoustic noise. Overall, these results show the importance of bottom-up resonance mechanism for event detection; however, they suggest that bottom-up processing of acoustic envelope is not able to perfectly detect events relevant to speech temporal segmentation, highlighting the potential and complementary role of top-down, predictive knowledge.

Index Terms: speech perception, oscillatory-based model, neuro-inspired computational model, syllable segmentation, event detection, isochronous speech, P-center

1. Introduction

Speech perception involves multiple hierarchical levels of processing, to go from the acoustic, continuous signal to speech unit identification, to, ultimately, the perceived meaning of utterances. Classical psycholinguistic models inspired by interaction-activation processes such as TRACE or SHORTLIST [1, 2, 3, 4], as well as automatic speech recognition (ASR) models such as Hidden Markov Models (HMMs) or Deep Neural Networks (DNNS) [5, 6, 7], directly decode the speech continuous stream from spectro-temporal information through a battery of computational processes associating phonetic-prosodic, lexical and syntactic-semantic knowledge. However, recent studies in speech neuroscience focus on cognitive processes that appear crucial for speech perception, and that perform temporal segmentation, that is, identifying in the speech signal temporally relevant events (e.g., syllabic boundaries). Through synchronization processes between different populations of neurons operating in different frequency bands, typically in the gamma band (40–100 Hz) for acoustic spectro-temporal analysis, in the theta band (3–8 Hz) for syllabic segmentation, and in the delta band (1–2 Hz) for rhythmic/syntactic binding, the human brain would exploit neuronal oscillations to perform this temporal segmentation of incoming acoustic signals [8, 9, 10].

An important consequence of the assumption that oscillatory processes play a role in syllabic event detection is that regular sequences of such events should be better able to evoke resonance phenomena in these oscillatory processes, and hence result in stronger outputs and possibly better detection. Therefore, in par with these oscillatory principles, it could be predicted that the syllabic segmentation of speech should be easier when the distribution of syllabic durations is rather isochronous than when it is not. Indeed, a recent study [11] studied the intelligibility of speech in noise in two languages differing by their rhythmic properties, i.e. French, a syllable-timed language, and English, a stress-timed language. The authors analyzed the departure from syllabic isochrony among the presented sentences in both languages, and showed that more syllabic-isochronous sentences were better decoded than non-isochronous ones. A possible interpretation is that syllabic isochrony enhanced acoustic event detection, which in turn enhanced comprehension.

If indeed cortical activity can entrain more or less to regular features of the speech input envelope, the question remains of the nature of events it tracks. Various proposals emerge in the literature. The first one is classically focused on the search for syllable boundaries which roughly correspond to energy valleys/troughs. Neurophysiological data and models suggest an alternative, considering instead energy peaks that correspond globally to vowel nuclei of syllables [12, 13, 14], or to the so-called P-centers, corresponding to the perceptual center of the syllabic units [15, 16] associated to the peak in envelope intensity increase at the vowel onset. Indeed, it has been recently shown that the brain neural activity is more robust to the tracking of P-centers than to the syllable boundaries [17]. Hence, the question of the relative efficiency of neural oscillatory processes to detect P-centers rather than syllabic onsets comes as an additional question of importance in the study of neural oscillatory speech segmentation processes.

To address these questions, various models of syllabic parsing based on neural oscillations in the cortex at the theta rhythm (3–8 Hz) have been proposed in the literature [18, 19, 20, 21]. While some of these neuro-computational models exploit sophisticated realistic neuronal processing principles, sometimes even at the spike level of representation, we focus here on the model developed by Räsänen, Doyle and Frank [20] which stands out as the simplest one, operating on simple processes of envelope detection and linear second-order oscillators. In the remainder, we refer to this model as the RDF model, after its authors. The RDF model has been shown to efficiently extract syllable onsets in various languages, exploiting realistic speech corpora. However, it is yet unclear how such an oscillatory-

10.21437/Interspeech.2022-10426

based model would perform relative to the isochrony of speech signals, and whether it is specific to syllable onset detection, or could be extended to other syllabic events such as P-centers.

In this paper, we evaluate the RDF model by using it for syllabic boundary detection on a French corpus, to widen its evaluation set and assess its generalizability, and extending it to the detection of P-centers on the same French corpus, to assess its performance relative to the nature of syllabic events. We then analyze the model's behavior with respect to the departure of natural speech from isochrony, to assess whether the model yields better robustness for more isochronous inputs. We also take the opportunity to evaluate the model's robustness to noise.

