How speech rate, syllabic complexity and diversity affect the emergence of speech rhythm in speeded syllable repetition

Leonardo Lancia[1], Jinyu Li[2], Cécile Fougeron[3]
[1]Laboratoire Parole et Langage (LPL), [2,3]Laboratoire de Phonétique et Phonologie (LPP)
[1]leonardo.lancia@sorbonne-nouvelle.fr, [2]jinyu.li@sorbonne.nouvelle.fr, [3]cecile.fougeron@sorbonne-nouvelle.fr

ABSTRACT

Speech rhythm has been related to the grouping of nearby syllables, which for several authors depends on features of the language's phonology. However, most evidence of stable relations between rhythmic features and structural features relies on perceptual judgments, which are affected by non-phonological factors like speech rate. We studied how speaker-specific performances and syllable-complexity affect the rhythmic structure emerging in the repetition of the same syllable or of different syllables at maximal speed for 177 French speakers. We found that the degree of coordination between the production of syllables and the production of word-level prominence, determining syllables grouping in a way that reflects French metrical structure, depends on syllabic complexity and on the presence of different syllables in the same sequence. Moreover, speakers reaching higher rates coordinate better syllables and supra-syllabic prominence, supporting the idea that speech rhythm emerges from the interaction between structural features and general coordination principles.

Keywords: Speech rhythm, amplitude modulation, phase coordination, diadochokinetic task.

1. INTRODUCTION

Although there is a lack of agreement on the way to define and characterize speech rhythm, it is unequivocally related to the grouping of phonetic events into larger units. More specifically several authors (see [1 - 4], for early proposals) suggest that speech rhythm is mainly characterized by the grouping of syllables into word-sized structures (e.g. prosodic words, [4]). In a few studies, this theoretical stance led to characterize speech rhythm by measuring the degree of coordination between syllabic activity and activity related to the production of suprasyllabic prominence at the level of the word, as captured by band-pass filtering the amplitude modulation signal at the appropriate frequencies [5, 6]. In this way, it was possible to quantify the similarity between the rhythmic patterns observed in different languages and to reproduce the patterns of cross linguistic similarity that were obtained via human perceptual judgments [6].

In the present work, we adopt this approach to study the rhythmic features emerging in speeded syllable repetition tasks. This choice is motivated by the observation that, when speakers repeatedly produce a sequence of speech sounds at a fast speech rate, we often observe changes in the way speakers coordinate in time consecutive phonetic events both inside a syllable [7, 8] and between consecutive syllables [9, 10]. These results suggest that speech rate pushes the sensorimotor system to explore coordinative strategies underlying different groupings of the speech stream by jumping from one possible coordinative strategy to the other in function of their stability (i.e. unstable coordinative strategies are unlikely to characterize the speaker's behaviour during long time intervals). Speeded repetition paradigms thus provide a powerful experimental tool permitting to study how the relative stability of different groupings of consecutive speech sounds is affected by the interplay between controllable factors. The aim of this study is to investigate how a number of factors that are known to change between rhythmically dissimilar languages affect the degree of coordination between syllables and suprasyllabic prominence in French (in which suprasyllabic prominence depends on accents production). The factors considered are the type of syllables to be repeated, in terms of complexity (CV vs. CCV) and diversity (same syllable vs. different syllables in sequence) and the speech rate reached by the speaker in the task.

2. MATERIAL AND METHODS

The productions of 177 French speakers (from France and Switzerland) in a diadochokinesic (DDK) task were extracted from the MonPaGe_HA database [11]. Speakers were instructed to repeat as fast and as accurately as possible sequences of syllables in a continuous manner in a single breath group. Sequences to be repeated varied in complexity and diversity: (a) repetitive CV syllables involving different places of articulation: /bababa/, /dedede/, /gogogo/, (b) repetitive CCV syllables: /klaklakla/, /tʁatʁatʁa/, (c) sequences of different CV syllables: /badego/.

