Running title: Speech rhythm entrainment in a dyadic reading task

SPEECH RHYTHM CONVERGENCE IN A DYADIC READING TASK

Karina Cerda-Oñate[1]
Gloria Toledo Vega[2]
Mikhail Ordin[3,4*]

[1] Departamento de Lengua Castellana y Literatura, Universidad Católica del Maule, Chile
[2] Facultad de Letras, Pontificia Universidad Católica de Chile
[3] BCBL- Basque Center on Cognition, Brain and Language
[4] IKERBASQUE- Basque Foundation for Science

Acknowledgments:

This research was supported by the Spanish Ministry of Economy and Competitiveness (MINECO), through the "Severo Ochoa" Programme for Centres of Excellence in R&D (SEV-2015-490). KC-O acknowledges financial support from the Program for the Training of Advanced Human Capital (PFCHA) at CONICYT (Scholarship n. 21151287/2015). The authors also thank C. García-Meza (UCSH Chile), F. Nocetti (UDEC Chile), P. Timofeeva (BCBL), J. Aguasvivas (BCBL) and Candice Frances (BCBL) for their help with the data collection and synchrony measurements, and Dr. Magda Altman (BCBL) for language support, proofreading and editing this manuscript. We are also thankful to the editors of this special issue, who highlighted the timely and important questions related to phonetic convergence in interaction.

*Address for correspondence: Mikhail Ordin, Basque Center on Cognition, Brain and Language, San Sebastián, Mikeletegi 69, 20009, Spain.

Abstract

We tested the effect of co-presence on entrainment to speech rhythm, examining differences in speech rhythm convergence during a reading task, in conditions where the reading partner was present or absent. Speech rhythm was operationalized as a two-level phenomenon. At a lower level, rhythm was operationalized as regularity in the distribution of salient acoustic events (vowel onsets) and regularity in the duration of speech intervals (consonantal and vocalic intervals). At a higher level, rhythm was operationalized in terms of meter, the distribution of salience that can group syllables into metrical structures, based on lexical stress and phrasal prominence. To assess the impact of the presence/absence of a co-speaker on speech entrainment, we asked a *model speaker* and *experimental participants* to sit side-by-side and read two texts aloud at the same time while recording each speaker separately. Later, we requested only *experimental participant*s to read in synchrony with the recording obtained from the *recurrent model speaker* during side-by-side reading and, afterwards, we asked *experimental participants* to read in synchrony with a solo recording from the same *recurrent model speaker*. We used a poetic text with strong meter and a narrative text with weak meter as reading materials. To assess the degree of speech rhythm convergence, we measured the degree of durational variability in vocalic and consonantal intervals across conditions and texts and the deviation in vowel onsets between chorusing readers. We found that participants make their speech more regular to facilitate chorusing. Importantly, inter-speaker synchronization was substantially improved during side-by-side reading compared to the condition where the speaker read in synchrony with a recording obtained during side-by-side synchronous reading, even though the acoustic signal received by the experimental participants was the same in both conditions. We also found that the text with strong meter modulated the success of speech rhythm convergence only in the most challenging condition, requiring reading in synchrony with a recording of the *model* speaker reading solo. These results show that co-presence plays a crucial role in inter-speaker entrainment, while meter can provide additional benefits for chorusing in more challenging conditions.

Keywords: phonetic convergence, speech rhythm, dyadic reading, rhythm convergence, rhythm synchronization

Introduction

Prosody is an essential aspect of the speech signal that is used to convey post-lexical meaning and is modulated by speaker gender, social status, native language and dialect, discourse, and emotional state, among others (Crystal, 1969; Ladd & Cutler, 1983; Labov, 2006; Szczepek Reed, 2010; Clopper & Smiljanic, 2011). Prosody includes all suprasegmental phenomena, such as stress, tone, intonation, and rhythm, produced by the interplay of pitch, intensity, and duration in stretches of speech larger than a segment (Lehiste, 1976; Ladd & Cutler, 1983; Nooteboom, 1997; Fletcher, 2010).

Rhythm as a suprasegmental speech phenomenon

As a suprasegmental speech phenomenon, speech rhythm is conceptualized and defined in different ways (Moore, 2012; Turk & Shattuck-Hufnagel, 2013; Smith et al., 2014; Nolan & Jeon, 2014; Cummins, 2015). Since we do not focus on establishing a theoretical definition for rhythm, we will not delve into terminological debates and will provide an operational definition of speech rhythm that emphasizes those aspects of the phenomenon that are relevant for our study. In this context, speech rhythm will be operationalized in terms of the duration and variability of sub-syllabic units: *consonantal intervals* (single or multiple consonants that happen consecutively and are not interrupted by vocalic segments) and *vocalic intervals* (vowels uninterrupted by pauses). Lower variability scores thus reflect higher regularity in the duration of vocalic and consonantal intervals, and higher regularity in the temporal distribution of *perceptual centers* – temporal reference points when an acoustic event (e.g., syllable, or vowel onset corresponding to a sharp increase in intensity level) is psychologically perceived to occur (Marcus, 1981) – in the speech signal. These perceptual centers (p-centres) determine the perception of regularity in the acoustic signal (Scott, 1998).

Entrainment and speech rhythm convergence

Tapping, speech-cycling, and dancing tasks reveal the human ability to entrain to a rhythmic sensory input signal (Cummins & Port, 1998; Repp, 2005; Eerola et al., 2006, Merchant et al., 2015). In the case of speech rhythm, it has been found that native speakers can entrain to perceptually salient acoustic events that may correspond to more or less regularly distributed stressed syllables or vowel onsets, (Marcus, 1981; Vos et al., 1995). These events are produced by another speaker when reading aloud. To facilitate synchronization by a partner, a speaker will (intentionally or unintentionally) make the distribution of these salient events more regular, hence more predictable; this adaptation does not require any prior training (Cummins, 2002, 2003, 2007, & 2009; Bowling et al., 2013). It has been suggested that side-by-side synchronous reading

allows for convergence of rhythmic patterns in speech (Pardo, 2010; for a full review of research on phonetic convergence in speech imitation tasks, see Pardo et al., 2017). Pardo (2010) and Pardo et al. (2017) consider phonetic convergence to be a general phenomenon, which occurs not only in regard to speech rhythm, but also with the formant structure of vowels, pitch, speech tempo, VOT on plosives, etc. In our study, we focus on the convergence of temporal patterns that a) result in specific rhythmic organization of speech; and b) lead to the rhythmic convergence, and c) allow speakers to use rhythmic predictability to synchronize speech in dyadic reading tasks.

Earlier studies have already reported convergence effects during synchronous reading (Cummins 2002; 2009). In Experiment 2, Cummins (2002) asked three participants to synchronize their reading with a recording made either in a synchronous reading condition – when two speakers were reading the same text together, and the speech of one of these speakers was recorded – or with a recording made when the speaker read the text alone. A higher degree of convergence was found for the synchronous reading condition for two out of three participants. In Cummins (2009), this effect was replicated with four participants who had to synchronize with two different model speakers. Three out of four participants synchronized better with the speech originally obtained in dyadic reading conditions.