2. Methods and Materials

2.1. Oscillatory model of syllable boundary detection

The RDF model is a neuro-computational model originally developed to study the pre-linguistic segmentation of syllables (that is, "accessible to an infant with no phonological or lexical knowledge"). The RDF model implements neural tracking based on an oscillatory system driven by energy fluctuations in the speech signal [13, 18, 22, 23]. Starting from the speech signal, a set of signal processing techniques are applied in order to obtain an estimate of the sonority of the signal. First, Gammatone filter-banks [24, 25] are applied to the speech input to get the amplitude envelope in 20 frequency bands. Their outputs are then low-pass filtered and down-sampled to have an overall sampling rate of 1,000 Hz. Each envelope of each frequency band is then passed to a harmonic oscillator which resonates at a central frequency f_0 within a bandwidth Δf. Together, these define the oscillator Q factor, $Q = f_0/\Delta f$. Finally, the N most energetic outputs (N, in the following, is set to 8, as in [20]) are combined by taking the sum of the logarithms of the amplitudes to obtain the sonority output; the final values are normalized between 0 and 1 over the stimulus duration. The resulting sonority function can be used in various ways, to identify speech relevant events.

In their work, RDF used the oscillatory model output to search for syllabic boundaries in the speech signal. To do so, they identified local minima (valleys) in the sonority output. After optimizing the model parameters, they evaluated its performance on various languages, namely Finnish, Estonian and English, and they showed that, globally, the model performs well for the three tested languages (see evaluation criterion later) on various speech corpora.

2.2. Adaptation of the RDF model to detect P-centers

P-centers correspond to "psychological moment of occurrence" of syllables [15, 26]. A precise acoustical landmark that corresponds to P-centers is still lacking; however, they are classically determined by looking at peaks of energy increase of the speech envelope [27, 28]. We therefore extended the RDF model to detect peaks in the first-order derivative of the sonority output. In the following, we will interpret these events as corresponding to the detection of P-centers.

2.3. Experimental corpus

In all the experiments, we used the *Fharvard* corpus, which is a resource equivalent to the Harvard corpus, but in French [29] [1]. It consists of phonemically-balanced natural spoken French sentences uttered by a male speaker. The initial dataset contains 700 sentences sharing similar structure. In this paper, we only used a subset of the overall dataset that was fully annotated by the original authors at various levels, namely at the word, syllable, P-center and phoneme levels. It amounts to 177 sentences composed of multi-syllabic words with a total of 646 distinct syllables. Figure 1 shows the distributions of syllable duration in the corpus.

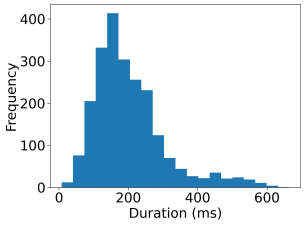

Figure 1: *Histogram of syllable duration in the Fharvard corpus [29] (mean is 203 ms, median is 180 ms, mode is 154 ms).*

To evaluate the model's robustness to noise, we added white Gaussian noise to the initial speech data, with varying signal to noise (SNR) ratio from -30 dB (very noisy) to 30 dB (almost noise free) by steps of 10 dB (totalling 7 SNR values). More realistic speech maskers such as long-term average speech spectrum noise [30] or multi-talker babble noise [31] may be used in future studies.

2.4. Performance measure

In order to evaluate performance, we define a performance measure indicative of the quality of event detection. For this purpose, we use the F-score measure [32, 33]. It is a trade-off between precision P (the proportion of events predicted by the algorithm that corresponds to real events) and recall R (the proportion of real events correctly predicted by the algorithm). It is calculated by $F = 2PR/(P + R)$.

In practice, we consider that a real event is correctly predicted by the model if it has detected an event in a time window of 50 ms around the real event. We calculate the performance measure for all the detected events by the model, adding to these the stimulus onset and offset events, which are supposed to be detected prior to the analysis by the RDF model itself. This configuration of performance evaluation is identical to the one used by RDF [20].

2.5. Parameter calibration

The RDF model has four free parameters, that require calibration to ensure optimal use of the oscillator algorithm. To obtain optimal values for each of these parameters, we performed a search on a predefined grid of values. To perform calibration, we optimised performance on a training dataset with 100 audio files within the 177 available ones, while all experimental results provided below were obtained from the remaining 77 sentences in the test set. We now recall the model parameters, and define our 4-dimensional calibration grid.