2.1. Parameterization of the filter's coefficients

To extract the acoustic amplitude modulation (henceforth AM(k) with k belonging to {1,…,K} and K being its length in number of samples) from each sequence of uninterrupted speech, we submit the acoustic signal to the Hilbert transform and compute its amplitude. By band-pass filtering the AM signal we capture its syllabic and supra-syllabic components (henceforth respectively referred to as *syllAM*(k) and *accAM*(k), see Figure 2). Usually, the filter cut-off frequencies are set at values coherent with oscillatory frequencies of ca. 5Hz (for *syllAM*) and 2Hz (for *accAM*) based on empirical observations (e.g., [12]) of cross-linguistic data and on theoretical considerations (c.f. [13]). However, such canonical cut-off frequencies reflect average tendencies and are not appropriate to model deviant behaviours such as that observed in diadochokinetic tasks (in which speakers easily reach speech rates higher than 7hz). Therefore, we adapted the cut-off frequencies to the analysed data. To do that, for each experimental trial, we set the three required cut-off frequencies at the midpoints between the following rates: segments rate, syllables rate, accents rate and the rate of the fastest rhythmic component slower than the accentual one (capturing, for example, modulation of activity due to the production of sentence-level features).

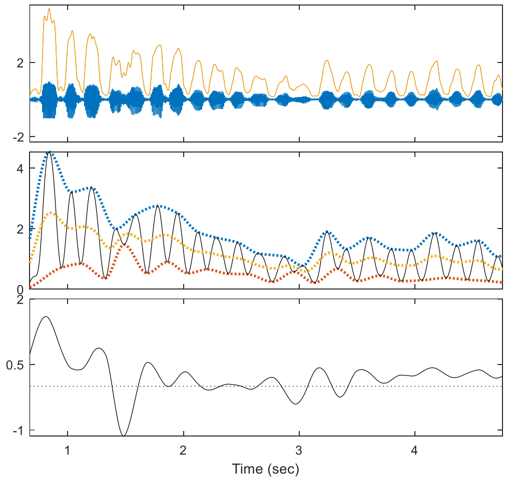

Figure 1: Derivation of accent rate. Topmost panel: waveform and its acoustic amplitude (red). Middle panel, continuous line: *lowPassAM*; dotted lines: maximum minimum and mean envelopes. Bottommost panel: syllabic cycles' extension as derived from *lowPassAM*'s envelopes.

Thanks to a manual segmentation of the syllabic boundaries available for our data, we could readily estimate the syllable and segment rates (via the number of surface segments per syllable). The procedure adopted to determine the accent rate is illustrated in Figure 1. Firstly, we low-pass filter the amplitude modulation signal with a cut-off frequency located halfway between the syllable and segment rates. This operation produces a signal expected to contain both the syllabic and suprasyllabic components of the acoustic energy signal (henceforth *lowPassAM*, see continuous line of the middle panel in Fig. 1). Our modelling strategy is motivated by the observation that speech activity displays a pulse-like shape and by the hypothesis that these pulses are mainly associated with syllables and modulated by one or more levels of suprasyllabic prominence.

To capture the changes of the pulses' amplitude due to suprasyllabic activity, we identify all the peaks and valleys in the *lowPassAM* signal. We then obtain the upper and lower envelope of *lowPassAM* by interpolating separately the values of peaks and valleys. The time-varying distance between these two envelopes minus their time-varying mean (bottommost panel in Fig. 1) is considered as an estimate of the syllabic cycles' extension which, according to our hypothesis, reflects suprasyllabic prominence (accents production). Therefore, we estimate the accent rate by dividing the number of peaks in the obtained signal by the duration of the considered time interval.