In the current study, we decided to run a hypothesis-driven experiment to confirm this finding in a larger sample of participants (N=30) and generalize the findings over different language populations. Importantly, we sampled our participants from a population with an ambient language (Spanish), which is rhythmically different from English – the native language of participants in Cummins' studies (2002; 2009) – so we could generalize results over a broader range of languages (Spanish exhibits more regular distribution of vowel onsets, stressed syllables, more equal vocalic and consonantal intervals, etc.; White & Mattys, 2007). We believe that rhythm provides cues that facilitate synchronization, and thus the effect should not be limited to languages with a specific type of temporal organization. This hypothesis, however, had not yet been tested on languages which are rhythmically different from English (where this effect was first established). Also, we decided to look at synchronization on different types of text materials – poetic and narrative texts – to understand whether this effect is limited to a certain type of reading material or whether it might increase when the material contains stronger inherent rhythm (in this case, a poetic text) that could potentially provide further cues for synchronization. Most importantly, we introduced

136 another condition: a live speaker. Cummins (2002; 2009) provided some empirical
137 evidence that at least English speakers achieve better rhythmic convergence with
138 recordings obtained in synchronous conditions than with recordings of solo readers. We
139 introduced a live speaker as an additional condition because a recorded model speaker
140 cannot make online adaptations to a partner while the task is being performed; only a
141 live speaker can adapt their speech online to achieve better convergence with their
142 partner. Thus, we posit that in dyadic reading both readers will dynamically change
143 their speech at the same time, (1) providing cues that help their live partner increase
144 synchrony, and (2) trying to converge their own speech with the partner's speech. These
145 dynamics might further enhance convergence of temporal patterns, as we aimed to test
146 in this study. A more technical formulation of our hypothesis is provided below. But
147 first, we introduce two core concepts/terms that are key to our theoretical and
148 methodological framework: beat and meter.

149 **Beat, meter, and their role in speech rhythm convergence**

150 Beats, or perceptually salient acoustic events in the speech stream – p-centres – may be
151 perceived as more or less regularly distributed (Marcus, 1981; Vos et al., 1995). A
152 regular beat licenses predictions for the timing of following beats. This facilitates
153 synchronization with speech as p-centers correspond to more prominent speech
154 segments, for example, vowel onsets (thus each syllable in Spanish would correspond to
155 one p-centre). Beats can be organized into hierarchical patterns based on their relative
156 strengths (London, 2004; Large & Snyder, 2009; Fitch, 2013) to create a hierarchical
157 temporal grid, which we refer to as meter. For example, alternations between stressed
158 and unstressed syllables can be organized into iambic, trochaic, or more complex
159 metrical patterns.

160 Meter has been found to play a significant role in the entrainment of motor
161 output to sensory input when dancing, marching, tapping, clapping, singing, and other
162 types of behavior that require the perception and use of expectancy cues (Hannon &
163 Johnson, 2005; Fitch, 2013). Pitt and Samuel (1990) showed that metrical expectancy
164 cues are weaker in the context of normal narrative speech and that meter only leads to
165 better performance when beat cues are insufficient, e.g., when reading a list of words, or
166 when context encourages the listener to pay closer attention to the meter.

167 The relationship between timing regularity – regular distribution and thus the
168 predictability of when the next syllable will occur – and metrical expectancy in speech
169 perception was also explored by Quené & Port (2005). In one experiment, participants

had to listen to a set of words in which the onsets of the stressed syllables were either regularly or irregularly distributed (all words within sets had the same meter). In the second experiment, participants had to listen to a set of words. In one condition, the words had either exclusively trochaic or exclusively iambic meter, making stressed syllables onsets predictable. In the other condition, iambic and trochaic meters were used interchangeably, making the stressed syllable onsets less predictable. Participants in both experiments had to do a phoneme monitoring task, which consisted of listening to a set of words or sentences and pressing a button as soon as the target phoneme was detected. Quené & Port (2005) found a main effect for timing regularity (meaning that reaction time was shorter when the syllable onsets were regularly distributed) but not for metrical expectancy (indicating that reaction time was not modulated by whether words had the same or different meter).

Meter can provide expectancy cues for the timing of the following stressed syllable, but the salience of these cues – as tested in phoneme monitoring tasks – remains unclear (Pitt & Samuel, 1990 found a facilitatory effect of meter consistency; Quené & Port, 2005 did not), and their role in the precision and ease of rhythmic convergence has not been established. We therefore decided to include materials with both weak and strong meter in our experiment, choosing a narrative text with weak meter (Text 1) and a poetic text with strong meter (Text 2). We believe that the organization of beats into metrical patterns may provide yet another set of expectancy cues and further facilitate the convergence of rhythmic patterns in speech. These ideas will be further explained in the next section.

Hypothesis and predictions

In this study, we explored speech rhythm as an acoustic phenomenon that is modulated by human co-presence and variation in the degree of meter. To explore human co-presence, we tested the effect of the presence/absence of an interlocutor during a synchronous reading aloud speech task. We posited that the presence of an interlocutor would induce speakers to provide better expectancy cues in an effort to encourage enhanced rhythmic synchronization; this more regular speech rhythm in speech should, in turn, be reflected in lower scores for the rhythm variability metrics and higher %V measures (Ramus et al., 1999; Grabe & Low, 2002; Dellwo, 2006; White & Mattys, 2007).

We adopted the premise that perception of rhythmicity in speech is based on the extraction of p-centers (Morton, Marcus, & Frankish, 1976), which can serve as

expectancy cues, allowing the listener to predict the onsets of vowels, syllables, and higher-level prosodic constituents (Lehiste, 1973). Better expectancy cues, such as the ones provided by a text with strong meter, should allow the listener to better entrain their speech production – motor output – to the speech produced by an interlocutor, that is, the sensory signal emitted by another biological organism. If the timing of p-centers encourages the perception of temporal regularity, then speech rhythm entrainment should also be enhanced (Large & Snyder, 2009; Vuust & Witek, 2014).

As for the role of meter, we expected that the text with stronger meter would allow for better timing convergence of stronger beats during speech production, further enhancing synchronization of verbal behavior between partners who read the poetic text compared to the narrative text.

To summarize, we hypothesized that co-presence would play a significant role in increasing rhythmic convergence. Bi-directional attempts at speech synchrony (when both speakers dynamically change their speech patterns) result in greater temporal convergence than unidirectional attempts to synchronize speech with a recording. We believe that regular distribution of vowel onsets (and thus syllables) will provide expectancy cues to align timing patterns in dyadic reading. Metrical prominence would affect the degree of speech rhythm convergence, by reinforcing expectancy cues – e.g., promoting higher regularity between vowel onsets in stressed syllables. We also predicted that, even if the text with stronger meter enhanced the degree of coupling between speakers, the co-presence of reading partners would be the strongest predictor for increased levels of synchronization.

To test these hypotheses, we set up an experimental study asking native Spanish speakers to read a narrative text with weak meter (Text 1) and a poetic text with strong meter (Text 2) in synchrony. The *experimental participant* either read together in a live session with the co-present *model speaker*, or read along with a recording of the model speaker obtained when the *model speaker* was reading the texts solo, and when the *model speaker* was reading in synchrony with the *experimental participant*. We measured the degree of speech rhythm convergence by estimating synchrony measures. Synchrony was operationalized in terms of temporal matching of onset starting points for consonantal and vocalic intervals. Vowel and consonants starting points were compared across conditions and texts for the *model speaker* and the *experimental participants*. All of these measurements are explained in the Methods and Materials section below.

Methods and materials

This experiment was conducted following the research policies of the Social Sciences and Humanities Ethics Committee at the Pontifical Catholic University of Chile (PUC) and the Chilean National Commission for Scientific and Technological Research. Participants provided written informed consent and were informed that they could withdraw from the experiment at any time without any consequences. All data were stored anonymously.