The first parameter is the central frequency f_0, that is, the resonant frequency of the oscillator, which varies in the theta frequency band, and is usually speaker dependent. We considered 7 calibration values (from 5 to 8 Hz with a .5 Hz step).

The second parameter is the quality factor Q, a function of the central frequency and the bandwidth of the oscillator $Q = f_0/\Delta f$. It measures the damping rate of the oscillator.

[1] Found online at `https://zenodo.org/record/1462854#.YitevozMLm4`

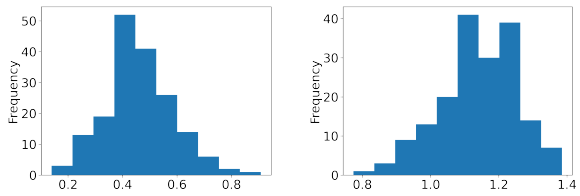

Figure 2: *Histograms of temporal distortion values in the Fharvard corpus [29], with respect to P-centers (left) and with respect to syllabic boundaries (right). The lower the value, the more isochronous the sentence.*

A notable value is $Q = .5$, for which the oscillator is critically damped, so that the oscillator would follow the envelope of the signal as closely as possible. For larger values of Q ($Q > .5$), under-damped oscillator), the oscillator resonates more around its central frequency, with a slower decay of its amplitude, even if it is no longer excited by a real signal. For smaller values of Q ($Q < .5$), over-damped oscillator), the oscillator performs more temporal smoothing, with little dependence on its central frequency. We considered, for calibration, an empirically defined set of 21 possible values for parameter Q: .15, .25, .5, .75, from 1 to 1.9 with a .1 step, and from 2 to 5 with a .5 step.

The third parameter is the minimum detection threshold thr, that is, the minimal difference between a local extremum and neighbour extrema enabling to consider the local extremum as meaningful. We considered 3 possible threshold values: .01, .025 and .5.

The fourth parameter is a fixed delay del, to shift all detected events, so as to mitigate artifacts introduced by signal processing techniques, in particular delays due to smoothing, filtering and windowing operations. For syllabic boundary detection, we considered 15 possible values (0 to 70 ms with a step of 5 ms); for P-center detection, we considered 7 possible values (0 to 30 ms, step of 5 ms).

2.6. Temporal distortion metric

To characterize the departure from isochrony in speech signals, we use a previously introduced temporal distortion metric noted δ [11]. It is computed for a given reference time series t (the initial temporal event series) which is transformed into a target time series t' (here an hypothetical isochronous time series with the same number of events) as the following: $\delta = \sqrt{\frac{\sum_{i=1}^{N}(\log \tau_i)^2 d_i}{\sum_{i=1}^{N} d_i}}$, with d the duration between successive reference events ($d_i = t_{i+1} - t_i$), τ_i the time-scale factor between the reference and target time series: $\tau_i = d'_i/d_i$ where $d'_i = (t_N - t_1)/(N-1)$. The lower δ is, the more a sequence of events is isochronous, that is, regularly spaced; on the contrary, the higher δ is, the more temporally distorted the sequence of events is.

We calculated temporal distortion both for P-centers and for syllable boundaries. The distributions of P-center and syllable distortion values for the 177 sentences in the experimental corpus are shown in Figure 2.

3. Results

3.1. Performance on syllabic event detection in French

Table 1 provides the optimal values, resulting from calibration, for model parameters, for both P-center and syllable boundary detection tasks. To recall, calibration was performed on the training set. We observe that optimal parameters, both for P-center and syllable boundary detection, correspond to under-damped oscillators ($Q = 1.4$ and 1.9, respectively). Interestingly, the optimal f_0 value at 7 Hz for syllable boundary detection is higher than the inverse value of the mean syllable duration (mean syllable duration is 203 ms, inverse is 4.9 Hz). This is also the case for all simulations in [20]. However, it is actually close to the inverse of the mode or of the median of the asymmetric distribution of syllable duration in Figure 1, which suggests that this statistic could better describe the overall speech rate in the corpus.