The rate of the fastest rhythmic component slower than the accentual component is estimated by submitting the estimated extension of the syllabic cycles to Empirical Mode Decomposition (EMD, [14]). This method returns a small number of independent oscillatory components that summed produce the original signal. The first component captures the fastest oscillation present in the input, which in the case of the signal representing the extension of the syllabic cycles (bottom panel in Fig. 1) is due to accent production. Therefore, by counting the peaks in the second component obtained through EMD, we have an estimate of the fastest possible component slower than the accent rate.

2.2. PLV as an estimate of syllabic and supra-syllabic coordination

The temporal coordination between two signals can be measured by computing the phase locking value (PLV, [15]), inversely related to the variability of the lag between the positions of the two signals in their cycles (as represented by their instantaneous phase values $\phi(\mathrm{k})_{syll}$ and $\phi(\mathrm{k})_{stress}$. Instantaneous phase signals are computed via the application of the Hilbert transform to *syllAM(k)* and *accAM(k)* signals after normalizing their cycles' amplitudes so that each cycle oscillates between -1 and 1 (to meet the requirements of the Hilbert transform [16], see the two bottommost panels of Figure 2).

These values permit computing at each analysis frame the generalized phase difference:

$$(1)\quad \Delta\Phi(\mathrm{k})_{m,n} = m \times \phi(\mathrm{k})_{syll} - n \times \phi(\mathrm{k})_{stress},$$

where m and n are two integers such that $m \times \omega(\mathrm{k})_{syll} = n \times \omega(\mathrm{k})_{stress}$, and $\omega(\mathrm{k})_{syll}$ and $\omega(\mathrm{k})_{stress}$ are the two frequencies of *syllAM* and *accAM* at frame *k*. The PLV corresponds to the first Fourier moment of the distribution of generalized phase differences observed in a given time window.

$$(2) \quad PLV = \langle \cos \Delta\Phi(\mathrm{k})_{m,n,} \rangle_k^2 + \langle \sin \Delta\Phi(\mathrm{k})_{m,n} \rangle_k^2$$

where the hooks indicate averaging over the time frames (indexed by *k*) belonging to the considered time window. A window width of 4 seconds is used, while the temporal resolution of the phase signals was set to 1ms. The window duration was chosen because it is the minimal duration of an uninterrupted chunk in our data base. For the same reason, we analysed only the first time-window of each recording. The values of *m* and *n* were determined separately for each analysis window by following [15], who compute one PLV for each combination of possible *m* and *n* values (where *m* and *n* can be chosen among the integer ranging from 1 to 10) and choose the combination that gives the highest PLV.

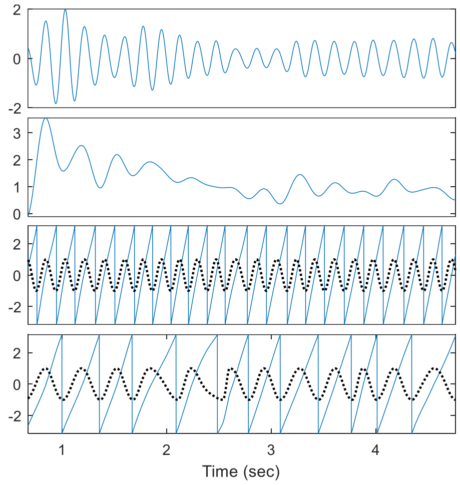

Figure 2: **Amplitude modulations and phases over time**. First panel from top: *syllAM*, second panel: *accAM*, third panel: amplitude normalized *syllAM* (dotted tick line) and phase (continuous thin line); fourth panel: amplitude normalized *accAM* and phase.

A potential issue with the computation of the PLV is due to its dependency on the variability of the oscillatory rates and on the number of cycles of the two signals included in the same analysis window. To factor out these effects from our analysis, for each PLV measurement, we also computed the PLV from 10 bivariate signals obtained by pairing the $\phi(\mathrm{k})_{stress}$ signal used to compute the PLV with 10 different versions of the $\phi(\mathrm{k})_{syll}$ signal obtained by randomly permuting the positions of chunks of instantaneous phase values corresponding to different cycles. These surrogate signals display the same features of the original $\phi(\mathrm{k})_{syll}$ signal but lose all temporal relation with the observed $\phi(\mathrm{k})_{stress}$ signal. The PLV obtained by the observed data is then normalized through division by the average of the PLVs obtained from the analysis of the surrogate data. Note that due to this normalization, a significant degree of coordination is observed when the normalized PLV is significantly higher than one.