Participants

Thirty experimental participants were recruited using different social media websites. 22 participants were female (mean age = 23.27; SD = 3.80) and 8 participants were male (mean age =24.5; SD = 3.11). All participants were native, monolingual speakers of Chilean Spanish and reported no previous experience with any activities that might train reading synchronization abilities, such as ensemble vocal music performances or theatrical performances requiring people to speak at the same time and in synchrony. None of the participants had ever resided outside of Chile for longer than six months. This background demographic information was obtained using a background questionnaire that was completed by each participant. The model speaker was a female native Chilean Spanish speaker.

Materials

The experiment included three reading materials in Spanish, which were read in 4 different conditions: 1) Warm-up text, Text 0, a list of 16 sentences; 2) A narrative text (text with weak meter), Text 1, which is a Spanish version of *The North Wind and the Sun* (Coloma, 2016); 3) A poetry text (text with strong meter), Text 2, which is a poem written in octosyllabic verse, consisting of two stanzas (Parra, 1988). The recordings obtained from participants reading Text 0 were not used as speech data, since Text 0 was only used for the warm-up phase. These materials are further explained in Table 1 below.

INSERT TABLE 1 SOMEWHERE HERE

Design

Data were obtained from the *model speaker* and *experimental participants* reading in 4 different conditions. After running the first condition, the order of other conditions was counter-balanced across participants. These four conditions are described in Table 2:

INSERT TABLE 2 SOMEWHERE HERE

In the first condition or *Synchrony-live condition* (1), participants read aloud and in synchrony with the *recurrent model speaker*, a naïve female monolingual native speaker of Spanish, who had previous experience in acting and was unaware of the objectives of the research. The *recurrent model speaker,* as the name indicates, did not vary across participants; in other words, the same *recurrent model speaker* read along with each one of the 30 *experimental participants*. In the second condition or *Synchrony with recording from live condition* (2), *experimental participants* listened to the recording of the *recurrent model speaker* reading with them in the first condition and read in synchrony with that recording. In the third condition or *Synchrony with recording from non-live condition* (3), all *experimental participants* read the texts in synchrony with a recording obtained from the model speaker reading the texts alone. The individual recording used in the third condition was the same for all 30 *experimental participants*. Finally, in the fourth condition or *Solo recording condition* (4), *experimental participants* read aloud individually. Therefore, three out of the four conditions were synchronic reading conditions.

Procedures

All recordings for each *experimental participant* were made during one session, which lasted around 40 minutes. *Experimental participants* were asked to read the texts in all four conditions. A warm-up material (Text 0) was used which was used to familiarize participants with the task (and was left unanalyzed); a narrative text with weak meter (Text 1); and a poetry text with a strong meter (Text 2). The recording for condition 4 (the model speaker reading alone) was made prior to the experiment. The *recurrent model speaker* read each of the three texts with the participants in conditions 1, 2 and 3 (condition 1 always preceded condition 2; otherwise the order of conditions and texts 1 and 2 within conditions was counterbalanced across participants). The only instruction given throughout the experiment to the *experimental participants* was to try to "read in synchrony with their partner and to avoid repeating after the other participant in all synchronous conditions." The model speaker, who was unaware of her experimental role, was never advised on correct pronunciation, timing, or any other aspect related to rhythm, beat, or periodicity. It is also important to note that the *recurrent model speaker* and the *experimental participants* were unfamiliar with any concepts related to speech rhythm, speech convergence, or synchronous speech.

As for the *Synchrony-live condition* (1), the *recurrent model speaker* and *experimental participant* sat side by side and read aloud synchronously. In this

condition, the reading partners were not able to look at each other because they were both focused on reading the texts that were placed in front of them, positioned at eye level on a book stand. They sat facing in the same direction, which excludes the possibility of seeing something peripherally without simultaneously turning their heads. Thus, we can assume that visual interaction had a minimal impact, if at all, on the results we report. For the *Synchrony with recording from live condition* (2), and *Synchrony with recording from non-live condition* (3), the *experimental participant* wore headphones to listen to the recording from the *recurrent model speaker*. The volume of the recording was adjusted to a comfortable threshold defined by each participant before starting the experimental task. Finally, during the *Solo-recording condition* (4), the participant read alone.

All recordings took place in a quiet room at PUC. Two head-mounted dynamic uni-directional (cardioid) WH20XLR Shure microphones were used to simultaneously record the speech of both the *recurrent model speaker* and the *experimental participant*; these microphones were chosen to avoid picking up any signal from the co-speaker. Also, to avoid any interference, participants sat side-by-side with a distance of about 60 cm between them. Sony MDR-NC8 Headphones (with noise canceling switched off) were used to play the recordings of the *model speaker* during the *Synchrony with recording from live condition* (2) and *Synchrony with recording from non-live condition* (3). The noise canceling was switched off. The noise cancelling would have cut participants off from sounds in the ambient auditory environment including their own voice, thus disrupting participants' ability to monitor their own speech production. With noise-cancelling switched off, participants heard their voice not through the headphones, but as a part of the ambient auditory environment. The experimental participants and the model speaker were recorded using the same recorder in a single WAV file as left and right channels, in all conditions; a TASCAM DR-40 recorder (sampling rate = 44.1 kHz, 16 bit) was used for this purpose. It is important to note that all recordings from the experimental participants were extracted during the same session; this helped to control for any unwanted noise or interference in the *Synchrony with recording from live condition* (2). We also used windscreens to minimize the possibility that close-up uni-directional microphones would pick up the partner's voice.

Data analysis

The assessment of rhythmic duration took both consonantal and vocalic intervals into consideration. The speech data was labeled using the orthographic transcription of the reading materials (Text 1 or weak meter; Text 2 or strong meter) as a reference. For this, the acoustic software analysis *Praat* (Boersma & Weenink, 2016) and the *Praat* plugin for forced text-to-speech alignment *EasyAlign* (Goldman, 2011) were used. Notably, the *EasyAlign* plugin was employed to segment and label the recordings into phonemic units that were based on the orthographic transcription of the texts the participants read during the tasks. EasyAlign is a Praat plug-in for semi-automatic aligning of segmented words, syllables, and phones to an acoustic signal based on the audio recording and its orthographic transcription. First, orthographic texts were manually segmented into utterances (identical segmentation of texts into utterances was applied to each recording). The second stage was a grapheme-to-phoneme conversion, in which the orthographic transcription for each utterance was converted into a sequence of phonemes (choosing from a set of possible pronunciation variants); optional phonemes were marked by an asterisk. After this, an expert phonetician (the first author) compared the sequence of phonemes against the audible recording and made the necessary adjustments for individual idiosyncrasies. At the third stage, a verified phonetic transcription was aligned to the acoustic signal using HTK-toolkit, an algorithm based on Markov Hidden Chains (http://htk.eng.cam.ac.uk/), to produce segmentation of the recording into phones, syllables, and words. The durations of the phones were used for further analysis. Details of the algorithm can be found in Goldman (2011).