Table 1: *Parameter values resulting from calibration on the training set, and resulting F-scores on the test set.*

	P-centers	Syllable boundaries
f_0 (Hz)	6.5	7
Q	1.4	1.9
thr	0.025	0.01
del (ms)	0	55
F-score	.89	.75

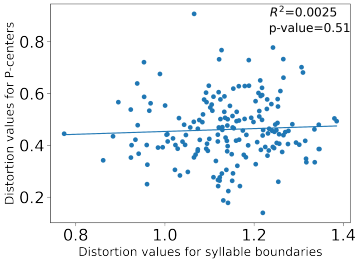

Figure 3: *Correlation between distortion values δ computed with respect to syllabic boundaries (x-axis) and P-centers (y-axis), for the 177 sentences of the experimental corpus. Linear regression (solid line) and corresponding squared correlation coefficient R^2 are indicated in the plot.*

Table 1 also reports detection performance on the test set for both tasks. Performance for syllable boundary detection is measured by an overall F-score of .75, which is comparable to previous experimental results in Finnish, Estonian and English [20]. In contrast, performance is quite higher for P-center detection, with an overall F-score of .89.

3.2. Role of isochrony in event detection

3.2.1. Relation between isochrony in the distribution of syllabic boundaries and P-centers

Figure 3 shows the correlation between lack of isochrony for syllabic boundaries and P-centers, for all sentences in the experimental corpus. We observe that there is no significant correlation: sentences with low distortion to synchrony in syllabic boundaries may have large distortion for P-centers, and vice-versa (Pearson correlation coefficient $R = 0.05$, p-value $p = 0.51$). In the following, we use only distortion to synchrony computed over the distribution of P-centers, in line with the experimental study by Aubanel & Schwartz [11].

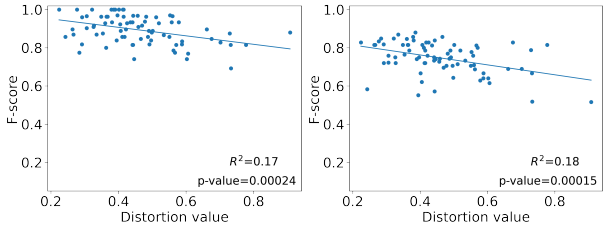

Figure 4: *Event detection performance (F-scores, y-axis) against P-center temporal distortion (δ, x-axis), for P-center detection (left) and syllable boundary detection (right). Linear regressions (solid lines) and corresponding squared correlation coefficients R^2 are indicated in the plots.*

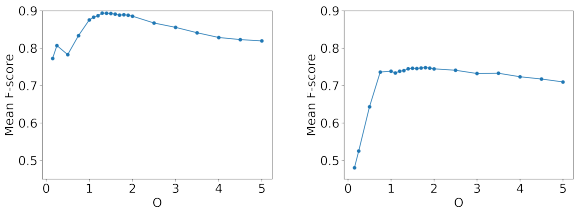

Figure 5: *Event detection performance (mean F-scores, y-axis) against the Q parameter value (x-axis), for P-center detection (left) and syllable boundary detection (right).*

3.2.2. Relation between distortion to P-center isochrony and event detection

Figure 4 shows the variations of event detection performance as a function of distortion to P-center isochrony, for P-center detection (left) and syllable boundary detection (right). We observe that for both P-center and syllable boundary detection, there is a statistically significant negative correlation between model performance and temporal distortion. In other words, model performance is higher, and events are better identified, when temporal distortion is small, that is to say, for natural sentences which happen to be more isochronous.

3.2.3. Role of the resonance factor in event detection

Figure 5 shows event detection performance as a function of the Q factor when all other model parameters are fixed, for P-centers (left) or syllable boundaries (right). Strikingly, the best performance is obtained for resonant systems with Q values much larger than the so-called critical damping value $Q = .5$ which corresponds to a system that essentially tracks the acoustic envelope with no additional resonance process. While the optimal value for the Q factor is similar for P-centers and syllable boundaries in the $1.2 - 1.5$ range, the adequate range is rather restricted for P-centers, with quasi optimal values between 1.1 and 1.8 and then a rapid decrease for too resonant systems; in contrast, a large range of Q values above 0.75 are adequate for syllable boundary detection, although detection performance is lower overall.

3.3. Event detection in noise

Figure 6 shows how model performance varies as a function of the signal to noise ratio (SNR). The RDF model appears to be rather robust to noise, with its performance almost unchanged up to a rather large level of noise (SNR at 0 dB), with performance sharply decreasing for lower values of SNR.