3. RESULTS AND DISCUSSION

Figure 3 displays the normalized PLVs obtained from the repetitions of different syllable sequences over speech rate (in syllables per sec).

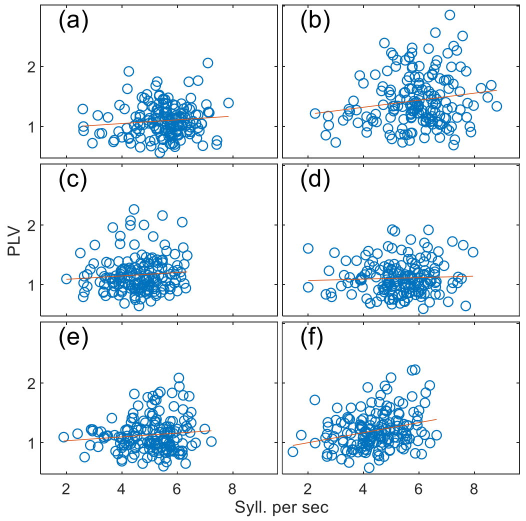

Figure 3: normalized PLV vs syllable rate for each syllable. Panel (a): /bababa/: panel (b): /badego/: panel (c): /klaklakla/; panel (d): /dedede/; panel (e): /gogogo/; panel (f): /tʁatʁatʁa/.

Higher PLVs indicates a stronger coordination between the syllabic and suprasyllabic rates. Pairwise comparisons via t-tests (corrected for multiple comparisons according to the False Discovery Rate criterion [17]) show higher PLVs for /badego/ than for other sequences and higher PLVs for /tʁatʁatʁa/ than for all the CVs (/ba/, /de/, /go/). Finally, PLV is higher during the repetition of /klaklakla/ than during the repetition of /bababa/. In summary, the coordination between the syllabic and supra-syllabic components is lower during the repetition of the same CV syllable than during the repetition of sequences containing different CV syllables or during the repetition of identical syllables with complex structure (CCV).

A linear mixed model with speaker-specific random intercepts and rate-related slopes was used to predict the effects of the syllabic content of the sequences (reference level: /badego/), of syllable rate (centered around its mean), and of their interaction on PLVs. Results show that the coordination of the two rhythmic components was stronger (higher PLV) for /badego/ sequences than for any other sequence (vs.

/bababa/: est.:-0.31, std. err.: 0.04, t val.: -8.39; vs. /klaklakla/: est.:-0.2, std. err.: 0.39 , t val.: -5.42; vs. /dedede/: est.:-0.27, std. err.: 0.036, t val.: -7.91; vs. /gogogo/: est.:-0.26, std. err.: 0.03604, t val.: -7.11; vs. /tʁatʁatʁa/: est.:-0.13, std. err.: 0.04, t val.: -3.22). The simple effect of speech rate (computed on /badego/) had a significant effect on PLV (est.: 0.07, std. err.: 0.023, t val.: 2.856). Interactions between rate and sequence types were not significant, although sequence-specific trends appear in Figure 3.