The labeled data for each participant, *Condition,* and *Text* was saved into a single Praat TextGrid file. This aligned phonetic labels with the acoustic waveform, using different tiers for each transcription. The transcriptions were all in SAMPA for Spanish, a machine-readable phonetic alphabet (Wells, 1996). The annotations were then manually checked to correct the most obvious technical errors. EasyAlign works reliably with Spanish speech produced by native speakers (both Castilian and Latin American Spanish varieties are targeted). The main motivation for using EasyAlign in our study was to automate grapheme-to-phoneme conversion. Automatizing this process improves the reproducibility of the results because the segmentation decisions are not affected by subjective choices, such as those related to the role of liquids and non-nuclear vocoids. Subjective decisions tend to differ across annotators and could be biased by an experimental hypothesis, adding to the inconsistency of the results across studies. Here instead, purely acoustic parameters were used to detect the phoneme boundaries.

To assess the temporal patterns of speech, we computed the durational variability of consonantal (C) and vocalic (V) intervals using *Correlatore* (automatic rhythm analysis software; Mairano & Romano, 2010). This variability is captured, among other measures, by the coefficients of variation, VarcoC and VarcoV (White & Mattys, 2007; Dellwo, 2006); the coefficient is the standard deviation divided by the mean duration of intervals for each sentence. Fluctuations in standard deviation depend on the mean duration of intervals, thus standard deviation will be modulated by individual speech rate or stylistic changes in speech rate even between different sentences produced by the same speaker (speech delivered at a faster rate will have shorter overall speech interval durations, thus smaller standard deviations). Varco measures thus reflect variability, while controlling for between- and within-speaker differences in speech rates (Dellwo, 2006). We also estimated %V, or percentage of vocalic material in a sentence (this measure is also reported to be robust to tempo variation). Varco and %V measures are referred to as global metrics because they capture durational variability averaged across the whole utterance. Local metrics, on the other hand, capture variability in the duration of C and V intervals in a pairwise fashion (respectively, CnPVI and VnPVI), assessing how the second interval differs from the first, the third from the second, and so on (Grabe & Low, 2002). The difference in interval durations between each consecutive pair is normalized by the mean duration of that pair, and the sum of these normalized differences is then divided by the number of consecutive pairs. The result is often multiplied by 100 to convert it into the percentage of overall variability within a sentence, in which interval durations fluctuate from one pair to the next. The formula is given below (1). If a consonantal or vocalic interval was interrupted by a pause exceeding 150ms, we considered them to be two separate intervals. These metrics were calculated for each sentence, and then an average was calculated for each speaker and text. %V, Varco, and nPVI measures have been widely used in studies on speech rhythm; using these metrics makes it possible to compare and discuss the results of this research in light of other studies on speech rhythm.

$$nPVI = 100 * (\sum_{1}^{m-1} \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/_2}) / (m-1) \qquad (1)$$

where m is the number of intervals, and k is the ordinal number of each consecutive interval.

To measure convergence, we estimated the difference between the onset of each speech segment delivered by the two readers. To assess synchronization between speakers, we computed synchronization metrics for segment onsets using an in-house Python 3.6

script (see Supplementary material) to compare and calculate vowel and consonant onset time differences between the Praat TextGrids of the model speaker and the experimental participant, both recorded by the same recorder in a single WAV file as left and right channels. The TextGrids from the model speaker and experimental participants were compared as two columns. If the segments in the columns were the same, the onset difference was calculated. We only included this data in further analyses if 85% of the compared segments matched, otherwise we considered synchronization had failed (this happened when convergence was too weak and readers could not align with onsets of at least 85% of the segments within a sentence; we discarded 5.5% of speech data per pair).

The rhythm metrics were calculated for each utterance and then averaged for each participant across utterances. The averages were used in a repeated-measures ANOVA (in SPSS v.19) to assess the differences between conditions: *Synchrony-live* (1); *Synchrony with recording from live* (2); *Synchrony with recording from non-live* (3); and *Solo-reading* (4). We applied Greenhouse-Geisser correction for sphericity violations and report corrected p-values and degrees of freedom, with the a-priori alpha level set to 0.05. *Condition* and *Text* (weak meter [Text 1] versus strong meter [Text 2]) were included as factors in each analysis. The interaction *Condition* by *Text* was included in the models. For pairwise comparisons, paired t-tests (with Bonferroni correction) were used.

Results

Vowels

Vowel percentage.

We found a significant effect of *Condition* and *Text* (refer to Table 3 for the statistical parameters). The effect size was substantial for both factors; the interaction was not significant. Further pairwise comparisons (with the Bonferroni correction) indicated that %V was significantly higher in the *Synchrony-live* (1) than in the other conditions. We also observed a significantly higher %V in the text with strong meter (Text 2) than in the text with weak meter (Text 1) across all conditions. Figure 1 displays the differences in %V between conditions and text types.

INSERT FIGURE 1 SOMEWHERE HERE

VarcoV.

The analysis showed a significant effect of *Condition* and *Text* with a substantial effect size for both factors; interaction between the factors was not significant. Pairwise

comparisons revealed that VarcoV was higher in the *Synchrony with recording from non-live* (3) than in the other conditions (this measure was not different between other pairs of conditions). The measure scores were lower in the text with strong meter (Text 2) than in the text with weak meter (Text 1). This pattern is displayed in Figure 2.

INSERT FIGURE 2 SOMEWHERE HERE

VnPVI.

We observed a significant but moderate effect of *Condition* and a substantial effect of *Text* on the nPVI scores with no significant interaction between the factors. Pairwise comparisons showed that the only difference for the factor *Condition* was between *Solo recording* (4) and *Synchrony with recording from non-live* (3) with higher scores in condition 3 than 4. Again, the nPVI scores were lower in the text with strong meter (Text 2) compared to the text with weak meter (Text 1). Figure 3 displays this pattern.

INSERT FIGURE 3 SOMEWHERE HERE

Consonants

VarcoC and CnPVI.

We did not observe a significant effect of *Condition* on VarcoC or CnPVI scores (see Table 3). However, the effect of the factor *Text* on VarcoC was significant and substantial. No significant interaction between the factors was observed. The effect on CnPVI was also statistically significant, but unsubstantial (η_p^2=.03). Durational variability of consonantal intervals at the timescale of whole utterances – captured by VarcoC – was higher on the text with strong meter (Text 2, poetic) than the text with weak meter (Text 1, narrative). Pairwise durational variability (changes in durational variability between successive pairs of consonantal clusters) was similar in the two texts (the effect of *Text* on consonantal PVI was negligible). This pattern is displayed in Figures 4 and 5.

INSERT FIGURE 4 SOMEWHERE HERE

INSERT FIGURE 5 SOMEWHERE HERE

INSERT TABLE 3

Synchrony data

Onset synchrony

To analyze the differences in vowel onset synchrony between poetic and narrative texts, and between the three different synchronic conditions (conditions 1, 2 & 3), we ran a repeated-measures ANOVA with *Condition* and *Text* as factors and onset differences as

the dependent variable (onset difference was calculated by subtracting the onset time of the participant from the onset time of the model speaker). The factor *Condition* had three levels: *Synchrony live* (1), *Synchrony with recording from live* (2), and *Synchrony with recording from non-live* (3). We report the corrected significance values and degrees of freedom in Table 4 below and display the pattern of results in Figure 6 (negative onset differences in conditions 2 and 3 confirmed that experimental participants were following the model). The analysis showed a significant effect of C*ondition*; pairwise comparisons (Bonferroni corrected) showed that vowel onset differences were smaller in the *Synchrony live condition* (1) than in the other two conditions, and synchronization was significantly stronger when speakers were shadowing the recording made in the synchrony reading condition than the recording made in the solo reading condition. We did not observe a significant effect of *Text*, however, we did observe a significant if weak interaction between *Text* and *Condition*, F (2, 42) = 3,402, p =. 04 (uncorrected), η_p^2 = .018 (noteworthy, p = .06 after applying Greenhouse-Geisser correction, which signals that further interpretation of this interaction should be considered with caution and requires verification in further studies). To understand this interaction, we performed 2-tailed paired t-tests comparing synchronization on segment onsets for poetic vs. narrative tests in three different conditions. The tests showed that in the *live* condition (p=1.0) and in *Synchrony with recording from live* condition (p=1.0), the difference was not significant, while in the *Synchrony with recording from non-live* condition, synchronization was stronger for the poetic than the narrative text (p=.033, corrected).