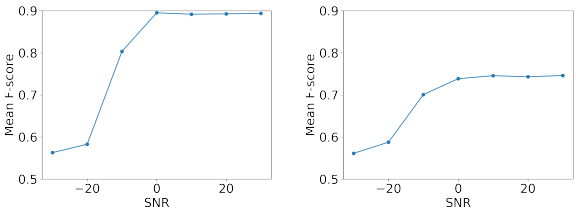

Figure 6: *Model performance (mean F-score, y-axis) with respect to the noise level (SNR, x-axis) for P-center detection (left) and for syllable boundary detection (right).*

4. Conclusion & Discussion

In this paper, we have evaluated the RDF oscillatory model of event detection [20] on a French corpus, and shown that it performs as well as previous evaluations on other languages.

Importantly, our results point to the role of resonance mechanisms in this process. Indeed, it appears that (1) the system performs better for resonant than for non-resonant characteristics of the proposed algorithm (see Figure 5) and (2) acoustic speech signals with higher inter-P-center isochrony lead to better event detection (see Figure 4). Furthermore, the detection process based on a resonant response to envelope modulations appears more efficient to detect P-centers than syllabic onsets (see Table 1). This is likely due to the fact that P-centers are more robust events within the speech envelope dynamics. It could lead to propose segmentation algorithms involving P-center detection as a complement signal to syllable boundary detection: although P-centers are not systematically related to syllable onsets (see Figure 3), P-center detection is a likely signal that a syllable boundary preceded, and was possibly missed.

The event detection system of the RDF model appears rather robust in acoustic noise. Still, our study, in line with results from previous experiments, suggests that performance is far from perfect (with 23 % missed events for syllabic onsets and 11 % for P-centers) without noise, and rapidly degraded for noise at SNR values under 0 dB. This suggests a potential role for top-down processes, exploiting statistics of sentence rhythms in relation to lexical, syntactic and prosodic knowledge. In their study on comprehension of speech in noise, Aubanel & Schwartz [11] showed that, while natural isochrony improved comprehension, anisochronous speech re-timed to become more isochronous is actually less well perceived, which points to the role of top-down predictive processes in speech segmentation. This is the core of the COSMO-Onset model we have previously developed [34] to model how bottom-up and top-down information could be combined for speech syllabic segmentation. The present study provides an important baseline: the RDF model is a purely bottom-up, signal driven event detection model, and current works aims at complementing it with top-down knowledge for syllabic event detection.

5. Acknowledgements

This work is supported by the French National Research Agency in the framework of the Investissements d'avenir program (ANR-15-IDEX-02; Ph.D. grant to MN from Université Grenoble Alpes ISP project Bio-Bayes Predictions). Authors also acknowledge additional support by the Auvergne-Rhône-Alpes (AURA) Region (PAI-19-008112-01 grant). This work has also been partially supported by the Multidisciplinary Institute of AI (MIAI) @ Grenoble Alpes (ANR-19-P3 IA-0003).

6. References

[1] J. L. McClelland and D. E. Rumelhart, "An interactive activation model of context effects in letter perception: I. an account of basic findings." *Psychological review*, vol. 88, no. 5, p. 375, 1981.

[2] J. L. McClelland and J. L. Elman, "The TRACE model of speech perception," *Cognitive Psychology*, vol. 18, no. 1, pp. 1–86, 1986.

[3] D. Norris, "Shortlist: A connectionist model of continuous speech recognition," *Cognition*, vol. 52, no. 3, pp. 189–234, 1994.

[4] D. Norris and J. M. McQueen, "Shortlist b: a bayesian model of continuous speech recognition." *Psychological review*, vol. 115, no. 2, p. 357, 2008.

[5] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[6] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The htk book," *Cambridge university engineering department*, vol. 3, no. 175, p. 12, 2002.

[7] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber, and X. Alameda-Pineda, "Dynamical variational autoencoders: A comprehensive review," *Foundations and Trends in Machine Learning*, vol. 15, no. 1–2, pp. 1–175, 2021.

[8] G. Buzsáki and A. Draguhn, "Neuronal oscillations in cortical networks," *Science*, vol. 304, no. 5679, pp. 1926–1929, 2004.

[9] G. Buzsaki, *Rhythms of the Brain*. Oxford University Press, 2006.

[10] P. Fries, "Rhythms for cognition: communication through coherence," *Neuron*, vol. 88, no. 1, pp. 220–235, 2015.

[11] V. Aubanel and J.-L. Schwartz, "The role of isochrony in speech perception in noise," *Scientific Reports (Nature Publisher Group)*, vol. 10, no. 1, 2020.

[12] O. Ghitza, "On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum," *Frontiers in psychology*, vol. 3, p. 238, 2012.