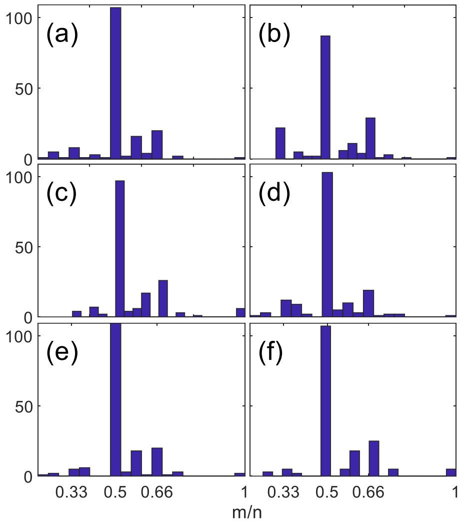

Figure 4: Counts of the optimal m/n ratio values obtained from the PLV analysis for each sequence inversely related to the number of syllables per accent. Same distribution over panels (a to f) as in Figure 3.

Figure 4 displays the distributions of the ratios between the *m* and *n* values estimated to compute the PLVs in Figure 3. As these ratios correspond to the ratios between the frequencies of the syllabic and the supra-syllabic rhythmic components, each value is inversely related to the number of syllables per accent that are produced by one speaker by sequence types. The most frequent ratio across distributions is 0.5. This means that for most of the time, speakers produce two syllables per accent. Most distributions display a peak around 0.66, compatible with the presence of two accents per group of three syllables. Finally, the peak at 0.33 is compatible with the presence of three syllables per accent. Let’s note that the number of speakers that adopt these last two frequency ratios, both compatible with a grouping in chunks of three syllables (with either one or 2 accents), increases in the repetition of /badego/. Moreover, the number of speakers producing two accents per 3 syllables (probably an initial and a final accent [18], ratio of .66) increases for /tʁatʁatʁa/ (c) and /klaklakla/ (f) when compared to CV sequences.

To sum up, coordinative strategies resulting in the grouping of three syllables are more frequent in sequences displaying relatively strong degrees of coordination (high PLVs) between syllables and suprasyllabic prominence (/badego/ (b), /klaklakla/ (c) and (f) /tʁatʁatʁa/).

If coordination between rhythmic components is related to the stability of the grouping pattern adopted, a relatively strong degree of coordination is to be expected in repetitions of sequences of different syllables (as /badego/), because these permit anticipation and overlap between syllables, thus favouring grouping. However, the relatively high degrees of coordination observed in the repetition of identical complex CCV syllables may be due to the fact that speakers reduce some CCV syllables in response to a loss of stability at fast speech rate. Coherently with this interpretation, in the production of /badego/ sequences speakers can choose between two ways of chunking the speech stream in groups of three syllables (both one and two accents per group of three syllables are observed) However, in the production of /tʁatʁatʁa/ or /klaklakla/ sequences, only a few speakers produce one accent per groups of three syllables (i.e. the coordinative relation between syllables and accents underlying this grouping strategy becomes unstable).

4. CONCLUSION

This study brings further evidence of the presence of global coordination patterns in speech production. Their role in speech sensorimotor control is quite crucial, as they tie the behaviour of the many different elements composing the vocal apparatus to the production of functional constituents including several syllables. Moreover, speakers that succeed in reaching faster speech rates, as required by the task, better coordinate the production of syllables with the production of accentual prominence. This result reconciles the presence of speech rate differences between languages with the presence of genuine rhythmic differences ([19, 20]), because the latter kind of differences may be induced by the former one. Further work aimed at including other dimensions (e.g. articulator movements and f0) and time scales in the characterization of the rhythmic aspects of speech production within and across languages will benefit from the observational and modelling paradigms introduced in this paper. The proposed analytical approach constitutes a major advancement in the modelling of the rhythmic hierarchy underlying speech, as it requires minimal linguistic information (syllables and segments rates), which often is available, or which can be estimated (e.g. [21, 22]).

5. ACKNOWLEDGMENTS

Research supported by grant ANR-21-CE28-0015-01 (PASDCODE) of the French National Research Agency.

6. REFERENCES

[1] Dauer, R. M. 1983. Stress-timing and syllable-timing reanalyzed. *Journal of phonetics*, *11*(1), 51-62.