INSERT TABLE 4 SOMEWHERE HERE

INSERT FIGURE 6 SOMEWHERE HERE

Effect of Condition versus model speaker training

As the model speaker performed the task multiple times, it is possible that her performance improved over the course of the experiment. If so, better synchronization in the live condition could have been caused by progressively better entrainment by the model speaker to the experimental participants rather than by co-presence of a live partner allowing for bi-directional alignment. To ascertain whether this was the case, we performed ranked correlations between the order of the participant in the experiment (the higher the order, the more practice could have benefitted the model speaker) and onset differences. Strong and significant correlations between the ordinal variable and

onset differences would signal that the precision of synchronization was associated with the duration of the experiment(i.e., longitudinal effect on the model speaker's skills). The correlations were run separately for poetic (ρ=-.11, p=.566) and narrative texts (ρ=.19, p=.299). We found weak and non-significant correlations, showing that the null hypothesis was as likely as the alternative hypothesis (in other words, precision of onset synchronization in the live condition was not associated with the progression of the experiment). To estimate support for the null hypothesis, we calculated Bayes factors for the correlation coefficients (BF_{10}=.23 for the poetic text and BF_{10}=.43 for the narrative text). This showed moderate but reliable evidence in favour of the null hypothesis and allowed us to exclude the possibility that better synchronization between readers in the live condition occurred because the model speaker progressively honed her speech entrainment skills over the course of the experiment.

Summary of the results

The data showed that the durational variability of vocalic intervals was consistently lower (i.e., VarcoV and nPVI scores were lower) in the text with strong meter (Text 2) than in the text with weak meter (Text 1). At the same time, the durational variability of consonantal intervals was significantly higher in the text with strong meter (Text 2). These patterns suggest that speakers tried to make the vowel durations more or less equal in the text with strong meter (Text 2), possibly for stylistic reasons, and vowels were stretched more naturally. This modulation is also reflected in the higher percentage of vocalic material in the text with strong meter compared to the text with weak meter – %V was lower in the text with weak meter (Text 1). However, this behavior also led to increased variability in the duration of inter-vocalic intervals. Thus, it seems that speakers tried to quasi-equalize the intervals that could form beats, i.e., intervals that typically had a higher level of intensity – vowels – rather than normalizing interbeat intervals.

We also observed unequal variance in the measures of vocalic variability, i.e., significant ANOVAs, between conditions, but we find this result difficult to interpret. We observed lower variability in the duration of vocalic intervals in the *Synchrony-live* condition (1) and the *Synchrony with recording from live condition* (2). It is possible that this result reflects the effort made to provide a partner with an opportunity for efficient synchronization, and from an attempt to synchronize with partners' utterances by changing vocalic durations (vowels can be stretched and compressed easily).

However, in the *Solo recording* condition (4), the durational variability of the vowels was lower than in *Synchrony with recording from non-live condition* (3), although we would not expect differences between these two conditions (and if we had expected any difference, it would have been in the opposite direction). The variability in the duration of consonantal clusters between conditions is not significantly different, indicating that people do not try to sync their inter-vowel intervals, but rather take advantage of the stretchability/compressibility of the vowel intervals during synchronization.

The segment onset differences did not vary significantly between text types, suggesting that the synchronization strategy is probably the same for texts with weak meter (Text 1) and strong meter (Text 2).

As for between-speaker synchrony, we observed substantial enhancement in onset convergence in the *Synchrony-live condition* (1) compared to synchronization with the recordings, irrespective of whether these recordings were made in the live (condition 2) or non-live condition (condition 3). This trend proves that the co-presence of a live partner is important because only in the *Synchrony-live condition* (1) did both partners work in synergy to both provide and perceive synchronization cues and use them to amend their speech production, probably, as suggested above, by manipulating the duration of vocalic intervals rather than consonantal clusters.

Teasing apart the significant interaction between *Condition* and *Text* factors on the synchronization measure, we showed that strong meter enhanced synchronization only in the most challenging condition, the condition which required reading in synchrony with a solo recording from the recurrent model speaker, and that it did not prove to be beneficial in other conditions.

Discussion

Our findings show that when speakers read with another speaker or with a recording of another speaker, they regularize the durations of vowels, consonantal intervals, and syllables, thus making the distribution of vowel and syllable onsets – p-centers or beats – more isochronous and thus more predictable. Enhanced expectancy cues at the beat level, in turn, allow for better inter-speaker synchronization at vowel onsets, as reflected in the higher synchrony measures found in our study.

We found that the temporal variability of vocalic intervals differed across conditions. This suggests that when people need to provide expectancy cues to enable better synchronization – even if they are unaware of their behavioral patterns – the durational variability of vowels becomes lower. Durational values for both the

Synchrony-live condition (1) and *Synchrony with recording from live condition* (2) were more or less maintained for both texts. Interestingly, the durational variability of consonantal intervals was not modulated by *Condition*. The fact that experimental participants did not manipulate consonant intervals is not surprising, since the nucleus of the syllable in Spanish is always a vocalic interval. Therefore, keeping the consonantal structure constant could be considered a way to increase the chances of achieving synchrony for vowel onsets.

The results indicate that the metric structure of Text 2 made a significant contribution to the regularity of beat distribution in the speech signal, as reflected by the higher values of %V and lower durational variability of vowels, both pairwise (VnPVI) and as measured across whole utterances (VarcoV). Interestingly, for VarcoC, we found the opposite effect. For CnPVI, we did not observe significant differences between the text with weak meter (Text 1) and the text with strong meter (Text 2). This suggests that a prominent meter (that is, regular and predictable distribution of stressed syllables) in a poetic text leads to better inter-speaker coherency in the pronunciation of consonantal clusters and preserves syllabic structure. These results indicate that the mere presence of strong metrical structure makes the distribution of beats more regular. This not only provides additional expectancy cues at the level of meter but also enhances expectancy cues at the level of beat.

It is important to note some limitations on generalizing our findings. Spanish is a language that does not have great durational contrasts between stressed and unstressed vowels. In this respect, it differs from languages like English, German, and Russian, whose native speakers likely do not benefit from the effect of meter in the same manner as Spanish speakers do. We believe that in these other languages, different strategies might be used to enhance synchronization. These could include different behaviors related to consonantal intervals or combining synchronization of vocalic and inter-vocalic intervals, something that was not observed in our data. Besides, our conclusions are based on two texts (a narrative and a poem; one text per reading style). Although we had a large sample of participants, which justifies generalization across speakers sharing the same language, the degree to which the results are generalizable across texts has not been tested. The differences that we attributed to the effect of meter might also be due to differences in the phonetic material – idiosyncrasies in the phonotactic composition of the sentences in the narrative and poetic texts. Thus, the effect of meter should be confirmed by future studies and for other languages.