[13] N. Ding and J. Z. Simon, "Cortical entrainment to continuous speech: functional roles and interpretations," *Frontiers in human neuroscience*, vol. 8, p. 311, 2014.

[14] K. B. Doelling, M. F. Assaneo, D. Bevilacqua, B. Pesaran, and D. Poeppel, "An oscillator model better predicts cortical entrainment to music," *Proceedings of the National Academy of Sciences*, vol. 116, no. 20, pp. 10 113–10 121, 2019.

[15] J. Morton, S. Marcus, and C. Frankish, "Perceptual centers (p-centers)." *Psychological review*, vol. 83, no. 5, p. 405, 1976.

[16] S. Scott and P. Howell, "Perceptual centers in speech: An acoustic analysis," *The Journal of the Acoustical Society of America*, vol. 92, no. 4, pp. 2443–2443, 1992.

[17] Y. Oganian and E. F. Chang, "A speech envelope landmark for syllable encoding in human superior temporal gyrus," *Science advances*, vol. 5, no. 11, p. eaay6279, 2019.

[18] O. Ghitza, "Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm," *Frontiers in Psychology*, vol. 2, p. 130, 2011.

[19] A. Hyafil, L. Fontolan, C. Kabdebon, B. Gutkin, and A.-L. Giraud, "Speech encoding by coupled cortical theta and gamma oscillations," *eLife*, vol. 4, p. e06213, 2015.

[20] O. Räsänen, G. Doyle, and M. C. Frank, "Pre-linguistic segmentation of speech into syllable-like units," *Cognition*, vol. 171, pp. 130–150, 2018.

[21] S. Hovsepyan, I. Olasagasti, and A.-L. Giraud, "Combining predictive coding and neural oscillations enables online syllable recognition in natural speech," *Nature Communications*, vol. 11, no. 1, pp. 1–12, 2020.

[22] A.-L. Giraud and D. Poeppel, "Cortical oscillations and speech processing: Emerging computational principles and operations," *Nature Neuroscience*, vol. 15, no. 4, p. 511, 2012.

[23] L. H. Arnal, "Predicting "when" using the motor system's beta-band oscillations," *Frontiers in Human Neuroscience*, vol. 6, p. 225, 2012.

[24] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, vol. 2, no. 7, 1987.

[25] J. Holdsworth, I. Nimmo-Smith, R. Patterson, and P. Rice, "Implementing a gammatone filter bank," *Annex C of the SVOS Final Report: Part A: The Auditory Filterbank*, vol. 1, pp. 1–5, 1988.

[26] A. Strauß and J.-L. Schwartz, "The syllable in the light of motor skills and neural oscillations," *Language, Cognition and Neuroscience*, vol. 32, no. 5, pp. 562–569, 2017.

[27] S. M. Marcus, "Acoustic determinants of perceptual center (p-center) location," *Perception & psychophysics*, vol. 30, no. 3, pp. 247–256, 1981.

[28] A. D. Patel, A. Löfqvist, and W. Naito, "The acoustics and kinematics of regularly timed speech: a database and method for the study of the p-center problem," in *Proceedings of the 14th international congress of phonetic sciences*, vol. 1. Linguistics Department, University of California Berkeley, 1999, pp. 405–408.

[29] V. Aubanel, C. Bayard, A. Strauß, and J.-L. Schwartz, "The fharvard corpus: A phonemically-balanced french sentence resource for audiology and intelligibility research," *Speech Communication*, vol. 124, pp. 68–74, 2020.

[30] R. Plomp and A. Mimpen, "Speech-reception threshold for sentences as a function of age and noise level," *The Journal of the Acoustical Society of America*, vol. 66, no. 5, pp. 1333–1342, 1979.

[31] K. J. Van Engen and A. R. Bradlow, "Sentence recognition in native-and foreign-language multi-talker background noise," *The Journal of the Acoustical Society of America*, vol. 121, no. 1, pp. 519–526, 2007.

[32] N. Chinchor, "Muc-4 evaluation metrics," *Proceedings of the Fourth Message Understanding Conference*, pp. 22–29, 1992.

[33] Y. Sasaki *et al.*, "The truth of the f-measure. 2007," *Manchester: School of Computer Science, University of Manchester*, 2007.

[34] M. Nabé, J.-L. Schwartz, and J. Diard, "Cosmo-onset: A neurally-inspired computational model of spoken word recognition, combining top-down prediction and bottom-up detection of syllabic onsets," *Frontiers in Systems Neuroscience*, p. 75, 2021.