[2] Dauer, R. M. 1987. Phonetic and phonological components of language rhythm. In *Proceedings of the 11th international congress of phonetic sciences* 5, 447-450.

[3] Bertinetto, P. M. 1989. Reflections on the dichotomy 'stress' vs.'syllable-timing'. *Revue de phonétique appliquée*, *91*(93), 99-130.

[4] Auer, P. 1993. Is a rhythm-based typology possible. A study of the role of prosody in phonological typology (KontRI Working Paper No. 21). Hamburg: Universität Hamburg.

[5] Leong, V., Stone, M. A., Turner, R. E., & Goswami, U. 2014. A role for amplitude modulation phase relationships in speech rhythm perception. *The Journal of the Acoustical Society of America*, *136*(1), 366-381.

[6] Lancia, L., Krasovitsky, G., & Stuntebeck, F. 2019. Coordinative patterns underlying cross-linguistic rhythmic differences. *Journal of Phonetics*, *72*, 66-80.

[7] Stetson, R. H. 1951. *Motor Phonetics*. North-Holland, Amsterdam, 2nd ed.

[8] Tuller, B., & Kelso, J. S. 1991. The production and perception of syllable structure. *Journal of Speech, Language, and Hearing Research*, *34*(3), 501-508.

[9] Rochet-Capellan, A., & Schwartz, J. L. 2007. An articulatory basis for the labial-to-coronal effect:/pata/seems a more stable articulatory pattern than/tapa. *The Journal of the Acoustical Society of America*, *121*(6), 3740-3754.

[10] Lancia, L., & Rosenbaum, B. 2018. Coupling relations underlying the production of speech articulator movements and their invariance to speech rate. *Biological cybernetics*, *112*(3), 253-276.

[11] Fougeron, Cécile, Véronique Delvaux, Lucie Menard, and Marina Laganaro. 2018. The MonPaGe_HA database for the documentation of spoken French throughout adulthood. Proceedings of the 11th LREC, 2018, Miyazaki, Japan, pp. 4301–6.

[12] Varnet, L., Ortiz-Barajas, M. C., Erra, R. G., Gervain, J., & Lorenzi, C. 2017. A cross-linguistic study of speech modulation spectra. *The Journal of the Acoustical Society of America*, *142*(4), 1976-1989.;

[13] Poeppel, D., & Assaneo, M. F. 2020. Speech rhythms and their neural foundations. *Nature reviews neuroscience*, *21*(6), 322-334.

[14] Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., ... & Liu, H. H. 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, 454(1971), 903-995.

[15] Rosenblum, M., Pikovsky, A., Kurths, J., Schäfer, C., & Tass, P. A. 2001. Phase synchronization: from theory to data analysis. In *Handbook of biological physics* (Vol. 4, pp. 279-321). North-Holland.

[16] Huang, N. E., Wu, Z., Long, S. R., Arnold, K. C., Chen, X., & Blank, K. (2009). On instantaneous frequency. Advances in adaptive data analysis, 1(02), 177-229.

[17] Benjamini, Y., & Yekutieli, D. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165-1188.

[18] Jun, S. A., & Fougeron, C. (2002). The Realizations of the Accentual Phrase in French Intonation. *Probus*, *14*(special issue on Intonation in the Romance Languages), 147-172.

[19] Arvaniti, A. 2009. Rhythm, timing and the timing of rhythm. *Phonetica*, *66*(1-2), 46-63.

[20] White, L., Mattys, S. L., & Wiget, L. 2012) Language categorization by adults is based on sensitivity to durational cues, not rhythm class. *Journal of Memory and Language*, *66*(4), 665-679.

[21] Wang, D., & Narayanan, S. S. 2007. Robust speech rate estimation for spontaneous speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(8), 2190-2201.;

[22] Hoang, D. T., & Wang, H. C. 2015. Blind phone segmentation based on spectral change detection using Legendre polynomial approximation. *The Journal of the Acoustical Society of America*, *137*(2), 797-805.