Our data demonstrated that vowel and consonant onset time differences were modulated by *Condition*. Inter-speaker alignment of vowel and consonant onsets was the highest in the *Synchrony-live condition* (1). Importantly, in the *Synchrony-live condition* (1), the deviation among values for onset timing difference was very small (Figure 6). These findings prove the boosting effect of a live partner on rhythmic synchronization with the speech signal. For example, in the case of the *Synchrony with recording from live condition* (2), participants were not able to achieve the same degree of synchrony that they achieved in the *Synchrony-live condition* (1). These results indicate that co-presence provided additional expectancy cues beyond the signal-inherent acoustic correlates of speech rhythm that did not differ between the *Synchrony-live condition* (1) and *Synchrony with recording from live condition* (2).

The boosting effect of co-presence on speech synchronization can be explained by the mutual efforts of the interactants, who dynamically modified their rhythmic patterns online and adjusted their speech onsets. In the case of the S*ynchrony from recording live condition* (2), participants had to synchronize their reading with a signal that did not change in real-time, while in the *Synchrony-live condition* (1), the signal of one interlocutor was modulated by the constantly changing signal of their partner. Our analysis of segment onset synchronization showed that, across conditions, when participants were aligning their speech with the recording, the recorded model led, while the experimental participants followed (with a larger onset difference in the condition, in which the recording was made during solo reading). The fact that experimental participants failed to take a leading position in dyadic reading with a recorded model suggests that they find it challenging to use expectancy cues, predict segment onsets, and produce an upcoming segment before it has been initiated by the model. To a large extent, in the conditions with a recorded model, participants relied on shortening an existing lag rather than predicting segment onsets based on expectancy cues and initiating corresponding segments before they actually occurred in the model speech. This result pattern indicates that the optimal use of expectancy cues occurs in a dynamically changing situation (*live* reading), when the cues are modified online, in the course of a reading-aloud task performed by both partners, allowing them to switch between leading and lagging positions interactively. We propose that during live co-presence, participants modify their prosodic timing patterns online and that this mutual effort exerts beneficial effects on task performance. This means convergence is an

638 interactive process, and rhythmic synchronization with a live partner is more precise
639 than unidirectional synchronization, e.g., with a recording of someone's speech.
640 It is interesting to note that the effect found by Cummins (2002, 2009) – that
641 people synch better with recordings made in a synchronous reading condition versus a
642 non-synchronous reading condition – was confirmed in our study only for the narrative
643 text; we did not observe this effect with the poetic text. Meter potentially provides
644 additional expectancy cues for synchronization. Indeed, these cues proved relevant in
645 the most challenging condition (synchronization to a recording made during solo
646 reading) because they allowed participants to predict when the next **stressed** syllable
647 would occur. Thus, meter allows for synchronization to the onset of **stressed** syllables.
648 As attention shifts to the onsets of the **stressed** syllables, convergence of vowel onsets
649 in inter-stress intervals is relaxed. Narrative text does not provide strong expectancy
650 cues for stress onsets, thus reducing synchronization in the *Synchrony with recording*
651 *from non-live condition* for the narrative compared to the poetic text.
652 Our data emphasize the relevance of the co-presence of a live partner who
653 attempts to sync with the other speaker by dynamically changing their speech. We have
654 established that people synchronize better in the *Synchrony-live condition* (1) due to the
655 expectancy cues provided by co-presence. We posit that people achieved better
656 performance when synchronizing with another interlocutor because they were trying to
657 increase cooperation with the model speaker, a point also raised by Bowling et al.
658 (2013). In this sense, the aptitude for joint speech in humans could be linked to the
659 ability for entrainment found in primate communication. Notably, coordination in
660 primate vocal behavior appears to provide a signal of group cohesion and cooperative
661 strength (Ravignani, Bowling, & Fitch, 2014), as also seems to be the case with human
662 beings (Polyanskaya, Samuel, & Ordin, 2019). If we look at the different spheres where
663 joint speech is used – education, religion, sports, among others – we find a clear
664 indication that joint speech is used for group cohesion and cooperative strength.
665 The social and cognitive implications of humans' ability to entrain to each other'
666 behavior (including vocal communication) are enormous, in the sense of both individual
667 and social performance. In other words, there is great potential for exploiting this ability
668 in all spheres of social life and cognition since human beings engage in individual and
669 social entrainment tasks (entrainment of one´s motor or vocal output to the signal
670 emitted by another individual) from birth. Furthermore, the effects of joint speech are
671 not only beneficial for achieving cooperation and cohesion. Recent research has found

that when two people are in a synchronous speech condition, neural networks that are suppressed during individual speech are activated (Jasmin et al., 2017). For example, activation in the auditory sensory cortices was usually enhanced during speech perception but inhibited during speech production. And yet, during chorusing, these sensory cortices were not suppressed. Moreover, activation of neural substrates in the right hemisphere, which do not constitute part of the classic speech production network, was observed. It seems that our brains detect that joint speech is, in fact, a different type of speech, perhaps due to its highly cooperative and interactive nature. The authors suggested that these neural changes in synchronous speech production overshadow the distinction between one's own and partner's speech, such that the partner's speech is perceived as if it were self-produced. This might facilitate extraction and use of expectancy cues from the speech stream and allow participants to take leading positions in dyadic reading tasks rather than relying solely on shortening the existing lag between the segment they hear and the segment they initiate during their own speech production (as the boundary between one's own speech production and the other's speech production is diffused). Perceiving another person to be more like oneself enhances group affiliation, facilitates joint activities, and encourages cooperation. Enhanced cooperation during joint speech production and inevitable convergence of speech patterns (and thus motor patterns, since speech is a motor activity) may be beneficial for cognitive training, in the case of speech pathologies or any other type of condition involving diminished speech and motor abilities, such as Parkinson's disease (Thaut et al., 2001; deDreu et al., 2012).

Conclusions

Our data showed that the co-presence of a live speaker has the strongest modulatory effect on rhythmic patterns and on the degree of synchronization between speakers in a chorus reading task. We assume that this facilitatory effect emerges from dynamic online changes resulting from two simultaneous drives: the drive to synchronize with a partner; and the drive to provide cues that allow this partner to synchronize better with oneself. The interaction of these two complementary drives – present in both partners – increases the complexity of cue exchange and speech modification. The exact mechanistic explanation for these interactions should be sought in the mathematics of complexity theory.

References

Boersma, P., & Weenink, D. (2016). *Praat: Doing Phonetics by Computer* [Computer software]. Version 6.0.19.

Bowling, D. L., Herbst, C. T., & Fitch, W. T. (2013). Social origins of rhythm? Synchrony and temporal regularity in human vocalization. *PLoS One*, *8*(11), e80402. doi: 10.1371/journal.pone.0080402

Clopper, C. G., & Smiljanic, R. (2011). Effects of gender and regional dialect on prosodic patterns in American English. *Journal of Phonetics*, 39(2), 237-245. doi: 10.1016/j.wocn.2011.02.006

Coloma, G. (2016). Una versión alternativa de "El viento norte y el sol" en español. *Revista de Investigación Lingüística*, *18*, 191-212.

Crystal, D. (1969). *Prosodic systems and Intonation in English* (Vol. 1). Cambridge: C.U.P.

Cummins, F., & Port, R. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics*, *26*(2), 145-171. doi: 10.1006/jpho.1998.0070

Cummins, F. (2002). On synchronous speech. *Acoustic Research Letters Online*, *3*(1), 7-11. doi: https://doi.org/10.1121/1.1416672

Cummins, F. (2003). Practice and performance in speech produced synchronously. *Journal of Phonetics*, *31*(2), 139-148. doi: https://doi.org/10.1016/s0095-4470(02)00082-7

Cummins, F. (2007). *The quantitative estimation of asynchrony among concurrent speakers*. Technical Report UCD-CSI-2007-2, School of Computer Science and Informatics, University College Dublin.

Cummins, F. (2009). Rhythm as entrainment: The case of synchronous speech. *Journal of Phonetics*, *37*(1), 16-28. doi: https://doi.org/10.1016/j.wocn.2008.08.003

Cummins, F. (2015). Rhythm and Speech. In M. Redford (Ed.), *The Handbook of Speech Production (*pp. 158-177). New York: Wiley.

deDreu, M. J., van der Wilk, A. S., Poppe, E., Kwakkel, G., and van Wegen, E. E. (2012). Rehabilitation, exercise therapy and music in patients with Parkinson's disease: a meta-analysis of the effects of music-based movement therapy on walking ability, balance, and quality of life. *Parkinsonism Relat. Disord*, *18*, 114–119. doi: 10.1016/S1353-8020(11)70036-0

Dellwo, V. (2006). Rhythm and speech rate: a variation coefficient for deltaC. In P. Karnowski & I. Szigeti (Eds.), *Language and Language Processing* (pp. 231-241). Frankfurt am Main: Peter Lang.

Eerola, T., Luck, G., & Toiviainen, P. (2006, August). An investigation of pre-schoolers' corporeal synchronization with music. In *Proceedings of the 9th International Conference on Music Perception and Cognition* (pp. 472-476). The Society for Music Perception and Cognition and European Society for the Cognitive Sciences of Music Bologna.

Fletcher, J. (2010). The prosody of speech: Timing and rhythm. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (pp. 521-602). John Wiley & Sons.

Fitch, W. (2013). Rhythmic cognition in humans and animals: distinguishing meter and pulse perception. *Frontiers in Systems Neuroscience*, *7*, 68. doi: 10.3389/fnsys.2013.00068

Goldman, J. (2011). EasyAlign: an automatic phonetic alignment tool under Praat. In *Interspeech '11, Proceedings of the 12th Annual Conference of the International Speech Communication Association* (pp. 3233-3236). Florence, Italy: International Speech Communication Association.

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in Laboratory Phonology 7* (pp. 515-546). Berlin: Mouton de Gruyter.

Hannon, E. E., & Johnson, S. P. (2005). Infants use meter to categorize rhythms and melodies: Implications for musical structure learning. *Cognitive Psychology*, *50*(4), 354-377. doi: 10.1016/j.cogpsych.2004.09.003

Jasmin, K., McGettigan, C., Agnew, Z., Lavan, N., Josephs, O., Cummins, F., Scott, S. (2017). Cohesion and Joint Speech: Right Hemisphere Contributions to Synchronized Vocal Production. *Journal of Neuroscience*, *36*(17), 4669–4680. doi: 10.1523/JNEUROSCI.4075-15.2016

Labov, W. (2006). *The social stratification of English in New York city* (3rd ed.). Cambridge University Press.

766 Ladd, D. R., & Cutler, A. (1983). Models and measurements in the study of prosody. In
767 A. Cutler, & D. R. Ladd (Eds.), *Prosody: Models and measurements* (pp. 1-10).
768 Heidelberg: Springer.

769 Large, E., & Snyder, J. (2009). Pulse and meter as neural resonances. *Annals of the New*
770 *York Academy of Sciences, 1169*(1), 46–57. doi: 10.1111/j.1749-
771 6632.2009.04550.x

772 Lehiste, I. (1973). Rhythmic units and syntactic units in production and perception. *The*
773 *Journal of the Acoustical Society of America, 54*(5), 1228-1234. doi:
774 10.1121/1.1914379

775 Lehiste, I. (1976). Suprasegmental features of speech. In N. Lass (Ed.), *Contemporary*
776 *issues in Experimental Phonetics: Perspectives in Neurolinguistics and*
777 *Psycholinguistics* (pp. 225-242*).* New York: Academic Press.

778 Low E., Grabe E., & Nolan, F. (2000). Quantitative Characterizations of Speech
779 Rhythm: Syllable-Timing in Singapore English. *Language and Speech, 43*(4),
780 377-401. doi: 10.1177/00238309000430040301.

781 London, J. (2004). *Hearing in Time: Psychological Aspects of Musical Meter*. Oxford
782 University Press. Oxford, U.K.

783 Mairano, P. & Romano, A. (2010). *Correlatore*. [Computer software]. Retrieved May
784 22, 2016 from http://www.lfsag. unito.it/correlatore/index.html

785 Marcus, S. M. (1981). Acoustic determinants of perceptual center (P-center) location.
786 *Perception & Psychophysics, 30*(3), 247-256. doi: 10.3758/BF03214280

787 Merchant, H., Grahn, J., Trainor, L., Rohrmeier, M., & Fitch, W. T. (2015). Finding the
788 beat: a neural perspective across humans and non-human primates. *Phil. Trans.*
789 *R. Soc. B, 370*(1664), 20140093. doi: 10.1098/rstb.2014.0093

790 Moore, R. (2012). Finding Rhythm in Speech: A response to Cummins. *Empirical*
791 *Musicology Review*, *7*(1-2), 36-44. doi: 10.18061/1811/52977

792 Morton, J., Marcus, S., & Frankish, C. (1976). Perceptual centers (P-centers).
793 *Psychological Review*, *83(*5), 405–408. doi: 10.1037/0033-295x.83.5.405

794 Nolan, F., & Jeon, H.S. (2014). Speech rhythm: a metaphor? *Phil. Trans. R. Soc. B,*
795 *369*(1658), 20130396. doi: 10.1098/rstb.2013.0396

796 Nooteboom, S. (1997). The prosody of speech: melody and rhythm. In W.J. Hardcastle
797 & J. Laver (Eds.), *The Handbook of Phonetic Sciences* (pp. 640-673). Oxford:
798 Blackwell.

799 Pardo, J. S. (2010). Expressing oneself in conversational interaction. In E. Morsella
800 (Ed.), Expressing oneself/expressing one's self: Communication, cognition,
801 language, and identity (pp.183-196). Hove, England: Psychology Press/Taylor &
802 Francis. doi: 10.4324/9780203848708

803 Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence
804 across multiple measures and model talkers. *Attention, Perception, &*
805 *Psychophysics*, *79*(2), 637-659. doi: 10.3758/s13414-016-1226-0

806 Parra, V. (1988). La Muerte. In V. Parra, *Décimas. Autobiografía en verso*. Buenos
807 Aires: Editorial Sudamericana.

808 Pitt, M. A., & Samuel, A. G. (1990). The use of rhythm in attending to speech. *Journal*
809 *of Experimental Psychology: Human Perception and Performance, 16*(3), 564-
810 573. doi: 10.1037/0096-1523.16.3.564

811 Polyanskaya, L., Samuel, A.G., & Ordin, M. (2019). Speech rhythm convergence as a
812 social coalition signal. *Evolutionary Psychology 17(3)*, 1-11. Doi:
813 10.1177/1474704919879335.

814 Quené, H. & Port, R. F. (2005) Effects of timing regularity and metrical expectancy on
815 spoken word perception. *Phonetica, 62*(1), 1-13. doi: 10.1159/000087222

816 Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the
817 speech signal. *Cognition*, *73*(3), 265-292. doi: 10.1016/S0010-0277(99)00058-X

818 Ravignani, A., Bowling, D. & Fitch, T. (2014). Chorusing, synchrony and the
819 evolutionary functions of rhythm. *Frontiers in Psychology*, *5*, 1118. doi:
820 10.3389/fpsyg.2014.01118

821 Repp, B. H. (2005). Sensorimotor synchronization: a review of the tapping literature.
822 *Psychonomic Bulletin & Review*, *12*(6), 969-992. doi: 10.3758/BF03206433

823 Scott, S. (1998). The point of P-centres. *Psychological Research 61*, 4–11.

824 Smith, R., Rathcke, T., Cummins, F., Overy, K. & Scott, S. (2014). Communicative
825 rhythms in brain and behaviour. *Phil. Trans. R. Soc. B, 369(1658),* 20130389.
826 doi: 10.1098/rstb.2013.0389

827 Szczepek Reed, B. (2010). Speech rhythm across turn transitions in cross-cultural talk-
828 in-interaction. *Journal of Pragmatics*, *42*(4), 1037-1059. doi:
829 10.1016/j.pragma.2009.09.002

830 Thaut, M. H., McIntosh, K. W., McIntosh, G. C., & Hoemberg, V. (2001). Auditory
831 rhythmicity enhances movement and speech motor control in patients with
832 Parkinson's disease. *Functional Neurology*, *16*(2), 163-172.

833 Turk, A., & Shattuck-Hufnagel, S. (2013). What is speech rhythm? A commentary on
834 Arvaniti and Rodriquez, Krivokapić, and Goswami and Leong. *Laboratory*
835 *Phonology*, *4*(1), 93-118. doi:10.1515/lp-2013-0005

836 Vos, P. G., Mates, J., & van Kruysbergen, N. W. (1995). The perceptual centre of a
837 stimulus as the cue for synchronization to a metronome: Evidence from
838 asynchronies. *The Quarterly Journal of Experimental Psychology Section A*,
839 *48*(4), 1024-1040. doi: /10.1080/14640749508401427

840 Vuust, P., & Witek, M. A. (2014). Rhythmic complexity and predictive coding: a novel
841 approach to modeling rhythm and meter perception in music. *Frontiers in*
842 *Psychology*, *5*, 1111. doi: 10.3389/fpsyg.2014.01111

843 Wells, J. (1996). *SAMPA for Spanish*. London: Division of Psychology and Language
844 Sciences, University College London.

845 White, L., and Mattys, S. (2007). Calibrating rhythm: First language and second
846 language studies. *Journal of Phonetics 35(4)*, 501–522.

847

FIGURE CAPTIONS

Figure 1. Vowel % across conditions and texts. Here and in subsequent figures, error bars show 2SE around the mean. Asterisks mark significant contrasts, * - p<.05, ** - p<.005, *** - p<.001.

Figure 2. VarcoV across conditions and texts

Figure 3. VnPVI across conditions and texts

Figure 4. VarcoC across conditions and texts

Figure 5. CnPVI across conditions and texts

Figure 6. Mean differences for onsets across synchrony conditions and texts

858 Table 1 *Experimental materials*

Text number	Long name	Short name	Description	Objective
Text 0	Warm-up text	Warm-up text	16 sentences	Warm-up: 16 sentences were created to habituate experimental participants to the synchrony-reading task. These sentences also introduced specific vocabulary found in Texts 1 and 2, so as to prevent any semantic or phonological interference during the experiment.
Text 1	Narrative text with weak meter (Text 1)	Text with weak meter (Text 1)	A Spanish version of *The North Wind and the Sun* (Coloma, 2016)	Experimental: A phonetically balanced text, with weak meter, which includes every segmental sound in Spanish.
Text 2	Poetic text with strong meter (Text 2)	Text with strong meter (Text 2)	A poem written in octosyllabic verse, consisting of two stanzas (Parra, 1988).	Experimental: A text with strong meter.

859
860

Table 2 *Description of Conditions*

Condition	Name	Description	Synchronic condition
(1)	Synchrony-live	Experimental participants were recorded reading aloud and in synchrony with a female native speaker of Spanish (henceforth, the model speaker)	Yes
(2)	Synchrony with a recording from live condition	Experimental participants were recorded reading aloud and in synchrony with the recording of the model speaker obtained from the *Synchrony-live condition* (1)	Yes
(3)	Synchrony with a recording from non-live condition	Experimental participants were recorded reading aloud and in synchrony with a recording of the model speaker obtained from a non-synchronous condition	Yes
(4)	Solo recording condition	Experimental participants were individually recorded reading aloud	No

Table 3 *Vocalic and consonantal interval measurements (none of the Condition*Text interactions were significant, and thus interactions are not reported).*

Measurement	**Factor**	**Intra-subject effects**	**Pairwise Contrasts ([2])**
V%	Conditions (4)	F(2.375, 64,116) = 14,556; $p < 0.0005$; $\eta_p^2 = 0.350$ (**[1]**)	Conditions 1 & 2 (p = .027); Conditions 1 & 3 (p < 0.0005); Conditions 1 & 4 (p = 0.001); Conditions 2 & 3 (p = .001)
	Text (2)	F(1, 27) = 281.132; $p<0.0005$; $\eta_p^2= 0.912$	Text 1 and Text 2 (p <0.0005).
VarcoV	Conditions (4)	F(3, 75) = 18.076; $p < 0.0005$; $\eta_p^2 = 0.420$	Conditions 1 & 3 (p<0.0005); Conditions 2 & 3 (p<0.0005); Conditions 3 & 4 (p=.002)
	Text (2)	F(1, 25) = 84.173; $p < 0.0005$; $\eta_p^2 = 0.771$	Text 1 and Text 2 (p < 0.0005).
VnPVI	Conditions (4)	F (3, 66) = 4.355; $p = 0.007$; $\eta_p^2 = 0.165$	Conditions 3 & 4 (p = 0.004)
	Text (2)	F(1, 22) = 176.295; $p < 0.0005$; $\eta_p^2 = 0.889$	Text 1 and Text 2 (p < 0.0005).
VarcoC	Conditions (4)	No significant differences	No significant differences
	Text (2)	F(1, 26) = 23.102;	Text 1 and Text 2 (p < 0.0005).

		$p < 0.0005$; $\eta_p^2 = 0.470$	
CnPVI	Conditions (4)	No significant differences	No significant differences
	Text (2)	$F(1, 26) = .792$; $p = 0.025$; $\eta_p^2 = 0.030$.	Text 1 and Text 2 ($p = 0.025$)

Notes. [1]Greenhouse-Geisser correction was applied to all repeated-measures ANOVAs in this study if the assumption of sphericity was violated; [2]Bonferroni correction was applied to all pairwise contrasts in the study.

Table 4 *Measurements for Speech Rhythm Synchrony*

Measurement	**Factor**	**Intra-subject effects**	**Pairwise contrasts**
Onset synchrony	Conditions (3)	$F(1.415, 28.302) = 21.933$; $p < 0.0005$; $\eta_p^2 = 0.523$	Conditions 1 & 2 ($p < 0.0005$); Conditions 1 & 3 ($p < 0.0005$); Conditions 2 and 3 ($p = .025$)
	Text (2)	No significant differences	No significant differences